Web Log Preprocessing: A Major Step in Web Usage Mining

Muskan
Research Scholar, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra
muskan21dudi@gmail.com
Dr. Kanwal Garg
Assistant Professor, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra
gargkanwal@kuk.ac.in

Abstract
The data captured in web logs is highly unstructured, incomplete, inconsistent and noisy in nature. So to get the quality results after analysis this raw data present in the log files need to be preprocessed. Web log preprocessing is the first and very important step in web usage mining because analysis can be done effectively only if data is clean, consistent and complete. This paper provides an overview of the various types of web log formats, various steps included in web log preprocessing- Data Cleaning, User and Session Identification, Path Completion and various issues and challenges in preprocessing of the log files.

Keywords—Data Cleaning, Path Completion, Preprocessing, Session Identification, User Identification

1. Introduction

Internet has become a crucial tool for people to satisfy their needs from the server. So web mining[1],[2] is attracting the attention of researchers for extracting the knowledge out of web data for improving the performance of websites. *Web mining*can be categorized into three streams as*Web Content Mining* that focuses on the content of the webpage, *Web Structure Mining*that focuses on the hyperlink structure of the webpage and*Web Usage Mining* that focuses on processing of web log files. This paper is focused on Web usage mining (WUM) also called as Web log mining.The term Web usage mining was first introduced by Cooley in 1997[3],[4] when a first attempt of taxonomy of Web Mining was done, in that they defined Web Mining as "the discovery and analysis of useful information from World Wide Web". When the resources are requested from server,user leaves abandon information behind and this information is captured by the log files. Analyzing these log files to extract some knowledge about the navigation behavior of the various users is called as WUM.Various steps performed for web usage mining are web log preprocessing, pattern discovery and pattern analysis. Basically *preprocessing* step consists of data cleaning, data integration, data reduction and data transformation. After preprocessing the next step in WUM is pattern discovery. *Pattern discovery* deals with extracting rules out of preprocessed data using various techniques like clustering, classification etc. The last step in WUM is pattern analysis. In *Pattern analysis* uninteresting rules out of the set of rules obtained from pattern discovery phase, are filtered out.

2. Sources of Web Logs

Log files can be of different types depending on their location of capturing data[3]:-

- Web Servers
- Web Proxy Servers
- Client Browsers

Web Server Log Files: - Aweb server log contains the most accurate and complete information. This information is regarding single site multiple users. But they do not record the cached pages. Web server log files contain very sensitive information, hence they are generally kept close by the operator.

Web Proxy Server Log Files: - Proxy server is also called as intermediate server that exists between client and server. When user request goes through a proxy server then an entry in the log file will be the information of the proxy server not the user. Hence proxy server maintains a separate log file that captures the information of users.

Client Browser Log Files: - This type of log files reside on the client browser. HTTP cookies are used for client browser. HTTP cookies are pieces of information generated by the server and stored on client for future use.

3. Types of Web Server Logs

Web Server log files comprises access logs, referrer logs, agent logs and error logs[3].

3.1 Access Logs

Access logs provide an extensive view of web servers and users. These logs can be used by the web server administrator and decision makers to characterize the users and usage patterns. These logs contain the record of files requested by the user from the server. There are various log formats available for access log files, which are discussed in section 4.

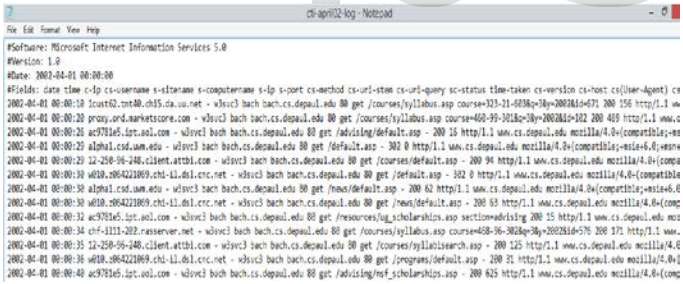

Figure 1:access log in w3c extended log format[1]

3.2 Error Logs

An entry in most of the error logs looks like as follows:-

[Sun Mar 7 21:16:17 2004] [error] [client 24.70.56.49] File does not exist: /home/httpd/twiki/view/Main/WebHome[5]

If an error occurs, when the client requests something from the server then an entry of the error is recorded in the error log file. The first part of this entry indicates the date and time of the occurrence of error. The second part indicates the error being reported. The third part indicates the IP of the client that had made the request. Next part is the message itself. Last part is the system path of the requested document generated by the server.

