# Web Usage Mining – Tools, Data Preparation Methods and Comparative Study on Implementation Techniques

K.Jayamalini
Research Scholar,
Computer Science Engineering,
Bharath University,
Chennai, India
*malini1301@gmail.com*

Dr.M.Ponnavaikko
Vice-Chancellor
Bharath University
Chennai, India
*vc@bharathuniv.ac.in*

***Abstract-*** Recent years are characterized by an exponential growth of the number of Web sites available on the Internet and the number of their users. This phenomenal growth has produced a huge quantity of data related to the users' interactions with Web sites, stored by Web servers in access log files. A log file is a text file which records the requests made to the Web server in chronological order. Generally, a log file contains: the client's host name or IP address, the request's date and time, the operation type, the requested resource name (URL) and the size of the requested page. Today, understanding the interests of users is becoming a fundamental need for Web sites' owners in order to better serve their visitors. The analysis of Web log files permits to identify useful patterns of the browsing behavior of users. To discover behavioral patterns from Web usage data, Web Usage Mining uses with different data mining techniques. However, there are several preprocessing tasks that must be performed prior to applying data mining algorithms to the data collected from server logs.

This paper presentsexplains tools for analyse data preparation techniques in order to identify unique users and user sessions. It also presents different data mining algorithms used to analyze the usage data. At the end, it presents a comparative study of different data mining algorithms.

***Keywords-*** *Web Usage Mimning (WUM), Data Preparation,Apriori Algorithm,FP Growth Algorithm.*

*****

## I. INTRODUCTION

Web mining [1], aims to discover useful information or knowledge from the Web hyperlink structure, page content and usage log. Based on the primary kind of data used in the mining process, Web mining tasks are categorized into three main types which are shown in fig 1:

- Web Structure Mining
- Web Content Mining
- Web Usage Mining(WUM)

Web structure mining is the process of analyzing the node and connection structure of a web site. Web content mining is process of mining, extraction and integration of useful data, information and knowledge from Web page contents. Web usage mining is the process of extracting useful information from server logs i.e. users' history or it is also defined as a process of finding out what users are looking for on Internet.

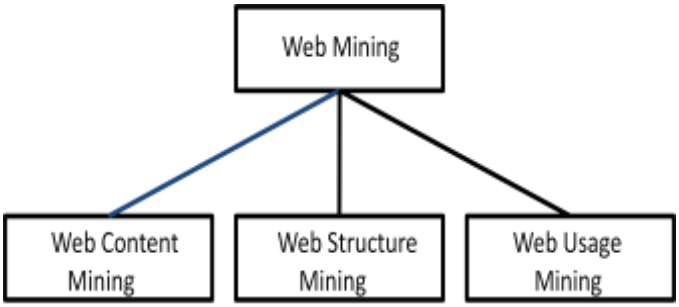


Fig.1. Web Data Mining Basic Classifications.

It is used to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity of Web users along with their browsing behavior at a Web site. There are different type of Web Usage Mining data, such as web server data, application server data and application level data depending on the kind of usage data considered. Web server data correspond to the user logs that are collected at Web server, which include IP addresses, page references, and access time of the users. The main goal of the WUM is to identify usage pattern from web log files of a website. Apriori and FP Growth algorithms are used for this purpose.

In first section, this paper explains various tools used for analyzing web server log files. In second section, it explains data preparation techniquesused for WUM and finally this paper explains algorithms likeApriori and FP Growth which are used to analyze the usage data. Towards the end, this paper gives the comparison among various algorithms used for data analysis.

## II. WEB SERVER LOG ANALYSIS TOOLS

There are several commercially available Web server log analysis tools [7] to analyze the server logs. Some of them are:

1. Deep Log Analyzer- is the best free Web analytics software or local log analysis tool. It works on your site logs without requiring any codes or bugs on your site.

