IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-1(1) E-ISSN: 2347-2693

Effectuation of Web Log Preprocessing and Page Access Frequency using Web Usage Mining

Brijesh Bakariya*1, Ghanshyam Singh Thakur2
*1Department of Computer Application, Maulana Azad National Institute of Technology, India
2 Department of Computer Application, Maulana Azad National Institute of Technology, India
www.ijcaonline.org

Received: 15/July/2013 Revised: 19/July/2013 Accepted: 04/August/2013 Published: 30/Sept/2013

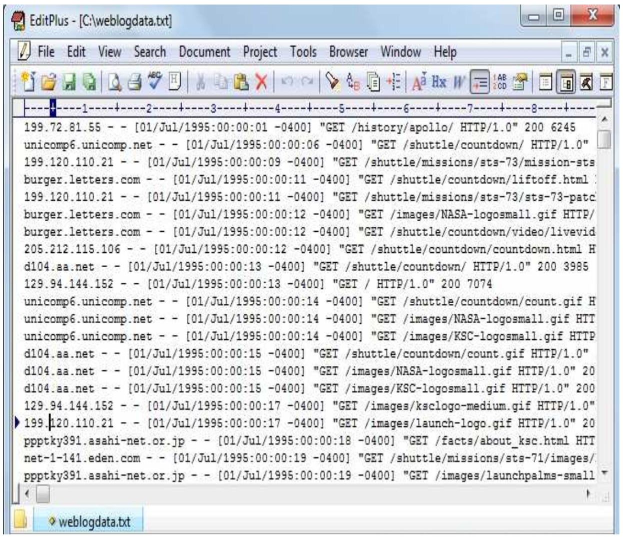
***Abstract*—** For accessing the information from web log, this is very important task and this task can be accomplished by web usage mining technique. Through web usage mining technique we can find out visitors behavior which can automatically and very fast access intrinsic information from huge amount of web log data, such as interesting access path, identify the user, accessing the web page group, web user clustering and web pre-fetching. Web usage mining is milestone for decision making process for an organization. Data preprocessing is very important concepts for the mining process. If our web log data is preprocessed then we can easily find out the desire information about visitor and also retrieve other hidden information from web log data. In this paper we focus on data preprocessing technique of web usage mining, after completion of data preprocessing, any king of irrelevant information can be sort out. We have also proposed an algorithm and its implementation for web log preprocessing in web usage mining. Every page has been assigned with an individual token. According to this token and frequency, data mining technique (Classification, Association Rules, and Clustering) can be applied. In this article we can easily find the highest and lowest value according to page access frequency.

***Keywords*—**Web Usage Mining, Preprocessing, Web Log Data, Frequency, Clustering.

I. INTRODUCTION

One of the names of web usage mining is web log mining; it is one of the techniques to find interesting information from web log data. All in the web log data is stored on the server. When server stored this kind of data which is semi-structured or unstructured; to find the relevant information it is very complex task. For understanding the visitor's behavior from different kind of website and containing different kind of log data using the web usage mining techniques [1]. Web log record contains various kinds of information like IP address, URL, Referrer, Time etc. In this figure we are showing a snippet of log data. Discovering and analyzing of web log could be helping us to find out lots of information and it is also necessary for good website design [1]. But the web log data contain lots of noisy data and we have to eliminate the irrelevant information, web log preprocessing is required because noisy data affect the desired outcome and our result would be wrong due to irrelevancy. Preprocessing is the first step required after that step we have to perform data mining techniques like association rules, classification and clustering for retrieving interesting information [4]. The data preprocessing includes data cleaning, user identification, session identification and path completion [17].

An efficient data preprocessing approach improves time required for data mining or pattern discovery. If our data preprocessing is efficient then it can be easily find out frequent pattern or interesting rule among web data with limited amount of time. In this paper we have analyze the web log data and differentiate its attributes. After analyzing the data we have to preprocess of that data and give an algorithm for data preprocessing. Using our preprocessing Algorithm, it is easier to find relation among web data within the limited amount of time.


