International Journal of Scientific Research and Modern Education (IJSRME)
Impact Factor: 7.137, ISSN (Online): 2455 – 5630
(www.rdmodernresearch.com) Volume 3, Issue 2, 2018

PATTERN DISCOVERY TECHNIQUES IN WEB USAGE MINING

J. Umarani* & Dr. S. Manikandan**
* Research Scholar, Research and Development Centre, Bharathiyar University, Coimbatore, Tamilnadu
** Head, Department of Computer Science and Engineering, Sriram Engineering College, Chennai, Tamilnadu

Cite This Article: J. Umarani & Dr. S. Manikandan, "Pattern Discovery Techniques in Web Usage Mining", International Journal of Scientific Research and Modern Education, Volume 3, Issue 2, Page Number 1-3, 2018.

Abstract:

WWW is a very popular and interactive medium for broadcasting information today. Due to the vast, diverse and lively nature of web it advancesthe scalability, multimedia data and temporal issues respectively. The development of the web has given rise to large quantity of data that is freely available for user access.Web Usage Mining enhances the user experience while browsing web pages by using past history of web data. It also used to improve the web site navigation. Web mining makes use of data mining techniques and deciphers potentially useful information from web data. Web usage mining is divided into three parts Preprocessing, Pattern discovery and Pattern analysis. Pattern analysis techniques are used to highlight overall patterns in data and to filter out uninteresting patterns. The techniques like Knowledge Query Mechanism, OLAP and visualization are used for pattern analysis.Web Usage mining deals with understanding the behavior of users by making use of Web Access Logs that are generated on the server while the user is accessing the website.This paper gives patterndiscovery techniques for web usage mining.

Key Words: Web mining, Web Usage Mining & Pattern Analysis

1. Introduction:

Web usage mining refers to the automatic discovery and analysis of patterns in click stream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites [1]. The goal is to capture, model, and analyze the behavioural patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests.

The overall Web usage mining process can be divided into three inter-dependent stages: data collection and pre-processing, pattern discovery, and pattern analysis. In the pre-processing stage, the click stream data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site. [2] In the pattern discovery stage, statistical, database, and machine learning operations are performed to obtain hidden patterns reflecting the typical behaviour of users, as well as summary statistics on Web resources, sessions, and users. In the final stage of the process, the discovered patterns and statistics are further processed, filtered, possibly resulting in aggregate user models that can be used as input to applications such as recommendation engines, visualization tools, and Web analytics and report generation tools. The overall process is depicted in figure 1.

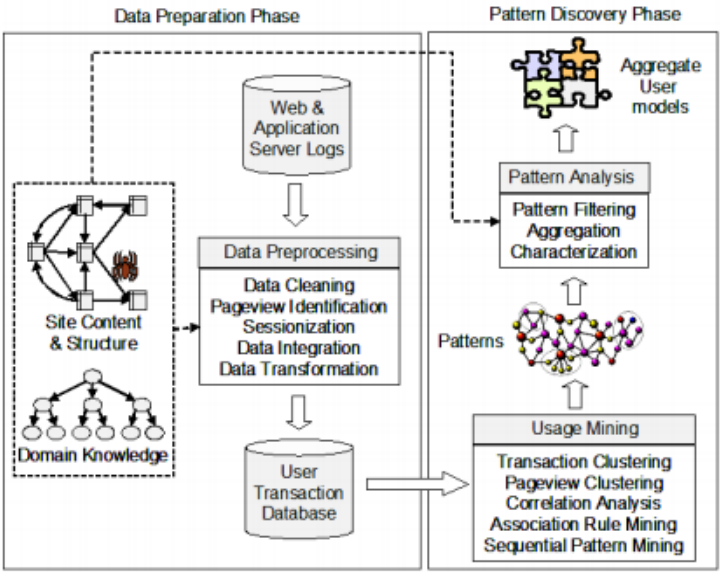

Figure 1: The Process of Web Usage Mining (WUM)

2. Pattern Discovery Techniques:

Pattern Discovery is used to extract patterns of usage from web data [3]. This method uses data mining techniques and algorithms to find out useful information. Knowledge extracted can be represented in many ways such as graphs, charts, tables, forms etc. Techniques used in pattern discovery are given as follows:

Association Rules: Association rule generation can be used to relate pages that are most often referenced together in a single server sessions [4]. In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. Association rule mining has been well studied in Data Mining, especially for basket transaction data analysis. Many association rule algorithms have been used, such as Apriori, Partition [5]. Aside from being applicable for e-Commerce, business intelligence and marketing applications, it can help web designers to restructure their web site. The results about the usefulness of such rules in supermarket transaction or in web application have not been reported. People also put some constraints over the mining process, and prune the extracted rules. The association rules may also serve as heuristic for pre fetching documents in order to reduce user-perceived latency when loading a page from a remote site. In electronic CRM, an existing customer can be retained by dynamically creating web offers based on associations with threshold support and/or confidence value [6].

