Learning Web Behavior through Classification by Extracting Web Log

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: *The society has impacted at the point of every way of our nation and it is also increasing generation by day discipline to its friendliness & scalability to the users. Since the number of Society sites and web pages has grown rapidly, discovering and genius web user's Behavior is consequential for the development of successful endorsement systems. WUM is the way of applying web mining method to extract useful information from the weblog. Web page classification is the supervised learning activity, which is a style of classifying web pages through user's visits. This classification helps to the visited user's behavior. In this current System, Apriori algorithm for association rule technique is implemented in endorsement system. In this technique execution time is more by the whole of large dataset and don't valuable accuracy Apriori algorithm. The Proposed System, Naïve Bayes classification used for classifying the web page Url & resolved the accuracy. The manner of the Naïve Bayes Algorithm for preprocessing web log and result the probability of web pages to endorsement web page navigation Url.*

Key Words: *Web logs, Web Mining, WUM, Classification, Associations technique .*

1. INTRODUCTION:

WM is a literally small part of the vast data mining field. It is an emerging considers area. This chapter will provide a complete background study containing basic concepts, technologies, methods, applications, and tools hand me down, etc. It will be uphold to be devoted by the whole of Data Mining & Web mining subjects [13].

WM manifest between and copes mutually semi structured data or unstructured data. Web mining calls for creative regard of data mining. Mining the WWW data is such of the challenging tasks for the DM and data management scholars inasmuch as there are full data that is less structured data available on the web and we can obviously get overwhelmed with data.

WM is the vast area of data mining methods to find interesting and information useful data from web data. It's freely considered that either the hyperlink node of the website, or its contents or web log data that are used in the mining process. It is used in data confirmation and Validation, data integrity and, content management, content generation and Survey System [13].

Application of Web Mining

Web Content Mining:

Web content mining is the mining, extraction and composite of convenient data, information and lifestyle from Web page content. The analyzes web content such as question, multimedia data, arrangement data. This is done to recognize the content of World Wide Web pages,

- demonstrative key language based indexing,
- web page ranking,
- web page content summarization
- Web search and analysis.

Web Structure Mining:

Web structure mining is the behavior of by graph the big idea to correlate the node and love arrangement of a World Wide Web site. According to the personality of Web structural data, web structure mining can be differentiates into two kinds: web structural data.

Extracting chain from hyperlinks in the web: a hyperlink is a structural factor that connects the web page to a diverse location.

Mining the file structure: Analysis of the treelike structure of page structures to decide HTML or XML haunt usage.

Web Usage Mining:

WUM is the procedure of extracting important information from logs. It helps to improve search efficiency and effectiveness. The search companies routinely conduct WUM to improve their good quality of service.

Method of Web Usage Mining

Pre Processing

The data pre-processing of Web usage mining is usually complex. The data pre-processing is to offer structural, reliable and integrated data source to pattern discovery. It consists of four steps: data cleaning, user identification, session identification, path completion.

Pattern Discovery

In this stage, DM methods are used in order to extract chain of usage form Web data. Patter discovery is the key process of the WM, which covers the algorithms and method from several research locations, such as DM, ML, statistics and pattern recognition. The

Method such as statistical analysis, association rules, clustering, classification, sequential chain and dependency modelling are used to discover rules and pattern.

Pattern Analysis
The last stage of the WUM is chain analysis. The goal of this procedure is to extract the interesting rules or patters from the output of the patter discovery process by eliminating the relative rules or patterns.

2. WEB LOG FILE

Web log file is log file automatically created and maintained by internet server. The raw web log prosecute format is for all practical purposes one line of demarcation of text for each hit to the web site. This contains information about who was visiting the site, to what place they came from, and actually what they were doing on the website.

Types of Web Log file
There are three types of log files which are as follows:

- **Web Server Logs**

The web page requests is maintained as a log file. These logs contain all field of C panel data, the request line exactly came from the client, etc. These data can be bound together as a single text file, or divided into different logs, like access log, referrer log, or error log.

- **Proxy Server Logs**

It acts as an intervening level of lies between client browser and web servers. Proxy caching is used to less loading time of a web page as well as the reducer's traffic at the server and client side. It is used as a data source for browsing behavior characterization of a group of unauthorized users sharing a common proxy server.

- **Browser Logs**

On client side gather data, user cooperation is needed. Here pre-processing discussed using Web Server Logs. Web server logs are used in the web page recommendation to improve the E-Commerce usability.

Format of Web Log file
Web log file is a simple plain text file which record information about each user.