EditPlus - [C:\weblogdata.txt]
File Edit View Search Document Project Tools Browser Window Help
weblogdata.txt

```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245
unicomp6.unicomp.net - - [01/Jul/1995:00:00:06 -0400] "GET /shuttle/countdown/ HTTP/1.0"
199.120.110.21 - - [01/Jul/1995:00:00:09 -0400] "GET /shuttle/missions/sts-73/mission-sts
burger.letters.com - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/countdown/liftoff.html
199.120.110.21 - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/missions/sts-73/sts-73-patc
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /images/NASA-logosmall.gif HTTP/
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/video/livevid
205.212.115.106 - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/countdown.html H
d104.aa.net - - [01/Jul/1995:00:00:13 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
129.94.144.152 - - [01/Jul/1995:00:00:13 -0400] "GET / HTTP/1.0" 200 7074
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /shuttle/countdown/count.gif H
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/NASA-logosmall.gif HTT
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/KSC-logosmall.gif HTTP
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0"
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 20
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200
129.94.144.152 - - [01/Jul/1995:00:00:17 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0"
199.120.110.21 - - [01/Jul/1995:00:00:17 -0400] "GET /images/launch-logo.gif HTTP/1.0" 20
ppptky391.asahi-net.or.jp - - [01/Jul/1995:00:00:18 -0400] "GET /facts/about_ksc.html HTT
net-1-141.eden.com - - [01/Jul/1995:00:00:19 -0400] "GET /shuttle/missions/sts-71/images/
ppptky391.asahi-net.or.jp - - [01/Jul/1995:00:00:19 -0400] "GET /images/launchpalms-small
```

Fig. 1 Sample of web log data

II. WEB USAGE MINING

Cooley et. al. [8] defined the term web usage mining and they also given the definition of web usage mining. Web usage mining is the automatic discovery of user access patterns web servers. An overall process of web usage mining is given the figure 1.

Corresponding Author: *Brijesh Bakariya*

A. *Data Preprocessing*

Lots of problems in data preprocessing like collection of data, integration of data and transaction identification etc [2][3]. It is a procedure to convert the various kinds of information like content information (text, image, audio, video, structure records), structure information (hyperlink, inter and intra document), usage information (web and application server logs) and also use for pattern discovery process.

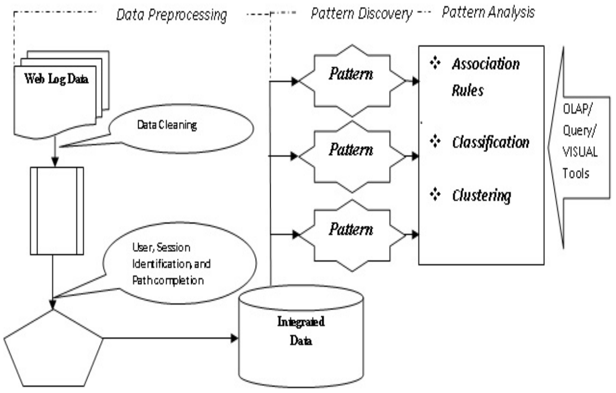


Fig. 2 Web Usage Mining architecture

Data preprocessing include data cleaning, user identification and session identification and path completion .This kinds of method convert our web data into this preprocessing covert the data into a unique format [15].

B. *Pattern Discovery*

It is a method which is use to find out pattern from web data but this method is only applicable when the data are preprocessed .In this process logs were taken from an NASA-HTTP website and log file is available at The Internet Traffic Archive sponsored by ACM SIGCOMM for the time period 1 July to 31 July1995[9]. We discover the statistical measure and pattern discover through sample data. This process may failed when our sample are in not appropriate format, it means that sample does not represent the complete data[5]. There are various techniques for pattern discovery process like clustering, statistical analysis, association rule, classification sequential pattern and dependency modeling.

