E-Learning Process Characterization using data driven approaches

Silvia Rita Viola

Dipartimento di Ingegneria Informatica, Gestionale e dell' Automazione, M. Panti",
Universita' Politecnica delle Marche
60100 Ancona, Italy
sr.viola@gmail.com

Abstract. This paper summarizes the outcomes of different data driven analyses. The data used are authentic data coming from an European E-Learning Project. The paper is aimed at presenting the approaches used for learners' profiles characterization. Learners' profiles characterization is here intended with respect to the learning strategies used by learners from one side; from the other, with respect to different ways of non linear navigation. In both cases the focus is on the effectiveness of data driven approaches in detecting individual differences. It is shown that, in both cases, data driven approaches are able to detect such individual differences. Therefore, it can be concluded that data driven approaches are effective for learners' profiling, and that their employment can be beneficial for improving personalization of learning environments.

Keywords: Usage mining, learners' profiles, principal component analysis, frequent episodes discovery

1 Introduction

In recent years E-Learning field has become an opportunity not only for thinking the role of technologies for learning, but for re-thinking the way of conceiving the learning process itself. E-Learning field presents, as an element of difference with respect to traditional educational settings, the possibility to track users' actions during navigation in the Electronic Learning Environments (ELEs): these data are fully authentic and expressed on a numeral scale.

Therefore, data driven approaches should be experimented to analyze such data. Looking at the Literature, it can be seen that an increasing attention is being dedicated to this topic inside different research communities [14] (Data Mining, User Modelling and Intelligent Tutoring Systems, E-Learning). An interesting recent survey on the topic is provided by [13].

Such approaches have been experimented for handling data coming from ELEs for different purposes, such as for providing adaptivity [6], for intelligent monitoring [10], and for investigating the impact of a program [11].

The main benefits coming from the introduction of data driven approaches can be seen in improving flexibity and authenticity of the learners models and in improving the costs/benefits ratio. Therefore, the personalization of the learning environments should be improved. Personalization is here meant as the ability of the system to adapt itself to preferences and the ability of characterizing the evolution of the learning process according with a suitable kind of representation of such process; in this case a "personalized" model could represent both individuals and groups. Moreover, because a successfull learning process implies a change in behaviours, a particular attention should be devoted to the evolution of the learning process, intended as the changes showed by the learners during time.

This paper summarizes the outcomes of different data driven analyses on authentic data coming from an European E-Learning Project. The attention is focused on establishing if and how far data driven methods can be applied to the learning process to attain information on the learning strategies and on interaction of learners.

The paper is structured as follows: in Section 2, the materials will be outlined; in section 3, the methods will be presented. Subsequently (section 4) some results are given. The conclusions will end the work.

2 Materials

The dataset used here comes from the V Framework European WINDS Project. The WINDS Advanced Learning Environment (ALE) contains 22 courses at all; the sample is composed by a subset of students selected from students geographically distributed over Europe attending 8 Courses. The whole dataset is made by 358 non dummy sessions realized by 57 European students.

The WINDS Project is inspired by active, collaborative and "meaningful learning" inspired pedagogical approaches. Accordingly, it provides different kind of learning resources, devoted to promote an efficient learning in Design and Architecture. Near to traditional learning resources containing lessons or self- evaluation tests, resources supporting both active and collaborative learning are provided. Such resources are:

- "Cases", which are aimed at supporting active learning. Cases are resources in which students are invited to analyze a real-world design task, realized by a famous practitioner, which is explained and commented in details.
- "Concepts", and "Maps", that are aimed at supporting "meaningful learning" [3] experiences. Concepts are definitions of keywords occurring in paragraphs objects; both the number and the objects themselves change according to each selected paragraphs. "Maps" are concepts maps provided by links accessible by concept pages, conceived to give a non linear and interdisciplinary view of each matter. By means of them learners can "jump" to other concepts, or to other paragraphs.
- "Annotations", and "Discussions", which are aimed at supporting collaborative learning. Annotations are a kind of "electronic notebooks" in which learners can put their observations, that can be viewed by anyone else and collaboratively edited and enriched. "Discussions" are kind of forums accessible during navigation.

