International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 04, April-2015

A Survey on Web usage Mining: Process, Techniques and Applications

Tanveer Kaur Dewgun[1]
[1]MTech Scholar,
Bansal Institute of Science and Technology,
Bhopal

Pushpraj Singh Chauhan[2]
[2]Professor,
Bansal Institute of Science and Technology,
Bhopal

***Abstract*— Due to the tremendous growth in the World Wide Web, there has been large amount of web data generated. This web data can be mined to provide results which can greatly improve the performance and help business to grow. In this paper, we have discussed about the type of web mining which uses the web server logs as data sources known as web usage mining. We have surveyed about the process of web usage mining and the different techniques which includes clustering, classification, association rule mining etc. We also studied about the various applications of web usage mining.**

***Keywords*— *Association Rule Mining, Clustering, Pre-fetching, Web Mining, Web Usage Mining*.**

I. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure, and web usage data [1].The World Wide Web is a huge repository of data which develops increasingly. This data produced by WWW can be discovered to extract information and acquire knowledge using the data mining techniques to further enhance the web.

The web provides different type of information that can be discovered using mining techniques:

- Web user activity, from server logs and Web browser activity tracking.
- Web graph structure, from links between pages, people and other data.
- Web content, for the data found on Web pages and inside of documents.

Extracting knowledge from the content of documents or their descriptions is known as web content mining. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs.

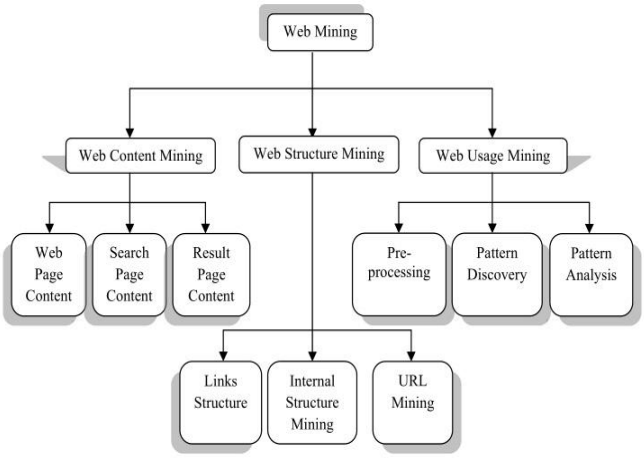

Fig.1. Web Mining Categories

II. WEB USAGE MINING

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications[2] [3].

A web user whenever interacts with the web server leaves behind traces of information which when analyzed helps improving the design of the system. Web servers record these user interactions which captures the identity or origin of web users with their browsing behavior in the access logs. The three main tasks for performing Web Usage Mining are Data Gathering, Data Preprocessing, Pattern Discovery and Pattern Analysis. Preprocessing consists of converting the usage, content and structure information contained in the various available data sources into the data abstraction necessary for pattern discovery [3].

III. WEB USAGE MINING PROCESS

A. Data Gathering:

Most of the web usage mining systems use the log data as their primary source of data. A Web log file records activity information when a Web user submits a request to a Web server. A log file can be located in three different places: i) Web servers, ii) Web proxy servers, and iii) client browsers.

B. Data Preprocessing:

Preprocessing consists of converting the usage, content and structure information contained in the various available data sources into the data abstraction necessary for pattern discovery [3]. This step includes: i) data cleaning, ii) identifying user, iii) building sessions and iv) path completion.

C. Pattern Discovery:

This step involves using statistical method to carry on the analysis and mine the processed data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

D. Pattern analysis:

Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user. Pattern analysis requires the knowledge of query languages such as SQL. The summarized data is loaded into a data cube and OLAP operations are performed to get better results.

IV. WEB USAGE MINING TECHNIQUES:

Pattern discovery phase of web usage mining can be achieved using different techniques such as association rule mining, sequential patterns, clustering and classification.

A. Association rule mining:

Association rule mining is a data mining technique which helps to predict the future events by studying the history of events. It is used to find the frequently accessed pages by the users which are referenced together. Association rule mining using Apriori algorithm [3] may find correlation between users who visited a page containing electronic products to those who access a page about sport related products. A typical example of frequent itemset mining is market basket analysis. This process analyzes customer buying habits by finding association between different items that customers place in their"shopping carts (baskets)". The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers [3].

B. Clustering:

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters [2]. In the Web Usage Mining, there are two types of interesting clusters to be discovered, usage clusters and page clusters. Clustering of users tends to find groups of users showing similar browsing patterns. Such knowledge is useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users. Also clustering of pages will discover groups of pages having related content. This information is very useful for Internet search engines and Web assistance providers.

