International Journal of Electronics Communication and Computer Engineering
Volume 2, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Web Mining: An Overview

P.V.G.S.Mudiraj[1],B. Jabber[2], K.David raju[3]

1 Associate Professor, Adams Engg. College,Paloncha,Khammam,A.P.
2 Associate Professor, Dept.of CSE,Pulipati Prasad college of Engg &Technology,Khammam,A.P
3 Associate Professor, Dept.of CSE,St.Peters Engg. College, Hyderabad, A.P, India

Abstract

Web usage mining is a main research area in Web mining focused on learning about Web users and their interactions with Web sites. The motive of mining is to find users' access models automatically and quickly from the vast Web log data, such as frequent access paths, frequent access page groups and user clustering. Through web usage mining, the server log, registration information and other relative information left by user provide foundation for decision making of organizations. This article provides a survey and analysis of current Web usage mining systems and technologies. There are generally three tasks in Web Usage Mining: Preprocessing, Pattern analysis and Knowledge discovery. Preprocessing cleans log file of server by removing log entries such as error or failure and repeated request for the same URL from the same host etc... The main task of Pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users. The statistics collected from the log file can help to discover the knowledge. This knowledge collected can be used to take decision on various factors like Excellent, Medium, Weak users and Excellent, Medium and Weak web pages based on hit counts of the web page in the web site. The design of the website is restructured based on user's behavior or hit counts which provides quick response to the web users, saves memory space of servers and thus reducing HTTP requests and bandwidth utilization. This paper addresses challenges in three phases of Web Usage mining along with Web Structure Mining.This paper also discusses an application of WUM, an online Recommender System that dynamically generates links to pages that have not yet been visited by a user and might be of his potential interest. Differently from the recommender systems proposed so far, **ONLINE MINER** does not make use of any off-line component, and is able to manage Web sites made up of pages dynamically generated.

Keywords: User/Session identification, Web Recommender, Web log.

1. INTRODUCTION

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. In order to better serve for the users, web mining applies the data mining, the artificial intelligence and the chart technology and so on to the web data and traces users' visiting characteristics, and then extracts the users' using pattern[l]. It has quickly 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 and filtering, and Web information systems.

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. 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 of extracting interesting patterns in web access logs.

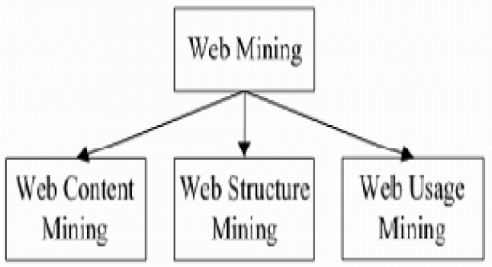

Figure 1: Taxonomy of Web Mining

2. WEB USAGE MINING

2.1. CONCEPT OF WEB USAGE MINING

Discovery of meaningful patterns from data generated by client-server transactions on one or more Web servers. Typical Sources of data:

1. Automatically generated data stored in server access logs, referrer logs, agent logs, and
client-side cookies
2. E-commerce and product-oriented user events (e.g. shopping cart changes, ad or product click-throughs, etc.)
3. User profiles and/or user ratings
4. Meta-data, page attributes page content, site structure

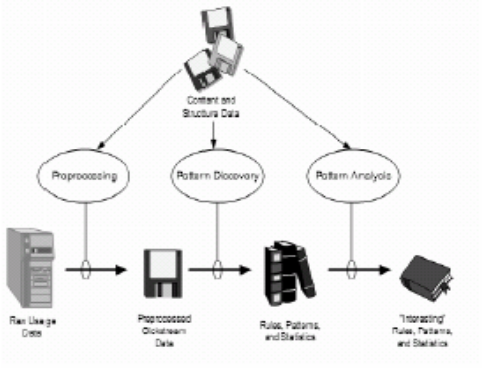

Figure 2: Web Usage Mining Process

2.2. WEB LOG FORMAT

A web server log file contains requests made to the web server, recorded in chronological order. The most popular 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.

