'Research Journey' International Multidisciplinary E- Research Journal
Special Issue I – Emerging Trends & Issues in Research & Development
Global Impact Factor – 0.565 (2014) & 0.676 (2015)
ISSN : 2348-7143 (Online)
April-May-June - 2016

WEB USAGE MINING FOR COMPARING USER ACCESS BEHAVIOUR USING CLUSTERING TECHNIQUE

Amit Dipchandji Kasliwal[1], Dr. Girish S. Katkar[2]

[1]Department of Computer Science, M.S.G. College, Malegaon Camp, Dist. Nashik – 423203
[2]Head, Department of Computer Science, Arts, Commerce and Science College Koradi, Nagpur, 441111
[1]amitkasliwal16@gmail.com, [2]girishkatkar2007@rediffmail.com

Abstract

Web mining has been implemented and developed into several research areas to satisfy the current era's need. The term web mining is prepared to symbolize the use of data mining techniques for automatically discovering hidden information from the data available over internet in the form of web content, website structure and web usage. This paper work represents the algorithm based on predictive approach of data mining technique used for web usage mining for comparing access behavior. Clustering technique is used to group items with similar characteristics. When applied to web access log data for comparing user access behavior, clustering technique used to group the accesses based on the characteristics. The main advantage that Clustering has it makes group of common characteristics like user's interest, behavior together to classify different user requests. In this research work, authors proposed an algorithm named ClusterView based on clustering technique.

Keyword: Web Usage Mining, Access Behaviour, Clustering, K-Means.

Introduction

Clustering can be supervised data mining or unsupervised data mining as given in [3] means that it does not rely on predefined training itemset or semisupervised data mining proposed in [4]. Clustering can be of two types either distance based or of model based. In distance based clustering technique, it involves determining distance computed between pairs of data items and then creating group of similar items putting together into clusters such as partitioned clustering and hierarchical clustering. Partitioned methods create partitions of data items into *K* groups and it is represented by k-means algorithm. A hierarchical method used to build hierarchical group (or set) of interleaved clusters with the top level cluster containing a single cluster of all data item. In model based clustering defined by Zhong et al and Ghosh et al [5], the model type is proposed as the model structure that can be determined by model selection techniques and parameters estimated using maximum likelihood algorithms such as Expectation Maximization (EM) technique. In proposed ClusterView algorithm, authors used partitioned and hierarchical clustering method to compare the user access behavior on internet. Clustering can also be used in data preprocessing task for classification technique that could operate on the detected clusters.

Proposed ClusterView Algorithm

To implement the ClusterView algorithm, authors applied partitioning algorithm to finds groups of strongly related (or correlated) pages by making partition according to its connected items those have the common characteristics that used for comparing the web access behavior using web access log data stored at server side. At first proposed algorithm performs Depth First Searching techniques in the web access log file that brings all related component in to matrix M by searching form the file. Once the item found, the algorithm checks if there are any items which not get visited or considered. It means that a related item has been split and therefore it needs to be identified to get into M. Proceeding further, in the proposed system authors used the two key aspects entropy and variance which are implemented in the proposed algorithm by counting the time spend on requestedURL and the occurrence in the manner of the number of users used a particular page. These two considerations used in identifying the relation between each pair of webpages. Therefore authors proposed a formula on statistical experiments to bring all those pages together which are accessed maximum number of times with the maximum number of time spent on it as shown below

$$Support_i = \frac{2 * (TimeTaken_i * Occurrence_i)}{(TimeTaken_i + Occurrence_i)}$$

Where,

$Support_i$ is calculated minimum support measure for webpage i
$TimeTaken_i$ is time taken by user that spent on accessing webpage i
$Occurrence_i$ is the total number of time that page i is accessed

Let us take this formula for counting support of each webpage in the web access log file. Considering a worst case, when all the URLs are in the same cluster, the output generated by the algorithm will have execution cost of this algorithm it the linear form of data available in the web access log data. Before putting items into the clusters for getting navigational pattern profile, the generated clusters are then ranked based on data available in the matrix M.

