International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 04, April-2015

Web usage Mining:Frequent Pattern Generation using Association Rule Mining and Clustering

Aarti M. Parekh[#1] , Anjali S. Patel[#2] , Sonal J. Parmar[#3],Prof.Vaishali R. Patel[#4]
Department of Information Technology
Shri S'ad Vidya Mandal Institute of Technology
Bharuch 392-001, Gujarat, India

***Abstract*— Analyzing the web log files through web usage mining is very important to discover the similar behavior users of particular website. Our paper discusses how to find useful knowledge from web log file using some data mining technique like Association rule mining and clustering. First we preprocess the web log file then apply association rule mining and clustering algorithm on web log file to discover usage pattern and same behavioral users.**

Keywords—Web usage mining, Web log files, Clustering , Association rule mining.

I. INTRODUCTION

Web mining is one of the application of data mining which is used to retrieve the useful information or knowledge from web data. Web mining is divided into 3 categories:

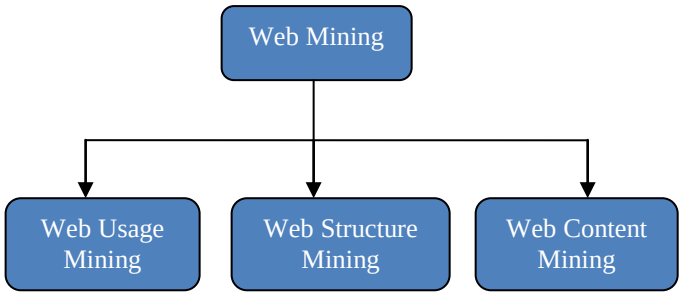

Fig 1: Categories of Web Mining

A. Web Usage Mining

Web Usage Mining is also known as web log mining which is used to discover the useful pattern from web log file. Web server log files is a primary data source of web usage mining. To understand the user needs and behavior is discover by analyzing web log file which is one type of textual file created by server automatically when user makes transaction on particular website [10]. The example of log file is given below:

```
213.135.131.79 - - [15/May/2002:19:21:49 -0400]
"GET /features.htm HTTP/1.1" 200 9955
```

B. Web Structure Mining

Web Structure Mining refers to mining the hyperlink structure of website. Hyperlink is one of the component which connects webpage to different location. To analyze the website structure we use graph theory. To analyze the HTML tags from web pages uses tree like structure to mine the structure of particular website[1].

C. Web Content Mining

Web Content Mining is also known as web text mining because it discover the useful information from audio, video, text, images in the website. Natural language processing and information retrieval technology are used to mine the content of website.

II. WEB USAGE MINING PROCESS

Web Usage Mining process is categorized into 3 phases: Preprocessing, Pattern Discovery and Pattern Analysis which is shown below:

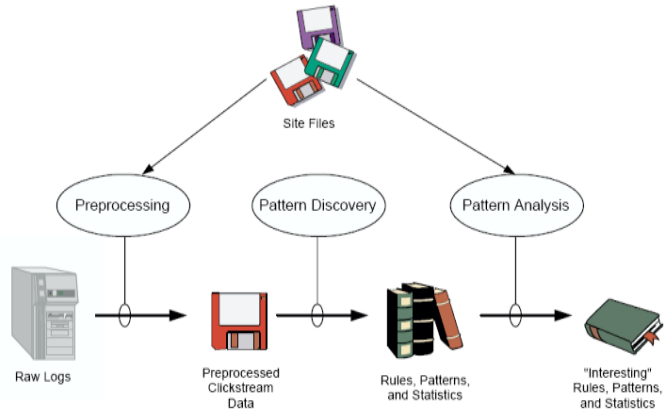

Fig 2: Process of Web Usage Mining

A. Preprocessing

Real world data may be noisy or inconsistent so we have to preprocess them to make them consistent and reliable. So preprocessing phase is very important step of web usage mining [2].

1. Cleaning

Data Cleaning refers to remove irrelevant entries from web log file. Remove the entries which has status code less than 200 and greater than 400.There are some redundant data to be removed like additional request and error entries.

