



Web Usage Mining: Finding Usage Patterns from Web Logs

Sanjeev Dhawan¹, Swati Goel²

¹Faculty of Computer Science & Engineering, ² Research Scholar,
University Institute of Engineering and Technology,
Kurukshetra University, Kurukshetra-136119, Haryana,
INDIA

Abstract: Web usage mining is an important research domain of Web mining which focuses on learning about the user's behavior while interacting with the websites. It is the application of data mining techniques on web log data in order to find the usage patterns of web users. A web log file represents activities of the users accessing a web site and hence provides vital information about user access patterns. Web usage mining includes three main phases namely: data collection and preprocessing, pattern discovery and pattern analysis. This paper gives an overview of web log file, describes its various types and gives a detail of the process of Web usage mining. In the end it also outlines future research directions in area of web usage mining.

Keywords: Data mining, Web Mining, Web Usage Mining, Web log file, Data preprocessing

I. Introduction

Web mining is one of the most significant fields in the area of data mining. Web Mining is the application of data mining techniques to Web data, which can be Web document content or hyperlink structure or Web log file, to find and extract the undiscovered knowledge and useful patterns [7]. Web mining can be categorized into three classes based on the origin of data to be mined: Web content, Web structure and Web usage mining. Web content mining includes discovering knowledge from the contents of a Web page. Web structure mining includes discovering knowledge from the structural material of Web page. Web usage mining includes discovering knowledge from Web log files. When a user accesses a website, its interactions are recorded in the log file. The mining of this kind of data has emerged as a necessity in today's world as users expect to find the relevant information on the web as quickly as possible and web administrators want to track user's behavior for many reasons. The sources of Web usage data are: server side, proxy side and client side log file. Server side logs are further divided into four types: Access log, Agent log, Error log and Referrer log. These logs can be mined individually or can be joined for mining. These log files comes in different formats namely: IIS, NCSA and W3C. All the three formats have their own distinct characteristics. Web usage mining includes three main phases namely: data collection and preprocessing, pattern discovery and pattern analysis. Data collection includes collecting data from different data sources which can be a single or more than one source. In the latter case, it may be required to merge all the data into a single file and then perform mining. Data preprocessing is a vital step in this process so as to avoid misleading results as data is often noisy and inconsistent. Pattern discovery is discovering knowledge by applying methods such as clustering, classification, association rule to the web data. Pattern analysis includes evaluating the patterns discovered in the previous stage and presenting in a human understandable format such as graphs, reports etc. Figure 1 describes the taxonomy of Web mining:

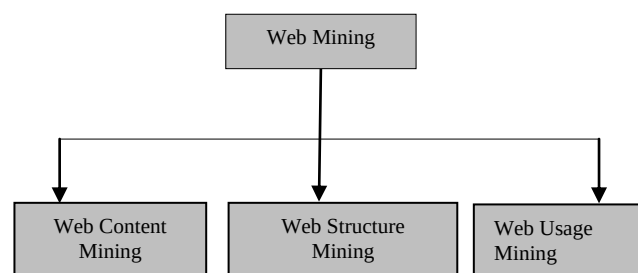


Figure 1 Web Mining taxonomy

This paper is placed into following sections: section 2.1 gives the overview of web log file, section 2.1.1 discusses different types of web server logs, section 3.0 describes the approach of Web Usage Mining in detail and finally section 4.0 concludes the paper and outlines some of the promising areas of the future research.

II. Web Usage Mining

Cooley *et al.* [10] were the first to put their efforts in the area of Web Usage Mining by proposing the system Webminer. It may be defined as the application of data mining techniques to web log data in order to find the usage patterns of web users which can be used for different activities such as web site optimization and personalization, designing cross marketing methods, evaluating the effectiveness of various campaigns, predicting future requests etc.

