Student Profile Ergonomically Adapted to e-Learning. A Data Clustering and Statistical Analysis based Survey

LIANA STANCA, RAMONA LACUREZEANU, VASILE PAUL BRESFELEAN
Business Information Systems Department
Babes-Bolyai University
Teodor Mihali street, 58-60, Cluj-Napoca
ROMANIA
{liana.stanca, ramona.lacurezeanu, paul.bresfelean} @econ.ubbcluj.ro

IOANA POP
University of Agricultural Sciences and Veterinary Medicine
Manastur Street, 1-3, Cluj-Napoca
ROMANIA
popioana@usamvcluj.ro

Abstract: - In the present work we carried out a study over a 4 years period in order to develop a student profile that matches computer-assisted learning. In our opinion, much of the teaching-learning effort will be reduced if the forms of education that fit each individual can be correctly identified. The ergonomics of teaching / learning comprises the correct identification of the student profile so as to connect with the right method and tools for learning. In the process of student profile identification we used the statistic analysis, association rules, and the data mining clustering techniques based on the K-means algorithm.

Key-Words: - Association rules, Clustering, K-means, Ergonomics, Student profile, e-Learning

1 Introduction

Ergonomics in the teaching-learning method is given by the correct identification of the student profile so as to be associated to the appropriate learning technique. Over the years there has been a concerted effort to channel the teaching-learning processes on the personal learning pace of each participant; more specifically, in order to introduce learning tempos / differentiated levels, because each individual has his own pace of assimilation. Until the advent of the Internet this objective has been difficult to achieve. The Web and its new media have made possible the folding of the learning process to the personal pace of each participant in this process.

The Internet in the teaching-learning process acts as a substitute teacher and determined the development of e-Learning, lifelong learning, blended learning. But this substitute experience gained in the field does not fit every individual who wishes to study within the form of distance education also known as blended learning. In this context, we consider that most of those who accede to this type of education barely realize that it does not suit their own learning manner.

We considered that, by identifying a student profile that will optimally assimilate the information taught in full time or on-line and distance education (ODL) named e-Learning form, we would simplify/optimize the teaching-learning effort. The purpose of this research is to be able to offer a starting point in the correct orientation of the individual towards distance or full time education forms (in Romania). In the recent literature the theme was approached by several researchers, such as: [2], [4] [5]. They refer to the issue addressed in terms of the methodological character of multimedia instruments, of 2D and 3D tools and last but not least, in terms of virtual environment. They proffer the possibility to synthesize structured data and information to meet each individual's capacity to assimilate.

In [7] authors support the need of using ergonomic principles in the new eLearning systems. This expression refers to ergonomic principles, communication and use of knowledge. Such systems will be aligned to the requirements and user profiles. In another paper [9], the authors discussed the need for ergonomic research in terms of a WBT system from two perspectives: use of content and performance tasks.

Based on these works and taking into account the incursion of new tools devoted to the learning process we structure this article as follows: in the first part we present the reasons and purpose of the paper; in the second part, we perform the statistical and data mining analysis; in the conclusions we penciled a student profile that suits e-Learning best.

2 Research Objectives

We initiated the research based on the following hypotheses:

1. Students following their own learning system/pace fit comfortably on computer assisted teaching/learning.
2. Students following a traditional or blended learning will not assimilate acceptably in computer assisted teaching/learning.

The premises of this study were to identify the students' profile who can properly develop their knowledge in a virtual environment.

3 Methodology

The research is based on the structured interview method [1] using the questionnaire tool. The questionnaire was used to identify the student profile that fits principally in a computer assisted teaching-learning system. We also applied several data mining techniques, such as data clustering and classification learning.

In our study we completed three phases:

1. Exploratory Study
2. Empirical Study
3. Defining requirements/needs.

The purpose of the exploratory study was as following: to challenge the comments, opinions, suggestions and the idea developed in other studies by practitioners and theoreticians. The study focused on gathering more ideas and suggestions of many experts who were encouraged to more freely express their ideas, opinions. At the end of this phase we developed a questionnaire to be completed during an oral interview. Responses collected and reformulated by the specialist were concentrated in the questionnaire‘s items. We then applied the questionnaire in the empirical study phase.

