ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
*Published By:* **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
March 2013,
Vol. 6, No. (1):
Pgs. 119-124

# Web Usage Mining and Business Intelligence

**H.B. BASANTH KUMAR**

Pooja Bhagavat Memorial Mahajana PG Centre, DoS in CS, KRS Road, Metagalli, Mysore, India.

(Received: March 02, 2013; Accepted: March 10, 2013)

**ABSTRACT**

Now a day's World Wide Web has become very popular and interactive for transferring of information. The web is huge, diverse and active and thus increases the scalability, multimedia data and temporal matters. The growth of the web has outcome in a huge amount of information that is now freely offered for user access. The several kinds of data have to be handled and organized in a manner that they can be accessed by several users effectively and efficiently. So the usage of data mining methods and knowledge discovery on the web is now on the spotlight of a boosting number of researchers. 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 access can be mined with the user access mode which will provide foundation for decision making of organizations. This paper presents brief overview of web usage mining and business intelligence.

**Key words:** web usage mining, web log, user / session identification, World Wide Web.

## 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. 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[1]. Researchers have identified 3 broad categories of web mining: web content mining, web structure mining and web usage mining.

### Web content mining

Web content mining is the application of data mining techniques to content published on the internet, usually as semi structured, unstructured, structured documents. Structured data extraction is one of most widely studied research topics of Web content mining. Structured data on the Web are often very important as they

represent their host pages essential information. Extracting such data allows one to provide value added services, e.g. shopping and meta-search. In contrast to unstructured texts, structured data is also easier to extract. This problem has been studied by researchers in AI, database and data mining.

**Web structure mining**

web structure mining operates on the web's hyperlink structure. This graph structure can provide information about pages ranking or authoritativeness and enhance search results through filtering.

**Web usage mining**

Web usage mining analyses result of user interaction with a web server, including web logs, click streams, and database transactions at a web site or a group of a related sites. By performing analysis on web usage log data, web mining systems can discover knowledge about a systems usage characteristics and users interest. Such knowledge has various applications, such as personalization and collaboration in web based systems, marketing, website design, and web site evaluation and decision support[2].

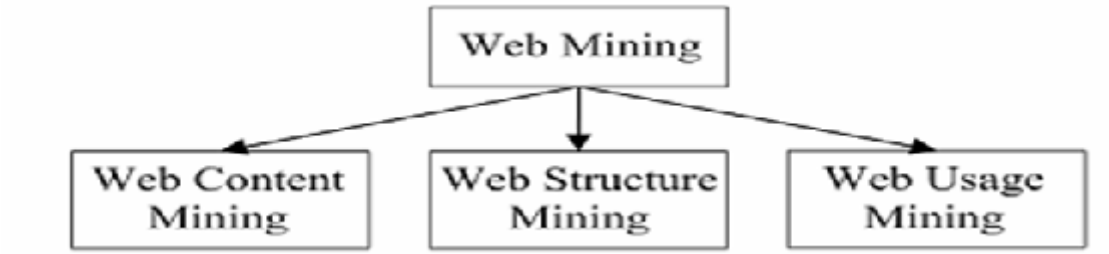


**Fig. 1: taxonomy of Web Mining**

**Requirements of Web Usage Mining**

It is necessary to examine what kind of features a Web usage mining system is expected to have in order to conduct effective and efficient Web usage mining, and what kind of challenges may be faced in the process of developing new Web usage mining techniques. A Web usage mining system should be able to:

- Gather useful usage data thoroughly,
- Filter out irrelevant usage data,
- Establish the actual usage data,
- Discover interesting navigation patterns,
- Display the navigation patterns clearly,
- Analyze and interpret the navigation patterns correctly, and
- Apply the mining results effectively[3].

**Web Usage Mining**

**Concept of web usage mining**

Discovering a meaningful pattern 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.

**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. A standard log file has the following format as shown in Figure 2.

<ip_addr><base_url> - <date><method><file><protocol><code><bytes><referrer><user_agent>

**Fig. 2: Common Web Log Format**

```
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:21 -0600] "GET /Calls/OWOM.html
HTTP/1.0" 200 3942 "http://www.lycos.com/cgi-
bin/pursuit?query=advertising+psychology&maxhits=20&cat=dir" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:23 -0600] "GET
/Calls/Images/earthani.gif HTTP/1.0" 200 10689 "http://www.acr-news.org/Calls/OWOM.html"
"Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:24 -0600] "GET /Calls/Images/line.gif
HTTP/1.0" 200 190 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en] (Win98; I)"

203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:25 -0600] "GET /Calls/Images/red.gif
HTTP/1.0" 200 104 "http://www.acr-news.org/Calls/OWOM.html" "Mozilla/4.5 [en] (Win98; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:31 -0600] "GET / HTTP/1.0" 200 4980 ""
"Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/line.gif
HTTP/1.0" 200 190 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/red.gif
HTTP/1.0" 200 104 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:35 -0600] "GET /Images/earthani.gif
HTTP/1.0" 200 10689 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"

203.252.234.33 www.acr-news.org - [01/Jun/1999:03:32:11 -0600] "GET /CP.html HTTP/1.0" 200
3218 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"
```

