Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue 4, April-2017

Web Mining: A Survey on Web Usage Mining

Kuldipsinh M. Chauhan[1], Prof. Virendra Barot[2]
Department of Information Technology (System and Network Security),
Sardar Vallabhbhai Institute of Technology, Vasad, Gujarat, India-388306

Abstract – *Web mining is the application of data mining techniques to extract knowledge from web that including web documents, hyperlinks between the documents, Usage logs of web sites, etc. People use the web for vast approach of applications like online shopping, online bill payment, entertainment, social network, education, marketing, data sharing, data storage, etc. Web Usage Mining is very popular technique to extract the knowledge from web logs. It uses three different phases Pre-processing, Pattern Discovery and Pattern Analysis. This Paper describes the Web Usage Mining and Research issues of Web mining Applications in detail.*

Keywords *– Web Mining, Web Usage Mining (WUM), Web Structure Mining, Web Content Mining, Pre-Processing, Pattern Discovery, Pattern Analysis.*

I. INTRODUCTION

Web mining lies in between and copes with semi structured data or unstructured data. Web mining calls for creative use of data mining and its distinctive approaches. Mining the web data is such of the most challenging tasks for the data mining and data management scholars because there are huge heterogeneous, less structured data available on the web and we can easily get overwhelmed with data [1].

Web mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data. It is normally considered that either the hyperlink structure of the website, or its contents or web log data that are used in the mining process [1]. The proper web mining applications are website raw material, web search, search engines, information retrieval, network management, e-commerce, business and simulated 3D environment, web supermarket stores and web communities [2].

II. APPLICATION OF WEB MINING

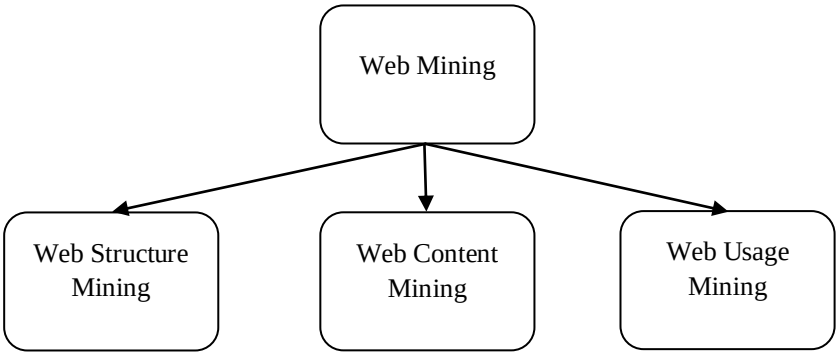

Figure 1. Application of Web mining

2.1 Web Structure Mining

Web Structure Mining applies to identifying the interesting pattern and information from the underlying process of the web. The structure of the web is realized through the links which suggest a page or resource from the referrer in which it resides. The main goal of web structure mining is to generate structural summary about the website and web page. Web Structure Mining discovering the structure of a web document [3].

Web Structure Mining finds out the popular and similar websites using different Algorithms. It is based on website Hyperlink and Document Structure.

Various techniques utilized for Web Structure Mining are PageRank, HITS, weighted PageRank & Topic Sensitive, ECLAT algorithm, PageRank [3].

2.2 Web Content Mining

Web Content Mining defines as informal search of information resource available online. Resources contains the text, audio, video, metadata and hyperlinks some of the semi-structured, such as HTML documents, or a more structured data like the database tables [4].

Web Content Mining is the Process to discover useful information from the content of the web page. The main goal of Web Content Mining is to get the structured information from unstructured website Contents [5].

Various techniques utilized for Web Content Mining are Spam Mail Filtering and Bayes' Method [6].

2.3 Web Usage Mining

Users uses the internet for different purposes. Thus, WUM establishes to see the what kind of activities is performed by the users are doing and in which object or resource they are interested [3].

Web Usage Mining is used to discover usage patterns from web data. Data is usually collected from user's interaction with the web like, user queries, registration data or data generated by the web server, proxy server, cache, cookies, interestingness and usual surfing habits of the users. There is three phases of WUM Pre-Processing, Pattern Discovery and last phase is Pattern Analysis that are discuss further in detail [3].

WUM has several tools to analyze the behavior of the user. They are KOINOTITES, web SIFT, web usage miner, INSITE, speed tracer, Archcollect, i-Miner, AWUSA, i-JADE web miner, Web Quilt, STRATDYN, SEWeP, webTool, MiDAS, web mate, WebLog Miner, DB2Intelligent miner of Data, Poly Analyst version 6.0, Clemetine, WEBMINER, WEBVIZ [2].

