An Overview of Web Usage Mining

Kamal, Assistant Professor, BRCMCET, Bahal (Hry.)
Dinesh Kumar, Assistant Professor, BRCMCET, Bahal (Hry.)

ABSTRACT

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data relates with the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining is deals with the discovery and analysis of usage patterns from Web data, specifically 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 these information for the specific needs.

Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data. Web server data correspond to the user level logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time as well as pattern of the users and is the main input to the present research. Web usage mining consists of three phases, namely pre-processing, pattern discovery, and pattern analysis.

This paper describes these phases in detail and the current applications of the technique.

Keywords: User/Session identification, Web Recommender, Web log, Web Usage Mining

1. Introduction

Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access. It has now become one of the most important areas in Computer and Information Sciences because of its direct applications in e-commerce, CRM, Web analytics, information retrieval at a faster pace. According to the differences of the mining objects, there are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions [4]. Web document text mining, resource discovery based on concepts indexing or agent; based technology may also fall in this category. 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. This paper provides an up-to-date survey of Web Usage mining, including both academic and industrial research. Section 2 describes the various kinds of Web data that can be useful for Web Usage mining. Section 3 discusses the challenges involved in discovering usage patterns from Web data. The three phases are preprocessing, pattern discovery, and patterns analysis. The paper delivers a better application of web usage mining and a taxonomical survey.

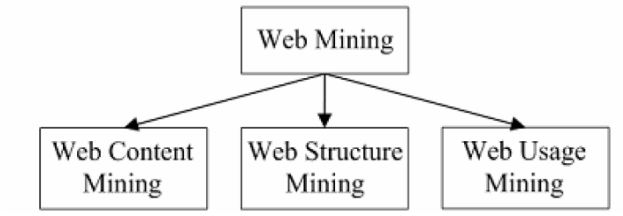


Figure 1.0 : Taxonomy of Web Mining

2. Web Usage Mining

2.1. Concept of web usage mining

Web usage mining is the process of extracting useful information from server logs e.g. users' history. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

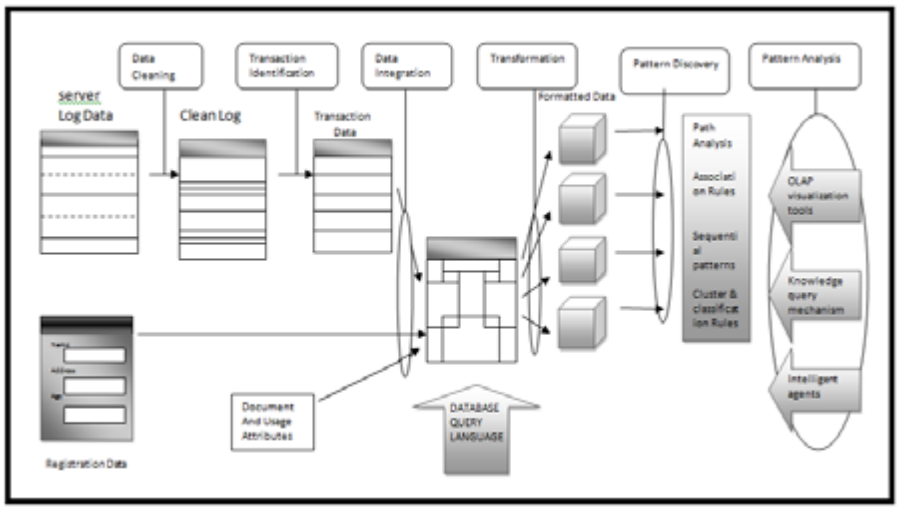
- Web Server Data: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.
- Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
- Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

2.2. Web Log Format

A **server log** is a log file (or several files) automatically created and maintained by a server of activity performed by it. Also anyone can see its History and visited places. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. A common log format file is hence forth created by the web server to keep track of the requests that occur on a web site.

2.3. Approach of Web usage mining

The whole procedure of using Web usage mining for Web recommendation consists of three steps, i.e. data collection and pre-processing, pattern mining (or knowledge discovery) as well as knowledge application. Fig 1.1 depicts the architecture of the web usage mining.



Fog2.0 : The Architecture of Web usage mining

2.3.1 Data collection:

Data collection is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. Therefore it must use scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data (agent server data and package detecting).

A Web server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data which is recorded in server logs contains the access of a Web site by multiple users. However, the site usage data recorded by server logs may not be entirely reliable due to the presence of various levels of caching within the Web environment. Cached page views are not recorded in a server log. A Web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy caching can be used to reduce the loading time of a Web page experienced by users as well as the networkload. at the server and client sides. The performance of proxy caches depends on their ability to predict future page requests correctly. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers.