C. *Pattern Analysis*

In this technique we have to extract frequent web data pattern from the all data pattern which is huge amount of web data. For this we have various kinds of methods like Apriori, FP-tree, and WAP-tree etc [10]. By using this method we can discover statistical measurement (mean, median, mode etc), interesting rules and most frequent patterns [16].

III. Data Preprocessing

The Quality of Data is a key part to understand when we are going to mine from it. Most of mining attempt use to progress quality of data. To analyze user behavior, web log data play a very important role for it. But web log data contain some unnecessary information like image access, failed entries, server load etc [6][14]., which will affect the precision of pattern discovery and pattern analysis. Because of this data Preprocessing is very important task in mining to find efficient patterns and getting efficient result.

A. *Data Collection*

In this article, the data source which is in IIS file format, for the finding hidden information of visitor is collected by NASA-HTTP. The log file is available at The Internet Traffic Archive sponsored by ACM SIGCOMM [9]. We use the part of the logs during the period of 1 July to 31 July1995. For session identification, set the maximum elapsed time to 30 min, which is used in many commercial applications [16]. There are following data structures of web logs are shown in figure3.

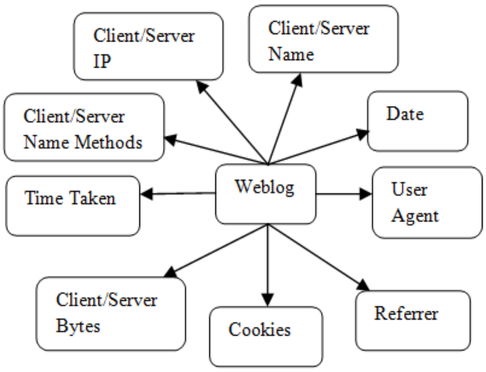


Fig. 3 Major attribute contain web log data

B. *Data Cleaning*

This is the first step to performed in the in the preprocessing of web log data. It is the procedure to eliminate irrelevant information from web log data. Lots of users can be accessed a website, but when HTTP status code failed with record it also maintain in web log data [12][13]. Data cleaning is generally site specific procedure; lots of extra information may not be significant for analyzing the web log data like style of file, sound (audio, video) files [7]. Therefore some useless entries may occur and it can be remove in data cleaning process. Through data cleaning technique, inconsistency, irrelevancy, noise can be find out to improve the quality of data. There is following algorithm to clean the noisy data from server log is illustrated below.

C. *An algorithm to clean noise in a data*

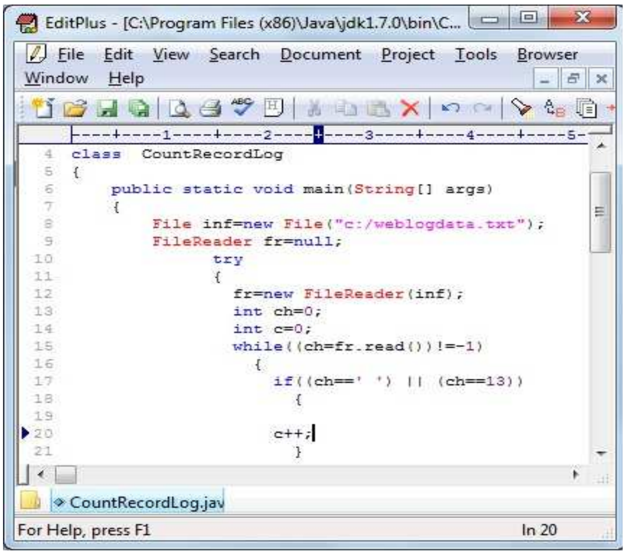
```
1) Start
2) Scan the Log Record in LF
3) For every record in LF
4) Read all the fields (Referrer, Methods, Status etc)
5) If status code = Success
6) then
```

```
7)  Take IP Address and URL
8)  If(Suffix of URL=*.txt, *.mpg,*.gif , *.css, *.jpg)
9)  then
10) Remove suffix from URL
11) Otherwise
12) Save records
13) End if
14) Fetch the next record
15) End if
16) Stop
```

This algorithm not only cleans the irrelevant data but can also eliminate the inconsistency and incomplete data. Error request are not in use of mining technique. These kinds of request can be eliminated after checking the request code, suppose the request code like 404 then this code is error generated code [14]. When this kind of code can occur while reading the file it can be remove according to error request code. All the log entries having this kind of code can be eliminated because it is useless. We have analyzed the log data and show the given table I below This table gives the information about data before cleaning process.

