EFFICIENT RESOURCE UTILIZATION OF WEB USING DATA CLUSTERING AND ASSOCIATION RULE MINING

ILAMPIRAY.P

M.Tech-Advanced Computing, School of Computing, SASTRA University, Tamil Nadu, India-613402

E-mail:ilampiray.vp@gmail.com

ABSTRACT

Data mining is a dominant technology which extorts the hidden information from the huge databases. Data mining techniques envisage future progressions and trends, acquiesce businesses to make practical, knowledge-actuate arbitration. Data clustering and association rule mining has attracted a lot of research interest in field of computational statistics and data mining. Web usage mining is an important application of data mining techniques and it is used to determine user navigation pattern from web log data. In this paper, the clustering technique is applied for grouping the users based on the ip address and association rule mining is used for finding the associations between the clustered groups. A hybrid Leaders complete linkage algorithm is used for clustering the users into groups and the Apriori algorithm is used for finding associations between the users. This approach reveals the individual access pattern of users.

Keywords: *Hierarchical Clustering, Incremental Clustering, Web Usage Mining, Association Rule Mining.*

1. INTRODUCTION

Data mining is an evolutionary process which analyse the data from different data sources and summarize those data into useful information. Data clustering is the most active area in data mining and it is used in many applications including business management, user profile analysis, and marketing. For many issues of computer science including statistical analysis, database compression, vector quantization and pattern recognisation the data clustering technique is needed. Clustering analysis aims at identifying group of similar objects and helps to discover distribution of patterns and interesting correlation in data sets. The data clustering algorithms are classified as partitional clustering, fuzzy clustering, hierarchical clustering and artificial neural network for clustering.

The clustering qualities of partitional algorithms are actually poorer and less effective than the agglomerative counterparts. In the framework of document retrieval, the hierarchical algorithm seems to process superior than the partitional algorithms for retrieving related documents. The Hierarchical clustering has the disadvantage that, once the clustering process is completed, one can't back track it. Incremental clustering algorithms are useful to handle huge data set due to their lower time and space complexity. Most of these algorithms are intended for dynamic databases. For improving the clustering quality the hierarchical complete linkage and incremental leaders algorithms are incorporated and a hybrid Leaders Complete Linkage algorithm (LCL) is proposed [2].

Data mining techniques are applied to the web logs for retrieving the user access patterns. The user web logs are clustered i.e., future behavior of user is forecasted. The web pages are assigned into a cluster based on likeness or other relationship measures in the web page clustering. The Association rule mining is used to determine the frequently visited web pages collectively in a Session and interesting associations between huge set of data items.

For grouping the user into different clusters a hybrid Leaders complete linkage algorithm (LCL) is used. The advantage of this algorithm is to avoid rearranging of all the objects. The data mining techniques such as association rules, page clusters and user clusters are applied to web server logs for identifying the user navigation patterns. Web server logs include the browsing information of web users.

Web server logs are clustered using the LCL algorithm and the APRIORI algorithm is used to determine several rules which identify individual user access pattern and the frequent web pages visited by the individual users. Based on these revealed information the bandwidth is limited according to the content of the usage. That is, if a person continuously views the entertainment sites in the web then the bandwidth is limited for those users.

2. RELATED WORKS

Data mining applications can employ a mixture of parameters to inspect the data. They comprise associations among different data objects, path where one data object leads to another data object, classification which identifies new patterns, clustering [1] which group the similar data objects. In each cluster the objects are similar between themselves and dissimilar to the objects in the other clusters. In this paper a hybrid clustering algorithm is used by combining the hierarchical and incremental clustering.

The Partitional clustering algorithms are compatible for clustering huge datasets owing to their relatively low computational requirements. The Partitional clustering algorithm splits the data objects into k partitions and each partition is known as a cluster. K-means [3] is one of the traditional partitional clustering techniques. In the first step, number of clusters is fixed to k.In the second step centers for the desired *k* clusters are identified and the data objects are assigned to the clusters which are nearer to the cluster center. In most of the partitional clustering algorithm i.e., k-mean types of algorithm such as K-medoid, K-median, Fuzzy c-means require the number of clusters k are to be specified in advance.