The questionnaire was applied at first in 2006 on a sample of 41 persons (15 gentlemen and 26 ladies), willing to attend the e-Learning. Within this form of education, the teacher–student communication is made by means of web resources. Consequently, the first important information for the development in good conditions of the teaching/learning process was to determine if the Internet (in a time was it considered a luxury in our country) can be optimally accessed by the student. The surveyed sample could access this resource in the following manner:7% at work, 42% at Internet Cafes, 51% at home.

Note that half plus one of the respondents affirmed that they had Internet access at home; therefore we consider that the act of education would take place in good conditions. In this context, we divided the sample into two groups:

Group A: those who completed high school in the period 1992-2004 (14 respondents).

Group B: those who completed high school in 2005-2006 (27 respondents).

The reasons for choosing the year 2004 as a reference split, were the following:

1. Internet access this year in Romania was extended to the large community by reducing prices and expansion of several companies. Until then, the Internet could be accessed mainly from public institutions and Internet cafés (a glowing business at the time).
2. Until 2004, Informatics courses had been taught in few high schools in the country. That year was the starting point in which Informatics was widely introduced in high schools, for both science and humanistic classes.

Group A characteristics:

- Science profile for the graduated high school: economic specialization (57%), informatics specialization (43%).
- 43% of them have Internet access at home, 43% at Internet cafes and 14% work.
- Within this group, 14% use the computer for access to information, 36% for communication and the remaining 50% do not use (or barely use) the computer.
- Some of them (43%) had used the Internet in computer-assisted learning, and the rest had not used it at all. Note that the reason for the orientation of 92% respondents in a computer-assisted education is the possibility of combining work with training. Only 8% want to follow this form of education to be able to assimilate information at an own pace.
- In this group, 14% of the respondents are the followers of classical education, 42% prefer the computer assisted education, and the remaining fancy a combination of tradition and computer, as it follows: □

  - 28% want classes to be held exclusively on-line and examination to take place in a traditional manner □

- 14% desire that both class courses and exams to take place on-line.

Group B characteristics:

- Science profile for the graduated high school: economic specialization (66%), informatics specialization (33%).
- 56% of them have Internet access at home, 40% at Internet cafes and 4% work.
- Within this group, 41% use the computer for access to information, 50% for communication and the remaining 9% offered an indecisive answer.
- The majority of them (62%) had used the Internet in computer-assisted learning, and the rest had not used it or were inconclusive. Note that the reason for the orientation of 81% respondents in a computer-assisted education is the possibility of combining work with training. 10% of them would chose this form of education due to financial reasons, while the rest (9%) because want to be able to assimilate information at an own pace.
- In this group, 18% of the respondents are the followers of classical education, 38% prefer the computer assisted education, and the remaining fancy a combination of tradition and computer, as it follows: □

- 26% want classes to be held exclusively on-line and examination to take place in a traditional manner □

- 19% desire traditional classes, but on-line exams.

3.1 Data clustering identification of student profile. K-means algorithm

Data clustering [10] represents the organization of a collection of patterns (frequently represented as a vector of measurements, or a point in a multidimensional space) into groups founded on similarity. K-means represents a traditional clustering technique consisting in the following steps [6]:

- first, specify in advance how many clusters are being required: parameter k.
- k points are chosen at random as cluster centers. All the instances are allocated to their closest cluster center in accordance with the ordinary Euclidean distance metric.
- the centroid, or mean, of the instances in each cluster is calculated—this is the "means" part. These centroids are taken to be new center values for their respective clusters.
- finally, the whole process is repeated with the new cluster centers. The process continues until the same points are assigned to each cluster in consecutive rounds.

The k-means algorithm purpose is to minimize an objective function – a squared error function, which is an indicator of the distance of the n data points from their respective cluster centers [11]:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

where $||x_i^{(j)}-c_j||^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster centers.

