Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 3 (2015) pp. 159-167

Discovering Web Server Logs Patterns Using Clustering and Association Rules Mining

Mohammed Hamed Ahmed Elhebir[1] and Ajith Abraham [2]

[1] Faculty of Mathematical and Computer Sciences, University of Gezira,
P.O. Box 20, Wad Medani, Sudan
elhibr@uofg.edu.sd

[2] Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and
Research Excellence, P.O. Box 2259, WA, USA
ajith.abraham@ieee.org

Abstract: **Recording server log data files are nowadays a commonplace practice. The server log data files capture useful information during the interaction of users with the online site, as well as the interaction among users during online sessions. The discovered patterns are used for improving web site organization and behavior. Clustering is a discovery process in data mining, which groups set of data items, in such a way that maximizes the similarity within clusters and minimizes the similarity between two different clusters. Various forms of clustering are used in a wide range of applications. In this paper, the server log files of the Website " www.sust.edu" is considered for overall study and analysis. This paper presents the discovering patterns of Web Usage Mining (WUM) using Clustering and Association Rule from web log data. In the first stage, the data pre-processing phase was performed, and in the second stage k-means and density-based clustering algorithms are used for clustering the log file into groups. Then these clusters are plotted on a plane using WEKA tool, where different clusters are distinguished by distinct colors and distinct symbols. Performance and accuracy of the clustering algorithms are presented and compared. The results of the experiment show that clustering with feature selection gives better Performance. In the third stage apriori algorithm is used to discover relationship among data. Eliminating redundant rules and clustering decreased the size of the generated rule set to obtain Interestingness rules.**

Keywords: Web Usage Mining, Pattern Discovery, Clustering and Association Rule Mining.

I. Introduction

Data mining is a process of knowledge discovery used to reveal relationships and patterns in large amount data. Also it is referred to as a process of extracting/mining of knowledge from large amounts of data set. Further, data mining can be applied to predict a result for a given entity. WUM supports knowing frequently accessed pages, prediction of user navigation, improvement of web site structure etc. The process of WUM consists of these steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis [1]. Pre-processed and cleaned data could be used for pattern discovery and pattern analysis. Clustering is useful in several fields such as pattern discovery, pattern analysis, machine-learning situation, pattern classification and many other fields. Clustering is an unsupervised classification technique widely used for WUM with a main objective of grouping a given collection of unlabeled objects into meaningful clusters [2]. Association rule is one of the data mining tasks that can be used to discover relationship among data. Association rule identifies specific association among data and its techniques are generally applied to a set of transactions in data. Since the amount of data handled is extremely large, current association rule techniques are trying to prune the search space according to support count [3]. Rule discovery finds common rules in the format A→B, meaning that, when page A is visited in a transaction, page B will also be visited in the same transaction. These rules may have different values of confidence and support [4].

II. Web Server Log: Pre-processing and Pattern Discovery

Web server log was the main data source for web log mining that records the user's request information to server. The log form is an expansion, each record contains IP address of a client, method, request page URL, timestamp, receiving bytes, protocol version, user agent and previous page of a visitor [5].

The data pre-processing phase was performed using a reduced log file, which was "cleaned" by removing all useless, irregular, and missing data from the original common log file. Error records, requests for images and multimedia files were also removed from Server log. User Identification was considered as the next step. In user identification, IP address and user agent were used. Meaning that, a combination of IP address and user agent was used to identify a unique user. In session construction, time-oriented approach was used for identifying user sessions. Session timeout threshold was set as 30 minutes.

After performing data pre-processing phase, the pattern discovery method should be applied. This phase consists of different techniques derived from various fields such as statistics, machine-learning method, mainly with Association

Rules, data mining, pattern recognition, etc. being applied to the web domain and to the available data [6]. Several methods and techniques have already been developed for this step [7]. Some of the frequently used solutions are clustering and association rules.

A. Clustering

Clustering has been widely used in WUM to group together similar sessions among large amounts of data based on a general idea of distance function which computes the similarity between groups [8].Clustering means the act of partitioning an unlabeled dataset into groups of similar objects. Each group, called a 'cluster', consists of objects that are similar between themselves and dissimilar to objects of other groups. In the past few decades, cluster analysis has played a central role in diverse domains of science and engineering [9], [10]. Two types of clusters can be found in WUM: user clusters and page clusters. User clusters will discover users having same browsing patterns whereas page clusters will discover pages possessing similar content [11]. Here we will briefly describe some techniques to discover patterns from processed data. Commonly used clustering algorithms are: K-means and Density based clustering.