C. Classification

It is the task of categorizing data items into one of several predefined classes. It can be done by using supervised learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines etc [3]. Weblog information can be integrated with Web content and Web linkage structure mining to help Web page ranking, Web document classification and the construction of multilayered Web information base as well. In a particular discipline, the documents need to be classified based on subject index classification standard such as to classify a set of Web documents automatically, Web linkage information to improve the quality of such classification, use Web usage information to improve the quality of such classification [4].

D. Sequential Patterns

Sequential patterns discovery is to find the inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. Web log files can record a set of transactions in time sequence. If the web-based companies can discover the sequential patterns of the visitors, the companies can predict users' visit patterns and target market on a group of users.

V. APPLICATIONS

The general goal of Web Usage Mining is to gather interesting information about users navigation patterns (i.e., to characterize web users). This information can be exploited later to improve the web site from the users' viewpoint. The results produced by the mining of web logs can use for various purposes : (i) to personalize the delivery of web content; (ii) to improve user navigation through prefetching and caching; (iii) to improve web design; or in e-commerce sites; (iv) to improve the customer satisfaction.

A. Personalization of Web Content

Web Usage Mining techniques can be used to provide personalized web user experience. For instance, it is possible to anticipate, in real time, the user behavior by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users [4]. Personalized Site Maps [5] are an example of recommendation system for links. [6] proposed an adaptive technique to reorganize the product catalog of the products according to the forecasted user profile. A survey on existing commercial recommendation systems, implemented in e-commerce web sites, is presented in [7].

B. *Pre-fetching and Caching*

The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. Typically, Web Usage Mining can be used to develop proper pre-fetching and caching strategies so as to reduce the server response time, as done in [8].

C. *Support to the Design.*

Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications. [9] Uses stratograms to evaluate the organization and the efficiency of web sites from the users' viewpoint.[10] exploits Web Usage Mining techniques to suggest proper modifications to web sites. Adaptive Web sites represents a further step. In this case, the content and the structure of the web site can be dynamically reorganized according to the data mined from the users' behavior [11].

D. *E-commerce*

Mining business intelligence from web usage data is dramatically important for e-commerce web-based companies. Customer Relationship Management (CRM) can have an effective advantage from the use of Web Usage Mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure [12].

VI. CONCLUSION

Web Mining is the process of extracting knowledge from the huge data repository the World Wide Web. In this paper, firstly we gave an introduction about the web miing and also studied the categories of web mining. Web Mining a part of data mining in terms of Web includes three categories, Web Content Mining, Web Structure Mining and Web Usage Mining. In the next sections, we focus on mainly Web Usage Mining and the process of web usage mining. Then we studied the techniques and different applications of web usage mining.

REFERENCES

[1] Web Mining— Concepts, Applications, and Research Directions Jaideep Srivastava, Prasanna Desikan, Vipin Kumar.

[2] Srivastava J., Cooley R., Deshpande, M. and Tan, P. N., Web Usage Mining: Discovery and applications of Usage Patterns from web data', SIGKDD Explorations 2000.

[3] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd Edition, Elsevier.

[4] S. Shiu C. Wong and S. Pal. Mining fuzzy association rules for web access case adaptation. In Case-Based Reasoning Research and Development: Proceedings of the Fourth International Conference on Case-Based Reasoning, 2001.

[5] Fergus Toolan and Nicholas Kushmerick. Mining web logs for personalized site maps.

[6] Hye-Young Paik, Boualem Benatallah, and Rachid Hamadi. Dynamic restructuring of e-catalog communities based on user interaction patterns. World Wide Web, 5(4):325–366, 2002.

[7] J. Ben Schafer, Joseph A. Konstan, and John Riedl. E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5(1-2):115–153, 2001.

[8] Bin Lan, Stephane Bressan, Beng Chin Ooi, and Kian-Lee Tan. Rule-assisted prefetching in web-server caching. In Proceedings of the ninth international conference on Information and knowledge management (CIKM 2000), pages 504–511. ACM Press, 2000.

[9] Bettina Berendt. Using site semantics to analyze, visualize, and support navigation. Data Mining and Knowledge Discovery, 6(1):37–59, 2002.

[10] Yongjian Fu, Mario Creado, and Chunhua Ju. Reorganizing web sites based on user access patterns. In Proceedings of the tenth international conference on Information and knowledge management, pages 583–585. ACM Press, 2001.

[11] Tapan Kamdar. Creating adaptive web servers using incremental web log mining. Master's thesis, Computer Science Department, University of Maryland, Baltimore County, 2001.

[12] Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. ACM Transactions on Internet Technology (TOIT), 3(1):1–27, 2003.