A Survey: Investigation for Apache Log Preprocessing in Web Usage Mining

Arpit Shah[1], Nilesh Kakade[2],
[1]M.Tech Student, Department of Information Technology , PIET Limda, Waghodia ,Vadodara, Gujarat, India
[2] Ass.Professor, Department of Information Technology , PIET Limda, Waghodia ,Vadodara, Gujarat, India
Email - shah.a.s015@gmail.com nileshkumarjkd@gmail.com

Abstract: *Web use mining is a key procedure to concentrate information from web. Web use mining gives use design and client route for the site. Utilizing WUM association advantageously extricates visit get to designs and likewise enhances the website composition. In this Paper, The Principle Procedure is 3 Phases, Information Cleaning, User Recognizable proof and Session Distinguishing proof. We are execution of the information cleaning procedure of web log information utilizing Web utilization mining Procedure. Web log preprocessing we are enhancing for exactness and proficiency.*

Keywords: *Preprocessing Web logs, Web usage mining, Classification, Clustering, Associations technique*

1. INTRODUCTION:

Web Use Mining gives use designs the client. Web utilization mining procedure is the most ideal approach to obtaining client designs for research or investigation. The principle Variety amongst WUM and other two systems is, the web content digging gives substance to pertinent data and web structure mining gives information's structure and grouping of information while WUM gives information with correct route visit designs utilizing the web log record. Web use mining is a field of study where Weblogs are broke down and mined to suspect client's route conduct.

Basic data mining technic like association rule mining, classification, and clustering, sequential rule mining, etc. The web usage mining used to discover useful patterns from Web logs.

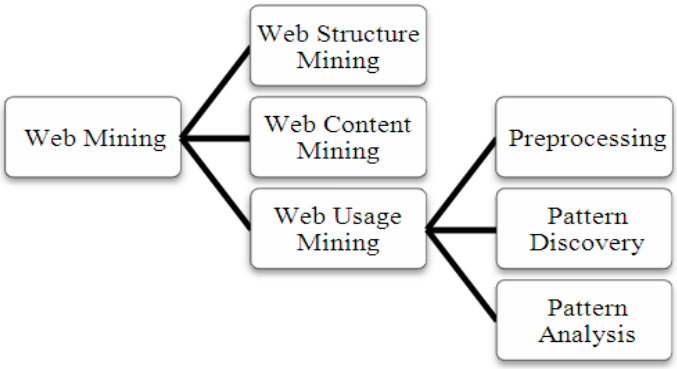

Fig. 1: Types of Web Mining [2]

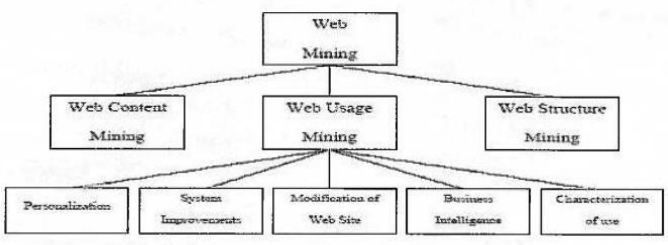

Fig. 2: Application of Web Usage Mining [3]

Web Content Mining:

Web content mining is the mining, extraction and joining of valuable information, data and learning from Site page content. The investigates web substance, for example, content, sight and sound information, and structure information. This is done to comprehend the substance of website pages; instructive catchphrase based ordering, site page positioning, site page content outline web inquiry and investigation.

Web Structure Mining:

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site.

Web Usage Mining:

Web utilization mining is the way toward separating valuable data shape server logs. It enhances seek proficiency and viability. Web seeker organizations routinely lead web utilization mining to enhance their nature of administration.

Web mining is the utilization of information mining methods to find examples, structures, and learning from the web. Web Information Mining Can be characterized as the Revelation and Examination of helpful data from the www-information [1]

- Collection of log data
- Preprocessing of server logs data, such as data cleaning and filtering, identification of user, identification of sessions, path completion, etc.
- Investigation of log data also referred as Web usage mining, to find out useful patterns.
- Evaluation of uncovering patterns.
- Keeping the evaluation of uncovering patterns

2. METHODOLOGIES:

Pre-processing of Web log is an important phase in web usage mining.it has been said it is complex process and it has been said it is a complex process and consumes 75% of the mining process.