Hierarchical clustering is one among the methods of cluster analysis and it constructs a hierarchy of clusters which avoids the number of desired clustered to be specified in advance. Hierarchical clustering algorithms are widely classified into two types: Agglomerative and Divisive. In Agglomerative: Clustering process starts in its own cluster, and pair of clusters are merged and moves up in the hierarchy. In Divisive: All Clustering process starts in one cluster, and the splits are performed continuously and move down the hierarchy.

The agglomerative clustering [4] methods are classified as: SLINK (single-linkage) and CLINK (complete-linkage) clustering. In the *complete-linkage* clustering the likeness of one cluster and the other cluster have to be identical to that of the maximum likeness measure from any member of one cluster to any other member of the other cluster. The Hierarchical clustering has the disadvantage that, once the clustering process is completed the objects in the cluster cannot be rearranged.

Incremental clustering algorithms [5] are useful to handle huge data set due to their lower time and space complexity. Most of these algorithms are intended for dynamic databases. Some applications need rearranging of data objects time to time, for these applications incremental clustering algorithms are desired. The incremental clustering algorithms are developed since 1970's. The LEADER algorithm [6] is one of the incremental clustering algorithms which use the threshold value to decide if an data object must be positioned into an existing cluster or a new cluster has to be produced.

Web usage mining [7] information provide recent clients, maintain existing clients, and track clients who are leaving the web site. The procedure of information retrieval from the secondary data such as web server access logs, registration data, user sessions or transactions is defined as web usage mining. Web Usage Mining is performed by three main tasks: Pre-processing, Pattern Discovery and Pattern Analysis.

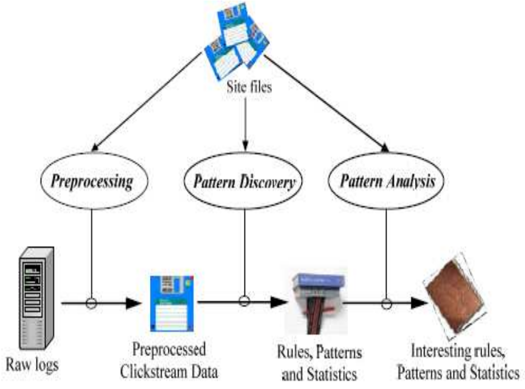


Figure 1. Association Rule Mining

The web pages [8] are assigned into a cluster based on likeness or other relationship measures in the web page clustering. Data mining techniques are applied to the web logs for retrieving the user access patterns. The user web logs are clustered i.e., future behavior of user is forecasted. In web usage mining, the association rules identify the sets of pages that are accessed together which are not directly correlated to one another through hyperlinks.

The Association rule mining [9] is used to determine the frequently visited web pages collectively in a Session and interesting associations between huge set of data items. The discovery of association rules is based on the sessions and each session is elucidated as a transaction. Association rule [10] can identify the series in which the web pages were viewed and these series of pages are called as paths. The association rule mining algorithms are applied to web mining for identifying associations between web pages and exciting access patterns. Association rules are measured to be exciting if both a minimum support and a minimum confidence is satisfied.

3. PROPOSED METHODOLOGY

A hybrid LCL algorithm is used for effective clustering of web server logs. The web logs of SASTRA UNIVERSITY are taken as the input dataset. Clustering of web logs are based on the IP address and it is clustered into two groups of clusters which identify the access patterns of those clustered groups. The association rule mining, Apriori algorithm is applied on the clustered logs. This algorithm is suitable for identifying the frequent web pages viewed by each of the grouped clusters. This method avoids the unnecessary usage of web in the educational institutions.

3.1 Leaders Complete Linkage Algorithm

Leaders Complete Linkage is a hybrid clustering algorithm. Several clustering techniques are proposed to merge the characteristics of two distinct clustering algorithms. The hybrid algorithm first divides the input data set into n sub clusters and then a hierarchical structure is constructed based on these n sub clusters. A hybrid clustering algorithm is used in this paper which is leaders complete linkage algorithm (LCL).It combines the leaders algorithm and complete-linkage algorithm.

