RESEARCH ARTICLE OPEN ACCESS

Use of Unsupervised Clustering to Characterize Graduate Students Profiles based on Educational Outcomes

Lotfi NAJDI[1], Dr. Brahim ER-RAHA[2]
[1](GMES Laboratory, ENSA, Ibn Zohr UniversityAgadir, Morocco)
[2] (GMES Laboratory, ENSA, Ibn Zohr University Agadir, Morocco)

----------**********----------

Abstract:

The identification of profiles and typologies of students plays interesting role in adapting educational approaches and improving academic outcomes. It is with this perspective that we will show, in this work, how unsupervised learning techniques can be applied to educational data for the extraction of typologies and student profiles. We will also implemented this Clustering analysis using K-means algorithm and R programming language, in order to identify homogeneous groups of students, according to their academic performance in combination with the length of studies of the bachelor program. The dataset used in this study consists of student's official grades for the six semesters and the final grade of the bachelor degree. The approach presented in this study will enrich the understanding of different learning characteristics of graduatestudents and could be used to adapt teaching approaches and strategies according to the identified student profiles.

Keywords **—Educational data mining, unsupervised learning, cluster analysis, K-means.**

----------**********----------

I. INTRODUCTION

During thelast decade wehave observed aphenomenalincrease in the amountof digitaldata accumulated intheMoroccan universitiesmanagement systems,due to theadoptionofinformation technologyto manage themonitoring process of students andteachingfrom registration totheawarding of the diploma. The challengethatuniversitiesface today, is transformingthesedataintouseful knowledgein order to supportdecision making. This challenge can be addressed through the use of knowledge discovery in databases or data mining in education called Educational Data Mining (EDM). EDM is concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students and the settings they learn in [1]. Educational data mining methods fall into the following general categories: prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment [2].

The aim of this work is to propose an integrated approach for the application of unsupervised learning techniques to university data with the aim of extracting clusters and profiles of graduate students. This approach implements K-means algorithm to identify homogeneous groups of students according to their academic performance in combination with the time to degree for the Whole bachelor program.

The rest of the paper is organized as follows: In Section 2, basic clustering techniques are introduced and related works in cluster analysis applied to educational setting are summarized. Section 3 is devoted to the methodology of this study. Section 4 gives a short description of data and tools used in this work. Section 5 presents the results in detail. Finally, this paper is concluded by a summary and an outlook for futures works.

II. CLUSTERING IN EDUCATIONAL DATA MINNG

Data mining methods may be categorized as either supervised or unsupervised. Clustering is the

most common unsupervised data mining method [3]. This method consists of assigning a set of observations into subsets called clusters, so that observations in the same cluster have some points in common [4]. The identifiedgroups mustsatisfytwo main conditions.

- Homogeneity: Elements of the same cluster are maximally close to each other;
- Separation: Data elements in separate clusters are maximally far apart from each other.

Clusteringmethodsare divided intofive main categories: hierarchicalmethods, partitioning methods, density-based methods, model-based clustering and grid-based methods [5].In this workwe willparticularlymake use ofclusteringbypartitioning, whichconsists of relocatingthe observationsby moving them from one clusterto another,from aninitial partition. This methodgenerally requiresthatthe number of clustersto beinformedin advance[6].

Clustering in the field of EDM can be used to group the students based on their similarity measures (marks, talents, practical knowledge in a particular field, family background) [7]. Groups of students can be created at several levels of granularity: schools could be clustered together (to explore similarities and differences between schools), students could be clustered together (to study similarities and differences between students), and student actions could be clustered together (to investigate patterns of behavior) [8].The application of various clustering algorithm has been applied in many cases to educational data set in diverse studies: K-means clustering algorithm has been used to discover interesting patterns that characterize the work of stronger and weaker students in collaboration tool for senior software development project[9]; Expectation Maximization, Hierarchical Clustering, and X-Means have been applied to determining students' learning behaviors by mining Moodle log data[10]; An approach based on Agglomerative hierarchical clustering has been suggested to model learner participation profiles in online discussion forums[11]; Latent Class Analysis has been compared to K-means Algorithms for Clustering Educational Digital Library Usage Data[12]; Clustering combined with classification and association analysis have been proposed for joint use in the mining of student's assessment data[13]. Clustering algorithms have been also proposed to group students according to their educational outcomes: K-means algorithm combined with factor analysis have been applied to discover the profiles of students from course evaluation data [14];K-means clustering with Decision tree have been used to improve Student's Academic Performance [15]; K-means clustering algorithm coupled with deterministic model has been described to evaluate the performance of students during specific semester [16]; Hierarchical clustering and k means clustering have been applied to characterize the students' academic performance[17].