1) K-Means Clustering

The k-means method partitions the data set to classify objects based on attributes into positive k cluster in which each observation belongs to the cluster with the nearest mean [12]. The clustering is done by minimizing the sum of squared distance in each cluster. Thus, the strength of K-means algorithm lies in its computational efficiency and the nature of easy to use. The procedure follows a simple way to classify a log file dataset. The basic step of k- means clustering is simple. In the beginning we determine number of clusters k and assume the centroid or center of clusters. We can take random objects as the initial centroid or the first k objects, which serves as an initial centroid. Then the k-means algorithm will carry out its steps until convergence. Iterate until stable (no move group) to group the objects based on minimum distance.

2) Density-based clustering

The basic idea of density-based clustering is that clusters are dense regions in the data space, separated by regions of lower object density [13]. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts) [14].

B. Feature Selection

Feature Selection is a term commonly used in data mining to describe the techniques available for reducing inputs to a manageable size for processing and analysis. Correlation-based Feature Selection (CFS) measures correlation between nominal features, so numeric features is first discretized. It is also an effective dimensionality reduction technique and is an essential preprocessing method to remove noise features [15]. The basic idea of feature selection algorithms is searching through all possible combinations of features in the data to find which subset of features works best for prediction. The selection is performed by reducing the number of features of the feature vectors, keeping the most meaningful discriminating ones, and removing irrelevant or redundant ones.

C. Association Rules

Association rule mining is one of the major techniques of data mining and it is the most common form of pattern discovery in unsupervised learning systems. It serves as a useful tool for finding correlations between items in large database [16]. Most common approaches to association discovery are based on the Apriori algorithm. This algorithm finds groups of items (namely; page-views appearing in the preprocessed log) occurring frequently together in many transactions (i.e. satisfying a user specified minimum support threshold) [17]. It finds rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. Two measurements in association rule mining are support and confidence. The support corresponds to the frequency of the pattern while confidence indicates rule's strength.

Support of a rule A → B = no. of instances with A and B / no. of all instances.

Confidence of a rule A → B = no. of instances with A and B / no. of instances with A = support (A & B) / support (A).

The goal of association rule mining is to find all rules having: support ≥ minsup threshold and confidence ≥ minconf threshold.

1) Lift

Lift is an interestingness measure of an association rule that compares the rule confidence to the expected rule confidence. Lift of a rule A → B = support (A & B) / [support (A) * support (B)].

2) 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 [18]. The minimum support threshold parameter needs to be set to the value that gives optimal results. If the support threshold is set too low, too many potentially not truly interesting rules are generated, cluttering the rule set and making it hard to understand for the final user. On the other hand, if the support threshold is set too high, there is a chance that too many potentially interesting rules are missed from the rule set. Eliminating Redundant rules and Clustering decreased the size of the generated rule set for obtain Interestingness rules.

3) Redundant rules

Deleting redundant rules from the result set: If you have A → B and A & C → B, the second rule is redundant.

4) Page cluster

Let us suppose that the set of all rules R contains the following rules: a→ b, conf (a→ b) ≈ 1, b→ a, conf (a→ b) ≈1, where a and b are items a ∈ I, b∈ I.

We define a cluster Cab = {a, b}.

III. Working Scheme

The web logs of SUST University were taken as the input dataset. Clustering of web logs was based on the two types of clusters that can be found in web usage mining: user clusters and page clusters. User clusters will discover users having the same browsing patterns whereas page clusters will discover pages possessing similar content. IP or Agent represented attributes were used for user clusters, where a Requested Page was an attribute used for page clusters. Feature extraction or selection is one of the most important steps in pattern clustering. It is also an effective dimensionality reduction technique and an essential preprocessing method to remove noise features. In this experiment, we performed cfsSubsetEval feature selection method over log file dataset to select relevant features. It evaluates the worth of subset of attributes by considering the individual ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation with other attributes are preferred.

The clustering algorithms were compared according to these factors: Cluster Instances, Number of clusters, Time taken to form clusters, Incorrect cluster Instances, No. of Iterations and Accuracy. Then the Apriori algorithm was applied on the log dataset. This algorithm was suitable for finding correlations between items in large database. The proposed Working Scheme is shown in Figure 1.

