Dharmendra Patel, Dr. Kalpesh Parikh, Atul Patel / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 1, Jan-Feb. 2012, pp. 327-330

Sessionization –A Vital Stage in Data Preprocessing of Web Usage Mining-A Survey

Dharmendra Patel*, Dr. Kalpesh Parikh, Atul Patel *****

*(Asst.Professor,Faculty of Computer Science and Applications, CHARUSAT, CHANGA,Gujarat,INDIA
** (Director,Intellisense IT,Ahmedabad,Gujarat,INDIA)
***(Principal,Faculty of Computer Science and Applications, CHARUSAT, CHANGA,Gujarat,INDIA)

ABSTRACT

The World Wide Web has impacted on almost ever aspects of our lives in modern era. The Web has many unique characteristics and which make mining useful information and knowledge a challenging task. Web mining uses many data mining techniques but it is not an application of traditional data mining due to heterogeneity and unstructured nature of the data on Web. Web mining tasks can be categorized into three types: Web Structure Mining,. Web Content Mining and Web Usage Mining. The goal of Web Usage Mining is to capture, model and analyze the behavioral patterns and profiles of users interacting with a Web Site. Web Usage Mining consists of many stages but this paper focuses on first stage i.e data preprocessing. Data Preprocessing consists of data cleaning, page view identification, sessionization, data integration and data transformation. This paper focuses on most complex part of data preprocessing and that is Sessionization.This paper covers many important aspects of sessionization stage which are very useful for research scholars who are doing research work in Web Usage mining field.

Keywords – **Preprocessing, Sessionization, Web Content Mining, Web Structure Mining, Web Usage Mining**

INTRODUCTION

Web Usage Mining Process, , if it uses the standard data mining process[12], can be divided into three inter dependent stages (i) Data Collection and Preprocessing (ii) Pattern Discovery and (iii) Pattern Analysis. An important task in Web Usage Mining is the creation of an effective target data set to which web mining algorithms can be applied. Usage data preparation presents a number of preprocessing tasks such as data fusion, data cleaning, user session identification, page view identification and episode identification [7]. The primary data sources used in the process of Web Usage Mining are the server log files. Server log files consist of Web server access logs and application server logs. In addition to log data it also requires some additional data sources like site files,meta data, operational data bases, application templates and knowledge of domain. [4]Describes primary groups of data obtained through various sources. Log files consist of large amount of irrelevant information so data from log files can not be directly use for procedures of Web Usage Mining. Preprocessing on web usage data is required to eliminate noisy data and make data effective for further analysis task. Preprocessing contains many stages like data cleaning, page view identification, sessionization, data integration and data transformation. Many authors have worked on data cleaning stage [3][11] but very less work has done on sessionization stage. Sessionization is very important stage of data preprocessing which is used to identify the behavior of user and that information is very crucial for many applications. The section I describes

the different strategies of sessionization in details. The section II deals with number of software tools available for the purpose of sessionization. The Section III deals with the problems of sessionization and their related solutions. The last section describes the conclusion of the paper.

I STRATEGIES OF SESSIONIZATION

Sessionization is the determination of the number of visitors to a Web site. The user session identification is very important for the traffic characterization purpose. Web users transaction can be transformed into number of sessions by different strategies describe in [10]. Table -1 describes characteristics of different sessionization strategies.

II SOFTWARE TOOLS FOR SESSIONIZATRION

This section of paper deals with many software tools available in market for the generations of sessions form raw log data.In addition to sessionization they are very useful for web data analysis.

[1] WebSpy Vantage: - This is a very powerful tool that transforms raw log data into manageable information. Sessionization activity is efficiently handled by this software.It supports 200 log formats from many different vendors. For more information you can visit http://www.webspy.com.

[2] Relax :- It is a free specialized web server log analysis tool. Relax supports logs in RefererLog, Apache combined, NCSA extended/combined, TransferLog, and WebSTAR format.It is distributed under GNU General Public License(GPL).You can download this software from http://ktmatu.com/software/relax

[3] Weblizer Xtended : - This is a very powerful tool for web analysis and produces many statistics related to traffic.It is also very good tool for sessionization purpose.For more detail you can refer http://www.patrickfrei.ch/webalizer/

[4] Awstats Advanced Web Statistics:- It is a web server log analyzer that generates graphical statistics from web logs. It has a large number of features, including the ability to count unique visitors, most popular page, domains/countries of visitors, rush hours, browsers, operating system used, robot visits, search engines, keywords used in search engines, HTTP errors etc. You can refer http://awstats.sourceforge.net/ for more details of this tool.

[5] W3Perl :- This tool consists of set of Perl Script that can analyze log files for IIS,Apache,FTP,mail etc. upports sessions (length of time visitors spend on your site), RSS stats, referrers, keywords used on search engines, list of error pages invoked, classifcation of your visitors by countries, browser stats, screen sizes, real-time statistics, etc. The software is free and licensed under the GNU GPL. http://www.w3perl.com/ describes more details about this tool.

[6] Logrep :- This tool analyze and present the data obtained from various logs (e. web log,mail log,event log etc) and generate statistics for the system.The source of this tool is licensed under the GNU General Public License(GPL).For more detail you can log in to http://itefix.no/cgi-bin/itefix/logrep

[7] Analog Log File Analyzer: - This tool is very popular and used on number of web hosts to produce exhaustive reports of web data. It works in any operating system and freely available. For more detail login to http://www.analog.cx/

[8] Web Lizer :- It is a fast, free web server log file analysis tool. It produces highly detailed, easily configurable usage reports in HTML format, for viewing with a standard web browser.The main thing is that it supports unlimited log file sizes and partial logs.

