ASSOCIATION RULE MINING ON WEB LOGS FOR EXTRACTING INTERESTING PATTERNS THROUGH WEKA TOOL

Ankit Kumar[1], Omkar singh[2], Vinay Rishiwal[3], Rakesh Kumar Dwivedi[4], Rakesh Kumar[5]

[1,4]Dept of CS/IT, Teerthanker Mahaveer University ,,Moradabad, (India)
[2,5]Dept of CS/IT, Dev Bhoomi Institute of Technology Dehradun, (India)
[3]Dept of CS/IT,IET, MJP Rohilkhand University, Bareilly, (India)

ABSTRACT

The Web Usage Mining is a process for extracting the navigation behavior of the user which are gathered by the web servers and stored in server access logs. The analysis of the server logs through WEKA that can provide organization with information on how to make better structure for the web site to effective use and benefit of the organization. The data is collected from the server access logs which are generated as a result of interaction between the client and the server. The two important tasks in the data preprocessing phase are data cleaning and feature extraction. The first phase consists of Data Fusion, Data Cleaning, User Identification, Session Identification, Path Completion, Formatting and Data Summarization. After data preprocessing some of the features are added from the log file this is called Feature Extraction. Then this file serves as an input to our program.

***Keywords: Web Usage Mining, Apriori, Weka*.**

I. INTRODUCTION

The WWW serves as a vast, widely distributed, global information service center for advertisement, consumer information, e-commerce, education, financial management, government, news and other services. So, it has become much more difficult to access relevant information from the web with the

Explosive growth of information available on the internet [1]. Therefore, further research work needs to be carried out on the existing web services as the services offered are not so adequate enough to satisfy the needs of different web users. Users accessed millions of web pages for business and individual transactions. The servers stored massive of amount of page access data.

II. WEB USAGE MINING

Web usage mining is a part of data mining in which the meaningful information is extracted from the Web Server Log for the various purpose such as for the business strategies, financial activities etc. through web mining we automatic detection the user access activity and access patterns from many web servers. The purpose of wen usage mining to find the user access patterns.. Web mining can be practices in three different domains i.e. the content mining, hyper link web structure mining and web usage mining [2]. These approaches effort to extract valuable information from the web which is then applied to some real world problems

Five major steps followed in web usage mining are

1. Data collection – Web log files, which keeps track of visits of all the visitors
2. Data Integration – Integrate multiple log files into a single file
3. Data preprocessing – Cleaning and structuring data to prepare for pattern extraction
4. Pattern extraction – Extracting interesting patterns
5. Pattern analysis and visualization – Analyze the extracted pattern

2.2 Steps in Web Usage Mining

2.2.1 Data Collection

The data is collected from the log file which supports three different formats. The log information can be collected from the Sever log file, Proxy server log file and client log. If the cookies are enabled then we can get the information of the user access on the client system.

2.2.2 Data Preparation

In this phase we restore the user patterns through web log in a consistent way. This phase should at a minimum achieve the following four major tasks:

- Removing undesirable entries
- Distinguishing among users
- Building sessions
- Restoring the contents of a session

2.2.3 Pattern Discovery/ Pattern Extraction

Here we find the interesting patterns from the web log file. Now we applied the pattern mining algo on the web log data. So preprocessing is important and it is carried out with proper care. Pattern Discovery is the key component of web usage mining [8,10].

III. PROBLEM STATEMENT

In the internet web sites on the internet are useful source of information for almost every activity. Web mining is the application of data mining, artificial intelligence, chart technology and so on to the web data and traces user’s visiting behaviors and extracts their interests using patterns. Because of its direct application in e-commerce, Web analytics, e-learning, information retrieval, web mining has become one of the important areas in computer and information science. There are several techniques like web usage mining exists. But all processes its own disadvantages. This study focuses on providing techniques for better data cleaning and transaction identification from the web log.

