DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5763

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A SYSTEMATIC REVIEW ON DATA PREPROCESSING AND PATTERN DISCOVERY OF WEB USAGE MINING

Neeru
PhD Scholar
University Institute of Engineering & Technology
Maharshi Dayanand University
Rohtak- 124001, India

Rainu Nandal
Assistant Professor
University Institute of Engineering & Technology
Maharshi Dayanand University
Rohtak- 124001, India

Abstract: To find valuable knowledge from web data is known as web mining. The growth of World Wide Web exceeded all expectations with the development of internet technology. Rapid growth of World Wide Web has affected a lot of both visitors and web site owners. Retrieving different information in different format has become a very difficult task. To solve this problem, one positive approach is web usage mining (WUM).Web mining that extracts patterns from user weblogs is known as web usage mining, that is an implementation part of data mining. The goal of web usage mining is to understand the behaviour of web site users by processing the data mining of web access data. The success of web usage mining depends upon efficient knowledge extracted from large amount of raw log data. Knowledge obtained from web usage mining does the task of finding the hidden important information about user behaviour, and facilitate more effective browsing, to enhance web design, its page surfing pattern and other valuable information which is used for various purposes. In this paper, we provide detailed review of work done for different phases of web usage mining.

Keywords: Web Usage Mining; Data Mining; Data Pre-processing; Pattern Discovery

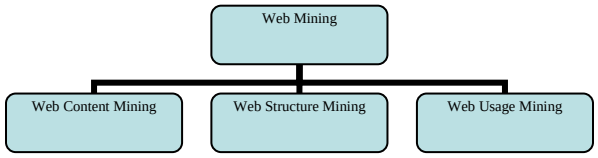
I. INTRODUCTION

The World Wide Web is interactive and popular medium for searching information today as millions of users are interacting daily with websites, so easy to use web systems is becoming a necessity for today. Web is a platform to provide freely available information to different users effectively and efficiently. So, a large number of researchers are putting their effort in the field of data mining techniques on the web [9]. Web mining can be broadly defined as the extract of useful information from the World Wide Web.

Web mining is the mining of data on www from which we derive meaningful and useful knowledge from contents, hyperlinks present in web pages and user usage logs. As web mining, is used to discover useful and interesting patterns from the web. It assists web surfers in browsing internet, so much attentions is needed in modeling web user's browsing patterns and making recommendation. Main objective of web mining is to develop more intelligent tools to help users in finding, extracting, evaluating and filtering valuable information and resources.

II. WEB USAGE MINING

Web mining is the technique of data mining to extract and discover interesting knowledge from web data. Web mining is of three types: web content mining-mining the data on the web, web structure mining-mining the web structure data and web usage mining-mining the web log data [2].



Web Content Mining
It deals with web data content e.g. HTML/XML, digital libraries and documents etc. It discovers useful knowledge or information from web page contents i.e. text, image, audio and video[15]

Web structure Mining
It deals with analyzing, structuring hyperlinks at inter document level of web pages or websites to generate structure summary. On this summary, Various techniques are applied to recreate, redesign the websites. It finally improves the structure quality of website.

Web usage mining
It deals with weblog data, to extract knowledge about user behaviour, while they are interacting with websites. This extracted knowledge helps in efficient reorganization of websites, improvement in links and navigation, better recommendation and personalization etc. It attracts more users towards website and generates more revenue out of it.
As web content and web structure mining mines the real or primary data on the web, whereas web usage mining mines the secondary data derived from various log files.

Web usage mining process consists of following phases:

1. Data collection
2. Data preprocessing
3. Pattern discovery

4. Pattern analysis

Data Collection
Data is collected from various resources such as web server, client machine and proxy servers.

Data Preprocessing
After collection of data, it becomes necessary to make data consistent, relevant and integrated. It includes various tasks: Data cleaning, session identification, user identification and path completion.

Pattern Discovery
It is the key field of web usage mining. There are various techniques to discover patterns from web usage data to extract knowledge i.e. statistical analysis, association rules, clustering, classification, sequential pattern analysis and dependency modeling.

Statistical analysis
It performs statistical operation like median, frequency etc on session files.

Association rules
It finds relationship amongst various set of items have basic fundamental of support and confidence of users.

Clustering
It is a grouping of items having similar properties. Page cluster and usage clusters are two kinds of clusters. Cluster of users is to setup group of users which show the similarity in browsing patterns.

