COMPARATION POSSIBILITIES OF K-MEANS AND HAC CLUSTERING IN ANALYSIS OF USERS' PATTERNS OF BEHAVIOR

Marija Blagojević[1]
[1]*Technical Faculty Čačak*

Abstract - *The paper presents a comparison of k-means and HAC clustering. Here is determined the applicability of these methods of clustering in the analysis of user behavior patterns. Also, in this paper, differences between them were observed.*

1. INTRODUCTION

The World Wide Web (WWW) is a vast resource of multiple types of information in varied formats. Need for discovering and analysis of new behavior patterns of the users has increased since the expansion of the web. Analysis of users' patterns of behavior can be used for new model designing that can be of high importance for understanding of users' behavior in virtual environment.

According to [1], clustering can be used for determination of users` patterns of behavior in e-learning and in e-commerce domain as well. They propose in that paper new algorithm based on sequence alignment to measure similarities between web sessions where sessions are chronologically ordered sequences of page accesses. Data mining techniques are applied on log files for the purpose of obtaining recommendations for efficiency improvement within electronic courses [2]. That paper [2] proposes a platform dependant framework for recording, processing and analyzing data from Learning Management Systems (LMS).

Data mining presents analysis of observational data sets with the purpose for detection of undetected links and data summing in a sophisticated manner, understandable and useful for data owner [3]. The relations that are obtained by the data mining process are defined as models or patterns. K-means [4] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or *agglomerate*) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called *hierarchical agglomerative clustering* or *HAC*. [5]

2. PURPOSE OF THE STUDY

Learning management system (LMS) is a software application for the administration, documentation, tracking, and reporting of training programs, classroom and online events, e-learning programs, and training content [6]. However, LMS doesn't allow detail monitoring of the users' activities nor the evaluation of the course contents' structure and its efficiency in the teaching process. In order to consider the complete teaching process that includes the usage of electronic courses within a specific LMS, a thorough analysis is a must. Bearing that in mind and other techniques that are used in electronic courses evaluation, a comparison of two clustering types has been conducted. The comparison was conducted in order to determine differences in the application of the mentioned techniques during the detection of users` patterns of behavior.

Tasks of the study:

- Data pre-processing: clean and prepare the Web server log file
- Application of k-means and HAC-clustering on pre-processed data
- Analysis of obtained results and evaluation of users` patterns of behavior

Purpose of the study:

- Determination of possibilities of the k-means and HAC clustering application in the analysis of users` behavior patterns and their comparison

3. METHODS

Clustering that is applied on log files, is also used for analysis of users` behavior patterns.

3.1 Participants

Data is collected on a sample that consists of 1789 bachelor and masters students at Technical Faculty in Cacak, Serbia. These students are users of Moodle learning management system. System with courses is available for overview at the address [7].

3.2 Tool

A tool that is used for application of clustering is called Tanagra 1.4 [8]. This tool provides a vast number of analysis possibilities in the data mining research domain. Module that relates to clustering is used for the needs of this specific research.

3.3 Procedure

Before the beginning of clustering process, pre-processing on log files has to be conducted. Raw log files contain data that has to be normalized. After pre-processing, log

files contain data that is staggered in the following columns: year, month, day, hour, minute, module, activity, and course.

	A	B	C	D	E	F	G
1	Month	Day	Hour	Minut	Modul	Activity	Course
2	04	17	01	45	course	view	1
3	04	16	01	43	forum	view discussion	68
4	04	16	01	43	resource	view	65
5	04	16	01	41	course	view	68

Figure 1. The illustration of log file after pre-processing

After the data importation, entry and target parameters are being determined, and their choice is defined by the selection of a specific clustering method.

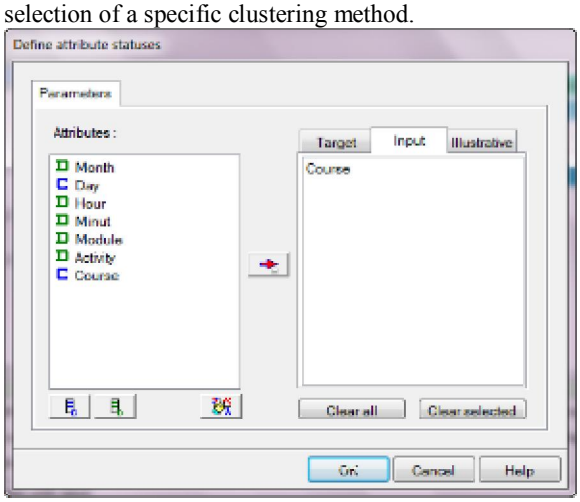

Figure 2: the selection of an entry and target parameters

After the selection of an entry and target parameters, a clustering method is being chosen. In this, specific study, the comparation of a k-means and HAC clustering method. The selection of the above mentioned clustering types is given in the Figure 3.