In our research of building a student profile that will optimally assimilate and comprehend best the information taught in traditional or ODL education, we divided the students into 3 clusters:

1. first cluster - students who prefer their own system/pace of learning (Table 1)
2. second cluster - who prefer traditional education
3. third cluster - who prefer the combined forms of education.

Table 1. Students who prefer their own system/pace of learning. Cluster centroids

	High school type	Specialization	Internet access mostly	Used before Internet in education	Course Preference	Exam preference
Group A 1st Cluster	Science	Informatics	Internet cafés	Yes- in learning at a rate of approx. 15%	on-line	on-line
Group B 1st Cluster	Science	Informatics	At home	Yes- in learning at a rate between 15%-50%	on-line	on-line

Some noteworthy conclusions of the survey are: 50% of those surveyed precisely indicated the preferred form of education with the aim of meeting their aspirations.

Another outcome of this study is that there were a small number of respondents who correctly associated e-Learning with the possibility of adapting the teaching/learning act to their own pace. Their profile is: science high school graduate, medium knowledge of Informatics, Internet and computer use in education, follower of full computer assisted learning. In the studied sample this rule had 7% support, 66% confidence.

For 2006, the minimum support set at 7% with 66% confidence was achieved. Our study was extended to the next period and according to the results presented below; this rule was confirmed to be valid.

- 2007 - the questionnaire was applied to a sample of 84 people [40% support, 66% confidence]
- 2008 - the questionnaire was applied to a sample of 84 people [50% support, confidence 66.7%]
- 2009 - the questionnaire was applied to a sample of 84 people [74% support, 85% confidence]
- 2010 - the questionnaire was applied to a sample of 84 people [80% support, 98% confidence].

As a result, we may affirm that while in 2006 it did not seem a compelling association rule [6] (between the high school graduate profile, computer use in high school learning environments, and computer knowledge) that would lead individuals to choose computer-assisted learning, after the analysis performed over several years, it proved to be a serious rule of association and identification of the Romanian student profile that appeals to e-Learning. There was no need for any modification of the questionnaire because all items have been validated.

3.2 Statistical analysis

In order to estimate the reliability of the questionnaire, based on the data collected, we used the following statistical analysis:

- *split-half* method – calculating the fidelity coefficient of the entire test, we obtained the value r = 0.82, and a confidence level p <0.001;
- analysis methods and *internal consistency* – *Cronbach α* internal consistency coefficient (0.80) was calculated, which indicates the unitary structure of the instrument used.
- *test-retest* method – it was performed after 1 year, respectively 4, r = 0.76, p <0.01, which indicates the questionnaire's stability in time.

Starting from these results we focused on the activities of those who wish to develop their learning in their own pace, on the Website pages. The aim was to follow the pages these users utilize in order to create a web application to deliver customized content, which is able to continuously adapt. Features of such an application are determined by its ability to anticipate users' needs and provide with information and content in the desired shape.

An adaptive application requires the ability to reorganize itself based on users' precedent behavior, so as to provide them with personalized information. According to the basic principles of teaching-learning act, the natural order of the pages on the site would be: courses materials, practical examples, bibliography and homework solving issues. Starting from this idea, we followed the students' behavior during the semester on our site, and the results can be seen in Fig. 1.