Classification
It makes the various data item sets into several predefined classes. Classes are defined by choice of features and features extracted.

Sequential pattern
It finds inter session patterns and web marketers predicts future visit patterns.

Dependency modeling,
It is a model showing important dependencies between various stages is developed.

Pattern Analysis
After discovery of pattern, this phase filters unwanted information and field [13,14]

III. LITERATURE REVIEW

A. Data Collection and Data Preprocessing

The data source of web usage mining is web log files to realize user's browsing pattern. According to Chintan R. Varnagar et al [11], L.Chen, et al [1], K. Sudheer Reddy et al [12] and Jaideep Srivastava et al [14], data is obtained from following sources-

Client side log file:-
At this level, data is gathered by recording of activities, events which happen on premises of client machines, such as mouse clicks, context selection and mouse wheel relation ,scrolling with a particular page etc. it requires users participation integrating java applet or by writing java scripts.

Browser side log file
In this method, data collection is done by modifying the source code of existing browser. Much versatile data is provided by this method as it considers the behaviour of single user on multiple sites.

Server logs:-
Server side log files are included in this, as they collect information about the user such as IP address, access time, linked visits etc. This log files can take two formats-common logs format or extended log format.

Proxy logs:-
A proxy serve acts as an agent to grant user requests while it is looking for resources from other severs .The access log files includes HTML files, graphics associated files with web pages.

Data Preprocessing
When data comes from various sources from combined and individual log file, it may this log data is unformatted, noisy and impure. So this data undergoes a complex process to make the data consistent, integrated, relevant and of superior quality. Superior quality of data gives reliable information.

According to Chintan R. Varnagar et al [11], K. Sudheer Reddy et al [12], Monika Dhandi et al [13] and P. Sukumar et al[15], steps for preprocessing are data cleaning, user identification, session identification and path completion

Data cleaning
It is the process of removing unnecessary or irrelevant field from raw log data. Web log files have a lot of attributes, from them unnecessary fields are removed by selecting necessary attributes.

User identification
By considering the IP address, it counts the total no. of users. Problem comes when a lot of users are accessing the same machine (IP address). Its solution is to gather information about operating system, browser and time series.

Session identification
It finds different user sessions from web access log file. User session is defined by click stream i.e. a set of user clicks on web server. Cookies and URL rewriting is used for session identification.

Path completion
It is a method that handles some important user accesses that are not being recorded in access log data.

B. Pattern Discovery

According to L.Chen, et al [1], three parameters i.e. visiting path, browsing frequency and relative length of access time help in clustering the client according to their interest. This information is valuable for business firms to provide personalized service to improve customer satisfaction, retain customer loyalty, and build competitive advantage. It also helps to sell items across category and convert visitor into purchaser.

According to A. Bhargav et al [2], three factors i.e. country based, site entry based and access time are used for user classification. It improves administration and does personalization of the websites. By this classification firm's profit increases.

According to S. G. Langhnoja et al [4], association rule mining technique is used on clustered data i.e. clustering will be applied first and then association rule technique is applied. It provides frequent accessed set of link and results accuracy is improved. Effective usage pattern is discovered.

According to Abhaysingh Saste et al [6], demographic information play great role in personalization of systems. Web data is used for the prediction of demographic attributes which is used for designing of business website and advertisement.

According to Suharjito et al [5], visitor activity is analyzed by naïve Bayesian classification and K-nearest neighbor technique. K-nearest neighbor shows good results in comparison of accuracy. Business firms can use the results for decision making.

According to Aditya, S. P et al [8], apriori algorithm uses more memory and time to generate frequent pattern. Improved Apriori Algorithm takes less time to generate frequent patterns.

According to Nandita Aggarwal et al [7], user behavior is predicted by considering the factors like page visited, time spent on pages, operating system used and browser used on the proposed algorithm.

