USER AND SESSION IDENTIFICATION FROM WEB DATA PREPROCESSING

Dr. UJWALA M. PATIL

R.C. Patel Institute of Technology, Shirpur, Maharashtra, India.
patilujwala2003@gmail.com

Web usage mining is the process to discover Web navigational patterns from Web log data. These navigational patterns are used to predict Web user behaviour. Whenever user interact with the Web live their footprints in the web log file at the server. Web server log data is in text format and each log entry is saved as a line of text. The Web log data file cannot be used as it is for pat-tern discovery algorithm. Therefore preprocessing is one of the important step before pattern discovery. The quality of Web log data after preprocessing increases the efficiency of the pattern discovery algorithm. In this paper we proposed algorithms for field extraction, data cleaning, user session, and transaction identification. Experimental results show the effectiveness of the proposed algorithms.

***Key words*:** Web usage mining, Web log data, Web user behaviour

1 Introduction

Web usage mining is an interesting area due to the popularity and widespread use in various fields. Day by day all the traditional applications are going to switch online. As an effect, there are increasing num-ber of Web sites and Web users. Web based applica-tions using web documents written in a standard formats like HTML and JavaScript. Web based ap-plications are supported by a variety of web browsers. It uses Hyper Text Transfer Protocol (HTTP) to visit set of Web pages through a Web browser. Through a Web browser, Web user generates a request to the Web server over the Internet. The web server re-sponds back to the request generated by the Web user in terms of a web page, image, video, etc. The Web user navigates through hyperlinks present in a Web page. When Web user navigates over the Web site, the Web server records all the actions followed by the Web user in Web log files.

Web usage mining is the subfield of Web mining. Web usage mining uses data mining techniques to discover information from navigation behaviour of World Wide Web users.

Web usage data is collected at Web server. Web server data is in text format and each log entry is saved as a line of text. The data in log file cannot be used as it is for pattern discovery algorithm. Therefore data preprocessing is the important step before pattern discovery. The quality of data after preprocessing increases the efficiency of the pattern discovery algorithm. Preprocessing includes basic steps like data cleaning, user identification, session identification to make data ready for the pattern discovery [1, 2, 3, 4, 5].

The paper presents four algorithms for preprocessing of Web usage log as field extraction, data cleaning, user and session identification, transaction identification. The main goal of preprocessing of

web usage log is to find Web user behavior. The ef-fectiveness of the preprocessing increases the accuracy in further steps of Web user behavior prediction.

This paper is organized as follows. Section 2 dis-cuss some related work. In section 3 input Web Log File Structure is discussed. In Section 4, proposed system architecture for preprocessing with algorithms field extraction, data cleaning, user and session iden-tification, and transaction identification. Section 5 describes experimental results with input dataset characteristics. Conclusions are given in section 6.

2 Current Practice and Research

Web usage mining analyzes Web log files to bring out interesting patterns. These derived patterns are useful for Web user behavior prediction [10, 11, 12].

Tasawar Hussain et al. made a survey on different preprocessing techniques. Along with basic steps involved, additional web session clustering is also used in preprocessing because similarity measure plays key role. He explained different web server log file formats like Common Log File Format (NCSA), Ex-tended Log Format (W3C), and IIS Log Format (Microsoft) and the attributes present in the log file. Web usage mining algorithms use different log data for their analysis. Based on the survey made by Tasawar Hussain et al. Web server log is more reliable to find the Web user behaviour [3].

Theint Aye presented two algorithms in data pre-processing. Every access made by Web user is not useful in the pattern discovery. Similarly Web server log records various attributes. Depending on the algorithm to be used in pattern discovery, all the attributes in the Web log file is not used. Therefore the first algorithm is used to extract fields from Web log file, and the second algorithm is used for data cleaning. He used Web log of department of engineering from university of Computer Studies, Mandalay and proved the efficiency of both algorithms [4].

Zakaria et al. designed a methodology, uses Web server log files to discover the knowledge from browsing patterns. This knowledge about users accessing web pages types is used for user profiling classes such as sports, economic, political, etc. There are certain situations occur while identification of users and/or sessions. This paper addressed the difficulties in users and/or sessions correctly due to single IP address with more sessions, more sessions with single IP address, more IP addresses with single User, more browsers with single user [5].

Bhuvaneswari et al. [6] made a comparative study on various tools used for prepressing of Web log files to analyse Web user behaviour. Each tool has its own features and limitations.

