COMPARISON OF DECISION AND RANDOM TREE ALGORITHMS ON A WEB LOG DATA FOR FINDING FREQUENT PATTERNS

A.Jameela[1], P.Revathy[2]

[1]*PG Student, Computer Science and Engineering, Rajalakshmi Engineering College, Tamil Nadu, India*
[2]*Associate Professor, Computer Science and Engineering, Rajalakshmi Engineering College, Tamil Nadu, India*

Abstract

Web mining is the process of analyzing and extracting useful knowledge from web data. Web mining is divided into three categories namely web content mining, web structure mining and web usage mining. Web content mining is the process of extracting useful information from the web document. Web structure mining is used to identify the relationship between the web pages. Web usage mining is the process of extracting knowledge from the web log data. Web log data is classified into three, namely Client Log, Proxy Log and Web Server Log. In this paper NASA Web log data for the first 10 days are considered which consist of attributes like date, time, client and server IP address, user authentication, server port and method and URI. The web log data is preprocessed. Two more attributes user and session are added. The preprocessed data is classified using classification algorithms like decision and random tree. The useful information like frequently used web pages and no. of unique users are mined. The unique users and web pages in the forenoon and afternoon are classified. The performance evaluation between decision tree and random tree are drawn graphically.

Keywords: web mining, web usage mining, classification.

----------***----------

1. INTRODUCTION

Web mining is the process of extracting useful knowledge from web data. Web mining is classified into three namely web content mining, web structure mining and web usage mining. Web content mining is extracting useful information from the content of the web document. Web structure mining is the process of finding the relationships between hyperlinks of the web pages. Web usage mining is the process of finding frequent pattern or gaining knowledge from web log data [17].

Client log file, Proxy log file and Web server log files are three different data collection in web log data. Client log file collects all the activities by the client. Proxy log file collects all the activities between the client browsers and web servers. Web server log files records the visitor'sbehavior [1].

The general methodology of the web usage mining is data preprocessing, pattern discovery and pattern analysis. Data preprocessing is process of removing unnecessary and noisy data from the web log data. Pattern discovery is the process of finding all interesting patterns. Pattern analysis is to analysis the important knowledge from data.

In this paper the web log data of NASA is taken. Web usage mining techniques are applied along with the classification algorithms, decision and random tree [16]. The performance valuation between the algorithms is compared.

Section 2 explains about the related work. Section 3 describes the system architecture of the paper. In Section 4 and 5 we discuss on the methodology and experimental results. Section 6 is conclusion.

2. RELATED WORK

K. R. Suneethaand R. Krishnamoorthi[2]proposed a work on web usage mining which deeply explains about the data preprocessing and its analysis. NASA Web log data of first 7 days were taken to find information about the user, web site, errors etc., which helps system admin and web designer to improve their system by determiningsystem errors, corrupted and broken links. In data preprocessing number of IP address, entries, hits, unique user and errors were found. It does not concentrate on the interesting users.

G T Raju and P S Satyanarayana[3]proposed KDWUD (Knowledge Discovery from Web Usage Data). Web log data, NASA and Academic Sites were taken and then merged together. Data preprocessing steps was done on the data. User and session identification were carried out.

ThanakornPamutha[4] proposed data preprocessingon NASA data set. User and session identification were implemented in the data set. Total unique IP's, total unique pages and frequently used pages were found.

MaratheDagaduMitharam[5] proposed web usage mining preprocessing. The preprocessing steps consist of data cleaning, page view identification, user identification, sessionization and path completion. The following were

carried out in the data set and found that it is more efficient than other processes.

Rahul Mishra [6] proposed Apriori-Growth based on Apriori algorithm and FP-Growth algorithm. These two algorithms were used to find the frequent pattern in the NASA data set. The frequently used web pages and images files are found. Apriori-Growth was not efficient for large dataset which is the major issue.

