IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

An Efficient Algorithm for Data Pre-Processing and Personalization in Web Usage Mining

Preeti Rathi[1*], Nipur Singh[2]

[1,2]Dept. of Computer Science, Kanya Gurukul Campus, Dehradun, India

Corresponding Author: mcapreeti.rathi@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i5.160164 | Available online at: www.ijcseonline.org

Accepted: 11/May/2019, Published: 31/May/2019

Abstract- With the huge amount of data in web, web mining is the process to extract useful data from web. Web usage mining is the type of web mining to retrieve data from web in form of logs, and it is also called web log mining. Web log mining extract useful pattern or information from log files and it help to determine user behaviour.
In this paper we proposed an algorithm for data pre-processing and personalization in web usage mining. Firstly collect the data from server and merge these log files into single log file. After collection of data separate each field using field separate algorithm, then cleaning the data to remove noise and unwanted data and after personalize these data for further used.

Keyword - Data Cleaning, Data field Extraction, Cluster, Session Identification, User Identification, Pre-processing, personalization.

I. INTRODUCTION

The process involved in the extraction of information pertaining from structured to unstructured or semi-structured in the form of web source is referred to as the web data mining. The information extracted from the web is also called as the web mining**.[9]** With the help of web data mining one can connect to a website's web pages and request information or pages, exactly as one's browser would do. In turn, the task of the web server is to send the html web page whose sole purpose is to extract the particular information from that web page. **[6]**

The growth of web is tremendous as approximately one million pages are added daily. Web Applications are increasing at giant speed and its users are increasing at exponential speed. Users' accesses are recorded in the web log files**. [1]**In today's period it has become important to know the user access mode. Because of the terrific usage of the web, the web log files are growing at a faster rate and the size is becoming huge. So to have a relevant data being resulted or analysed we can take help of the concept which is known as Web Mining**.[7]** Web mining involves exploration of web server logs of a website whereas data mining involves techniques to find relationships in huge amount of data (server logs) **[5].**

There are three category of web mining-

1.1 WEB CONTENT MINING

Extracting the knowledge behaviour from the contents present in the documents.

1.2 WEB STRUCTURE MINING

Obtaining the knowledge from Internet links.

1.3 WEB USAGE MINING

Obtaining the patterns which are of interest from web log access.

The technique of web usage mining involves the mining of data that extract usage patterns and identify the behaviour from Web log data. As a whole, web usage mining is divided into pre-processing, discovery of pattern, and analysing the same for pattern identification. **[12]** This process is used to find interesting pattern from log files and it helps to access information from log files. **[2]** The task of pre-processing involves the processing of site files that are untreated and convert the profile of the user data into page classification, site location and server session files **[3]**.The pattern discovery treats a server session file into session rules, patterns, and statistical information. The analysis of pattern identifies the rules, patterns, and statistical information obtained from the pattern discovery process**. [4]** In data mining process, learning can be categorized as supervised and unsupervised learning technique. In supervised learning a trainer is available, mean to say the training data includes the attributes and their outcomes. On the other way in unsupervised classification the data contains only attributes there are not any class labels exist**. [11]** In clustering data can be divided into different cluster and design of cluster according to user preference. **[10]**

II. RELATED WORK

There are many research paper discuss the pre-processing techniques. Pre-processing of data is the first step of personalization of data. Personalization means retrieve the data from log files according to user preference. Several techniques of web usage mining to personalization web log i.e. data cleaning, session identification, user identification.
Author, Jiang Chang-bin **[14]** suggested web log pre-processing algorithm based on collaborative filtering. The main purpose of web log data pre-processing to retrieve the quality of data from data sources and improve efficiency of usage pattern mining. Collaborative filtering is based on hypothesis i.e. user preference and user rating. This algorithm is same as k-nearest neighbour classifier, calculates the similarity between target user and former user. It can perform user session identification fast and gives better result.

Author, K.S.R. Pawan Kumar **[15]** suggest that three types of web mining i.e. web content mining, web structure mining, web usage mining. Author also discuss the some techniques of web usage mining and also consider the phases of web usage mining- pre-processing, pattern analysis, and pattern discovery. Pre-processing of data consist of data cleaning, user identification, session identification. In pre-processing of data reduce the size of data and remove the unwanted data and improve the searching time and reduced the memory consumption, personalization of data according to user specification.