In this work,the objective isto apply theclusteringbypartitioning,asunsupervisedddata miningtechnique forthe identification oftypologiesandprofilesof graduate studentsbasedon academicperformance andthelength of studies.

III. METHODOLOGY

The proposed approachforthe identification ofstudent profiles makes use ofknowledgediscovery techniquesfromeducational dataand integratesthe necessary stepsfor performingcluster analysis,beginning withdata preprocessingto the interpretation of identified clusters.This approach is describedas follows:

A. Data cleaning and pre-processing

This stage involves the implementation of necessary treatments for improving data quality and transformations required by the K-means clustering. The different proposed functions and services implemented by this approach are summarized as following:

1) *Missing values:*The first routine implemented by this work deal with missing values. These observations correspond to tuples that have no recorded value for several attributes. Replacement by attribute mean value was chosen for this pre-processing purpose.

2) *Outliers:* The second task is concerned with the detection and replacement of outliers by the nearest non suspect data. Outliers are defined as

observations which appear to be inconsistent with a set of data.

3) *Scaling data:* The thirdfunction madein this step isfeatures scaling. Scaling is a data transformationwhich seeks to restrictthevariation rangeof thenumeric attributes. The objective of implementing this routine is to not favor attributeswith the greatestareas ofvariation inthe clustering step.

4) *Features selection*: In terms of types of data considered for the clustering task, K-means is restricted to continuous values, so only numerical variables must be selected form the original dataset. The clustering algorithm is applied to scalled data.

B. Determining the appropriate parameter for the clustering task

The number of clusters is a priori unknown; hence the ELBOW method has been used to determine the optimal number of groups. The total within-cluster sum of square (WSS) measures the compactness of the clustering and we want it to be as small as possible. The algorithm of Elbow method is defined as follow:

1) Compute K-means different values of k.
2) For each iteration, apply K-means and calculate the total within-cluster sum of square (WSS)
3) Plot the curve of WSS according to the number of clusters k.
4) The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

C. Performing students clustering using K-means

The result of ELBOW method is passed as parameter K to the K-means clustering algorithm. This algorithm partitions the data into K clusters represented by their centers. The center of each cluster is calculated as the mean of all the instances belonging to that cluster. The algorithm K-means is performed as follows:

1) The dataset is partitioned randomly into K clusters.
2) For each data point: Calculate the distance from the data point to each cluster.
3) If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
4) Repeat the above step until a complete pass through all the data points, results in no data point moving from one cluster to another.