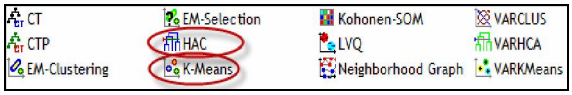

Figure 3: Illustration of clustering types within Tanagra program

After the selection of clustering type, a selected clustering type is being conducted by selecting the option Execute. Apart from that, Group Characterization is also done in order to present differences between groups.

4. RESULTS

4.1 Results that are obtained with the application of k-means clustering

K-Means parameters	
Clusters	4
Max Iteration	10
Trials	5
Distance normalization	variance
Average computation	McQueen
Seed random generator	Standard

Figure 4: Illustration of k-means clustering parameters

Within this figure data about parameters of k-means clustering are given. These data include cluster number, maximal number of iterations, distance normalization, average computation and seed random generator. Cluster number is 4, maximal number of iteration is 10, and number of attempts is 5.

Cluster size and WSS

Clusters	4		
Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	4570	99.5483
cluster n°2	c_kmeans_2	8170	140.5143
cluster n°3	c_kmeans_3	16646	4.9063
cluster n°4	c_kmeans_4	36090	730.1378

Figure 5: Cluster size and WSS

Figure 5 presents cluster size as well as the vector of length which contains the within sum of squares foe each cluster. According to figure 5, the smallest cluster is cluster 1, and the biggest one is cluster 4.

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4
Course	12.377462	34.362913	1.056891	63.835827

Figure 6: Illustration of cluster centroids

The illustration of cluster centroids in relation to attribute course is presented in Figure 6.

Results

Description of "Month"

Month=04				Month=02				Month=05				Month=10				Month=03			
Examples			[50.6 %] 33126	Examples			[3.3 %] 2134	Examples			[45.2 %] 29568	Examples			[0.3 %] 223	Examples			[0.6 %] 425
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
Course	-0.06	38.96 (28.09)	40.61 (27.69)	Course	0.17	45.24 (23.84)	40.61 (27.69)	Course	0.05	42.00 (27.29)	40.61 (27.69)	Course	0.55	55.90 (33.46)	40.61 (27.69)	Course	0.00	40.70 (29.24)	40.61 (27.69)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			

Figure 7: Illustration of results that are obtained with k-means clustering

Results that are obtained with the application of k-means clustering are given in the Figure 7. Description of the following months is given. Those months are February, March, April, May and October. For each month the percentage amount is given for the present instance, and the standard deviation for a class that is continuous attribute.

4.2 Results obtained with the application of HAC clustering

Clustering results

Clusters	4	
Cluster	Description	Size
cluster n°1	c_hac_1	33126
cluster n°2	c_hac_2	223
cluster n°3	c_hac_3	2134
cluster n°4	c_hac_4	29993

Figure 8: Initial results of HAC clustering

Figure 8 presents application of clusters and their sizes. According to Figure 8, the smallest cluster is cluster 2, and the biggest is cluster 1.

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4
Course	38.959518	55.896861	45.244142	41.981429

Figure 9: Illustration of cluster centroids

Figure 9 presents cluster centroids in relation to attribute course.

Figure 10 presents HAC dendrogram that is obtained with the application of HAC clustering on baseline data. Set of embedded clusters is organized with the help of a tree. Based on figure and data that are obtained in a program Tanagra, clusters that resemble the most are the ones that relate to months 3 and 5, following (3, 5) and 2, then, (3, 5, 2) and 10, and at the very end (3, 5, 2, 10) and 4.