2. Google Analytics - is one of the best free Web log analysis tools available.
3. AWStats - it is a free Web analysis tool that works as a CGI script on your Web server or from the command line.
4. W3Perl - W3Perl is a CGI based free Web analytics tool. It offers the ability to use a page bug to track page data without looking at log files or the ability to read the log files and report across them.
5. Power Phlogger - is a free Web analytics tool that is based on PHP to track information you can offer to other users on your site.But it can be slow.
6. Visitors - Visitors is a command line free log analysis tool. It can generate both HTML and text reports by simply running the tool over your log file.
7. BBClone - PHP based Web analytics tool or Web counter for your Web page which provides information about the last visitors to your site tracking things like: IP address, OS, browser, referring URL and more.
8. Analog - widely used free Web log analysis tool. It works on any Web server and it is easy to install and run if you understand how your server is administered.

The biggest impediments to collecting reliable usage data are

a) Local caching - In order to improve performance and minimize network traffic, most Web browsers cache the pages that have been requested. As a result, when a user hits the \back" button, the cached page is displayed and the Web server is not aware of the repeat page access.
b) Proxy Servers - Proxy servers provide an intermediate level of caching and create even more problems with identifying site usage. In a Web server log, all requests from a proxy server have the same identifier, even though the requests potentially represent more than one user.
c) Cookies and Remote agents - Cookies are used to tag and track site visitors automatically. Instead of sending a cookie, sends a Java agent that is run on the client side browser in order to send back accurate usage information to the Web server. The major disadvantage of the methods is that rely on implicit user cooperation stem from privacy issues. Many users choose to disable the browser features that enable these methods.

## III. PREPROCESSING

Preprocessing is used to convert raw Web server logs into user session files in order to perform Web Usage Mining. Figure.2 shows how the preprocessing tasks of Web Usage Mining in greater detail. The inputs to preprocessing phase are the server logs, site files, and optionally usage statistics from a previous analysis. The outputs are the user session file, transaction file, site topology, and page classifications.

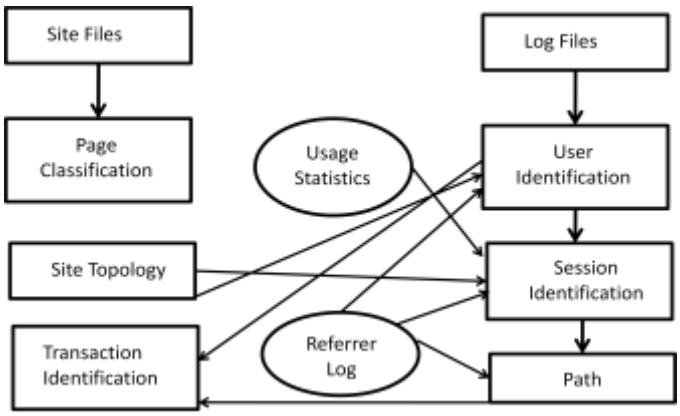


Fig.2. Web Usage Mining Process

The major impediments in creating a reliable user session file are

a) Browser and Proxy server caching- Current methods to collect information about cached references include the use of cookies and cache busting. Cache busting is the practice of preventing browsers from using stored local versions of a page, forcing a new down-load of a page from the server every time it is viewed. Cookies can be deleted by the user and cache busting defeats the speed advantage that caching was created to provide, and is likely to be disabled by the user.
b) User Registration – It is another method to identify users However, due to privacy concerns, many users choose not to browse sites that require registration and logins, or provide false information.