A. Information Cleaning

Information will be cleaned in light of Url example. Information cleaning gives information according to client designs. This procedure will take out subpages and download pages from web log record, likewise erase the extra passages. The cleaning procedure is vital to pack the information and likewise give to lessen information repetition.

Algorithm 1: Information Cleaning
Input: l_IP1

Rows in the Web log file	Rows after Preprocessing	Total No. of users	Total No. of Session
273	46	28	45

Output: refine_log_T Begin

1. Read All Tuple in l_IP1
2. For each Tuple in l_IP1
3. Read fields (Status code)
4. If Status code=200, Then Get all fields Retrieve.
5. If suffix.URL_L = {*.gif,*.jpg,*.css,*.ico} then,
6. Remove suffix.URL_L
7. Save fields in new table.

End if

Else

8. Next record

End if

End

B. User Recognizable

In light of the client IP framework can without much of a stretch locate the most went by clients. Client distinguishing proof gives particular and one of kind clients. A log record contains also a huge number of passages, so the framework needs to check every last client's entrance. Since that framework needs to get distinctive clients and their IP address. Client's distinguishing proof is, to recognize who get to a site and which pages are gotten to.

Algorithm 2: User Recognizable

Input: refine_log_T

Output: identification of user Begin

1. Read records in l_IP1
2. for each record in dataset do
3. If current IP is not in L_IP1 then add the current IP in L_IP1 mark whole record as a new user and assign u_ID
4. else assign the old u_ID.

End else

End if

C. User Session Distinguishing

In view of time taken field effortlessly tally the quantity of sessions for a solitary client. Additionally, the framework can locate a specific session to get information from each page. A grouping of pages seen by a client amid one visit is known as the Session. The session is recorded in the log document. In pre-handling it is important to discover session of every client. It characterizes the quantity of times the client has gotten to a site page. It takes all the page reference of a given client in a log and partitions them into client sessions. These sessions can be utilized as an info information vector in order, grouping, forecast and different undertakings.

Algorithm 3: User Session Distinguishing

Input: user identified table

Output: identified sessions Begin

1. Read records in l_IP1
2. for each record in dataset do

3 if time_required > 30 Minitues assign new s_ID for that log entry

4. increments_ID
5. else assign the old s_ID.