3 Outline of the approaches

The following subsections will give an insight of the different approaches used for characterizing learners' profiles with respect to learning strategies (subsection 3.1) and with respect to non linear navigation (subsection 3.2).

3.1 Characterizing individual preferences with respect to learning strategies

This analysis is devoted to answer the question: can data driven approaches give information on the learning strategies of learners, looking at how learners use the learning resources provided by the ALE?

At this step, the focus is learning strategies detection. Learning strategies are a part of response of the individual to the environment *stimuli*, and can be seen as cognitive tools helpful to the individual to perform a given task [12]. Therefore, learning strategies are developed (and thus changing) during interactions with the environment (and thus depending on the environment assets). Environment assets are here to be intended as the resources available to perform learning according to given pedagogical models. As a consequence, it can be assumed that the kinds of resources used by learners are as expressive of the environment assets of the learning environment.

In this analysis, the matrix of data contains individuals in rows and the kinds of learning objects in columns.

Principal Component Analysis - PCA, [7], which is a well-known statistical technique, has been used. It consists of finding a basis, that maximizes the the total variance of the dataset, on which data are subsequently projected. That basis is usually found by a Singular Value Decomposition [8] of the matrix of data. After the projection, a subset of linear combinations is selected to give a low dimensional representation. The cardinality of this subset can be at most equal to the rank of the data matrix. This low dimensional representation allow detecting some features, given by linear combinations of data, that are unobservable in the original data; furthermore, these features are uncorrelated each others.

Notice that session data are heterogeneous, both for length and for number belonging to each individual. Being PCA a variance based methods, heterogeneity needs to be addressed for avoiding affecting the results. Here, the epsilon-delta rank criterion to select the low dimensional feature space has been used [5]. The epsilon delta rank criterion looks at the differences in order of magnitude between subsequent singular values. These differences in order of magnitude are proportional to the variance expressed by each linear combination. According to this criterion, a low dimensional space made by 6 linear combinations has been selected.

Two views are considered:

- "profiles view", which is focused on detecting individual differences. Each row of the profiles view matrix contains a learner profile, while each colum contains a different kind of learning resource. A learner profile is given by the average number of the different learning resources.
- "sessions view", which is focused on profiling the evolution of individual differences during interaction. Eac row of the sessions view matrix contains a

session profile, while each colum contains a different kind of learning resource. A session profile is given by the number of different learning resources used within a session.

Moreover, a proximity measure is needed for detecting learners' profiles. Before choosing a measure, different measures, such as angle-based measures, as well as scattering measures, have been tested.

For profiles view, the proximity measure used for profiling is the ratio between the square 2-norm of the projection on each component, over the sum of the projection over all components of the selected model. Therefore, for i=1,..., 57 and for j=2,...,7, $r_{i,j} = \left\| y_{i,j} \right\|_2^2 / \left\| y_{.,j} \right\|_2^2$ while $y_{i,j}$ indicates the projection of the i-th student vector on the jth component, and $y_{.,j}$ indicates the jth component. This measure represents the part of the total length of each student vector represented on each component. As threshold, it has been used the proportion of the total length of each student vector over the maximum value of r achieved on each axis, that is $\max(r(y_{.,j}))$. $\max(r(y_{.,j}))$ has been divided into four equal parts, and these parts used as thresholds.

For sessions view, the average euclidean distance from the mean on data 2-normalized in row has been used. This measure is defined, for the i-th student, as $d_i = 1/k \sum_{i=1}^{k} d(ynorm_k^{(i)}, cnorm^{(i)})$, being k the whole number of sessions made the the i-th learner, $ynorm_k^{(i)}$ the k-th session of the i-th student after being normalized in row, $cnorm^{(i)}$ the mean of the sessions of the i-th student normalized in rows, and d the euclidean distance. Both the measures have been chosen after a comparison with other measures; in particular, d has been chosen looking at the insensitiveness with respect to the different number of sessions made by each learner.