A. Web log file: An overview

It is a simple text file which keeps record of the requests that are submitted by the user to the server while accessing a website. Key fields contained in log file are: username, IP address, user agent, date, time, number of bytes transferred etc.[4] Example of a log file entry stored on IIS server taken from www.uietkuk.org is shown in figure 2:

```
2013-03-10 12:57:40 208.91.198.202 GET /downloads.html- 80 - 37.228.106.69 HTTP/1.1
Opera/9.80 200 5444 31
```

Figure 2: Sample log file entry

Various fields are explained below:

- Date: The date of access is recorded for each hit. The format is YYYY MM-DD[3]. In fig. 2, date is 2013-03-10.
- Time: It denotes the time at which access was made. The time format is HH:MM:SS. In fig. 2 time is 12:57:40.
- Server IP: It is the server IP address. Its value in fig. 2 is 208.91.198.202.
- Server method: It is the method by which servers send information to the client. It can be: GET, POST, or Head. In fig. 2, GET method is used.
- Uri-stem: URI-Stem is a path of the page accessed from the host. Its value in fig. 2 is /downloads.html.
- Server Port number: It is a communication port used for transmission of data between client and server. Usually port number 80 is used .
- Client IP: It is the IP address of the client. Its value in fig. 2 is 37.228.106.69.
- Version and User-Agent: It denotes internet protocol/version and Browser type/version respectively. HTTP is the protocol with version 1.1 and browser used is Opera with version 9.80 in the above figure.
- Status: This indicates the status code returned by the web server to the client to indicate the success or failure of the transaction. 200 is returned by the server in above figure, denoting success.
- Number of Bytes transferred: It indicates the number of bytes transferred by the server to the client. 5444 bytes are send by the server.
- Time taken: It indicates the time taken for the complete request and response cycle.

1. Types of web server logs

Web server logs keeps track of pages visited by a user as well as details related to the accesses such as IP address of client, request date and time, page that is requested, HTTP code, number of bytes transferred, user - agent, referrer etc. This data can be saved into a single or multiple files. These files are usually not be easily retrieved by the general internet users.

Different types of server log files are:

1. Access log: It keeps track of all the requests that are sent by the client and hence processed by the server [4]. Information about the user is then processed to determine user behavior [8]. Commonly indicated fields in this log are: Date and time of access, client IP, User authentication, Server name, its IP address and port [3]. Its format is same as in fig 2.

2.Error log: Whenever an error occurs in accessing a page, which is being requested by client to web server, the entry is made in this log [2]. For example, error log file records an entry when a user clicks on a particular link that does not locate the promised page or web site, as a result of which a message is displayed “Error 404 File

Not Found’’. It is beneficial for the web page designers to optimize the links of web site [8]. Error code 404 is clearly shown in Figure 3 below.

```
2012-03-25 00:29:44 W3SVC49 PLESK-WEB14 208.91.198.202 GET /robots.txt - 80 -
88.131.106.22 HTTP/1.0 Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+en-
US)+Speedy+Spider+(http://www.entireweb.com/about/search_tech/speedy_spider/) - -
www.uietkuk.org 404 0 2 5413 443 2480
```

Figure 3 Sample Error Log entry

3. Agent log file: It contains the information about user’s browser, its version and operating system and its version[4]. This information is used by the web site designer and administrator for the analysis of which specific browser is used by users. Also, this can be used to find out the most popular browsers and operating systems among the users. In figure 4 the browser used is Mozilla with version 5.0 and Operating system used is Windows 7.

```
Mozilla/5.0/Win 7
```

Figure 4 Sample Agent Log entry

4. Referrer log file: It contains the detail about the referrer. For example, as someone jumps from www.google.com to any site by clicking the link, referrer log file of that web server will record a referrer entry that a user came from www.google.com [8]. The Referrer Log tells what web sites link to a server[3]. In figure 5 referrer is www.google.com.

```
http://google.com/search;_ylt=AiDy..QmB7VT2jNWjbB
WPVmutIF;_ylc=X1MDOTc2ODQxNDIEX3IDMgRmcgN5ZnA
tdC03MDQEbl9ncHMDMTAEb3JpZ2luA2luLnhaG9vLmNvbQ
- - ?p=india&toggle=1&cop=mss&ei=UTF-8&fr=yfp-t-704
```

Figure 5 Sample Referrer log Entry [3]