Visualizing Result

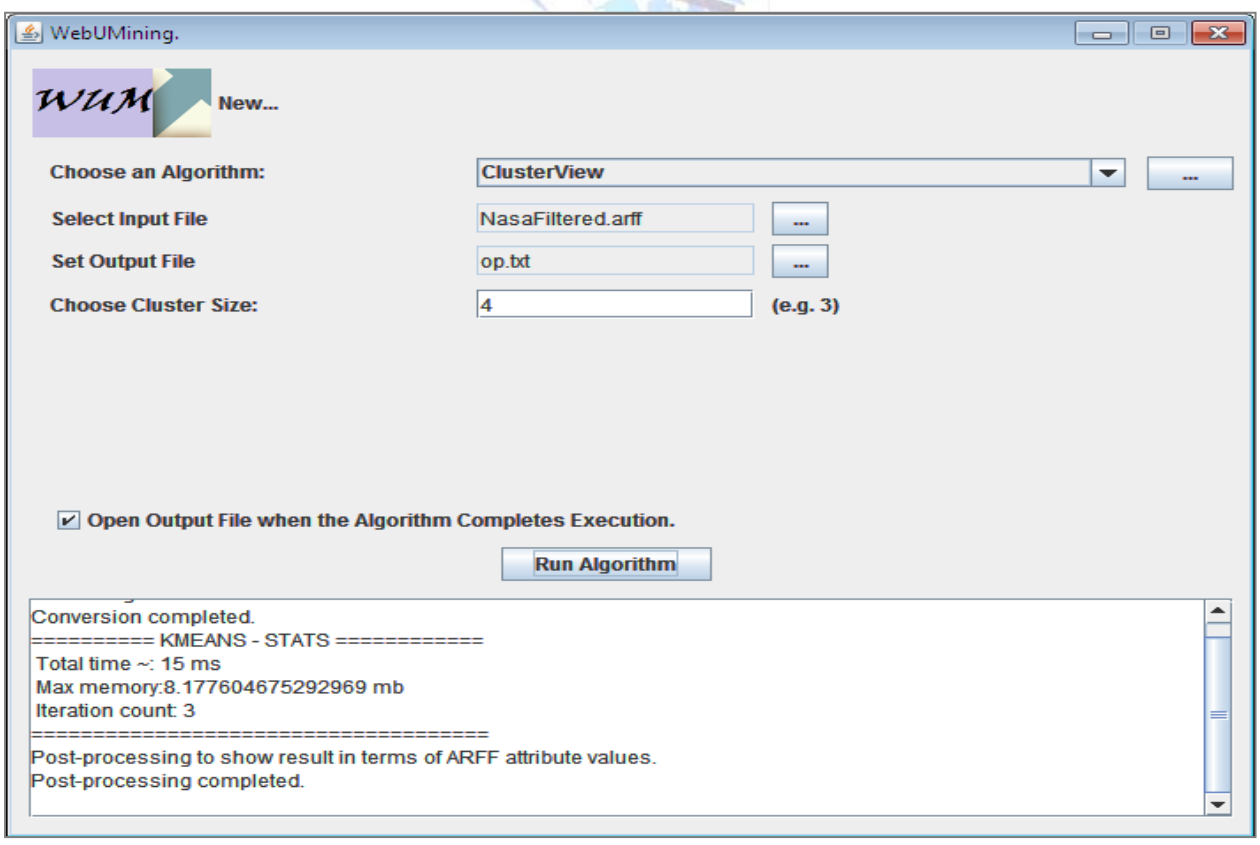
Here, authors worked on the web access log data 'NASALog.arff' of NASA's Kennedy Space Centre's website available at the archive.org. On this input file, experiment performed to find out the clusters base on the web pages and the maximum times spend over that webpage. The proposed ClusterView algorithm can also be used to compare and for predicting user access behavior through web access log data and it is implemented in WebUMining model. During experiment, the focus is on looking and searching for the webpages that are accessed by the users who had the

common characteristics about their behavior access from web log data. This file is converted from text file to arff file format using WEKA as shown in below:

