IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 4 (Mar. - Apr. 2013), PP 76-79
www.iosrjournals.org

A Survey on Data Preprocessing in Web Usage Mining

Murti Punjani[1], Mr.Vinitkumar Gupta[2]
[1](Department of Computer Engineering,Hasmukh Goswami College of Engineering, India
[2] (Department of Computer Engineering,Hasmukh Goswami College of Engineering, India

Abstract : *With the abundant use of Internet and constant growth of users, the World Wide Web has a huge storage of data and these data serves as an important medium for the getting information of the users access to web sites which are data stored in Web server Logs. Today people are interested in analyzing logs file as they show actual usage of web site. But the data is not accurate so preprocessing of Web log files are essential then after that data are suitable for knowledge discovery or mining tasks. Web Usage Mining, a part of Web mining and application of data mining is used for automatic discovery of patterns in clickstreams and associated data collected or generated as a result of user interactions with one or more Web Sites. This survey paper gives the literature review and also overview of various steps needed for preprocessing phase.*

Keywords *– Data Fusion, Path Completion, Pre processing, Session Identification, Web usage, Web Server Log file.*

I. INTRODUCTION

With the fast growth of Internet technology, preprocessing is necessary to get useful information about user access and is of the most important research topics. With the explosive use of growth of information available on WWW (World Wide Web), discovery and analysis ofuseful information has become necessity. The Web has become an important medium to communicate ideas, transact business and promote entertainment. The discovery and analysis of useful information from the Web documents is referred to as Web mining [1].The data is stored in web server log and it is in heterogeneous form. So we need to preprocess these data to extract useful information. Web Mining is divided into three categories [11]1.Web Content Mining 2.Web Structure Mining 3.Web Usage Mining. Web Content Mining is process to extract useful information from the contents of web documents. Web Structure Mining is the process of discovering structure information from the web. Structure represents hyperlinks and document structure. Web Usage Mining is application of data mining used to extract user access from web server log files.

A. Web Usage Mining

Also known as Web Log Mining is used to discover patterns from web server logs. The primary source of data for web usage mining consists of textual logs collected from several web servers all around the world. There are four phases in web usage mining. [4]

1. Data Collection- User Logs are collected from client and server side servers, proxy servers, application servers etc.
2. Data Preprocessing- Consists phases like data fusion and cleaning, user identification, session identification, path completion
3. Pattern Discovery- Discovering patterns from preprocessed data using various data mining techniques like statistical analysis, association, clustering, and pattern matching and so on.
4. Pattern Analysis-Once patterns are discovered, analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

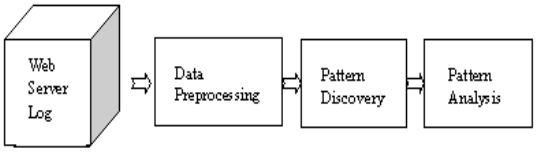

Fig. 1 Phases of Web Usage Mining

1.1 Motivation

Thousands of users access multiple web sites all over the world. When the different users access the websites, huge amount of data is gathered in the web log files which is very much useful many times as we can know how many times user access the same page frequently. These data can be further used to get user access

pattern and user behavior. As the data cannot be directly used in WUM, Preprocessing is necessary. Preprocessing of the web log file is tedious job and it takes 80% of total time of web usage mining process as whole [12].Seeing the advantages and disadvantages, we conclude that preprocessing is significant phase and which also improves quality of the data [13].

II. Literature Review

The aim of literature review is to study and compare the various available techniques for preprocessing. Due to huge amount of extraneous and inaccurate entries in web log file, log file cannot be directly used in WUM process so preprocessing is must.

According to Ravindra Gupta and Prateek Gupta [17], in which two main tasks are done which are customized web log preprocessing and improved FP Tree algorithm. Raw web log file was taken as input. The authors modified the algorithm FP tree and proposed improved FP tree algorithm. The proposed algorithm was divided into two main processes: creation of modified FP tree and mining. In modified FP tree algorithm, structure items were stored in descending order of their frequency. Customized preprocessing steps were Customization in which log cleaning was performed on basis of user requirement, next steps were Data Cleaning, User Identification, Session Identification and last step was database of cleaned log.After applying these steps compressed log file having user access behaviour in numeric form was generated and which can be further sent for mining using modified FP tree algorithm.

According to Wahab, et al, [16] discussed different types of log files in detail. Also discussed all the 19 attributes of web log file as well as different log file formats in detail. They proposed an algorithm for reading server logs and also algorithm for transferring the log file to database was proposed. After reading the log web files of any one type out of three formats, various attributes were ignored because they were considered not significant for the analysis. Data filtering was performed to remove unwanted attributes of web log file. The web server log file was containing 18 attributes, out of which 17 attributes were removed considering them as unwanted and only one attribute was known i.e. "URL" and was stored in the database. Some important attributes were not considered, so reliability was not maintained. So seeing the pros and cons, the proposed algorithms need to be modified.

