The Descriptive Study of Knowledge Discovery from Web Usage Mining

Yogish H K[1] Dr. G T Raju[2] Manjunath T N[3]

**[1]Department of Computer Science and Engineering
REVA Institute of Technology and Management,
Yelahanka. Bangalore-560064, Karnataka, India.
(Research Scholar - Bharathiar University Coimbatore-641046)**

**[2]Department of Computer Science and Engineering
RNS Institute of Technology, Bangalore -560061,
Karnataka, India.**

**[3]Wipro Technologies
(Research Scholar - Bharathiar University Coimbatore-641046)**

Abstract

The World Wide Web serves as huge, widely distributed, global information service centre for news, advertisements, consumer information, financial management, education, government, e-commerce and many other information services. The web also contains a rich and dynamic collection of hyperlink information and web page access and usage information, providing rich sources of data for data mining. The Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from web logs, in order to improve web based applications. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use this information for the specific needs.

Keywords*: Data mining, Knowledge Discovery, bot, Preprocessing, associations, clustering, web data.*

1. Introduction

Web Usage Mining is a part of Web Mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of web data.

The web data is:

1. Content: 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).
2. Structure: Data which describes the organization of the website. It is divided into two types. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. The principal kind of inter-page structure information is the hyper-links used for site navigation.
3. Usage: Data 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.

In Web Usage Mining[3,5,7], data can be collected in server logs, application server logs, browsers logs, user's profiles, user's queries, book marked data, mouse clicks and scrolls, registration data, cookies, user sessions or transactions data. 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.

The logs can be examined from either server or client perspective. When evaluated from a server perspective, mining uncovers informs about a sites where the service resides, it can be used to improved the design of the sites. By evaluating client's sequence of clicks information about the users or group of users or detected. This could be used to perform pre-fetching and caching of pages.

For Example: The web master at ABC [3] corporation learns that a high percentage of users have the following patterns of reference to pages :(A, B, A, C). This means that user access page A, then page B, then back to page A and finally to page C. Based on this observation he determines that a link is needed directly to page C from B. He then adds this link.

Web Usage mining involves mining the usage characteristics of 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 is often regarded as a part of the Business Intelligence in an organization rather than the technical aspect. It is used for deciding business strategies through the efficient use of Web Applications. It is also crucial for the Customer Relationship Management (CRM) as it can ensure customer satisfaction as far as the interaction between the customer and the organization is concerned.

The major problem with Web Mining in general and Web Usage Mining in particular is the nature of the data they deal with. With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format so that it needs a lot of preprocessing and parsing before the actual extraction of the required information.

The Web Usage Mining process consists three phases, namely preprocessing, pattern, and discovery and pattern analysis. This paper describes each of these phases in detail.

1.1 Motivation

In the current era, we are witnessing a surge of Web Usage around the globe. A large volume of data is constantly being accessed and shared among a varied type of users; both humans and intelligent machines. Thus, taking up a structured approach to control this information exchange, has what made Web Mining one of the hot topics in the field of Information Technology.

2 Data Sources [1]

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

Web Server Logs [7] - 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.

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.

Browser Logs - Various browsers like Mozilla, Internet Explorer Opera 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 both the bot and session identification problems.

3. Information Obtained [3]

i. ***Number of Hits*****:** This number usually signifies the number of times any resource is accessed in a Website. A hit is a request to a web server for a file (web page, image, JavaScript, etc.). When a web page is uploaded from a server the number of "hits" or "page hits" is equal to the number of files requested. Therefore, one page load does not always equal one hit because often pages are made up of other images and other files which stack up the number of hits counted.

ii. ***Number of Visitors****:* A "visitor" is exactly what it sounds like. It's a human who navigates to your website and browses one or more pages on your site.

iii. ***Visitor Referring Website****:* The referring website gives the information or URL of the website which referred the particular website in consideration.

iv. ***Visitor Referral Website:*** The referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

v. Time and Duration: This information in the server logs give the time and duration for how long the Website was accessed by a particular user.
vi. Path Analysis: Path analysis gives the analysis of the path a particular user has followed in accessing contents of a Website.
vii. Visitor IP address: This information gives the Internet Protocol (I.P.) address of the visitors who visited the Website in consideration.
viii. Browser Type: This information gives the information of the type of browser that was used for accessing the Website.
ix. Cookies: A message given to a Web browser by a Web server. The browser stores the message in a text file called cookie. The message is then sent back to the server each time the browser requests a page from the server. The main purpose of cookies is to identify users and possibly prepare customized Web pages for them. When you enter a Web site using cookies, you may be asked to fill out a form providing such information as your name and interests. This information is packaged into a cookie and sent to your Web browser which stores it for later use. The next time you go to the same Web site, your browser will send the cookie to the Web server. The server can use this information to present you with custom Web pages. So, for example, instead of seeing just a generic welcome page you might see a welcome page with your name on it.

x. ***Platform:*** This information gives the type of Operating System etc. that was used to access the Website.

