ISSN:0975-9646

A Survey of Web Usage Mining Techniques

Parth Suthar#1, Prof. Bhavesh Oza*2

#M.E., Computer Science and Engineering, L. D. College of Engineering
Ahmedabad, Gujarat, India

*Assistant Professor, Computer Engineering Department, L. D. College of Engineering
Ahmedabad, Gujarat, India

***Abstract*— Web is a very wide and well reached phenomenon. Its enormous popularity stems from the fact that it provides an enormous wealth of information on almost every conceivable subject. People use Web for vast amount of applications like online shopping, online bill payment, entertainment, social networks, education, marketing, data sharing, data storage etc. Result to these activities, web is flooded with large amount of data in the form of the access logs at web servers on daily basis. Leveraging this data and harnessing hidden information from it proves to be a very useful analysis task to improvise the services provided over the Web to better serve the users and also for the longer sustainability of the web itself. This analysis is called web usage mining which is a part of a broader concept called web mining subsequently is part of data mining. In this paper, we describe various techniques, classified based on their nature, that have been developed to find useful information from the Web.**

***Keywords*— Web usage mining, web mining techniques, web usage mining techniques, frequent pattern mining, clustering, classification**

I. INTRODUCTION

Data mining is a process of identifying useful patterns from large amount of data. Data could be stored in databases, data warehouses, data marts or any other data repositories. [15] All data stored in these stores is raw and represents no useful information. Intelligent techniques are applied on this data to process it and to find useful knowledge & pattern from it. Data mining is a discipline representing all those techniques collectively. The techniques used to mine the data are [15] Statistical analysis, Association rules generation, Classification and Clustering.

Data mining is successfully applied in various fields like science, health, marketing, finance etc. One of the applications of data mining is processing of large web repositories to identify hidden patterns which is called the web mining. Web mining, its taxonomy and techniques are explained in further sections. Rest of the paper is organized as follows: Section II contains Web Mining and its taxonomy. Section III contains Web Usage Mining process and techniques and finally the paper is concluded in section IV.

II. WEB MINING

The World Wide Web is a large collection of information in various forms such as text, images, links, videos etc. Most of the data available on the WWW is unstructured. Finding useful patterns from this data is what Web mining all about. Web mining is a sub part of Data mining in which various mining techniques are applied on the data generated as well as residing on the web to find out previously unknown interesting and useful patterns. Applications of web mining are market segmentation, business intelligence, measuring returns of online campaigns, E-commerce etc. [13] The taxonomy of web mining is as shown in the figure 1. Web mining is classified in to 3 broad areas

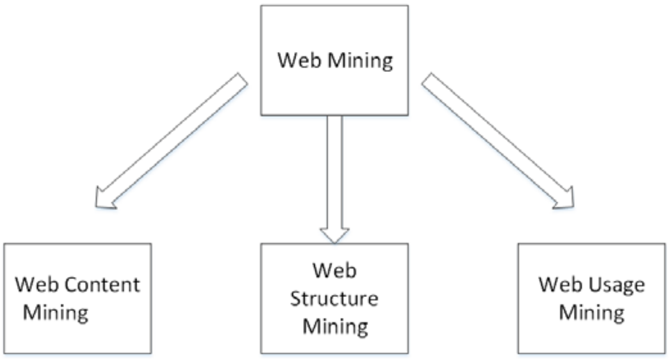

Figure 1 Web Mining

Web structure mining

Web structure mining refers to identifying the interesting pattern and information from the underlying structure of the web. The structure of web is realized through the links which points to a page or a resource from the referrer in which it resides. By having such links on every pages on the web, the web can be viewed as a directed graph having nodes as pages and resources and links as the edges among them. Link mining had produced some agitation on some of the traditional data mining tasks. [8] Various techniques utilized for web structure mining are PageRank, HITS, weighted PageRank & Topic Sensitive PageRank. [7]

Web content mining

Web contains large amount of texts, images, videos, and links. All these data keep generating on daily basis. Web content mining refers to identifying useful information from the content of the Web. Content can reside in the database, web server or in an html page. It could be structured, semi-structured or unstructured. Every day new web apps are being deployed hence web is flooded with newer and previously unknown data every day. Various techniques are employed to mine this data to find useful information based on whether the data is structured or unstructured. For unstructured data, the techniques are Information Extraction, topic tracking, summarization, categorization, clustering & Information visualization. [9] For structured data, the techniques are wrapper generation, web crawling and page content mining. [9] Also for semi structured data, the mining techniques implemented are Object Exchange Model, top down extraction & web data extraction language. [9]

Web usage mining

Every user uses the Web differently for his/her different purposes. Thus it is required to find from the web that what kind of activities the users are doing and in which object or resource they are interested. This is done in the web usage mining in which various mining techniques are used to mine the data generated by the Web servers, proxy servers, caches and cookies to find out the navigation pattern, interestingness and usual surfing habits of the users. This is also called secondary mining as it mines the log data generated by the web.

