DCUIS: An Exhaustive Algorithm for Pre-Processing of Web Log File

Sowmya H.K., Dr. R.J. Anandhi
*Department of Computer Science and Engineering, #Department of Information Science and Engineering
*The Oxford College of Engineering, Affiliated to VTU, Bangalore, India
#New Horizon College of Engineering, Affiliated to VTU, Bangalore, India
hk.sowmyakiran@gmail.com, rjanandhi@hotmail.com

Abstract— The recent growth of Internet and World Wide Web lead to exponential development of Internet usage for various purposes such as online shopping, social media, education etc. Every hit to the web site is recorded in log file which includes user request, IP address, date and time of page demanded etc. This information can be utilized to derive favorable perceptions. Web Usage Mining is one such approach, which is applied to log file to automatically discover user navigational pattern. This research work, presented a novel algorithm termed DCUIS: Data Cleaning, User Identification and Sessionization. It is an exhaustive algorithm which considers all stages of pre-processing phase. The proposed algorithm has taken raw log file as input, and performed cleaning operation to obtain data of superior quality. This data is used in the next step of algorithm to uniquely identify users and which in succession assists to find user sessions.

Keywords— Web Usage Mining, User Identification, Sessionization, Pre-processing, log file

I. INTRODUCTION

World-wide usage of internet for retrieving facts and details produces large volume of data daily. This overflowing data cannot be applied for analysis straightly. Thus Web Mining came into picture, which is the exercising of techniques available in data mining area to explore patterns from the data present on the internet. It is used to generate impressive outcomes on investigation which may possibly support the enterprise in captivating additional customers. Web mining can be classified into Web Content, Web Usage and Web Structure mining. Web Content mining implies, drawing out knowledge from web page which includes text, images, videos etc. Web Structure Mining finds knowledge from hyper link structure. Whereas Web Usage Mining discovers access pattern of user from web logs.

Web Usage mining is determining and analyzing visitors’ navigational patterns by realizing different data mining strategies on log files of server or proxy, client. These navigational patterns are well realized in diversified policies such as website restructuring, web page recommendation and web content personalization, and improvisation of server activities. The process of Web usage mining can be divided into three major phases: Data pre-processing, Pattern Discovery and Pattern analysis. Data pre-processing phase is a intricate job and spends longer time in web usage mining process due to the large volume of log data and its unstructured nature. It includes sub tasks such as data cleaning, user and session identification. This phase takes log file as input and identifies visitors of the web site and produces sessions which can then be used for pattern extraction and evaluation.

The entire paper is divided in to 4 sections. First section describes literature survey of existing approaches for pre-processing. Second section presents problem formulation of various stages of pre-processing. Third section explains various phases of Web Usage Mining and proposed novel exhaustive algorithm. Fourth section implements a new algorithm and compares it with existing algorithm which shows that our proposed algorithm produces better result.

II. LITERATURE SURVEY

P. Sukumar et.al [4] proposed De-Spidering heuristic algorithms was used to remove web robots. This algorithm used for data cleaning isolated the entries with file extension .css, .gif, .jpeg etc. It retained only the entries with status code in the range [200-299]. Additionally, the appeal initiated by web Crawlers, Robot or Spider is detached.

Sudheer Reddy et. al. [7], proposed preprocessing methods used to remove irrelevant entries with file extension .gif,.jpeg,.css and error codes. They gave algorithm for user identification, which determines new user by checking IP address. In case IP address is identical at that point, it compares with web browser and Operating System.

Patel et.al. [3], gave an algorithm for data cleaning which removed entries having extension .js, .css, .gif, .png, .jpg, .svg and error status codes. They also gave visitor identification algorithm which is developed on IP address and user agent to uniquely determine user. Session identification algorithm identifies user session based upon IP address and threshold time of 30 minutes.

Srivatsava et. al. [5], presented algorithm that considered unsuccessful inquiry, appeals of multimedia and other irrelevant files, and HTTP mechanism except GET comprising demands. This also removes failed status codes and links for pdf and image files. This algorithm eliminate records of URL ending like jpg, gif and css file.

Michal Munk et.al. [2], presented an approach to pre-process educational data and identified phases which are needed in case of pre-processing for increased usage of understanding analytics methods. Outcome of their experiment revealed that session identification algorithm with reference length had a notable influence on grade of extricated series conventions. Path completion technique had remarkable effect on quantity of extracted sequence rules.