IV. Implementation of Data Preprocessing Algorithm

We have implemented the above algorithm in java programming language. First of all to clean the log data, read the web log file and count all the record. The logic behind that the procedure is that, we have read character by character from a file compare the character from ASCII value of space and enter key and count all the record from web log file. We are showing the given figure4.

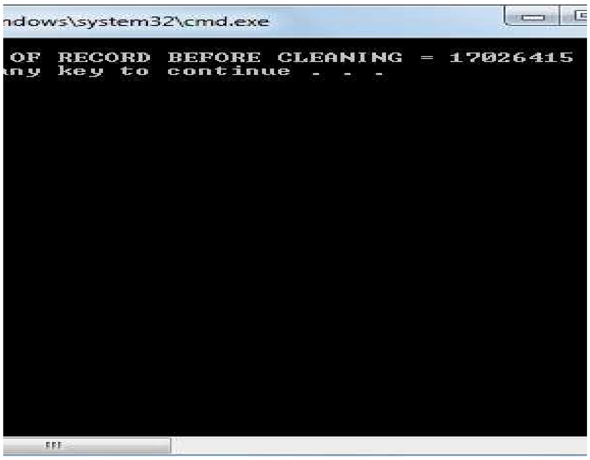

EditPlus - [C:\Program Files (x86)\Java\jdk1.7.0\bin\C...
File Edit View Search Document Project Tools Browser
Window Help
CountRecordLog.jav
For Help, press F1
In 20

```
class  CountRecordLog
{
    public static void main(String[] args)
    {
        File inf=new File("c:/weblogdata.txt");
        FileReader fr=null;
              try
              {
                fr=new FileReader(inf);
                int ch=0;
                int c=0;
                while((ch=fr.read())!=-1)
                  {
                    if((ch==' ') || (ch==13))
                      {

                    c++;
                      }
```

Fig. 4 A code for counting number of records from web file

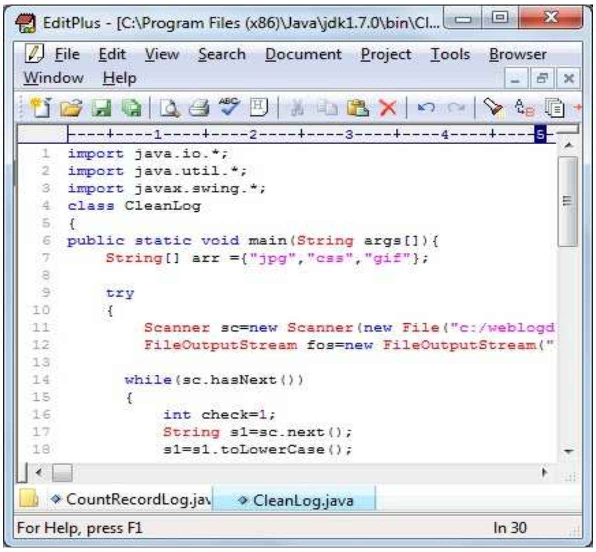
The output return all number of records from a file, we are showing the given figure 5.



```
OF RECORD BEFORE CLEANING = 17026415
ny key to continue . . .
```

Fig. 5 Result for showing number of records in a log file

After counting all number of records, we will have to preprocess (clean) our data. In this procedure we have to remove the entire suffix (*.jpg,*.css,*.gif etc.) which is not necessary in a file. The file size is also reduced after cleaning the data. The code and result show the below figure6.