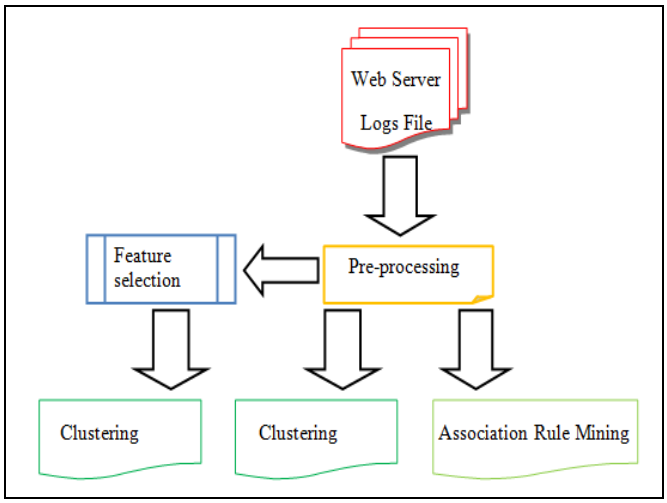

Figure 1. Architectural overview of Working Scheme

IV. Experimental Setup and Results

For experimentation, data from Server log file of the Web site *www.sust.edu* collected from 07-11-2008 to 20-08-2009 was considered for the overall study and analysis. After cleaning the collected data, the number of requests declined from 291642 to 122122. Table 1 shows the detailed changes in data Pre-processing.

Table 1. Statistical Summery

Number of record before cleaning	291642
Number of unique Users	23524
Number of Unique IP address	11030
Number of unique pages	8861
Number of sessions	13869

A. Clustering

K-mean algorithm and Density-based clustering were used to obtain the clusters using WEKA Clustering Tool on a set of Pre-processed log file. The output for the data set for user cluster with two clusters using K-mean is shown in Figure 2. Figure 3 shows output for the same data set with two clusters using Density-based clustering .Figurers 4 and 5 show the content of each cluster.

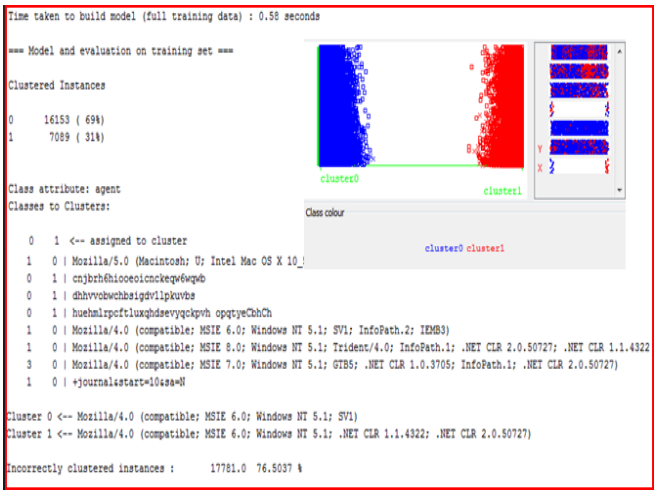

Figure 2. User Cluster using K-means Clustering

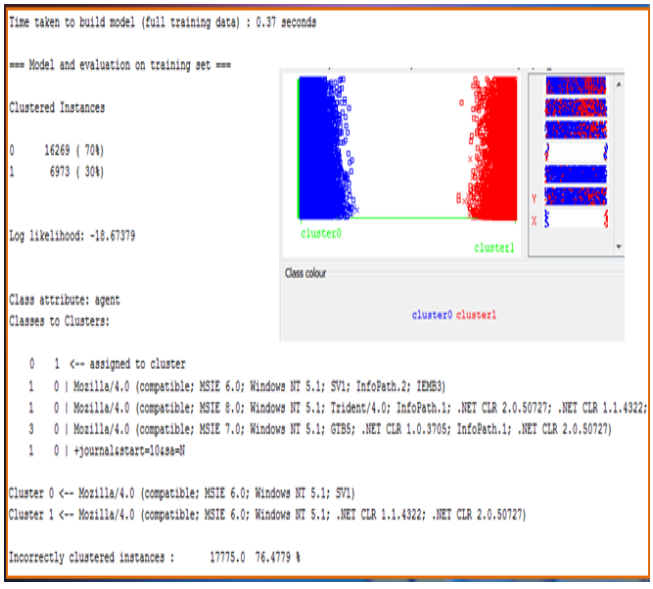

Figure 3. User Cluster using Density Based Algorithm

- Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1)
- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; InfoPath.2)
- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)
- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 2.0.50727)
- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
- Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.10) Gecko/2009042316 Firefox/3.0.10
- Opera/9.80 (Windows NT 5.1; U; en) Presto/2.2.15 Version/10.00
- Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 2.0.5072
- Mozilla/5.0 (compatible; heritrix/1.14.3 +http://www.accelobot.com)
- Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.1.2) Gecko/20090729 Firefox/3.5.2
- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; GTB6)
- Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.8) Gecko/2009032609 Firefox/3.0.