III. Approach of Web usage mining

The following steps are included in web usage mining;

a. Data Collection: It is the first step of Web usage mining. Data collection includes collecting data from various data sources which can be a single or more than one source. In the latter case, it may be required to fuse multiple data files into a single file and then perform mining. In order to ensure authenticity and integrity of data, scientific and advanced techniques should be used for data collection. There are mainly three sources of data namely: server data, client level data and proxy level data [6].

b. Preprocessing: Data needs to be treated before mining as log files are inconsistent and noisy [6]. Mining directly from such a log file called as raw log file, would lead to misleading results. Hence the following preprocessing methods should be adopted to remove noise and inconsistency:

- i. Data cleaning: The main purpose of this step is to remove irrelevant, noisy and unnecessary entries from the log file. Usually user requests to the web server results in many useless entries. These can be:
 - Requests that are performed by automated programs such as web robots plus spiders .The traffic that these programs generate can create false results. Example of such a file is robots.txt.
 - Usually icons(t.ico), images(a.jpg, b.png, c.gif), javascripts(d.js) and cascaded style sheets(e.css) are also retrieved along with requested pages which results in several useless log entries while users are just interested in what they requested explicitly.
 - Entries with unsuccessful HTTP status codes. Only codes between 200 and 299 are treated as successful. Hence, codes other than 200 series code are removed.
 - Log entries with request methods apart from “Get” and “Post”.
- ii. User identification: Identifying an individual user uniquely who access a web site is one of the key stages of data preprocessing. The easiest approach for user identification is the use of cookies. But the

main problem is that not all websites use cookies due to privacy reasons. Also users can disable cookies. Another approach can be the use of IP addresses for user identification. Users with different IP addresses can be treated as unique but it is not the perfect solution as ISPs assign rotating addresses to users [9]. Hence, for those with same IP, combination of IP and browser or operating system can be considered as unique.

- iii. Session identification: Session is the time elapsed since the user starts accessing a website to the time he leaves. Session identification is to divide the page accesses of each user identified in previous step into sessions. Two heuristics are used for session identification: time based and structure based. Time oriented generally uses timeout mechanism. The most common one is setting a threshold on the time limit which is generally 30 minutes. If a user's session exceeds this limit, he is treated as a different user. On the other hand, structure oriented uses static structure of a website [9].
 - iv. Path completion: It is determining if some of the significant accesses are missed from being recorded. Main reasons resulting in path incompleteness are local caching and use of back button. These types of accesses are not recorded in the logs. Such incomplete paths result in missing some of the important entries to be recorded in log file. These must be appended in order to mine useful traversal patterns [6]. The methods used for user identification to a certain extent, can be used for session identification.
- c. Knowledge discovery: Pattern discovery includes discovering knowledge by applying methods such as clustering, classification, association rule etc. to the web data. Various methods that can be applied to web data in order to find useful patterns are:
- Statistical analysis: This is the most commonly used method in discovering knowledge about web users. Presently, there are many traffic analysis tools which generate a report depicting statistics such as mean length of path accessed, mean time of page viewing, most frequently accessed pages etc [1].
 - Association rules: This method can be used to find group of pages which are frequently accessed together with support exceeding a threshold. It is not necessary that the pages are connected directly. For example, Apriori algorithm [5] can be used to find relation between users who access a faculty page of a college and those who access a syllabus download page. It is not only applicable to marketing as well as business domains but also is very useful for web designers and administrators in deciding the layout of website contents based on the correlations generated [1].
 - Classification: It is mapping a page to a set of predefined groups or classes. This is applicable in web domain as one may be interested in class or category of users having stated characteristics. Various algorithms that can be used for classification include decision tree, naïve Bayesian, k-nearest neighbor algorithm etc. For example, rules of the form 40% of the users who placed an order for iPhone online are youngsters between the age group of 20-40 years with majority living in United States of America [1].
 - Clustering: It is grouping together a set of items having similar features. Two types of clusters can be found in web usage mining: user clusters and page clusters. User clusters will discover users having same browsing patterns whereas page clusters will discover pages possessing similar content [1].
 - Sequential patterns: It is same as association rule with the difference of time ordering. It finds patterns such that one or set of pages are accessed after the another set but in a time sequence. Its application is the prediction of future visitors so as to target advertising on a group of users.
- d. Pattern analysis: It includes filtering uninteresting patterns and to present the interesting patterns to the user in a human understandable format such as graphs, table, pie charts, reports, rules generated from discovered patterns etc. [6].