3.2 Characterizing individual differences in non linear navigation

This analysis is devoted to answer the question: can data driven approaches characterize individual differences in the ways in which learners use the learning environment, especially looking at non-linear ways of navigating?

This analysis is performed on a subset made by 254 sessions, belonging to 53 learners, made by at least 10 items.

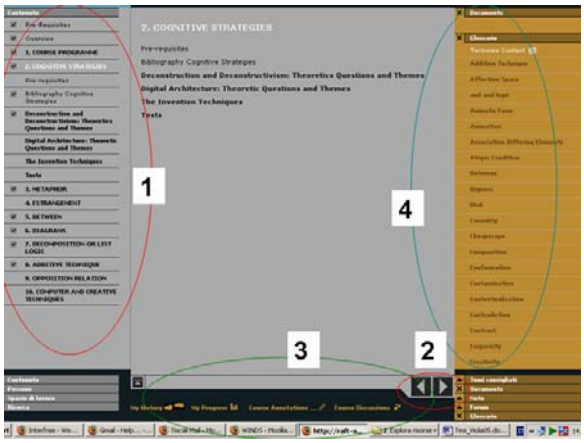
Frequent episodes discovery algorithms (FED) [9] have been used on sessions treated as sequence data. The frequencies of paths going from one kind of object to another are investigated. Here non sequential patterns have been mainly considered, that is, the ones in which there is no strict sequence of steps between objects.

Frequent discovery algorithms use a sliding window – of size *win* - over sequences to detect episodes, once they are defined, and returns the fraction of windows in which an episode occurs.

To detect the episodes to be searched, an analysis of the topology of the WINDS ALE (Figure 1) has been initially done. Such an anlysis leads to the conclusion that

two ways of construction of non sequential patterns are available: the first one using the left tree menu that allow to jump from one (traditional) page to another; the second one using concepts and maps to navigate non-linearly between the contents.

Fig. 1. The WINDS Advanced Learning Environment. 1. The left tree menu. 2. The next/previous buttons. 3. The collaborative objects. 4. Concepts.



Therefore, the episodes of interest have been defined from one side the ones involving the left tree menu to navigate non linearly between materials; from the other the ones involving concepts and maps to navigate non-linearly between the materials.

Belongs to the first group the episode made by the co-occurrence of a paragraph, followed by a unit, then followed by another paragraph (episode α). Belong to the second group the episodes made by the co-occurrence of a unit, followed by a concept, followed by a map (episode β); the episode made by the co-occurrence of a paragraph, followed by a concept, followed by a map (episode γ); the episode made by the co-occurrence of a concept, followed by a map, followed by a paragarph (episode δ) [16]. All these episodes have size 3. Therefore, *win* has also been set to 3 in order to avoid biased results.

Both the average occurrence of each episode per student, and the evolution in time of the occurrences have been investigated. The results are also cross-validated and explored using Mann-Kendall statistics, both for cross-validation and for achieving synthetic indexes of the evolution of each learner profile along time [17].

4 Results

The following subsections will report the results respectively for learning strategies (subsection 3.1) and for non linear navigation characteristics (subsection 3.2).

4.1 Learning strategies characterization

According to the epsilon-delta criterion, 6 components, the ones going from the second to the sixth, have been selected. The component are made by the right singular values af the matrix of data, that make the basis on which data are projected. In table 1 are collected these components for users view. In evidence are the absolute values greater than 10^{-1}.

Table 1. Factor loading on each Singular Value in Users View and in Sessions View. In italic the absolute values greater than 0.1.

