A Web Based Recommendation Using Association Rule and Clustering

Vidhu Singhal[#1], Gopal Pandey[*2]

[#]*ME Scholar, Department of Information Technology*
Shantilal Shah Engineering College, Gujarat Technological University, Bhavnagar, Gujarat, India
[1]vidhu04@yahoo.com

[*]*Incharge HOD of Information Technology Department*
Sir Bhavsinhji Polytechnic Institute, Bhavnagar, Gujarat, India
[2]mr.gopal.pandey@gmail.com

Abstract: **In the era of intense competition amongst organizations, it's a real world challenging task to complement the need of user and maintain their interest in their web site. That is why the problem of predicting a user's behavior on a web-site has gained importance. A web based recommendation system include browsing history database enclosed with information related to the web pages that a user browsed. This paper presents the Prediction of User navigation patterns of WUM using Association Rule and Clustering from web log data. In the first stage, separating the potential users is processed, and in the second stage clustering process is used to group the users with similar interest, and in the third stage association and clustering is used to navigate the user future requests. The experimental results are really encouraging and produce valuable information.**

Keywords**- Web Usage Mining, Web Log File, Association Rule, Clustering, Weka Tool**

I. INTRODUCTION

Data mining research interest is the consequence of the immense amount of data that forms part of our daily activities and is the process of analyzing data from different perspectives. Web data mining involves applying various data mining techniques.[1] It usually focuses on the Web pages link structure, their content and their usage. Web mining is categorized in three forms: Web Usage mining, Web Content and Web Structure mining. The main focus of this paper is Web usage mining. WUM concentrates on tools and techniques used to predict users' navigational paths by discovering their Web access patterns. It includes three stages: preprocessing, pattern discovery and pattern analysis.[1][2]

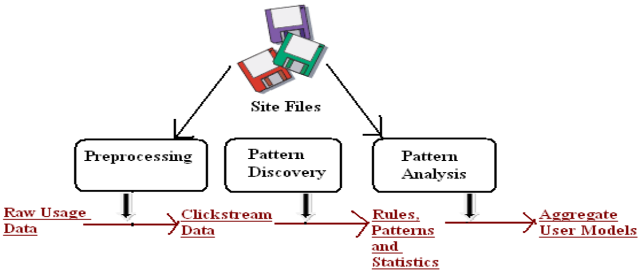


Figure 1: Web Usage Mining Process

Since the knowledge of the user's browsing history on the site gives the valuable information as to which one of the most frequently accessed pages will be accessed next.

Pattern discovery techniques achieve valuable information by extracting useful knowledge and patterns applying different tools and techniques. Some of these tools are association rules and clustering. Each pattern discovery techniques has its own strong points and limitations. Discovered patterns of accessed Web pages help to envisage the next page to be navigated by the user.

The main motivation behind this paper is the correlation between Web usage mining and Web personalisation. The work of Web usage mining is interpreting patterns of user access to web information systems by mining the data collected from user interactions with the system. The ultimate goal of Web personalisation is to provide Web users with the next page they will access in a browsing session. [3] This achieved by evaluating their browsing patterns and comparing the discovered patterns to similar patterns in history.

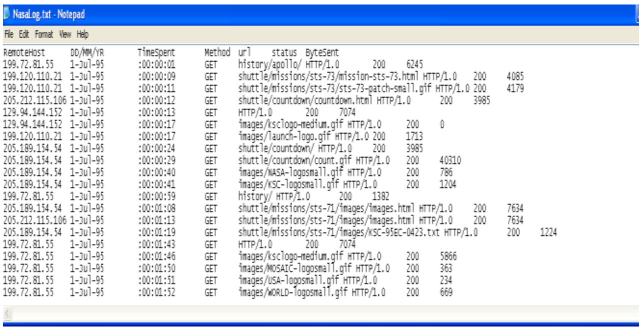
Traditionally, this has been used to support the decision making process by Web site operators in order to add better understanding of their visitors and to create a more efficient structure of the Web sites.