Mary et.al. [8], proposed a method to enhance the performance of session identification to find accurate user navigational behavior. To identify a new session, it has considered set of pages that are shared between different sessions of same user. Suppose shared pattern does not exists, at that point ,such session of same user will be rejected. This improves quality of session.

Mitali Srivastava et.al. [1], developed a new algorithm for user identification based on MapReduce method. They identified user by IP address and user agent information. The proposed algorithm is free to address machine recognition and expandability affairs. Author have not considered details of referrer and site layout for identifying visitors.

III. PROBLEM FORMULATION

Each time user demands for a web page from a web site, an access is documented against a web server log file. Server log file exists in different formats like IIS standard/Extended, NCSA Common/Combined, and Netscape Flexible etc. NCSA extended common log format is most popular log format. Data cleaning, User identification and Session identification problem formation for ECLF format is in this manner.

Consider IP={ip_1,ip_2,….ip_n} is set of each respective IP addresses of n users, who browsed website. WR={$wr_{1,}wr_2$,….wr_n} denotes the web resources of website, UA={$ua_{1,}ua_2$,….ua_n} represents user agents of visitor of the web and EL={el_1,el_2,….el_n} is a set of external links. Now log record in ECLF possibly specified as LE=<ip_i;t;m;v;sc;bt;[Ref_i];[ua_i];[cookies]> ,where ip_iƐIP , wr_iƐ WR, Ref_iƐRUEL, ua_iƐUA, t denotes timestamp, m represents access method, sc represents status code and bt depicts amount of bytes moved. Ref_i and ua_i are attributes. Web server log file comprises WL= {wl_1,wl_2,….wl_n}. So data cleaning issue perhaps depicted in this manner: Taking into account a web server log file WL, remove all irrelevant entries and prepared log file can be represented as CL= {cl_1, cl_2,….cl_n},which contains only relevant entries and cl_i=<ip_i;wr_i;[Ref_i];[ua_i]>. Suppose U={u_1,u_2,….u_n} is a group of users who browsed website. A user's visit can be defined as v_i=<u_i,E_i> , where E_i=<(t_1,wr_1,[Ref_1]);(t_2,wr_2,[Ref_2]);…. (t_n,wr_n,[Ref_n])>; t_{i+1}>t_i. Now user identification problem can be defined as follows: From the cleaned log file CL, identify set of visitors V={v_1,v_2,….v_n}. Enter V inside user activity file. Further, session identification question possibly specified in this fashion: From the user activity file, identify set of sessions S={(v_1=<s_1,s_2,….s_k>);(v_2=<s_1,s_2,….s_k>);...(v_n=<s_1,s_2,….s_k>)}, where <s_1,s_2,….s_k> represents k different session of the visitor. Write this into user session activity file.

IV. PROPOSED METHODOLOGY

Web Usage Mining deals with approaches that may potentially speculate the user attitude when they are communicating with the WWW. Web usage mining uncover browsing patterns of visitor and attempt to discover the appropriate information from web log file.

The entire Web usage mining system can be splitted into three vital stages as shown in Figure 1. The pre-processing stage, the log file having click stream data is prepared to remove very noisy and ambiguous bunch of user activities during their visit to the site.

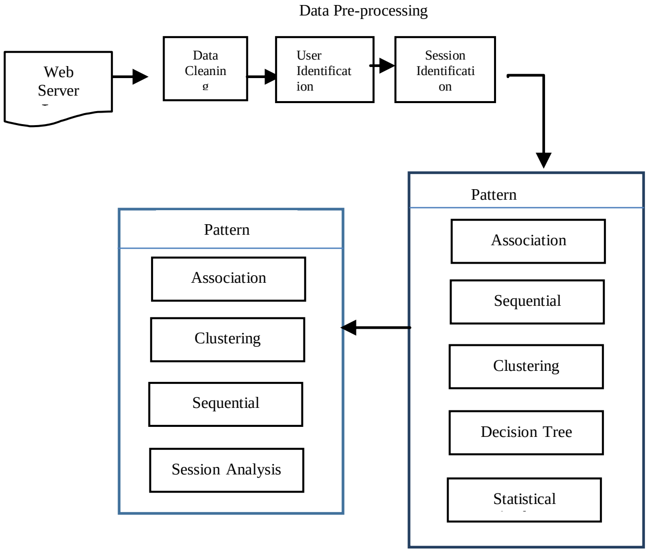
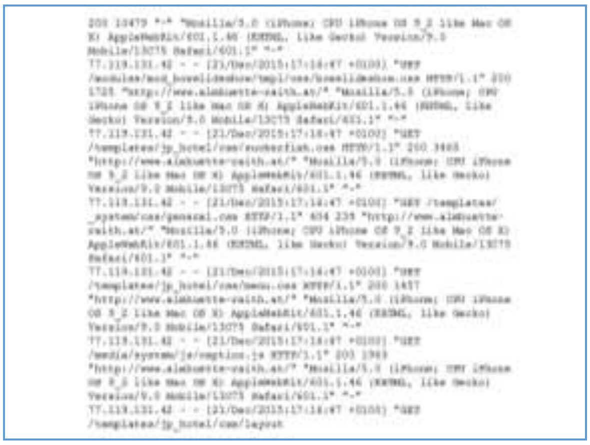