4. Possible Actions [9]

i. Shortening Paths of High visit Pages: The pages which are frequently accessed by the users can be seen as to follow a particular path. These pages can be included in an easily accessible part of the Website thus resulting in the decrease in the navigation path length.

ii. Eliminating or Combining Low Visit Pages: The pages which are not frequently accessed by users can be either removed or their content can be merged with pages with frequent access.

iii. Redesigning Pages to help User Navigation: To help the user to navigate through the website in the best possible manner, the information obtained can be used to redesign the structure of the Website. Redesigning Pages For Search Engine Optimization: The content as well as other information in the website can be improved from analyzing user patterns and this information can be used to redesign pages for Search Engine Optimization so that the search engines index the website at a proper rank.

iv. Help Evaluating Effectiveness of Advertising Campaigns*:* Important and business critical advertisements can be put up on pages that are frequently accessed.

5. Web Usage Mining Process [1]:

The main processes in Web Usage Mining are:

Preprocessing: Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

Pattern Discovery: Web Usage mining can be used to uncover patterns in server logs but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. The following are the pattern discovery methods.

1. Statistical Analysis
2. Association Rules
3. Clustering
4. Classification
5. Sequential Patterns

Pattern Analysis [11]: This is the final step in the Web Usage Mining process. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information. The methods like SQL
(Structured Query Language) processing and OLAP (Online Analytical Processing) can be used.

5.1 Web Usage Mining Areas [15]

1. Personalization
2. System Improvement
3. Site Modification
4. Business Intelligence
5. Usage Characterization

5.2 Web Usage Mining Applications [3]

i. ***Letizia [4]:*** Letizia is an application that assists a user browsing the Internet. As the user operates a conventional Web browser such as Mozilla, the application tracks usage patterns and attempts to predict items of interest by performing concurrent and autonomous exploration of links from the user's current position. The application uses a best-first search augmented by heuristics inferring user interest from browsing behavior.

ii. WebSift[1]: The WebSIFT (Web Site Information Filter) system is another application which performs Web Usage Mining from server logs recorded in the extended NSCA format (includes referrer and agent fields. The preprocessing algorithms include identifying users, server sessions, and identifying cached page references through the use of the referrer field. It identifies interesting information and frequent item sets from mining usage data.

iii. Adaptive Websites: An adaptive website adjusts the structure, content, or presentation of information in response to measured user interaction with the site, with the objective of optimizing future user interactions. Adaptive websites are web sites that automatically improve their organization and presentation by learning from their user access patterns. User interaction patterns may be collected directly on the website or may be mined from Web server logs. A model or models are created of user interaction using artificial intelligence and statistical methods. The models are used as the basis for tailoring the website for known and specific patterns of user interaction.

6. Analysis of Web Server Logs

We used different web server log analyzers like Web Expert Lite 6.1 and Analog6.0 to analyze various sample web server logs obtained. The key information obtained was:
Total Hits, Visitor Hits, Average Hits per Day, Average Hits per Visitor, Failed Requests, Page Views Total Page Views, Average Page Views per Day , Average Page Views per Visitor, Visitors Total Visitors Average Visitors per Day, Total Unique IPs , Bandwidth, Total Bandwidth , Visitor Bandwidth , Average Bandwidth per Day, Average Bandwidth per Hit, Average Bandwidth per Visitor. Access Data like files, images etc., Referrers, User Agents etc. Analysis of above obtained information proved Web Usage Mining as a powerful technique in Web Site Management and improvement.

Fig: shows the generalized model of web usage mining [8]

