*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 13, Issue 1 (Jul. - Aug. 2013), PP 22-26*
*www.iosrjournals.org*

# Web Data mining-A Research area in Web usage mining

[1]V.S.Thiyagarajan, [2]Dr.K.Venkatachalapathy
*[1,2]Annamalai University Department of Computer Science*

***Abstract: -*** *Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. The data mining technology normally adopts data integration method to generate data warehouse, on which to gather all data into a central site, and then run an algorithm against that data to extract the useful module prediction and knowledge evaluation. 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 article provides a survey and analysis of current Web usage mining systems and technologies.*
***Keywords:*** *Web log, Session model, path completion.*

## I. 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. It is one of the application of Data mining techniques. 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]. According to analysis target, web mining is classified into three categories, Web content mining, Web structure mining and Web usage mining.

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 known as Web Log Mining, is the process of extracting interesting patterns in web access logs.

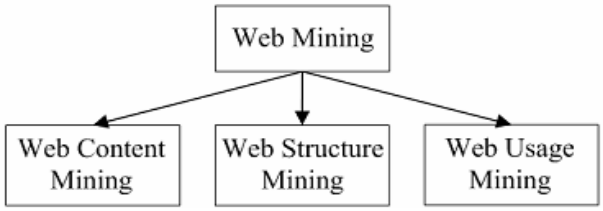

Fig1: Taxonomy of Web Mining

## II. 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. In other words; web usage mining is the process of extracting the required information from the server logs.

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 productclick-throughs, etc.)
3. User profiles and/or user ratings
4. Meta-data, page attributes page content, site structure.