Factor Loading on the Right Singular Values – Users view						
	2	**3**	**4**	**5**	**6**	**7**
units	*0,890*	*0,185*	-0,070	0,006	-0,000	0,022
paragraphs	*-0,379*	*-0,170*	0,016	-0,002	-0,000	-0,019
cases	*-0,239*	*0,966*	-0,024	-0,014	-0,031	-0,024
exercises	-0,017	0,006	-0,025	*-0,737*	*0,236*	*0,631*
concepts	0,055	0,049	*0,918*	0,082	*0,378*	-0,008
annotations	-0,028	0,014	0,024	*0,586*	-0,159	*0,753*
discussions	-0,004	0,006	0,010	*0,155*	-0,081	*0,177*
maps	0,033	-0,015	*0,386*	*-0,284*	*-0,875*	0,013

The second component shows cases (+) as represented in the opposite direction of paragraphs (-): the model was able to recognize the difference active objects/traditional objects.

On the third component concepts and maps are mainly represented: a cross validation with correlation coefficients showed that the objects were highly correlated (.72): this factor seems to reveal the hypertextual dimension of learning embedded in usage; according to frequencies of usage, the total variance is low (5.06).

The fourth component shows relationships between exercises and collaborative objects, in particular annotations: a manual verification showed that annotations were mainly used in order to express difficulties arising in exercises; very few annotations have been used in order to share knowledge. More uncertain is the relationship between exercises and maps: it can be supposed that maps were used as a kind of glossary during exercitations.

The fifth component shows again the relationships between maps usage and concepts on one side and exercises on the other, underlying another collaborative dimension of learning.

The sixth component points out again the relationship between exercises and collaborative objects.

It can be noticed that these unobservable dimensions reflect the pedagogical approaches inspiring the WINDS ALE (active learning, meaningful learning, collaborative learning). Accordingly to the unobservable dimensions in table 1, the learners profiles are arranged. The following results are drawn grouping learners according to *r*, considering 4 equally spaced thresholds ranging from the minimum to the maximum of *r* for each component.

Table 2. Learners' profiles given by r in users view.

Learners' profiles – Users view						
$r_{i,j}$	**2**	**3**	**4**	**5**	**6**	**7**
$\geq \frac{3}{4}\max(abs(r_{.,j}))$	49%	9%	1.5%	2%	2%	3.5%
$\geq \frac{1}{2}\max(abs(r_{.,j}))$	17.5%	17.5%	1.5%	0%	0%	0%
$\geq \frac{1}{4}\max(abs(r_{.,j}))$	21%	15.5%	1.5%	5%	0%	0%
$< \frac{1}{4}\max(abs(r_{.,j}))$	12.5%	58%	95.5%	93%	98%	96.5%

It can be seen that about 66% of sample shows consistent preferences (>1/2) for written traditional resources (component 2), such as paragraphs or units. This percentage decreases to 26.5% when objects supporting active learning are analyzed (component 3); for what concerns tools supporting hypertextual and collaborative tools (components 4, 5, 6 and 7) this the percentage decreases to 2%-3.5%.

Regarding sessions view, the focus has been put on the scattering of the points representing sessions, which indicates a preference for learning resources coherent with more than one latent dimension. The students have been grouped according to the value of d_i. The following table summarizes the results.

Table 3. Learners' profiles given by d in sessions view.

Learners' profiles – Sessions view			
Thresholds for *d*	**d < .5**	**d > .5, d < .8**	**d < .8**
Percentage of students	38.5%	37%	24.5%

The results of table 3 show that according to the most high threshold (.8), that expresses consistent variations in proportion of usage of each learning resource, only 24.5% of the learners utilize fully potentials of the ELE in order to create personalized routes. About 37% of profiles, the ones corresponding to the thresholds going between .5 and .8, shows moderate variations in usage of the various learning resources; students that use massively only few objects and occasionally the others belong to this group. The other students, that present at most variations of percentages of usage of a few learning resources, are about 38.5%.