3.3 Referrer Logs

These logs contain the URL's of the webpagesfrom where a user got redirected to a particular webpage.

Sample referrer log entry: -

[07/Mar/2004:16:50:54-0800] "http://www.snapdeal.com/index.html"

Various fields of this log entry are described as follows:-

- *[09/Jan/2008:16:48:54-0800]*- This field represents the time in the format [day/month/year:hour:minute:second-zone].
- *"http://www.snapdeal.com/index.html"*- This entry indicates the URL of a webpage from where the user got redirected to a particular webpage.

3.4 Agent Logs

The agent logs provide the information about client's browser and platform using which a client made a request to the server. This is the major information, because it depends on the type of the browser and the platform that what a user can access on a website.

Sample Agent Log entry: -

[20/Dec/2011:18:15:06-0800] "Microsoft Internet Explorer-5.0"

Various fields of this log entry are described as follows:-

- *[20/Dec/2011:18:15:06-0800](%t)*- This entry represents time in the format [day/month/year:hour:minute:second-zone].
- *"Microsoft Internet Explorer-5.0"*- This field presents the details of browser which is used by the user to request the resource from the server.

4. Formats of Access Logs

4.1 Apache Log Formats: -Three types of log formats are considered in Apache servers, which are discussed as follows:-

4.1.1 Common Log Format (CLF)

The configuration of the common log format is given below:-

LogFormat "%h %l %u %t \"%r\" %>s %b"

Sample entry of a log file in CLF is shown below:-

64.242.88.10--[07/Mar/2004:16:50:54-0800]"GET/dccstats/stats-spam-ratio.1week.png HTTP/1.1" 200 7654

Various parts of this log entry are described as follows:-

- *64.242.88.10 (%h)* - This represents the IP address of the client requesting the resources from the server.
- *- (%l)* – This represents the login name of the person who owns the account that is making the request.
- *- (%u)* – This entry represents the full name of the user who owns the account that is making the request to the server.
- *[07/Mar/2004:16:50:54-0800] (%t)* – This represents the time in the format [day/month/year:hour:minute:second-zone].
- *"GET /dccstats/stats-spam-ratio.1week.pngHTTP/1.1" (\"%r\")* – This indicates the request sent by the client to the server. *Get* is the method used. *dccstats/stats-spam-ratio.1week.png* is the item requested by the client. *HTTP/1.1* is the protocol used by the client.
- *200 (%>s)* – This represents the status code sent from the server. Code starting with 2xx represents the successful response, starting with 3xx represents the redirection, starting with 4xx represents the error in the client and starting with 5xx represents the error in the server.
- *7654 (%b)* – This entry indicates the size of the object returned by the server to the client. If nothing is returned, then "-" will be recorded.

4.1.2 Combined Log Format

This is the customized log format used by Apache server. The configuration of the combined log format is given below:-

LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\""

Entry from a sample log file in combined log format is given below:-

64.242.88.10--[07/Mar/2004:16:50:54-0800]"GET/dccstats/stats-spam-ratio.1week.png HTTP/1.1" 200 2341"http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"

Additional fields in combined log format of this log entry are described as follows:-

- "http://www.example.com/start.html" *(\"%{Referer}i\")* – This field represents the URL of the webpage from where the client is redirected to a particular webpage. In this example this should be the page that links to or include *stats-spam-ratio.1week.png.*
- *"Mozilla/4.08 [en] (Win98; I ;Nav)" (\"%{User-agent}i\")* – This field gives the information about the client browser that made request to the server.