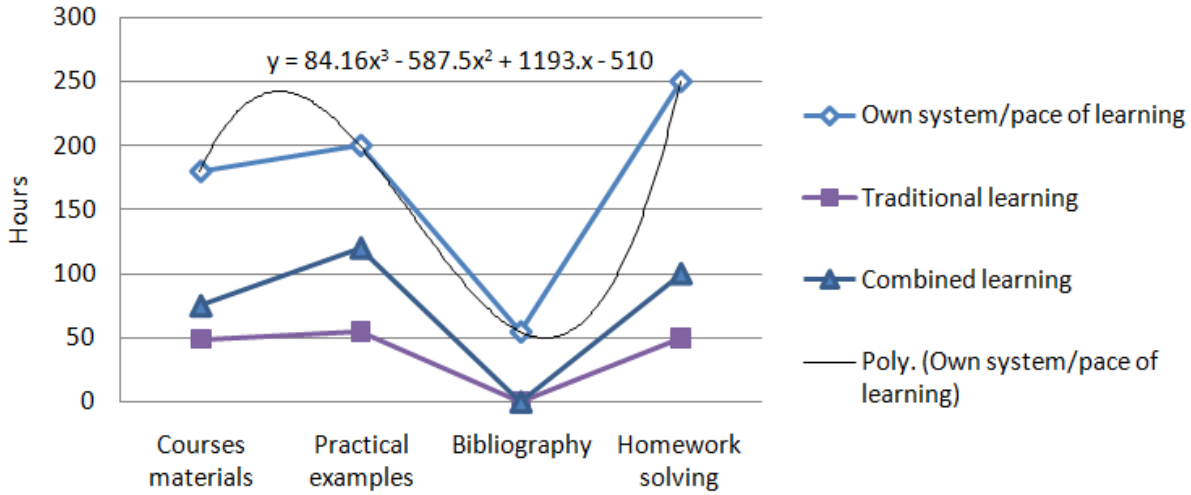

Fig. 1 Students' behavior on our website during the semester

3.3 Coefficients of correlation/determination

In our study we also examined whether there is a correlation between the number of access hours to pages according to basic principles of teaching-learning act. After calculating the Bravais-Pearson correlation coefficient ρ_{BP} and the coefficient of determination R^2 (Table 2),we reached the following conclusion:

1st Cluster
ρ_{BP} = 0.75 - means a good correlation with high degree of association [3], as interpreted by Colton (1974), between the hours of pages access and the studio group.
R^2 = 0.55 - means that 55 of 100 students are identified by the number of hours spent on the site.
2nd Cluster
ρ_{BP} = 0.37 - weak correlation with a low association degree [3], as interpreted by Colton (1974).
R^2 = 0.14 - means only 14 of 100 students are identified by the number of hours spent on the site.
3rd Cluster
ρ_{BP} = 0.50 - good correlation with a medium association degree [3], as interpreted by Colton (1974).
R^2 = 0.25 - only 25 of 100 students are identified by the number of hours spent on the site.

Table 2. Coefficients per cluster

	Bravais-Pearson correlation coefficient ρ_{BP}	**Coefficient of determination R^2**
1st cluster	0.75	0.55
2nd cluster	0.37	0.14
3rd cluster	0.50	0.25

The Bravais-Pearson correlation coefficient ρ_{BP} [12] represents an appropriate measure of association when n couples of continuous data ((y_i, x_i) with i=1,2,…n), collected on the same experimental unit, follow a bivariate normal distribution. In this case the relationship that can be postulated is the linear one [12].

The coefficient of determination R^2 [13] represents the proportion of variability in a data set that is accounted for by the statistical model. It evaluates how well future outcomes are expected to be predicted by the model, and is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information [13].

For the reason that in the number of access hours to pages P-value is 0.03 <0.05, the analysis performed on the number of hours is statistically significant, so it may be that the number of hours spent on the e-Learning site is determined by a linear regression model in the studied groups.

Founded on the access algorithms (Appendix I,II,III), it results the first cluster's algorithm is the one that respects the principles of e-Learning. The remaining groups ought to pursue other forms of education that suits them in order to assimilate as much information.

4 Conclusions

Based on the statistics results of the study we can affirm that the determination of a student profile that will assimilate and comprehend best the information in e-Learning - is prepared on association rules between the high school graduated specialization, the use of computer in the learning process in high school environments and the level of computer knowledge. Following our research carried out over the years, we have shown that it is a serious rule association and identification of the Romanian student profile that appeals to a form of e-Learning. For these individuals, the number of hours spent on the web site meets a linear regression model with their profile; and the algorithm for accessing pages is in accordance with principles of the teaching/learning act.