In order to provide another verification, some profiles that presented very high and very low scattering measure were randomly selected and explored graphically [15]. In general, a low *d* correspond to a linear dependence pattern due to the rank deficiency of the submatrix belonging to that profile. This indicates that the profiles with a low *d* use in general only a subset of the learning resources provided by the learning environment. A high *d* indicates instead a profile that use all, or many of the learning resources of the learning environment. Moreover, a low *d* indicates a learner profile with a little variation of usage of different learning resources during different

sessions. A high *d* indicates a learner profile that show consistent variations of the usage of resources during different sessions. In particular, this characteristic is made more evident in the linear combinations that represent mostly collaborative or hypertextual objects.

From these results it can concluded that the data driven approach used here has detected individual differences in learning strategies according to the usage of different learning resources provided by a learning environment, and their evolution during time.

4.2 Characterization of differences in non linear navigation

Table 4 provides, for each episode, the mean number of occurrences. It can be seen that the differences between episode α from one side, and episodes β, γ and δ from the other, is always of one order of magnitude. Therefore, the usage of maps and concepts is much less frequent than the usage of hypertextual structure.

Table 4. Learners' profiles for non linear navigation episodes.

	Student Profiles			
episodes	**α**	**β**	**γ**	**δ**
mean	.097	.0015	.0022	.0031
profiles > mean	23	7	8	11
profiles < mean	30	46	45	42
max	.363	.0132	.0385	.0279
min	0	0	0	0

The learners profiles that show a preference for the usage of maps and concepts (episodes β, γ, δ) are clustered above all around a single course, while learner profiles that show a preference for episode α are more sparse. Furthermore, the preference for episode α seems to be in general mutually exclusive with the preference of one or more episodes β, γ or δ (only in four cases all the means of profiles are greater than the sample mean). Therefore, it seems that a latent variable, that is, the interaction with the teacher, enacts on learners' profiles. In particular it seems that the usage of complex objects, such as concepts maps, has to be *learned* and that teacher's influence is determinant.

According to these results, two set, one of them containing students that show profiles higher than the mean in episode α, the other containing students that show profiles higher than the mean in at least two of episodes between β, γ and δ have been selected, in both cases irrespective for the distance from the mean. The first group contains 19 students, the second one 7 students. Students that shows a mean profile higher in all episodes (4/53) have not been considered. The whole cardinality is 26. The analysis of the significance of the differences of the means of the two groups of students has been performed using chi-square test and *p* values at the level of significance .05. Results show that the difference is significant. In particular *p* value is near to 0 (less than 10^{-4}) and chi-square value is 36.116 on 3 d.o.f., 2 groups, 26

individuals. Therefore, the differences in usage of the two kind of non sequential patterns are statistically significant.

Furthermore, the evolution in time of these profiles has been investigated. For the analysis of the behaviour of the patterns within the two above mentioned groups – sessions have been grouped according to the step in which they have been realized, that is, all the first sessions (irrespective with the time in which have been realized) have been grouped together; all the second have been grouped and so on. The frequency of the four (α β γ and δ) episodes during the first six steps have been considered. The first six steps reach about 65% of the total number of sessions.

The results show that when all the episodes are nonzero, the two patterns belonging to the two groups behave in opposite ways during time. Moreover, episode α in the first group shows a slow increase, although not monotonically (while the other episodes in the same group are equal to zero). Eventually, episode α in the second group shows a slow decrease, although not monotonically (while all the other episodes for the same group are nonzero).All the other episodes do not show a clear trend [17].

To detect if there is a trend in the series, the Mann-Kendall test has been used. The Mann-Kendall test is a nonparametric test for detecting increasing or decreasing trends in time series made by at least 4 observations, and for testing for their significance (e.g. [2]). The Mann-Kendall statistics, referred as S, is calculated by comparing sequentially every observation in the serie to all the subsequent observations. An increasing trend is given by a positive S, while a decreasing trend is given by a negative S. The significance of S is tested against an absolute critical value corresponding to a given coefficient. The literature suggest to consider a coefficient greater than .20 significant [2]. With respect to a serie made by 6 observation, the critical value is 6 with α=.20.