II. WEB LOGS AND WEB USAGE MINING

1) Web Logs- Log files are a standard tool for computer systems developers and administrators. They record the, "what happened when by whom" of the system. Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. [4] The log files are maintained by the web servers. By analyzing these log files gives a neat idea about the user. A Web log is a file to which the Web server writes information each time a user requests a website from that particular server. A log file can be located in three different places:

- Web Servers
- Web proxy Servers
- Client browsers

The most popular log file formats are the Common Log Format (CLF) and the extended CLF. The following is an example form of the server log file which gives significant information: (NASA_access_log_jul95).



```
RemoteHost      DD/MM/YR        TimeSpent       Method  url     status  ByteSent
199.72.81.55    1-Jul-95        :00:00:01       GET     history/apollo/ HTTP/1.0        200     6245
199.120.110.21  1-Jul-95        :00:00:09       GET     shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0    200     4085
199.120.110.21  1-Jul-95        :00:00:11       GET     shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/1.0 200     4179
205.212.115.106 1-Jul-95        :00:00:12       GET     shuttle/countdown/countdown.html HTTP/1.0       200     3985
129.94.144.152  1-Jul-95        :00:00:13       GET     HTTP/1.0        200     7074
129.94.144.152  1-Jul-95        :00:00:17       GET     images/ksclogo-medium.gif HTTP/1.0      200     0
199.120.110.21  1-Jul-95        :00:00:17       GET     images/launch-logo.gif HTTP/1.0 200     1713
205.189.154.54  1-Jul-95        :00:00:24       GET     shuttle/countdown/ HTTP/1.0     200     3985
205.189.154.54  1-Jul-95        :00:00:29       GET     shuttle/countdown/count.gif HTTP/1.0    200     40310
205.189.154.54  1-Jul-95        :00:00:40       GET     images/NASA-logosmall.gif HTTP/1.0      200     786
205.189.154.54  1-Jul-95        :00:00:41       GET     images/KSC-logosmall.gif HTTP/1.0       200     1204
199.72.81.55    1-Jul-95        :00:00:59       GET     history/ HTTP/1.0       200     1382
205.189.154.54  1-Jul-95        :00:01:08       GET     shuttle/missions/sts-71/images/images.html HTTP/1.0     200     7634
205.212.115.106 1-Jul-95        :00:01:13       GET     shuttle/missions/sts-71/images/images.html HTTP/1.0     200     7634
205.189.154.54  1-Jul-95        :00:01:19       GET     shuttle/missions/sts-71/images/KSC-95EC-0423.txt HTTP/1.0       200     1224
199.72.81.55    1-Jul-95        :00:01:43       GET     HTTP/1.0        200     7074
199.72.81.55    1-Jul-95        :00:01:46       GET     images/ksclogo-medium.gif HTTP/1.0      200     5866
199.72.81.55    1-Jul-95        :00:01:50       GET     images/MOSAIC-logosmall.gif HTTP/1.0    200     363
199.72.81.55    1-Jul-95        :00:01:51       GET     images/USA-logosmall.gif HTTP/1.0       200     234
199.72.81.55    1-Jul-95        :00:01:52       GET     images/WORLD-logosmall.gif HTTP/1.0     200     669
```

Figure 2: A text file with logs

2) Fundamental of Web Usage Mining-Web usage mining also acknowledged as web log mining in various fields. It is the application for an easier facility website design tasks. Web servers collect large volumes of data from Web sites usage.
This data is stored in Web access log files. Together with the Web access log files, other data can be used in Web Usage Mining like the Web structure information, user profiles, Web page content, etc. WUM is classified into three main steps: preprocessing, pattern discovery and pattern analysis. [2][5]

2.2.1 Pre-processing

Before starting any mining technique, Web data has to be cleaned and pre-processed. Preprocessing prepares data for the pattern discovery stage. It transforms Web log files into Web transaction data that can be processed by data mining tasks. Web data could take many forms. The primary data sources are the server log files that include Web server access logs and application server logs. Also, additional data sources may include operational databases, domain knowledge, site files and meta-data. This additional data can be available from client-side or proxy level data collection as well as from external clickstream or demographic data sources. The most important and the most easily accessed data is the Web server log report that keeps track of every single user access to the server. In general, the log entries include information like date, time, client IP, URL of the source, name of the script or file requested and the server status. There are three types of preprocessing: usage preprocessing, content preprocessing and structure preprocessing.