4.1.3 Multiple Access Logs

In this type of logs three log files are created. It can be said as a combination of common log format and combined log format.

The configuration of the multiple access log is given below:

LogFormat "%h %l %u %t \"%r\" %>s %b" common

CustomLog logs/access_log common

CustomLog logs/referer_log "%{Referer}i -> %U"

CustomLog logs/agent_log "%{User-agent}i"

The first line contains the basic CLF information, while the last two lines contain referrer and browser information.

4.2 IIS Server Log Formats:-Two types of log formats are considered in IIS servers, which are discussed as follows:-

4.2.1 Microsoft IIS (Internet Information Services) log file format

Microsoft IIS log file format is non customizable ASCII Format. It records less information than W3C

format. Configuration of an entry in the log file is shown below: -

ip username date time status_code received_byteselapsed_timesent_bytes action target_file

4.2.2 W3C(World Wide Web Consortium) Extended Log File Format

IIS servers use this log format by default. In this ASCII text format is used and the time is recorded as UTC. This format is customizable. Configuration of an entry in the log file is shown below: -

date time c-ipcs-username s-sitename s-computername s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status time-taken cs-version cs-host cs(User-Agent) cs(Referrer)

Sample Entry is shown below: -

2002-04-01 00:00:10 1cust62.tnt40.chi5.da.uu.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=323-21-603&q=3&y=2002&id=671 200 156 http/1.1 www.cs.depaul.edumozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21) http://www.cs.depaul.edu/courses/syllabilist.asp

5. Data Preprocessing

Raw web log exist generally in unstructured form consisting of noise and missing values. It is difficult to use it directly for the analyses purpose. Only quality data can generate quality results[3]. The features of quality data are complete, accurate, and consistent. To generate this quality data we need to pre-process the raw data. Many researchers ignore this step, but it can be assumed as one of the most important part of the analysis. It is observed that 80% of the time during analysis is consumed in the preprocessing stage. Since preprocessing removes the irrelevant information, hence it reduces the size of data to be analyzed. Web log preprocessing phases comprisesdata cleaning, user identification, session identification, path completion[1]. These steps are displayed in the figure shown belowand discussed inthe next section in detail.

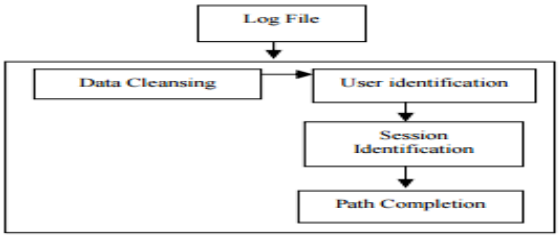

Figure2:block diagram for data preprocessing[1]

5.1 Data Cleaning

In this step irrelevant entries are removed from the web log file. This will considerably reduce the size of web log file making it more manageable and useful for the analysis purpose.

Irrelevant information in web log includes: -

*i. Graphics, video, and format information.*The records having .JPEG, .PNG, .GIF, .MPEG, .CSS extensions in cs_uri_stem field are removed.

ii. The records with failed HTTP status codes. The records having status less than 200 and greater than 299 are removed.

iii. The records with robot entries. These entries are made by the software (called as robots) used by search engines for updating their indexes after every fixed interval of time. Most of the robots declares themselves in the user_agent field of log file.

5.2 User and Session Identification

In this step different users and different sessions of various users are identified. Different sessions are identified by using referrer field. Different users are identified by different IP address.

*i. If the IP addresses are same***,** the operating system and browser information can be used to distinguish different users. Different browsers and different operating systems represent different users. This information is contained in user agent field.

*ii. If all of the IP addresses, browsers, and operating system details are same***,** the referrer information should be checked. A new session is identified if URL in ReferURI (cs_referer) has not been accessed previously or there is a large time gap (usually more than 30 mins) between the accessing times of these records.