2.3.2 Data preprocessing:

Some databases are insufficient, inconsistent and including noise. The data pretreatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pretreatment work, mainly include data cleaning, user identification, session identification and path completion.

A) Data Cleaning:

The purpose of data cleaning is to eliminate irrelevant(noisy data) items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. Following two kinds of records are unnecessary and should be removed:

1. The records of graphics, videos and the format information The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record;
2. The records with the failed HTTP status code.

B) User and Session Identification:

The task of user and session identification is to check out the different user sessions from the original web access log. User's identification is, to identify who access web site and which

pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study.

C) Path completion

Another important step in data preprocessing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. Using the local caching and proxy servers also produces the difficulties for path completion because users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log. As a result, the user access paths are incompletely preserved in the web access log.

To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time. It is especially important to web log files, in respect that the structure of web log files are not the same as the data in database or data warehouse. They are not structured and complete due to various causations. So it is especially necessary to pre-process web log files in web usage mining. Through data pre-processing, web log can be transformed into another data structure, which is easy to be mined.

2.3.3 Knowledge Discovery

Uses the statistical method to carry on the analysis and mine the pretreated data. 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 significance and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

2.3.4 Pattern analysis

Challenges of Pattern Analysis is to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse. Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

Pattern Discovery

3. TAXONOMY AND PROJECT SURVEY

Since 1996 there have been several research projects and commercial products that have analyzed Web usage data for a number of different purposes. This section describes the dimensions and application areas that can be used to classify Web Usage Mining projects.

3.1 Taxonomy Aspects

While the number of candidate dimensions that can be used to classify Web Usage Mining projects is many, there are five major aspects that apply to every project.

The data sources used to gather input, the types of input data, the number of users represented in each data set, the number of Web sites represented in each data set, and the application area focused on by the project. The algorithms for a project can be designed to work on inputs representing one or many users and one or many Web sites. Single user projects are generally involved in the personalization application area. The projects that provide multi-site analysis use either client or proxy level input data in order to easily access usage data from more than one Web site. Most Web Usage Mining projects take single-site, multi-user, serverside usage data (Web server logs) as input.

3.2 Project Survey

Projects such as [14; 16] have focused on Web Usage Mining in general, without extensive tailoring of the process towards one of the various sub-categories. The Web SIFT project is discussed in more detail in the next section. Chen et al. [13] introduced the concept of maximal forward reference to characterize user episodes for the mining of traversal patterns. A maximal forward reference is the sequence of pages requested by a user up to the last page before backtracking occurs during a particular server session. The Speed Tracer project from IBM Watson is built on the work originally reported in [13]. In addition to episode identification, Speed Tracer makes use of referrer and agent information in the preprocessing can be used to classify Web Usage Mining projects.

3.1 Taxonomy Aspects

While the number of candidate dimensions that can be used to classify Web Usage Mining projects is many, there are five major aspects that apply to every project. The data sources used to gather input, the types of input data, the number of users represented in each data set, the number of Web sites represented in each data set, and the application area focused on by the project. The algorithms for a project can be designed to work on inputs representing one or many users and one or many Web sites. Single user projects are generally involved in the personalization application area. The projects that provide multi-site analysis use either client or proxy level input data in order to easily access usage data from more than one Web site. Most Web Usage Mining projects take single-site, multi-user, serverside usage data (Web server logs) as input.

3.2 Project Survey

Projects such as [14; 16] have focused on Web Usage Mining in general, without extensive tailoring of the process towards one of the various sub-categories. The Web SIFT project is discussed in more detail in the next section. Chen et al. [13] introduced the concept of maximal forward reference to characterize user episodes for the mining of traversal patterns. A maximal forward reference is the sequence of pages requested by a user up to the last page before backtracking occurs during a particular server session. The Speed Tracer project from IBM Watson is built on the work originally reported in [13]. In addition to episode identification, routines to identify users and server sessions in the absence of additional client side information. The Web Utilization Miner (WUM) system provides a robust mining language in order to specify characteristics of discovered frequent paths that are interesting to the analyst. In their approach, individual navigation paths, called trails, are combined into an aggregated tree structure. Queries can be answered by mapping them into the intermediate nodes of the tree structure. Han et al. [18] have loaded Web server logs into a data cube structure in order to perform data mining as well as On- Line Analytical Processing (OLAP) activities such as roll-up and drill-down of the data. Their WebLogMiner system has been used to discover association rules, perform classification and time- series analysis (such as event sequence analysis, transition analysis and trend analysis). Shahabi et. al. [53;59] have one of the few Web Usage mining systems that relies on client side data collection. The client side agent sends back page request and time information to the server every time a page containing the Java applet (either a new page or a previously cached page) is loaded or destroyed.