a) Usage preprocessing: is the most difficult task as it deals with the incomplete log entries and the wide usage of local caches and proxy servers. Often there is a need for using more accurate data from other sources like cookies or a client side collection method. With usage preprocessing, the data usually needs to be transformed and aggregated at different levels of abstraction. The most basic level of data abstraction is pageview which represents a collection of Web objects displayed as a result of a single user action. A collection of pageview for a single user during a single visit forms a session. Sessions may be used to analyse the user's behavioural browsing patterns.

b) Content preprocessing: involves preparing text and multimedia files using classification and clustering techniques. Static Web pages can be easily pre-processed by parsing the HTML and reformatting the information. However, dynamic Web pages that are the result of database accesses or personalisation algorithms are usually more difficult to pre-process.

c) Structure preprocessing: consists of preprocessing the inter-page structure information or the Hyperlinks that connect one page to another. Again, pages that have a predefined structure are easily pre-processed. However, dynamically structured pages can be more difficult. Dynamic structure sometimes creates problems since a different site structure may have to be build up for each server session.

2.2.2 Pattern Discovery

In this, WUM can be able to unearth patterns in server logs and carried out only on samples of data. Interpretation and evaluation of results be done on samples of data. The various pattern discovery methods are— Statistical Analysis, Association Rules, Clustering, Classification, Sequential Patterns, and Dependency Modeling. During this stage, algorithms are run on the data and patterns are extracted from it. Pattern discovery involves the employment of sophisticated techniques from artificial intelligence, data mining techniques, psychology and information theory in order to extract knowledge from collected and pre-processed data.

2.2.3 Pattern Analysis

The main motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase and log files. Content and structure information can be employ to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure. Pattern analysis involves removing out the unneeded patterns or rules discovered through the pattern discovery phase. The most common pattern analysis technique is the use of query language such as SQL.

III. WORKING SCHEME FOR WEB PAGE PREDICTION

Personalising the Web users' content and recommending appropriate Web pages imply that users achieve with what they require based on their previous interactions within the same Web site. This task is viewed as a prediction task since it predict the users' level of interest in specific pages and rank these pages according to their predicted values.
The working scheme of this paper involves pattern discovery techniques of web usage mining to make predictions: Association Rule and Clustering

1) Association Rules-Association rule mining are one of the major techniques of data mining and it is the most common form of local-pattern discovery in unsupervised learning systems.[10][11] It serves as a useful tool for finding correlations between items in large databases. In Web mining context, association rules help optimize the organisation and structure of Web sites. The terms used in these rule are:

Support: The support supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.
supp(X) = no. of transactions which contain the itemset X / total no. of transactions

Confidence: The measurement of certainty coupled with each and every discovered pattern. The confidence for an association rule X implies Y is the ratio of the number of transaction that contains X U Y to the number of transaction that contains X.
con f (X->Y) = supp (XUY) / supp(X)

Large Item Set: A large item set is an item set whose number of occurrences is above a threshold or support.
The minimum support requirement dictates the efficiency of association rule mining. One major motivation for using the support factor comes from the fact that we are usually interested only in rules with certain popularity. Support corresponds to statistical significance while confidence is a measure of the rules strength. Confidence represents the conditional probability that item *pa* occurs in a transaction given that item *pb* occurred in the same transaction. Support and confidence are the most commonly used metrics when using association based approaches to personalisation.

The most common approach to finding association rules is to break up the problem into two parts:
1. Finding every occurred frequent itemsets.
2. Generating strong association rules commencing the frequent itemsets satisfying minimum support and minimum confidence.