Fig. 1. Phases of Web Usage Mining Process

Data pre-processing is the basic step in data preparation step. Its objective is to reformat the web log file to identify users and user sessions. Web server makes an entry in the log file for each user hit to request resource from the website. Log data can be saved as Common Log Format (CLF) or Extended Log Format (ECLF). These files have fields such as User's IP address, User Identification, Authentication, Access date and time, Request method, Status code, Number of bytes transmitted , Referrer and User agent Field.



```
200 10479 "-" "Mozilla/5.0 (iPhone; CPU iPhone OS 9_2 like Mac OS
X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0
Mobile/13C75 Safari/601.1" "-"
77.119.131.42 - - [21/Dec/2015:17:16:47 +0100] "GET
/modules/mod_bowslideshow/tmpl/css/bowslideshow.css HTTP/1.1" 200
1725 "http://www.almhuette-raith.at/" "Mozilla/5.0 (iPhone; CPU
iPhone OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like
Gecko) Version/9.0 Mobile/13C75 Safari/601.1" "-"
77.119.131.42 - - [21/Dec/2015:17:16:47 +0100] "GET
/templates/jp_hotel/css/suckerfish.css HTTP/1.1" 200 3465
"http://www.almhuette-raith.at/" "Mozilla/5.0 (iPhone; CPU iPhone
OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko)
Version/9.0 Mobile/13C75 Safari/601.1" "-"
77.119.131.42 - - [21/Dec/2015:17:16:47 +0100] "GET /templates/
_system/css/general.css HTTP/1.1" 404 239 "http://www.almhuette-
raith.at/" "Mozilla/5.0 (iPhone; CPU iPhone OS 9_2 like Mac OS X)
AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13C75
Safari/601.1" "-"
77.119.131.42 - - [21/Dec/2015:17:16:47 +0100] "GET
/templates/jp_hotel/css/menu.css HTTP/1.1" 200 1457
"http://www.almhuette-raith.at/" "Mozilla/5.0 (iPhone; CPU iPhone
OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko)
Version/9.0 Mobile/13C75 Safari/601.1" "-"
77.119.131.42 - - [21/Dec/2015:17:16:47 +0100] "GET
/media/system/js/caption.js HTTP/1.1" 200 1963
"http://www.almhuette-raith.at/" "Mozilla/5.0 (iPhone; CPU iPhone
OS 9_2 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko)
Version/9.0 Mobile/13C75 Safari/601.1" "-"
77.119.131.42 - - [21/Dec/2015:17:16:47 +0100] "GET
/templates/jp_hotel/css/layout
```

Fig. 2. A Sample Web Server Log File

Data pre-processing feasibly accomplished in numerous levels like data fusion, data extraction, data cleaning, session identification and user identification. Pattern discovery can be done with the help of methods such as association rule mining, clustering and classification of data mining approach. Irrelevant rules and patterns are eliminated by executing pattern analysis method and outcome of this operation is treated as input to implementations namely visualization tools and generation tools.

From all the above phases, data pre-processing is major and critical step in web usage mining process.

A. Data Pre-processing

Data Pre-processing step remove irrelevant and noisy data from the web log file. It improves the quality of data and organizes only relevant and appropriate data, which will be used as input in various web mining algorithm. It also enhance the performance and expedience of Pattern discovery and Pattern analysis phases by providing pre-processed data.