EditPlus - [C:\Program Files (x86)\Java\jdk1.7.0\bin\Cl...
File Edit View Search Document Project Tools Browser
Window Help
CountRecordLog.jav
CleanLog.java
For Help, press F1
In 30

```
import java.io.*;
import java.util.*;
import javax.swing.*;
class CleanLog
{
public static void main(String args[]){
    String[] arr ={"jpg","css","gif"};

    try
    {
        Scanner sc=new Scanner(new File("c:/weblogd
        FileOutputStream fos=new FileOutputStream("

      while(sc.hasNext())
      {
          int check=1;
          String s1=sc.next();
          s1=s1.toLowerCase();
```

Fig. 6 A code for cleaning suffix from web file

After the cleaning process has been done then information is describe below the table I.

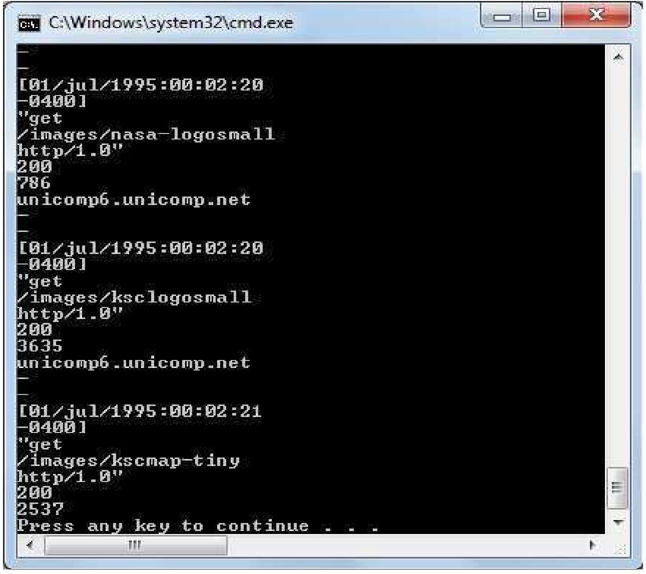
TABLE I.

Evaluation of Log Data

Cleaning Process	*Log File*	*Size(Bytes)*	*Number of Records*
Before Cleaning	LF	205,242,368	17026415
After Cleaning	LF	122,548,815	10315667

Page Access Frequency

Log data consists of different types of attribute like IP Address, User Name, Timestamp, access Request, Status code, Byte Transferred, Referrer, User Agent etc. After studying all these different types of attributes we experiment on the access request. In this access request, the information regarding that



```
-
-
[01/jul/1995:00:02:20
-0400]
"get
/images/nasa-logosmall
http/1.0"
200
786
unicomp6.unicomp.net
-
-
[01/jul/1995:00:02:20
-0400]
"get
/images/ksclogosmall
http/1.0"
200
3635
unicomp6.unicomp.net
-
-
[01/jul/1995:00:02:21
-0400]
"get
/images/kscmap-tiny
http/1.0"
200
2537
Press any key to continue . . .
```