K. R. Suneetha and R. Krishnamoorthi [7] proposed classification of NASA data set. The first phase is that data were preprocessed. Then decision tree algorithm is used for classification the data to find the group of interested users.

A. K. Santra [8] proposed model using enhanced version of decision tree algorithm C4.5 to classify the interested users. It showed good result in the improvement of time and memory utilization, it can be applied to any web log files.

3. SYSTEM ARCHITECTURE

Initially the web log data is downloaded from the internet[9]. The web log data is then preprocessed by data cleaning, user identification and session identification. Data cleaning is the process of removing irrelevant and incomplete data. User identification is finding the unique users from web log data. Session identification is the set of pages visited in the time period. Classification algorithm such as decision tree and random tree are applied on the preprocessed data. Using classification some patterns are derived in pattern discovery and pattern analysis is done to attain needed information from the log data. Fig-1 explains the system architecture.

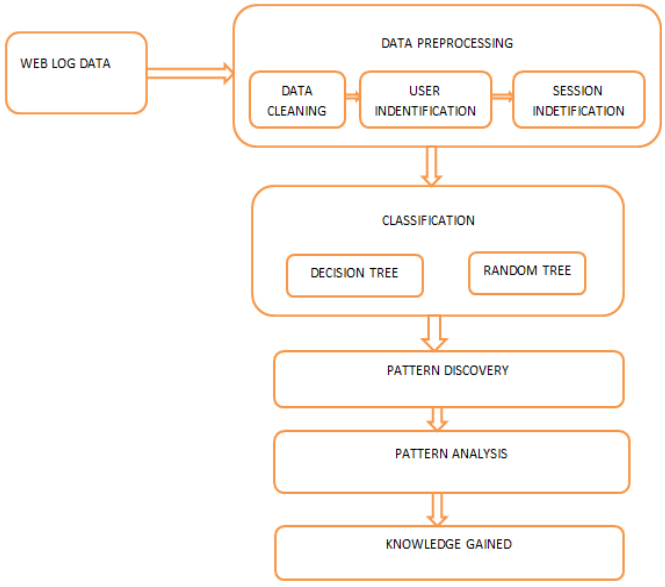

Fig-1: System Architecture

4. METHODOLOGY

4.1 Data Set

The data is collected from NASA web server log of AUG 1995. Total number of records is 1569898(1-31 AUG 1995). Algorithms are implemented on available log files on HTTP requests to the NASA Kennedy Space Center WWW server in Florida. In this project the logs from 00:00:00 Aug 1, 1995 through 23:59:59 Aug 10, 1995, a total of 10 days is considered. The records from 1 AUG 14:52:01 to 3 AUG 04:36:13, 1995 are not recorded due to web server shut down because of Hurricane Erin. The sample web log data is given below.

In24.inetnebr.com - - [01/Aug/1995:00:00:01 -0400] "GET /shuttle/missions/sts-68/news/sts-68 mcc-05.txt HTTP/1.0" 200 1839

Attributes in the data log data

- *Server IP Address:* Server IP is used for the access of information from the server.
- *User name and Password:* The two hyphen presents the user name and password.
- *Date:* The date of these entries is recorded in the format DD/MM/YYYY.
- *Time:* Time of transactions in the format HH:MM:SS
- *Server Port:* Port used in the data transmission.
- *Server Method (HTTP Request):* The word request refers to an image, movie, sound, pdf, .txt, HTML file and more.
- *URI:* Path from the host.
- *User Authentication:* Security features to enter user name and password. Authorized user can access the data.

4.2 Data Preprocessing

The first step in this paper is data preprocessing. Data preprocessing is the process of removing unwanted data from the web log data. This process is useful to extract the information we need from the web log data for further processes as it consist of many unwanted noisy data [10]. Data preprocessing comprises of three steps namely data cleaning, user identification and session identification.