III. WEB LOG FILES

3.1 Web Server Log

Web page access history maintained as a log file. Web servers are most common and costly data source. Log files collect the large volume of information like date, time, IP address, Client information, etc [7]. This data can be a single text file or in different files as given below [2]:-

3.1.1. Access Log

Access log is used to discover the information about the user and it has many numbers of attributes. It will record each click event, hits, and access of the user.

3.1.2. Agent Log

Agent logs are capture the information about online user behavior, browser used by user, browser version and Operating System used by user. It is a standard log file compare to access log.

3.1.3. Error Log

When user click on a particular link and the browser does not display the particular page or website then the user receives error 404 not found.

3.1.4. Referrer Log

Referrer log is used to store the information of the URLs of web pages on other sites that link to web pages.

3.2 Proxy Server Log

Intermediate level of catching information that lies between the client browser and web server. Proxy Server logs reduce the loading time of web page and reduce the network traffic at the server and client side. The proxy server log is used as a data source for browsing behavior characterization of a group of unauthorized users sharing a common proxy server [7].

3.3 Browser Log

Browsing history collected through JavaScript and Java Applet. To implement data collection from client side, user cooperation is needed. It is in form of Cookies and Cache [7].

IV. PROCESS OF WEB USAGE MINING

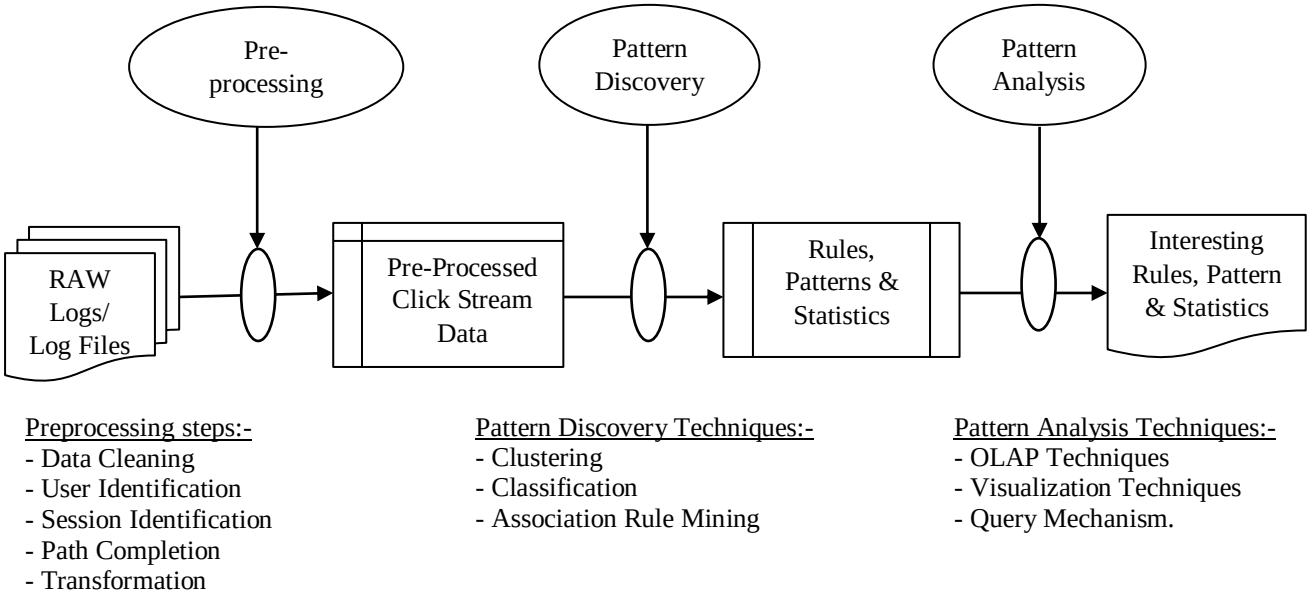

Figure 2. Process of Web Usage Mining

4.1 Pre-Processing

Usually Web data is noisy, inconsistent and irrelevant by nature therefore, this data is not suitable for mining and analysis, so Pre-Processing is needed and it contains the information about each resource access like image, audio, video, links, etc. To identify relevant data from those bulky data and organizing it in terms of users as well as sessions is what pre-processing performs [3]. Steps of Pre-processing [8]:

4.1.1. Data Cleaning

Usually web log contains inappropriate entries, which consequently sink data quality, and increases file size. The use of data cleaning procedure is to remove all the unwanted data used in data analysis and mining. Data Cleaning is also increase mining efficiency [7]. There are various techniques to clean the data.

4.1.2. User Identification

It identifies the single unique user from a set of the user.