IV. APRIORI ALGORITHM

Apriori is an algo which is given by R. Agrawal and R Srikant in 1994 [3] for extracting thr frequent item sets for Boolean association rule. Apriori make use of an iterative approach known as Level-wise search, where k item set are used to explore (k+1) item sets. Each iteration consists of two steps [4].

i. Makes the candidate item sets.

ii. Now we count the existence of the item set in the database. And Prunes disqualified

Now we apply the Apriori: Apriori uses two pruning techniques

i. First is based on Support Count (Greater than User specified support threshold)

ii. Item set to be frequent, all its subset should be in last frequent item set.

According the algorithm if a set of items is frequent, [4] then all its proper subsets is also frequent.

Apriori algorithm in pseudocode

```
$L_1$= {frequent items};
for (k= 2; $L_{k-1}$ !=∅; k++) do begin
$C_k$= candidates generated from $L_{k-1}$ (that is: Cartesian product $L_{k-1}$ x $L_{k-1}$ and eliminating any k-1 size itemset that is not frequent);
for each transaction t in database do increment the count of all candidates in $C_k$ that are contained in t
$L_k$ = candidates in $C_k$ with min_sup end
return $L_k$;
```

V. MATERIAL & METHODOLOGIES

We apply the experiment on filtered data set of the web log file. The concept of data mining onto the filtered data. We use data mining technique for finding the meaningful knowledge from the big dataset. This is also known as Knowledge Discovery [5,10].

Step 1

Client sends the request to the server for the web pages he/she wants to access on the web.

Step 2

Server log file contains the data which is generated as of an interaction between the client-server. The server log file contains nine attributed. This info is raw and most of the data are garbage.

Step 3

The raw data is preprocessed. During the preprocessing the unwanted data is removed.

Step 4

Now we apply the data mining techniques on the extracting information. Again the data is filtered for applying finding the frequent sequence pattern.

Step 5

Now, apply the Sequential data mining technique based on Apriori rule. This algorithm mines the filtered database and it looks for frequent patterns which is also known as frequent sequences which afterwards used by end user for finding the relation between the different [10]events.

Step 6

After applying Recommendation Rule generator take into consideration the currently accessed web pages based on some threshold value defined and the pattern that are discovered after applying sequential pattern mining based on apriori algorithm into consideration and generate the pages [7,10] that are frequently accessed by the user.

Step 7

Now we sent the rule to the web server which recommended by the by the recommendation rule generator system then client accesses any web page. The web page contains those links that are of his/her interest [7, 8].

VI. INTRODUCTION TO WEKA

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains [9,10]. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from our own Java code. Weka contains tools for

- Pre-Processing

  Preprocessing is used to choose the data file to be used by the application.
- Classification

  Classification is used to test and train different learning schemes on the preprocessed data file under experimentation
- Clustering

  Clustering is used to apply different tools that identify clusters within the data file
- Association Rule Extraction

  Association rule extraction used to apply different rules to the data file that identify association within the data
- Select Attributes

  Select attributes is used to apply different rules to reveal changes based on selected attributes inclusion or exclusion from the experiment
- Visualize

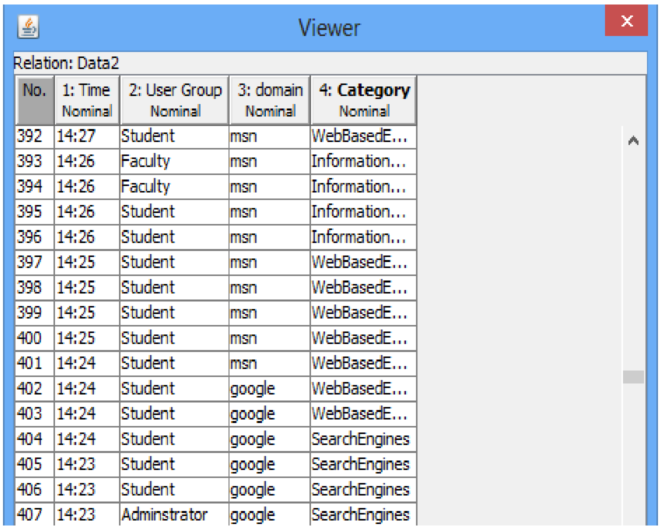
  Visualize is used to see what the various manipulation produced on the data set in a 2D format, in scatter plot and bar graph output [10,11].