5.3 PathCompletion[6]

This is an important and difficult phase. Path completion is used to access the complete user access path. The incomplete user access path is recognized on the basis of user session identification. There are chances of missing path because of proxy and caching problems. Various algorithms can be used for path completion like Maximal Forward Reference (MFR) and Reference Length (RL) algorithm. At the end of this step we will get the user session file.

6. Literature Review

After reviewing the research papers, it has been observed that the concept of web usage mining was introduced in late 90's to extract knowledge from the data present in the web logs for improving the websites as per user's convenience. Preprocessing is a major step in web usage mining because analysis could be better with preprocessed data. To carry on the present review work the researcher had studied the research papers from 2000 to 2015. In the upcoming paragraphs some of the important outcomes of the research are explored.

Naga Lakshmi et.al.[3]explained the different source of data and various log formats of Apache and IIS servers. Steps included in web log preprocessing and problems associated with user identification were discussed. Experimental results of web log preprocessing are displayed in the form of statistics.Ms. Deepa Dixit et. al.[1] suggested two approaches: pre-processing using XML and pre-processing using text file. Both the approaches were applied on a sample IIS server log file having extended log file format. Then results from these two approaches were compared and pros and cons of both the approaches were discussed.L.K. Joshila Grace et. al.[7] provided list of various types of web server logs andstatus codes sent from the server. Steps of web usage mining were discussed and also gave a brief overview on creation of extended log file.Theint aye[8]proposed algorithms for data cleaning and field extraction in web usage mining process. The statistical results obtained after applying the proposed algorithms were displayed.T. Revathiet. al.[9]explained various data fields contained in a web log file. The author has explained various steps included in preprocessing of a web log file – Log unification, Data Cleaning, User Identification, Transaction Identification, Data Integration and Transformation. To conduct preprocessing experiments on data, author has considered data from an offline website. And missing values are replaced with the most probable value for that attribute. Mitali Srivastava et.al.[10]described different technique for each step in preprocessing. For robot identificationin cleaning phase, various normal and some heuristic techniques were explained. Techniques for user identification such as user identification by IP address, user identification by authentication data and user identification using site topology were discussed. For session identification there exist two types of techniques reactive and proactive. Generally reactive techniques are preferred because proactive techniques require user's cooperation. Three reactive techniques- by the time gap, by the referrer attribute and by the time spent on observing page were explained.Tanasaet.al.[11]explained four steps of preprocessing: Data fusion, Data cleaning, Data structuring and Data summarization.In data cleaning accessorial resources such as css, multimedia files - images, audio, video etc. and robot's generated requests are removed. In Data structuringuser identification is done by authentication data or IP address, Session identification is done by host and agent, Page view identification is done by site map. At last in Data summarization pattern analysis is performed by using data generalization and aggregation. The author did not remove unsuccessful requests in data cleaning phase which is an important step in preprocessing of weblogs to reduce the number of records so that analysis can be performed well with less calculations.G. T Raju et. al.[12]discussed a preprocessing methodology to transform any collection of web server logs into structured collection of tables. The authorperformed a comparative analysis of techniques given by other researchers for each step in preprocessingand showed experiment results likes reducing size of log file for preprocessing, day wise unique visitors, user session identifications.ShailyLanghnojaet.al.[13]proposed algorithms for data cleaning, user and session identification. The author had used MS SQL server 2008 for loading the preprocessed data. The results after preprocessing were displayed.V.Chitraa&Dr. Antony SelvadossDavamani[6] presented an efficient

technique for path completion. They combined various individual path completion techniques like Maximal Forward Reference (MFR), Reference Length (RL) and Time window concept.

The author concludes that preprocessing is a very important step in web usage mining and if it is performed well then rest of the analysis can go smoother. The major challenge is to deal with unstructured or semi-structured data in efficient manner.The potential issues &challenges inpreprocessing of web logs are explored in the upcoming sections.

7. Potential Issues & Challenges

1. Web log files are unstructured in nature and cleaning for semi structured and unstructured data is a very difficult task. More research has to be done on cleaning of semi structured and unstructured data.