- **W3C Extended log file**

The W3C log format is the default log file format on the IIS server. All fields is separated by space; time is recorded as GMT. It can be customized, that is administrators can add or remove fields depending on what information want to record.

- **NCSA common log file**

The NCSA Common log file format is a static ASCII text-based format, so you can't customize it. The Web sites and for SMTP and NNTP services, but it is not available for FTP sites. Because HTTP.sys handles the log file format, in this format records HTTP.sys kernel-mode cache hits.

- **IIS log file**

The IIS log file format is a static ASCII text-based format, so you can't customize it. Because HTTP.sys handles log file format, this format records HTTP.sys kernel-mode cache hits.

- **Apache log file**

The Apache log file format is a static ASCII text based format, so you can't customize it.

Example of Web Log file
2016-09-09 9:30:10 10.10.10.12 GET /default.aspx - 80 – 172.16.0.0.1 Wget/1.12+ 200 0 0 18182

TABLE: 1 Log Format

Field Name	Field Description	Example
Date	The date that the activity occurred	2016-09-09
Time	The time that the activity Occurred	9:30:10
s-ip:	The IP address of the server	10.10.10.12
cs-method	The action of client was trying to perform	GET
cs-uri-stem	The Resource accessed	/default.aspx
cs-uri-query	The query, if any, the client was performing	-
s-port	The port number of client is connected to	80
cs-username	The name of the authenticated user who accessed your server. This doesn't include anonymous users, who are represented by a hyphen.	-
c-ip	The ip address of the client	172.16.0.0.1
cs(User-Agent)	The browser used on the client	Wget/1.12+
sc-status	The statues of action ,in http or ftp terms	200
Sc-substatus	Is the sub status e.g. for a 503.19 HTTP status it would be the 19 part	0

Sc-win32-status	The Status of the action , in terms used by Microsoft windows	0
time taken	The duration of time, in Millisecond, that the action Consumed	18182

3. METHODOLOGIES:

Pre-processing of Web log is an important phase in web usage mining.

Information Cleaning

The use of data cleaning procedure is to unfasten all the unwanted data used in data analysis and mining. To increase the mining efficiency data cleaning is very important. The cleaned data include removal of local and global noise, elimination of videos, graphic records and the format efficiency, elimination of HTTP status code records [4].

Algorithm 1: Information Cleaning [4]

```
Input: l_IP1
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
```

User identification

Each different user accessing the website is identified in the user recognition process. The aim of this process is to retrieve every user's access characteristics, then make user clustering and provide endorsement service for the users. Different users are identified by different IP addresses.

Algorithm 2: User Identification [4]

```
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
```

User Session Identification

A Structure of pages viewed by a user during one visit is known as the Session. The session is stored in the log file. In pre-processing it is necessary to search session of each user. It's defines the number of times the user has Visited a web page. It takes all the page reference of a given user in a log and divides them into user sessions. These sessions can be used as an input data vector in classification, clustering, prediction and other tasks. Based on a uniform fixed timeout a traditional session recognition algorithm is used. A new session is identified when the interval between two sequential requests exceeds the one hour.

Algorithm 3: User Session Identification [4]

```
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.   increment s_ID
5.   else  assign the old s_ID.
            End else
                 End if
                    End
```

Algorithm 4: Apriori Algorithm [4]

1. Scan all of the transactions to count the number of occurrences of each item.
2. Determine L1- set of Frequent 1 – item set.
3. Generate C2 – 2 –item set candidate and scan D for sup Count.
4. Determine L2 – set of Frequent 2 – item set.
5. Generate C3 – 3 Item set candidate.
6. Find the support count of l5.
7. Determine l3 – the set of frequent 3 item set.
8. Determine C4 – 4 item sets Candidate.Null

Existing System Limitation

In this existing system, Web page Navigation **Url** recommendation using Apriori algorithm is done, but accuracy are still issues for further work. The web navigation link recommendation is more computation time.

Purpose System

From the Literature review, it is concluded that there are many algorithms that are already working to for increase the accuracy and efficiency of web page navigation recommendation system. Very common problem with earlier work is the scalability problem.

Proposed Work will be based on classification of Naïve Bayes algorithm using web page navigation Url recommendation system. Using classification techniques with the suitable dataset we will get good accuracy from the proposed system. Applying Naive Bayes Algorithm to generate the strong rule with the classification algorithm will improve the system result. Solve the common problem like accuracy and efficiency problem.