Figure 4.Content of cluster 0

- Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.9) Gecko/2009040821 Firefox/3.0.9
- Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1;1813)
- Mozilla/3.0 (compatible; Indy Library)
- Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; GTB6; .NET CLR 1.1.4322; .NET CLR 2.0.50727; InfoPath.1)
- msnbot/2.0b(+http://search.msn.com/msnbot.htm)
- Mozilla/5.0 (compatible; Yahoo! Slurp/3.0;

Figure 5.Content of cluster 1

Figures 6, 7 and 8 show the clusters according to the page request using K-mean algorithm with 2, 3 and 4 clusters respectively. Figure 9 shows the Page Clustering using Density-based Algorithm with 2 Clusters.

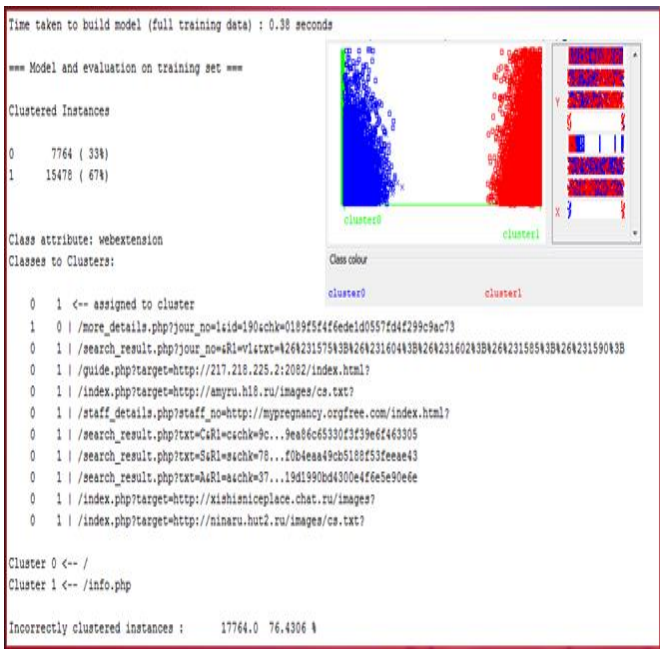

Figure 6. Page Cluster using K-mean with 2 Clusters

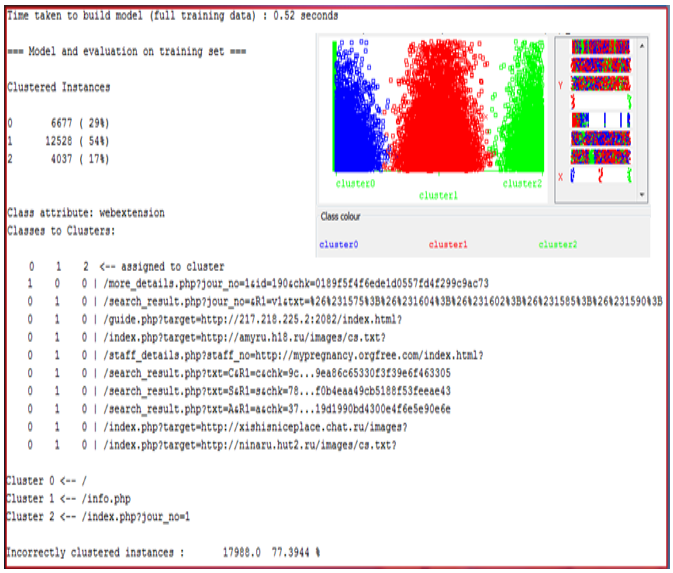

Figure 7. Page Cluster using K-mean with 3 Clusters

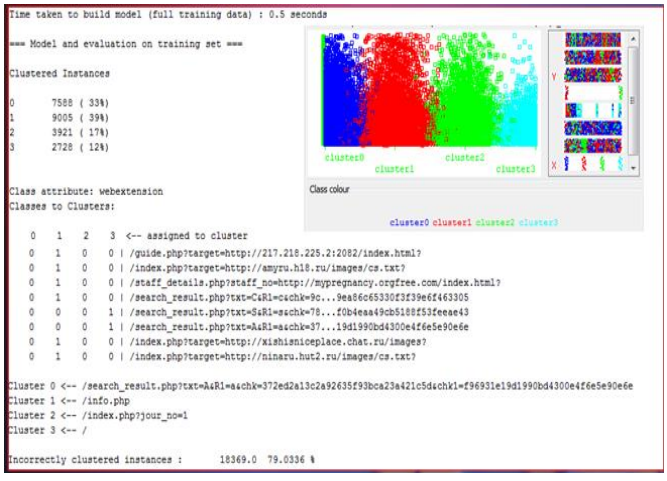

Figure 8. Page Cluster using K-mean with 4 Clusters

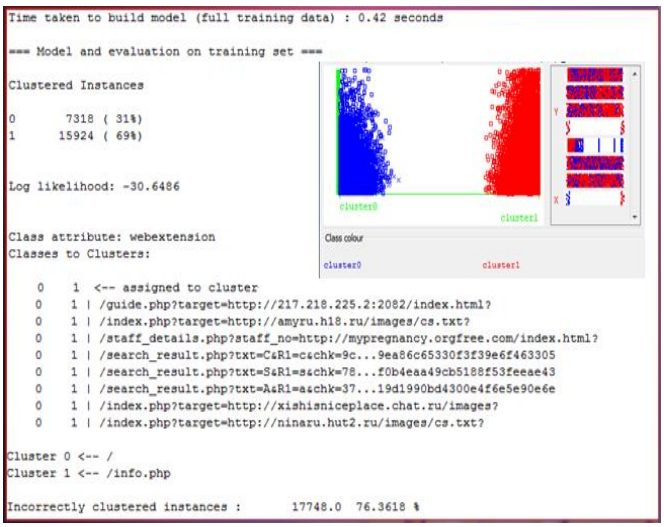

Figure 9. Page Cluster using Density based with 2 Clusters.

According to the previous implementation of the data clustering techniques, the two clustering algorithms are compared according to these factors: Cluster Instances, Number of clusters, Time taken to form clusters, Incorrect cluster Instances, Number of Iterations and Accuracy. It is useful to summarize the results and present some comparison of performances. A summary of the best-achieved results for each of the two techniques is presented in Table 2.

