



## Web Usage Mining

**Pranit Bari**, Student, Department of Computer Engineering and Information Technology, VJTI, Mumbai  
**P.M. Chawan**, Associate Professor, Department of Computer Engineering and Information Technology, VJTI, Mumbai

### ABSTRACT

*The paper discusses about web usage mining involves the automatic discovery of user access patterns from one or more Web servers. This article provides a survey and analysis of current Web usage mining systems and technologies. The paper also confers about the procedure in which the web usage mining of the data sets is carried out. Finally the paper concludes with the areas in which web usage mining is implemented.*

**Keywords:** User/Session identification, Web Recommender, Web log, server log

### Introduction

Web mining - is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artefacts or activity related to the World Wide Web.

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 content mining is the process to discover useful information from text, image, audio or video data in the web. Web content mining sometimes is called web text mining, because the text content is the most widely researched area. 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.

### Web Usage Mining

Web usage mining is the type of Web mining activity that involves the automatic discovery of user access patterns from one or more Web servers. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by Web servers and collected in server access logs.

Other sources of user information include referrer logs which contains information about the referring pages for each page reference, and user registration or survey data.

Web Usage Mining is to mine data from log record on web page. Log records lots useful information such as URL, IP address and time and so on. Analysing and discovering Log could help us to find more potential customers and trace service quality and so on. The web usage mining is the process of applying the data mining technology to the web data and is the pattern of extracting something that the users are interested in from their network behaviours. When people visit one website, they leave some data such as IP address, visiting pages, visiting time and so on, web usage mining will collect, analyse and process the log and recording data. Through these, utilize some mathematic method to establish users' behaviour and the interest models, and use these models to understand the user behaviour, thus to improve the website structure. Then finally provides a better characteristic information service for the user. The content and structure of a Web site are used as inputs to every major step of the process.

#### 1. Content

The real data in the Web pages, that is, the data the Web page was designed to convey to the users. This usually consists of, but is not limited to, text and graphics.

#### 2. Structure

Data that describes the organization of the content. Intrapage structure information includes the arrangement of various HTML or XML tags within a given page. The principal kind of interpage structure information is hyperlinks connecting one page to another.

### 3. Usage

Data that describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of accesses. Typically, the usage data comes from an Extended Common Log Format (ECLF) Server log.

## Approach Of Web Usage Mining

### 1. Requirement Analysis

Web access logs are the files that record the users' browsing information on the server. Many kinds of formats are available for web log files.

#### a.) Common 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

remotehost	logname	username	date	request	status	bytes
------------	---------	----------	------	---------	--------	-------

Format of Standard Log File

- remotehost is the remote hostname or its IP address,
- logname is the remote logname of the user,
- username is the username as which the user has authenticated himself,
- date is the date and time of the request,
- request is the exact request line as it came from the client,
- status is the HTTP status code returned to the client, and
- Bytes is the content-length of the document transferred.

#### b.) Extended log file format

An extended common log format file is a variant of the common log format file simply adding two additional fields to the end of the line, the referrer Universal Resource Locator (URL) and the user agent fields:

- Referrer URL is the page the visitor was on when they clicked to come to this page.
- User Agent is whatever software the visitor used to access this site. It's usually a browser, but it could equally be a web robot, a link checker, a File Transfer Protocol (FTP) client or an offline browser. The format of the extended common

log file line has the following fields as shown in figure

Remote host	Log name	User name	Date	Request	Status	bytes	referrer	User agent
-------------	----------	-----------	------	---------	--------	-------	----------	------------

Format of Extended Log File

#### c.) Apache log file format

Field	Meaning
219.144.222.253	Users' IP address (UIP)
[16/Aug/2004...	The date and time of the request (Date)
GET	The method of the request (Method)
/images/1_r3...	The URL of the current request (URI)
HTTP/1.1	The version of transport protocol (Version)
200	The HTTP status code returned to the client (Status)
418	The content-length of the page transferred (Bytes)
http://202.11...	The URL requested just before (Referer)
Mozilla/4.0 (...)	Browser & OS (BrowserOS)

Apache Log File Format

An example of a record in web access log stored in Apache log format is:

```
219.144.222.253 - - [16/Aug/2004:15:36:11
+0800]
"GET /images/1_r3_c2.jpg HTTP/1.1"
200 418
"http://202.117.16.119:8089/index.html"
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT
5.1)"
```

## Steps In Web Usage Mining (WUM)

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

### 1.) Data collection:

Data collection is the first step of web usage mining, the data authenticity and integrity 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).

