Survey of Tasks and Application Areas of Web Usage Mining

A.Nirali Honest, B.Dharmendra Patel, C.Atul Patel

Abstract-The World Wide Web (WWW) provides a simple yet effective media for users to search, browse, and retrieve information in the Web. Application of Data Mining Techniques to the World Wide Web is referred to as Web Mining. Web mining can be classified into three ways (i) Web Structure Mining (ii) Web Content Mining and (iii) Web Usage Mining. Web Usage Mining aims to discover interesting patterns generated by user's interaction on web usage data. Web usage mining has increased interest from both research and business communities. This paper presents the survey of tasks of web usage mining and describes each task in detail.

Index Terms— Association Rules, Data Cleaning, Preprocessing, Web Mining.

1 INTRODUCTION

Mining [2] information and knowledge from large database has been recognized by many researchers, so it has emerged as a field of research in recent years. Modern era is of internet, so data on web are very important for both business community and research group. The data on web has given the birth of new research area and that is Web Mining. Web Mining [4] is the application of data mining to World Wide Web. As shown in Fig.1 there are three main sub categories of it based on which part of web to mine: Web Content Mining, Web Structure Mining and Web Usage Mining.

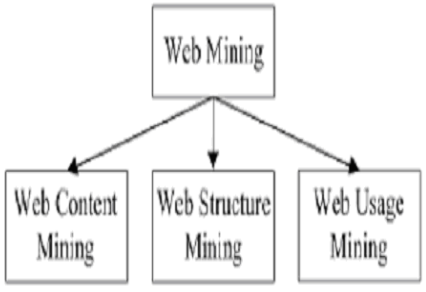

Fig. 1. Categories of Web Mining.

Web Content Mining describes the discovery of useful information from web content/data/document. Web Structure Mining tries to discover the model underlying the link structure of the web. Web Usage Mining focuses on techniques that could predict the user behavior while user interacts with the web. It aims to discover interesting patterns generated by user's interaction on the web usage data.).

This paper is structured as follows. In section 2 we give an overview of Web Usage mining with the details of steps involved in the process, i.e. Data Preprocessing, Pattern Discovery and Pattern Analysis, in section 3 we give the application areas that can gain benefit because of web usage mining and finally we conclude in section 4.

2 WEB USAGE MINING

2.1 Notion of Web usage mining

Web usage mining involves automatic discovery of user access patterns from Web data. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by web servers and collected in server access logs. WUM can be decomposed into the following subtasks as shown in Fig. 2

Data Preprocessing: It is necessary to perform a data preparation to convert the raw data for further process. The data taken as input are web server logs, referral logs, registration files, index server logs & optionally statistics from a previous analysis, and the output generated are User session file, transaction file, site topology and page classification.

Pattern Discovery: Pattern discovery covers the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition.

Pattern Analysis: The main goal of Pattern Analysis is to filter uninteresting information and to visualize and interpret the interesting patterns to the user

- *F.A. Author is with the Charotar Institute of Computer Application,, Charotar University of Science & Technology,Changa, Gujarat, India.*
- *S.B. Author Jr. is with th Charotar Institute of Computer Application,, Charotar University of Science & Technology,Changa, Gujarat, India.*
- *T.C. Author is with the Charotar Institute of Computer Application,, Charotar University of Science & Technology,Changa, Gujarat, India.*

© 2011 JICT
www.jict.co.uk

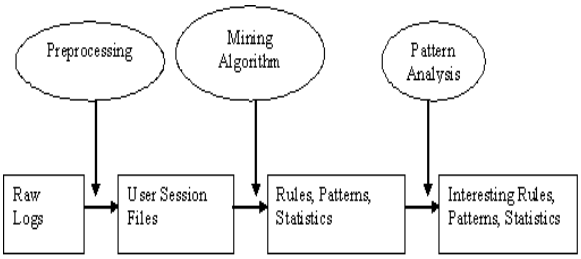

Fig. 2. Web Usage Mining Phases

2.2 Data Preprocessing

Data Collection: The first step in web usage mining is Data Collection. The data integrity and authenticity will affect the next phases of work so it must be collected accurately. The main data source for web usage mining is a log file. The key function of log file is to record the browsing behavior of site visitor(s). These logs can be collected by web servers, intermediate proxies or client browser. Web Servers record access information as a click-stream data into log files. Whenever a Web page is clicked, corresponding data will be generated and recorded. The most accepted log file formats are the Common Log Format (CLF) and the extended CLF. A common log format file is created by the web server to keep track of the requests that occur on a web site. A standard log file has the following format as shown in Fig. 3.