Figure 6 describes the complete process of web usage mining.

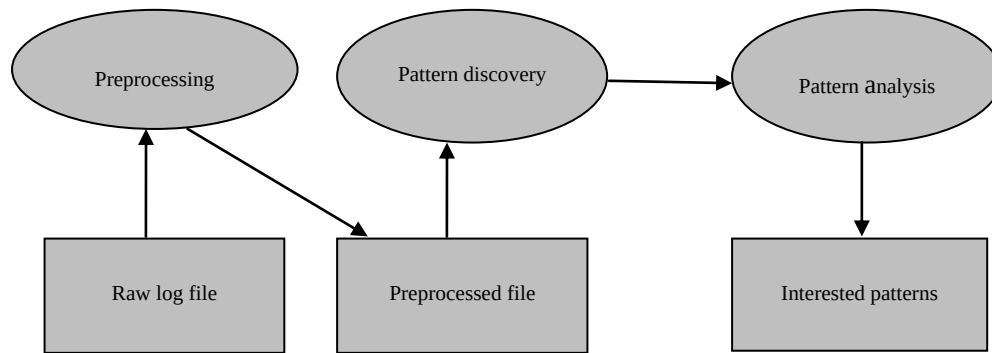


Figure 6 illustrates the complete process of Web Usage Mining.

IV. Conclusions and future research directions

This paper focuses on web log file overview, its types and the process of web usage mining. The web log files records information of each click by the user. Log files usually contain noisy and irrelevant data. Preprocessing is done to remove unnecessary data from log file. User and session identification is also done as a preprocessing step which includes identifying users and sessions. Pattern discovery techniques such as association rule mining (such as Apriori), clustering and classification are applied on the reduced log file. Finally, these are presented to the user in a suitable format. One of the challenges in web usage mining is efficient mining of web server logs. The quality of an algorithm is judged usually in terms of efficiency. Therefore, one of the future research problems is to increase the efficiency of association rule mining technique, namely Apriori, when used for discovery of patterns from Web server logs. This can be done by reducing the size of log file thereby reducing the time of log file scans. Others improved methods could be efficient implementation of other pattern discovery methods such as classification, clustering etc. on Web server logs.

References

- [1] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang -Ning Tan. "Web usage Mining: Discovery and applications of usage patterns from web data". ACM SIGKDD. Jan 2000.
- [2] L.K. Joshila Grace, V. Maheshwari, Dhinaharan nagamalai. "Analysis of web logs and web user in web Mining". International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.
- [3] K Sudheer Reddy, Dr. G. Partha Saradhi Varma, Dr. I. Ramesh Babu. "Preprocessing the Web Server Logs – An illustrative approach for effective usage mining". ACM SIGSOFT. Volume 37 Number 3. May 2012.
- [4] Priyanka Patil and Ujwala Patil. "Preprocessing of web server log file for web mining". Proceedings of National Conference on Emerging Trends in Computer Technology. April 21, 2012.
- [5] R.Agarwal and R.Srikant. "Fast Algorithms for mining association rules in large databases". In the proceedings of the 20th International conference on very large databases. pages 487-499, Santiago,Chile, 1994.
- [6] Rajni Pamnani, Pramila Chawan, "Web Usage Mining: A Research Area in Web Mining", ", International Conference on Recent Trends in Computer Engineering, ISCET 2010.
- [7] Sanjeev Dhawan, Swati Goel. "Analytical study for web mining and content negotiation". Proceedings of National Conference on Advanced Computing Technologies, March 30, 2013.
- [8] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood, "Web Usage Mining: A Survey on Preprocessing of Web Log File", In the proceedings of International Conference on Information and Emerging Technologies, 2010
- [9] Bamshad Mobasher, B. Liu and Olfa Nasraoui, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Data-Centric Systems and Applications, DOI, Springer-Verlag Berlin Heidelberg, 2011",
- [10] Cooley, R., Mobasher, B., & Srivastava, J. "Data preparation for mining World Wide Web browsing patterns". Knowledge Information Systems, pp. 5-32, 1999.