4.1.3. Session Identification

A Session refers to sequence of page view by a particular user during particular visit. User Session identification finds out which page belongs to which session.

4.1.4. Path Completion

It fills the missing page references in a session that are important for mining purpose.

4.1.5. Transformation

Format the sessions according to the type of data mining to be accomplished.

4.2 Pattern Discovery

Pattern Discovery belongs to methods and algorithms developed from several fields such as statistics, data mining, machine learning, and Pattern recognition [3].

These methods represent the approaches that often appear in the data mining literature such as discovery of Association rules, Sequential Patterns, Clustering, Classification etc [9].

4.2.1. Association Rule Mining

Association Rule generation can be used to involve pages that are most often referenced together in a single server session. In the framework of WUM, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via links. Different techniques are association rule mining is Apriori and F-P Growth Algorithm.

4.2.2. Sequential Patterns

It is same as the Association Rule mining with the difference of time ordering. It finds the sequence of pages are accessed after the other set, but in time sequence. Application is to predict the future visitor of the website. Algorithms of Sequential Patterns are Hashing, Pruning, WAP tree, etc.

4.2.3. Clustering

Clustering is a technique to group together a set of items having similar characteristics. In the web usage domain, there are two kinds of interesting clusters to be discovered: Usage Cluster and Page Cluster. Through Clustering, we can find browsing similar patterns. Several algorithms for clustering are K-Means, DBSCAN, Fuzzy C Means, etc.

4.2.4. Classification

Classification is the job of mapping a data item into one of the number of predefined classes or labels. In the Web Usage mining, one is interested in generating a user profile belonging to a particular class or category. Classification uses Supervised-learning algorithms in which class label is known in advance. Various Classification algorithms are naïve Bayesian, KNN, Decision Tree, etc.

4.3 Pattern Analysis

Pattern Analysis is the last step of WUM to discover the interesting rules or filter out the uninteresting rules. The most common form of pattern analysis is the query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data [9].

V. RESEARCH ISSUES

5.1 Research Issues in Web Structure Mining

- Reducing irrelevant search results.
- Indexing information on the web [2].

5.2 Research Issues in Web Content Mining

- The web contains large amount of data, each website accepts similar information in a different way, So Similar data discovery is an important problem for web application.
- Opinion extraction from online sources.
- Detecting noise and automatically segmenting the web pages that aims to website could not have any intermediate advertisement or Navigation links it directly redirect to main contents [2].

5.3 Research Issues in Web Usage Mining

- Session Identification
- CGI data
- Catching
- Dynamic pages
- Robot Detection and Filtering
- Transaction Identification [2].

VI. CONCLUSION

Web Usage Mining is useful for analysis of users and their behavior towards the website and its contents. As we discuss in the paper there is various techniques, tools, algorithms used in pattern discovery and discuss the issues of the web mining applications. Web Usage Mining can be used in Personalization, System Improvements, Website Modifications, Business Intelligence Usage Characterization, etc.

VII. ACKNOWLEDGEMENT

Second author of the paper offers the sincere gratitude to the Head of the Department and faculty members in Department of Information Technology of Sardar Vallabhbhai Patel Institute of Technology for the constant encouragement and unparalleled support provided throughout the period of this research work.

REFERENCES

[1] Brijendra Singh, Hemant Kumar Singh, "Web Data Mining Research: A Survey," in IEEE-2010.

[2] Dr.S. Vijiyarani and Ms. E. Suganya, "Research Issues In Web Mining," in International Journal of Computer-Aided Technologies (IJCAx) Vol.2, No.3, July 2015.

[3] Parth Suthar, Prof. Bhavesh Oza, "A Survey of Web Usage Mining Techniques," International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6 (6), 2015.

[4] Hui Yu, Zhongmin Lu, "Analysis of Web Usage Mining"

[5] Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining" International Conference on Artificial Intelligence and Embedded Systems, 2012.

[6] Arne Pottharst, "Web Mining Examples and Applications" TU Darmstadt, Germany 2008.

[7] G. Neelima, Dr. Sireesha Rodda, "Predicting user behavior through Sessions using the Web log mining," in March 2016.

[8] Dr. Sanjay Kumar Dwivedi, Bhupesh Rawat, "A Review Paper on Data Preprocessing: A Critical Phase in Web Usage Mining Process" in IEEE 2015.

[9] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" SIGKDD(ACM Simple Interest Group KDD) Explorations 2000.

[10] https://en.wikipedia.org/wiki/Web_mining

[11] http://www.web-datamining.net/usage/

[12] https://en.wikipedia.org/wiki/Data_pre-processing

[13] https://www.researchgate.net/publication/236673264_Web_Usage_Mining_Process_and_Techniques