Naive Bayes Algorithm

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
The Naive Bayes algorithm is good accuracy and scalability. In this algorithm is implemented in xl miner tools which is analysis of the whole data set and we are easily understand about dataset. In Naive Bayes accuracy good based on Purposed system.
NB Algorithm is easy to build and particularly used for large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.
Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

P(C/X) = P(X/C) P(C) / P(X)

P(c/x) = P(x1/c) * P(x2/c)*... * P (x n/c) * p(c)

P(C) Class Prior Probability

P(X) Predictor Prior Probability

P(*c*|*x*) is the posterior probability of *class* (c, target) given predictor (x, attributes).

P(*c*) is the prior probability of class.

P(*x*|*c*) is the likelihood which is the probability of predictor given class.

P(*x*) is the prior probability of predictor.

Purpose System Flow

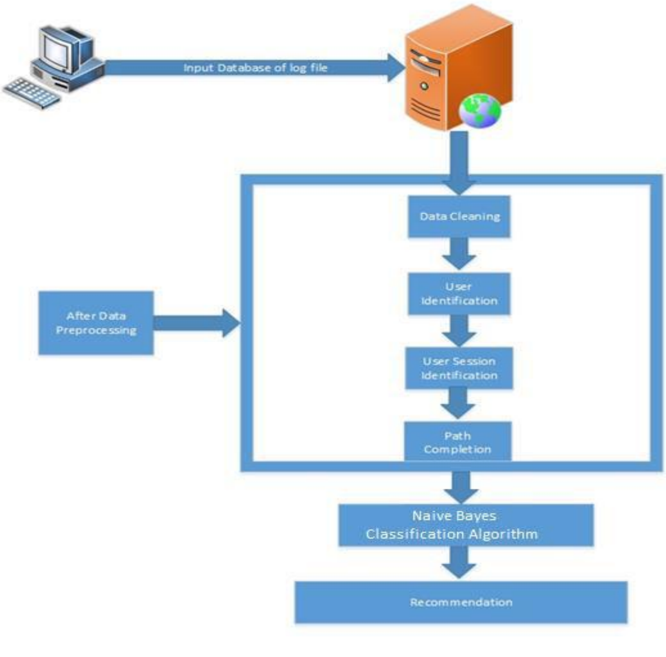

Input Database of log file
Data Cleaning
User Identification
User Session Identification
Path Completion
After Data Preprocessing
Naive Bayes Classification Algorithm
Recommendation

Fig 1. Proposed System Flow

Implementation Process

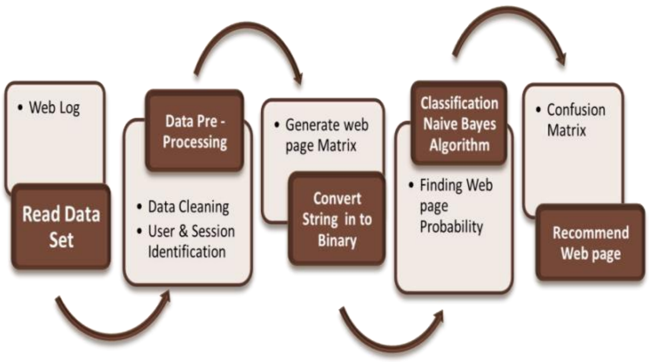
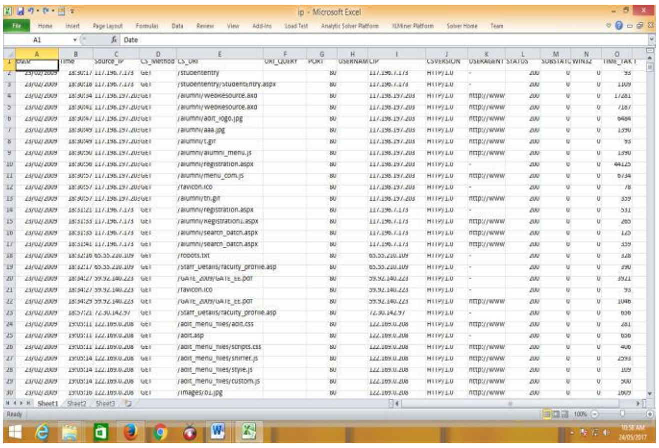

• Web Log
Read Data Set
Data Pre - Processing
• Data Cleaning
• User & Session Identification
• Generate web page Matrix
Convert String in to Binary
Classification Naive Bayes Algorithm
• Finding Web page Probability
• Confusion Matrix
Recommend Web page