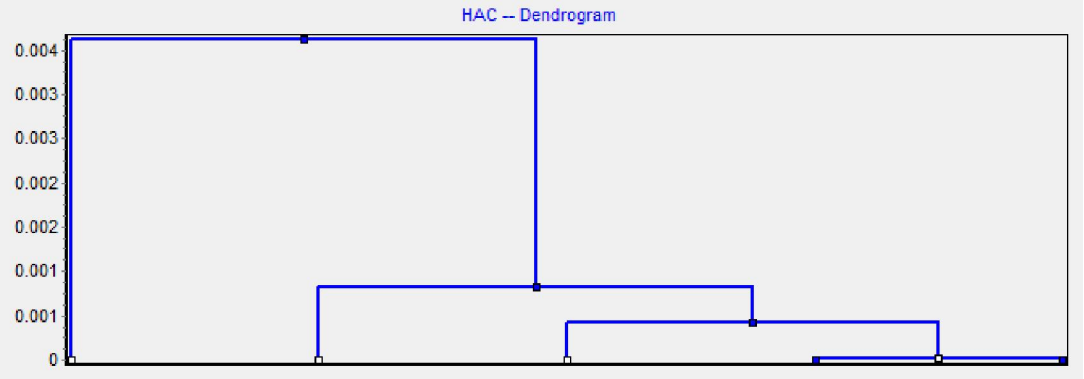

Figure 10: Illustration of dendrogram obtained by the HAC clustering method

5. DISCUSSION

Having in mind figures given in the Chapter Results, one can conclude following things. Based on these results we can conclude that K means and HAC clustering can be used to analyse user behaviour patterns. These conclusions relate to the determination of differences between k-means and HAC clustering in the analysis of users` patterns of behaviour.

As it can be seen, in figure 4 and 8, the number of clusters is the same. The only difference is that in the method of k-means clustering the number has to be stated, whilst in HAC clustering, it isn't recommendable to give any assumptions about the number of clusters.

According to figures 6 and 9, centroids that are different for k-means and HAC clustering are defined for the same attribute. That is the thing that indicates different

algorithm for centroids choosing in these two clustering methods. In both of these methods determination of centroids is conducted with the assistance of repeat/until loop. However, HAC clustering in being updated by the resemblance matrix and k-means clustering is being recomputed for centroid for each cluster in every step. There is one more difference between these two clustering types, and it relates to dendrogram formatting. Dendrogram formatting can be done in HAC clustering and it is presented in Figure 10. Graphic illustration that is given in Figure 10, enables better perception of clusters that are organised with the assistance of a tree. Unlike dendrogram within HAC clustering, data about clusters are presented in figure 7 in a form of a table.

Both of the mentioned methods are found to be very useful in the users` profile analysis, where log filer records are grouped in clusters. When analysing users` profiles it is essential to choose clustering method according to specific research demands, a way of results obtention and selection of the number of clusters. The following study relates to the analysis of the other clustering types in the analysis of users` behaviour patterns.

REFERENCES

[1] Wang, W and O. Zaiane, *Clustering Web Sessions by Sequence Alignment*, Retrieved from: http://webdocs.cs.ualberta.ca/~zaiane/postscript/dexa2002.pdf, 2002.

[2] Kazanidis, I., Valsamidis, S., Theodosiu, T. *Proposed framework for data mining in e-learning: The case of open e-class*, retrived from: http://utopia.duth.gr/~skontog/papers/iadis2009.pdf, 2009.

[3] Hend D., Mannila H., Smyth P.: *Principles of Data Mining*, e-book, retrived from: http://books.google.com/books?hl=en&lr=&id=SdZ-bhVhZGYC&oi=fnd&pg=PR17&dq=Hend+D.,+Mannila+H.,+Smyth+P.:+Principles+of+Data+Mining&ots=yvT8zjstk1&sig=3fZxLzXOZR-mqJxAcGV6NTZTlbE#v=onepage&q&f=false

[4] MacQueen, J. *Some methods for classification and analysis of multivariate observations.* In Le Cam, L. M. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281-297, Berkeley, Califonia. University of California Press, 1967.

[5] *HAC clustering*, retrieved from: http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html, 2009.

[6] Ellis, R: *A Field Guide to Learning Management Systems,* ASTD Learning Circuits, from http://www.astd.org/NR/rdonlyres/12ECDB99-3B91-403E-9B15 7E597444645D/23395/LMS_fieldguide_20091.pdf, 2009.

[7] LMS Moodle on the Technical Faculty: http://itlab.tfc.kg.ac.rs/moodle/

[8] Software Tanagra, Retrieved from: http://eric.univlyon2.fr/~ricco/tanagra/en/tanagra.html