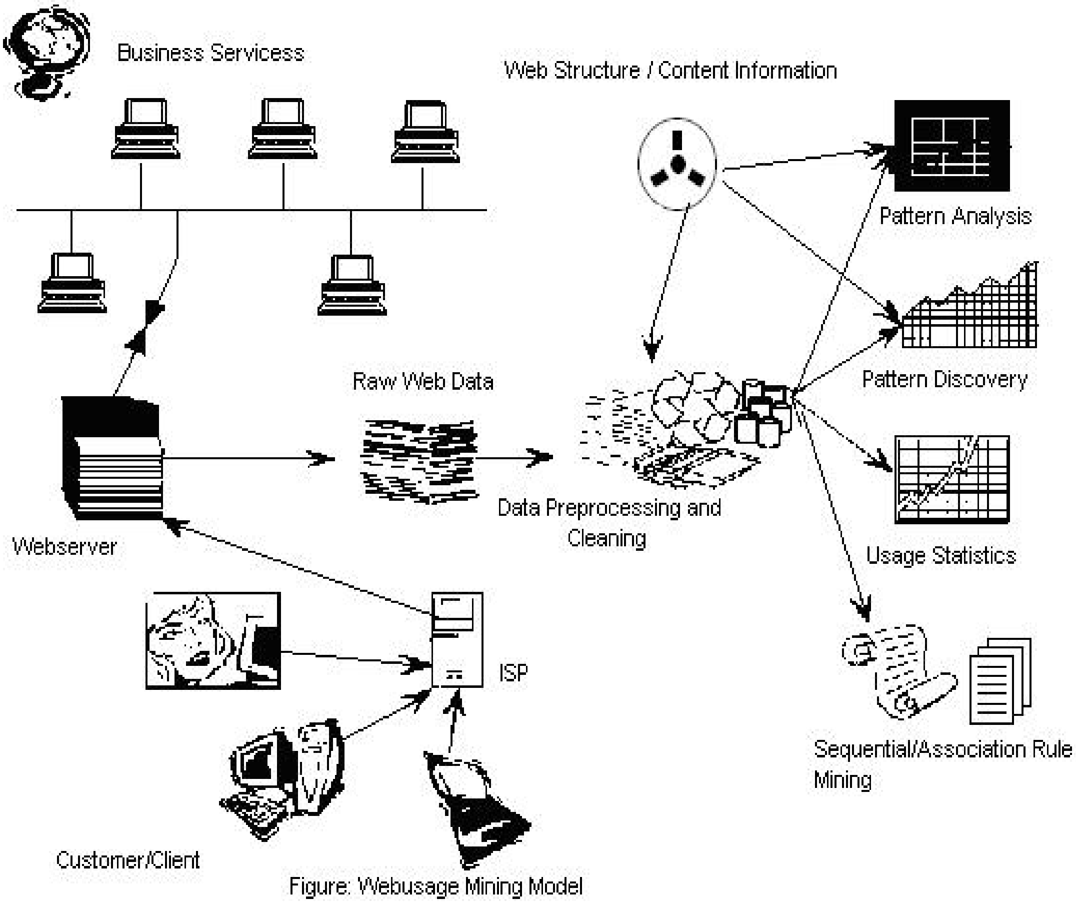

Figure: Webusage Mining Model

7. Data Preprocessing for Mining [2] [7]

Web data is collected in various ways [7], each mechanism collecting attributes relevant for its purpose. There is a need to preprocess the data to make it easier to mine for knowledge, specifically; we believe that issues such as instrumentation and data collection, data integration and transaction identification need to be addressed.

Clearly improved data quality can improve the quality of any analysis on it. A problem in the Web domain is the inherent conflict between the analysis needs of the analysts, who want more detailed usage data collected, and the privacy needs of users, who want as little data collected as possible. This has lead to the development of cookie files on one side and cache busting on the other, the emerging OPS standard on collecting profile data may be a compromise on what can and will be collected. However, it is not clear how much compliance to this can be expected. Hence, there will be a continual need to develop better instrumentation and data collection techniques, based on whatever is possible and allowable at any point in time

Web usage data collected in various logs is at a very fine granularity, Therefore, while it has the advantage of being extremely general and fairly detailed, it also has the corresponding drawback that it cannot be analyzed directly, since the analysis may start focusing on micro trends rather than on the macro trends, On the other hand, the issue of whether a trend is micro or macro depends on the purpose of a specific analysis.

Hence, we believe there is a need to group individual data collection events into groups, called Web transactions, before feeding it to the mining system.

8. The Mining Process [5][8][9]

The key component of Web mining is the mining process itself. The Web mining has adapted techniques from the field of data mining, databases, and information retrieval, as well as developing some techniques of its own.

Web mining studies reported to data have mined for association rules, temporal sequences, clusters, and path expressions. As the manner in which the Web is used continues to expand, there is a continual need to figure out new kinds of knowledge about user behavior that needs to be mined.

The quality of a mining algorithm can be measured both in terms of how effective it is in mining for knowledge and how efficient it is in computational terms. There will always be a need to improve the performance of mining algorithms along both these dimensions.

The data collection on the Web is incremental in nature. Hence, there is a need to develop mining algorithms that take as input the existing data mined knowledge, and the new data, and develop a new model in an efficient manner.

The data collection on the Web is also distributed by its very nature. If all the data were to be integrated before mining, a lot of valuable information could be extracted.

9. Analysis of Mined Knowledge [3]

The output of knowledge mining algorithms is often not in a form suitable for direct human consumption, and hence there is a need to develop techniques and tools for helping an analyst better assimilate it. Issues that need to be addressed in this area include usage analysis tools and interpretation of mined knowledge.