[9] RU-Software Log Analyzer :- It is very powerful tool that provides quick and easy way to know visitors activity like hits,hosts, sessions etc. from raw log file. It generates lots of statistical reports not only for all sites but also for any directory of your site. More detail

about this tool is given in the web site http://www.download32.com/ru-software-log-analyzer-i38937.html.

III SESSIONIZATION PROBLEMS AND RELATED SOLUTIONS

Sessionization stage of web data preprocessing is very complex stage of Web Usage Mining process. This section deals with main problems arises in this phase and available solution related to problem. Following are different problems description which is arising in the phase of sessionization.

[1] Problem due to caching which is performed either by proxy server or browser. Caching problem causes a single IP address to be associated with different user sessions so it is quite difficult to identify user based on IP address. This problem can be solved by two main perspective one is use of cookies [8] and second one is URL rewriting or by requiring the user to log in when entering the web site [2].

[2] HTTP protocol is stateless so it is impossible to determine when a user actually leaves the web site in order to determine session finish time. Heuristic based solution of above mentioned problem is given by many authors [1][5][9].

[3] Important information passed through POST method will not be available in server log so it is difficult to form a session. Packet Sniffing is an alternative method to collecting usage data through server logs.

[4] User may visit the site more than once so the server logs records multiple sessions for each user. The solution of above mentioned problem is given by [6].

[5] Dynamic IP addresses can create a problem during sessionization process.The solution of the problem is to establish cookie and URL encoding kind of mechanism to identify unique user from transaction.

IV CONCLUSION

In this paper we have presented most critical issues of most complex stage of data preprocessing and that is sessionization. We provided different strategies of sessionization in details so any one can select appropriate strategy based on an his application area. This paper also deals with different software tools available in market for any kind of web analytics purpose including sessionization from raw log data. Finally this paper describes different problems may arises in the stage of sessionization and what should be the related solution of that problem so any one can easily resolve it.

REFERENCES

[1]BETTINA BERENDT, BAMSHAD MOBASHER, MIKI NAKAGAWA, AND MYRA SPILIOPOULOU. The impact of site structure and user environment on session reconstruction in web usage analysis. In Proceedings of the 4th WebKDD 2002 Workshop, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2002), 2002.

[2] Corin R. Anderson. A Machine Learning Approach to Web Personalization. PhD thesis, University of Washington, 2002.

[3]Hussain T., Asghar S., Masood N, Web log cleaning for mining of web usage patterns,IEEE ICCRD,2011,pp.490-494.

[4] J.Srivastava,R.Cooley,M.Despande and P.Tan,Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,SIGKDD Explorations,pp.12-23,2000.

[5]Mao Chen, Andrea S. LaPaugh, and Jaswinder Pal Singh. Predicting category accesses for a user in a structured information space. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in in- formation retrieval, pages 65–72, 2002.

[6]Myra Spiliopoulou, Bamshad Mobasher, Bettina Berendt, Miki Nakagawa, A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis.

[7]R.Colley,B.Mobasher and J.Srivastava,Data Preparation for Mining World Wide Web Browsing Patterns,Knowledge and Information Systems,pp.5-32,1999.

[8] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD thesis, University of Minnesota, 2000.

[9]Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1(1):5–32, 1999.

[10]Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analyis. INFORMS Journal on Computing.

[11]Tanasa D, Trousse B, Advanced Data Preprocessing for intersites Web Usage Ming,IEEE Intelligence System, pp.59-65. [12]U.M.Fayyad,G.Piatetsky-Shapiro and P.Smyth.From Data Mining to Knowledge Discovery: An Overview.In Advances in knowledge discovery and data mining,AAAI/MIT press,pp.1-34,1996.

Sr.No	*Strategies*	*Description*	*Advantages*	*Disadvantages*	*When Useful*
1.	*IP Address+Agent +Sessionization Heuristic.*	*Each unique pair of IP address/Agent is a unique user. Heuristic in terms of time is applied to generate session.*	*(i)Always available.* *(ii) No Additional Technology is required to determine unique user.*	*(i)Defeated by Rotating IPs.* *(ii) Selecting time for Heuristic is quite challenging.*	*Web User identity does not appear explicitly in the registers*
2.	*Embedded Session Ids.*	*It uses dynamically generated pages to associate ID with every hyperlink.*	*(i)Always available.* *(ii) Independent of IP address.*	*(i) Additional overhead for dynamic page* *(ii) It is not possible to capture repeat visitors.*	*When Session Ids are used by an application to uniquely identification of user.*
3.	*Cookie+* *Sessionization Heuristic*	*Save ID on client machine to identify unique user. Heuristic in terms of time or ID is applied to generate session.*	*Able to track repeat visits from same browser.*	*It is possible that cookie may be turned off by user.*	*When coolie is available for further analysis.*
4.	*Registration*	*User explicitly logs on site.*	*Able to track individual not only browser.*	*We can't force the user for registration.*	*When explicit registration is available.*
5.	*Software Agents*	*According to this technique program loaded into browser it sends back usage data.*	*Accurate data for a single site.*	*It may likely to be rejected by users.*	*When software agent is loaded with browser.*

(Table-1 Characteristics of Sessionization Strategies)