Author, Gajendra Singh **[16]** investigated the new algorithm for web log mining. In this paper author discuss the log mining techniques through basic rules and optimize the execution time and comparison of proposed algorithm with existing apriori algorithm and gives better result in terms of throughput and execution time. This algorithm support the accuracy of log record from accessing data from log files. Author also discuss the firstly apply pre-processing techniques, after reduce the size apply algorithm to find minimum throughput and low CPU utilization.

Author, Gajendra Singh Chandel **[17]** proposed an approach for web usage mining using Fuzzy C- Mean algorithm. Fuzzy C-Mean algorithm start with the initial value of C i.e. the value of each cluster or group which is belong to training data points it is called centre of the cluster. FCM redefines these values again & again hence gives the better performance of consideration. Cluster centre value with its membership value of the training points to distance of training centre. Membership value of training point is the best value of cluster centre according to consideration value. In this paper comparison of k-mean and FCM algorithm based on performance including execution time and accuracy. K- Mean algorithm disadvantage achieve by fuzzy c-mean algorithm.

Author, Doddegowda B J **[18]** proposed a new algorithm for web personalization through integration of web user profiles and behavioural patterns. In this paper author find the set of user profiles and set of behavioural patterns according to prediction pages. Along those prediction find the similarity between user profiles and behavioural patterns, most significant behavioural pattern and user profile is used for further find rank of each page, top number of pages with highest page rank for recommended web personalization, for result analysis used KDDCup 99 datasets. This algorithm gives best result to customized web pages for web user effectively.

III. PROPOSED WORK

The flow of whole process of data pre-processing and personalization shown this following figure-

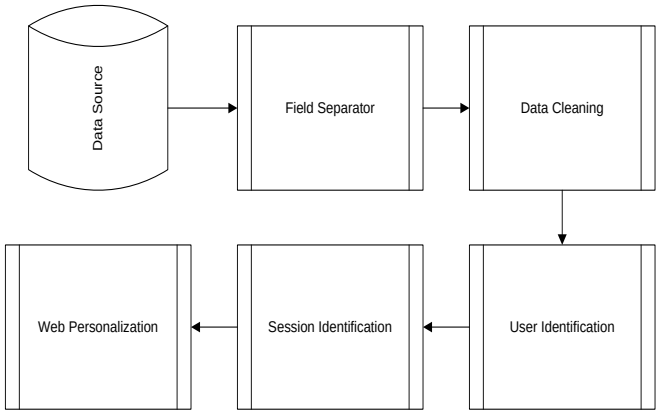

FIGURE-1 ARCHITECTURE OF PROPOSED WORK

IV. DATA COLLECTION

We have collected log data from website server. Web log data contains information about website visitors, IP-address, host name, Username, timestamp, method, path, protocol, status code and user information**.**

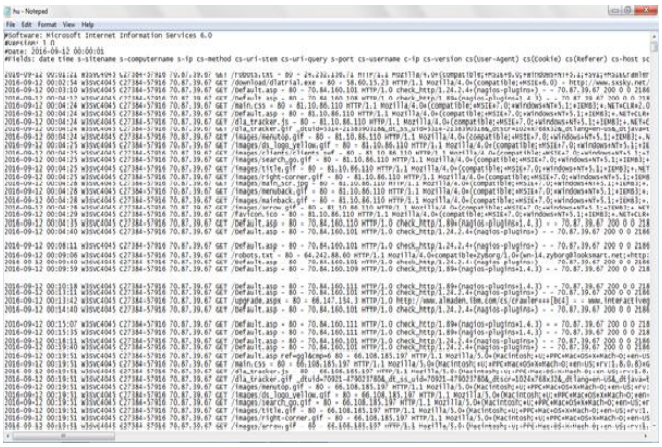

FIGURE -2 SAMPLE OF LOG FILE

The sample web log file has recorded information for each access as:

a) IP Address: Remote hostname or IPaddress number.
b) Remote User: The remote login name of the user. (If Remote user name is not present the – sign is normally used)
c) User: The username as which the user has authenticated himself. This is available when using password protected WWW pages. (If not exits - sign is normally used)
d) Timestamp: Date and time of the request.
e) Request: The request line exactly as it came from the client.
f) Method: The method is used to retrieve from request line.
g) Status code: The HTTP response code returned to the client. Indicates whether or not the successfully retrieved, and if not, what error message was returned. (i.e. 200,400)
h) Bytes: The number of bytes transferred.
i) Referrer: The URL, the client was on before requesting your URL. (If it not exits then – sign be placed for that place)
j) User Agent: The User Agent is whatever software the visitor used to access this site. (i.e. Mozilla)

V. FIELD SEPARATOR

Field separator is the part of data pre-processing, in this phase we separate the each field of log files. Through data field separator (i.e. delimiters), we separate each field using delimiters.