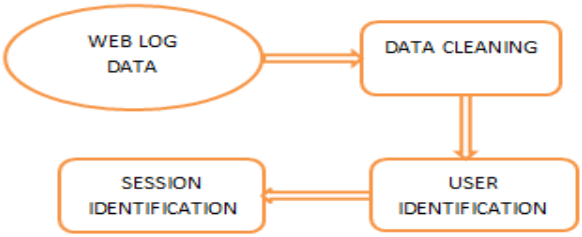

Fig-2: Data Preprocessing

4.3 Data Cleaning

The log entries for the first 10 days is converted into excel sheet columns. First step in data cleaning is to remove all the incomplete and noisy data from the logs. Then the attribute values like method, http request and server port are removed since all entries have the same values and nowhere used in the mining process. All the multimedia files like .gif, .jpg, .jpeg, etc., are also removed. The authentication status attribute values are scanned. The scanned values are 200, 302, 304, 403, 404, 500 and 501.

Table-1: Status Definition

200	Ok
302	Moved temporarily
304	Not modified
403	Forbidden
404	Not found
500	Server error
501	Not implemented

From Table-1 it is clear that only attribute value 200 is successful and remaining all are error codes. The bytes consumed by the error status are zero which is unwanted information in web data. So, all the error codes are removed from the data. More than 75% of the data are removed from the original data. Table-2 explains about the no. of entries before and after data cleaning.

Table-2: Entries before and after Data Cleaning

DATE	BEFORE DATA CLEANING	AFTER DATA CLEANING
1	33291	7681
3	40425	9190
4	58209	13313
5	31184	7114
6	31738	7619
7	56215	13227
8	59072	13688
9	59397	13864
10	60052	13658
Total	429583	99354
Percentage of Preprocessed Data		23.12%

Only 23.12% of original data remains after data preprocessing. The attributes after data cleaning are server IP, date, time, web pages, authentication status code and bytes consumed.

4.4 User Identification

User identification is the process of finding the unique user. Each IP is considered as a unique user. Whenever a new IP is found the user is incremented by one. The user with same IP address with different user agent is considered as different users. But in this data set no user agent is used. Table-3 gives the number of user in each day. Total Number of users for 10 days are 21021.

Table-3: No. of Users

DAY	NO OF USERS
1	2097
3	2551
4	3408
5	2062
6	2046
7	3382
8	3662
9	3609
10	3720

4.5 Session Identification

After user identification, the pages visited by each user are categorized into different session called as session identification. In this paper two sessions are identified namely forenoon and afternoon.From 00:00:00 to 11:59:59 is considered as forenoon and from 12:00:00 to 23:59:59 is afternoon. Table-4 explains the number of entries in the forenoon and afternoon.

Table-4: Entries in Forenoon and Afternoon

SESSION	NO OF ENTRIES
FORENOON	38346
AFTERNOON	61008

4.6 Classification

Classification is the automated process of assigning a class label and mapping a user based on the browsing history. The data are classified according to the predefined attributes. Decision tree, random tree, naïve Bayesian classifiers, support vector machine etc., are some of the classification algorithm. Based on the classification the user personalization and recommendations are done efficiently. In this paper we consider on two algorithms namely random tree and decision tree.

4.6.1 Decision Tree

Decision tree is the predictive machine-learning model that classifies the required information from the data. Each internal

node of a tree is considered as attributes and branches between the nodes are possible values. Decision tree is drawn on the data set using algorithm. The major disadvantage of decision tree is that repeated users are given different sub trees which reduce the accuracy of the classification. From the tree the frequently used users and web pages are classified.

4.6.2 Random Tree

Random tree is a classification algorithm which solves the regression problem in decision tree. It is the collection of tree predictors called forest. Each data is a tree in forest. In regression, the classifier response is the average of all the trees in the forest. Random tree algorithm is also applied on the data to gain the same information as decision tree.

