EXPLORATION OF WEBMINER SYSTEM

Navjot Kaur*

Dr. Himanshu Aggarwal*

ABSTRACT

The World Wide Web is an immense source of data. It provides abundance of information for the Internet users. Users' accesses are recorded in web logs. Because of the tremendous usage of web, the web log files are growing at a faster rate and the size is becoming huge. Web usage mining is the area of data mining which 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 includes three process, namely,preprocessing, pattern discovery and pattern analysis.This paper reviews existing work done in the preprocessing stage. Web log data is usually noisy and ambiguous and preprocessing is an important process before mining. This paper presents several data preparation techniques in order to identify unique users and user sessions.Finally an overview of various applications of web usage mining is also presented.

Keywords:*World Wide Web ,data mining, web usage mining, information retrieval, information extraction.*

*Assistant Professor, Department of Computer Engineering, University College of Engineering, Punjabi University Patiala.

**Department of Computer Engineering, University College of Engineering, Punjabi University Patiala.

I INTRODUCTION

Web mining [1] is the application of data mining techniques to discover patterns from the Web. The World Wide Web is an immense source of data that can come either from the Web content, represented by the billions of pages publicly available, or from the Web usage, represented by the log information daily collected by all the servers around the world. Web Mining is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web.

According to analysis targets, web mining can be divided into three different types[2], which areWeb content mining, Web structure mining and Web usage mining. The World Wide Web is an immense source of data that can come either from the Web Content, represented by the billions of pages publicly available or from the Web usage represented by the log information daily collected by all the servers around the world.

Web Usage Mining is often regarded as a part of the Business Intelligence in anorganization rather than the technical aspect. It is used for deciding business strategiesthrough the efficient use of Web Applications. It is also crucial for the CustomerRelationship Management (CRM) as it can ensure customer satisfaction as far as theinteraction 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.

Web Usage mining is the application of data mining techniques to large Web data repositories in order to produce results that can be used in the designing tasks of any website. Some of the data mining algorithms that are commonly used in Web Usage Mining are association rule generation, sequential pattern generation, and clustering. Association Rule mining techniques discover unordered correlations between items found in a database of transactions. In the context of Web Usage Mining a transaction is a group of Web page accesses, with an item being a single page access.

A Web usage mining system performs five major tasks:

- Data gathering,
- Data preparation,

- Navigation pattern discovery,
- Pattern analysis and visualization, and
- Pattern applications.

II DATA SOURCES

In Web Usage Mining, data can be collected in server logs(Access Log, Referrer Log, Agent Log), proxy logs,Web clients or obtained from an organization's database. 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.

There are many kinds of data that can be used in Web Mining.

- **Content:** The visible data in the Web pages or the information which was meant tobe imparted to the users. A major part of it includes text and graphics (images). Typical applications are Content-based categorization and content based ranking of Web pages.
- **Structure:** Data which describes the organization(or Structure) of the website. Source data mainly consist of the structural information present in web pages(e.g., links to other pages); It is divided intotwo types. First is Intra-page structure mining,evaluates the arrangement of the various HTML or XML tags within a page . Second is Inter-page structurerefers to hyper-links connecting one page to another. Applications are link-based categorization of Web pages, ranking of Web pages through a combination of content and structure and reverse engineering of Web site models.
- **Usage:** Data that describes the usage patterns of Web pages. Data is extracted from server log files Usage pattern ,such as nameand IP addresses of the remote host,page references, and the date and time of accesses and various other informationdepending on the log format(Common Log Format[4],Extended Log Format[5], Log ML [6]). Applications are those based on user modelling techniques, such as Web personalization, adaptive Web sites and user modelling.

Web Usage Mining applications are based on data collected from following main sources:

Web Server Logs – These are logs which maintain a history of page requests. TheW3C maintains a standard format for web server log files, but other proprietaryformats exist. More recent entries are typically appended to the end of the file.As shown in Fig1 information about the request[3], include client IP address, request date/time, pagerequested, 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 asan access log, error log, or referrer log. However, server logs typically do not collectuser-specific information. These files are usually

not accessible to general Internetusers, only to the webmaster or other administrative person. When exploiting log information from Web servers, the major issue is the identification ofusers_sessions, i.e., how to group all the users_ page requests so to clearly identify the paths that users followed during navigation through the web site. This task is usually quite difficult and it depends on the type of information available in log files. The most common approach is to use cookies to track down the sequence of users_ page requests (see [7] for an overview of cookie standards). If cookies are not available, various heuristics [8] can be employed to reliably identify users_ sessions. Note however that, even if cookies are used, it is still impossible to identify the exact navigation paths since the use of the back button is not tracked at the serverlevel [9].

