Ethos of Web Usage Mining - A Survey

Simmi Bagga[1*] & Satinder Kaur[2]

[1]Assistant Professor, Sant Hira Dass Kanya Maha Vidyalaya, Kala Sanghian, INDIA

[2]Assistant Professor, GNDU, RC, Sathiala, INDIA

Abstract: Nowadays, web mining is the highest era to acquire information and knowledge. Web mining can be broadly defined as the search and measure of useful information from the World Wide Web. It provides automatic search of information resources available on-line, and the discovery of user access information from Web servers, i.e., Web usage mining. Web Usage Mining is the application of data mining methods which analyze the recordings of Web usage, which are stored in the form of Web server logs. The goal of web usage mining is to find the user's access patterns quickly in the form of frequent traversal paths, frequent access page-sets, and user clustering. Web usage mining mines the data which are generated by the web users while interacting with the web. The mining processes include data preparation, mining process, and process analysis. It provides automatic search of information resources available on-line, and the discovery of user access information from Web servers, i.e., Web usage mining. This paper presents research issues, techniques and application areas of web usage mining. The paper presents the detailed taxonomy and survey of the existing efforts in Web Usage mining and elaborates potential application areas of web usage mining.

Keywords: Component; Formatting; Style; Styling; Insert

I. INTRODUCTION

Web Mining is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data, Web Usage mining involves mining the usage characteristics from the users of Web Applications. This extracted information can then be used in a variety of ways such as, improvement of the application, checking of fraudulent elements etc. Web usage mining has then become a necessary task in order to provide meaningful information to web administrators about users and usage patterns for a website in order to improve the quality of web information and service performance. Successful websites may be those that are customized to meet user preferences in two aspects i.e. information presentation and content relevance [1-2].

To implement web usage mining, data can be collected from server logs, browser logs, proxy logs, or obtained from an organization's database. Web mining tries to extract interesting, potentially useful and hidden information from the documents and logs on the Web in order to help people so that they can abstract knowledge from WWW. Web usage mining is similar to data mining in some extent. It is a cross field of database, data mining, artificial intelligence, information retrieval, natural language understanding and so on. These data collections differ in terms of the location of the data source, the kinds of data available, the segment of population from which the data was collected, and methods of implementation so have a lot of heterogeneity [3-7].

II. METHODS OF WEB MINING

Web mining can be done in various ways. Each way depends on need of information one wants from web mining.

a. ***Web Content Mining (WCM)***: Content means the visible data in the web pages or the information which was meant to be imparted to the users. A major part of it includes text and graphics (images). Because a text document presents no machine-readable semantic so some approaches have been suggested which are used to restructure content of document in a representation so that machines can understand it such as Free texts, HTML Files, XML Files, Dynamic Content, Multimedia Files. The usual approach to exploit known structure in documents is to use wrapper to map documents to particular data model. There are mainly two groups for web content mining strategies, one which directly mine the document's contents and other which improve on the search contents of other tools like search engines as shown in figure 1 [1-7].

b. ***Web Structure Mining (WSM)***: It is the mining method which provides information about the organization of the website. It is further divided into two types.
   - *Intra-page structure information* includes the arrangement of various HTML or XML tags within a given page.
   - *Inter-page structure information* is the hyper-links used for site navigation.

c. ***Web Usage Mining (WUM):*** It provides the information that describes the usage patterns of Web pages, such as IP addresses, page references, and the date and time of accesses and various other information depending on the log format. Free texts, HTML Files, XML Files, Dynamic Content, Multimedia Files [1-7].

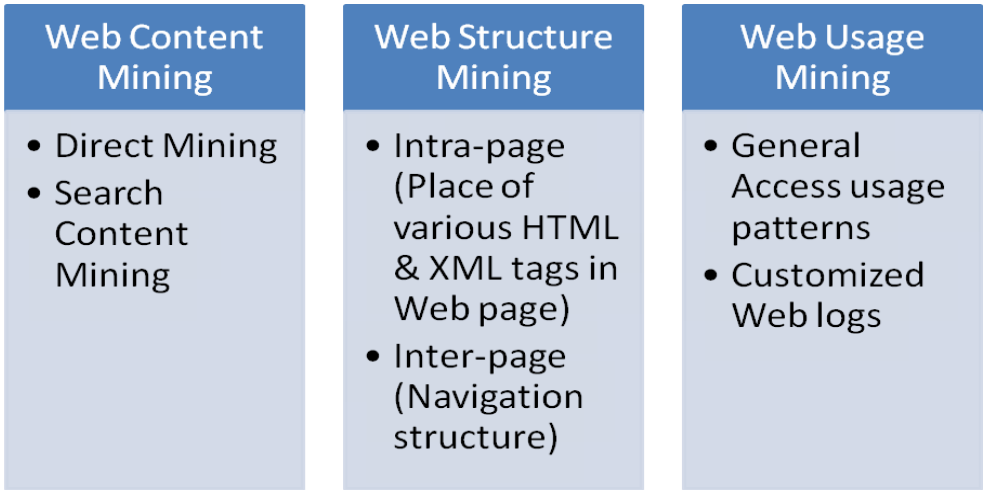

Figure 1: Types of Web Mining

III. WEB USAGE MINING DATA SOURCES

The data sources used in Web Usage Mining include web data repositories like:

a. ***Web Server Logs*** – These are logs which maintain a history of page requests. The W3C maintains a standard format for web server log files, but other proprietary formats exist. More

recent entries are typically appended to the end of the file. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added. These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information. These files are usually not accessible to general Internet users, only to the webmaster or other administrative person. A statistical analysis of the server log may be used to examine traffic patterns by time of day, day of week, referrer, or user agent. Efficient web site administration, adequate hosting resources and the fine tuning of sales efforts can be aided by analysis of the web server logs. Marketing departments of any organization that owns a website should be trained to understand these powerful tools.