VII. EXPERIMENTS AND RESULTS

In the following tables, we are making the sequences of the web sites accessed by the users by making the groups of the websites into categories. The Apriori algorithm calculates the support count of each category.

Filtered Data in Microsoft Excel that serves as an input to Weka.

Input File Name – Data_Set.csv.



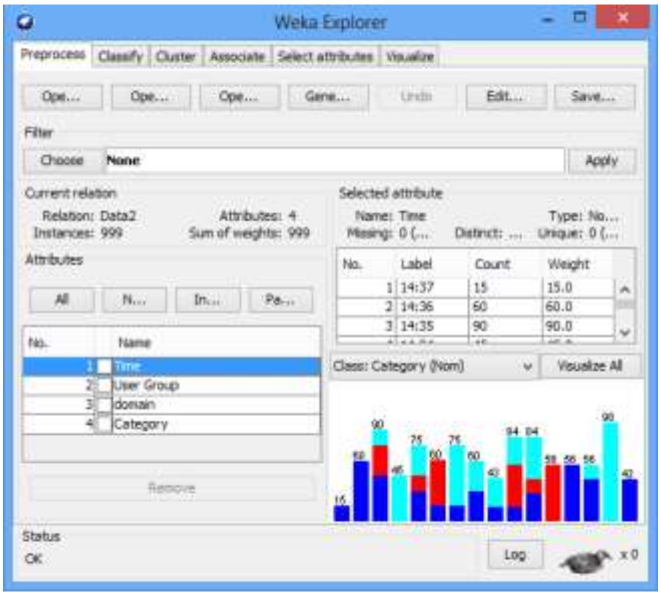
Relation: Data2

No.	1: Time Nominal	2: User Group Nominal	3: domain Nominal	4: Category Nominal
392	14:27	Student	msn	WebBasedE...
393	14:26	Faculty	msn	Information...
394	14:26	Faculty	msn	Information...
395	14:26	Student	msn	Information...
396	14:26	Student	msn	Information...
397	14:25	Student	msn	WebBasedE...
398	14:25	Student	msn	WebBasedE...
399	14:25	Student	msn	WebBasedE...
400	14:25	Student	msn	WebBasedE...
401	14:24	Student	msn	WebBasedE...
402	14:24	Student	google	WebBasedE...
403	14:24	Student	google	WebBasedE...
404	14:24	Student	google	SearchEngines
405	14:23	Student	google	SearchEngines
406	14:23	Student	google	SearchEngines
407	14:23	Adminstrator	google	SearchEngines

Preprocessing – Pre-processing tools in WEKA are called "filters"

Data can be imported from a file in various formats – .csv format is used in this experiment

WEKA contains filters for: Discretization, normalization, resampling, attribute selection, transforming and combining attributes. Now apply the Preprocessing technique we use the weka tool.



Filter

Choose None

Current relation

Relation: Data2 Attributes: 4
Instances: 999 Sum of weights: 999

Selected attribute

Name: Time Type: No...
Missing: 0 (... Distinct: ... Unique: 0 (...

No.	Label	Count	Weight
1	14:37	15	15.0
2	14:36	60	60.0
3	14:35	90	90.0

Attributes

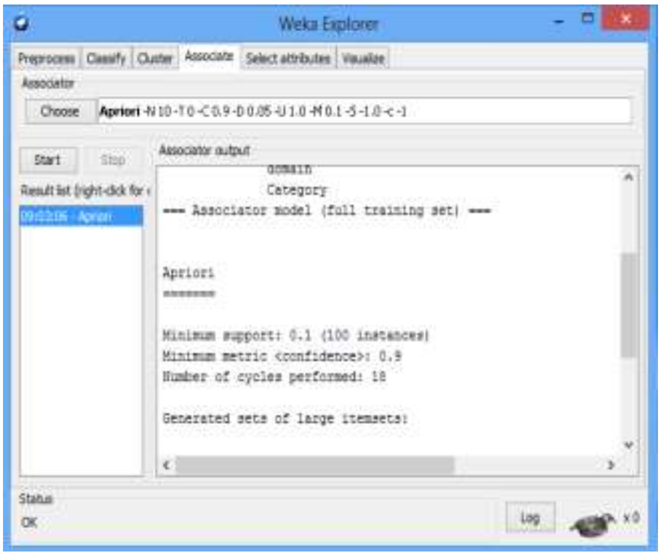
No.		Name
1	□	Time
2	□	User Group
3	□	domain
4	□	Category