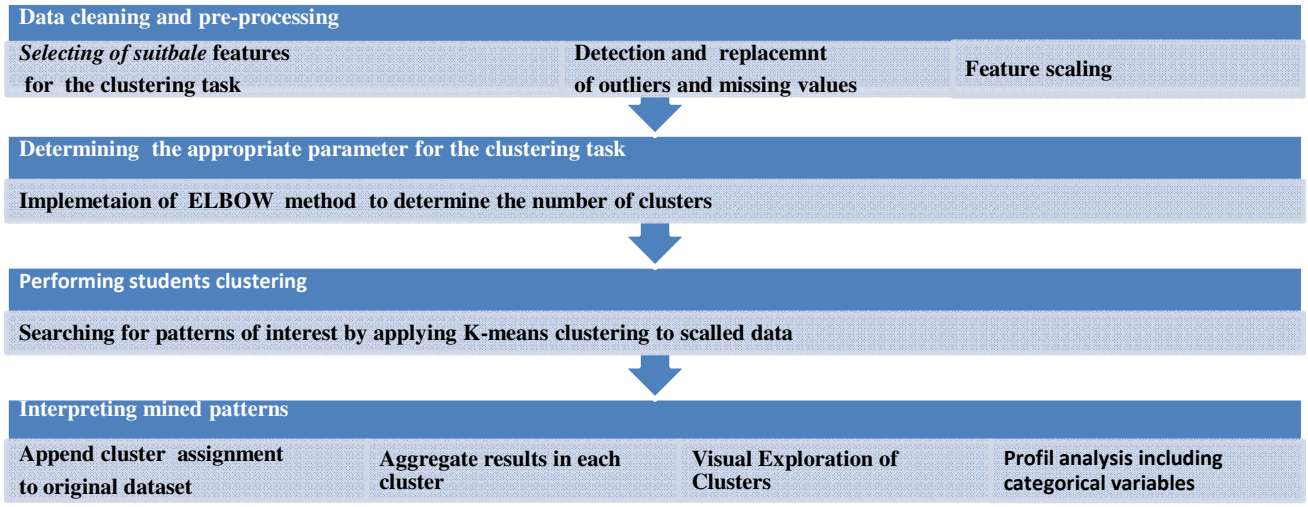


Figure 1: Steps Of the proposed Clustering approach

D. Interpreting mined patterns

After performing K-means, cluster assignment has been appended to original dataset that contains both numeric and categorical variables before the scaling processing.Characteristics of students in each cluster are examined by aggregating results and calculating the means of each of the variables across different clusters. In order to gain a deeper insight clustering results and understands the structures of the identified clusters,a visual exploratory data analysis of results have been performed. After forming student clusters, a profile analysis was carried out so as to examine the variation of other student characteristics in different student segments. These characteristics are variables that are not included in clustering task.

IV. DATA DESCRIPTION AND TOOLS USED FOR CLUSTERING

The original data set is obtained from the Faculty of Law, Economic and Social Sciences at University IBNZOHR.The clustering analysis is being performed on the basis of marks obtained by students of batch 2008-2012, in combination with the number of enrollment from the six semesters forming the bachelor degree program. The attributes selected are indicated in table1. This dataset contains 2600 records of students of four consecutive years.

All the steps of the clustering analysis, as described in section 3, were implemented with the help of R language.R is a programming language and data analysis software.R is used by the scientific and academic community in order to make sense of data by performing statistical analysis, data visualization, and predictive modeling. Data analysis in R can be done by writing scripts and functions in the R programming language. This language provides objects, operators and functions that make the process of exploring, modeling, and visualizing data a natural one.

Table 1. Data attributes and description

Attribute	Type	Description
SSG	Nominal	Secondary school grade (before enroll university program). Values(AB,P,B,TB)
CGPA	Real	The overall student average with the whole years of studies
MS1	Real	Student mark at semester 1
MS2	Real	Student mark at semester 2
MS3	Real	Student mark at semester 3
MS4	Real	Student mark at semester 4
MS5	Real	Student mark at semester 5
MS6	Real	Student mark at semester 6
NES1	integer	Student Number of Enrollment at semester 1
NES1	Integer	Student Number of Enrollment at semester 2
NES2	Integer	Student Number of Enrollment at semester 3
NES3	Integer	Student Number of Enrollment at semester 4
NES4	Integer	Student Number of Enrollment at semester 5
NES5	Integer	Student Number of Enrollment at semester 6
NES6	Integer	Student Number of Enrollment at semester 1
CGPA	Real	The overall student average with the whole years of studies
GBG	Nominal	Global grade of bachelor degree. Values(P,AB,TB)
NEB	Integer	Number of Enrollment with the whole years of studies in the bachelor program

V. EXPERIMENTAL RESULTS

A. BEST NUMBER OF CLUSTERS

In order to determine the optimal number of clusters, we iterate through wss array 10 times. For every iteration the intra-clusters sum of square is stored in wss. After we plot each iteration in order to display the elbow graph. As we can see from in Figure 2 the slope of the graph changes majorly in the third iteration, hence we consider the optimized number of cluster as three.

The result of elbow method is then passed as parameter K to the K-means clustering algorithm.