For α episode in G1, a value S=+7 is obtained; for α episode in G2, a value S=-7 is obtained. These results show that the two series exhibit a trend and that this trend is opposite in the two cases. Moreover, being the critical value 6, the results can be considered significant in both cases.

From these results it can concluded that the data driven approach used here has detected individual differences in non linear navigation which are statistically significant. Moreover, it can be confirmed the hypothesis that these ways of navigating are learned, because they reinforce during time. Eventually, it can be further hypothesized that the influence of the teacher can be determinant for such a learning.

4 Conclusions and future work

In this paper the outcomes of different data driven approaches for learners' profiles characterization are summarized. Learners' profiles characterization is here investigated with respect to the learning strategies used by learners from one side; from the other, with respect to different ways of non linear navigation.

The results show that data driven approaches can be considered effective for learners' profiling as well as detecting the evolution of the profiles during time.

Therefore, the employment of such methods can be beneficial for improving personalization of learning environments.

Future work will deal with the investigation of the effectiveness of different data driven approaches, and with the comparison with the ones presented here.

References

1. Cherkassky, V. and Mulier, F.: Learning from data. Wiley Interscience (1998)
2. Conover, W. J.: Practical Nonparametric Statistics. New York (1971)
3. De Grassi, M., Giretti, A., and Natale, F.: Meaningful Learning in Web-Based Design Teaching Environments. In Proceeding of CELDA 05 Conference, Porto, Portugal, December 14-16, IADIS Press, pp. 333-342 (2005)
4. Ford, N. and Chen, S. Y. Individual Differences, Hypermedia Navigation and Learning: An Empirical Study. Journal of Educational Multimedia and Hypermedia. 9(4), 281-312 (2000)
5. Golub, G. H., Klema, V., Stewart, G. W.: Rank degeneracy and least squares problems. Technical Report Stan-CS-76-599, Standfor University (1976)
6. Graf S., and Kinshuk: Considering Learning Styles in Learning Management Systems: Investigating the Behavior of Students in an Online Course. Proceedings of the First IEEE International Workshop on Semantic Media Adaptation and Personalization (SMAP 06), pp. 25-30 (2006)
7. Jolliffe, T. I.: Principal Component Analysis. Springer (1986)
8. Kalman, D.: A Singularly Valuable Decomposition: the SVD of a Matrix. Preprint from College Mathematics Journal (2002)
9. Mannila, H., Toivonen, H. & Verkamo, A. I.: Discovery of Frequent Episodes in Event Sequences, Data Mining and Knowledge Discovery, 1:259-289 (1997)
10. Merceron, A., and Yacef, K.: TADA-Ed for Educational Data Mining. Interactive Multimedia Electronic Journal of Technology-Enhanced Learning. Available online at http://imej.wfu.edu/articles/2005/1/03/index.asp, accessed July 11, 2007 (2005)
11. Monk, D.: Using data mining for E-Learning Decision Making. Electronic Journal of E-Learning, 3(1):41-54 (2005)
12. Rinding, R., and Rayner, S.: Cognitive styles and learning strategies. David Fulton Publisher. (1998)
13. Romero, C., and Ventura, S. Eds: Data mining in E-Learning. Wit Press (2006)
14. Romero, C., and Ventura S.: Educational Data Mining. A survey from 1995 to 2005. Expert Systems with Applications, 33(1):135-146 (2007)
15. Viola, S. R., Giretti, A.& Leo, T: Discovering learning process patterns by multivariate analysis of usage frequencies data in e-learning courses. ICL 2005 Proceedings, Kassel University Press .(2005)
16. Viola, S. R., Giretti, A.& Leo, T.: Differences in meaningful learning strategies of navigation: an empirical model. ICALT 2006 Proceedings, IEEE Press, pp. 441-445 (2006)
17. Viola, S. R., Giretti A., and Leo T. (submitted): Detecting differences in "meaningful learning" behaviours and their evolution: a data-driven approach. Invited paper currently under review on Int. J. of Computing & Information Sciences