ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

A Survey on Web Usage Mining Preprocessing

[1] Dr. P.Sumathi, [2]B.Uma Maheswari

[1]Asst. Professor, Govt. Arts College, Coimbatore, India

[2]Research Scholar, Dept. of Computer Science & Applications, Bharathiyar University, Coimbatore, India

ABSTRACT: Web mining is to discover and extract useful information. In the internet era web applications are increasing at enormous speed and the web users are increasing at exponential speed. As number of users grows, web site publishers are having increasing their information for attracting and satisfying users. it is possible to trace the users' essence and interactions with web applications through web server log file and Web log file contains only (.txt) file. The data stored in the web log file consist of large amount of eroded, incomplete, and unnecessary information. Because of large amount of irrelevant data's available in the web log file, an original log file cannot be directly used in the web usage mining. So preprocessing technique is applied to improve the quality and efficiency of a web log file. Different techniques are applied in preprocessing that is data cleaning, data fusion, data integration. In this paper we will survey different preprocessing technique to identify the issues in web log file and to improve web usage mining preprocessing for pattern mining and analysis.

KEYWORDS: Web Usage mining, Data pre processing, Data mining, Server logs, Users and User sessions.

I . INTRODUCTION

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It consists of following tasks: 1. Resource finding: It involves the task of retrieving intended web documents. It is the process by which we extract the data either online or offline resources available on web. 2. Information selection and pre-processing: It involves the automatic selection and pre processing of specific information from retrieved web resources. This process transforms the original retrieved data into information. The data is transformed into useful information by using suitable transformation. The transformation could be renewal of stop words, or it may be aimed for obtaining the desired representation such as finding particular format of data. 3. Generalization: It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization 4. Analysis: It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web.

II. WEB MINING TAXONOMY

Web mining is broadly classified into three types based on the type of the data to be mined:
1. Web Content Mining:
Web content mining is the process of extracting useful and valuable information from the contents of web documents. Content data is the collection of data from which a web page is designed. It may consist of text, images, audio, video, or structured records such as lists and tables.

2. Web Structure Mining::
Web structure mining is the process of discovering structured information from the web. The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. This can be further divided into two kinds based on the kind of structure information used. First is Hyperlink Structure that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and the hyperlink that connects two different web pages is called an inter-document hyperlink. And the second one is Document Structure that contains the content within a Web page that can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

3. Web Usage Mining:
Web usage mining is the application of data mining techniques for discovering interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity and origin of web users along with their browsing behavior at a web site. Web usage mining tries to make sense of data generated by the web surfer"s session or behaviour . Web usage mining itself can be classified further depending on the kind of usage data considered. First one is Web Server Data in which user logs are collected by the web server and typically include IP address, page reference and access time. Second is Application Server Data which track various kinds of business events and log them in application server logs. And third one is Application Level Data in which new kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these specially defined events.

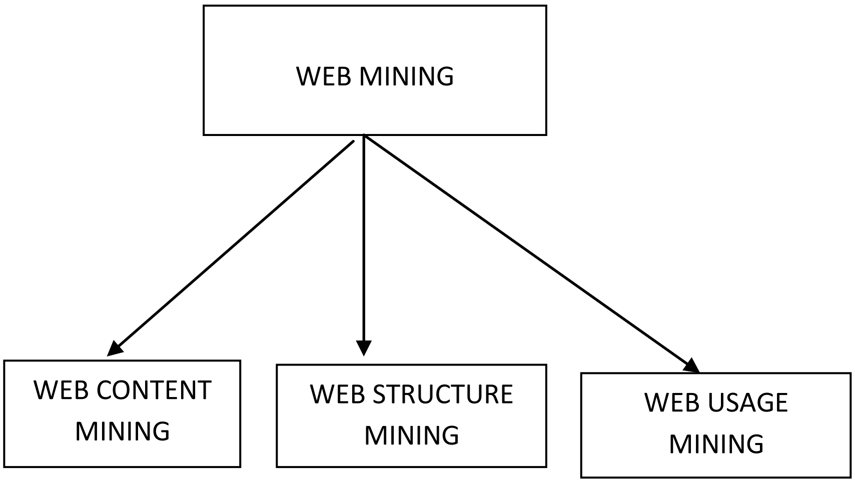

Fig.1. Kinds of Web Mining

The three important phases of web usage mining are explained here.