Fig. 7 Result after cleaning process

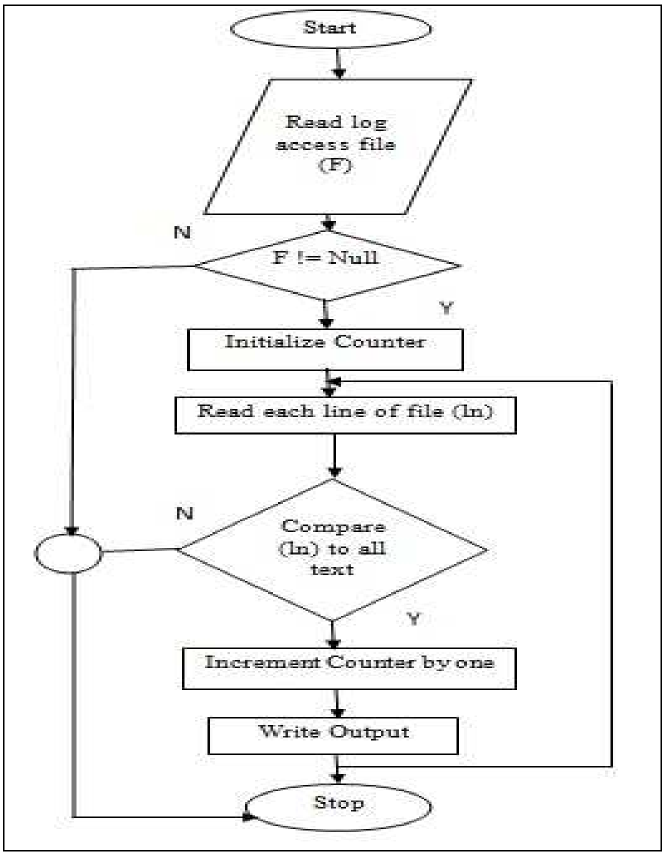


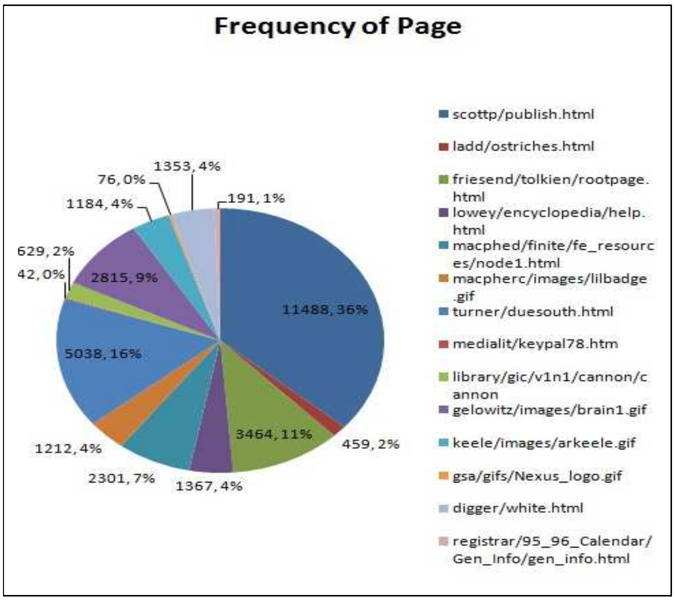
Fig. 8 Flow diagram for frequency of page

In this article we use the dataset is available at The Internet Traffic Archive sponsored by ACM SIGCOMM for the time period 1 July to 31 July1995[9], and divide that data into different types according to its attribute.

Pages	Frequency
scottp/publish.html	11488
ladd/ostriches.html	459
friesend/tolkien/rootpage.html	3464
lowey/encyclopedia/help.html	1367
macphed/finite/fe_resources/node1.html	2301
macpherc/images/lilbadge.gif	1212
turner/duesouth.html	5038
medialit/keypal78.htm	42
library/gic/v1n1/cannon/cannon	629
gelowitz/images/brain1.gif	2815
keele/images/arkeele.gif	1184
gsa/gifs/Nexus_logo.gif	76
digger/white.html	1353
registrar/95_96_Calendar/Gen_Info/gen_info.html	191

Fig. 9 Frequency of an individual page

Out of which we access one part those name is Access Page. In this part we applied data preprocessing technique at the starting. Then we counted the frequency of the access page, the term Frequency of page means that the numbers of visibility of that page in a particular file.



We implement this process for which we read the string line by line and checked it with other string s present in the file. If the string gets matched again and again then we increment its counting by one each time and this counting only shows its frequency. This process is repeated until we reach the end of file. We can easily understand that process to above figure.

V. CONCLUSION

The important task of Web Usage Mining is data preprocessing. For applying data mining techniques our data must be preprocessed, after the can be use data mining techniques like classification, clustering and association rule mining etc. Generally data preprocessing is a time consuming process. In this article we proposed an algorithm for web log data cleaning and also implemented this algorithm in to programming language. When our data would be preprocessed then we can easily detect the user's behavior that used the website. The term user behavior means that how much time user use a particular website? What they surf on that? What are their interests? According to this kind of information we can judge the interesting information applying some data mining techniques. Here we counted the page access frequency and defined different pages. So that highest number of accessed page and lowest number of accessed page can be estimated.