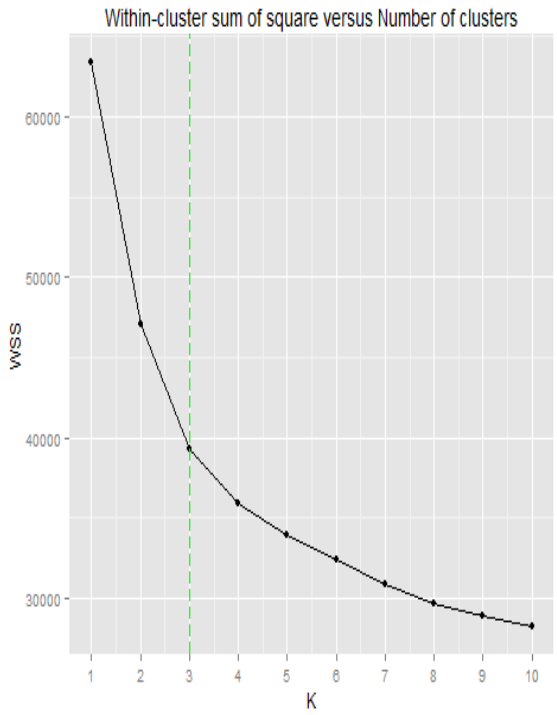


Figure 2: Assessing the optimal number of clusters using the Elbow method

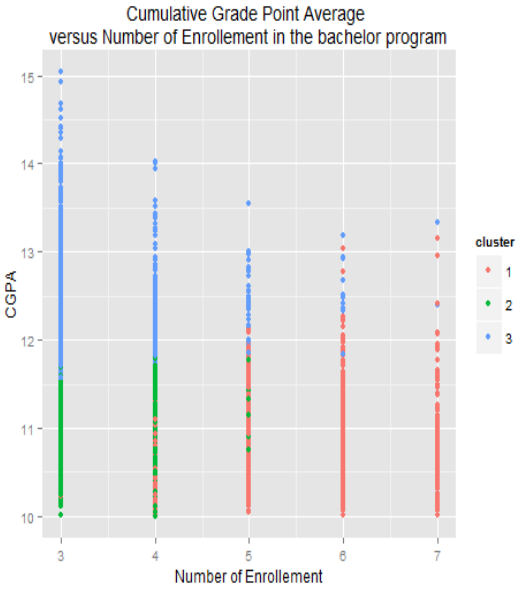
B. MEANING OF THE CLUSTERS

As the result of K-means clustering analysis, we would examine the means for each cluster on each dimension to assess how distinct the three clusters are.The summary of results (see Table 2), shows the means of each of the variables across each cluster. The first two rows show the count or number of observations in the cluster and the percentage in the population.

After this summary of results of cluster analysis, a visual data exploration has been done using the ggplot and ggparcoord R packages. For example the plot of the data for the cumulative Global Point Average CGPA versus the length of studies in the bachelor cycle (or Number of enrollments) with different clusters shown in different colors is presented in Figure 3

Table 2. Clustering result summary

	Cluster 1	Cluster 2	Cluster 3
Size	728	1300	572
Percent	28%	50%	22%
NES1	2.60	1.39	1.21
NES2	2.54	1.36	1.18
NES3	2.64	1.38	1.16
NES4	2.77	1.42	1.12
NES5	2.33	1.34	1.12
NES5	2.24	1.34	1.11
Bachelor cycle	5.32	3.57	3.25
MS1	10.56	10.82	11.83
MS2	10.61	10.82	12.09
MS3	10.70	10.87	12.10
MS4	10.58	10.86	12.13
MS5	10.63	10.91	11.99
MS5	11.48	11.86	13.07
CGPA	10.76	11.02	12.20



.

Figure 3: Cumulative grade point average versus the number of enrollment

Another solution that could be very promising in this exploratory data analysis is to make a multivariate Data Visualization. Using the ggparcoord packages the following two parallel-coordinate plots have been made: The first one show the mean of the length of studies from semester 1 to semester 6 among the tree identified

clusters (see Figure 4); the second plot shows the mean of the obtained marks from semester 1 to semester 6 among the tree identified clusters (see Figure 5).