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)

Fig 1. Sample Web Server Log

Apart from Web logs, users_ behaviour can also be tracked down on the server side by means ofTCP/IP packet sniffers. Even in this case the identification of users_ sessions is still an issue, but theuse of packet sniffers provides some advantages [10]. In fact: (i) data are collected in real time; (ii)information coming from different Web servers can be easily merged together into a unique log;(iii) the use of special buttons (e.g., the stop button) can be detected so to collect information usuallyunavailable in log files. Notwithstanding the many advantages, packet sniffers are rarely usedin practice. Packet sniffers raise scalability issues on Web servers with high traffic [10].

Proxy Server Logs - A Web proxy is a caching mechanism which lies betweenclient browsers and Web servers. side. Many Internet Service Providers (ISPs) give to their customer proxy server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. 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. The main difference in this case is that proxy servers collects data of groups of users accessing huge groups of web servers.Even in this case, session reconstruction is difficult and not all users_ navigation paths can be identified.However, when there is no other caching between the proxy server and the clients, the identificationof users_ sessions is easier.

Client side– Usage data can be tracked also on the client side by using Javascript, Java applets[11], or even modified browsers [12]. These techniques avoid the problems of users_ sessions identification and the problems caused by caching (like the use of the back button). In addition, they provide detailed information about actual user behaviours[9]. However, these approaches rely heavily on the users cooperation and raise many issues concerning the privacy laws, which are quite strict. Client-side collection scores over server-sidecollection because it reduces both the bot and session identification problems.

III THE WEBMINER SYSTEM

The WEBMINER system [13,14] divides theWeb Usage Mining process into three main parts, as shown in Figs1. Input data consists of the three serverlogs - access, referrer, and agent, the HTML files that make up the site, and anyoptional data such as registration data or remote agent logs. The first part ofWeb Usage Mining, called preprocessing, includes the domain dependent tasksof data cleaning, user identi_cation, session identi_cation, and path completion.Data preprocessing has a fundamental role in Web Usage Mining application. The preprocessing of Web logs is usually complex and time demanding .It comprises of several steps: data collection, data pretreatment, knowledge discovery and pattern analysis.

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

Data preprocessing: Some databases are insufficient, inconsistent andincluding noise. The data pretreatment is to carry on aunification transformation to those databases. The result isthat the database will to become integrate and consistent,thus establish the database which may mine. In the datapretreatment work, mainly include data cleaning, useridentification, session identification and path completion.

- **Data Cleaning:** The purpose of data cleaning is to eliminate irrelevantitems, and these kinds of techniques are of importance forany type of web log analysis not only data mining.According to the purposes of different miningapplications, irrelevant records in web access log will beeliminated during data cleaning. Since the target of WebUsage Mining is to get the user's travel patterns,following two kinds of records are unnecessary andshould be removed:

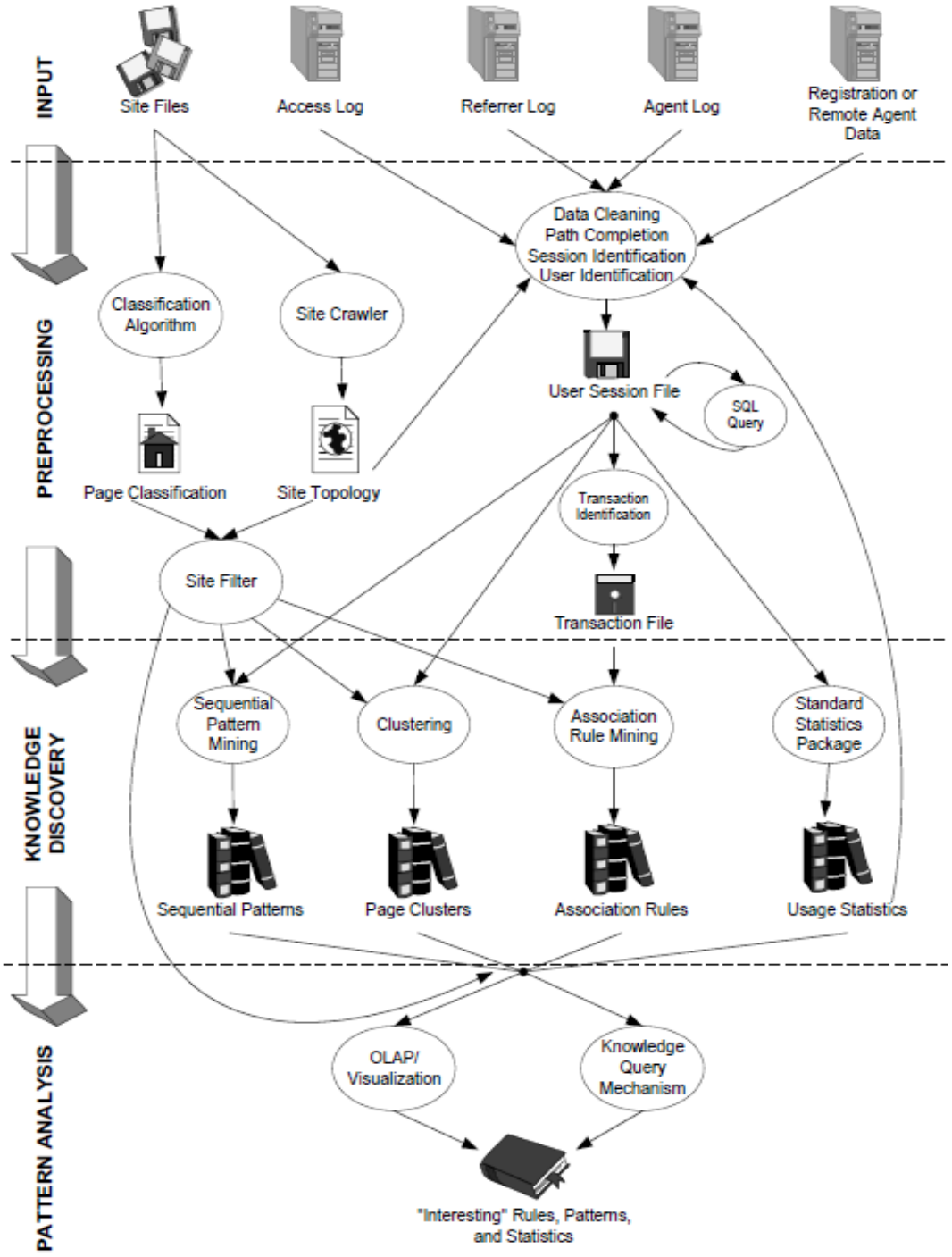

Fig.2 Architecture of WebMiner System

- 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;
- The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

➢ **User and Session Identification:** The task of user and session identification is find out thedifferent user sessions from the original web access log.User's identification is, to identify who access web siteand which pages are accessed. The goal of sessionidentification is to divide the page accesses of each user ata time into individual sessions. A session is a series ofweb pages user browse in a single access. The difficultiesto accomplish this step are introduced by using proxyservers, e.g. different users may have same IP address inthe log. A referrer-based method is proposed to solvethese problems in this study. The rules adopted todistinguish user sessions can be described as follows:

The different IP addresses distinguish different users;

- If the IP addresses are same, the different browsers and operation systems indicate different users;
- If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified if the URL in the Refer URI field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds) between the accessing time of this record and the previous one if theRefer URI field is empty;
- The session identified by rule 3 may contains more than one visit by the same user at different time, the time oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

➢ **Path completion:** Another critical 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 the mined.

Knowledge Discovery: Use statistical method to carry on the analysis and mine the pretreated 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 Association Rules, Clustering, Classification, Sequential Patterns , Dependency Modeling ,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**.**

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.

IV WEB USAGE MINING APPLICATIONS

The general goal of Web Usage Mining is to gather interesting information about users navigation patterns.Users' behavior is used in different applications such as [3]Personalization, e-commerce, to improve the system and toimprove the system design as per their interest etc..