2. User Identification

User identification refers to identify unique users. Users with different ip address are consider as unique users. It is very important to mine the users access characteristics.

3. Session Identification

Session identification refers to differentiate the web log entries into different user sessions by a session timeout. We have used 20 minute timeout for session's timeout property [9].

B. Pattern Discovery

In Pattern Discovery phase, data mining techniques like association rule mining and clustering applied on web log files after preprocessing to discover the useful pattern.

1. Association rule mining

Association rule mining problem was specified by Agrawal [3]. Association rule mining is one of the data mining technique which is used to discover useful pattern. It works on generating frequent pattern and rules. In web log file number of URL visit by number of users so we can identify frequently accessed web pages by users which can help to understand user needs. Two basic parameters of association rule are support and confidence.

$$\text{Support (XY)} = \frac{\text{Support sum of XY}}{\text{Overall records in the database D}}$$

$$Confidence(X/Y) = \frac{Support\ (XY)}{Support\ (X)}$$

2. Clustering

Clustering is unsupervised learning technique. Clustering analysis defined as similar characteristics users are group together without knowledge of group defination. Clustering will help us to find group of common behavior users. Clustering of webpages are very important for internet service provider to analyze the behavior of users [5]. Many clustering algorithms have been developed and are categorized such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods.

C. Pattern Analysis

In Pattern Analysis phase, irrelevant pattern are remove from the pattern which identified during pattern discovery phase. The main purpose of pattern analysis is to analyze the pattern which is identified during pattern discovery phase.

III. PROPOSED SYSTEM

We would like to propose a system which discover the useful pattern from web server log file. In the case of web transactions, association rules finds the relationships among page views based on the navigation patterns of users. So we implement the apriori algorithm on the web log files which gives frequently accessed webpages and unique users. Then we apply clustering k-means algorithm on web log file so we can predict better result. Our proposed approach is shown below:

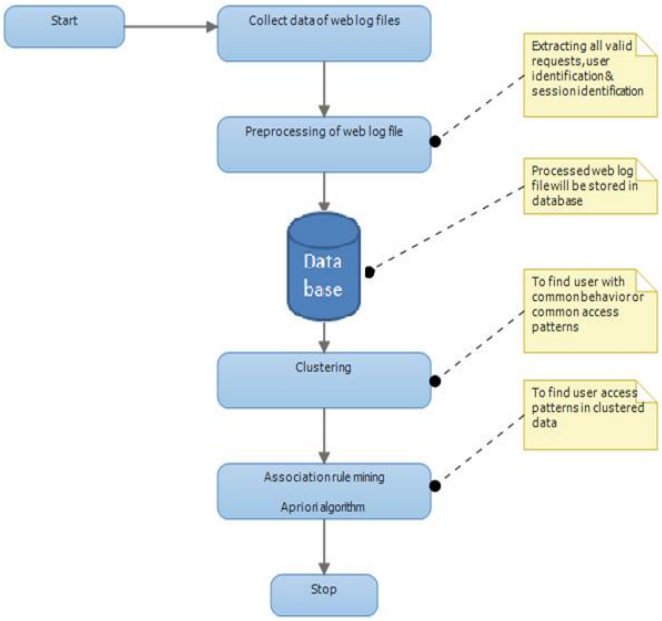

Fig 3 : Proposed approach for our system

A. Apriori algorithm

The Apriori algorithm is an effective algorithm for finding all frequent item sets from web log files [8]. Apriori works in iterative approach known as a level-wise search, where k-itemsets are used to find (k+1) itemsets. First of all frequent 1-itemsets is found. This is defined as L1. L1 is used to find L2, the frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found.