Fig 2. Implimentation process Diagram

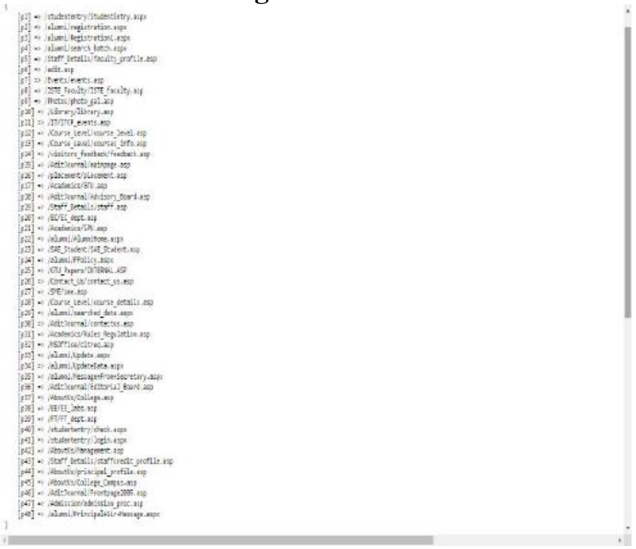
Step 1: Read Dataset [5525 data]
Step 2: After Preprocessing
Rows after Preprocessing: 780
Total No. of Users: 148
Total No. of Web Pages: 47
Step 3: create a Web Page Matrix where user are visited or not visited.1 means visited page and 0 means Not Visited
Step 4: Finding all web page probability
Step 5: Based on Probability, which is not visited pages find &highest probability web page to recommend to user.
Step 6: Implemented Naive Bayes Algorithm.

I. EXPERIMENTAL RESULTS


ip - Microsoft Excel
File Home Insert Page Layout Formulas Data Review View Add-Ins Load Test Analytic Solver Platform XLMiner Platform Solver Home Team
A1
Date
Sheet1 Sheet2 Sheet3
Ready
100%
10:58 AM
24/05/2017

