Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Web Usage Mining For Web Personalization Using Clustering Techniques

Kuldeep Singh Rathore*
M.Tech Student, MITS, Gwalior,
Madhya Pradesh, India

Sanjiv Sharma
Assistant Prof., MITS, Gwalior,
Madhya Pradesh, India

Abstract– The routine of grouping similar objects together is termed as 'Clustering'. The objects presents in the same cluster are similar and those presents to different cluster are dissimilar. The techniques for clustering are World Wide known so that they can be used in many applications for example biological, financial applications and many other fields. Web clustering is one of these application types where various heterogeneous types of objects can be clustered into different groups for various purposes and usage. This survey paper deals with the different aspects of Web data mining and provides an overview about the various techniques incorporated in the field of web mining.

Keywords– Data Mining, Web Mining, Web Usage Data, Clustering, Web Personalization

I. INTRODUCTION

The volatile expansion of online data because of the Internet and the common use of databases have formed huge need for KDD methodologies. Knowledge Discovery and Data Mining (KDD) is an emerging area focused on the methods for mining useful information or extract knowledge from raw data [1]. Here users uses navigation traces, which can be pulled based on the user behavior analysis. The applications of web mining similar analyses have been successfully executed by methods of Web Usage Mining [2]. The challenge of finding out knowledge from data extract from research web user behavior and checking high performance computing, to provide advanced business intelligence in statistics, machine learning, data visualization, optimization, databases, pattern recognition, and web discovery solutions. It is a influential technology with high prospective to help different kind of industries to focus on the significant information in their data warehouse. Data Mining trends can be ordered into two classifications, Descriptive Mining and Predictive Mining [2]. The various Descriptive Mining strategies, for patterns, Clustering, Association Rule Discovery, Sequential Pattern Discovery, is utilized to discover human-interpretable patterns that depict the data. The Predictive Mining procedures like Classification, Regression, Deviation Detection, utilize a few variables to predict not know or future scoped values of different variables.

II. WEB MINING

Web mining is the sub field of data mining techniques to extract knowledge from Web data, Web log data is followed in the mining process. Researchers have found in three different forms of Web mining [3, 4, 5] .

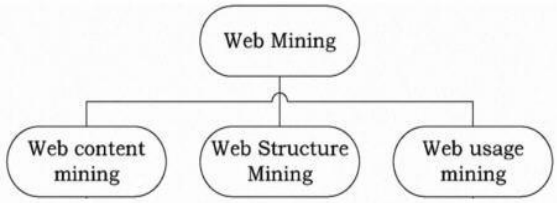

Fig. 1 Web Mining Categories

1. *Web Content Mining* – (find the content of web pages and results of web Searching) The data present on the Web page was designed to provide the users. This usually contained images, graphs, charts, but is not limited to, text and graphics only.
2. *Web Structure Mining* - (Hyperlink Structure) Data that explains the organization of the content. Intra page (pages from same file) structure information incorporates the arrangement of various HTML or XML tags within a given page. The inter page structure information is hyperlinks connecting pages one by one to another pages.
3. *Web Usage Mining* - (analyzing user web navigation) Data that gives information about the pattern of Web pages, like IP addresses, page references, and the date and time of data access. Typically, the use of data comes from an Extended Common Log Format (ECLF) Server log files.

III. WEB USAGE MINING

Web usage mining is used to searching out the patterns of user activities in order to full fill the needs of the users for example by dynamic link handling, by page recommendation, etc. The future scope of a Web site or Web portal is to

supply the user the information which is useful for him. There is competition between the various commercial portals and Web sites because every user means ultimately money (through advertisements, etc.). Thus the goal of each owner of a portal is to give nice look and feel for the user. Due to this reason the response time of each single site have to be kept below 2s. Some extras have to be provided like supplying dynamic content, links and recommended pages for the user that are according to user interest. Clustering or grouping of the user activities stored in different types of log files is in the form of key that is issued in the Web community.