End else

End if

End

3. EXPERIMENTAL RESULTS:

Table 1.Final Result

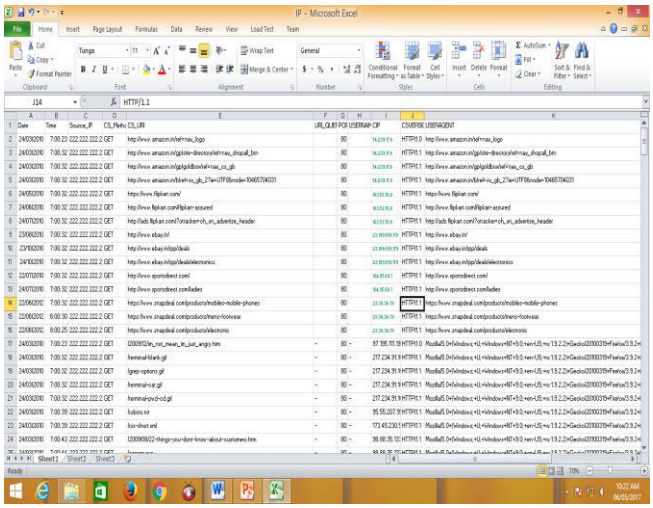

Fig. 3: Dataset

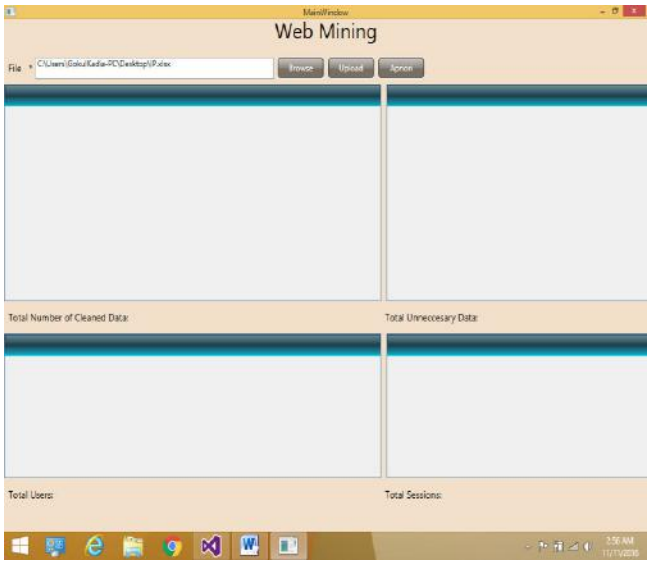

Fig. 4: Web Mining Screen

Fig. 5: Data Preprocessing

4. CONCLUSION:

Web Usage Mining is the one of the large areas of research and improves the sub domain of data mining and method. It is important to carry out

preprocessing stage is efficient. In data Preprocessing are the various stages like Information Cleaning, User Recognizable, Session Distinguishing and path Completion. There are many techniques like association technique; Clustering and classification technique must be applied in a web log.

So we are enhancing for exactness and proficiency are improving for Web log Preprocessing. In Feature various algorithms can be applied like Apriori algorithm, Naïve Bayes Classification Algorithm, K mean Algorithm, Decision Tree Algorithm can apply in a web log.

5. ACKNOWLEDGEMENT:

Second author of the paper offers the sincere gratitude to the faculty members in Department of Information Technology of Parul Institute of Technology for the constant encouragement and unparalleled support provided throughout the period of this research work.

6. REFERENCES:

1. Chhavi Rana, "A Study of Web Usage Mining Research Tools" 2012 Advance Networking and Applications
2. Mitali Srivastava, Rakhi Garg, P K Mishra, "Analysis of Data Extraction and Data Cleaning In Web Usage Mining" 2015 ICARCSET
3. Hengshan Wang, Cheng Yang, Hua Zeng, "Design and Implementation of a Web Usage Mining Model Based on Upgrowth and preflxspan" 2006 Communication of the IIMA
4. Greg Linden, Brent Smith, Jeremy York, "Amazon.com Recommendations: Item-to-item Collaborative Filtering," IEEE Internet Computing,Vol 7,no.1,pp.76-80,jan./feb.2003
5. Tobias Schnabel, Paul N.Bennett, and Thorsten joachims, "Using Shortlist to Support Decision Making and Improve Recommender System Performance" 2016 IW3C2
6. Hao Ma, Dengyong Zhou, Chao Liu, Michael R.Lyu, Irwin King, "Recommender Systems with Social Regularization, "2011 WSDM
7. G.Neelima, Dr.Sireesha Rodda, " Predicting user behavior through Sessions using the web log mining,"2016 IEEE International Conference on Advances in Human Machine Interaction"
8. Neha Goel, Dr.C.K.Jha, "Preprocessing web log: A critical Phase in web usage Mining, "2015 IEEE International Conference on Advance In computer Engineering and application
9. Wichian Premchaiswadi, Walisa Romasaiyud," Extracting Weblog of Siam University For Learning User Behavior On Map Reduce,"2011 IEEE
10. Yuqi Wang, Wenquiam Shang," Personalized News Recommendation Based on Consumer's Click Behavior,"2015 IEEE
11. Ying Han, Kejian Xia," Data Preprocessing Method Based On user characteristic of interest for web log mining,"2014 IEEE
12. Liu Kewen," Analysis of Preprocessing Methods for web Usage data,"2012 IEEE

i