III. Web Usage Mining

The steps involved in the process of web usage mining are as shown in the figure 2.

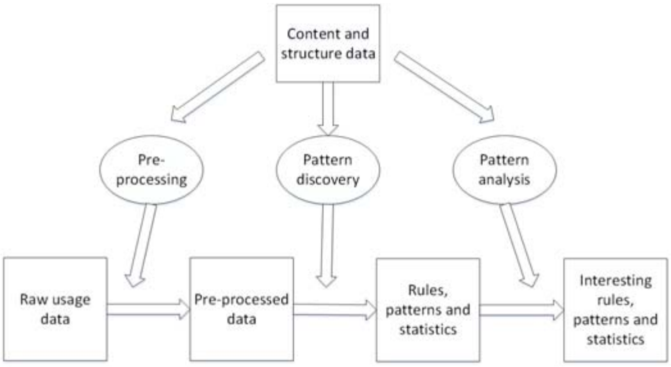
Content and structure data
Pre-processing
Pattern discovery
Pattern analysis
Raw usage data
Pre-processed data
Rules, patterns and statistics
Interesting rules, patterns and statistics

Figure 2 Web Usage Mining

A. *Pre-processing*

Data generated by servers is very bulky and noisy. It contains information about every resource access like image, video, web page etc. Identifying relevant data from this and organizing it in terms of users as well as sessions is what pre-processing performs. [1] [2] Steps of pre-processing are as follows.

1) Data cleaning
2) User identification
3) Session identification
4) Path completion

B. *Pattern discovery*

This step is performed to identify frequent patterns from server generated data. User accesses many resources through clicking the hyperlinks. By identifying the sequence of those click streams pattern about user's interests can be realized. These patterns if constrained by time threshold, sessions of requests can be identified. Mining those sessions, behaviour & interestingness of users can be identified. These tasks are performed in pattern discovery step. Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. [6] Various techniques involved in this step are frequent itemset mining, clustering, statistical analysis, classification and sequential analysis.

1) *Frequent itemset mining:*

This method can be used to find group of pages which are frequently accessed together with support exceeding a threshold. Frequent itemset mining also used to discover information like

- Set of pages repeatedly accessed together by web users
- The next page that will be fetched
- Frequently accessed paths by web users

Algorithms utilized for frequent itemset mining are as follows [11]

TABLE I Frequent Itemset Mining Algorithms

Algorithm	Storage	Pros	Cons
Apriori	Array	- Precise - Follows Apriori property - Easy & simple implementation [10]	- Multiple scans of the Database - Large time and space complexity
Apriori Tid	Array	- Number of entries are smaller than number of transactions	- Large time and space complexity
FP - Growth [10]	Tree (FP-tree)	- Only 2 scans of the Database - Iterative rule mining - Less memory requirements - Reduces total no of candidate itemsets [10]	- Large execution time
Custom built Apriori	Array	- Efficient and effective pattern analysis	- Large time and space complexities
FAP – Growth	Tree (FAP-tree)	- Mines both long and short patterns	- No sequence among the elements of data
K-Apriori [12]	Matrix (Binary data)	-Large datasets are partitioned -More efficient than Apriori	-Implementation complexity is increased
Multi objective Association Rule Mining with Evolutionary Algorithm [6]	Array	-Reduces the number of comparisons -Reduces time complexity -Performance improvement	-Works with only Boolean dataset rather than categorical and numerical data sets so conversion is required
Rapid Association Rule Mining (RARM) [18]	Support Ordered Trie Item set (SOTrieIT)	-Fast -Efficient -Scalable	-Difficult in incremental rule mining and iterative mining process
PD-FARM [19]	Tree (FP-tree)	-Efficient -Reduced no of database scans	-Large space complexity -Complex processing