Table I. Pattern Discovery

Sr. No.	*Title*	*Authors*	*Year*	*Technique*	*Results*
1	User-Based Approach For Finding Various Results In Web Usage Mining	Nandita Agrawal,.Anand Jawdekar	2016	Algorithm based on sessions	Page visited and time spent on page is calculated to improve the design
2	Discovering user's interest at E-commerce site using clickstream data	L. Chen and Q. Su	2013	Leader clustering algorithm	Find cluster users of same interest and improve cross selling ability of a website
3	Pattern discovery and users classification through web usage mining	A. Bhargav and M. Bhargav	2014	Classification	Firm's net profit increases. Use three factors: country based, site entry based and access time
4	An approach for frequent access pattern identification in web usage mining	M. M. Sharma and A. Bala	2014	classification	Improve administration and personalization hence increase profit of website
5	Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern,"	S. G. Langhnoja, M. P. Barot, and D. B. Mehta	2013	DBSCAN	Combining clustering and association rule to get effective usage pattern
6	Implementation of Classification Technique in Web Usage Mining of Banking Company	Suharjito, Diana and Herianto	2016	K-nearest neighbor	Gets good results in comparison to classification
7	Predicting Demographic Attributes from Web Usage: Purpose and Methodologies	Abhaysingh Saste, Mangesh Bedekar and Pranali Kosamkar	2017	Classification Clustering Association rule	Demographic information help in designing business strategies and advertisement
8	Effective Algorithm for Frequent Pattern Mining	Aditya, S. P., Hemanth, M., Lakshmikanth, C. K. and Suneetha, K. R.	2017	Improved Apriori	Find interested pattern in less time

IV. CONCLUSION

This paper has provided the review of data preprocessing and pattern discovery of web usage mining. Web usage mining provides the hidden knowledge present in various log files. A proper extraction of knowledge from web log data increase the profit of business firm by redesigning the website according to user's interest.

V. REFERENCES

[1] L. Chen and Q. Su, "Discovering user's interest at E-commerce site using clickstream data," in IEEE 10th International conference on Service systems and service management (ICSSSM), 2013, pp. 124–129

[2] A. Bhargav and M. Bhargav, "Pattern discovery and users classification through web usage mining," in IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 632–636.

[3] M. M. Sharma and A. Bala, "An approach for frequent access pattern identification in web usage mining," in IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 730–735.

[4] S. G. Langhnoja, M. P. Barot, and D. B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for

Pattern," Int. J.Data Min. Tech. Appl., vol. 2, no. 1, pp. 141–150, 2013.

[5] Suharjito, Diana and Herianto "Implementation of Classification Technique in Web Usage Mining of Banking Company" International Seminar on Intelligent Technology and Its Application, pp. 211-218,2016.

[6] Abhaysingh Saste, Mangesh Bedekar and Pranali Kosamkar, "Predicting Demographic Attributes from Web Usage: Purpose and Methodologies" International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017) pp.381-386,2017

[7] Nandita Agrawal,.Anand Jawdekar, "User-Based approach For Finding Various Results In Web Usage Mining" Symposium on Colossal Data Analysis and Networking (CDAN), 2016

[8] Aditya, S. P., Hemanth, M., Lakshmikanth, C. K. and Suneetha, K. R," Effective Algorithm for Frequent Pattern Mining" IEEE International Conference on IoT and Application (ICIOT) - Nagapattinam, Ind pp.1-5,2017

[9] Nina Shahnaz Parvin, Rahman Mahmudur, Bhuiyan, Khairul Islam, "Pattern Discovery of Web Usage Mining" International Conference on Computer Technology and Development - Kota Kinabalu,pp. 499-503,IEEE2009

[10] Ruili Geng, Member and Jeff Tian," Improving Web Navigation Usability by Comparing Actual and Anticipated Usage" IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 45, NO. 1, 2015

[11] Chintan R. Varnagar; Nirali N. Madhak; Trupti M. Kodinariya; Jayesh N. Rathod," Web usage mining: A review on process, methods and techniques", IEEE International Conference on Information Communication and Embedded Systems (ICICES),pp,40-46,2013

[12] K. Sudheer Reddy; M. Kantha Reddy; V. Sitaramulu," An effective data preprocessing method for Web Usage Mining", IEEE-International Conference on Information Communication and Embedded Systems (ICICES),pp.7-10,2013

[13] Monika Dhandi; Rajesh Kumar Chakrawarti," A comprehensive study of web usage mining" IEEE-Symposium on Colossal Data Analysis and Networking (CDAN),pp.1-5,2016

[14] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan," Web usage mining: discovery and applications of usage patterns from Web data", ACM SIGKDD Explorations Newsletter: Volume 1 Issue 2, 2000

[15] P. Sukumar; L. Robert; S. Yuvaraj," Review on modern Data Preprocessing techniques in Web usage mining (WUM)" IEEE-International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS) pp.64-69, 2016