<ip address>< base url><date><method><file> <protocol><code><bytes><referrer><user agent>

Fig. 3. Common Web Log Format

Data Cleaning: Data cleaning means eliminate the irrelevant information (information which is not useful for the further process, or even misleading or faulty)from the original Web log file and exchange the Web log as database which is convenient for further processing. This irrelevant information includes that:

1. Unless the Web site mainly includes picture pages, the request for picture pages, such as .gif, .jpeg, .jpg, .css, and .is deleted because these image file auto download with our requested pages.
2. Transmission error such as 404(not found), 301 (deleted perpetually), 500(inner server error).
3. The request finished by auto search engine, such as Crawler, Spider, Robot etc., which exceed the range of Web usage mining.
4. Use a request method except "Get". There also need to delete some irrelevant fields while deleting some irrelevant records.

User Identification: User identification means identifying each user accessing Web site, whose goal is to mine every user's access characteristic, and then make user clustering and provide personal service for the users. We may assume that each user has unique IP address and each IP address represents one user. But in fact there are three conditions:

1. Some user has unique IP address.
2. Some user has two or more IP addresses.
3. Due to proxy server, some user may share one IP address.

Session Identification: For Web logs that span long periods of time, it is very likely that users will visit the Web site more than once. After user identification, these pages access of each user must be divided into individual session, which is session identification. The goal session identification is to find each user's access pattern and frequency path. The simplest method is using a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout, and [3] established a timeout of 25.5 minutes based on empirical data

Data Formatting and Summarization: This is the last step of data preprocessing. Here, the structured file containing sessions and visits are transformed to a relational database model. Then, the data generalization method is applied at the request level and aggregated for visits and user sessions to completely fill in the database. Two tables are designed in the relational database model – one for storing the log data and the other for storing the sessions. The data summarization concerns with the computation of aggregated variables at different abstraction levels (for example request, visit, and user session).These aggregated variables are later used in the data mining step. They represent statistical values that characterize the objects analyzed. Depending on the objective of the analysis, the analyst can decide to compute additional and more complex variables.

2.3 Pattern Discovery

Pattern discovery covers the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. It has separate subsections as follows,

Statistical Analysis: Statistical analysts may perform different kinds of descriptive statistical analyses based on different variables when analyzing the session file. By analyzing the statistical information contained in the periodic web system report, the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitation the site modification task, and providing support for marketing decisions

Association Rules: In the web domain, the pages, which are most often referenced together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions.

Clustering: Clustering analysis is a technique to group together users or data items (pages) with the similar characteristics. Clustering of user information or pages can

facilitate the development and execution of future marketing strategies.

Classification: Classification is the technique to map a data item into one of several predefined classes. The classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifier, Support Vector Machines etc.

Sequential Patterns: The technique of sequential pattern discovery attempts to find inter transactions patterns such that the presence of a set of items is followed by another item in a time stamp ordered set of sessions or episodes. In Web server transaction logs, a visit by a client is recorded over a period of time. The time stamp associated with a transaction in this case will be a time interval, which is determined and attached to the transaction during the data cleaning or transaction identification processes. The discovery of sequential patterns in Web server access logs allows Web-based organizations to predict user visit patterns and helps in targeting advertising aimed at groups of users based on these patterns. By analyzing this information, the Web mining system can determine temporal relationships among data items. Another important kind of data dependency that can be discovered, using the temporal characteristics of the data, is similar time sequences. For example, we may be interested in finding common characteristics of all clients that visited a particular file within the time period. Or, conversely, we may be interested in a time interval (within a day, or within a week, etc.) in which a particular file is most accessed. Other types of temporal analysis that can be performed on sequential patterns include trend analysis, change point detection, or similarity analysis.

Dependency Modeling: The goal of this technique is to establish a model that is able to represent significant dependencies among the various variables in the web domain. The modeling technique provides a theoretical framework for analyzing the behavior of users, and is potentially useful for predicting future web resource consumption.

2.4 Pattern Analysis

Pattern Analysis is a final stage of the whole web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transform to a format can be assimilate easily. There are two most common approaches for the patter analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multidimensional data cube before perform OLAP operations. All these methods assume the output of the previous phase has been structured.

3 Application areas of Usage Mining

Personalization: Web Usage Mining techniques can be used to provide personalized web user knowledge. The science behind personalization has undergone tremendous changes, and several Web-based personalization systems have been proposed in recent years. Although personalization can be accomplished in numerous ways, most Web personalization techniques so far fall into three major categories: decision rule-based filtering, content-based filtering, and collaborative filtering [5]. Web usage mining uses data mining algorithms to automatically discover and extract patterns from Web usage data and predict user behavior while users interact with the Web. Recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users [6],[7],[8],[9] . Personalized site maps [10] are an example of recommendation system for links.