## 2.) Data preprocessing:

Some databases are insufficient, inconsistent and including noise. The data pre-treatment 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 pre-treatment work, mainly include data cleaning, user identification, session identification and path completion.

### i. 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.

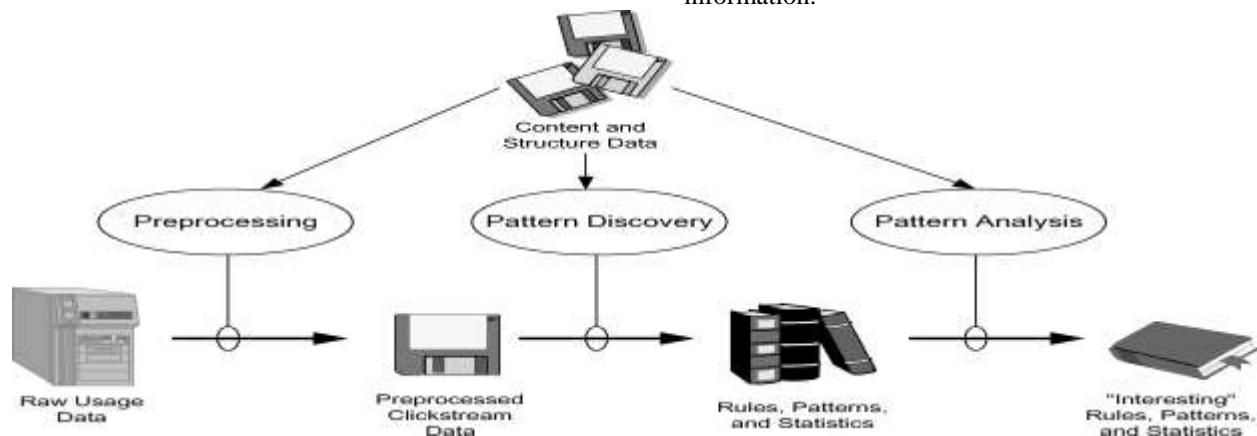


Fig 1. High level Web Usage Mining process

### ii. 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:

- 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 the Refer 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.

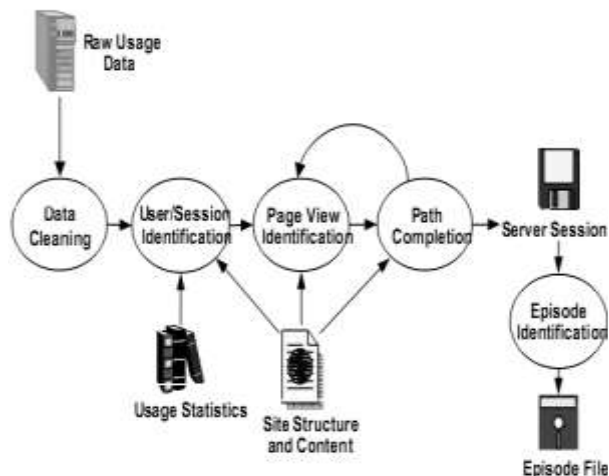


Fig. 2 Preprocessing of Web usage data

### iii. Path completion

Another critical step in data preprocessing is path completion. There are some reasons that result in path's incompleteness, 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 preprocessing, 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.

### 3.) 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.

### 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.

### Wum Tools

In the past, several WUM projects have been proposed to foresee users' preference and their navigation behaviour, as well as many recent results improved separately the quality of the personalization or the user profiling phase.

#### 1. SpeedTracer

In SpeedTracer, a usage mining and analysis tool, is described. Its goal is to understand the surfing behaviour of users. Also in this case the analysis is done by exploring the server log entries. The main characteristic of SpeedTracer is that it does not require cookies or user registration for session identification. In fact, it uses five kind of information: IP, Timestamp, URL of the requested page, Referral, and Agent to identify user sessions. Advanced mining algorithms uncover user's movement through a Web site. The final result is a collection of valuable browsing patterns which help webmasters better understand user's behaviour. SpeedTracer generates three types of statistics: user-based, path-based and group-based. User-based statistics pinpoint reference counts and durations of accesses. Pathbased statistics identify frequent traversal paths in Web presentations. Group-based statistics provide information on groups of Web site pages most frequently visited.