Leader clustering is an efficient incremental clustering algorithm in which each of the N number of clusters are represented by a leader. N clusters are generated by using a threshold value. In the first step, pattern is selected as the leader of a cluster and the remaining patterns are classified depending on the existing leaders (Lds) or may become leader of a new cluster. In first step of LCL algorithm, the Leaders algorithm is applied to construct the sub clusters which are represented by leaders.

In second step, complete linkage algorithm is implemented on the predefined sub clusters. The cohesion which is the measure of similarity in between two sub clusters is based on the joinability of clusters. The joinability of two clusters (Cx and Cy) is based on data point 'P' with location 'L' is defined in 3.1..a.

$$(p, Cx, Cy) = \min (fx (l), fy (l)) \quad \text{------ (3.1.a)}$$

$$Ch (Cx, Cy)= \frac{\sum_{P\text{€}Cx,\, Cy} J(P,Cx,Cy)}{|Cx|+|Cy|} \quad \text{---- (3.1.b)}$$

Joinability J of the clusters Cx and Cy are produced by identifying the minimum probability density function fx and fy of point 'P' at location 'L'.By finding the joinability, the cohesion between the clusters are measured using 3.1.b.The cohesion of all possible combinations are determined and fit into the same cluster or a new cluster.

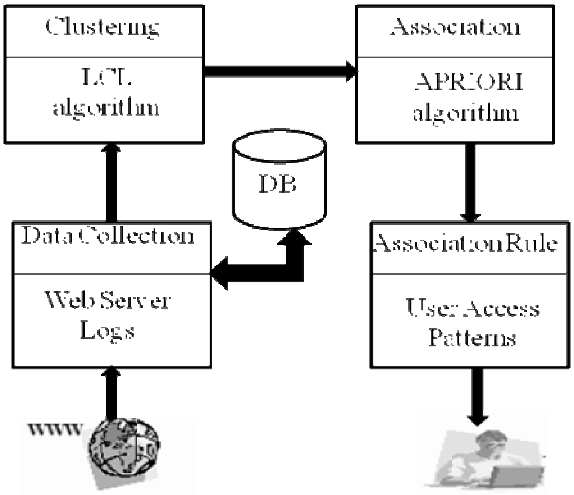


Figure 2. Proposed Methodology

LCL Algorithm

Step1. Select one data object and assume it as leader.
Step2. Choose all the data objects one by one and compare it with leader.
Step3. Obtain leader for each sub cluster.

Step4. Repeat step 2 and 3 until there is no points change between clusters.
Step5. Achieve the cohesions, Ch (Cx, Cy) for all pairs of sub clusters, Cx and Cy.
Step6. A heap is constructed, QCh, with the cohesion of all possible combinations.
Step7. The maximal cohesion is extorted, say Ch (Cx, Cy), from, QCh.
Step8. If Cx and Cy do not fit in to the similar sub cluster, subsequently combine the two sub cluster so that they fit into a new sub cluster.

In LCL algorithm the excellence of the partition is examined at each step, if the excellence of partition is not enhanced then the clustering will be concluded and the recent partition is the final result. The high excellence of clustering is achieved using this algorithm. Web server log data is used as a input to the LCL algorithm for determining web page clusters. The clusters having arbitrary shapes are obtained using the LCL algorithm and enhanced clustering results are also achieved using this algorithm.LCL handles huge data set, accuracy and efficiency of the algorithm is remarkably high when compared to other clustering algorithms.

3.2 APRIORI Algorithm

Association rule mining is used to discover which web pages are frequently viewed by the client. Exciting associations and relationship between web logs are determined by association rules. The Apriori algorithm is intended for the data items which include transactions. Web log contains huge amount of data items hence the Apriori algorithm is choosed for handling these logs. This algorithm has two important terms: Support and Confidence. The support is defined as the amount of transactions that includes all items in the predecessor X and Successor Y is divided by the sum of transactions. There may be some transactions which contain both X and Y then the association rule is defined as X→Y. The confidence is defined as the amount of transactions which comprise every items in both successor as well as the predecessor to the amount of transactions which comprise every items in the predecessor.