Row No.	IPADDRESS	DateTime	Link string	Protocol	Status	Bytes
1	endeavor.fujitsu.co.jp	[01/Aug/2009:00:01:51-0400]	GET/shuttle/missions/sts-68/ksc-srl-image.html	HTTP/1.0	200	1404
2	www-d3.proxy.aol.com	[01/Aug/2009:00:01:52-0400]	GET/shuttle/missions/sts-71/mission-sts-71.html	HTTP/1.0	200	13450
3	in24.inetnebr.com	[01/Aug/2009:00:01:54-0400]	GET/shuttle/missions/sts-68/news/sts-68-mcc-09.txt	HTTP/1.0	200	2166
4	205.163.36.61	[01/Aug/2009:00:01:55-0400]	GET/shuttle/countdown/countdown.html	HTTP/1.0	200	4324
5	rpgopher.aist.go.jp	[01/Aug/2009:00:01:58-0400]	GET/ksc.html	HTTP/1.0	200	7280
6	139.230.35.135	[01/Aug/2009:00:02:02-0400]	GET/shuttle/missions/sts-49/mission-sts-49.html	HTTP/1.0	200	9271
7	54.teleport.com	[01/Aug/2009:00:02:03-0400]	GET/history/apollo/apollo-13/apollo-13.html	HTTP/1.0	200	18556
8	piweba4y.prodigy.com	[01/Aug/2009:00:02:04-0400]	GET/history/apollo/apollo.html	HTTP/1.0	200	3260
9	205.163.36.61	[01/Aug/2009:00:02:10-0400]	GET/cgi-bin/imagemap/countdown70?342 281	HTTP/1.0	302	98
10	piweba1y.prodigy.com	[01/Aug/2009:00:02:13-0400]	GET/	HTTP/1.0	200	7280
11	165.213.131.21	[01/Aug/2009:00:02:15-0400]	GET/procurement/procurement.html	HTTP/1.0	200	3646
12	205.163.36.61	[01/Aug/2009:00:02:15-0400]	GET/shuttle/countdown/liftoff.html	HTTP/1.0	304	0
13	haraway.ucet.ufl.edu	[01/Aug/2009:00:02:20-0400]	GET/facilities/mlp.html	HTTP/1.0	200	2653
14	rpgopher.aist.go.jp	[01/Aug/2009:00:02:27-0400]	GET/shuttle/countdown/	HTTP/1.0	200	4324
15	slppp6.intermind.net	[01/Aug/2009:00:02:30-0400]	GET/history/skylab/skylab-2.html	HTTP/1.0	200	1478
16	in24.inetnebr.com	[01/Aug/2009:00:02:32-0400]	GET/shuttle/missions/sts-68/news/sts-68-mcc-10.txt	HTTP/1.0	200	1712
17	rpgopher.aist.go.jp	[01/Aug/2009:00:02:45-0400]	GET/cgi-bin/imagemap/countdown70?181 275	HTTP/1.0	302	110
18	rpgopher.aist.go.jp	[01/Aug/2009:00:02:47-0400]	GET/shuttle/missions/sts-70/movies/movies.html	HTTP/1.0	200	2979
19	www-d3.proxy.aol.com	[01/Aug/2009:00:02:54-0400]	GET/shuttle/countdown/	HTTP/1.0	200	4324
20	in24.inetnebr.com	[01/Aug/2009:00:02:57-0400]	GET/shuttle/missions/sts-68/news/sts-68-mcc-11.txt	HTTP/1.0	200	2187
21	rpgopher.aist.go.jp	[01/Aug/2009:00:03:14-0400]	GET/shuttle/missions/sts-70/movies/woodpecker.mpg	HTTP/1.0	200	190269
22	piweba1y.prodigy.com	[01/Aug/2009:00:03:22-0400]	GET/history/history.html	HTTP/1.0	200	1602
23	in24.inetnebr.com	[01/Aug/2009:00:03:28-0400]	GET/shuttle/missions/sts-68/news/sts-68-mcc-12.txt	HTTP/1.0	200	1881
24	gw1.att.com	[01/Aug/2009:00:03:33-0400]	GET/shuttle/missions/sts-73/mission-sts-73.html	HTTP/1.0	304	0
25	haraway.ucet.ufl.edu	[01/Aug/2009:00:03:39-0400]	GET/shuttle/missions/sts-71/mission-sts-71.html	HTTP/1.0	200	13450
26	ai.asu.edu	[01/Aug/2009:00:03:45-0400]	GET/	HTTP/1.0	200	7280