Personalization of web Content . Web Usage mining techniques can be used to providepersonalized web user experience. Webpersonalizationoffers many functions such as

simple usersalutation to more complicate such as content delivery as perusers interests. It is possible to anticipate the user behaviourbyanalyzing the current navigation patterns with patterns whichwere extracted from past web log. Recommendation systemsare the most common application. Personalized sites areexample for recommendation systems.

*E-Commerce*applications need customer details for Customer RelationshipManagement. Mining Business intelligence from Web usage data is dramatically important for e-commerce Web-based companies. Usage mining techniques are very useful tofocus customer attraction, customer retention, cross sales andcustomer departure.

System Improvement is done byunderstanding the web traffic 247ehaviour by mining log data sothat policies are developed for Web caching, load balancing,network transmission and data distribution. Patterns fordetecting intrusion fraud, attempted break-ins are also providedby mining. Performance is improved to satisfy users.

SiteModification is a process of modifying the web site andimproving the quality of design and contents on knowing theinterest of users. Pages are re-linked as per customer 247ehaviour.

V CONCLUSION

This paper presented the preprocessing tasks that are necessary for performing Web usage Mining. The results of mining can be used to improve thewebsite design and increase satisfaction which helps in variousapplications. Log files are the best source of information to know userbehavior. But the raw log files contains unnecessary detailslike image access, failed entries etc., which will affect theaccuracy of pattern discovery and analysis. So preprocessingstage is an important work in mining to make efficient patternanalysis. To get accurate mining results user's session detailsare to be known. The survey was performed on a selection ofweb usage methodologies in preprocessing proposed byresearch community. More concentration is done onpreprocessing stages like session identification and pathcompletion.Thevarious Works done bydifferent researchers are also presented.

REFRENCES

[1]O. Etzioni, The world-wide Web: quagmire or gold mine? Communications of the ACM 39 (11) (1996) 65–68.

[2] R. Kosala, H. Blockeel, Web mining research: a survey, SIGKDD: SIGKDD explorations: newsletter of thespecial interest group (SIG) on knowledge discovery & data mining, ACM 2 (1) (2000) 1–15.

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

[4] Configuration file of W3C httpd, http://www.w3.org/Daemon/User/Config/ (1995).

[5] W3C Extended Log File Format, http://www.w3.org/TR/WD-logfile.html (1996).

[6] J.R. Punin, M.S. Krishnamoorthy, M.J. Zaki, Logml: Log markup language for web usage mining, in: R. Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava (Eds.), WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised Papers, vol. 2356 of Lecture Notes in Computer Science, Springer, 2002, pp. 88–112

[7] D.M. Kristol, Http cookies: standards, privacy, and politics, ACM Transactions on Internet Technology (TOIT) 1 (2) (2001) 151–198.

[8] B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou, The impact of site structure and user environment onsession reconstruction in web usage analysis, in: Proceedings of the 4th WebKDD 2002 Workshop, at the ACMSIGKDDConference on Knowledge Discovery in Databases (KDD_2002), 2002.

[9] K.D. Fenstermacher, M. Ginsburg, Mining client-side activity for personalization, in: Fourth IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS_02), 2002, pp. 205–212..

[10] Pilot Software, Web site analysis, Going Beyond Traffic Analysis http://www.marketwave.com/productssolutions/hitlist.html (2002).

[11] C. Shahabi, F. Banaei-Kashani, A framework for efficient and anonymous web usage mining based on client-side tracking, in: R. Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava (Eds.), WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised papers, vol. 2356 of Lecture Notes in Computer Science, Springer, 2002, pp. 113–144. F.M. Facca, P.L. Lanzi / Data & Knowledge Engineering 53 (2005) 225–241 237

[12] L.D. Catledge, J.E. Pitkow, Characterizing browsing strategies in the World-Wide Web, Computer Networks and ISDN Systems 27 (6) (1995) 1065–1073

[13] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the World Wide Web. In International Conference on Tools with Arti_cial Intelligence, pages 558{567, Newport Beach, CA, 1997

[14] B. Mobasher, N. Jain, E. Han, and J. Srivastava.Web Mining: Pattern discovery from World Wide Web transactions.Technical Report TR 96-050, University of Minnesota, Dept. of Computer Science, Minneapolis, 1996.