4.7 Pattern Discovery

Pattern discovery isthe knowledge or information attained from the above steps. Here the classification result based on session identification is found. Number of user in forenoon and afternoon are analyzed. Number of unique user in each day is found. Top most pages visited are also analyzed. The users and web pages in different sessions are classified. All the information about the web log data are discovered here.

4.8 Pattern Analysis

Pattern analysis is the process of gaining information from pattern discovery. In pattern discovery all information about the web log data are found in which some may be unrelated to us. In pattern analysis the required information are only analyzed and removed from the complete pattern. The information found in the pattern analysis is listed out in the section 5.

5. EXPERIMENTAL RESULTS

NASA web log data of first 10 days are considered. In data preprocessing more than 76 % of the original data are removed. Then all procedures are applied through programs and WEKA tool. The following informationare gained from the data. Gained information is represented graphically. Fig-3gives the top most pages visited in the web data. Fig-4 gives the top text documents visited. Fig-5 lists the top five perl file viewed.

Fig-6 is the number of unique users in forenoon and afternoon for each day. Fig-7 is the number of unique users for each day in graphical representations. Fig-8 gives the graphically representation of the total number of users in the forenoon and afternoon. It is clear that number of users in the afternoon is higher.

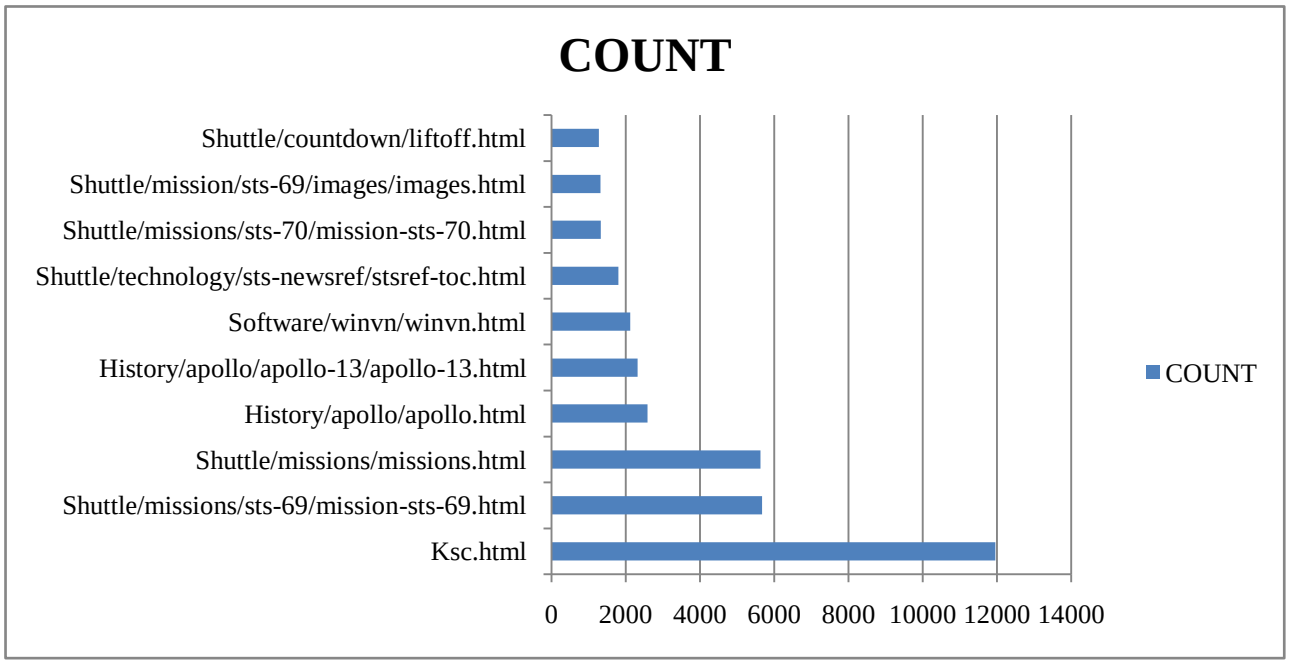

Fig-3: Top pages visited

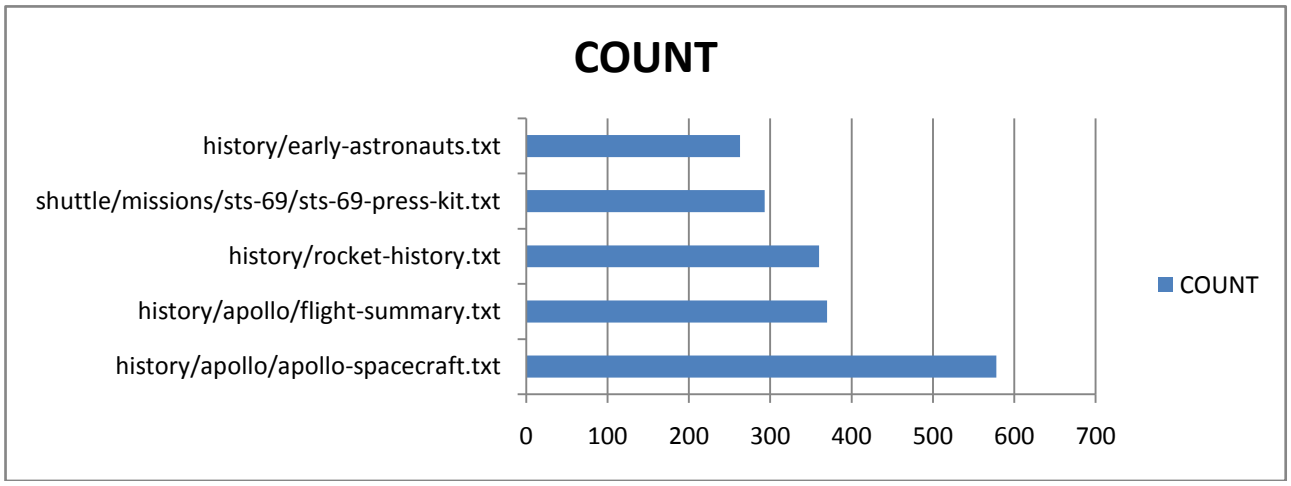

Fig-4: Top Text Documents visited