2) *Clustering*-The chief stimulus behind the use of clustering as a model-based pattern discovery algorithm in Web usage mining stage of Web mining is to improve the efficiency and scalability of the real-time personalisation tasks. Mainly clustering aims at dividing the data set into groups (clusters) where the inter-cluster similarities are minimised while the similarities within each cluster are maximized.[8] Clustering Web sessions can be achieved through page clustering or user clustering. Web page clustering is performed by grouping pages having similar content.
In this, K-means algorithm using session based similarity measures for clustering is used for the user session into k-clusters. If two users accessed the same pages in sessions, they might have some similar interests in the sense that they are interested in the same information. The number of common pages they accessed can measure this similarity. The measure is defined by using the equation mentioned:

$$Sim1(pi,sj) = \Sigma_k(use(p_{k,}s_j)*use((p_{k,}s_j)) / \sqrt{\sum_k(use(p_{k,}s_j)*\sum_k use((p_{k,}s_j))}$$

Where Σ_k (use($p_{k,}s_j$) is the total number of pages that were accessed by the user of session s and$\sqrt{\sum_k(use(p_{k,}s_j)*\sum_k use((p_{k,}s_j))}$ is the number of common pages accessed. By Assuming k=2 and two of the user sessions as centre. By using the similarity matrix we cluster the user sessions which are highly similar to each other. So by using the clustering better prediction is performed as we have to apply prediction algorithm to a specific cluster rather than whole data set.

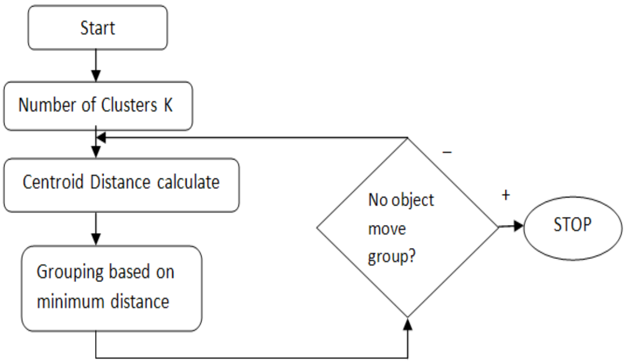

Start
Number of Clusters K
Centroid Distance calculate
Grouping based on minimum distance
No object move group?
-
+
STOP

Figure 3: K-means clustering algorithm flowchart

IV. EXPERIMENTAL SETUP AND RESULT

Experiments are run on a system with a 1.60 GHz Intel(R) Core(TM) and 2 GB RAM running Windows XP. The web access log was collected from the web server at NASA Kennedy space Centre form 00:00:00 Aug 1, 1995 through 23:59:59 Aug 31, 1995, a total of 31 days. In this period there are totally 1,891,700 requests recorded by the log file. Our implementation is developed using WEKA tool.
"*Waikato Environment for Knowledge Analysis*"[9] is a collection of machine learning algorithms for data mining and it's an open Source Machine Learning Software in Java. The WEKA data mining tool requires that the input data set should be represented in the ARFF/CSV format.

1) Data Preprocessing-Before using the log files data, it is vital to perform data preprocessing. Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step. Pre-processing may include cleaning data of inconsistencies, filtering out irrelevant information according to the goal of analysis.

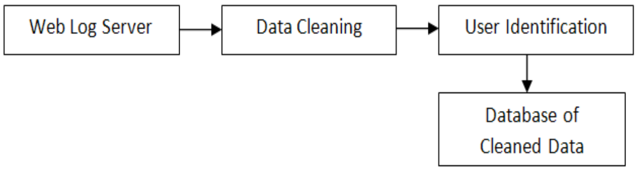


Figure 4: Overview of Data preprocessing

Table 1: Summary of Log Data

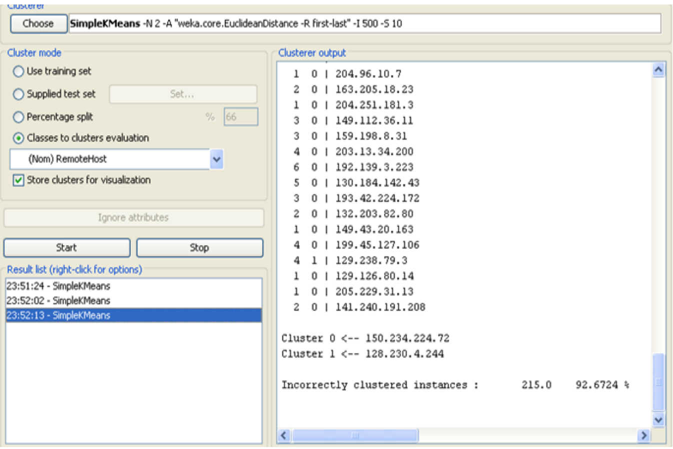
Hits	
Total Hits	1,891,700
Visitor Hits	1,891,700
Spider Hits	0
Average Hits per Day	65,231
Average Hits per Visitor	11.62
Cached Requests	132,627
Failed Requests	10,966
Page Views	
Total Page Views	613,118
Average Page Views per Day	21,142
Average Page Views per Visitor	3.77
Visitors	