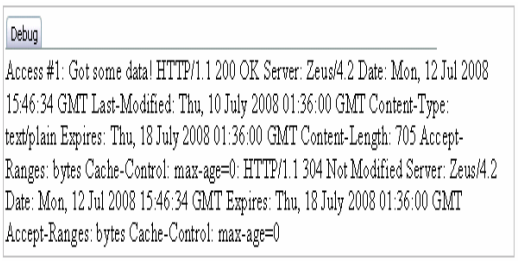

Figure 3: Example of typical server log

2.3. APPROACH OF WEB USAGE MINING

The web usage mining generally includes the following several steps: data collection, data pretreatment, knowledge discovery and pattern analysis.

A] 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).

B] 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 isthat 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.

1) Data Cleaning:

The purpose of data cleaning is to eliminate irrelevant 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. Since the target of Web Usage Mining is to get the user's travel patterns, 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. 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.

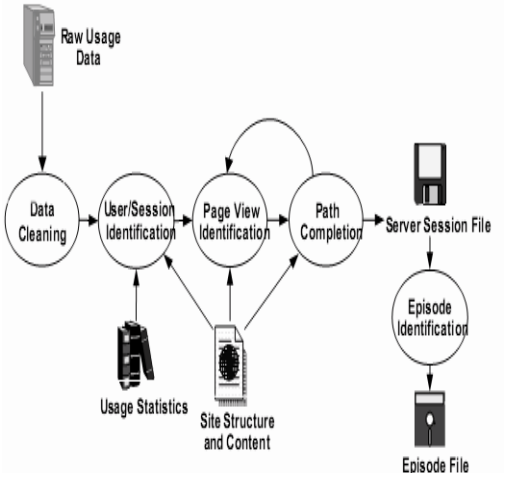

Figure 4 : Preprocessing of Web Usage Data

2) User and Session Identification:

The task of user and session identification is find 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. The rules adopted to distinguish user sessions can be described as follows:

a. The different IP addresses distinguish different users;
b. If the IP addresses are same, the different browsers and operation systems indicate different users;
c. 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 the Refer URI field is empty;
d. The session identified by rule 3 may contains more than one visit by the same user at different time, the timeoriented 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.

3) 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 he mined.

C] 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 clustering, classifying, 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.

D] 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.

2.3 ONLINE WEB PERSONALIZATION SYSTEM

The main limitation of traditional Personalization systems is the loosely coupled integration of the Web personalization system with the Web server ordinary activity. A new web usage-mining tool named **"Online Miner"** which is a user-friendly tool collects the user behavior and stores them in the respective category defined by the administrator, with no complicated queries it generates rapid reports and also maintains accuracy in reports. The key idea of **Online Miner** is to collect the required data from the live source of user behavior on the web with the help of dynamic configuration of filters and transfers data by applying transformations, in to the web usage repository for generating reports with no complicated queries and less processing time. The tool behavior is inherited from Cluster Model and implements the customized usage tracking trend**.**

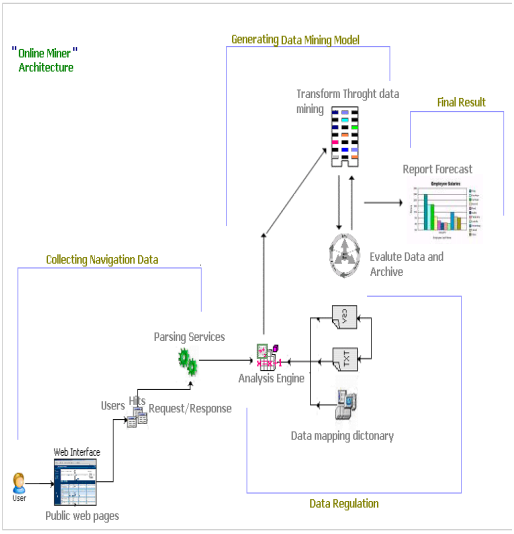

Figure 5: Online Miner High Level Architecture

The proposed Online Miner architecture covers all the activities of Web Usage mining ranging from collecting the navigational data, Data Regulation, Generating Data Mining Model, Final Report generation.