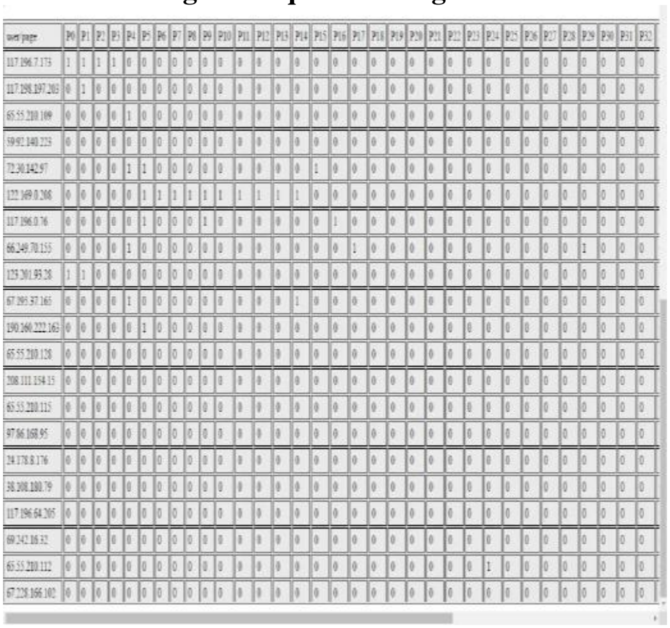
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Date	Time	Source_IP	CS_Method	CS_URI	URI_QUERY	PORT	USERNAME	CIP	CSVERSION	USERAGENT	STATUS	SUBSTATUS	WIN32	TIME_TAK
2	23/02/2009	18:30:17	117.196.7.173	GET	/studententry		80		117.196.7.173	HTTP/1.0	-	200	0	0	93
3	23/02/2009	18:30:18	117.196.7.173	GET	/studententry/studententry.aspx		80		117.196.7.173	HTTP/1.0	-	200	0	0	1109
4	23/02/2009	18:30:34	117.198.197.20	GET	/alumni/webresource.axd		80		117.198.197.203	HTTP/1.0	http://www	200	0	0	17281
5	23/02/2009	18:30:41	117.198.197.20	GET	/alumni/webresource.axd		80		117.198.197.203	HTTP/1.0	http://www	200	0	0	7187
6	23/02/2009	18:30:47	117.198.197.20	GET	/alumni/adit_logo.jpg		80		117.198.197.203	HTTP/1.0	http://www	200	0	0	6484
7	23/02/2009	18:30:49	117.198.197.20	GET	/alumni/aaa.jpg		80		117.198.197.203	HTTP/1.0	http://www	200	0	0	1390
8	23/02/2009	18:30:49	117.198.197.20	GET	/alumni/t.gif		80		117.198.197.203	HTTP/1.0	http://www	200	0	0	93
9	23/02/2009	18:30:50	117.198.197.20	GET	/alumni/alumni_menu.js		80		117.198.197.203	HTTP/1.0	http://www	200	0	0	1390
10	23/02/2009	18:30:56	117.198.197.20	GET	/alumni/registration.aspx		80		117.198.197.203	HTTP/1.0	-	200	0	0	44125
11	23/02/2009	18:30:57	117.198.197.20	GET	/alumni/menu_com.js		80		117.198.197.203	HTTP/1.0	http://www	200	0	0	6734
12	23/02/2009	18:30:57	117.198.197.20	GET	/favicon.ico		80		117.198.197.203	HTTP/1.0	-	200	0	0	78
13	23/02/2009	18:30:57	117.198.197.20	GET	/alumni/tn.gif		80		117.198.197.203	HTTP/1.0	http://www	200	0	0	359
14	23/02/2009	18:31:21	117.196.7.173	GET	/alumni/registration.aspx		80		117.196.7.173	HTTP/1.0	-	200	0	0	531
15	23/02/2009	18:31:33	117.196.7.173	GET	/alumni/Registration1.aspx		80		117.196.7.173	HTTP/1.0	http://www	200	0	0	265
16	23/02/2009	18:31:35	117.196.7.173	GET	/alumni/search_batch.aspx		80		117.196.7.173	HTTP/1.0	http://www	200	0	0	125
17	23/02/2009	18:31:41	117.196.7.173	GET	/alumni/search_batch.aspx		80		117.196.7.173	HTTP/1.0	http://www	200	0	0	359
18	23/02/2009	18:32:16	65.55.210.109	GET	/robots.txt		80		65.55.210.109	HTTP/1.0	-	200	0	0	328
19	23/02/2009	18:32:17	65.55.210.109	GET	/Staff_Details/faculty_profile.asp		80		65.55.210.109	HTTP/1.0	-	200	0	0	390
20	23/02/2009	18:34:27	59.92.140.223	GET	/GATE_2009/GATE_EE.pdf		80		59.92.140.223	HTTP/1.0	-	200	0	0	3921
21	23/02/2009	18:34:27	59.92.140.223	GET	/favicon.ico		80		59.92.140.223	HTTP/1.0	-	200	0	0	93
22	23/02/2009	18:34:29	59.92.140.223	GET	/GATE_2009/GATE_EE.pdf		80		59.92.140.223	HTTP/1.0	http://www	200	0	0	1046
23	23/02/2009	18:57:21	72.30.142.97	GET	/Staff_Details/faculty_profile.asp		80		72.30.142.97	HTTP/1.0	-	200	0	0	656
24	23/02/2009	19:05:11	122.169.0.208	GET	/adit_menu_files/adit.css		80		122.169.0.208	HTTP/1.0	http://www	200	0	0	281
25	23/02/2009	19:05:11	122.169.0.208	GET	/adit.asp		80		122.169.0.208	HTTP/1.0	-	200	0	0	656
26	23/02/2009	19:05:11	122.169.0.208	GET	/adit_menu_files/scripts.css		80		122.169.0.208	HTTP/1.0	http://www	200	0	0	406
27	23/02/2009	19:05:14	122.169.0.208	GET	/adit_menu_files/shifter.js		80		122.169.0.208	HTTP/1.0	http://www	200	0	0	2593
28	23/02/2009	19:05:14	122.169.0.208	GET	/adit_menu_files/style.js		80		122.169.0.208	HTTP/1.0	http://www	200	0	0	109
29	23/02/2009	19:05:14	122.169.0.208	GET	/adit_menu_files/custom.js		80		122.169.0.208	HTTP/1.0	http://www	200	0	0	500
30	23/02/2009	19:05:16	122.169.0.208	GET	/images/01.jpg		80		122.169.0.208	HTTP/1.0	http://www	200	0	0	1609