System Improvement: Performance is a crucial attribute to user satisfaction. The results produced by Web Usage Mining can be used to improve performance of web servers and web based applications. Web Usage Mining can be used to develop proper prefetching and caching strategies to reduce the server response time as done in [11], [12],[13],[14],15]. Web usage mining provides the key to understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission [1], load balancing, or data distribution.

Site Modification: Attractiveness and usability are 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. [16] Uses stratograms to evaluate the organization and the efficiency of web sites from the users point of view. [17] Exploits techniques to suggest proper modifications to web sites. Adaptive web sites can be built, in this case the content and the structure of the web site can be dynamically reorganized according to the data minied from the users behavior [18], [19] .

Business Intelligence : Mining business intelligence from web usage data is important for e-commerce web-based companies. Customer Relationship Management (CRM) can have an effective advantage from the use of Web Usage Mining Techniques. Here the focus is on business specific issues such as customer attraction, customer retention, cross sales and customer departure [20], [21], [22].

4 Conclusion

World Wide Web is the hub of information in modern era; everything is available on web so there is a growing need to develop tools and techniques that will help to improve its overall usefulness. Web mining has been used to refer to different kind of techniques that encompasses a broad range of issues related to web. In this paper we survey the different tasks and application areas of web mining which focuses on issues related to understanding

the behavior of web users, called Web Usage Mining. We have provided the detail survey of different tasks that are very useful to understand the area of Web Usage Mining. This paper also explored the different applications areas of Web Usage Mining. This paper is very useful for persons who require further research in field of Web Usage Mining.

REFERENCES:

[1] E. Cohen, B. Krishnamurthy, and J. Rexford. Improving end-to-end performance of the web using server volumes and proxy filters. In Proc. ACM SIGCOMM, 1998, pp. 241-253.

[2] Ming-Syan Chen,Jiawei Han,Philip S., Data Mining : An overview from database perspective,IEEE transactions on knowledge and data engineering,Vol.8,No.6, December-1996.

[3] Perkowitz M, Etzioni 0. Adaptive sites: Automatically learning from user access patterns. In:Proc of 61h Int'l World Wide Web Conf. Santa Clara, California, 1997

[4] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A survey, SIGKDD Explanation,ACM SIGKDD,July-2000. {2}

[5] Special Issue on Personalization and Privacy, IEEE Internet Computing, vol. 5, Nov./Dec. 2001, pp. 29–62.

[6] Gediminas Adomavicius and Alexander Tuzhilin. Extending recommender systems: A multidimensional approach.

[7] 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.

[8] Debra VanderMeer, Kaushik Dutta, and Anindya Datta. Enabling scalable online personalization on the web. In Proceedings of the 2nd ACM E-Commerce Conference (EC'00), pages 185–196. ACM Press, 2000.

[9] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Effective personalization based on association rule discovery from web usage data. In Web Information and Data Management, pages 9–15, 2001.

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

[11] Cheng-Yue Chang and Ming-Syan Chen. A new cache replacement algorithm for the integration of web caching and prefectching. In Proceedings of the eleventhinternational conference on Information and knowledge management, pages 632–634. ACM Press, 2002.

[12] 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.

[13] Tianyi Li. Web-document prediction and presending using association rule sequential classifiers. Master's thesis, Simon Fraser University, 2001.

[14] Alexandros Nanopoulos, Dimitrios Katsaros, and Yannis Manolopoulos. Exploiting web log mining for web cache enhancement. In R. Kohavi, B. Masand, M. Spiliopoulou, and J. Srivastava, editors, WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Third International Workshop, SanFrancisco, CA, USA, August 26, 2001. Revised Papers, volume 2356 of Lecture Notes in Computer Science, pages 68–87. Springer, 2002.

[15] Yi-Hung Wu and Arbee L. P. Chen. Prediction of web page accesses by proxy server log. World Wide Web, 5(1):67–88, 2002.

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

[17] 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.

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

[19] Osmar R. Za¨ıane. Web usage mining for a better web-based learning environment.In Proceedings of Conference on Advanced Technology for Education, pages 450–455, 2001.

[20] Catherine Bounsaythip and Esa Rinta-Runsala. Overview of data mining for customer behavior modeling. Technical Report TTE1-2001-18, VTT Information Technology, 2001.

[21] Suhail Ansari, Ron Kohavi, Llew Mason, and Zijian Zheng. Integrating e-commerce and data mining: Architecture and challenges. In Nick Cercone, Tsau Young Lin, and Xindong Wu, editors, Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001). IEEE Computer Society, 2001.

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