J.X. Yu et al. [7] described how to identify customers' behavior and classify them as visitors with purchase interest, visitors without purchase interest, and network robots. Classification is on the basis of different attributes derived from Web log file. Attributes are like the number of time of the visit, total session time, number of child pages accessed from a single page, the depth of the number of pages accessed from a single page, visitors stay time, use of GET, POST or Head access mode, and the frequency of access of images and graphic files. For experiment, a Berkeley log from the Computer Science Department at the University of California was used. It contains 431,066 records gathered

during 28-30 Sept. 2001 after preprocessing of Web log, 39,033 sessions are derived. Then from selected sessions, some records are used as a training dataset and some are used as a testing dataset.

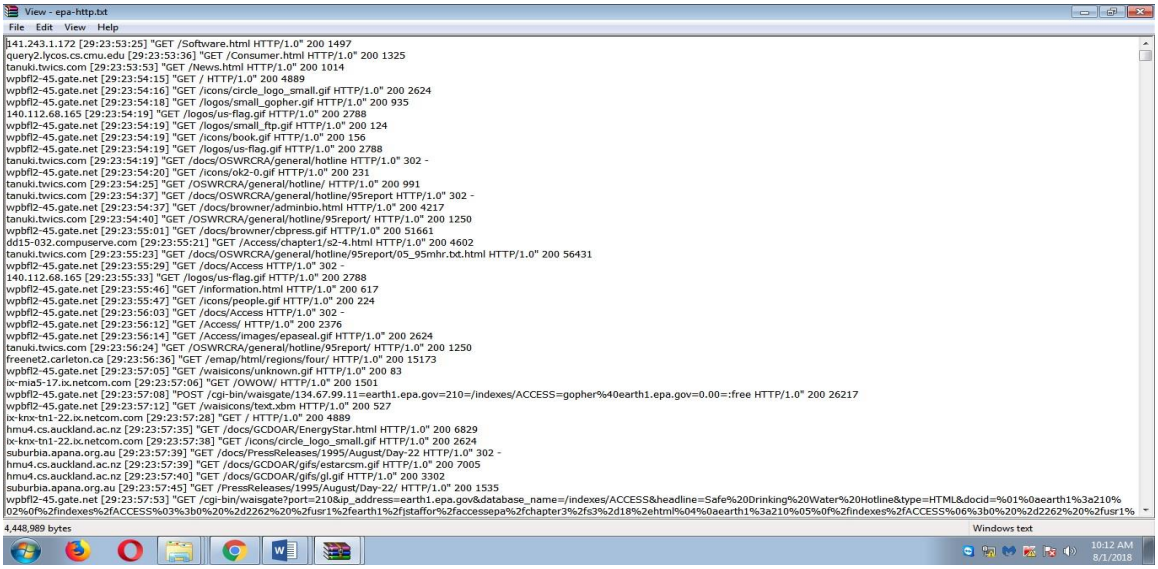

Figure 1. EPA-HTTP.txt file.

Table 1. EPA Dataset Parameters of Log File

Sr. No.	Name of the Parameter	Description
1	Host	Identify users Who visited the Website by its IP address of the client (remote host) which made the request to the server like 141.243.1.172
2	Time stamp	Date and time when user browsed like 29/Aug/1995:23:53:25
3	Request	Method used by the client is either GET, POST, HEAD.
		The Resource requested by client like Software.html
		The Protocol used by a client like HTTP.
4	Status code	HTTP reply code, whether the request resulted in a successful response or not.
5	Bytes	Bytes in reply indicate size of the object returned to the client.

2.1 Web Log File Structure

A Web log dataset used, is a real-world dataset from the United States EPA. It has a quantitative feature as the time spent viewing the Web page. The EPA dataset was a Web server log collected from a 24-hour period of HTTP requests. The location of the Web server is Research Triangle Park, NC, USA. The EPA dataset log was recorded from 23:53:25 EDT 29th August 1995 to 23:53:07 30th August 1995. It has 47748 requests: 46014 GET requests, 1622 POST requests, 107 HEAD requests

and 6 invalid requests. Figure 1 shows sample records from EPA-HTTP.txt file. It contains five parameters. The description of each parameter is given in the table 1.

3 Research Approach

In order to extract knowledge from Web log file, there are several problems exist to extract useful information from Web log file. Before applying preprocessing steps on input Web log file, various fields should be separated. A server log file is a text file where all the fields are space separated. By taking this text file as an input the first algorithm i.e. algorithm 1, separates all the fields and transfer the contents of Web log files to database table. The algorithm 1 used for field extraction is given below.