This algorithm level-wise searching using frequent web pages. It handles the web log which contains large amount of transactions in it. Apriori algorithm is useful for identifying the web pages viewed by each unique user. The algorithm for apriori is given below:

```
Apriorialgo()
{
F = ∅;
Lk = {frequent 1-itemsets};
k = 2; /* k represents the pass number. */
while (Lk-1 != ∅){
F = F ∪ Lk ;
Ck = New candidates of size k generated
from Lk-1 ;
for all transactions t ∈ D
increment the count of all candidates in Ck
that are contained in t ;
Lk = All candidates in Ck with minimum
support ;
k++ ;
}return ( F ) ;
}
```

B. Modified K-Means algorithm

K-means is a clustering algorithm in data mining. It was one of the simple and un-supervised learning algorithms. It is a partitioning clustering algorithm. The original k-means algorithm consists of two limitation that we have to define the k clusters and there may be empty cluster when we take large value of k because of problem of initial centroid [4]. Instead of choosing initial centroid randomly, Proposed algorithm determines the initial centroid. So we remove this two limitation in our modified k-Means. The algorithm for modified K-Means is shown below:

Input:

D = {d1, d2,.....,d} //set of n data items.

Output:

A set of k clusters.

Steps:

1) Determine the value of K using following formula.

$$k \approx \sqrt{n/2}$$

Where n is number of objects.

2) Here we have to select k data objects from dataset D as initial cluster centers as follow:

- For each column of the data set, determine the range as the variation between the maximum and the minimum element;
- Identify the column with maximum range;
- Sort the entire data set increasing order based on the column having the maximum range;
- The sorted data set are partitioned into k equal parts;
- Determine the arithmetic mean of each part obtained in Step 4 as a1, a2,….ak; Take these mean values as the initial centroids.

3) Then calculate the distance between each data object di (1 <= i<=n) and all k cluster centers aj(1<=j<=k).

4) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

5) For each cluster, recalculate the cluster center.

6) Until no changing in the center of clusters

IV. IMPLEMENTATION RESULTS

Web Usage Mining is implemented on sample web server log files as input. Then apply preprocessing on web log file and store into the database. We can generate useful pattern from web log file by association rule mining and clustering algorithm. The following figure shows step wise implementation:

Step 1: Raw web log files are choose from where it is stored.

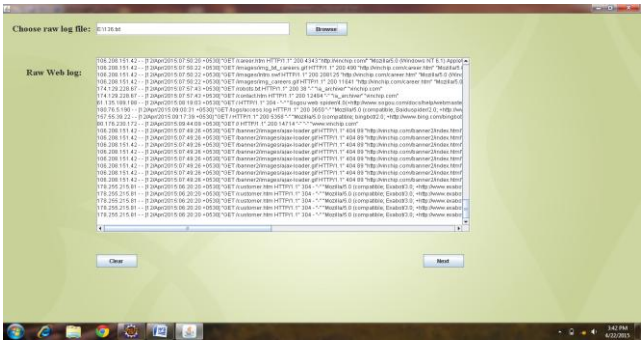

Fig 4: Choose raw web log file

Step 2: Apply the preprocessing on web log files and store them into the database.

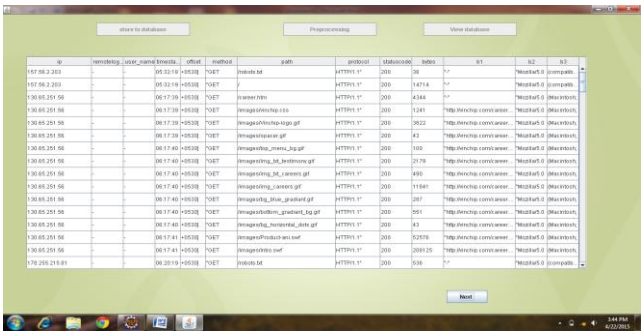

Fig 5: Raw web log file after preprocessing

Step 3: Unique users and webpages are identified from web log files.

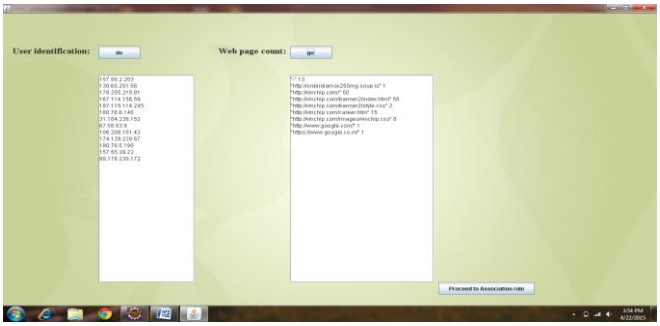

Fig 6: Unique users and web pages

Step 4: Apriori and k-means clustering algorithm is applied on web log files and get frequently accessed webpages.