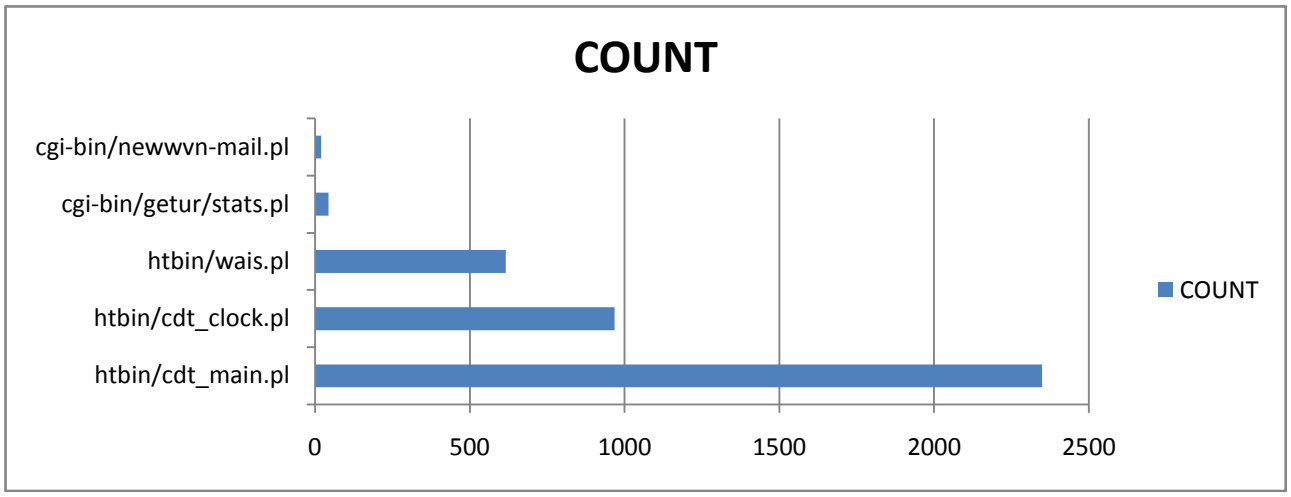

Fig-5: Top five perl file viewed

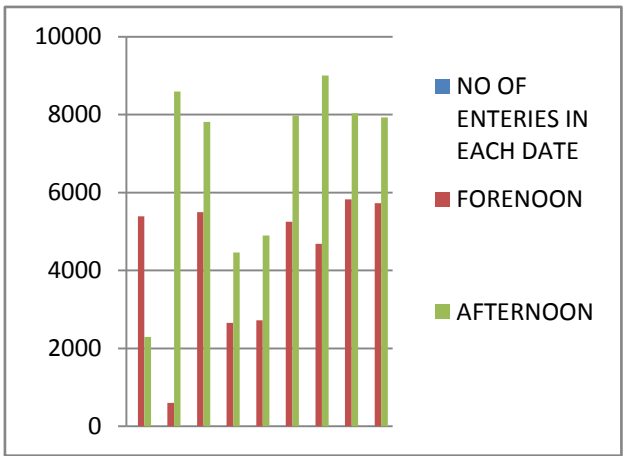

Fig -6: Number of Users in Forenoon and Afternoon

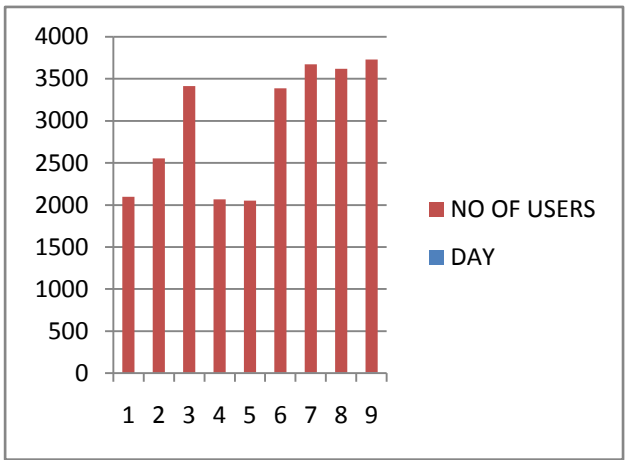

Fig -7: Number of Users per day

The performance evaluation between the decision tree and random tree are found by the number of users and web pages classified in different sessions by algorithms and WEKA tool. From Table-5 random tree results and tool results coincides. This means that random tree is more accurate than decision tree.

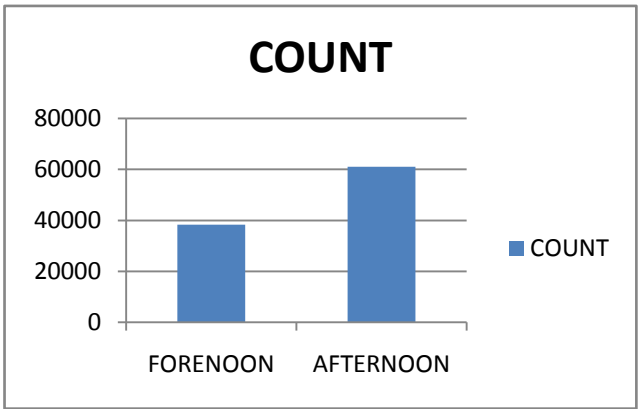

Fig -8: Users Count in each session

Table-5: Number of Users and Web Pages in Different Session

ATTIBUTES	NO OF USERS			NO OF WEB PAGES		
	DECISION TREE	RANDOM TREE	TOOL	DECISION TREE	RANDOM TREE	TOOL
FORENOON	38346	8883	8883	38346	1090	1090
AFTERNOON	61008	14074	14074	61008	1090	1090