Algorithm 1 *FieldExtraction*
Input- tf: Web Log Text File
Output- lt: Log Table

Create a table lt with five attributes as host, timestamp, request, reply_code, reply_bytes to store log data from Web log text file.
For each record i $\in$ tf **do**
Begin
Read all fields contain in record i and separate out the all the fields
Add all the fields into the lt Log Table
End

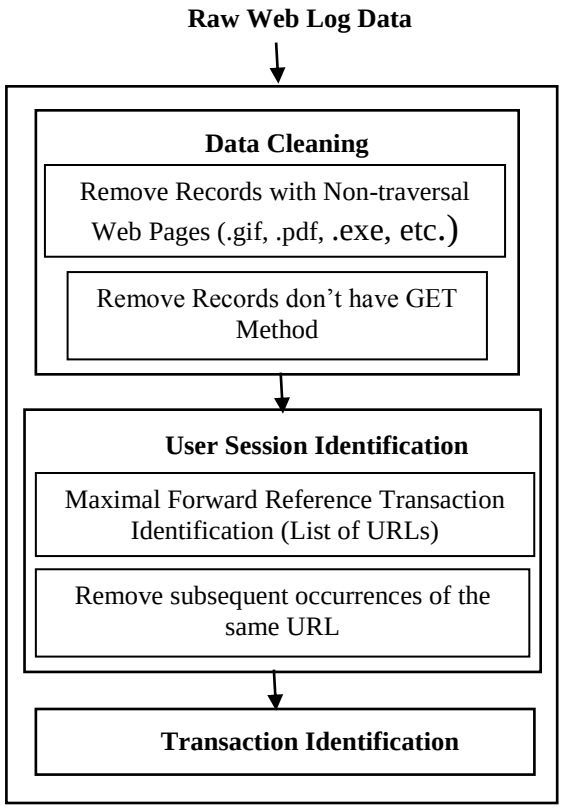

Figure 2. Preprocessing of Web Log Data

In order to extract valid information from Web log file necessary for pattern mining, apply general phases of data preprocessing. The phases include data cleaning, user and session identification, and transaction identification as shown in figure 2.

3.1. Data Cleaning

All the records in the Web log files are not used. It contains several irrelevant information like image files, document files, compressed and executable files, etc. After extracting fields from plain text file further processing was used to separate out all distinct suffixes of client resources in HTTP requests. The algorithm 2 was used for data cleaning.

Algorithm 2: *DataCleaning*

Input- lt: Log Table
Output- clt: Cleaned Log Table

Create a table clt with three attributes as host, timestamp, request to store cleaned log data from lt log table
For each record i $\in$ lt **do**
Begin
If reply_code = 200 and request = GET **then**
Begin
If URL.suffix not in {gif, xbm, zip, pdf, exe, gz, wpd, wp, jpg, imf} **then**
Begin
save record in clt i.e. $clt \leftarrow i$
End
End
End

3.2 User and Session Identification

Once the data are cleaned the next phase is user identification. For EPA dataset hostname of the requester is used for user identification.

Algorithm 3: *UserandSessionIdentification*

Input- clt: Cleaned Log Table, t: timeout threshold
Output- usid: User with sessions Identification Table

Create a table usid with three attributes as session_id, timestamp, request, time_spent to store user session information.

For each record i ∈ clt **do**
Begin
copy each record i to temp table.
End.
Sort the temp table by host and then by timestamp.
tot=0
For each $i.\text{host} \in temp$ **do**
Begin
XX:assign new session id
calculate the time difference between two consecutive log entries as time_spent
if tot + time_spent < t **then**
Begin
Add new session id, timestamp, request, time_spent to usid
End
else if tot + time_spent > t **then**
Begin
goto XX
End

End

There are various problems occur in exact user identification because of proxy servers and cache [8, 9]. The proxy servers hides the identity of the individual client or user to the outside world. All the request generated within organization or companies have the same IP address. Similarly, due to the server cache, requests generated may refer the cached Web page instead of requesting to the Web servers. Therefore a request made by the client will not appear on the Web server. Here, time window based user and session transaction identification was used. Along with time window, maximal forward reference identification was used for session identification. Algorithm 3 was used for user and session identification.

3.3 Transaction Identification

Once the sessions are identified, transactions are created over the sessions. Algorithm 4 was used for transaction identification.

Algorithm 4: *TransactionIdentification*

Input- usid: User with Sessions Identification Table
Output- tid: Transaction Identification Table

Create a table tid with three attributes as transaction_id, set_of_URLs, set_of_time_spent to store transaction information.
For each record $i \in usid$ **do**
Begin

transaction_id = i. session_id
set_of_URLs = all URLs in that session separated by space.
set_of_time_spent= all URLs time_spent in that session separated by space.