There are basically three types of log files which can be in usage for Web usage mining [6]. Log files are kept on various places like server side, client side and on the proxy servers. Due to the presence of one or more places for storing the information of navigation patterns of the users the mining process become more crucial. Reliable result obtained only on single condition if data composed from all types of log files. It happens because of server side don't have records of those Web page that are accesses and stored on the proxy servers or on the client side safely. Besides the log file on the server, that present on the proxy server gives additional information.

Web usage mining consists of three main steps: (i) preprocessing, (ii) pattern discovery and (iii) pattern analysis [7]. Figure shows the block diagram of the Web usage mining.

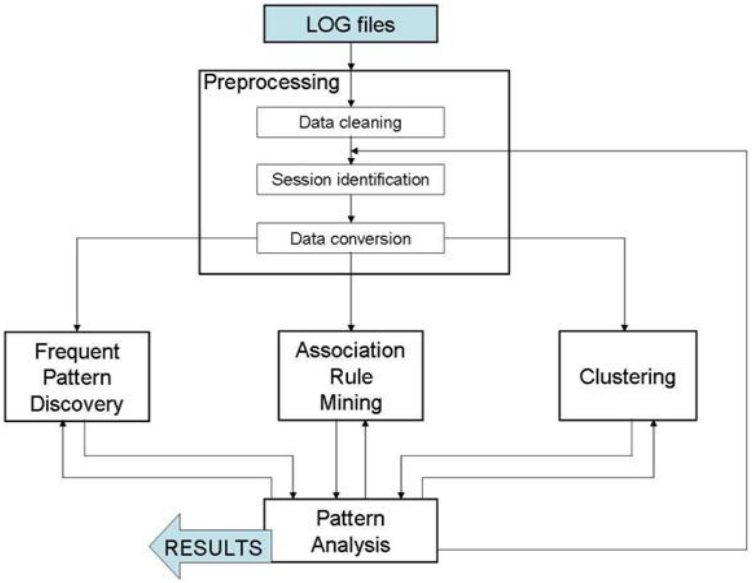

Fig. 2 Process of Web usage mining

Approaches

A. Data collection

There are three main sources from where data can be collected i.e. a) Server log files b) Proxy log files c) Client Log files.

1 *Server log file:* The most popular source for extract web usage mining is basically web server log data [8]. The web log data is automatically generated through web server when a user request arrives, which consist all information or data about user's activity. The frequent server log file types are access log, agent log, error log and referrer log.

2 *Proxy log file:* A Proxy server is an intermediate server that exists between the client and the Web server. Therefore if the Web server received a request of the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. To gather the information of the user, these web proxy servers maintain a separate log file.

3 *Client log file:* This type of log fie is present at the client side but still, all the entries are made at the server side. Client log files include activities and events that happen within the premises of client machine.

B. Data preprocessing

The raw data collected from the previous steps may contain noise, impurities or may be unformatted. So, data pre-processing is performed. The various data pre-processing methods are:

1. *Data cleaning and feature selection:* This method is used to remove unnecessary/irrelevant fields from the raw data. For example- removing JPEG, GIF, JPG files and audio/video files because they are not executed on the basis of user's request. Even the entries which are made to mark the unavailability of any resource or page are also removed. The entries occurred from crawlers or spiders also required to be removed because they do not show the way how users navigate through the web sites.
2. *User Identification:* User Identification means identifying a unique user. In most cases log files provides IP addresses. In other cases, log files also contain user login [9]. When user login is not available then IP addresses are used along with type of OS and browsing software.
3. *Session Identification:* A user session is distinct as a set of pages visited by the same user inside the duration of one particular visit to a website [10]. A user may be used a single or multiple sessions along a period. When a user has been recognized, the click stream of each user is divided into logical clusters.

C. Knowledge Discovery

It is the decisive stage where some useful knowledge will be derived by applying various statistical and/or data mining techniques at hand from various research areas like data mining, machine learning, statistical method and pattern recognition. Following methods are generally used for pattern discovery process.