6. CONCLUSIONS

The work focused on finding the user pattern from web log data. It uses the web designer to develop the web pages. In data preprocessing all the irrelevant and incomplete data are removed. Classification algorithm is applied and experimental results are given. The frequently used web pages in forenoon and afternoon are found. The number of users is found. Performance evaluation between the algorithmsis calculated. The result shows that random tree is more efficient than decision tree.

ACKNOWLEDGEMENTS

This work was supported by all my colleagues and faculties of Rajalakshmi Engineering College.

REFERENCES

[1] Ida Mele(February 4-8 2013), "Web Usage Mining for Enhancing Search-Result Delivery and Helping Users Sandra Stendahl and Andreas Andersson, "Web Mining" 2002.

[2] K.R.Suneetha and Dr. R. Krishnamoorthi, "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.

[3] G.T.Raju and P.S.Satyanarayana, "Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology," IJCSNS International Journal of Computer Science and Network Security, Vol. 8, No. 1, January 2008.

[4] ThanakornPamutha, "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns," International Journal of Research and Reviews in Wireless Communications, Vol. 2, No. 2, June 2012.

[5] MaratheDagaduMitharam, "Preprocessing in Web Usage Mining," International Journal of Scientific and Engineering Research, Vol. 3, Issue 2, February 2012.

[6] Rahul Mishra, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 9, September 2012.

[7] K.R.Suneetha and R. Krishnamoorthi, "Classification of Web Log Data to Identify Interested Users Using Decision Trees."

[8] A. K. Santra, "Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification" International Journal of Computer Science Issues, Vol. 9, Issue 1, No. 2, January 2012.

[9] http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html

[10] Sanjay BapuThakare, "A Effective and Complete Preprocessing for Web Usage Mining," International Journal on Computer Science and Engineering, Vol. 2, No. 3, 2010.

[11] to Find Interesting Web Content" WSDM'13, Rome, Italy.

[12] MehrdadJalali,Norwati Mustapha, Md. Nasir B Sulaiman and Ali Mamat "A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems" 12th International Conference Information Visualisation.

[13] Mehdi Heydari, Raed Ali Helal, Khairil Imran Ghauth(August 2009) "A Graph-Based Web Usage Mining Method Considering Client Side Data" 2009 International Conference on Electrical Engineering and Informatics 5-7, Selangor, Malaysia.

[14] Wang Yan, Le Jiajin and Le Jiajin(2010) "A Method for Privacy Preserving Mining of Association Rules

basedonWeb Usage Mining" International Conference on Web Information Systems and Mining.
[15] Sanjay Kumar Malik(2011) "Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation" ,International Conference on Computational Intelligence and Communication Systems.
[16] en.wikipedia.org.
[17] www.google.co.in.

BIOGRAPHIES

Ms. A.Jameela obtained her B.E. computer science degree from University College of Engineering Tindivanam, Tindivanam, Tamilnadu in 2012. She is currently pursuing her M.E. degree in compter science and engineering in Rajalakshmi Engineering College, Thandalam,Tamilnadu. She is doing her project in Web Usage Mining for Finding Knowledge from Web Log Data.

Mrs. P.Revathy completed her B.E from St.Peter's Engineering College, Chennai and M.E from Sathyabama Deemed University, Chennai. She has 13 years of experience in teaching as well as industry. She is presently working as Assosiate Professor in Rajalakshmi engineering College and pursuing her Ph.D (Part Time) in Computer Science and Engineering at Rajalakshmi Engineering College under faculty of Information and Communication Engineering, affiliated to Anna University, Chennai. Her areas of interest include Data Mining, Database Management Systems, Data Structures. She has published many papers in Refereed International Conferences and books in Data Structure.