Table 2. Performance results comparison

Algorithm	No. of clusters	Cluster Instances	Time taken to build model	Incorrect cluster Instances	Accuracy	No. of Iterations	Within clusters sum of squared errors
K-Means	2	23242	0.38	17764.0 (76.4306%)	(23.5694%)	8	59696.305
	3	23242	0.52	17988.0 (77.3944%)	(22.6056%)	10	55395.901
	4	23242	0.5	18369.0 (79.0336%)	(20.9664%)	12	54661.442
Density based Clustering	2	23242	0.42	177748.0 (76.3618%)	(23.6382%)	8	59696.305

From this comparison we can conclude that Density-based clustering with 2 clusters produces fairly higher accuracy and lower RMSE than K-means technique with 2 clusters and requires significant computation time.

B. Clustering with and without Feature Selection

In this experiment we performed cfsSubsetEval feature selection method over log file dataset to select relevant features. This method evaluated a subset of attributes, which are more relevant for the requested page (webextension) attribute. It selected only two attributes: IP address (IPADD) and (URL) from 6 attributes (see Figure 10). Then we performed K-means, and Density-based clustering methods on this subset (see Figures 11 and 12). Then we compared the result of clustering method with and without feature selection, as shown in Table.3.

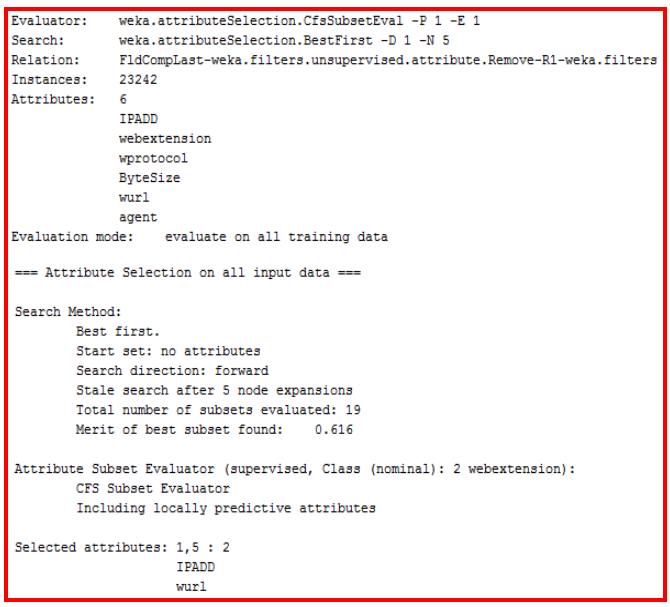

Figure 10. Features selected using filtering technique

Figure 11. K-means with feature selection

Figure 12. Density based with feature selection

Table 3. .Clustering method with and without feature selection.