3. IMPLEMENTATION

Online Miner is an add-on software component that can inspect traffic at a deeper level than any other web-mining tool does. It is a software component that can be hosted online and can inspect the data before it allows to the web-mining repository. This tool starts its activities of gathering, filtering and categorization of data when the user moves or clicks the mouse button or key in the data into the web pages. This tool provides the transparency to transactional analysis on user behavior. Online Miner is an Asp.net technology based frame work with C# coding to avoid common problems associated with processing Server Logs and to capture additional and more detailed data. The core elements of this frame work are clearly mentioned in the Online miner architecture.

In order to generate data mining model, Online miner uses a sophisticated algorithm known as " **Mining Repository Algorithm"** .

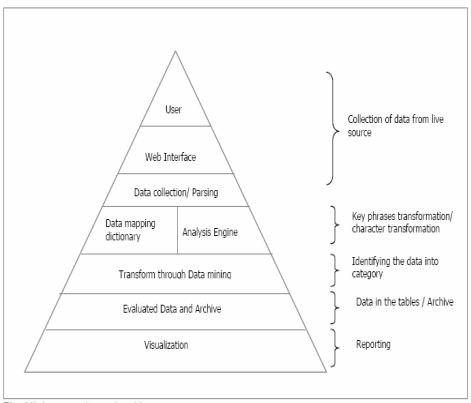

Fig: Mining repository algorithm

4. CONCLUSION

Web usage mining model is a kind of mining to server logs. Web Usage Mining plays an important role in realizing enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. This paper has attempted to cover all the activities of the rapidly growing area of Web usage mining. The proposed frame work "Online Miner "seems to work well for developing prediction models to analyze the web traffic volume. However ,Web usage mining raises some hard scientific questions that must answered before robust tools can be developed. Web usage patterns and data mining will be the basis for a great deal in future research.. Future research will also incorporate data mining algorithms to improve knowledge discovery.

5. REFERENCES

1. Qingtian Han, Xiaoyan Gao, Wenguo Wu, "Study on Web Mining Algorithm Based on Usage Mining", Computer- Aided Industrial Design and Conceptual Design, 2008. CAID/CD 2008. 9th International Conference on 22-25 Nov. 2008

2. Qingtian Han, Xiaoyan Gao, "Research of Distributed Algorithm Based on Usage Mining", Knowledge Discovery and Data Mining, 2009, WKDD 2009, Second International Workshop on 23-25 Jan. 2009

3. Ranieri Baraglia and Fabrizio Silvestri, "An Online Recommender System for LargeWeb Sites", Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on 20-24 Sept. 2004

4 .Yan Li, Boqin Feng, Qinjiao Mao, "Research on Path Completion Technique in Web Usage Mining", Computer Science and Computational Technology, 2008. ISCSCT '08. International Symposium on Volume 1, 20-22 Dec. 2008 5 Yi Dong, Huiying Zhang, Linnan Jiao, "Research on Application of User Navigation Pattern Mining Recommendation", Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress ,Volume 2.

5 .Prof K.Mrithyumjaya Rao, D.Suresh Babu, Online Miner- A tool for discovering business intelligence from web data. Library Progress (International) Vol.29(No.1) (P.33-56),2009.

6. Cooley R. (2000), Web Usage Mining: Discovery and Application of Interesting patterns from Web Data, Ph. D. Thesis, Department of Computer Science, University of Minnesota.

7 Spiliopoulou, M., Faulstich, L.C. (1999): WUM: A Web Utilization Miner. Proceedings of EDBT Workshop on the Web and Data Bases (WebDB'98), Springer Verlag, pp. 109-115.

8 Mobasher B., Cooley R. and Srivastava J. (1999), Creating Adaptive Web Sites through Usage-based Clustering of URLs, In Proceedings of 1999 Workshop on Knowledge and Data Engineering Exchange, USA, pp.19-25.

9 Olfa Nasraoui and Christopher Petens: Combining Web usage Mining and Fuzzy Inference for Website personalization.