Input web access log data file from NASA's Kennedy Space Centre

The developed model named WebUMining accepts this input file and after selecting the proposed ClusterView algorithm that performs its process as discussed earlier. Overall process of the algorithm implemented with the WebUMining model is represented with the following figures.



WebUMining user Interface for selecting ClusterView Algorithm

Result from the ClusterView algorithm achieved as shown in figure 6.4 and the visualization is represented using the same technique as authors visualized the AssociationRule algorithm to represent the associated webpages. In the results shown in figure 6.4 is generated with the minimum cluster size 4 in that the numbers represent the webpages that are accessed by the users and we can simply compare these after looking towards the each cluster combined in between [ ] (square brackets). The web pages that form part of a specific cluster are then Listed with the specific cluster as shown in class with some colored similar to show their class.

Clusters Based on IPAddress

Time taken to build model (full training data) : 122 seconds

Clustered Instances
0 9 (32%)
1 19 (68%)

Class attribute: IPADDRESS
Classes to Clusters:
0 1 <-- assigned to cluster

0 1 | endeavor.fujitsu.co.jp 1 1 | www-d3.proxy.aol.com 1 0 | in24.inetnebr.com 1 2 | 205.163.36.61 0 1 | rpgopher.aist.go.jp 0 1 | 139.230.35.135 0 1 | 54.teleport.com 0 1 | piweba4y.prodigy.com 1 0 | piweba1y.prodigy.com	0 1 | 165.213.131.21 0 3 | haraway.ucet.ufl.edu 1 3 | rpgopher.aist.go.jp 0 1 | slppp6.intermind.net 2 2 | in24.inetnebr.com 0 1 | piweba1y.prodigy.com 1 0 | gw1.att.com 1 0 | ai.asu.edu

Cluster 0 <-- in24.inetnebr.com
Cluster 1 <-- haraway.ucet.ufl.edu

Cluster Based on the IPAddress

Table 6.1 shows the cluster for IPAddress that spend the maximum time on the particular page. The cluster 0 represents the common items containing the IPAddresses of the user who followed a sequential path and during that they made the visits to more than one webpage in a session. This clusters represented with the numbers 0 and 1 so that it will be easy for website administrator to compare the users those are accessing the website more than once and creating sequential navigational path during access to website. The items in cluster 0 contain 32% of the overall access made to the website while cluster 1 contains 68% total access.

Clusters Based on Requested URL

Time taken to build model (full training data) : 217 seconds

Clustered Instances
0 11 (39%)
1 17 (61%)