Algorithm / Factor	K-means Without feature selection	K-means With feature selection	Density based Without feature selection	Density based With feature selection
Incorrectly clustered instance	17764.0 (76.4306%)	16925.0 (72.8208%)	17748.0 (76.3618%)	16887.0 (72.6537%)
Accuracy	(23.5694%)	(27.1792%)	(23.6382%)	(27.3463%)
Time taken to build model (seconds)	0.52	0.21	0.42	0.16
Number of iterations	8	3	8	3
Within cluster sum of squared errors	59696.305	37504.0	59696.305	37504.0

The accuracy of clustering algorithms in terms of correctly classified instances for Dataset is shown in Figure 13.

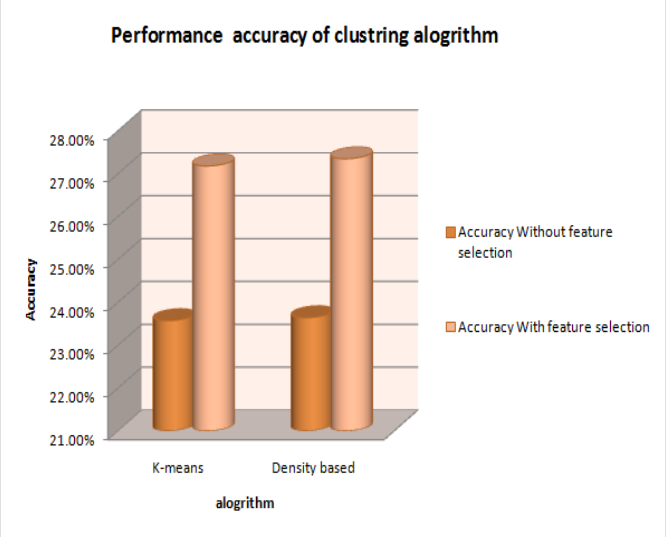

Figure 13.Clustering Performance

For K-means and Density-based clustering, we observed that time was reduced to 0.21 and 0.16 and accuracy increased to 27.18% and 27.35% respectively. K-means and Density-based clustering method, within cluster sum of square errors, was reduced to 37504.0. Also for two algorithms, the number of iterations was reduced to 3.

C. Association Rule Mining

Association rule mining aims to extract interesting correlations, frequent patterns and associations or casual structures among sets of items in the SUST log file. Setting parameter values in the right way, eliminating redundant rules and identify page Clusters lead to an interesting rule, useful for analysis.

1) Setting Parameter Values

While conducting the experiments, we noticed that a lot of interesting rules contained item sets with support of less than 0.1, a default value in WEKA tool. Based on our empirical research, we chose to set the minimum support of an item set to 0.07.

2) The Generated Rule Set

In accordance with our expectations, the initially generated association rule set contained many rules that had very high confidence. There were 10 (out of 50) rules with confidence equal to 1.0, while 29 (out of 50) rules had confidence greater than 0.85. This can be explained by the fact that many web pages are strongly correlated due to the link structure of the website. Figure.14 shows some results of association rule mining

.

```
Best rules found:

 1. wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 ==> wurl
 2. webextension=/info.php agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2346 ==>
 3. webextension=/info.php wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2
 4. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> wurl=- 2577    conf:(1)
 5. webextension=/info.php agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2346 ==>
 6. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> wprotocol=HTTP/1.1 wurl
 7. webextension=/info.php agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2346 ==>
 8. webextension=/info.php wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 23
 9. wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2577 ==> wprotocol=HTTP/1
10. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> wprotocol=HTTP/1.1 2573
11. webextension=/iepngfix.htc 1922 ==> wurl=- 1876    conf:(0.98) lift:(3.16) lev:(0.06) [1282] < conv:(28.26)>
12. webextension=/iepngfix.htc wprotocol=HTTP/1.1 1754 ==> wurl=- 1708    conf:(0.97) lift:(3.15) lev:(0.05) [1166] < conv:(25.79)>
13. webextension=/info.php wprotocol=HTTP/1.1 wurl=- 2456 ==> agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.
14. webextension=/info.php wurl=- 2525 ==> agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.
15. webextension=/info.php wurl=- 2525 ==> wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.
16. wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 ==> webe
17. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> webextension=/info.php
18. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> webextension=/info.php
19. wprotocol=HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 ==> webe
20. wprotocol=HTTP/1.1 wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 =
21. wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2577 ==> webextension=/in
22. wurl=- agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2577 ==> webextension=/in
23. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> webextension=/info.php
24. agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2579 ==> webextension=/info.php
25. webextension=/iepngfix.htc 1922 ==> wprotocol=HTTP/1.1 wurl=- 1708    conf:(0.89) lift:(3.62) lev:(0.05) [1236] < conv:(6.75)>
```