Total Visitors	162,754
Total Unique IPs	81,982
Bandwidth	
Total Bandwidth	36.04 GB
Average Bandwidth per Visitor	232.18 KB

Table 2: Results of Preprocessed data

Server Log File	NASA Jul-95
Duration	1 - 7days
Original Size	50.84MB
Reduced Size After Preprocessing	13.07MB
Percentage in Reduction	70.47
Total No. of Unique Users	8575

2) Clustering-There are several clustering algorithms available in WEKA. These are all part of the WEKA.CLUSTERS package. [8] In this K - means clustering is used, which is defined by the class SimpleKMeans. It clusters data using k-means; the number of clusters is specified by a parameter. The user can choose between the Euclidean and Manhattan distance metrics.



Clusterer
Choose SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode
○ Use training set
○ Supplied test set
○ Percentage split % 66
● Classes to clusters evaluation
(Nom) RemoteHost
☑ Store clusters for visualization

Result list (right-click for options)
23:51:24 - SimpleKMeans
23:52:02 - SimpleKMeans
23:52:13 - SimpleKMeans

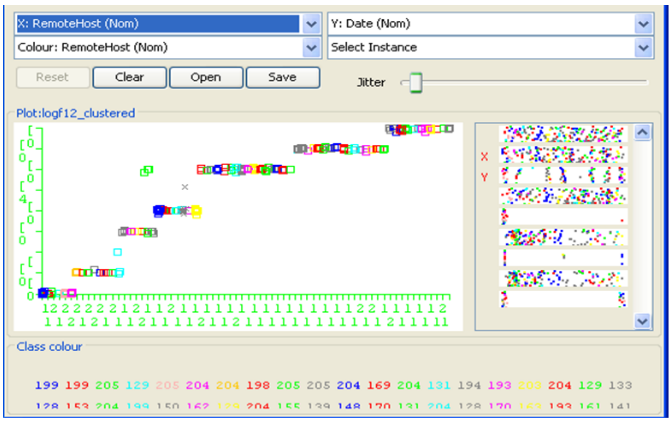
Clusterer output

```
  1  0 | 204.96.10.7
  2  0 | 163.205.18.23
  1  0 | 204.251.181.3
  3  0 | 149.112.36.11
  3  0 | 159.198.8.31
  4  0 | 203.13.34.200
  6  0 | 192.139.3.223
  5  0 | 130.184.142.43
  3  0 | 193.42.224.172
  2  0 | 132.203.82.80
  1  0 | 149.43.20.163
  4  0 | 199.45.127.106
  4  1 | 129.238.79.3
  1  0 | 129.126.80.14
  1  0 | 205.229.31.13
  2  0 | 141.240.191.208

Cluster 0 <-- 150.234.224.72
Cluster 1 <-- 128.230.4.244

Incorrectly clustered instances :       215.0     92.6724 %
```

Figure 5: Cluster evaluation by attribute: Remote Host

The above snapshot shows a summary of the processing of the k-means algorithm on the loaded data. It follows the training set and test set process. Clustering involves partitioning pages or sessions into similar groups. Prediction takes place based on these groups. It is distance based, unsupervised and partitional.

K-means clustering algorithm is the simplest and most commonly used clustering algorithm, especially with large data sets. It involves:
1. Define a set of items (n-by-p data matrix) to be clustered.
2. Define a chosen number of clusters (k).
3. Randomly assign a number of items to each cluster.