B. Data Cleaning

Data cleaning mechanism clear out irrelevant and missing data from web log file. It also excludes log entry having failed status code and access to image files such as jpeg, GIF etc. It eliminates requests made by the crawlers, records with POST or HEAD method and blank lines.

C. User Identification

User identification is the most complex and demanding task in pre-processing phase due to local cache and proxy servers. User identification task, identifies users and group their activities and store them into user activity file. Proactive and reactive methods are two major procedures available at present to perform user identification operation. Proactive method identifies users from their earlier or present interaction with the site. Proactive approach integrates procedures namely user validation, stimulation of cookies on the customer side. Even though, these proactive methods are more accurate and reliable, due to privacy concern they cannot be used in all situations. Reactive method uses IP address and agent field to uniquely identify users.

D. Session Identification

Session Identification task recognizes the group of sessions by each user. Identifying sessions from log file is again an intricate task, since the server log files always contain limited set of information. The role of sessionization approach is to figure out more feasible and relevant information such as user's inclination and his intention. There are two important methods used for identifying sessions namely time based and referrer based. Most generally applied session duration is 30 min (maximum) and page view duration is 10 min (maximum) time. One of the limitation of this approach is that, here just the same session is partitioned into multiple session or either more than one session counted as a single session. In referrer based method, if referrer of the current request matches with the URL of the previous request then the current request is considered in the same session, else fresh session is created. But this method too has a constraint i.e. mostly the referrer field of log file contain null.

This paper addresses the issues of various stages of pre-processing phase. It proposes an enhanced algorithm for data cleaning that greatly reduces the size of web log file. It also proposes user identification and session identification algorithm.

E. Proposed Algorithm (DCUIS): Data Cleaning, User Identification and Sessionization Algorithm

The new proposed algorithm identifies which data should be considered as noise and has to be removed. It also determines in what manner users need to be detected and which specific time span desired to be applied for session identification. The DCUIS (Data Cleaning, User Identification and Sessionization) algorithm includes important features for various stages of data preprocessing such as data cleaning, user identification, and session identification. This algorithm takes web log file as input and produces user sessions as output. This user session file can be passed as input to pattern discovery and analysis algorithms for further analysis.

```
Proposed Algorithm: DCUIS
Input: Web log file
Output: User Session file
Step
    1. For each entry of web log file , remove the
       following records
       i.    Status code not in the range[200-
             299]
       ii.   Multimedia, audio and video files
       iii.  Web robot requests
       iv.   Methods other than GET
       v.    Script files
       vi.   Blank lines
    2. Initialize user count UID=0
```

```
Extract IP address, User agent and Referrer field of the entry
For each entry do the following
If (Two successive IP addresses are identical)
Then
        Check web browser and OS
        If ( Two successive entries IP address and User Agent are identical)
        Then
                Examine Referrer field
                If( Web page can be accessed with Referrer link)
                Same User
Else
 New User, UID++
For new User create next Session,
Inside User Session
If ( referrer ==NULL)
        If(Time of current request – time of previous request)<Page Stay Time Threshold
        Then
                Same Session
        Else if((Time of current request – time of first request)<Session Time Threshold
        Then
                Same Session
        Else
                New Session
Else
        If referrer page is same as previous request
        Then
                Same Session
        Else if(Time of current request – time of previous request)<Page Stay Time Threshold
        Then
                Same Session
        Else if (Time of current request – time of first request)<Session time Threshold
        Then
                Same Session
        Else
                New Session
```

Above DCUIS algorithm provides complete solution to pre-processing of web log file by performing complete steps of pre-processing starting from data cleaning to session identification. It commences from data cleaning, then performs user identification by verifying IP address and user agent. Identified user details are stored in a file. Soon afterward sessions are spotted by matching up the duration of current request and previous request with threshold time. Output of algorithm is a set of user sessions which are stored in session file.

V. RESULTS AND DISCUSSION

The exhaustive DCUIS algorithm has implemented in java Eclipse Platform and experiments run on a laptop machine equipped with Intel I3 processor and 8GB memory.

Input applied to algorithm is taken from the link http://www.almhuette-raith.at/apache-log/access.log which contains 9874 records. It holds browsing history of web user from 12th December 2015 to 21st December 2015. Algorithm is implemented in Java Eclipse platform using Java programming language. Results of data cleaning is stored in a separate file which is used as input for user identification process. The outcome of user identification is saved in user activity file. Further, this file is used as input for session identification, which produces set of user sessions, which are then saved in user session file.