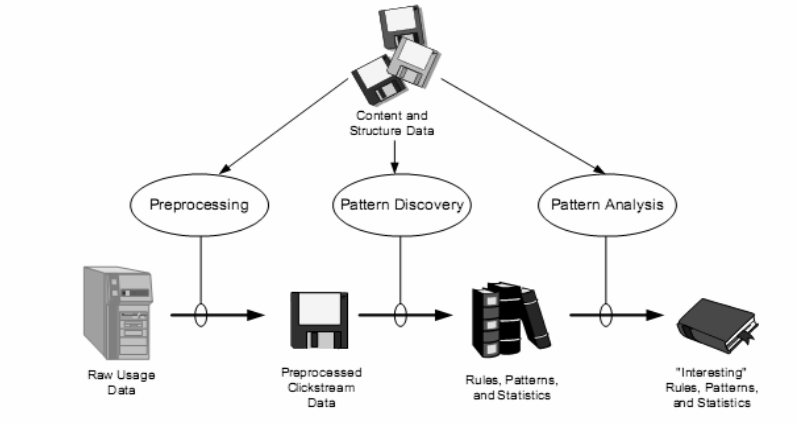


Fig 2: Web Usage Mining Process

**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. 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 3: 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/1999:03:33:11 -0600] "GET /CP.html HTTP/1.0" 200
3218 "http://www.acr-news.org/" "Mozilla/4.06 [en] (Win95; I)"
```

Fig 4: 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 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.

**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. A dirty data (irrelevant data) may cause confusion for the mining procedure, resulting in unreliable output. Data cleaning routines work to "clean" these dirty data by filling in missing values, smoothing noisy data, identifying or removing fake entities, and resolving inconsistencies. 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:

a) 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
b) 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 above 299 or below 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.

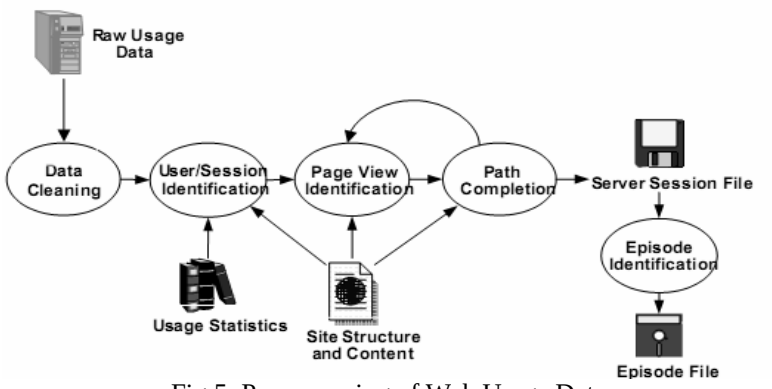


Fig 5: Preprocessing of Web Usage Data

**2) 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.

**3) Path Completion**

Another critical step in data pre-processing 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.

**C) Knowledge Discovery**

Use statistical method to carry on the analysis and mine the pre-treated 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 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.

## III. 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. SUGGEST is completely online and incremental, and it is aimed at providing the users with information about the pages they may find of interest. It bases personalization on a user's classification that evolves according to the user's requests. Usage information is represented by means of an undirected graph whose nodes are associated to the identifiers of the accessed pages, and each edge is associated to a measure of the correlation existing between nodes (pages). This graph is incrementally modified to keep the user model up-to-date. In the model the "interest" in a page does not depend on its contents but on the order by which a page is visited during a session. Therefore, to weight each edge of the graph we introduced a novel formula:

$$Wij=Nij/max(Ni, Nj). \quad (1)$$

Where Nij is the number of sessions containing both pages i and j, Ni and Nj are the number of sessions containing only page i or j, respectively. Dividing Nij by the maximum between single occurrences of the two pages has the effect of discriminating internal pages from the so-called index pages. Index pages are those that do not generally contain useful content and are only used as a starting point for a browsing session. We decided to consider index pages to be of too little interest as potential suggestions because they are very likely to be included in too many sessions. Index pages are used in other works to present the results of the personalization phase. In these cases index pages are not actually used to identify potentially useful information but just to present the personalization results.

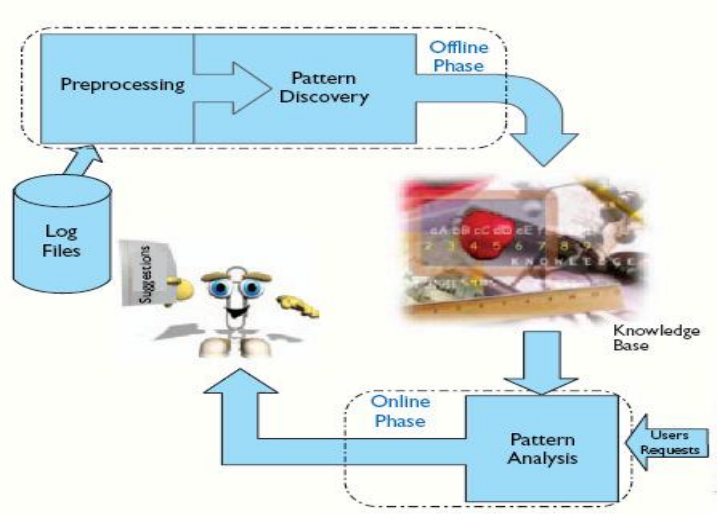


Fig 6: Architecture of the SUGGEST online Recommender System

**3.1 Session Model**

SUGGEST user sessions (identified by means of a cookie-based protocol) are used to build "Session Clusters" eventually leading to a list of suggestions. It finds groups of strongly correlated pages by partitioning the graph according to its connected components. Each component in turn represents a different class, or cluster, of users. The connected components are obtained in an incremental way by using a derivation of the well-known Breadth-First Search (BFS) visit limited to the nodes involved in the request. Basically, we start from the current page identifier and we explore the component to which it belongs. If there are any nodes not considered in the visit a previously connected component has been split and needs to be identified. We simply apply the BFS again, starting from one of the nodes not visited. Furthermore, in order to limit the number of edges of the graph we applied a threshold.

**3.2 Implementation**

The data structure used to store the weights is an adjacency matrix where each entry contains the weight related to a pair of accessed pages. In order to manage Web sites with a number of pages not known, such as Web sites that intensively use dynamic pages, a very innovative solution is applied in SUGGEST, which indexes a page when required. To allow the adjacency matrix to become manageable in size, a LRU algorithm is applied. The Web master of a site may adjust the matrix size according to predetermined constraints such as available resources and performance level. Smaller matrix size values, however, may lead to poor system performance due to frequent page replacements.

After the model has been updated SUGGEST prepares the list of suggestions on the basis of a classification of the user session. This is made in a straightforward way by finding the clusters having the largest intersection with the pages belonging to the current session. The final suggestions are composed by the most relevant pages in the cluster, according to the ranking determined by the clustering phase. The suggestions are then inserted as a list of links in the requested page. Visited pages are not included in the suggestions therefore users belonging to the same class could have different sets of suggestions, depending on which pages have been visited in their active session.

SUGGEST is implemented as a single Apache Web server module in order to allow easy deployment on potentially any kind of Web site currently available, without changing the site itself. Experimental results demonstrate that SUGGEST is able to provide significant suggestions as well as good system performance.

## IV. 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 discussed SUGGEST, an online Recommender System that is based on an incremental procedure, that is able to update incrementally and automatically the knowledge base obtained from historical usage data and to generate a list of links to pages (suggestions) of potentially interest for the user.

## 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