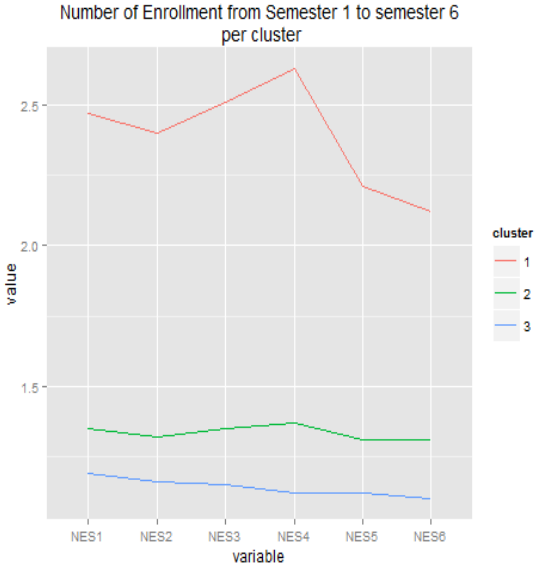


Figure 4: Number of Enrollment from Semester 1 to semester 6 per cluster

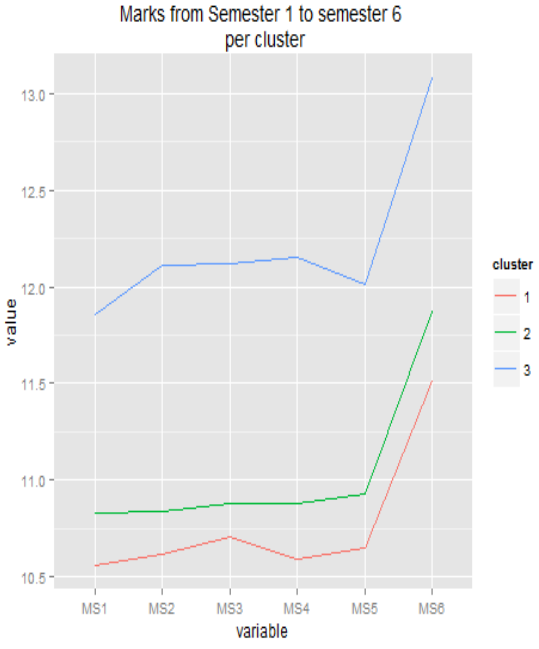


Figure 5: Student's marks from Semester 1 to semester 6 per cluster

As we can see in Figure 4 and Figure 5, Student mark and the number of enrolment at the first semester are the most important variables in separating graduate student's groups according to educational achievement.

Table 3. Characteristics of Students Group

Cluster	Description
Cluster 1	This group consists of 28% of the population. It can be described as the group of weak student. This group shows a very poor in boot index for booth educational outcomes and lengths of study. In this group almost students have a poor CGPA level for all semesters with a highest number of enrollments.
Cluster 2	This group consist the highest number of students in the group. It represents 50 % of the graduate students. This group has a good level of lengths of study with medium outcomes.
Cluster 3	In comparing with other groups, Cluster 3 can be considered as a group of most successful students. It represents 22% of the graduate students. This group has very good index for booth educational outcomes and lengths of study.

Based on student's segments described in table 3, the profiles of students including categorical variables such as educational background and high school types are determined. The majority of students from Cluster 1 (group of weak students) and Cluster 2 (group of medium level students), belong to grade P (see Table 4) corresponding to medium gradation performance level of graduation. However the difference between the two groups is that students from Cluster 1 take more time in the bachelor program before getting graduated.

Table 4. Cross Tabulation of Bachelor Grade and Student Clusters

		Cluster 1	Cluster 2	Cluster 3
Bachelor Grade	P	99,32%	100%	43,87%
	AB	0,68%	0	55,09%
	TB	0	0	1,04%

Table 5. Cross Tabulation of Secondary school grade and Student Clusters

		Cluster 1	Cluster 2	Cluster 3
SSG	**P**	77%	74,63%	54,62%
	AB	22,50%	23,94%	38,68%

	B	0,50%	1,33%	6,51%

Students from Cluster 3 (group of most successful students) are distributed in a balanced way, between grade P and grade AB. As we can see in table 5, all identified clusters contain students with different Secondary school grade. But as the percentage of students who belongs to AB grade increase (before they enroll university program), students become more likely to belong to the Cluster 2 and Cluster 3.