When the user appeals for web page, if it succeeds then status code entered in web log file is in the range 200-299. Other kind of codes represent, unsuccessful request. Details of status codes present in raw log file is shown in the Table I.

TABLE I. SUMMARY OF STATUS CODES

Status Code	Number of Records in web log file	Percentage
200	9180	92.97
300	450	4.55
400	229	2.319

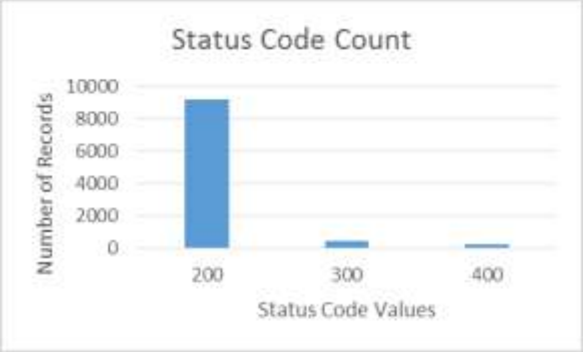


Fig. 3. Status Codes of Web Log File

The Raw web log file contains, enclosed image files, audio and multimedia files, which are not preferred and has to be cleaned. Our algorithm has cleaned the various types of irrelevant files as shown in the Table II.

TABLE II. SUMMARY OF FILE TYPES

File Type	Number of Records	Percentage
Script	330	3.34
Multimedia	79	7.73
Image	764	0.80

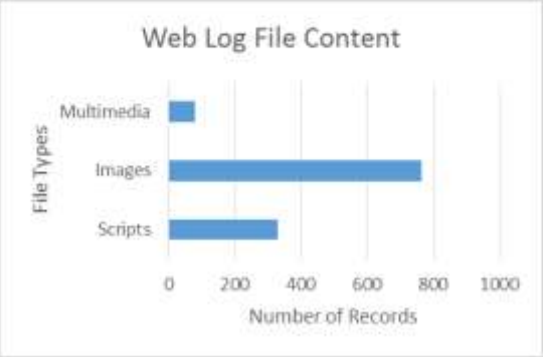


Fig. 4. Various File Types in Web Log File

Proposed data cleaning algorithm performs elimination of all unwanted records and retains, only valid records for further action. Table III shows result of the experiment after data cleaning operation.

TABLE III. RESULT OF DATA CLEANING PROCESS

Days	Size of Raw log file(KB)	Number of Raw Log Record	Number of valid records	% of valid Record
10 Days (12/12/2015 to 21/12/2015)	1870	9874	5104	52%

After accomplishing data cleaning operation, users are identified by our algorithm. Details about user identification is given in Table IV.

TABLE IV. RESULT AFTER IDENTIFYING UNIQUE USERS

Days	Number of records in a web log file	Number of User sessions
10 Days (12/12/2015 to 21/12/2015)	9874	1264

Sessions are identified for the unique users based on session threshold time and page stay time. Table V depicts the number of user sessions for various values of threshold time.

TABLE V. RESULT AFTER IDENTIFYING USER SESSIONS

Session Threshold Time (minutes)	Page Stay Threshold Time(minutes)	Original Records	Number of Sessions Identified
15	5	9874	3019
20	7	9874	3017
25	9	9874	3015

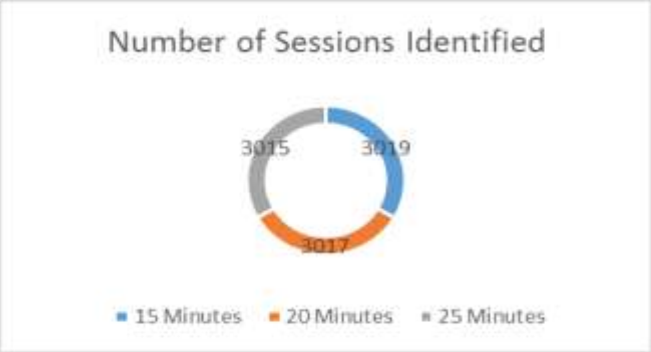


Fig. 5. Sessions Identified for Various Threshold Values

VI. CONCLUSION

Preprocessing of web log file is a major and complex task in web usage mining process which needs significant algorithms to perform data cleaning, identification of user and session. Several authors have proposed various algorithms for the different phases of preprocessing, but a complete algorithm for entire process is not available. So a new algorithm named DCUIS (Data cleaning, User Identification and Sessionization) is proposed for entire preprocessing phase. Hypothetical and experimental analysis have shown productiveness and relevance of empirical rooted preprocessing methods and algorithm policies. It shows that proposed novel algorithm is more efficient one which can be evolved as tool to administer exhaustive formula for preprocessing stage. The outcome of proposed algorithm can be used as input for subsequent phases of web usage mining.