2. Data Transformation means converting a set of data values from one format to other. But no exact tools are available for this. So transformation becomes a tedious task.

3. There is no direct algorithm for identifying the different users accessing the website with same IP address. So this can be considered as a future area of research[14].

4. Since due to proxy servers, different users can have same IP address. So identifying a new session for a user is difficult.

5. There is no exact method available to find when a user has left the website. So it can be a potential area of research. A default 30 minutes timeout method is used to break a user's click into sessions [15].

6. Another important issue is inferring cached page references. For better performance browsers cache a webpages when it is referenced for the first time. The next time whenever user requests the same page browser may fetch the cached copy. This entry will not be captured by the log file.One proposed solution for this problem is referrer based method[15]. More research has to be done regarding this issue.

7. Query processing is difficult in case of heterogeneous data.

8. Conclusion

Web usage mining is one of the emerging areas of research. In this paper the researcher describes the potential issues and challenges in preprocessing of web logs. In order to take full advantage of web usage mining, it is important to carry out preprocessing stage effectively. Steps of preprocessing comprises ofdata cleaning, user identification, session identification and path completion. Once preprocessing stage is well-performed, techniques like clustering, association, classification etc.can be used efficiently for pattern discovery and pattern analysis phase of web usage mining. Hence web log preprocessing is a very important and integral part of WUM.

References

[1] D. Dixit and M. Kiruthika, "Preprocessing of Web Logs," *International Journal on Computer Science and Engineering*, vol. 2, pp. 2447-2452, 2010.

[2] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," *SIGKDD Explorations*, vol. 2, no. 1, pp. 1-15, 2000.

[3] N. Lakshmi, R. S. Rao and S. S. Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, no. 4, pp. 274-279, 2013.

[4] R. Cooley, B. Mobasher and J. Srivastava, "Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns," in *Knowledge and Data Engineering Exchange Workshop*, Newport Beach, CA, 1997.

[5] "Apache (UNIX) Log Samples," Monitorware, [Online]. Available: http://www.monitorware.com/en/logsamples/apache.php?&PrinterVersion=1.

[6] V. Chitraa and A. S. Davamani, "An Efficient Path Completion Technique for web log mining," in *International Conference on Computational Intelligence and Computing*

Research, 2010.

[7] L. K. J. Grace and D. Nagamalai, "Analysis of Web Logs and Web User in Web Mining," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 3, pp. 1-12, 2011.

[8] T. T. Aye, "Web Log Cleaning for Mining of Web Usage Patterns," in *International Conference on Computer Research and Development (ICCRD)*, Shanghai, 2011.

[9] T. Revathi, M. M. Rao, C. S. Sasanka, K. J. Kumar and B. U. Kiran, "An Enhanced Pre-Processing Research Framework for Web Log Data," *International Journal of Advanced Research in Computer Science and Software Engineering*, pp. 358-363, 2012.

[10] M. Srivatava, R. Garg and P. Mishra, "Preprocessing Techniques in Web Usage Mining: A Survey," *International Journal of Computer Applications*, vol. 97, pp. 1-9, 2014.

[11] D. Tanasa and B. Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining," *IEEE Intelligent Systems*, pp. 59 - 65, 2004.

[12] G. T. Raju and P. S. Satyanarayana, "Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology," *International Journal of Computer Science and Network Security*, vol. 8, pp. 179-186, 2008.

[13] S. Langhnoja, M. Barot and D. Mehta, "Pre-Processing: Procedure on Web Log File for Web Usage Mining," *International Journal of Emerging Technology and Advanced Engineering (IJETAE)*, vol. 2, 2012.

[14] Jitendra, B. Upadhyay and S. V. Patel, "A Review Analysis of Preprocessing Techniques in Web usage Mining," *International Journal of Engineering Research & Technology (IJERT)*, vol. 4, no. 4, pp. 1160-1166, 2015.

[15] J. Srivastava, R. Cooley and M. Deshpande, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, vol. 1, no. 2, pp. 12-23, 2000.