Field Separator Algorithm
Input: - Log Data File (LDF)
Output: - Extracted Log File (ELF)
Step 1: Read text data from log data file in read mode
Step 2: Open another file named Extracted Log File in write mode to write the extracted data.
Step 3: While {
Read log data from log file until end of file.
If next line {
Read one line and write in extracted log file.
}
}
Step 4: Calculate size of extracted log file and number of records.
Step 5: Close both log and extracted files.

VI. DATA CLEANING

Data pre-processing to remove the noise or irrelevant i.e. status code 200,300 or 400 and also delete unwanted extensions as well as methods from log files. It is an important phase of proposed methodology. The main objective of data cleaning is to improve the quality of data and increase the accuracy and reduce the time. In our proposed data cleaning algorithm of log files reduced the size of file 36.0MB to 15.4 MB. We have determined different type of status code, methods, and suffix in log files after cleaning we clean this record from log file and write into another log file. In pre-processing reduced the access time and reduced the memory consumption. Before cleaning access time is more in comparison to after cleaning. These are the following parameters we used in data cleaning algorithm-

TABLE-1 PARAMETER DESCRIPTION

S.No	Parameter	Value	Consumption in %
1	Status Code	400 or 200	0.04%
2	Method	GET or POST	-
3	File_Ext	jpeg,png,bmp,mp3,xml,js,cgi	0.56%

Data Cleaning Algorithm
Input: Extracted Log File (ELF)
Output: Summarized Log File (SLF)
Step 1: Read data from server log file i.e. extracted log file
Step 2: While (read until EOF) {
Read data.status_code
Read data.method
Read data.File_Ext
If (data.status_code= 400||200 && data.method= GET||POST && data.File_Ext! = css || xml|| js|| png||jpeg||gif)
{
Write data in Summarized log file
}
Else remove another records from extracted log file
}
Step 3: above two step repeat until end of file
Step 4: Calculate size of CLF and number of records.
Step 5: Close Summarized log file.

VII. USER IDENTIFICATION

User identification means identified the each user with using their IP Address. While read each entry from SLF (Summarized log file) if IP address is not exist then it consider as a new user, another case if IP exists but access from another browser as well as operating system then it is also consider as different existing. Following algorithm is need to identify the user.

User Identification Algorithm
Input- Summarized log file
Output- User Identified
Step1: Read two consecutive IP's from SLF
If (IP of first log entry = = IP of successive log entry)
Then user is an exists user
Step2: if user IP exists then browser and OS of that IP is not exists then it is consider as a new user
Step3: Repeat step 1 and 2 until each entry of SLF
Step4: else it consider as a new user.
Step5: Exit

VIII. SESSION IDENTIFICATION

After identification of user next phase is to identify session of each user. Time interval which is access or spend on web page is known as session. Using referrer URL we find the session either it is new or old, we set the interval during 30 minutes between accessing time. The algorithmic representation for the session identification is given below.

Session Identification Algorithm
Input- Identified User Table
Output- User session identification
Step1: Sort the IUT in ascending order by IP address
Step2: After sorting of table next step is calculated to browsing time of each user. Exit time of user is, entry time of another user.
Browsing Time = EXT-ENT
EXT= exit time of web page
ENT= entry time of web page
Step3: Compare in and out surfing time of each web page.
Step4: Repeat Step 2 & 3 till last entry of IUT.
Step5: if (browsing time < minimum time of WP)
{
Set value = Zero
}
Else if (browsing time > maximum time of WP)
{
Set value=one
}
Else if (browsing time is between minimum and maximum time)
{
Set value=two
}
Else
{
Set value= undefined
}
Step6: if (referrer URL is not found)
{
Set fix value = 100
}
Step7: If the same page is access by the user again in IUT, then increasing corresponding entry by 1.
Step8: Repeat step 5 & 6 & 7 until EOF
Step9: Values are stored in the matrix form i.e. row and column form.