3.2.1 Personalization

Personalizing the Web experience for a user is the holy grail of many Web-based applications, e.g. individualized marketing for e-commerce[4]. Making dynamic recommendations to a Web user, based on her/his profile in addition to usage behaviour is very attractive to many applications, e.g. cross-sales and up-sales in e-commerce. Web usage mining is an excellent approach for achieving this goal, as illustrated in [4] Existing recommendation systems, such as [8; 6], do not currently use data mining for

recommendations, though there have been some recent proposals [16]. The Web Watcher, SiteHelper , Letizia, and clustering work by Mobasher et. al.and Yan et. al.[17;4] have all concentrated on providing Web Site personalization based on usage information. Web server logs were used by Yan et. al. to discover clusters of users having similar access patterns. The system proposed in [17] consists of an online module that will perform cluster analysis and an online module which is responsible for dynamic link generation of Web pages. Every site user will be assigned to a single cluster based on their current traversal pattern. The links that are presented to a given user are dynamically selected based on what pages other users assigned to the same cluster have visited.

3.2.2 System Improvement

Performance and other service quality attributes are crucial to user satisfaction from services such as databases, net- works, etc. Similar qualities are expected from the users of Web services. Web usage mining provides the key to under- standing Web tracer behavior, which can in turn be used for developing policies for Web caching, network transmission [11], load balancing, or data distribution. Security is an acutely growing concern for Web-based services, especially as electronic commerce continues to grow at an exponential rate [14]. Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted breakings, etc. Almeida et al.propose models for predicting the locality, both temporal as well as spatial, amongst Web pages requested from a particular user or a group of users accessing from the same proxy server. The locality measure can then be used for deciding pre-fetching and caching strategies for the proxy server,

3.2.3 Site Alteration

The attractiveness of a Web site, in terms of both content and structure, is crucial to some recent proposals [16]. The Web Watcher, SiteHelper , Letizia, and clustering work by Mobasher et. al.and Yan et. al.[17;4] have all concentrated on providing Web Site personalization based on usage information. Web server logs were used by Yan et. al. to discover clusters of users having similar access patterns. The system proposed in [17] consists of an online module that will perform cluster analysis and an online module which is responsible for dynamic link generation of Web pages. Every site user will be assigned to a single cluster based on their current traversal pattern. The links that are presented to a given user are dynamically selected based on what pages other users assigned to the same cluster have visited.

3.2.2 System Improvement

Performance and other service quality attributes are crucial to user satisfaction from services such as databases, net- works, etc. Similar qualities are expected from the users of Web services. Web usage mining provides the key to under- standing Web tracer behavior, which can in turn be used for developing policies for Web caching, network transmission [11], load balancing, or data distribution. Security is an acutely growing concern for Web-based services, especially as electronic commerce continues to grow at an exponential rate [14]. Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted breakings, etc. Almeida et al.propose models for predicting the locality, both temporal as well as spatial, amongst Web pages requested from a particular user or a group of users accessing from the same proxy server. The locality measure can then be used for deciding pre-fetching and caching strategies for the proxy server,

3.2.3 Site Alteration

The attractiveness of a Web site, in terms of both content and structure, is crucial to many applications, e.g. a product catalog for e-commerce. Web usage mining provides detailed feedback on user behavior, providing the Web site designer information on which to base redesign decisions. While the results of any of the projects could lead to re-designing the structure and content of a site, the adaptive Web site project (SCML algorithm) [48; 49]

focuses on automatically changing the structure of a site based on usage patterns discovered from server logs. Clustering of pages is used to determine which pages should be directly linked.

3.2.4 Business Intelligence

Information on how customers are using a Web site is critical information for marketers of e-tailing businesses. Buchner et al [22] have presented a knowledge discovery process in order to discover marketing intelligence from Web data. They define a Web log data hypercube that will consolidate Web usage data along with marketing data for e-commerce applications. They identified four distinct steps in customer relationship life cycle that can be supported by their knowledge discovery techniques: customer attraction, customer retention, cross sales and customer departure. There are several commercial products, such as SurfAid [11], Accrue [1], Net-Genesis [7], Aria [3], Hitlist [5], and WebTrends [13] that provide Web track analysis mainly for the purpose of gathering business intelligence. Accrue, NetGenesis, and Aria are designed to analyze e-commerce events such as products bought and advertisement click-through rates in addition to straight forward usage statistics. Accrue provides apath analysis visualization tool and IBM's SurfAid provides OLAP through a data cube and clustering of users in addition to page view statistics. Padmanabhan et. al. [46] use Web server logs to generate beliefs about the access patterns of Web pages at a given Web site. Algorithms for finding interesting rules based on the unexpectedness of the rule were also developed.