Fig 3. Data Set



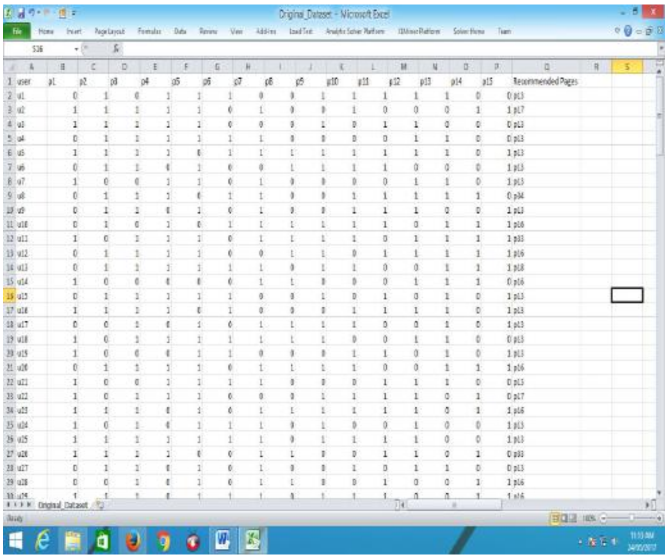
```
{
  [p1] => /studententry/StudentEntry.aspx
  [p2] => /alumni/registration.aspx
  [p3] => /alumni/Registration1.aspx
  [p4] => /alumni/search_batch.aspx
  [p5] => /Staff_Details/faculty_profile.asp
  [p6] => /adit.asp
  [p7] => /Events/events.asp
  [p8] => /ISTE_Faculty/ISTE_faculty.asp
  [p9] => /Photos/photo_gal.asp
  [p10] => /Library/library.asp
  [p11] => /IT/ITCP_events.asp
  [p12] => /Course_Level/course_level.asp
  [p13] => /Course_Level/course1_info.asp
  [p14] => /visitors_feedback/feedback.asp
  [p15] => /AditJournal/mainpage.asp
  [p16] => /placement/placement.asp
  [p17] => /Academics/GTU.asp
  [p18] => /AditJournal/Advisory_Board.asp
  [p19] => /Staff_Details/staff.asp
  [p20] => /EC/EC_dept.asp
  [p21] => /Academics/SPU.asp
  [p22] => /alumni/AlumniHome.aspx
  [p23] => /SAE_Student/SAE_Student.asp
  [p24] => /alumni/PPolicy.aspx
  [p25] => /GTU_Papers/INTERNAL.ASP
  [p26] => /Contact_Us/contact_us.asp
  [p27] => /SPE/see.asp
  [p28] => /Course_Level/course_details.asp
  [p29] => /alumni/searched_data.aspx
  [p30] => /AditJournal/contactus.asp
  [p31] => /Academics/Rules_Regulation.asp
  [p32] => /MSOffice/cltreq.asp
  [p33] => /alumni/Update.aspx
  [p34] => /alumni/UpdateData.aspx
  [p35] => /alumni/MessageFromSecretary.aspx
  [p36] => /AditJournal/Editorial_Board.asp
  [p37] => /AboutUs/College.asp
  [p38] => /EE/EE_labs.asp
  [p39] => /FT/FT_dept.asp
  [p40] => /studententry/check.aspx
  [p41] => /studententry/login.aspx
  [p42] => /AboutUs/Management.asp
  [p43] => /Staff_Details/staffcredit_profile.asp
  [p44] => /AboutUs/principal_profile.asp
  [p45] => /AboutUs/College_Campus.asp
  [p46] => /AditJournal/Frontpage2006.asp
  [p47] => /Admission/admission_proc.asp
  [p48] => /alumni/Principal&Sir-Message.aspx
}
```

Fig 4. Unique Web Page Url



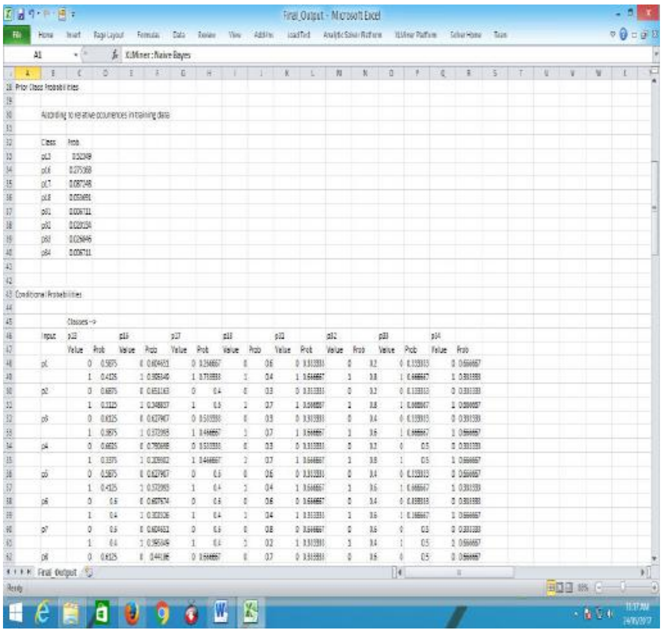
user/page	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30	P31	P32
117.196.7.173	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
117.198.197.203	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
65.55.210.109	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
59.92.140.223	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
72.30.142.97	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
122.169.0.208	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
117.196.0.76	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
66.249.70.155	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
123.201.93.28	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
67.195.37.165	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
190.160.222.163	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
65.55.210.128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
208.111.154.15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
65.55.210.115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
97.86.168.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24.178.8.176	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38.108.180.79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
117.196.64.205	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69.242.16.32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
65.55.210.112	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
67.228.166.102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig 5. Web Page Matrix