IX. WEB PERSONALIZATION

In web personalization we proposed an algorithm for personalization. Firstly we collect data from log files and apply data pre- processing techniques to reduce the data size and remove noise and irrelevant data and generate the user identification and session identification of each user. After pre-processing we discover the patterns from reduced data set apply A* and Dijkstra algorithm to find path between websites. Then we personalize the data according to user requirement we apply apriori algorithm, and we proposed an improved personalization algorithm.

Personalization Algorithm
Input- Cleaned Dataset
Output- Personalized Data
Step-1 Suppose U_T user detail set with size T
F_T frequent detail set with size T
Step-2 Set U_1 is large frequent set
Step-3 for each detail set T=0 and user set is not null
Step-4 Increment T to T+1
Step-5 If detail set = User detail set then it store into personalized data set
Step-6 Calculate Support and Confidence with Min_Support
Step-7 Support is an indication of how frequently the detail set appears in the dataset

$$\text{Supp}(X) = \frac{|\{t \in T ; X \text{ is subset to } t\}|}{|T|}$$

Where t is a transaction and T is set of transaction and X is detail set
Step-8 Confidence is an indication of how often the rule has been found to be true.

$$\text{Con}(X \Rightarrow Y) = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)}$$

Step- 9 Return frequent used dataset for personalization.

X. EXPERIMENTAL EVALUATION AND RESULT

We collected the data from sever in log file format. After collection data firstly we extracted the each field from the data set and apply data cleaning algorithm for remove noise or irrelevant data from log files. Cleaning is the important phase of data pre-processing and take 75% of the total mining process. The method is implemented in Java. For field separation regular expression is used. In Algorithm 1 data field extraction from log files, data size is 37765942 bytes and number of records are 142568. Algorithm 2 reduced the data size from 36 MB to 15.4 MB. We analysis status code (200) for successful requests is 84.25%, status code (400) is 19.05%, above 500 request is 4.5%. Highest number of requests containing GET (99.84%) method.

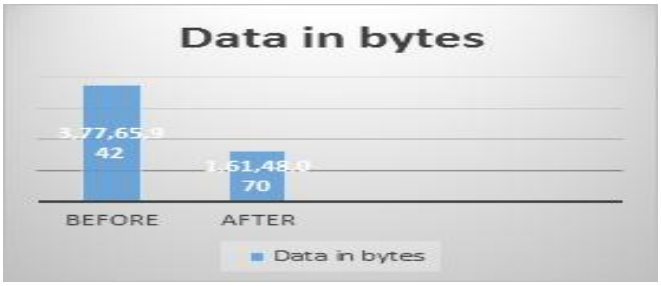

Figure 3 Result of Data Cleaning algorithm

In user identification we identify the user used user identification algorithm and check two consecutive IP's and identify the user.

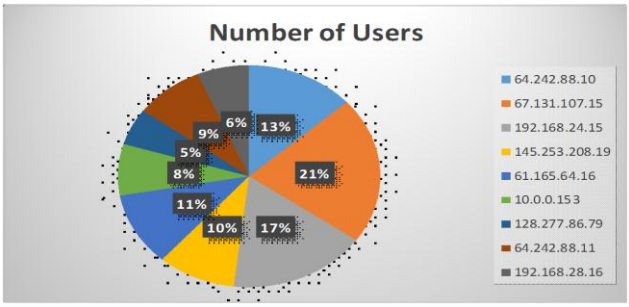

Figure- 4 Result of User Identification algorithm

In session identification we identify the session of each user and this is used for data personalization. Identification of user Result after identification of session of each user store in matrix form. We show this matrix in the form of chart.
Browsing time = Entry time – Exit time
Minimum time = 0 sec
Maximum time = 30 min
In one entry of matrix is given below-
Browsing time = 25 min
Then
BT is between minimum and maximum time then set the value is 2.
And manage the session of first user based on browsing time of web page. Accuracy of webpage is maximum and memory consumption is minimum.