X: RemoteHost (Nom)
Y: Date (Nom)
Colour: RemoteHost (Nom)
Select Instance

Figure 6: Graph X: Remote Host Y: Date

Figure 6 shows the clusters according to the remote host traversed frequently by the user with respect to date. A simple way to interpret this data is to consider the centroids as representative values for each cluster. Proper grouping and clustering of Web sessions helps increase the Web page access prediction accuracy. Thus, this will reduce the ambiguity

3) Association Rules

Relation: NASA Log
Instances: 67
Attributes: 7
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.15 (2 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level-wise full search algorithm using anti-monotonicity of itemsets, "*if an itemset is not frequent, any of its superset is never frequent*". By principle, Apriori assume that items contained by transaction or itemset are sorted in lexicographic strategy. After evaluating it shows that the user's access history can be used to use for mining user access patterns.

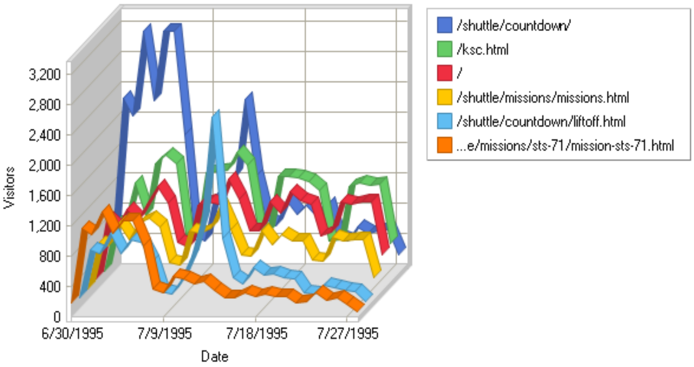


Figure 7: Graph of frequently visited page by user

V CONCLUSION

The main objective of the paper is to help achieve better prediction accuracy for Web page access. Recommending a next page, the Web user will access is very vital for diverse Web applications.

Web page access prediction is addressed by many literature publications. The main technology implemented for this purpose is through using Web usage mining pattern discovery techniques. By keeping the models limitations to a minimum and relying on their advantages according to different constraints, it become possible to achieve more accurate prediction results.

ACKNOWLEDGMENT

Author desire to show gratitude to guide Prof.Gopal Pandey for giving their valuable time and resources that helped me to implement research work. This research was carried out as part of the activities of, dissertation work in Post-Graduation.

REFERENCES

[1] Jiawei Han, Micheline Kamber, *Data mining concepts and techniques*, Elsevier Inc., Second Edition, San Francisco, 2006

[2] Jaideep Srivastava, Robert Cooleyz, Mukund Deshpande, Pang-Ning Tan proposed *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*, 2000.

[3] Yogita S. Pagar, Vishakha. R. Mote, Rahul S. Bramhane, *Web Personalization using Web Mining Technique*, Emerging Trends in Computer Science and Information Technol2012 (ETCSIT2012)

[4] Thanakorn Pamutha, Siriporn Chimphlee, Chom Kimpan1 and Parinya Sanguansat, *Data Preprocessing on Web Server Log Files for Mining Users Access Patterns*, International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol. 2, No. 2, June 2012, ISSN: 2046-6447

[5] FENG ZHANG,HUI-YOU CHANG, *Research And Development In Web Usage Mining System--Key Issues And Proposed Solutions: A Survey*, Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002

[6] Mehrdad Jalali, Norwati Mustapha, *WebPUM: A Web-based recommendation system to predict user future movements*, Expert Systems with Applications 37 (2010) 6201–6212

[7] K. R. Suneetha, Dr. R. Krishnamoorthi *Identifying User Behavior by Analyzing Web Server Access Log File,* IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009

[8] Narendra Sharma , Aman Bajpai , Mr. Ratnesh Litoriya *Comparison the various clustering algorithms of weka tools,* International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, Volume 2, Issue 5, May 2012)

[9] Dr. Wenjia Wang *Tutorial for DM tool Weka,* CMP: Data Mining and Statistics within the Health Services 19/02/2010

[10] R. Suguna, D. Sharmila *Association Rule Mining for Web Recommendation*, R. Suguna et al. / International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 4 No. 10 Oct 2012

[11] S. K. Pani, L. Panigrahy, and V.H.Sankar *Web Usage Mining: A Survey on Pattern Extraction from Web Logs,* International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011