TABLE II CLUSTERING ALGORITHMS

Algorithm	Pros	Cons
K-Means	-Feasible and Scalable	-Sensitive to Initial Parameter K -Unable to Handle Noise
Greedy clustering using belief function [4]	-Provide Efficient Results Using Dempster-Shafer's Belief Function	-Lacking in Scalability
Improved Fuzzy C Means	-Able to Identify Initial Cluster and Works on Irregular Datasets	-Sensitive to Noise
CLIQUE (Clustering In Quest)	-Measure Similarity Between Clustering	-More Time and Space Required
Cluster Optimization Using Fuzzy Cluster Chase	-Less Memory Utilization and Less Run Time	-Lacking in Scalability
K means with genetic algorithm [6]	-Minimises objective function	-Not the fastest algorithm but performance is comparable
Hierarchical Agglomerative Clustering [6]	-Can handle large dataset -Reduced execution time due to parallelism -Increase in efficiency	-Increase in efficiency is only linear
Cluster Optimization using Ant-Nestmate Approach [4] [6]	-Flexible -Robust -Optimize The Cluster by Increases Precision And Coverage	-Scalability -High performance
EB-DBSCAN (Entropy-Based DBSCAN) [6]	-great prospect in clustering of high-speed and massive data stream of arbitrary shape -uses batch data Processing so size of the data processing is effectively reduced and it can greatly reduce the time complexity -quick-building detecting models for high-speed, huge amount of stream data	-has a bit lower average purity than DBSCAN algorithm, but almost Equivalent -Parameter selection is very critical -The size of window is a direct factor affecting the average clustering precision
DBSCAN [16]	-Finds clusters of arbitrary shape -Doesn't need pre specified number of clusters -Insensitive to the ordering of the points in the Database	-Quality of the clusters depends on the distance measure used -Can't cluster datasets with large differences in the densities -border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data is processed

2) *Clustering*

It is a process of grouping together a set of items having similar features. Two types of clusters can be found in web usage mining: user clusters and page clusters. User clusters will discover users having same browsing patterns whereas page clusters will discover pages possessing similar content. The techniques of finding user clusters and page clusters are called usage based clustering and content based clustering respectively. [14] Agglomerative clustering [5] with single linkage, complete linkage, average linkage and centroid linkage are used for this task. Apart from that, various algorithms for clustering the web data are as shown in the table II.

3) *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. For example, naïve Bayesian classifiers, decision tree classifiers, k-nearest neighbour classifiers etc. [4]

TABLE III CLASSIFICATION ALGORITHMS

Algorithm	Pros	Cons
Naïve Bayesian [17]	-Higher time complexity -Low memory consumption	-High error rate -No of irrelevant attributes are more -Generated decision tree has more no of levels
CART	-Only relevant attributes are selected -Easy handling of missing values -High accuracy	-Generates only binary decision tree
C4.5	-Builds models that can be easily interpreted -Easy to implement -Deals with noise	-Small variation in data can lead to different decision trees -Does not work very well on a small training set
SVM	-Most optimal classifier -Works well on higher dimensions -Efficient utilization of memory	-kernel models can be quite sensitive to over-fitting the model selection criterion -Selection of kernel is crucial for precise classification
Backpropagation	-Efficiency -Can approximate any function reasonably well	-Convergence time is high -Sensitive to the value of learning rate -Sensitive to the number of hidden layers and neurons

4) Sequential analysis

It is same as association rule with the difference of time ordering. It finds patterns such that one or set of pages are accessed after the another set but in a time sequence. Its application is the prediction of future visitors so as to target advertising on a group of users. [3] Techniques utilized for sequential analysis are shown in table IV. [6]

TABLE IV SEQUENTIAL ANALYSIS ALGORITHMS

Algorithm	**Pros**	**Cons**
Hashing and pruning based algorithm	-Scalability	- Suffers from collisions
WAP tree association rule algorithm	-Scalability	-Recursively reconstructs the WAP tree
High utility sequential Patterns	-Scalability	-Large space complexity
Prefix span algorithm	-Scalability	- Constructs a projected database for every sequential pattern in worst case [20]
Transaction matrix comparison algorithm	-Scalability	- In some areas, the data available is insufficient to estimate reliable probability or transfer rates, especially for rare transitions

5) Statistical analysis

This is the most commonly used method in discovering knowledge about web users. Presently, there are many traffic analysis tools which generates a report depicting statistics such as mean length of path accessed, mean time of page viewing, most frequently accessed pages etc. [3]

C. Pattern analysis

The need behind pattern analysis is to filter out uninteresting rules or patterns from the set. The common techniques used for pattern analysis are visualization techniques, OLAP techniques, Data & Knowledge Querying, and Usability Analysis. Visualization techniques are useful to help application domains expert analyse the discovered patterns. [3]

IV. CONCLUSIONS

Web usage mining is very useful area in terms of analysis of users and their behaviour regarding the web contents. As stated in previous sections, there are vast amount of techniques available for web usage mining. Each technique has its advantages and disadvantages. These techniques can further be studied to identify the reason for their drawbacks and can be improved. Every technique is unique and efficient for specific nature of web data and application. Their combinations and improvements lead to successful results.