APRIORI Algorithm

```
Cn: Candidate item set of size n
Ln: Frequent itemset of size n
L1 = {frequent items};
Begin
For (n = 1; Ln! =∅; n++) do begin
        Cn+1 = candidates generated from Ln;
End For
For each transaction t in database do
        increment the count of all candidates in Cn+1
        that are contained in t
        Ln+1 = candidates in Cn+1 with min_support
End For
Return ∪n Ln;
End
```

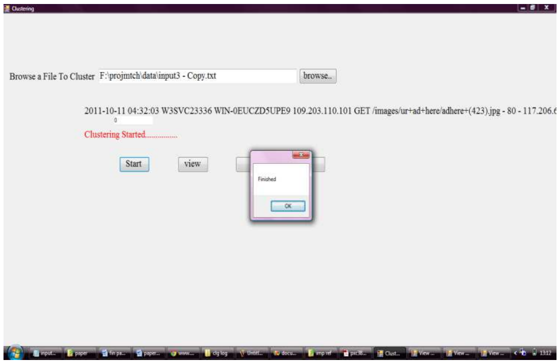
The Apriori algorithm is a proficient algorithm for determining all frequent web pages. The Frequent web pages form the association rules. This algorithm equips a level-wise investigation using frequent web pages and it could be additionally optimized. The Apriori algorithm is best in handling the web log which contains huge amount of transactions in it. Apriori algorithm is capable of identifying the web pages viewed by each individual user and it is parallelized and implemented.

4. ADVANTAGES OF PROPOSED METHODOLOGY

The K-means algorithm is usually applied for effective clustering of data. In this algorithm the numbers of desired clusters are to be specified in advance and all the data objects are processed essentially at the same time. The K-means algorithm could not handle the huge amount of data. Where as in the LCL algorithm which is used in our approach, the dynamic and huge web logs are processed frequently within the period of time. The time and space complexity of LCL algorithm is low.

5. EXPERIMENTAL RESULT

The above methodology is applied for the SASTRA university web logs. The web logs are clustered based on the ip address and the individual user access pattern of those clustered groups are identified. LCL algorithm is applied on these web logs for efficient clustering. Figure 3 defines the clustering process of the web logs.



Browse a File To Cluster F:\projmtch\data\input3 - Copy.txt

```
2011-10-11 04:32:03 W3SVC23336 WIN-0EUCZD5UPE9 109.203.110.101 GET /images/ur+ad+here/adhere+(423).jpg - 80 - 117.206.6
```

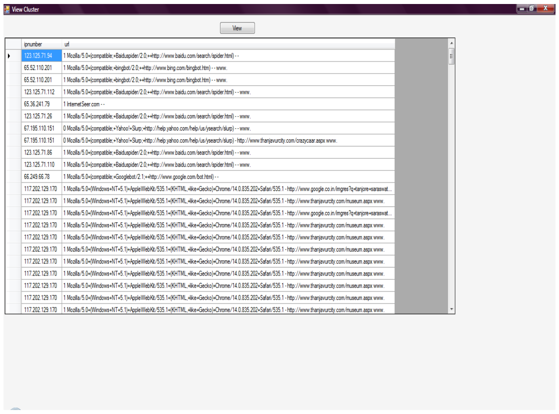
0

Clustering Started...............