End

Table 2. Data after Field Extraction Algorithm

ID	Host	Date time	request	reply code	bytes
1	141.243.1.172	[29:23:53:25]	GET /Software.html HTTP/1.0	200	1497
2	query2.lycos.cs.cmu.edu	[29:23:53:36]	GET /Consumer.html HTTP/1.0	200	1325
3	tanuki.twics.com	[29:23:53:53]	GET /News.html HTTP/1.0	200	1014
4	wpbfl2-45.gate.net	[29:23:54:15]	GET / HTTP/1.0	200	4889
5	wpbfl2-45.gate.net	[29:23:54:16]	GET /icons/circle_logo_small.gif HTTP/1.0	200	2624
6	wpbfl2-45.gate.net	[29:23:54:18]	GET /logos/small_gopher.gif HTTP/1.0	200	935
7	140.112.68.165	[29:23:54:19]	GET /logos/us-flag.gif HTTP/1.0	200	2788
8	wpbfl2-45.gate.net	[29:23:54:19]	GET /logos/small_ftp.gif HTTP/1.0	200	124
9	wpbfl2-45.gate.net	[29:23:54:19]	GET /icons/book.gif HTTP/1.0	200	156
10	wpbfl2-45.gate.net	[29:23:54:19]	GET /logos/us-flag.gif HTTP/1.0	200	2788
11	tanuki.twics.com	[29:23:54:19]	GET /docs/OSWRCRA/general/ hotline HTTP/1.0	302	-

4 Experimental Results

All the experimental results are carried out on core i3 processor with 4GB memory. Input to the proposed system is a raw Web log file. In order to retrieve useful data out of raw log file, series of algorithmic steps are applied. Table 2 shows set of sample records from the EPA dataset after separate out the fields from raw log data file.

Table 3 shows the records after data cleaning algorithm. Once the data is cleaned then user and sessions are identified. After sessions identification the transactions are constructed over the sessions.

Table 3. Data after Cleaning Algorithm

8 Preprocessing of Web Log to Predict User Behaviour

ID	Host	Date Time	Request
1	128.104.66.114	[30:12:26:42]	GET / HTTP/1.0
2	128.120.153.224	[30:18:29:58]	GET / HTTP/1.0
3	128.120.153.224	[30:18:37:03]	GET /Software.html HTTP/1.0
4	128.120.153.224	[30:18:37:24]	GET /enviro/html/ef_home.html HTTP/1.0
5	128.120.153.224	[30:18:39:00]	GET /enviro/html/ef_overview.html HTTP/1.0
6	128.120.153.224	[30:18:39:49]	GET /enviro/html/ef_query.html HTTP/1.0
7	128.120.153.224	[30:18:40:13]	GET /enviro/html/pcs/pcs_overview.html HTTP/1.0
8	128.120.153.224	[30:18:40:23]	GET /enviro/html/emci/emci_overview.html HTTP/1.0

Figure 3 shows the session details which includes session length vs number of occurrences of the session. There are short as well as long sessions present in the dataset.