Clustering: Clustering is a technique to group together a set of items having similar characteristics [7]. Clustering can be performed on either the users or the page views. Clustering analysis in web usage mining intends to find the cluster of user, page, or sessions from web log file, where each cluster represents a group of objects with common interesting or characteristic. User clustering is designed to find user groups that have common interests based on their behaviours, and it is critical for user community construction. Page clustering is the process of clustering pages according to the users' access over them. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in e-Commerce applications or provide personalized web content to the users. On the other hand, clustering of pages will discover groups of pages having related content. This information is useful for the Internet search engines and Web assistance providers. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information needs. The intuition is that if the probability of visiting page, given page has also been visited, is high, then maybe they can be grouped into one cluster. For session clustering, all the sessions are processed to find some interesting session clusters. Each session cluster may be one interesting topic within the web site. Mobasher et al [8] generated recommendations from URL clusters to build an adaptive web site by using ARHP (Association Rule Hypergraph Partitioning).

Sequential Pattern Mining: Sequential patterns in Web usage data capture the web page trails that are often visited by users, in the order that they were visited. These are sequences of items that occur in a sufficiently large proportion of (sequence) transactions. The view of web transactions as sequences of pageviews paved way to a number of useful and well-studied models in discovering user navigation patterns. One such approach is to model the navigational activities in the website as a Markov Model (MM): each pageview in this model can be represented as a state and the transition probability between two states can represent the likelihood that a user will navigate from one state to the other. This representation allows for the computation of a number of useful user or site metrics. Lower order markovian model lack accuracy because of its limitation of covering enough browsing history. Higher-order Markov models generally provide a higher prediction accuracy but result in much higher model complexity due to the larger number of states. Pitkow et al. [9] proposed all-kth-order Markov models (for coverage improvement) and a new state reduction technique, called longest repeating sub sequences to overcome the coverage and space complexity problems (for reducing model size). The use of all-kth-order Markov models generally requires the generation of separate models for each of the k orders. If the model cannot make a prediction using the kth order, it will attempt to make a prediction by incrementally decreasing the model order. This scheme can easily lead to even higher space complexity since it requires the representation of all possible states for each k. Deshpande et al. [10] proposed selective markov models in which they proposed three different techniques to overcome the space complexity of existing all-kth-order Markov models.

k-Nearest Neighbour: k-Nearest Neighbour (kNN) is based on the principle that the instances within a dataset will generally exist in close proximity to other instances that have similar properties. As kNN does not make any assumptions on the underlying data distribution and does not use the training data points to do any generalization, it is called as non parametric lazy learning algorithm. Guo et al. [11] proposed a novel kNN type method for classification that is aimed at overcoming the drawback of its dependency on the selection of a "good value" for k. Yu & Liu [12] addressed the problem of determining which of the available input features should be used in modeling via feature selection because it could improve the classification accuracy and scale down the required classification time.

Collaborative Filtering (CF): Collaborative filtering (CF) is a technique utilized primarily to predict individuals' preferences, has its origin in information filtering. This technique guides an active user depending on the preferences shared by like users. Once a database of preferences of like users is accumulated, a similarity measure is used to identify individuals with similar past preferences with the active user. A preference function

is applied on the database to guide or recommend the active user [13]. This technique is easy to comprehend and implement, but requires a large sample to make meaningful recommendations. Erroneous recommendations result when close neighbors don't exist. Content information and customer profile or behavior information is not used for making recommendations. As database size increases, the recommendation computation becomes computationally more intensive. These also suffer from a fundamental problem, called sparsity problem. Since the set of all possible available items in a system is very large, most users may have rated very few items, and, hence, it is difficult to find the active user‟s neighborhood with high similarity. As a result the accuracy of the recommendations may be poor. To overcome the above disadvantages classification and prediction had its application in the web domain of collaborative filtering. Lin et al. [14] proposed a collaborative recommendation system using association rules. Zhong hang Xia et al. [15] proposed a collaborative filtering system with SVM. Koji Miyahara and Michael J. Pazzani [16] proposed a collaborative filtering system with Bayesian classifier. Miha Grcar et al. [17] presented experimental results of confronting the k-Nearest Neighbor (kNN) algorithm with SVM in the collaborative filtering framework using datasets with different properties. Dhruv Gupta et al. [18] emphasized on a new, principal component analysis and clustering-based linear time collaborative filtering algorithm for efficient and effective personalized information retrieval.