Figure14. Some results of association rule mining

3) Eliminating Redundant Rules

As a first step, and after removing redundant rules, our rule set contained 43 rules out of the 50 rules generated originally. Some redundant rules are selected in Table 4.

Table 4. Some redundant rules.

The rule	Redundant Rule
/iepngfix.htc ==> wurl=-	/iepngfix.htc ,HTTP/1.1 ==> wurl=-
/iepngfix.htc ==> HTTP/1.1	/iepngfix.htc, wurl=- ==> HTTP/1.1
/info.php ==> wurl=-	/info.php ,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> wurl=-
wurl=- ==> /info.php	wurl=- ,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php
wurl=- ==> /info.php ,HTTP/1.1	wurl=- ,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php ,HTTP/1.1

4) Identifying Page Clusters

We eliminated 10 rules and introduced 5 rules as their cluster presentation (1 for each cluster), thus decreasing the size of the rule set by 38 rules (out of 43). For example, we eliminated four rules and introduced their cluster representatives, as shown in Table 5. The confidence of all eliminated rules was close to 1.

5) Interestingness of the Resulting Association Rules

Pruning our rule set according to redundant rules and using Clustering to decrease the size of the rule set from 50 to only 38 rules, we identified a webmaster to enhance the website structure and improve its browsing experience for the visitors. We identified 8 truly interesting rules out of the 38 rules in the rule set (21%).There were numerous rules generated with this algorithm as shown in Figure 14. Some of the interesting rules are shown in Table 6.

The first interesting rule found was that the information page is accessed with the most using agents: Mozilla/4.0 + Windows NT 5.1 and the referrer was '-'. This indicates that the request made to the information page was from regular visitors who know the website well. The second association rule shows that if a user referrer is: http://www.sustech.edu/sudannewar/staff_publicationsAR.php, then they will very likely request web extension=/.

The third association rule was found by a priori algorithm. It is an interesting rule which can be stated as: if visitors visit the "/iepngfix.htc" page with platform Mozilla/4.0, then they will be referrer'-'. This means that the request made to the /iepngfix.htc page is from the regular visitors who use Mozilla/4.0 as agent.

Table 5. Rules and Clusters.

Number of Clusters	Rules and Cluster
1	info.php, HTTP/1.1 ==> Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) conf:(0.8) Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php ,HTTP/1.1 conf:(0.91)
2	/info.php ,HTTP/1.1, wurl=- ==> Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) conf:(0.95) Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php, HTTP/1.1, wurl=- conf:(0.91)
3	/info.php, wurl=- ==> Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) conf:(0.93) Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php, wurl=- conf:(0.91)
4	/info.php ,wurl=- ==> HTTP/1.1 ,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) conf:(0.93) HTTP/1.1,Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) ==> /info.php ,wurl=- conf:(0.91).

Table 6. Some association rules

Number	association rule of homepage
1	webextension=/info.php agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2346 ==> wurl=- 2346 <conf:(1)> lift:(3.24)
2	wurl=http://www.sustech.edu/sudannewar/staff_publicationsAR.php 169 ==> web extension=/ 169 <conf:(1)> lift:(5.06)
3	web extension=/iepngfix.htc agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1) 282 ==> web url=- 274 <conf:(0.97)> lift:(3.15)

4	HTTP/1.1 agent=Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727) 2573 ==> web extension=/info.php 2343 <conf:(0.91)> lift:(6.89)

The fourth association rule shows that, the number of requests made to /info.php page was from web protocol=HTTP/1.1, agent=Mozilla/4.0. This indicates that these users visited the information home page almost use platforms Mozilla/4.0 and Windows NT 5.1 and also used the HTTP/1.1 protocol.