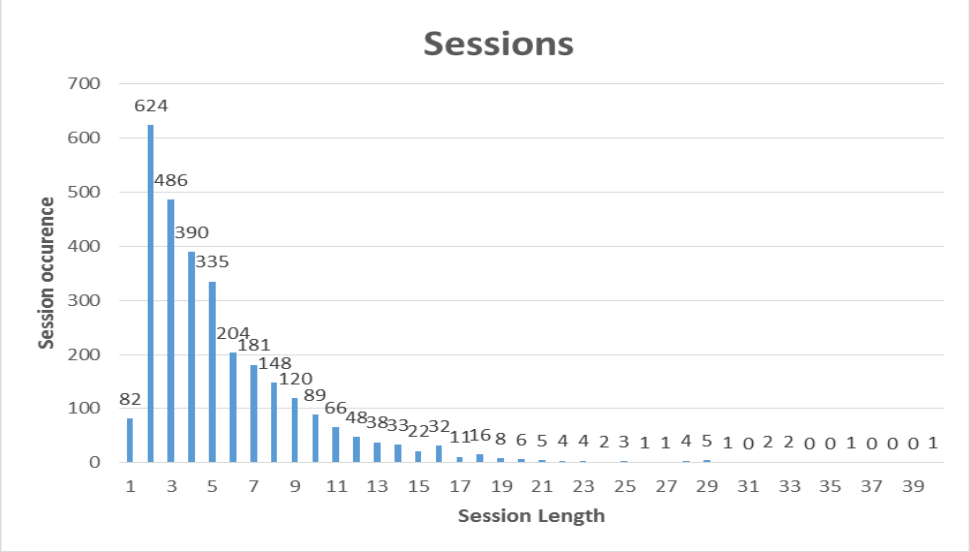

Figure 3. Session Length vs Session Occurrences

Table 4 shows overall EPA dataset characteristics before and after preprocessing. Initially, when we have downloaded the real-world EPA dataset, it has 47,748 records. In data cleaning, we have removed all the records with POST, HEAD and invalid requests. Along with that, we have also removed all the records with non-traversable pages. There are variations in the transaction size. Some transactions are shorter and some are longer. The maximum size of the transaction is 40 whereas the shortest transactions are of size 1. The transactions of size 1 are not useful therefore we have removed such transactions. Finally only 18,143 records are useful.

Table 4. EPA Dataset Characteristics

Before Preprocessing		After Preprocessing	
Data Volume	4.4 MB	Unique Host	1,299
Total Requests	47,748	Total Requests	18,143
GET Requests	46,014	Total URLs	5,147
POST Requests	1,622	Total Transactions	2688
HEAD Requests	107	Maximum Transactions Size	40
Invalid Requests	6	Average Transactions Size	6

5 Conclusions and Future Work

Data preprocessing is the one of the major phase in the Web usage mining process. This paper proposed algorithms for field extraction, data cleaning, user and session identification, and transaction identification. The proposed algorithms are effective to remove unnecessary data from raw Web log file. Most of the data irrelevant to the Web user prediction are removed. This reduction in data will effectively increases the efficiency of the further steps in Web user behaviour prediction.

Acknowledgements

This work is supported by "Vice Chancellor Re-search Motivation Scheme (VCRMS), NMU/11A/VCRMS/budget-2015-16/Engg. & Tech.-13/66/2016", to college teachers through a university fund of North Maharashtra University, Jalgaon.

References

1. R. Cooley, B. Mobasher, and J. Srivastava , "Data Preparation for Mining World Wide Web Browsing Patterns", Knowledge and Informati on Systems, vol. 1, no. 1, pp.5-32, 1999.
2. L.K. Joshila Grace, V. Maheswari, and D. Nagamalai , "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security Its Applications, vol. 3, no. 1, pp.99-110, January 2011.
3. T. Hussain, S. Asghar, and N. Masood, "Web usage mining: A survey on preprocessing of web log file", In 2010 International Conference on Information and Emerging Technologies (ICIET), pp.1-6, 2010.
4. T. T. Aye, "Web Log Cleaning for Mining of Web Usage Patterns", In 3rd International Conference on Computer Research and Development (ICCRD, pp.490494, March 2011.
5. Z.S. Zubi, M.S.E. Raiani, "Using Web Logs Dataset Via Web Mining for User Behavior Understanding", International Journal of Computers and Communications , vol. 8, pp.103-111, 2014.
6. Bhuvaneswari and S. Anand, "A Comparative Study of Different Log Analyzer Tools to Analyze User Behaviors", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 3, no. 5, pp. 2997-3002, May 2015.
7. J.X. Yu, Y. Ou, C. Zhang and S. Zhang, "Identifying Interesting Visitors through Web Log Classification", IEEE Intelligent Systems, vol. 20, no. 3, pp. 55-59, 2005.

8. R. Cooley, B. Mobasher and J. Srivastava, “Web Mining: Information and Pattern Discovery on the World Wide Web”, In Ninth IEEE International Conference on Tools with Artificial Intelligence, pp. 558 – 567, Nov. 1997.
9. R. Cooley, B. Mobasher, and J. Srivastava, “Data Preparation for Mining World Wide Web Browsing Patterns,” Knowledge and Information Systems, vol. 1, no. 1, pp. 5-32, 1999.
10. R. Gupta and P. Gupta, "Application specific web log pre-processing", International Journal of Computer Technology & Applications, vol. 3, no. 1, pp. 160-162, Jan-Feb 2012.
11. Jaswinder Kaur, and Dr. Kanwal Garg, "Analyzing the Different Attributes of Web Log Files To Have An Effective Web Mining", International Journal of Advanced Scientific and Technical Research, vol. 3, no. 5, pp. 127-134, May-June 2015.
12. O. M. Kumar and P. Bhargav, "Analysis of Web Server Log by Web Usage Mining for Extracting Users Patterns", International Journal of Computer Science Engineering, vol. 3, no. 2, pp. 123-136, Jun 2013.