According to Raju and Satyanarayana,[6], input was raw web log file collected from NASA Web site during July 1995. Customized preprocessing steps generated compressed log file having user behavior in numeric form which was further given for mining process using modified FP tree algorithm. It outputs complete relational database model for storing the structured information about the Web site, its usage and its users.

As web log file contains important data related to website, Suneetha and Krishnamoorthi[15], the input was the web log data of NASA website. Here the authors discussed the sources of web logs, web log structure and status codes of HTTP in detail. They performed preprocessing techniques on web server log file and first step was Data cleaning in which the irrelevant entries were removed like the entries that having status error or failure and images pages were removed next step was user identification in which three attributes were used from log file which are IP Address, Operating System, and User Agent. The output which can be further used to increase the effectiveness of the website. The authors did not apply session identification phase.

Author Name	**Source of log file**	**Preprocessing Technique**	**Algorithm applied**
Ravindra Gupta, Prateek Gupta	Raw web log file	Data Cleaning User Identification Session Identification Formatting	Improved FP Tree Algorithm
Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd and Mohamad Farhan Mohamad Mohsin	Server Log File	File Reading Data Cleaning Data Filtering	Proposed
Raju and Satyanarayana	Server Log File	Data Merging Data Cleaning User Identification Session Identification	NA
Suneetha, K. R. and D. R. Krishnamoorthi	Server Log File	Data Cleaning User Identification	NA

TABLE 1 Summary of Literature Review

III. Data Preprocessing Tasks

Fig 2 shows the phases of Data Preprocessing in Web Usage Mining**.** The goal of preprocessing is to transform the raw click stream data into a set of user profiles [5]. Data preprocessing presents a number of unique challenges which led to a variety of algorithms and heuristic techniques for preprocessing tasks such as merging and cleaning, user and session identification etc [6]. Input to the preprocessing stage is web server log file. Web Server Log contains 19 attributes such as *Date*, Time, Client IP, AuthUser, ServerName, ServerIP, ServerPort, Request Method, URI-Stem, URI-Query, Protocol Status, Time Taken, Bytes Sent, Bytes Received, Protocol Version, Host, User Agent, Cookies, Referer.

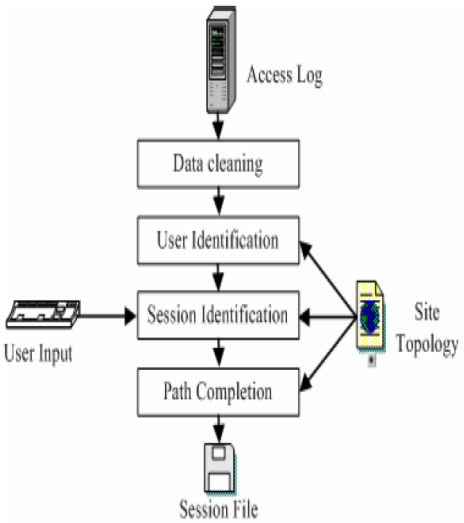

Fig 2: Phases of Data Preprocessing in Web Usage Mining[3]

Sample Log file is given below [3]:

2007-12-06 05:22:16 ::1 GET /iisstart.htm - 80 - ::1 Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+6.0;+SLCC1;+.NET+CLR+2.0.50727;+Media+Center+PC+5.0;+InfoPath.1;+.NET+CLR+1.1.4322;+.NET+CLR+3.5.21022;+.NET+CLR+3.0.04506) 200 0 0 296 336

3.1 Data Fusion and Cleaning

Merging of the log files from various Web and application servers is done at the Data Fusion phase.

```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2011-03-08 00:31:51
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem
cs-uri-query s-port cs-username c-ip cs-version cs(User-Agent) cs(Cookie)
cs(Referer) cs-host sc-status sc-substatus sc-win32-status sc-bytes cs-
bytes time-taken
2011-03-08 00:31:50 W3SVC23804 C36280-2214 74.52.2.99 GET /robots.txt -
80 - 77.88.29.246 HTTP/1.1
Mozilla/5.0+(compatible;+YandexBot/3.0;++http://yandex.com/bots) - -
www.abc.org 200 0 0 615 182 203
2011-03-08 00:38:24 W3SVC23804 C36280-2214 74.52.2.99 GET /Index.php - 80
- 114.80.93.57 HTTP/1.0 Sosospider+(+http://help.soso.com/webspider.htm)
- - www.abc.org 200 0 0 28541 161 1968
2011-03-08 00:45:18 W3SVC23804 C36280-2214 74.52.2.99 GET /md10-11.pdf -
80 - 66.249.71.201 HTTP/1.1
Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)
- - www.abc.org 304 0 0 274 287 46
```

Fig. 3 Web Log File in Text format [2]

The goal of data cleaning phase is to remove the extraneous and redundant log entries. Important fields like date, time, Client IP, User Agent, URL requested, URL referred, time taken, Referrer or browser used are considered for further processing. Extraneous or redundant data is to be removed which are[2] i)As we want only the log information related to user access so as HTTP is stateless protocol, graphics and scripts are also recorded. So extensions of the files are checked and files having extensions like .css, .gif, .jpeg, .gif, .jpg etc files are eliminated. ii) Removal of Robots request iii) some entries will be having errors. Eliminate the entries having status code less than 200 and greater than 299 as they are failure entries.