A. *Preprocessing*

*P*re-processing of web log data is the foremost important step before doing analysis on any kind of data. Web Log Pre-processing is a process of applying pre-processing steps on web logs. Web Log Pre-processing includes few of the steps like Data Cleaning, User Identification, Session Identification, Data Integration, Data Transformation and Path Completion. Data Pre-processing is a necessary foremost step before analyzing anything to improve the quality of results.

B. *Pattern Discovery*

Once after the web data is pre-processed in pre-processing step, the required interesting patterns or rules can be discovered out of web data by applying statistical methods and as well as data mining methods like Clustering, Association Rules, Classification Rules, Path Analysis, etc. Patterns discovered can be represented in some visual form, graphs, charts and tables, etc. Algorithms like FP-Growth and Apriori are used in this phase.

C. *Pattern Analysis*

In this phase the uninteresting patterns from the patterns discovered in previous pattern discovery phase are removed. And also the discovered patterns are analyzed by making use of some of OLAP tools or by SQL query mechanism.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

III. WEB LOG PRE-PROCESSING

Web logs will be containing raw, unnecessary and some irrelevant data. For further processing all these kinds of data has to be removed in Web Log Pre-processing phase so as to improve the quality of the data. In general Web Log Pre-processing involves the following steps .

A. *Data Cleaning*

Data Cleaning is a first important step in web log pre-processing to remove out all the irrelevant data from the web log files. This step helps reduce the size of the data by removing out unwanted log data and also improve the quality of data. Then further operations can only be applied on filtered data . For example the records referring images, graphics or video etc. are removed. And also the records with failed HTTP status codes are eliminated.

B. *User Identification*

The very next step after the data cleaning is identification of users. Here users can be identified by unique IP address or the unique ID assigned by the server (cookies). User Identification is required to categorize User accesses of websites or pages. Each user will be identified by a unique IP address and if different users will have same IP address then the different browser and different operating system represent different user.

C. *Session Identification*

Session Identification step of web log pre-processing phase is for identifying each individual user session. The goal of this step is to group page accesses or activities of each user into individual session. One of the common methods used to identify session is timeout mechanism.

D. *Data Transformation*

Web log records will be containing much number of fields each of which represents different kind of data. As it will become difficult to deal with different kinds of data the transformation step can be applied on web log records to transform the data into the format which is easy and relevant to process in further phases.

E. *Path Completion*

Path Completion step of web pre-processing is used to acquire complete user access path. This step is used to fill in missing page references. In some situations like when the user click on browser‟s back button while browsing it may result into incomplete user access path which will remain not entered in web logs, so to work with this kind of situations the Path Completion step is necessary.

DOI: 10.15680/IJIRCCE.2015. 0311219

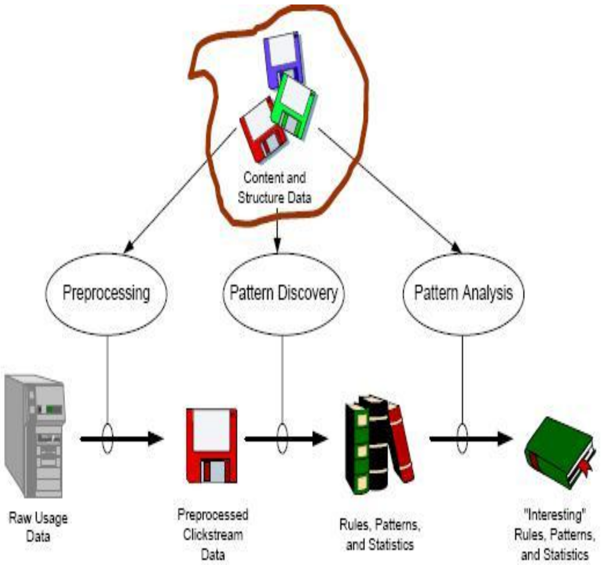
Content and Structure Data
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Usage Data
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Fig.2. Phases of Web Usage Mining

IV. EXPRIMENT

To effectiveness and efficiency of our methodology mentioned above, with valid we have to use web server logs. November 12, 2005 Initial data source for our experiment to 25 November 2005, which size is around 36 MB. As shown in Table 1, after data cleaning, the number of requests declined from **92168** to **26584**. Figure 2 shows the detail changes in data cleaning.