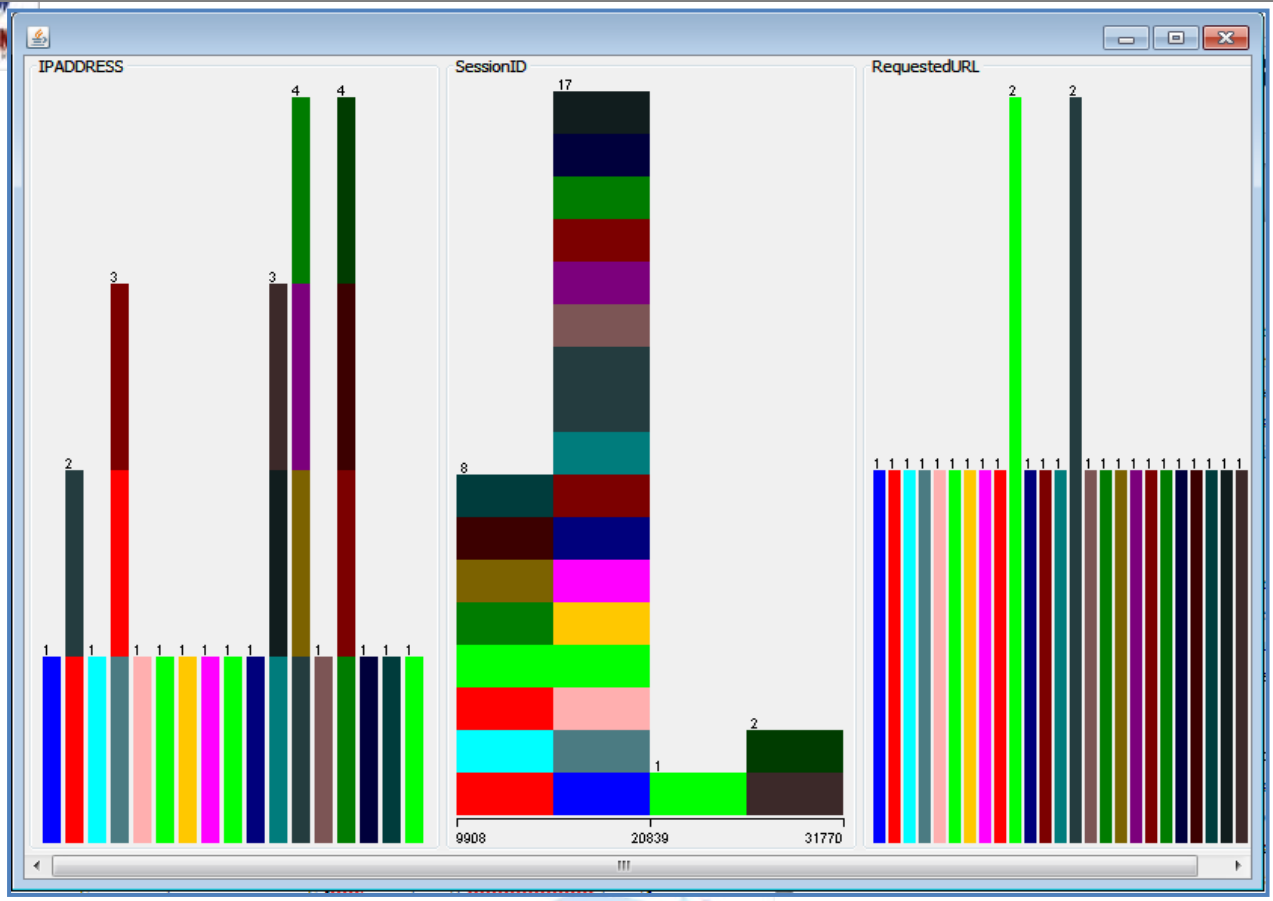
Class attribute: RequestedURL
Classes to Clusters:
0 1 <-- assigned to cluster