**Fig. 3: Example of Server Log**

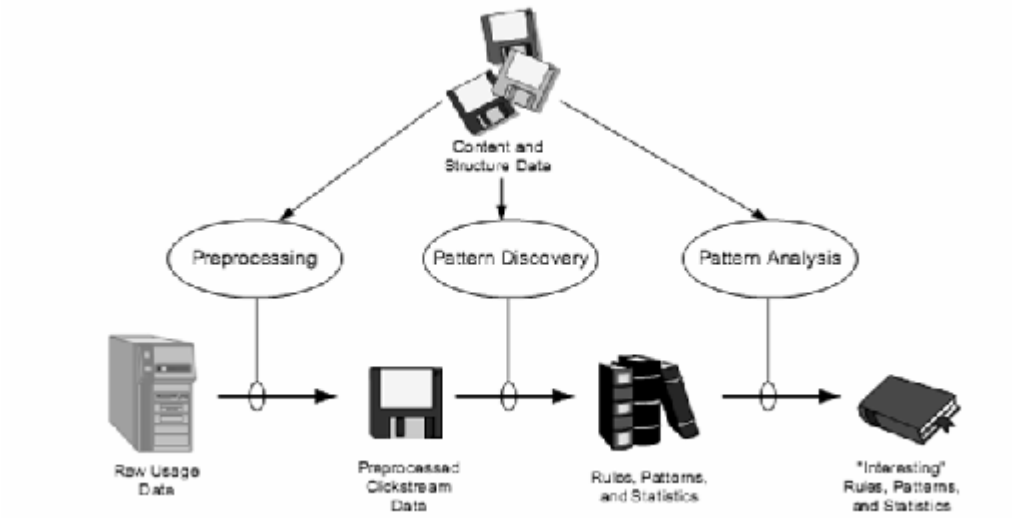


**Fig. 4: Web Usage Mining Architecture**

**Web Usage Mining Architecture**

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).

**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.

**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.

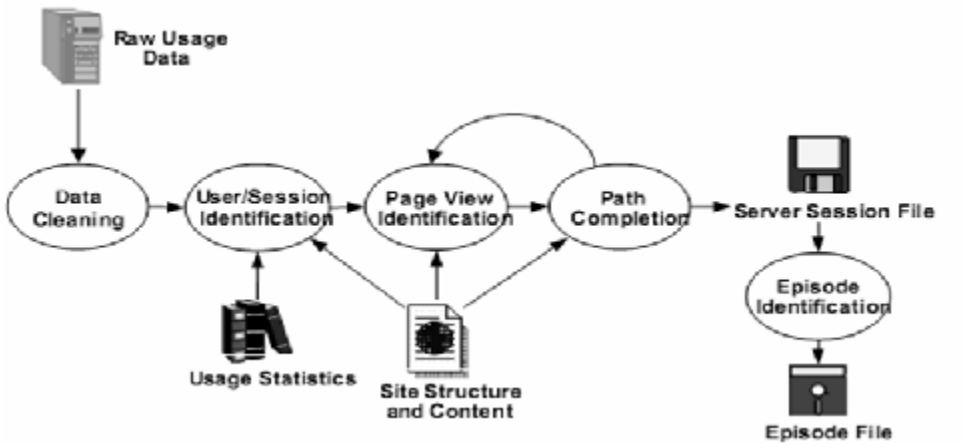


**Fig. 5: preprocessing web usage data**

**User and Session Identification**

The task of user and session identification is to 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 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 be 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 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.

**Pattern analysis**

Challenges of Pattern Analysis are 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.

**Web Usage Mining and Business Intelligence**

The rapid e-commerce growth has made both business community and customers face a new situation. Due to intense competition on the one hand and the customer's option to choose from several alternatives, the business community has realized the necessity of intelligent marketing strategies and relationship management.

Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the Web access logs can help understand the user behavior and the web structure. From the business and applications point of view, knowledge obtained from the web usage patterns could be directly applied to efficiently manage activities related to e-business, eservices and e-education. Accurate web usage information could help to attract new customers, retain current customers, improve cross marketing/sales, effectiveness of promotional campaigns, tracking leaving customers etc.

The usage information can be exploited to improve the performance of Web servers by developing proper prefetching and caching strategies so as to reduce the server response time. User profiles could be built by combining users' navigation paths with other data features, such as page viewing time, hyperlink structure, and page content[5].

What makes the discovered knowledge interesting had been addressed by several works. Results previously known are very often considered as not interesting. So the key concept to make the discovered knowledge interesting will be its novelty or unexpected appearance.

Web Usage Mining techniques can be used to anticipate the user behavior in real time by comparing the current navigation pattern with typical patterns which were extracted from past Web log. Recommendation systems could be developed to recommend interesting links to products which could be interesting to users.

One of the major issues in web log mining is to group all the users' page requests so to clearly identify the paths that users followed during navigation through the web site. The most common approach is to use cookies to track down the sequence of users' page requests or by using some heuristic methods. Session reconstruction is also difficult from proxy server log file data and sometimes not all users' navigation paths can be identified[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.

## REFERENCES

1. Qingtian Han, Xiaoyan Gao, Wenguo Wu, "Study on Web Mining Algorithm Based on Usage Mining", Computer- Aided Industrial Design and Conceptual Design, CAID / CD 2008. 9th International Conference on 22-25 (2008).
2. B.Naveena Devi, "Dynamic modelling approach For web usage mining using Open Web Resources", *International Journal of Engineering Science and Technology* **2**(10): 5605-5610 (2010).
3. Wen-Chen Hu, "World Wide Web Usage Mining Systems and Technologies", Department of Computer Science, University of North Dakota, Grand Forks.
4. Abraham. A., Business Intelligence from Web Usage Mining, Journal of Information & Knowledge Management (JIKM), World Scientific Publishing Co., Singapore, **2**(4): pp. 375-390 (2003).
5. F. Masseglia, P. Poncelet, and R. Cicchetti, "An Efficient Algorithm for Web Usage Mining", *Networking and Information Systems Journal (NIS),* **2**(5-6): 571-603, (1999).