REFERENCES

[1] K. Sudheer Reddy, G. Partha Saradhi Varma, and M. Kantha Reddy, “An Effective Pre-processing Method for Web Usage Mining”, *International Journal of Computer Theory and Engineering*, vol. 06, no. 05, October 2014.

[2] Sujith Jayaprakash, Balamurugan E., “A Comprehensive Survey on Data Preprocessing Methods in Web Usage Mining”, *International Journal of Computer Science and Information Technologies*, vol. 06, no 03, pp. 3170-3174, 2015.

[3] Sanjeev Dhawan, Swati Goel, “Web Usage Mining: Finding Usage Patterns from Web Logs”, *American International Journal of Research in Science, Technology, Engineering & Mathematics*.

[4] Nirali H.Panchal, Ompriya Kale, “A Survey on Web Usage Mining”, *International Journal of Computer Trends and Technology (IJCTT)*, vol. 17, no. 04, Nov. 2014.

[5] Karuna Katariya, Rajanikanth Aluvalu, “Agglomerative Clustering in Web Usage Mining: A Survey”, *International Journal of Computer Applications*, vol. 89, no. 08, March 2014.

[6] D. Jayalatchumy, Dr. P.Thambidurai, “Web Mining Research Issues and Future Directions – A Survey”, *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 14, issue 03, pp. 20-27, Sep. - Oct. 2013.

[7] Preeti Chopra, Md. Ataullah, “A Survey on Improving the Efficiency of Different Web Structure Mining Algorithms”, *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 02, issue 03, Feb. 2013.

[8] Ms. B. Nagarathna, Dr. M. Moorthi,, “A study on web content mining and web structure mining”, *International Journal of Modern Trends in Engineering and Research (IJMTER)*, vol. 02, issue 05, May 2015.

[9] Faustina Johnson, Santosh Kumar Gupta, “Web Content Mining Techniques: A Survey”, *International Journal of Computer Applications*, vol. 47, no.11, June 2012.

[10] Aanum Shaikh, “Web Usage Mining Using Apriori and FP Growth Alogrithm”, *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 06, no. 01, 2015.

[11] Samiksha Kankane, Vikram Garg, “A Survey Paper on: Frequent Pattern Analysis Algorithm from the Web Log Data”, *International Journal of Computer Applications*, vol. 119, no.13, June 2015.

[12] Ashok Kumar D, Loraine Charlet Annie M.C.,” Web Log Mining using K-Apriori Algorithm”, *International Journal of Computer Applications*, vol. 41, no.11, March 2012.

[13] Jinguang Liu, Roopa Datla, “Web usage mining”.

[14] Mehak, Naveen Aggarwal, “Web Usage Mining: An Analysis”, *Journal of emerging technologies in web intelligence*, vol. 5, no. 3, august 2013.

[15] Jiawei Han, Michaline Kamber, Jian Pei, *Data mining concepts and techniques*, 3rd ed., Elsevier Inc., 2012.

[16] Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta, “Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery”, *International Journal of Data Mining Techniques and Applications*, vol. 02, issue 01, June 2013.

[17] A. K. Santra, S. Jayasudha, “Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification”, *International Journal of Computer Science Issues (IJCSI)*, Vol. 09, issue 01, no. 02, Jan. 2012.

[18] Amitabha Das, WeeKeong Ng, YewKwong Woon, Rapid Association Rule Mining.

[19] Maryam Jafari, Farzad Soleymani Sabzchi, Shahram Jamali, “Discovering Users` Access Patterns for Web Usage Mining from Web Log Files”, *Journal of Advances in Computer Research*, vol. 04, no. 03, pp. 25-32, Aug. 2013.

[20] Nidhi Grover, “Comparative Study of Various Sequential Pattern Mining Algorithms”, *International Journal of Computer Applications*, vol. 90, no. 17, March 2014