b. ***Proxy Server Logs*** - A Web proxy is a caching mechanism which lies between client browsers and Web servers. It helps to reduce the load time of Web pages as well as the network traffic load at the server and client side. Proxy server logs contain the HTTP requests from multiple clients to multiple Web servers. This may serve as a data source to discover the usage pattern of a group of anonymous users, sharing a common proxy server.

c. **Browser Logs** – Various browsers like Mozilla, Internet Explorer etc. can be modified or various JavaScript and Java applets can be used to collect client side data. This implementation of client-side data collection requires user cooperation, either in enabling the functionality of the JavaScript and Java applets, or to voluntarily use the modified browser. Client-side collection scores over server-side collection because it reduces the session identification problems [1-7].

IV. APPLICATION AREAS OF WEB USAGE MINING

The data sources used in Web Usage Mining include web data repositories like:

a. ***Web Server Logs*** – These are logs which maintain a history of page requests. The W3C maintains a standard format for web server log files, but other proprietary formats exist. More recent entries are typically appended to the end of the file. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added. These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information. These files are usually not accessible to general Internet users, only to the webmaster or other administrative person. A statistical analysis of the server log may be used to examine traffic patterns by time of day, day of week, referrer, or user agent. Efficient web site administration, adequate hosting resources and the fine tuning of sales efforts can be aided by analysis of the web server logs. Marketing departments of any organization that owns a website should be trained to understand these powerful tools.

b. ***Proxy Server Logs*** - A Web proxy is a caching mechanism which lies between client browsers and Web servers. It helps to reduce the load time of Web pages as well as the network traffic load at the server and client side. Proxy server logs contain the HTTP requests from multiple clients to multiple Web servers. This may serve as a data source to discover the usage pattern of a group of anonymous users, sharing a common proxy server.

c. ***Browser Logs*** – Various browsers like Mozilla, Internet Explorer etc. can be modified or various JavaScript and Java applets can be used to collect client side data. This implementation of client-

side data collection requires user cooperation, either in enabling the functionality of the JavaScript and Java applets, or to voluntarily use the modified browser. Client-side collection scores over server-side collection because it reduces the session identification problems [1-7].

V. ADVANTAGES OF WEB MINING

Web mining is attractive for companies because of several advantages. In the most general sense it can contribute to the increase of profits, be it by actually selling more products or services, or by minimizing the costs. In order to do this, marketing intelligence is required. This intelligence can focus on marketing strategies and competitive analyses or on the relationship with the customers. The different kinds of web data that are somehow related to customers will then be categorized and clustered to build detailed customer profiles. Ultimately, Web usage mining is the key to enhance the web site performance [1-7].

VI. CONCLUSION

Web mining techniques seek to extract knowledge from Web data. Web mining can be broadly defined as to discover and analyze useful information from the World Wide Web. So, we can say that web mining provide facility to extract important information from World Wide Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs. The paper is mainly focused on Web usage analysis, partly because of its applicability in web based application. In this paper we discuss web usage mining, the process tasks associates with it. We also described how web usage mining as applicable in various web based application.

DECLARATION

We declare that this paper has no conflict of interest.
We also declare that we have followed ethical responsibilities.

REFERENCES

[1] J. Srivastava, P. Desikan, and V. Kumar, “Web Mining: Accomplishments and Future Directions,” Proc. US Nat’l Science Foundation Workshop on Next-Generation Data Mining (NGDM), Nat’l Science Foundation, 2002.
[2] Abha Moitra (Scotia, NY, US) Steven Matt Gustafson (Niskayuna, NY, US) Feng Xue (Clifton Park, NY, US), GENERAL ELECTRIC COMPANY
[3] Eirinaki, M., and Vazirgiannis, M.,Web Mining for Web Personalization. ACM Transactions on Internet Technology, 2002.
[4] Soumen Chakrabarti, "Mining the Web: Analysis of Hypertext and Semi Structured Data", Morgan Kaufmann, 2002
[5] Van Dongen, B. F., de Medeiros, A. K. A., Verbeek, H. M. W., Weijters, A. J. M. M., & Van Der Aalst, W. M. (2005). The ProM framework: A new era in process mining tool support. In Applications and Theory of Petri Nets 2005 (pp. 444-454). Springer Berlin Heidelberg.
[6] Jesus Mena, "Data Mining Your Website", Digital Press, 1999
[7] Soumen Chakrabarti, "Mining the Web: Analysis of Hypertext and Semi Structured Data", Morgan Kaufmann, 2002