XI. CONCLUSION

Due to huge amount of log data, it is necessary to perform pre-processing of data applying mining algorithm on web log data. The goal of data pre-processing is to prepare structured data. Our main focused on data field extraction and data cleaning algorithm. We proposed an algorithm for data cleaning and data field extraction, and extract useful patterns from log files. In this paper we also Proposed algorithm improve the performance of the webpages is based on accuracy and time and reduced the memory consumption and we personalized data for user requirement.

REFERENCES

[1]. Bhupendra Kumar Malviya, Jitendra Agrawal, "A Study on Web Usage Mining: Theory and Applications", Fifth International Conference on Communication Systems and Network Technologies, IEEE, Page: 935-939, April 2015, ISBN (Print) 978-1-4799-1797-6/15

[2]. Dr. Girish S. Katkar, Amit Dipchandji Kasliwal," Use of Log Data for Predictive Analytics through Data Mining", Current Trends in Technology and Science, page-217-222, ISSN: 2279-0535. Volume: 3, Issue: 3 (Apr-May. 2014).
International Journal of Computer Applications (0975 – 8887) Volume 103 – No.6, October 2014

[3]. M.Praveen Kumar," An Effective Analysis of Weblog Files to improve Website Performance", International Journal of Computer Science & Communication Networks, Vol. 2(1), Page: 55-60, 2011, ISSN: 2249-5789.

[4]. Mr. Jitendra B. Upadhyay, Dr. S. V. Patel," A Review Analysis of Preprocessing Techniques in Web usage Mining", International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 04, April-2015, page -1160-1166,ISSN: 2278-0181

[5]. Nehal G. Karelia, Prof. Shweta Shukla," Data Preprocessing: A Pre requisite for Web Log Files", International Journal of Engineering Research & Technology (IJERT), page-1571-1574, Vol. 3 Issue 4, April – 2014, ISSN: 2278-0181

[6]. Oren Etzioni," The World-Wide Web: Quagmire or Gold Mine?" ACM, Vol. 39, No. 11, November 1996, Page: 66-68.

[7]. Sameer Dixit, Navjot Gwal," An Implementation of Data Pre-Processing for Small Dataset",

[8]. Saurabh Choudhry, Prof A. K Solanki " Errors in Internet Log files for Website Improvement and Interaction", International Journal of Advanced Research in Computer Science and Software Engineering, Page-365-371, Volume 4, Issue 10, October 2014, ISSN- 2277 128X

[9]. Shakti Kundu, "An Intelligent approach of web data mining", International Journal on Computer Science and Engineering, page-919-928, Vol. 4 No. 05 May 2012, ISSN: 0975-3397.

[10]. Sheetal A. Raiyani, Rakesh Pandey, Shivkumar Singh Tomar, "Performance Enhancement of Web Server log for Distinct User Identification through different Factors", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 6, June 2014, Page: 7262-7267, ISSN (Online) : 2278-1021, ISSN (Print) : 2319-5940.

[11]. Shivaprasad G., N.V. Subba Reddy, U. Dinesh Acharya," Knowledge Discovery from Web Usage Data: An Efficient Implementation of Web Log Preprocessing Techniques", International Journal of Computer Applications (0975 – 8887) Volume 111 – No 13, February 2015

[12]. Surbhi Anand , Rinkle Rani Aggarwal "An Efficient Algorithm for Data Cleaning of Log File using File Extensions ", International Journal of Computer Applications (0975 – 888)Volume 48– No.8, June 2012

[13]. V.Chitraa, Dr.Antony Selvadoss Thanamani ," A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011

[14]. Jiang Chang-bin, "Web Log Data Preprocessing Based on Collaborative Filtering", 2010 Second International Workshop on Education Technology and Computer Science.

[15]. K. S. R. Pawan Kumar,"A Critique on Web Usage Mining", International Journal of Computer Science and Information Technologies, Vol. 3 (5) , 2012,5276-5279.

[16]. Gajendra Singh, "A New Algorithm for Web Log Mining", International Journal of Computer Applications (0975 – 8887) Volume 90 – No 17, March 2014 20

[17]. Gajendra Singh Chandel, "A Result Evolution Approach for Web usage mining using Fuzzy C-Mean Clustering Algorithm", IJCSNS International Journal of Computer Science and Network Security, VOL.16 No.1, January 2016

[18]. Doddegowda B J," A Novel Algorithm for Web Personalization through Integration of Web User Profiles and Behavioural Patterns", *IRACST* - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555, Vol.7, No.2, Mar-April 2017