user	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	Recommended Pages
u1	0	1	0	1	1	1	0	0	1	1	1	1	1	0	0	p13
u2	1	1	1	1	1	0	1	0	0	1	0	0	0	1	1	p17
u3	1	1	1	1	1	0	0	0	1	0	1	1	0	0	0	p13
u4	0	1	1	1	1	1	1	0	0	0	0	1	1	0	0	p13
u5	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	p13
u6	0	1	1	0	1	0	0	1	1	1	1	0	0	0	1	p13
u7	1	0	0	1	1	0	1	0	0	0	0	1	1	0	1	p13
u8	0	1	1	1	0	1	1	0	0	1	1	1	1	1	0	p34
u9	0	1	1	0	1	0	1	0	0	1	1	1	0	0	1	p13
u10	0	1	0	1	0	1	1	1	1	1	1	0	1	1	1	p16
u11	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	p33
u12	0	1	1	1	1	0	0	1	1	0	1	1	1	1	1	p16
u13	0	1	1	1	1	1	1	0	1	1	0	0	1	1	1	p18
u14	1	0	0	0	0	0	1	1	0	0	0	1	1	1	0	p16
u15	0	1	1	1	1	1	0	0	1	0	1	0	1	0	1	p13
u16	1	1	1	1	0	1	0	0	0	1	1	1	1	0	1	p13
u17	0	0	1	0	1	0	1	1	1	1	0	0	1	0	1	p13
u18	1	0	1	1	1	1	1	1	1	0	0	1	1	0	0	p13
u19	1	0	0	0	1	1	0	0	0	1	1	0	1	0	1	p13
u20	0	1	1	1	1	0	1	1	1	1	0	0	1	1	1	p16
u21	1	0	0	1	1	1	1	0	0	0	1	1	1	0	0	p13
u22	1	0	1	1	1	0	0	0	1	1	1	1	0	1	0	p17
u23	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	p16
u24	1	0	1	0	1	1	1	0	1	0	0	1	0	0	1	p13
u25	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	p13
u26	1	1	1	1	0	0	1	1	0	0	1	1	0	1	0	p33
u27	0	1	1	0	1	0	1	0	0	1	0	1	1	0	0	p13
u28	0	0	1	0	1	0	1	1	0	0	1	0	0	1	1	p16
u29	1	1	1	0	1	1	1	0	1	1	1	0	0	1	1	p16

Fig 6. Recommended Web Pages



Prior Class Probabilities

According to relative occurrences in training data

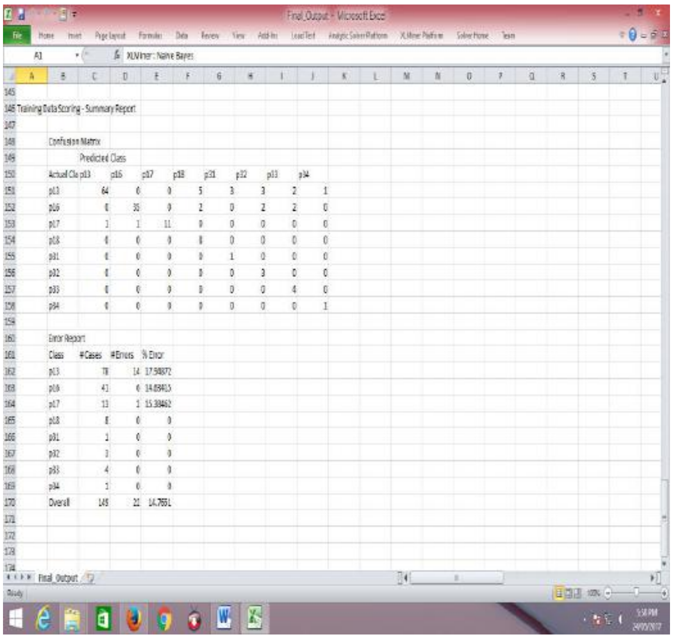
Class	Prob
p13	0.52349
p16	0.275168
p17	0.087248
p18	0.053691
p31	0.006711
p32	0.020134
p33	0.026846
p34	0.006711