The following steps are performed [8] in preprocessing of web server log files.

a) Data Cleaning
b) User Identification
c) Session Identification

*a)Data Cleaning*

It important to clean the server logs to eliminate irrelevant items for any type of Web log analysis. The following are not required for WUM:

a) Since the main intent of Web Usage Mining is to get a picture of the user's behavior, it does not make sense to include file requests that the user did not explicitly request.
b) All log entries with filename suffixes such as, gif, jpeg, GIF, JPEG, jpg, JPG, and map is removed.
c) Data cleaning is usually site-specific, and involves tasks such as, removing extraneous references to

embedded objects that may not be important for the purpose of analysis, including references to style files, graphics, or sound files.

d) The cleaning process also may involve the removal of at least some of the data fields (e.g. number of bytes transferred or version of HTTP protocol used, etc.) that may not provide useful information in analysis or data mining tasks.

e) Data cleaning also entails the removal of references due to crawler navigations.

*b) User Identification*

User Identification task is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. The Web Usage Mining methods that rely on user cooperation are the easiest ways to deal with this problem. It is possible to accurately identify unique users through a combination of IP addresses and other information such as user agents and referrers.

Consider, for instance, the example in Figure3. On the left, the figure depicts a portion of a partly preprocessed log file (the time stamps are given as hours and minutes only). Using a combination of IP and Agent fields in the log file, we are able to partition the log into activity records for three separate users (depicted on the right).

| Time | IP | URL | Ref | Agent |
|---|---|---|---|---|
| 0:01 | 1.2.3.4 | A | - | IE5;Win2k |
| 0:09 | 1.2.3.4 | B | A | IE5;Win2k |
| 0:10 | 2.3.4.5 | C | - | IE6;WinXP;SP1 |
| 0:12 | 2.3.4.5 | B | C | IE6;WinXP;SP1 |
| 0:15 | 2.3.4.5 | E | C | IE6;WinXP;SP1 |
| 0:19 | 1.2.3.4 | C | A | IE5;Win2k |
| 0:22 | 2.3.4.5 | D | B | IE6;WinXP;SP1 |
| 0:22 | 1.2.3.4 | A | - | IE5;WinXP;SP2 |
| 0:25 | 1.2.3.4 | E | C | IE5;Win2k |
| 0:25 | 1.2.3.4 | C | A | IE6;WinXP;SP2 |
| 0:33 | 1.2.3.4 | B | C | IE6;WinXP;SP2 |
| 0:58 | 1.2.3.4 | D | B | IE6;WinXP;SP2 |
| 1:10 | 1.2.3.4 | E | D | IE6;WinXP;SP2 |
| 1:15 | 1.2.3.4 | A | - | IE5;Win2k |
| 1:16 | 1.2.3.4 | C | A | IE5;Win2k |
| 1:17 | 1.2.3.4 | F | C | IE6;WinXP;SP2 |
| 1:26 | 1.2.3.4 | F | C | IE5;Win2k |
| 1:30 | 1.2.3.4 | B | A | IE5;Win2k |
| 1:36 | 1.2.3.4 | D | B | IE5;Win2k |

User 1

| | | | |
|---|---|---|---|
| 0:01 | 1.2.3.4 | A | - |
| 0:09 | 1.2.3.4 | B | A |
| 0:19 | 1.2.3.4 | C | A |
| 0:25 | 1.2.3.4 | E | C |
| 1:15 | 1.2.3.4 | A | - |
| 1:26 | 1.2.3.4 | F | C |
| 1:30 | 1.2.3.4 | B | A |
| 1:36 | 1.2.3.4 | D | B |

User 2

| | | | |
|---|---|---|---|
| 0:10 | 2.3.4.5 | C | - |
| 0:12 | 2.3.4.5 | B | C |
| 0:15 | 2.3.4.5 | E | C |
| 0:22 | 2.3.4.5 | D | B |

User 3

| | | | |
|---|---|---|---|
| 0:22 | 1.2.3.4 | A | - |
| 0:25 | 1.2.3.4 | C | A |
| 0:33 | 1.2.3.4 | B | C |
| 0:58 | 1.2.3.4 | D | B |
| 1:10 | 1.2.3.4 | E | D |
| 1:17 | 1.2.3.4 | F | C |

Fig.3: Example of user Identification using IP + Agent

*c) Session Identification*

The goal of Session Identification is to reconstruct, from the clickstream data, the actual sequence of actions performed by one user during one visit to the site and to divide the page accesses of each user into individual sessions. The simplest method used to achieve this is through a timeout. If the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session. Many commercial products use 30 minutes as a default timeout.