On this foundation we could correctly identify the profile of candidates for e-Learning; as for the rest of the students we could recommend other types of education. Our research will continue on the segment of students diverted to other forms of education to determine if they selected the best way to extend their knowledge.

Acknowledgements

This paper was supported by grant IDEI 2596, coordinated by Ramona Lacurezeanu.

References:

[1] Shahzard K., Elias M. , Johannesson P., Requirements for a business process model repository: A stakeholders' perspective, BIS, 158-170, LNBIP47(2010)
[2] Hoffmann H., Schirra R., Westner P., Meinken K. and Dangelmaieri M., Teach: Ergonomic Evaluation Using Avatars in Immersive

Environments, LNCS, Volume 4554/2007, 365-373
[3] Drugan T., Achimas A., Tigan S., Biostatistica, Editura SRIMA,Cluj-Napoca, 2005
[4] Rebelo F., Filgueiras E., Effectiveness of Multimedia Systems in Children's Education, LNCS, Volume 4566/2007, 274-283,
[5] Cuayáhuitl H., Dethlefs N., Frommberger L., Richter K.-F., Bateman J., Generating Adaptive Route Instructions Using Hierarchical Reinforcement Learning. In Spatial Cognition VII. Springer, LNAI 6222, 2010, 319-334
[6] Witten I.H., E. Frank, Hall M.A., Data mining : practical machine learning tools and techniques.—3rd ed. , Elsevier, 2011
[7] Millard D.E., Howard Y., Towards an Ergonomics of Knowledge Systems: Improving the Design of Technology Enhanced Learning , LNCS, Volume 6383/2010, 566-571
[8] ***http://www.seniorlearning.eu/site/Output usability and ergonomics report.pdf
[9] Gude D., Branahl E., Kawalek P., Prions A., Laurig W., Evaluation of a virtual reality-based ergonomics tutorial, http://www.ergonetz.de/integral/downloads/virtual-reality-ergonomics-elearning.pdf (2007)
[10] Jain A. K., Murty M. N., and Flynn P. J., Data clustering: a review. ACM Comput. Surv. 31, 3 September 1999, 264-323
[11] Matteucci M., A Tutorial on Clustering Algorithms, Politecnico di Milano, http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html
[12] Artusi R. , Verderio P. , Marubini E. , Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval, The International Journal of Biological Markers, Vol. 17 no. 2, 2002, 148-151
[13] Steel, R.G.D., Torrie, J.H., Principles and Procedures of Statistics, New York: McGraw-Hill, 1960

APPENDIX

APPENDIX I. Algorithm utilized to go through the web pages within the 1st cluster (who prefer an own system/pace of learning), is as following:

```
...subalgorihtm pseudocode
// codecourse is the unique identifier of the course, available on the server
select Codecourse, Name from materialsCourse
select Bibliography from cur_bibliography  where codecourse_bib=codecourse
select exemple_practice from course_example where codecourse_ex=codecourse
select homework from curs_homework where codecourse_homework=codecourse
// solve homework
insert intro homework ...
//closing connection
End for
```

APPENDIX II. Algorithm utilized to go through the web pages within the 2nd cluster (who prefer traditional education):

```
 ... subalgorihtm pseudocode
// codecourse is the unique identifier of the course, available on the server
select codecourse_homework, homework from course_homework
select exemple_practice from course_exemple where codecourse_ex=codecourse_homework
select Codecourse, Denumire from materialsCourse where codecourse_ex=codecourse_homework
// solve homework
insert intro homework ...
// closing connection
End for
```

APPENDIX III. Algorithm utilized to go through the web pages within the 3rd cluster (who prefer the combined forms of education):

```
 ... subalgorihtm pseudocode
//// codecourse is the unique identifier of the course, available on the server
select codecourse, Denumire from materialsCourse
select exemple_practice from course_exemple where  codecourse_ex=codecourse
select homework from course_homework where codecourse_homework=codecourse
// solve homework
insert intro homework ...
// closing connection
End for
```