0 1 | GET/shuttle/missions/sts-68/ ksc-srl-image.html	0 1 | GET/history/apollo/ apollo.htm
1 0 | GET/shuttle/missions/sts-71/ mission-sts-71.html	1 0 | GET/cgi-in/imagemap/ countdown70?342 281
1 0 | GET/shuttle/missions/sts-68/ news/sts-68-mcc-09.txt	0 1| GET/procure/procurement.htn
0 1 | GET/shuttle/countdown / countdown.html	0 1 | GET/shuttle/countdown/ liftoff.htn
0 1 | GET/ksc.html	0 1 | GET/facilities/mlp.html
0 1 | GET/shuttle/missions/sts-49/ mission-sts-49.html	0 2 | GET/shuttle/countdown/
0 1 | GET/history/apollo/apollo13/ apollo-13.html	0 1 | GET/history/skylab/ skylab- 2.html
0 1 | 54.teleport.com	1 0 | GET/shuttle/missions/sts-68/news/sts-68-mcc-10.txt
0 1 | piweba4y.prodigy.com	1 0 | GET/cgi-bin/imagemap/ countdown70?181 275
1 0 | piweba1y.prodigy.com	0 1 | GET/shuttle/missions/sts-70/movies/movies.html
0 1 | GET/shuttle/technology/sts-newsref/sts_asm.html	1 0 | GET/shuttle/missions/sts-68/news/sts-68-mcc-11.txt
1 0 | GET/shuttle/missions/sts-68/ news/sts-68-mcc-13.txt	1 0 | GET/shuttle/missions/sts-70/movies/woodpecker.mpg
0 1 | GET/shuttle/missions/sts-71/ mission-sts-71.html	0 1 | GET/history/history.html
1 0 | GET/shuttle/missions/sts-73/ mission-sts-73.html	1 0 | GET/shuttle/missions/sts-68/news/sts-68-mcc-12.txt

Cluster 0 <-- GET/shuttle/missions/sts-71/mission-sts-71.html
Cluster 1 <-- GET/shuttle/countdown/

Cluster Based on the Requested URL

Table 6.2 represents the clusters created on RequestedURL information of the web access log file. The cluster 0 represents the class of those webpages which are accessed more than once while cluster 1 represents the class that contains the webpages those are accessed at least once. Figure 6.6 shows the groups of the filtered information which used for clustering as discussed and shown in the figure. The visualization of these information and the generated results are shown in the following figures.



Result of WebUMining using ClusterView algorithm

The figures given above are the graphical representation of the clusters created with the proposed ClusterView and visualizes with the WEKA software for interpretation.

Conclusion

The result can be interpreted and determined by the requirement criteria when given as inputs after selecting ClusterView algorithm. At the initial stage of the visualization, it involves the visualization software as already discussed in chapter 3and then for specific form result authors used WEKA and SPSS software for same. These software provides the advantage of being accessible for any process of analysis over log data hence it makes sensible for making decision about the access to internet so that for any type of comparison since it is in demand to use for web usage mining process.

References

[1] Perkowitz M. and Etzioni O., "Adaptive websites: Conceptual cluster mining", International Joint Conference on Artificial Intelligence, 1999.

[2] Eick C., Zeidat N. and Zhao Z., "Supervised clustering algorithms and benefits", International Conference on Tools with Artificial Intelligence, 2004.

[3] Thomas F. and Joachims T., "Supervised k-Means Clustering", 2003.

[4] Albanese M., Picariello A. and Sansone C., "A Web Personalization System based on Web Usage Mining Techniques", ACM WWW, 2004.

[5] Basu S., Bilenko M. and Mooney R., "A probabilistic framework for semi-supervised clustering", ACM SIGKDD, 2004.

[6] Zhong S. and Ghosh J., "A Unified Framework for Model-based Clustering", Journal of Machine Learning Research, 2003.

[7] Tribus M. and McIrvine E., "Energy and Information", Scientific America, 1971.

[8] Valsamidis S., Kontogiannis S., Kazanidis I., Theodosiou T. and Karakos A, "A Clustering Methodology of Web Log Data for Learning Management Systems", Educational Technology & Society, 2012.

[9] Guo J., Keselj V. and Gao Q., "Integrating Web Content Clustering into Web Log Association Rule Mining", Advances in Artificial Intelligence, 2005.

[10] Etzioni O., "The World Wide Web: Quagmire or gold mine", Communications of the ACM, 1996.

[11] Rajimol, "Data Mining For Web Intelligence", Thesis, MG University, 2013.

[12] http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html