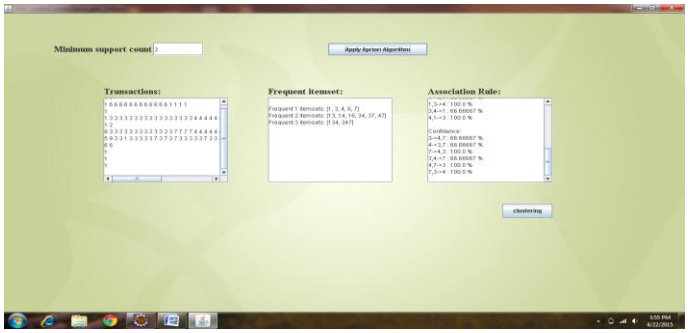

Fig 7: Frequent itemset generation using apriori algorithm

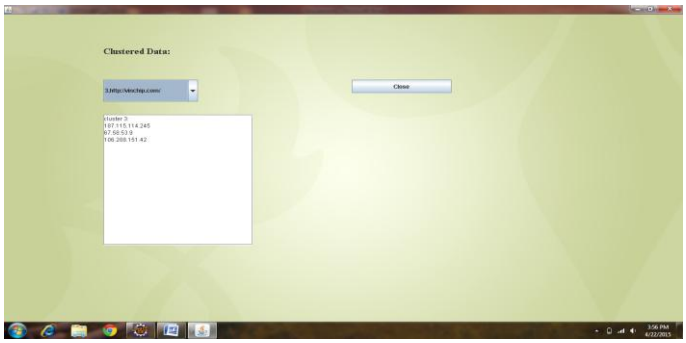

Fig 8: Clustered web pages

V. CONCLUSION

Web Usage Mining is a great research area in these days. In this paper, implementation of a system for pattern discovery using association rules and clustering is to introduce the process of web log mining, and to show how to find frequent pattern from the web log data in order to obtain useful information about the user's navigation behavior.The approach used in this paper, helps the website designers to improve their website usability.

REFERENCES

[1] Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery", International Journal of Data Mining Techniques and Applications, Vol 02, Issue 01, June 2013

[2] 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

[3] Ms Kiruthika M, Mr Rahul Jadhav,Ms Rashmi J ,Ms Dipa Dixit, "Pattern Discovery Using Association Rules", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 12, 2011

[4] M.Vijayalakshmi,MCA,M.Phil M.Renuka Devi,MCA,M.Phil,(Phd), "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012

[5] Deepti Sisodia, Sheetal Sisodia, Lokesh Singh, "Clustering Techniques: A Brief Survey of Different Clustering Algorithms" International Journal of Latest Trends in Engineering and Technology (IJLTET)

[6] Rupinder Kaur1, Simarjeet Kaur2, "A Review: Techniques for Clustering of Web Usage Mining", International Journal of Science and Research (IJSR)

[7] T. Karthikeyan and N. Ravikumar, "A Survey on Association Rule Mining" , International Journal of Advanced Research in Computer and Communication Engineering,Vol. 3, Issue 1, January 2014

[8] Nilesh Prajapati, Premal Patel, Ankit R Kharwar1 Viral Kapadia, "Implementing APRIORI Algorithm on Web serve log" ,National Conference on Recent Trends in Engineering & Technology

[9] Dr. Sanjeev Dhawan, Mamta Lathwal, "Study of Preprocessing Methods in Web Server Logs", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013

[10] C.P.Sumathi,R. Padmaja Valli and T. Santhanam, "an overview of preprocessing of web log files for web usage mining". Journal of Theoretical and Applied Information Technology, Vol. 34 No.2, December 2011

[11] Shaily Langhnoja1, Mehul Barot2, Darshak Mehta3, "Pre-Processing: Procedure on Web Log FileforWeb Usage Mining", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 12, December 2012.