3.2.5 Usage Characterization

While most projects that work on characterizing the usage, content, and structure of the Web don't necessarily consider themselves to be engaged in data mining, there is a large amount of overlap between Web characterization re- search and Web Usage mining. Catledge et al. [12] discuss the results of a study conducted at the Georgia Institute of Technology, in which the Web browser. Xmosaic was modified to log client side activity. The results collected provide detailed information about the user's interaction with the browser interface as well as the navigational strategy used to browse a particular site. The project also provides detailed statistics about occurrence of the various client side events such as the clicking the back/forward buttons, saving tale, adding to bookmarks etc. Pitkow et al. [36] propose a model which can be used to predict the probability distribution for various pages a user might visit on a given site. This model works by assigning a value to all the pages on a site based on various attributes of that page.

4. CONCLUSIONS

This paper has been an attempt to provide an upto- date survey of the rapidly growing area of Web Usage mining. With the rapid and intense growth of Web-based applications, specifically in field of electronic commerce, there is a significant interest in analyzing Web usage data to better understand Web usage, and apply the knowledge to better serve users. This has led to a number of commercial offerings for doing such analysis. However, Web Usage mining raises some probabilistic issues and some scientific questions that should be answered before robust tools can be developed. This article has been destined to enumerate such challenges, and the hope is that the research community will take up the challenge of addressing them.

5. REFERENCES

[1] Agrawal R., Imielinski T., and Swami A. (1993). Mining Associations between sets of items in Massive Databases. In Proceeding of the ACM-SIGMOD International Conference on Management of Data, pp. 207-216, Washingtom D.c USA.

[2] Kosala R., Blockeel H., (2000). Web mining research: a survey. SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2(1), pp. 1–15.
[3] Cooley R., Mobasher B., & Srivastava J. (1997). Web mining: Information and pattern discovery on the World Wide Web. In Proceeding of the IEEE International Conference on Tools with AI. pp. 558-567.
[4] Masseglia F., Poncelet P., and Teisseire M. (1999). Using data mining techniques on web access logs to dynamically improve hypertext structure. In ACM SigWeb Letters, 8(3): pp. 13-19.
[5] Zhang Huiying, Liang Wei.An (2004). Intelligent Algorithm of Data Pre-processing in Web Usage Mining. In Proceeding of the 5th World Congress on Intelligent Control and Automation. pp. 15-19. Hangzhou, P.R. China.
[6] Yinghui Yang and Balaji Padmanabhan. (2005). GHIC: A Hierarchical Pattern-Based Clustering Algorithm for Grouping Web Transactions. IEEE Transactions on Knowledge and Data Engineering, Vol 17, No. 9.
[7] Yi Dong, Huiying Zhang and Linnan Jiao. (2006). Research on Application of User Navigation Pattern Mining Recommentation. In Proceeding. of the 6th World Cogress on Intelligent Control and Automation. Dalian, China.
[8] Hannah Inbarani H., Thangavel K., and Pethalakshmi A. (2007). Rough Set based Feature Selection for Web Usage Mining. International Conference on Computational Intelligence and Multimedia Applications.
[9] Suneetha K. R., and Krishnamoorthi R. (2009). Identifying User Behavior by Analysizing Web Server Access Log File. IJCSNS International Journal of Computer Science and Network Security, Vol 9, No.4.
[10] Manoj Bahel and Chhay Dule. (2010). Analysis of Frequent Itemset generation process in Apriori and RCS (Reduced Candidate Set) Algorithm. International Journal of Advanced Networking and algorithms. Vol 02, Issue 02. pp. 539-543.
[11] Veeramalai S., Jaisankar S., and Kannan A. (2010). Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy. International. Journal of Computer Science and Information Technology (IJCSIT) Vol.2. No.4.
[12] Maja Dimitrijevic and Zita Bosnjak. (2011). Web Usage Association Rule Mining System. Interdiciplinary Journal of Information, Knowledge and Management, Vol 6.
[13] Maja Dimitrijevic and Zita Bosnjak. (2010). Discovering interesting association rules in the web log usage data. Interdisciplinary Journal of Information, Knowledge, and Management, 5,pp. 191-207.