3.2 User Identification

This phase identifies individual user by using their Client IP address. If new IP address, there is new user. If IP address is same but browser version or operating system is different then it represents different user. [7]

3.3 Session Identification

Session of a particular user means how much time the user is connected to particular website. It tells us total page accesses of particular user. The following rules we use to identify user session in our experiment: [3]
1) If there is a new user, there is a new session;
2) In one user session, if the refer page is null, there is a new session;
3) If the time between page requests exceeds a certain limit (30 or 25.5mintes), it is assumed that the user is starting a new session.

3.4 Path Completion

After session identification, path completion comes. As the client uses proxy servers and cache version of the pages using 'Back', the sessions which are identified have many lost pages. So this phase is used to identify lost pages.

IV. Conclusion And Future Work

Preprocessing of web log file is mandatory step for web usage mining. After data cleaning step, we can go for preprocessing step by which we can extract user access pattern and also can be used further for pattern analysis. In this paper, various current preprocessing techniques are outlined. In this paper also I have explained the various tasks needed for preprocessing of the data in web usage mining. My future work is to increase the performance of the web server by getting meaningful and useful information quickly. Analyzing web server log files, we can easily understand the user behaviors in web structure to get better design of web components and web applications.

References

[1] O.Etzioni, The World Wide Web: Quagmire or gold mine.Communications of the ACM, 39(11):65–68, 1996.
[2] Vijayashri Losarwar, Dr. Madhuri Joshi,Data Preprocessing in Web Usage Mining, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore
[3] Li Chaofeng,Research and Development of Data Preprocessing in Web Usage Mining, School of Management, South-Central University for Nationalities, Wuhan 430074, P.R. China
[4] V.Chitraa, Dr. Antony Selvdoss Davamani, A Survey on Preprocessing Methods for Web Usage Data, (IJCSIS) International Journal of Computer Science and Information Security,Vol. 7, No. 3, 2010,p.78-83
[5] Demin Dong,Exploration on Web Usage Mining and its Application, IEEE, 2009.
[6] Raju G.T. and Sathyanarayana P. Knowledge discovery from Web Usage Data : Complete Preprocessing Methodology, ", IJCSNS 2008
[7] Priyanka Patil,Ujwala Patil, Preprocessing of web server log file for web mining ,World Journal of Science and Technology 2012, 2(3):14-18 ISSN: 2231 – 2587
[8] Marathe Dagadu Mitharam ,Preprocessing in Web Usage mining, International Journal of Scientific & Engineering Research, Volume 3, Issue 2, February -2012 1 ISSN 2229-5518
[9] C.P. Sumathi, R. Padmaja Valli , T. Santhanam, "An Overview of Preprocessing of Web Log Files for Web Usage Mining", Journal Of Theoretical And Applied Information Technology 31st December 2011. Vol. 34 No.2,P.178-185
[10] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan Web usage mining: Discovery and applications of usage patterns from web data.SIGKDD Explorations, 1(2):12–23, 2000
[11] Alam, S., G. Dobbie, et al. (2008). Particle Swarm Optimization Based Clustering Of Web Usage Data. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 978-0-7695-3496-1/08 DOI 10.1109/WIIAT.2008.292 IEEE/WIC/ACM International Conference on Web.
[12] Pabarskaite, Z. (2002). Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining. 24th Int. Conf. information Technology Interfaces /TI 2002, June 24-27, 2002, Cavtat, Croatia
[13] Han, J. and M. Kamber (2006). Data Mining: Concepts and Techniques. A. Stephan. San Francisco,, Morgan Kaufmann Publishers is an imprint of Elsevier.
[14] Yuan, F., L.-J. Wang, et al. (2003). Study on Data Preprocessing Algorithm in Web Log Mining. Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.
[15] Suneetha, K. R. and D. R. Krishnamoorthi (2009)."Identifying User Behavior by Analyzing Web Server Access Log File." IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.
[16] Wahab, M. H. A., M. N. H. Mohd, et al. (2008). Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm. World Academy of Science, Engineering and Technology 48 2008.
[17] Ravindra Gupta and Prateek Gupta,Application Oriented Web Usage Mining with Customized Web Log Preprocessing & Frequent Pattern Tree, International Journal of Engineering Research and Applications, ISSN: 2248-9622,Vol. 2, Issue 1,Jan-Feb 2012, pp.596-598