3. Conclusion:

This paper gives an insight into the possible data mining techniques with Web usage data for achieving a synergetic effect of Web usage mining. Association rules are used to discover pages that are visited together quite often. Discovering sequential patterns from web access logs can be used for predicting future visits of the users. Clustering discovers groups of users or pages, based on their similarities. Classification classifies the new user into one of the predefined groups based on their maximum likelihood. It is hard, if not impossible, to declare that one data mining algorithm is the best in general, because the possible outcomes of WUM process always depend on the problem in hand.

4. References:

1. Brijendra Singh, Hemant Kumar Singh: Web Data Mining Research: A Survey, IEEE, 2010.
2. G.K. Gupta, Introduction to Data Mining with Case Studies: Web Data Mining, PHI Learning Private Limited, pp. 231-233, 2011.
3. Theint Aye: Web Log Cleaning for Mining of Web Usage Patterns, IEEE, 2011.
4. Mobasher B., Dai H., Luo T. and Nakagawa M., "Discovery of aggregate usage profiles for web personalization", WebKDD‟00, USA pp. 61–82, 2000.
5. Yong Wang, Zhanhuai Li and Yang Zhang, “Mining Sequential Association-Rule For Improving Web Document Prediction", ICCIMA‟05, pp. 146–151, 2005.
6. Yu L. and Liu H., "Efficient Feature Selection via Analysis of Relevance and Redundancy", JMLR, 5(Oct):1205-1224, 2004.
7. Bamshad Mobasher, Robert Cooley, Jaideep Srivastava," Creating Adaptive Web Sites Through Usage-Based Clustering of URLs", proceedings of the 1999 workshop on knowledge and data engineering, pp 19, 1999.
8. Pitkow J. and Pirolli P., "Mining Longest Repeating Subsequences to Predict WWW Surfing", Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems, 1999.
9. Deshpande M., Karypis G., "Selective Markov Models for Predicting Web-Page Accesses", Proceedings of the 1st SIAM International Conference on Data Mining, 2004.
10. Zhong S. and Ghosh J., "A unified framework for model based clustering", Machine Learning Research 4, 1001– 1037, 2003.
11. Guo G., Wang H., Bell D., Bi Y., and Greer K. "KNN Model-Based Approach in Classification", Lecture Notes in Computer Science, Vol 2888, Pages 986 – 996, 2003.
12. Yu L. and Liu H. , "Efficient Feature Selection via Analysis of Relevance and Redundancy", JMLR, 5(Oct):1205-1224, 2004.
13. http://www.fico.com/en/Communities/AnalyticTechnologies/Pages/CollaborativeFiltering.aspx.
14. Lin W., Alvarez S. A., and Ruiz C., "Efficient Adaptive Support Association Rule Mining for Recommender Systems”, Data Mining and Knowledge Discovery, vol. 6, pp. 83–105, 2002.
15. Zhonghang Xia Yulin Dong and Guangming Xing, "Support Vector Machines For Collaborative Filtering", ACM SE’06,Melbourne, Florida, USA., pp 10-12, March 2006.
16. Koji Miyahara and Michael J. Pazzani, "Collaborative Filtering with the Simple Bayesian Classifier", Proceedings of the 6th Pacific Rim International conference on Artificial intelligence, pp. 679689 ,Springer-Verlag Berlin, Heidelberg, 2000.
17. Miha Grcar, Miha Grcar, and Dunja Mladenic, "kNN Versus SVM in the Collaborative Filtering Framework", WebKDD „05, August 21, Chicago, Illinois, ACM 159593-214-3, 2005.
18. Dhruv Gupta, Mark Digiovanni, Hiro Narita, and Ken Goldberg, "Jester 2.0 : Evaluation of a New Linear Time Collaborative Filtering Algorithm", SIGIR „99 Berkley, CA, USA ,ACM, 1999.