1. *Sequential Pattern Mining:* It is used to mine the frequently occurred patterns related to the order of items in a sequence database. The discovered frequent data is useful for broad application, such as retail business, disease treatments, market analysis, etc. The task of sequential pattern mining is helpful for different applications, including market analysis, decision support, and business administration. One vital issue is to find continuous sequential patterns in a sequence database. Also, the enormous majority of the past works have concentrated on the order of the times. Sequential pattern mining is an important task in data mining.
2. *Association Rules:* It is a procedure for finding frequent patterns, correlations and associations among sets of stuffs and it is used to relate pages that are most frequently located together in a single server session. Association rules are basically used in order to disclose correlations among pages accessed together throughout a server session. Those types of rules point out the possible relationship between pages that are often viewed together even if they are not directly connected, and can disclose associations between groups of users with specific interests.
3. *Clustering:* Clustering is used for grouping of items that have same distinctiveness. In the Web Usage Mining, there are two main different types of interesting clusters to find user clusters and page clusters. User clustering results in clusters of users that reacts same when navigating through a Web site and Page clustering identifies clusters of pages that appear to be conceptually it is correlated according to the user's perception.

D. Pattern Analysis

Results of pattern discovery might not be in the form for real interpretation. Pattern analysis provides ways to contrast the results and to extract interesting rule or pattern from output of Pattern discovery. For this purpose, various visualization and presentation tools are used which represent data in 2D, 3D pictorial representation. These tools compare and characterize result in terms of charts, graphs, tables, Wien diagram and so many others visual presentations. Most of the times result generated or data itself are stored in data cubes or in data ware house.

IV. CLUSTERING

Clustering is basically performed for grouping set of objects in a manner that object belongs from same cluster are more similar to each other than other clusters. It is mainly used for exploring in the field of data mining and most popular technique for statistical data analysis. It has several applications information retrieval and biomedical data information, like machine learning, pattern discovery.

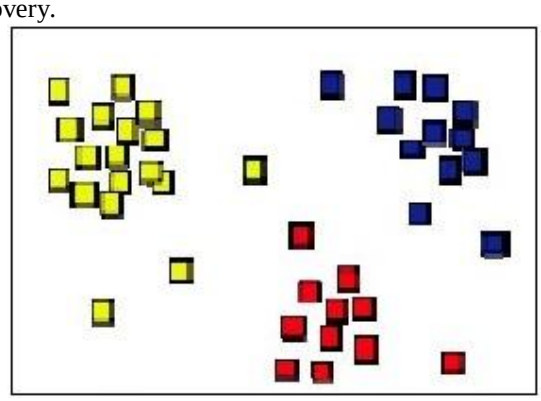

Fig. 3 Clusters of Similar Objects

A. Requirements of Clustering

Cluster analysis work efficiently and effectively, as many, Literature surveys have presented, there are the following typical Requirements for performing clustering in data mining.

1 Minimal requirements for domain knowledge to searched input parameters.
2 Insensitivity to the order of input records.
3 High dimensionality.
4 Scalability.
5 Quality to deal with different types of attributes.
6 Discovery of several clusters with arbitrary shape.

The researchers are focused on finding user behavior by using several good cluster analysis techniques.

B. Clustering Algorithms

1 *Hierarchical algorithms:* provide a level wise grouping of the objects. There exist two approaches, the bottom-up and the top-down approach. In case of bottom-up approach, at the starting of the algorithm each object presents a different group and at last all objects belong to the identical cluster. In top-down approach at the beginning of the algorithm all objects belong to the homogenous cluster which is split, until each object constitute a heterogeneous cluster. The steps of the algorithms can be represented using a dendrogram (tree diagram). The resulting clusters are

understand by cutting the dendrogram by a certain level or node. The main purpose of such algorithm is performing the distance identification between the objects and clusters. Many algorithms used to measure distance between the objects, for example Euclidian, City-Block, Minkowski and so on. The distance can be find between the clusters. Clusters can be searched between two near clusters, farthest clusters or the medians of clusters. The disadvantage of the hierarchical algorithm is that once an object is assigned to a given cluster it cannot be modified or changed in later phase. Furthermore, as in partition-based case, here also only spherical clusters can be obtained. The pros of the hierarchical algorithms is that the validation indices include correlation, inconsistency measure, that can be defined on the clusters. It can be used for finding the number of the clusters. The best known hierarchical clustering methods are CHAMELEON [11], BIRCH [12] and CURE [13].