Acknowledgment

I would like to thank to Dr. R J Anandhi, Professor and Head, Department of Information Science and Engineering, New Horizon College of Engineering, Bengaluru, my research supervisor, for her valuable guidance and encouragement towards this research work. I would like to express my special thanks of gratitude to Dr. A S Arvind, Principal, The Oxford College of Engineering, Affiliated to VTU, Bengaluru, for his advice and assistance that helped me during this research work. I would also like to thank Visvesvaraya Technological University for supporting me to do research work. Finally, a special thanks to my family for their support throughout my research work.

References

[1] Mitali Srivastava, Rakhi Garg,and P K Mishra,"*A Map Reduced-Based User Identification Algorithm for Web Usage Mining*", International Journal of Information Technology and Web Engineering.Volume 13, Issue 3, April – June 2018.

[2] Michal Munk, Martin Drlik,Benko and Reichel, "*Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniues*", IEEE Access 5, 8989-9004,2017.

[3] Patel,Parikh,” *Preprocessing on Web Server Log Data for Web Usage Pattern Discovery*”, International Journal of Computer Applications, Volume 165 , May 2017

[4] Sukumar, Robert, Yuvraj, “*Review on Modern Data Preprocessing Techniques in Web Usage Mining*”, International Conference on Computational System and Information Systems for Sustainable Solutions, 2016.

[5] Mitali Srivastava, Rakhi Garg,and P K Mishra, “*Analysis of Data Extraction and Data Cleaning in Web Usage Mining*”, ICARCSET ’15, March 06 - 07, 2015, Unnao, India.

[6] Anupama and Gowda, “*Clustering Of Web User Sessions to Maintain Occurrence of Sequence in Navigation Pattern*”, Second International Symposium on Computer Vision and the Internet,vol. 58, pp. 558–564,2015.

[7] K. Sudheer Reddy, G. Partha Saradhi Varma, and M. Kantha Reddy,”*An Effective Preprocessing Method for Web Usage Mining* “International Journal of Computer Theory and Engineering, Vol. 6, No. 5, October 2014.

[8] Mary,Baburaj,” *Performance Enhancement in Session Identification”*, International Conference on Control,Instrumentation ,Communicational and Computational Technology(ICCICCT), 2014.

[9] Ramya and Kavitha, “*An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network*” , Fifth International Conference on Information Processing, pp. 1–6,2011.

[10] http://www.almhuette-raith.at/apache-log/access.log

[11] CU, O., and P. Bhargavi. "*Analysis of Web Server log by web usage mining for extracting users patterns*." International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) 3.2 (2013): 123-136.

[12] DA, Adeniyi. "*Design and Realization of On-Line, Real Time Web Usage Data Mining and Recommendation System Using Bayesian Classification Method*." *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)6.3 (2016):19-38*.

[13] DHARMARAJAN, K., and K. ABIRAMI. "*DATA PREPROCESSING ALGORITHMIC APPROACH TO IDENTIFYING USER PATTERN BEHAVIOR FROM WEB SERVER LOG FILE."* International Journal of Mechanical and Production Engineering Research and Development (IJMPERD) 8, Special Issue 3 (2018):1434-1446

[14] SOLANKI, NIDHI, and AMAN DUREJA. "*COMPUTING FOR INNOVATION IN TECHNOLOGY*." International Journal of Industrial Engineering & Technology (IJIET) 3.2 (2013):87-94

[15] LATHA, N. PUSHPA, KV N. BHANU PRAKASH, and K. VENKATESWARA REDDY. "*HYBRID METHODOLOGY TO ANALYZE WEB USER BEHAVIOR IN WEB MINING AND FUZZY NETWORKS*." International Journal of Computer Science and Engineering (IJCSE) 3.3 (2014):157-166

[16] KUMAR, T. VIJAYA, and HS GURUPRASAD. "*Clustering of web usage data using fuzzy tolerance rough set similarity and table filling algorithm.*" *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)* 3.2 (2013): 143-152.