Anexample for session identification is given in Fig 4. In this figure, the heuristic h1, described above, with $\theta = 30$ minutes has been used to partition a user activity record (from the example of Fig. 2) into two separate sessions.

User 1

| Time | IP | URL | Ref |
|---|---|---|---|
| 0:01 | 1.2.3.4 | A | - |
| 0:09 | 1.2.3.4 | B | A |
| 0:19 | 1.2.3.4 | C | A |
| 0:25 | 1.2.3.4 | E | C |
| 1:15 | 1.2.3.4 | A | - |
| 1:26 | 1.2.3.4 | F | C |
| 1:30 | 1.2.3.4 | B | A |
| 1:36 | 1.2.3.4 | D | B |

Session 1

| | | | |
|---|---|---|---|
| 0:01 | 1.2.3.4 | A | - |
| 0:09 | 1.2.3.4 | B | A |
| 0:19 | 1.2.3.4 | C | A |
| 0:25 | 1.2.3.4 | E | C |

Session 2

| | | | |
|---|---|---|---|
| 1:15 | 1.2.3.4 | A | - |
| 1:26 | 1.2.3.4 | F | C |
| 1:30 | 1.2.3.4 | B | A |
| 1:36 | 1.2.3.4 | D | B |

Fig.4. Example of session Identification with a time-oriented heuristic

*d) Formatting*

Once the appropriate preprocessing steps have been applied to the server log, a final preparation module can be used to properly format the sessions or transactions for the type of data mining to be accomplished. For example, since temporal information is not needed for the mining of association rules, a final association rule preparation module would strip out the time for each reference, and do any other formatting of the data necessary for the specific data mining algorithm to be used.

## IV. FUTURE WORK

In future, to meet people's demand of intelligent information, we will plan to reduce the manual involvement, and developcomplete automatic intelligent extraction systems. We will also plan to developsystems which can extract data from Webpages which have different in structures.

## ACKNOWLEDGMENTS

We wish to express our sincere thanks and deep sense of gratitude to all the staff members of computer department of Bharath University, Chennai for their support and co-ordination. We also wish to thank to our family members for their support to complete this work.

## REFERENCES

[1] Shuyan Bai, Qingtian Han, Qiming Liu, Xiaoyan Gao "Research of an Algorithm Based on Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering, 2009, pp.1-4.

[2] P.Nithya, Dr. P.Sumathi, "A Survey on Web Usage Mining: Theory and Applications", Int.J.Computer Technology & Applications, pp. 1625-1629, August 2012.

[3] V. Sathiyamoorthi1, Dr. V. Murali Bhaskaran, "Data Preparation Techniques for Web Usage Mining in World Wide Web-An Approach", International Journal of Recent Trends in Engineering, pp.1-4, 2009.

[4] http://webdesign.about.com/od/loganalysis/tp/free_web_log_analysis_tools.htm(Accessed: October, 2012)

[5] www.maya.cs.depaul.edu/~mobasher/papers/12-web-usage-ining.pdf (Accessed: October, 2012)

[6] B.Santhosh Kumar, K.V.Rukmani, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms", Int. J. of Advanced Networking and Applications, pp. 400-404, 2010.

[7] http://webdesign.about.com/od/loganalysis/tp/free_web_log_analysis_tools.htm(Accessed: October, 2012)

[8] V. Sathiyamoorthi1, Dr. V. Murali Bhaskaran, "Data Preparation Techniques for Web Usage Mining in World Wide Web-An Approach", International Journal of Recent Trends in Engineering, pp.1-4, 2009.