Class: Category (Nom)

Status
OK

Fig 1: Use the Data in .csv format

In Fig 1 we use the data in .csv format after this weka tool show many result according to the data property and also show in the GUI.



Associator

Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

09:02:06 - Apriori

Associator output

```
              domain
              Category
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (100 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:
```

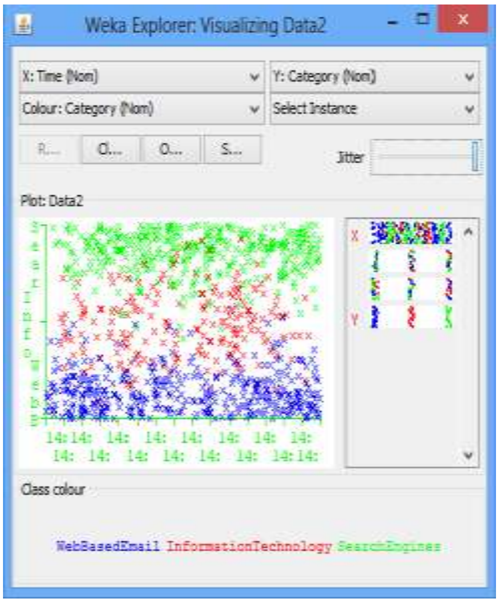
Status
OK

Fig 2: Click on Associate choose Apriori

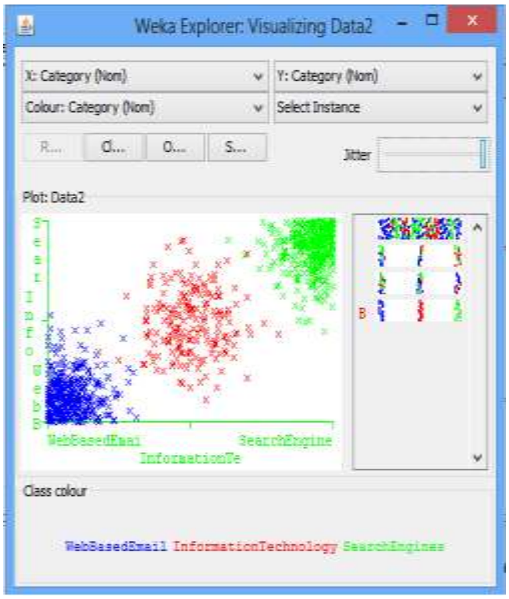
In Fig 2 we apply the apriori algo on the data set and show the result.

Visualize graph of data set which is given below.

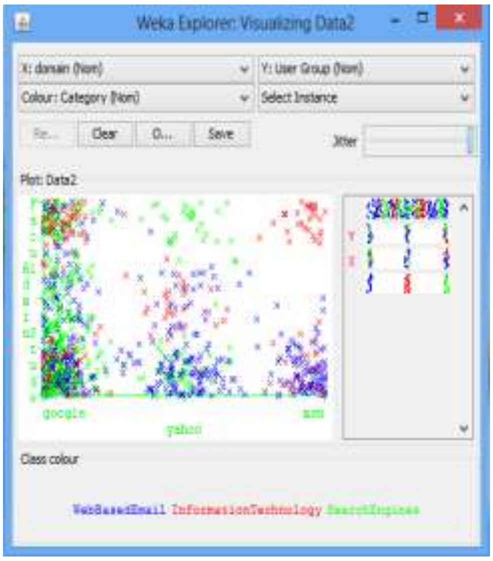
Fig: 3,4,5,6 Show the data accessibility.