#### 2. Suggest 3.0

SUGGEST 3.0 is a recommender system aimed to provide the users with useful information about pages they may find of their interest. The personalization is achieved by means of a set of dynamically generated



page links. Differently from the recommender systems proposed so far, SUGGEST 3.0 does not make use of any off-line component, and is able to manage Web sites made up of pages dynamically generated. To this purpose SUGGEST 3.0 incrementally builds and maintains historical information by means of an incremental graph partitioning algorithm, requiring no off-line component.

### 3. Collaborative filtering

Collaborative filtering is the most successful technology for building personalized recommendation system and is extensively used in many fields. Most web personalization recommendation system adopt two types of techniques: a content-based approach and a collaborative filtering (CF) approach. In the content-based approach, it recommends web objects that are similar to what the user has been interested in the past. In the collaborative filtering approach, it finds other users that have shown similar tendency to the given users and recommends what they have liked. The collaborative filtering recommendation acts according to other users' viewpoint to produce recommendation tabulates to the goal user. Its basic thought is based on a supposition: If user grade to some product quite similarly, then they grade to other product also similar.

### 4. Web Personalizer

System which provides dynamic recommendation, as a list of hyperlinks to users. Analysis is based on the usage data combined with structured data formed by hyperlinks of site. Aggregated usage profile is obtained by applying data mining technology in preprocessing phase. The online phase considers the active user session in order to find match among user activities and discover usage profile. Matching entries are then used to compute a set of recommendations which will be inserted into last requested page as a list of hypertext links. This system helps in personalization.

### 5. Classifying User Navigation Patterns Using Longest Common Subsequence Algorithm

Prediction of user future movements and intentions based on the users' clickstream data is a main challenging problem in Web based recommendation systems. Web usage mining based on the users' clickstream data has become the subject of exhaustive research, as its potential for web based personalized services, predicting user near future intentions, adaptive Web sites and customer profiling

is recognized. This model provides online prediction through web usage mining system and propose a novel approach for classifying user navigation patterns to predict users' future intentions. The approach is based on using longest common subsequence algorithm to classify current user activities to predict user next Movement.

### Conclusion

Web Usage Mining (WUM) systems are specifically designed to carry out this task by analysing the data representing usage data about a particular Web Site. WUM can model user behaviour and, therefore, to forecast their future movements. Online prediction is one web usage mining application Web usage mining model is a kind of mining to server logs. And its aim is getting useful users' access information in logs to make sites can perfect themselves with pertinence, serve users better and get more economy benefit.

### References

- [1] R.Lakshmipathy, V.Mohanraj, J.Senthilkumar, Y.Suresh, " Capturing Intuition of Online Users using a Web Usage Mining", International Conference on Advance Computing of IEEE,2009.
- [2] Discovering Task-Oriented Usage Pattern for Web Recommendation, Guandong Xu, Yanchun Zhang, Xiaofang Zhou, Proceedings of 17th Australasian Database Conference (ADC2006). 17th Australasian Database Conference (ADC2006), Tasmania, Australia, (167-174). 16-19 January, 2006.
- [3] An Online Recommender System for LargeWeb Sites, Ranieri Baraglia and Fabrizio Silvestri
- [4] Mehrdad Jalali , Norwati Mustaphan ,Ali Mamat,Nasir B Sulaiman " A Recommender System Approach for Classifying User Navigation Patterns Using Longest Common Subsequence Algorithm ", American Journal of Scientific Research ISSN 1450-223X Issue 4 (2009), pp 17-27 © EuroJournals Publishing, Inc. 2009
- [5] Subhash K Shinde, Dr U V Kulkarni "A New Approach for Online Recommender System in Web Usage Mining", International Conference on Advance Computing Theory and Engineering of IEEE, 2008.
- [6] Research on Personalized Recommendation Based on Web Usage Mining using Collaborative Filtering Technique, Taowei Wang, Yibo Ren
- [7] "Page Interest Estimation Based on the User's Browsing Behaviour", Bo-qin FENG, Feng WANG