V. Conclusion

Data mining framework, which incorporates the clustering technique along with association rule mining was implemented in this paper. Tow clustering techniques were reviewed in this work, namely: K-means clustering and Density-based clustering. The clustering solved the problem of categorizing data by partitioning a data set into a number of clusters based on some similarity measure so that the similarity in each cluster was larger than among clusters. Clustering algorithms was applied with and without feature selection for SUST log file dataset on WEKA tools. Performance of the clustering method was measured by the percentage of the incorrectly classified instances. As the percentage of the incorrectly classified attribute was low, the performance of the clustering was good. Density-based clustering gave better performance compared to k means clustering without feature selection. Density-based Clustering recognized characters with higher accuracy and minimum amount of time compared to k-means algorithm. Clustering algorithms was applied with feature selection. It was concluded that choice of a good feature can contribute a lot to clustering techniques. Also with feature selection the performance of Density-based clustering was better than K-Means algorithm. In this paper, implementation of a system for pattern discovery using association rules was discussed as a method for Web Usage Mining. Analysis results show that using an association rules in WUM can model the rules for managing and optimizing the website structure and advised to be used by users. This helps the web designers to improve website usability by determining related link connections in the website.

References

[1] Singhal, Vidhu, and Gopal Pandey. "A Web Based Recommendation Using Association Rule and Clustering." *International Journal of Computer & Communication Engineering Research (IJCCER)*, vol. 1, Issue. 1, pp. 1–5, 2013.

[2] H. Rana and M. Patel. "A Study of Web Log Analysis Using Clustering Techniques." *International Journal of Innovative Research in Computer and Communication Engineering* (IJIRCCE),Vol. 1, Issue 4, pp. 925–929, 2013.

[3] Kiruthika, D. Dixit, et al. "Pattern Discovery Using Association Rules." *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 2, No. 12, 2011.

[4] M. Henri Briand, M. Fabrice Guillet, M. Patrick Gallinari, M. Osmar Zaaiane, "Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support", World Academy of Science, Engineering and Technology 48 2008.

[5] Zhong, Xiu-yu. "The research and application of web log mining based on the platform weka." *Procedia engineering* 15, pp. 4073-4078, 2011.

[6] A. Upadhyay and B. Purswani, "Web Usage Mining has Pattern Discovery," *International Journal of Scientific and Research Publications*, vol. 3, no. 2, pp. 1–4, 2013.

[7] K. Etminani, "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method," *IFSA-EUSFLAT*, pp. 396–401, 2009.

[8] Y. Xie, V.V. Phoha, "web user clustering from access log using belief function*", in: proceedings of the first international conference on knowledge capture (k-cap 2001), ACM press,* , pp. 202–208, 2001.

[9] S. Das, a. Abraham, and a. Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 38, no. 1, pp. 218–237, Jan. 2008.

[10] H. Izakian and A. Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem*," Expert Systems with Applications*, vol. 38, no. 3, pp. 1835–1838, Mar. 2011.

[11] S. Dhawan and S. Goel, "Web Usage Mining: Finding Usage Patterns from Web Logs," *American International Journal of Research in Science, Technology, Engineering & Mathematics*, pp. 203–207, 2013.

[12] I. Frades and R. Matthiesen, "Overview on techniques in cluster analysis*," Bioinformatics Methods in Clinical Research*, pp. 81–107, Springer, 2010.

[13] J. Hwang, "A top-down approach for density-based clustering using multidimensional indexes," *The Journal of Systems and Software*, vol. 73, pp. 169–180, 2004

[14] B. Chaudhari and M. Parikh, "A Comparative Study of clustering algorithms Using weka tools," *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, vol. 1, no. 2, pp. 154–158, 2012.

[15] K. Selvakuberan, M. Indradevi, and R. Rajaram. "Combined Feature Selection and classification–A novel approach for the categorization of web pages." *Journal of Information and Computing Science*, vol. 3, No. 2, pp. 083-089, 2008.

[16] S. D. Serasiya and N. Chaudhary, "Simulation of Various Classifications Results using WEKA," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 1, no. 3, pp. 155–162, 2012.

[17] M. K. M and R. Jadhav, "Pattern Discovery Using Association Rules," *(IJACSA) International Journal*

of Advanced Computer Science and Applications, vol. 2, no. 12, pp. 69–74, 2011.

[18] V. Singhal and G. Pandey, "A Web Based Recommendation Using Association Rule and Clustering," *International Journal of Computer & Communication Engineering Research (IJCCER)*, vol. 1, no. 1, pp. 1–5, 2013.