Finished

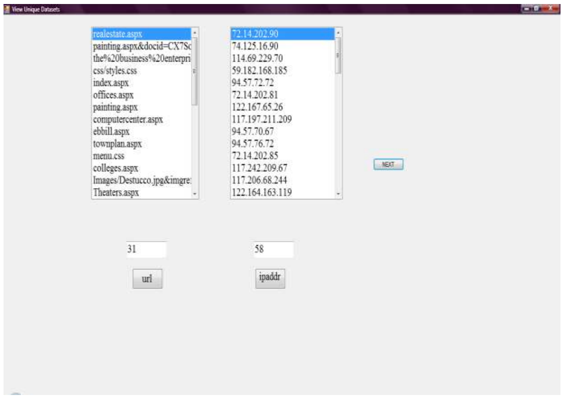
Figure 3. Clustering Process

The clustered web logs based on ip address are listed in the Figure 4. The distinct url and ip address are clustered into groups have been listed in the Figure 5. The Apriori algorithm is applied on the clustered groups and the urls been visited by the users are identifed. This aspect is shown in the Figure 6.



ipnumber	url
123.125.71.94	1 Mozilla/5.0+(compatible;+Baiduspider/2.0;++http://www.baidu.com/search/spider.html) - -
65.52.110.201	1 Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm) - - www.
65.52.110.201	1 Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm) - - www.
123.125.71.112	1 Mozilla/5.0+(compatible;+Baiduspider/2.0;++http://www.baidu.com/search/spider.html) - - www.
65.36.241.79	1 InternetSeer.com - -
123.125.71.26	1 Mozilla/5.0+(compatible;+Baiduspider/2.0;++http://www.baidu.com/search/spider.html) - - www.
67.195.110.151	0 Mozilla/5.0+(compatible;+Yahoo!+Slurp;+http://help.yahoo.com/help/us/ysearch/slurp) - - www.
67.195.110.151	0 Mozilla/5.0+(compatible;+Yahoo!+Slurp;+http://help.yahoo.com/help/us/ysearch/slurp) - http://www.thanjavurcity.com/crazycaar.aspx www.
123.125.71.86	1 Mozilla/5.0+(compatible;+Baiduspider/2.0;++http://www.baidu.com/search/spider.html) - - www.
123.125.71.110	1 Mozilla/5.0+(compatible;+Baiduspider/2.0;++http://www.baidu.com/search/spider.html) - - www.
66.249.66.78	1 Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html) - -
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.google.co.in/imgres?q=tanjore+saraswat...
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.thanjavurcity.com/museum.aspx www.
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.google.co.in/imgres?q=tanjore+saraswat...
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.thanjavurcity.com/museum.aspx www.
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.thanjavurcity.com/museum.aspx www.
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.thanjavurcity.com/museum.aspx www.
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.thanjavurcity.com/museum.aspx www.
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.thanjavurcity.com/museum.aspx www.
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.thanjavurcity.com/museum.aspx www.
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.thanjavurcity.com/museum.aspx www.
117.202.129.170	1 Mozilla/5.0+(Windows+NT+5.1)+AppleWebKit/535.1+(KHTML,+like+Gecko)+Chrome/14.0.835.202+Safari/535.1 - http://www.thanjavurcity.com/museum.aspx www.

Figure 4. Clustered Web Logs



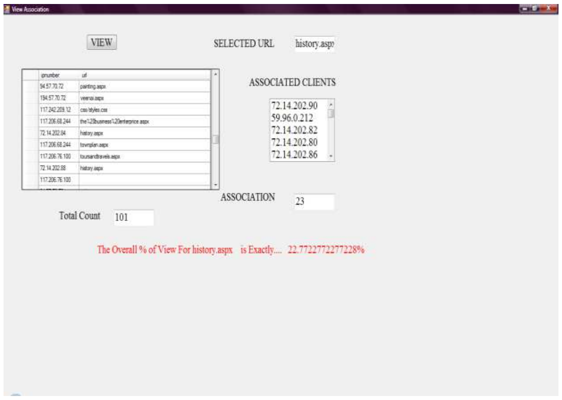
realestate.aspx
painting.aspx&docid=CX7Sc
the%20business%20enterpri
css/styles.css
index.aspx
offices.aspx
painting.aspx
computercenter.aspx
ebbill.aspx
townplan.aspx
menu.css
colleges.aspx
Images/Destucco.jpg&imgre:
Theaters.aspx

72.14.202.90
74.125.16.90
114.69.229.70
59.182.168.185
94.57.72.72
72.14.202.81
122.167.65.26
117.197.211.209
94.57.70.67
94.57.76.72
72.14.202.85
117.242.209.67
117.206.68.244
122.164.163.119