2 *Density-based algorithms:* start by finding the core objects, and they are rising the clusters based on these cores and by searching for those objects that are in locality within a radius of a given object. The advantage of these type of algorithms is that they can detect arbitrary form of clusters and it can filter out the noise. DBSCAN [14] and OPTICS [15] are density-based algorithms.

3 *Grid-based algorithms:* The algorithms based on grids use a hierarchical grid structure to broken down the object space into countable number of cells. Statistical information of each cell is stored about the objects and the clustering is achieved on these cells. The pros of this method is the fast processing time that is in general independent of the number of data objects. Grid-based algorithms Wave Cluster [16].

4 *Model-based algorithms:* use different Distribution models for the clusters which must be verified during the clustering algorithm. A model-based clustering method is MCLUST [17].

5 *Fuzzy algorithms:* let us consider that no hard clusters exist on the set of various objects; one object can be assigned to more than one cluster. The best known fuzzy clustering algorithm is FCM (Fuzzy CMEANS) [18].

V. APPLICATION AREAS

1. *Personalization of web content:* This is the generalized application of web usage mining as the website owners can personalize their web services based on the user interests. Recommendation Systems are the most ordinary application in this area as their aim is to recommend interesting links to the users.
2. *Pre-fetching and caching:* The results of web usage mining used to boost the performance of web servers and its applications. Web Usage mining also be taken into account to develop Perfecting and caching strategies to reduce server response time.
3. *Design Support:* Usability is very important for websites. Web Usage Mining methods can be used to help the website owners to tackle the design and implementation issues of websites.
4. *E-commerce:* Mining business intelligence to the Web usage data is very important for ecommerce. Web Usage Mining techniques help in Customer Relationship Management (CRM). In this case, the spotlight is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure.

VI. WEB PERSONALIZATION

Web personalization is the way of converting a website [19] to the need of every particular user or set of users, taking benefit of the information attain through the exploration of the customer's navigational performance". Personalization used to supply services to each specific user in a well manner. The broadcast of information on the internet has finished the personalization system is an compulsion. The personalization has capability to solve the additional data problems and let the customers practice at least action to find the data they require [20]. When the customers work on the website their actions can be categorized into two wide sets: browsing and searching. If the customers want to search data on internet, user have to distribute the web scheme with a search demand according to need. If the customers don't provide the web system according to precise search request, the system would have huge volume of in appropriate data which not be able to fulfill the requirement of the user. If the customers' request is explicit, searching data on the web might turn out to be simple. To deliver better option to the user the web personalization system's developers must recognize what the customers' behavior when they access the web.

VII. RELATED WORK

The focus of literature review is to study the various steps to be performed in web usage mining and the various threats to web usage mining. Websites attacks are also very common. These attacks are implemented using information available from the web log files. In 2000,

Jaideep Srivastava et al. [21] gave detailed overview of web usage mining with the help of web SIFT which performs web usage mining in the standard NSCA format. The web SIFT system provides the option of converting server session into episodes. The authors also explained the possible ways to protect the data and for maintaining individual identity.

Chintan R. Varnagar et al. [22] also discussed the work done so far on data collection and pre-processing stage of web usage mining. The authors presented the main sources of data i.e. Client Log File, Proxy Log File, Web Server Log File. The various data pre-processing techniques such as data cleaning/feature extraction user identification and session identification were also by the authors.

Soumi Ghosh and Sanjay Kumar Dubey. [24] This paper shows the result of the clustering process and competence of its domain application are generally found through algorithms. In this two main clustering algorithms namely centric based K-Means' and 'representative object based FCM clustering' algorithms are compared. These two algorithms are

applied and throughput is evaluated on the basis of the efficiency of clustering output. The number of dataset points as well the number of clusters is the factors on which the behavior patterns of first and second algorithms are analyzed. FCM gives results like K-Means clustering but it still requires more computation time than K-Means clustering.