VI. DISCUSSION AND CONCLUSION

Through this work, Clustering analysis as an EDM method has been discussed and implemented using K-means algorithm and R language, in order to group graduate students from the bachelor program The presented approach has successfully identified three groups of students, with according similar learning behavior in term of academic achievement and duration of studies for bath graduation and each semester of the bachelor program. Therefore, this clustering analysis has produced relevant results and will assist universities decision makers to gain a deeper insight in students' learning characteristics, to monitor student's performance during each semester of the bachelor degree program and to adapt pedagogical approaches and strategies to improve the performance and the quality of education in Moroccan universities.

In order to make the clustering approach, described in this paper, more accessible for university decision maker, an interactive web-based tool for clustering student's data could be designed. This tool must be adapted to non-expert in data mining requirements (intuitive, easy to use with good possibility of visualization). This tool must allow the user to select variables and the desired number of clusters to be created interactively. In addition, it must give them the possibility to interact with the resulting visualization of the clustering analysis. This way, Clusters of similar students can be visualized and characterized immediately and the assessment and refinement of the obtained results could be done instantly by decision maker.

REFERENCES

[1] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 40, no. 6, pp. 601–618, Nov. 2010.

[2] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," JEDM-J. Educ. Data Min., vol. 1, no. 1, pp. 3–17, 2009.

[3] Data Mining and Predictive Analytics, 2 edition. Hoboken, New Jersey: Wiley, 2015.

[4] C. Romesburg, Cluster Analysis for Researchers. Lulu.com, 2004.

[5] J. Han and M. Kamber, Data Mining: Concepts and Techniques. 2000.

[6] L. Rokach and O. Maimon, "Clustering methods," in Data mining and knowledge discovery handbook, Springer, 2005, pp. 321–352.

[7] D. Neha and B. M. Vidyavathi, "A Survey on Applications of Data Mining using Clustering Techniques," Int. J. Comput. Appl., vol. 126, no. 2, 2015.

[8] R. Baker and others, "Data mining for education," Int. Encycl. Educ., vol. 7, pp. 112–118, 2010.

[9] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaïane, "Clustering and Sequential Pattern Mining of Online Collaborative Learning Data," IEEE Trans Knowl Data Eng, vol. 21, no. 6, pp. 759–772, Jun. 2009.

[10] A. Bovo, S. Sanchez, O. Héguy, and Y. Duthen, "Analysis of students clustering results based on Moodle log data," in 6th International Conference on Educational Data Mining-EDM 2013, 2013, p. pp–306.

[11] G. Cobo, D. García-Solórzano, J. A. Morán, E. Santamaría, C. Monzo, and J. Melenchón, "Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums," in Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 2012, pp. 248–251.

[12] N. Flann, B. Xu, M. Recker, X. Qi, and L. Ye, "Clustering Educational Digital Library Usage Data: A Comparison of Latent Class Analysis and K-means Algorithms," ResearchGate, vol. 5, no. 2, Aug. 2013.

[13] A. Banumathi and A. Pethalakshmi, "A novel approach for upgrading Indian education by using data mining techniques," in 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), 2012, pp. 1–5.

[14] O. Darcan and B. Badur, "Student Profiling on Academic Performance Using Cluster Analysis," J. E-Learn. High. Educ., pp. 1–8, Jan. 2012.

[15] M. Shovon, H. Islam, and M. Haque, "An Approach of Improving Students Academic Performance by using k means clustering algorithm and Decision tree," Int. J. Adv. Comput. Sci. Appl., vol. 3, no. 8, 2012.

[16] Rakesh Kumar Arora and Dharmendra Badal, "Evaluating Student's Performance Using K-means Clustering," Int. J. Comput. Sci. Technol., vol. 4, no. 2, 2013.

[17] N. NorSyazwaniRasid and N. Ahmad, "GROUPING STUDENTS ACADEMIC PERFORMANCE USING ONE-WAY CLUSTERING," Int. J. Sci. Commer. Humanit., vol. 2, 2014.