31

58

Figure 5. Distinct IP And URLs Are Listed



SELECTED URL history.aspx

ipnumber	url
94.57.70.72	painting.aspx
194.57.70.72	veenai.aspx
117.242.209.12	css/styles.css
117.206.68.244	the%20business%20enterprice.aspx
72.14.202.84	history.aspx
117.206.68.244	townplan.aspx
117.206.76.100	toursandtravels.aspx
72.14.202.88	history.aspx
117.206.76.100	

ASSOCIATED CLIENTS

72.14.202.90
59.96.0.212
72.14.202.82
72.14.202.80
72.14.202.86

ASSOCIATION 23

Total Count 101

The Overall % of View For history.aspx is Exactly.... 22.7722772277228%

Figure 6.IPs Associated with each URL

The band width utilization of each user(ip address) is identified from the web server logs. Figure 7 illusterates the amount of data packets been sent and received by the experimental ip address.The experimental result shows that the amount of bandwidth received is comparitively higher than the sent packets.Using this result ,the web administrator of educatitional institutions allocate the banwidth to individual users based on the content viewed by those users.

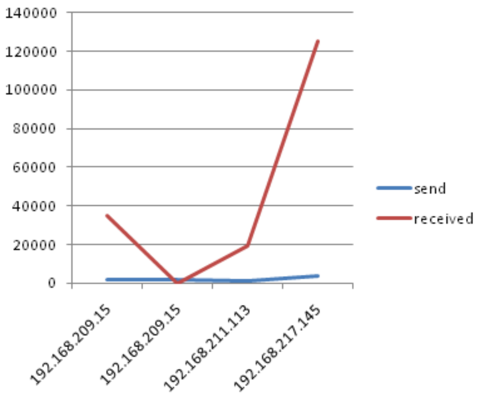


Figure 7.Bandwidth Utilization Of Each Users

6. CONCLUSION

A new data mining framework which incorporates the clustering technique along with association rule mining is implemented in this paper. The LCL clustering algorithm and the APRIORI algorithm is used for this purpose. This data mining technique is applied for SASTRA UNIVERSITY web server logs. These logs are clustered into different clusters and then the associations in each individual clusters are determined. The web usage of individual users is identified for proper allocation of bandwidth to that user in future. The information revealed from this methodology provides efficient utilization of web in educational institutions. This work can be extended by providing privacy for user web logs.

REFERENCES

[1] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A review. ACM Computing Survey (CSUR), 31(3):264-323, 1999.

[2] Srinivas and C. Krishna Mohan, Efficient Clustering Approach using Incremental and Hierarchical Clustering Methods.IEEE, 2010.

[3] K.Alsabti, S.Ranka and V.Singh. An Efficient K-Means Clustering Algorithm http://www.cise.ufl.edu/ ranka/, 1997.

[4]Distances between Clustering, Hierarchical Clustering.www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pd 36-350, Data Mining 14 September 2009.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Michael Wimmer, Xiaowei Xu. Incremental Clustering for Mining in a Data Warehousing Environment. Institute for Computer Science, University of Munich.

[6] Dalong Li, Steven Simske.Training Set Compression by Incremental Clustering. Journal of Pattern Recognition Research, Vol 6, No 1 (2011).

[7] Kobra Etminani, Mohammad-R, Akbarzadeh-T, Noorali Raeeji Yanehsari. Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method. IFSA-EUSFLAT 2009.

[8] Mrs. Kiruthika M and Mrs. Dipa Dixit.Mining Access Patterns Using Clustering. International Journal of Computer Applications (0975 – 8887) Volume 4– No.11, August 2010.

[9] Resul DAŞ and İbrahim Türkoğlu.Extraction of Interesting Patterns Through Association Rule Mining For Improvement Of Website Usability. Journal Of Electrical & Electronics Engineering, vol 9, No 2(2009).

[10] Bamshad Mobasher and Robert Cooley, Jaideep Srivastava.Creating Adaptive Web Sites Through Usage-Based Clustering of URLs.