Weka Explorer: Visualizing Data2
R...
Cl...
O...
S...
Jitter
Plot: Data2
14: 14: 14: 14: 14: 14: 14: 14:
X
Y
Class colour
WebBasedEmail InformationTechnology SearchEngines

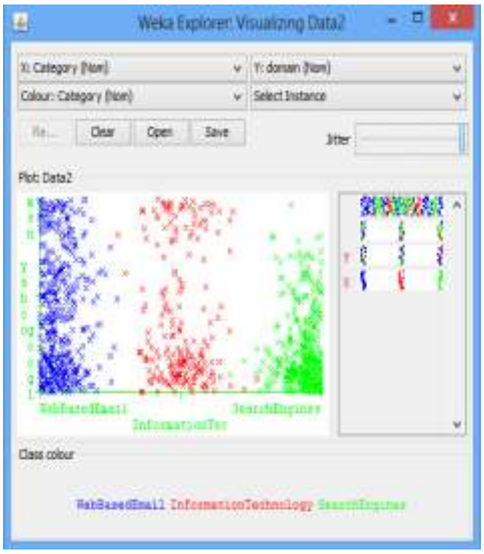
X: Time (Nom)
Y: Category (Nom)
Colour: Category (Nom)
Select Instance



X: Category (Nom)
Y: Category (Nom)
Colour: Category (Nom)
Select Instance


Weka Explorer: Visualizing Data2
Re...
Clear
O...
Save
Jitter
Plot: Data2
google
yahoo
msn
Y
X
Class colour
WebBasedEmail InformationTechnology SearchEngines

X: domain (Nom)
Y: User Group (Nom)
Colour: Category (Nom)
Select Instance



X: Category (Nom)
Y: domain (Nom)
Colour: Category (Nom)
Select Instance

In the above graphs the values of the data is according to the access pattern of the user and show by the different color and these graph according to the attribute. the blue color data set show the more frequent data set and other color show the less frequent data set. According to the above result we say that Web based email is more access by the student and faculty.

VII. CONCLUSION

Here we find the useful information about the user.' accesses is obtained from analysis of navigation behavior from the web logs, where all accesses to web pages are recorded. This paper adopted an efficient sequential pattern mining techniques using the Apriori, algorithm for the filtered data set.the algo help to find the user access behavior on the previous visits and also shows the comparison of the techniques adopted for predicting user access behavior.

REFERENCES

[1] M.Bharati, M. Ramageri "Data Mining Techniques And Applications" Indian Journal of Computer Science and Engineering 2008, Vol. 1 No. 4.pp. 301-305.

[2] S. K. Pani, "Web Usage Mining: A Survey on Pattern Extraction from Web Logs", International Journal of Instrumentation, Control & Automation (IJICA), Vol. 1, Issue 1, 2011.

[3] M.J.Zaki, S.Parthasarathy, and Wei Li: "Visual Interface for Online Watching of Frequent Itemset Generation in Apriori and Eclat" Machine Learning and Applications Proceedings. Fourth International Conference on 15-17 Dec. 2005.

[4] R.Srikant and R. Agrawal"Mining Sequential Patterns: Generalizations and Performance Improvements" IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120.

[5] Han & Kamber: "Data Mining: Concepts and Techniques" 2006.

[7] B.S. Kuma, K.V.Rukmani :"Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms Web" Int. J. of Advanced Networking and Applications Vol.01, Issue 06, pp.400-404,2012.

[8] S.Rao, Priyanka Gupta: "Implementing Improved Algorithm over APRIORI Data Mining Association Rule Algorithm" International Journal of Computer Science and technology Vol. 3, Issue 1, 489, Jan. - March 2012.

[9] M. Hall: "The WEKA Data Mining Software: An Update" SIGKDD Explorations Vol. 11, Issue 1 pp. 11, 2006.

[10] Ankit Kumar, Bhasker Pant," Predicting User Behavior and Comparison in Sequential Data Mining Techniques" IJCST Vol. 3, Iss ue 4, Oct - Dec 2012

[11] WEKA available at http://www.cs.waikato.ac.nz/ml/weka/arff.html