ACKNOWLEDGMENT

We wish to thank the anonymous Reviewers and the authors of references.

REFERENCES

[1] Theint Theint Aye, "Web Log Cleaning of Web Usage Patterns," IEEE, **2011**.

[2] Ms.Dipa Dixit and Ms. M. Kiruthika, "Preprocessing of Web Logs," International Journal on Computer Science and Engineering,vol. **02**, **2010**.

[3] Arshi Shamsi, Rahul Nayak, Pankaj Pratap Singh and Mahesh Kumar Tiwari , "Web Usage Mining by Data Preprocessing," IJCST, vol. 3, **2012**.

[4] Mahendra Pratap Yadav,Pankaj Kumar Keserwani and Shefalika Ghosh Samaddar, "An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner," IEEE, **2012**.

[5] Shaimaa Ezzat Salama, Mohamed I. Marie, "Web Server Logs preprocessing for Web Intrusion Detection," Computer and Information Science, vol. **4**, **2011**.

[6] Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations, vol. **1, 2000**.

[7] Liu Kewen, "Analysis of Preprocessing Methods for Web Usage Data," International Conference on Measurement , Information and control(MIC),IEEE,**2012**.

[8] R. Cooley,B. Mobasher and J Shrivastava, "Web Mining:information and pattern discoveryon the World Wide web," Ninth International Conference, **2011**.

[9] Web Log Data, "http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html,".

[10] Zhuang Like, Kou Zhongbao and Zhang Changshui, "Session identification based on time intervals in Web log mining," Journal of Tsinghua University (Science and Technology), **2005**.

[11] N. Zhang and W. F. Lu, " An Efficient Data Preprocessing Method for Mining Customer Survey Data," IEEE, **2007**.

[12] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood, " Web Usage Mining: A Survey on Preprocessing of Web Log File," IEEE, **2010**.

[13] T. Murata and K. Saito, "Extracting Users' Interests from Web Log Data," Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI **2006** Main Conference Proceedings, **2006**.

[14] Ling Zheng , Hui Gui and Feng Li, "Optimized Data Preprocessing Technology for Web Log Mining," International Conference On Computer Design And Appliations ICCDA, **2010**.

[15] R. Cooley, B. Mobasher and J. Srivastava, "Data preparation for mining world wide web browsing patterns," Knowledge and Information System, **1999**.

[16] Brijesh Bakariya and G.S.Thakur, "Preprocessing on Web Log Data in Web Usage Mining," International Conference on Intelligent Computing and Information System ICICIS, **2012**.

[17] Thi Thanh Sang Nguyen, Hai Yan Lu and Jie Lu, "Web-page Recommendation based on Web Usage and Domain Knowledge," IEEE, **2013**.

AUTHORS PROFILE

Brijesh Bakariya received Graduation degree From Barkatullah University Bhopal MP in 2005, and Post Graduation Degree in Computer Applications From DAVV Indore in year 2009. He is Currently Pursuing the PhD. Degree in the Department of Computer Application s, Maulana Azad National Institute of Technology Bhopal. M.P. His Research interests include Web Mining and Clustering.

Dr. Ghanshyam Singh Thakur has received BSc degree from Dr. Hari Singh Gour University Sagar M.P. in 2000. He has received MCA degree in 2003 from Pt. RaviShankar Shukal University Raipur C.G. and PhD degree from Barkhatullah University, Bhopal M.P. in year 2009. He is Assistant Professor in the department of Computer Applications, Maulana Azad National Institute of technology, Bhopal, M.P. India. He has eight year teaching and research experience. He has 26 publications in national and international journals. His research interests include Text Mining, Document clustering, Information Retrieval, Data Warehousing. He is a member of the CSI, IAENG, and IACSIT.