VIII. CONCLUSIONS

This paper enlightened some points with the difficulty of finds hidden information from large amount of log data, namely web log (also called as database), with Web usage mining. The main point of focus is clustering with the various different mining processes. After describing the process of clustering, the various techniques of clustering with their approaches are discussed. These algorithms serve as basis for the web usage clustering which was the main focus of the paper. The every aspects of classifying the web usage the algorithms of clustering were described and a classification based on these algorithms is provided as well. Web usage clustering algorithms are briefly described with classification based on the various aspects of such algorithms.

REFERENCES

[1] Ajith Abraham,"Business Intelligence from Web Usage Mining" Journal of Information & Knowledge Management, Vol. 2, No. 4 (2003) 375-390.
[2] M.N. Murty, A.K. Jain, P.J. Flynn, "Data clustering: a review", ACM Computer. Survey. 31 (3) (1999) 64– 323.
[3] Shaily G. Langhnoja, Mehul P . Bardot, Dashiki B. Meth a, "Web Usage Mining to Discover Visitor Group with Common Behavior Using DBSCAN Clustering Algorithm", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.
[4] Monika Yadav, Mr. P radeep Mitt al, "Web Mining: An Introduction".
[5] International Journal of Advanced Research in Computer Science and Software Engineering, March 2013.
[6] J Srivastava, R. Cooley, M Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *SIGKDD Explorations*, vol. 1, no. 2, pp. 12–23, 2000.
[7] P Batista, M. arid, and J. Silva, "Mining web access logs of an on-line newspaper," 2002.
[8] J Borges and M. Levene, "Data mining of user navigation patterns," in *WEBKDD*, pp. 92–111, 1999.
[9] M. Balabanovic and Y. Shoham, "Learning information retrieval agents: Experiments with automated web browsing," in Proceedings of the AAAI Spring Symposium on Information Gathering from Het-erogenous, Distributed Resources, pp.13-18,1995.
[10] H Han and R. Elmasri, "Learning rules for conceptual structure on the web," *J. Intel. Inf. Syst.*, vol. 22, no. 3, pp. 237–256, 2004.
[11] G Karypis, E.-H. S. Han, and V. K. NEWS,"Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75,1999.
[12] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," pp. 103–114, 1996.
[13] S. Guha, R. Rastogi, and K. Shim, "Cure: An efficient clustering algorithm for large databases," in SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA (L. M. Haas and A. Tiwary, eds.), pp. 73–84, ACM Press, 1998.
[14] M Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *KDD*, pp. 226–231, 1996.
[15] W. Wang, J. Yang, and R. Muntz, "Sting: A statistical information grid approach to spatial data mining," 1997.
[16] R Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," pp. 94–105, 1998.
[17] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A multi-resolution clustering approach for very large spatial databases,".
[18] C. Fraley and A. Raftery, "Mclust: Software for model-based cluster and discriminant analysis," 1999.
[19] M. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," Kluwer Academic Publisher's Machine Learning, vol. 42, pp. 31-60, 2001.
[20] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.C. Hsu, "FreeSpan: Frequent patternprojected sequential pattern mining," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '00), pp. 355-359, Aug. 2000.
[21] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", Volume 1, Issue 2 - page 12, ACM, 2000.
[22] Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya, Jayesh N. Rathod, "Web Usage Mining: A Review on Process, Methods and Techniques", IEEE, 2013.
[23] Soumi Ghosh , Sanjay Kumar Dubey (Department of Computer Science and Engineering, Amity University, Uttar Pradesh, Noida, India)" Comparative Analysis of K-Means and Fuzzy C-Means Algorithms",Vol. 4, 2013.
[24] K.Suresh R.Madana Mohana and A.RamaMohanReddy " Improved FCM algorithm for Clustering on Web Usage Mining",vol. 8,2011.