In Figure 3, Bar chart 1 represents the initial requests in raw web log. From Bar chart 2 represents the row data after preprocessing method. From Bar chart 3 represent the image & status code Entry Remove. From Bar chart 4 represent the Robots Entry Remove.

TABLE I
The Processes and Results of Data Preprocessing in Web Usage Mining.

Total Entries in web log	92168
Entries after data cleaning	26584
Image And other Data Remove with status code	59981
Identify robots.txt and Remove	1527
Identify Robots User Agent and Remove	4670
Identify Spider/crawler IP and Remove	608
Identify robots.txt & Robot User Agent on same Record and Remove	879
Identify robots.txt & Spider/crawler IP on same Record and Remove	71
Identify Robot User Agent & Spider/crawler IP on same Record and Remove	283
All There Bots on Same Record and Remove	31

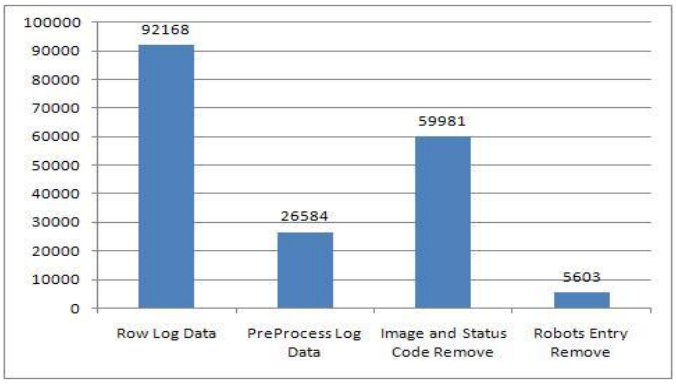

Fig.3. Processes of Data Cleaning

As Table 1 shows, the results obtained with the three methods employed for Web robot detection overlapped as shown in the Figure 4.

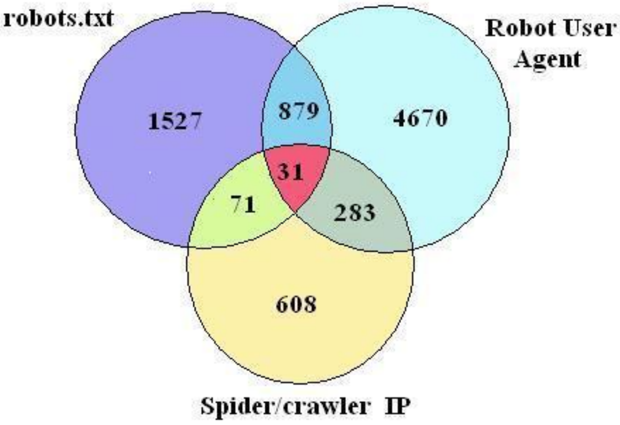

Fig.4. Number of Web Robot Hosts Identified Using Each Method

In General Preprocessing can take up to 60-80% of the times spend analyzing the data. Incomplete preprocessing task can easily result invalid pattern and wrong conclusions. Size of original log File before apply preprocessing is **37765942** Byte (**36.01** MB) and after apply the preprocessing is **4251968** Byte (**4.06** MB) so the Reduction in log File is **88.741263226** %

Finally, we have identified **3546** unique user on the basis of user identification's results, we have identified **4319** sessions by a threshold of **25.5** minutes and path completion.

V. CONCLUSION

We preprocessing in order to design and apply them easily at every stage of data to give some rules. Our experiments are important to us and our practices effectiveness data preprocessing estimates. This not only reduces log file size, but also increases the quality of data available. However, many problems remain such as data collection, applications of some heuristics in some phase of data preprocessing, the accuracy of user identification and session identification, applying the results of data preprocessing to patterns discovery and so on. We'll focus on solving these issues in the future.