There is a need to develop tools which incorporate statistical methods, visualization, and human factors to help better understand the mined knowledge.

In general one of the open issues in Web mining in particular, is the creation of intelligent tools that can assist in the interpretation of mined knowledge. Clearly, these tools need to have specific knowledge about the particular problem domain to do any more than filtering based on statistical attributes of the discovered rules or patterns. In Web mining, for example, intelligent agents could be developed that based on discovered access patterns, the topology of the Web locality, and certain heuristics derived from user behavior models, could give recommendations about changing the physical link structure of a particular site.

10. Conclusion

The term Web mining has been used to refer to techniques that encompass a broad range of issues However, while meaningful and attractive. This very broadness has caused Web mining to mean different things to different people and there is a need to develop a common vocabulary. Towards this goal we proposed a definition of Web mining and developed taxonomy of the various ongoing efforts related to it. Next, we presented a survey of the research in this area. We provided a detailed survey of the efforts in this area, even though the survey is short because of the area's newness. This paper is useful for researcher exclusively for doing research on web mining.

11. References

[1] Arun k Pujari, "Data Mining Techniques", University press, edition 2001.

[2] Jaiwei Han, Michelinne Kamber, "Data Mining: Concepts and Techniques ".

[3] Margaret H. Dunham,"Data Mining: Introductory and advanced topics".

[4] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P, Web usage mining: Discovery and applications of usage patterns from Web data, SIGKDD Explorations, Vol. 1(2), 12-23, 2000.

[5] Knowledge Discovery from Web Usage Data: Extraction of Sequential Patterns through ART1 Neural Network based Clustering Algorithm, International Conference on Computational Intelligence and Multimedia Applications 2007

[6] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar .,Web Mining – Accomplishments & Future Directions

[7] Mining Web logs for Prediction in Prefetching and Caching - Third 2008 International Conferences on Convergence and Hybrid Information Technology.

[8] Ajith Abraham – Business Intelligence from Web Usage Mining - Journal of Information & Knowledge Management, Vol. 2, No. 4 (2003) 375{390c iKMS & World Scientific Publishing Co.

[9] Adel T. Rahmani and B. Hoda Helmi, EIN-WUM an AIS-based lgorithm for Web Usage Mining, Proceedings of GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA, ACM978-1-60558-130-9/08/07 (Pages 291-292)

[10] Dr. G. K. Gupta, Introduction to Data Mining with Case Studies, PHI Publication.

[11] http://web.media.mit.edu/~lieber/Lieberary/Letizia/Letizia-Intro.html

[12] http://en.wikipedia.org/wiki/Web_crawler.

[13] Prefetching based on Web Usage Mining - Daby M. Sow, David P. Olshefski, Mandis Beigi, and Guruduth Banavar, IBM T. J. Watson Research Center, Hawthorne NY, 10532, USA.

AUTHORS' BIBLIOGRAPHY

YOGISH H. K received his Bachelor's Degree in Computer Science and Engineering from PES College of Engineering, Mandya, Mysore University, Karnataka, India during the year 1998 and M. Tech in Computer Engineering from Sri Jaya Chama Rajendra College of Engineering Mysore, Karnataka, India during the year 2004. Currently pursuing PhD degree in Bharathiar University, Coimbatore. He has total 13 years of Industry and Teaching experience. His areas of interests are Data Warehouse, multimedia, Databases and Operating Systems. He has published and presented various papers in Journals, National Conferences and an author of two text books.

Dr. G T Raju received his Bachelor's Degree in Computer Science and Engineering from Kalpataru Institute Of technology, Tiptur, Karnataka, India, during the year 1992 and M. E in Computer Science and Engineering from B.M.S College of Engineering, Bangalore, Karnataka, India during the year 1995 and Doctorate of Philosophy Ph.D. in the year 2008 in Computer Science and Engineering from Visveswaraya Technological University, Belgaum, Karnataka; He has 18 years of Experience. He has visited overseas to various Universities. His area of interests is Data Mining, Data Warehousing, Image Processing, Databases, Artificial Intelligence and Computer Graphics. He has published and presented papers in journals, international and national level Conferences and published a text book.

Manjunath T N. received his Bachelor's Degree in Computer Science and Engg from Bangalore, University, Karnataka, India during The year 2001 and M. Tech in Computer Science and Engineering VTU, Belgaum, Karnataka, India during the year 2004. Currently pursuing PhD degree in Bharathiar University, Coimbatore.
He has having total 10 years of Industry and teaching experience. His areas of interests are Data Warehouse & Business Intelligence, multimedia and Databases. He has published and Presented several papers.