Conditional Probabilities

Classes -->

Input	p13		p16		p17		p18		p31		p32		p33		p34	
	Value	Prob	Value	Prob	Value	Prob	Value	Prob	Value	Prob	Value	Prob	Value	Prob	Value	Prob
p1	0	0.5875	0	0.604651	0	0.266667	0	0.6	0	0.333333	0	0.2	0	0.333333	0	0.666667
	1	0.4125	1	0.395349	1	0.733333	1	0.4	1	0.666667	1	0.8	1	0.666667	1	0.333333
p2	0	0.6875	0	0.651163	0	0.4	0	0.3	0	0.333333	0	0.2	0	0.333333	0	0.333333
	1	0.3125	1	0.348837	1	0.6	1	0.7	1	0.666667	1	0.8	1	0.666667	1	0.666667
p3	0	0.0125	0	0.627907	0	0.533333	0	0.3	0	0.333333	0	0.4	0	0.333333	0	0.333333
	1	0.9875	1	0.372093	1	0.466667	1	0.7	1	0.666667	1	0.6	1	0.666667	1	0.666667
p4	0	0.6625	0	0.790698	0	0.533333	0	0.3	0	0.333333	0	0.2	0	0.5	0	0.333333
	1	0.3375	1	0.209302	1	0.466667	1	0.7	1	0.666667	1	0.8	1	0.5	1	0.666667
p5	0	0.5875	0	0.627907	0	0.6	0	0.6	0	0.333333	0	0.4	0	0.333333	0	0.666667
	1	0.4125	1	0.372093	1	0.4	1	0.4	1	0.666667	1	0.6	1	0.666667	1	0.333333
p6	0	0.6	0	0.697674	0	0.6	0	0.6	0	0.666667	0	0.4	0	0.833333	0	0.333333
	1	0.4	1	0.302326	1	0.4	1	0.4	1	0.333333	1	0.6	1	0.166667	1	0.666667
p7	0	0.6	0	0.604651	0	0.6	0	0.8	0	0.666667	0	0.6	0	0.5	0	0.333333
	1	0.4	1	0.395349	1	0.4	1	0.2	1	0.333333	1	0.4	1	0.5	1	0.666667
p8	0	0.6125	0	0.44186	0	0.666667	0	0.7	0	0.333333	0	0.6	0	0.5	0	0.666667

Fig 7. Prior Probability



Training Data Scoring - Summary Report

Confusion Matrix

Predicted Class

Actual Class	p13	p16	p17	p18	p31	p32	p33	p34
p13	64	0	0	5	3	3	2	1
p16	0	35	0	2	0	2	2	0
p17	1	1	11	0	0	0	0	0
p18	0	0	0	8	0	0	0	0
p31	0	0	0	0	1	0	0	0
p32	0	0	0	0	0	3	0	0
p33	0	0	0	0	0	0	4	0
p34	0	0	0	0	0	0	0	1

Error Report

Class	#Cases	#Errors	% Error
p13	78	14	17.94872
p16	41	6	14.63415
p17	13	2	15.38462
p18	8	0	0
p31	1	0	0
p32	3	0	0
p33	4	0	0
p34	1	0	0
Overall	149	22	14.7651

Fig 8. Confusion Matrix

TABLE: 2 Comparison Parameter

Algorithm	Accuracy	Efficiency	Execution Time
Apriori Algorithm	80%	Good	More
KNN Algorithm	78%	Bad	Medium
Naïve Bayes Algorithm	86%	Good	Less

4. CONCLUSION:

Web Usage Mining is the one of the large areas of research and improves the sub domain of data mining and method. Web Navigation Recommendation systems provide valuable suggestions to Recommend Url to users. There is the Accuracy problem in existing system . To solve this problem ,here we will Used Classification of Naïve Bayes Algorithm to increase the 87% accuracy of the Purpose system.

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 K mean Algorithm, Decision Tree Algorithm can apply in a web log.

ACKNOWLEDGEMENT

In this paper Second author is sincerely intelligent to the faculty members in Department of Information Technology of Parul Institute of Technology for the constant encouragement and Motivation to us provided throughout the period of this research work.

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
9. International Conference on Advance In computer Engineering and application
10. Wichian Premchaiswadi, Walisa Romasaiyud,” Extracting Weblog of Siam University For Learning User Behavior On Map Reduce,”2011 IEEE
11. Yuqi Wang, Wenquiam Shang,” Personalized News Recommendation Based on Consumer’s Click Behavior,”2015 IEEE
12. Ying Han, Kejian Xia,” Data Preprocessing Method Based On user characteristic of interest for web log mining,”2014 IEEE
13. Liu Kewen,” Analysis of Preprocessing Methods for web Usage data,”2012 IEEE