REFERENCES

1. Ankit R Kharwar1, Chandni A Naik2, Niyanta K Desai3, 1Assistant Professor, Department of Computer, Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli , 2,3Student of M.Tech Computer Engineering in

2. Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli, “A Complete Pre-Processing Method for Web Usage Mining”, International Journal of

3. Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 10, October 2013)

4. Surbhi Anand, Surbhi Anand, Department of Computer Science & Engineering, Thapar University, Patiala-147004 (India), “An Efficient Algorithm for Data Cleaning of Log File using File Extensions”, International

5. Journal of Computer Applications (0975 – 888) Volume 48– No.8, June 2012

6. Gopal Pandey, Swati Patel, Vidhu Singhal, Akshay Kansara, “A Process Oriented Perception of Personalization Techniques in Web Mining”,

7. International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Volume-1, Issue-2, January 2013

8. V. Shanmuga Priya1, S. Sakthivel, Department of computer science, Periyar University, TamilNadu, India, “An Implementation Of Web Personalization Using Web Mining Techniques”, International Journal of

9. Computer Science and Mobile Computing, ISSN 2320–088X, IJCSMC, Vol. 2, Issue. 6, June 2013, pg.145 – 150

10. V. Sathiyamoorthi, Department of CSE, Sona College of Technology, Salemi-5, and Dr.Murali Bhaskaran, Principal,Paavai College of Engineering,

11. Paachal, 637018, Tamil Nadu, India, “Data Preprocessing Techniques for Pre-Fetching and Caching of Web Data through Proxy Server”, IJCSNS

12. International Journal of Computer Science and Network Security, VOL.11 No.11, November 2011

13. Ramya C, Dr. Shreedhara K S and Kavitha G, M.Tech (Final Year), Professor & Chairman and Lecturer, Dept. of Studies in CS&E, U.B.D.T College of Engineering, Davangere Davangere University, Karnataka, INDIA cramyac@gmail.com and ks_shreedhara@yahoo.com, “Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process”,

14. International Conference on Communication and Electronics Information (ICCEI 2011)

15. Abdul Rahaman Wahab Sait, and Dr.T.Meyappan, “Data Preprocessing and Transformation Technique to Generate Pattern from the Web Log”,International conference on Computer Science and Information Systems (ICSIS‟2014) Oct 17-18, 2014 Dubai (UAE)

17. Wasvand Chandrama, Prof. P.R.Devale, Prof. Ravindra Murumkar, Department of Information technology, Research scholar of Bharati Vidyapeeth University College of Engineering, Pune, Maharashtra 411046, India., ISSN 2348 – 7968, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, December 2014.

18. Sheetal A. Raiyani, Shailendra Jain, Dept. of CSE(SS),TIT,Bhopal],

19. “Enhance Preprocessing Technique Distinct User Identification using Web Log Usage data”, ISSN:2249-5789, International Journal of Computer Science & Communication Networks,Vol 2(4), 526-530

20. Michal Munk, Jozef Kapusta, Peter Švec, Constantine the Philosopher

21. University in Nitra, Department of Informatics, Tr. A.Hlinku 1, 949 74 Nitra, Slovakia, “Data Preprocessing Evaluation for Web Log Mining:Reconstruction of Activities of a Web Visitor”, International Conference on Computational Science, ICCS 2010

24. Mr. Shivkumar Khosla, Mrs. Varunakshi Bhojane, Department of Computer Engineering, Mumbai University, India, “Capturing Web Log and Performing Preprocessing of the User‟s Accessing Distance Education System”, International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.2, Issue.5, Sep.-Oct. 2012 pp-3128-3130 ISSN: 2249-6645

26. Bamshad Mobasher, “A Web Usage Mining”, http://maya.cs.depaul.edu/~mobasher/webminer/survey/node6.html. 1997.

27. Li Chaofeng , “Research and Development of Data Preprocessing in Web Usage Mining ,"

28. Rajni Pamnani, Pramila Chawan , " Web Usage Mining: A Research Area in Web Mining "

29. Andrew Shen , “Http User Agent List”, http://www.httpuseragent.org/list/

30. Andreas Staeding , “User-Agents (Spiders, Robots, Crawler, Browser)”, http://www.user-agents.org/

31. “Robots Ip Address”, http://chceme.info/ips/

32. “Volatile Graphix, Inc.”,http://www.iplists.com/nw/