Communication and Resource Usage Analysis in Online Environments

An Integrated Social Network Analysis and Data Mining Perspective

Álvaro Figueira
CRACS / INESC TEC
University of Porto
Porto, Portugal
arf@dcc.fc.up.pt

***Abstract*—Predicting whether a student will pass or fail is one of the most important actions to take while giving lectures. Usually, the experienced teacher is able to detect problematic situations at early stages. However, this is only true for classes up to a hundred students. For bigger ones, automatic methods are needed. In this paper, we present a predictive system based on three criteria retrieved and computed from the logs of the learning management system. We built fast frugal decision trees to help predict and prevent student failures, using data retrieved from their resource usage patterns. Evaluation of the decision system shows that the system's accuracy is very high both in train and test phases, surpassing logistic regression and CART.**

Keywords*— Fail prediction, Data mining, Feature selection, Fast Frugal Decision trees, Moodle logs.

I. INTRODUCTION

Characterizing interactions between students that usually communicate through an online context is frequently not a simple task. We consider that the difficult achievement of analyzing online interactions is, in part, related to the ways we have of communicating beyond the written word. Considering a face-to-face context, we not only communicate by words, but also by facial/body expressions, intonation, and essentially all elements that complement messages, as well as individual personalities.

However, the method that we are introducing is based on the relational aspects of interactions rather than on its information or content. Our proposed system is able to detect patterns of the communication that students and tutors create in their online communications, as well as patterns for pedagogical resources' usage. Obtaining this kind of information at an early stage might then lead to a following analysis of message content.

We do recognize that, currently, students that use Learning Management Systems (LMS) have a diversity of pedagogical resources either directly placed in the system or accessible through a link. Nowadays, it is usual for the teacher to place lecture notes and slides in the LMS, as well as exercises, quizzes or even videos. The diverse spectrum of pedagogical material is then accessed by the students according to their needs or to the schedule provided by the teacher. Either way, ultimately, each student is responsible for accessing these materials and to use them when needed, eventually more than once. As consequence, in order to better understand the preparation level of the class, several questions arise: are there any typical usage patterns? Are there patterns that lead to better results? Which are they?

In these settings, we propose a methodology which, (i) collects data from interactions between class participants; (ii) as well as, between them and the available pedagogical resources placed by the teacher in the LMS; (iii) performs a thorough analysis in the collected data, in order to characterize the class, to predict results and, ultimately, to prevent ill-learning conditions for students.

Predicting future grades for students, particularly when done in early stages, has been a recent concern in the academia, as it can be used to give important advises to students based on "lessons" learned from past experiences. Recent research on learning analytics has taken different approaches for that goal [1,2].

In this work, we present an approach which is based on the analysis of the Moodle logs (similarly to the method adopted in [3]). This analysis is consolidated through a case study of a higher education course with the participation of more than 300 students. In the course – "Technical Communication" – the students had three small tests, and the activities of writing part of an article, assessing three other articles, creating slides and presenting them orally.

Our motivation for this research is to have an expert system based on previous experiences that can trigger alarms whenever the systems detects, with a high percentage of confidence, that a student is following a path will lead him to a failure, so that he/she can be helped to mitigate and solve the detected problem as early as possible, as proposed in [4].

We formulate our research question as: is it possible to use the Moodle logs, i.e. interaction patterns, to trigger such alarms?

The remaining of the article is structured as follows: in the next two sections, we explain how communication and resource usage can be modeled as a directed graph. In section iv we describe our experiment and the main features that led us to the resulting predicting model. In section v to vii we analyze the retrieved data from Moodle logs, create a predictive system and evaluate it, respectively. Finally, in section viii we draw our conclusions.

II. Communication as a Graph

Communication between course participants can be considered directed and target when it is a comment to someone's post, or a reply. In this case, we can draw an arrow from emitter node to the receiver node. On the other hand, if a particular message is not target to anyone in particular, it is not a reply, nor a comment we do not consider it as an explicit interaction. This definition allows us to create a directed graph whose vertices are people, and edges are exchanged messages.

Having the graph, we can withdraw important conclusions from the computed properties of the graph itself, the nodes and the edges, by the framing it into social network analysis theory.

In particular, we considered locality and centrality measures: the centrality degree of a node, the index of centrality of the net and, the density. Other parameters like the existence of sub-communities and their size, and the identification of sources and sinks of information, contribute to provide a more enlightened vision of the group. Hence, not only the whole network has particular metrics associated with it, but also every node on the graph. While the graph provides a first view on the communication pattern of the network participants, the computed metrics provide to the teacher a second layer of insights about the interactions.

This approach has been used in the last years to dig Moodle forums and representing the corresponding interactions by graphs in order to apply SNA and draw conclusions [5]. Therefore, despite being a very promising and fruitful research area, we do not explore that subject more in this article.

III. Resource Usage as a Graph

While it is important to understand how students interact with each other and with the teacher, it is also important to analyze how they use available online resources (at their disposal through the LMS). Our vision is that learners that interact with the same materials are expected to have similar academic behavior and, eventually, will also have a correlation in their final grades [1]. Therefore, valuable information can be gained from the way students interact with learning materials. Moreover, it is also particularly important to extract knowledge from the order in which the material is being used. Data mining algorithms have been and are widely applied to discover patterns on resource usage and, consequently for further assessment of their practices [2]. Resource usage has been proposed as a method for clustering learning objects in a way to reflect their semantic similarity [3]. We take an approach like the one taken by Ziebarth *et al* [4] when they intended to understand if students portray the same behavior in respect to resource usage during exam preparation. Our proposal goes further in the direction of predicting their final results from usage patterns [6][7].

A previous research work by [8], conducted under these settings was able to predict with a slightly above 60% of accuracy the final grades by using decisions trees. In our article, we address this hypothesis of predicting results based on the interaction patterns but, we also add a dynamic validating process comparing final predictions against the real classifications, in a 10-fold cross validation, and using the obtained feedback to fine-tune future predictions

To analyze learning resources, we data mined the Moodle log data, which provided information about learning resources being used by students and the respective time of use. We, then, detected the associations between resources that co-occur in the same online session, using a version of the *apriori* algorithm.

After this step, we identified the sequences of activities each student performs. This procedure led us to the creation of a sequence map of activities in respect to a particular resource. This map can be seen as a graph where the nodes are the resources and the edges are the ordering connectors. From this graph, we can compare the activity of each student with the activity of his class mates.

IV. The Experiment

We used the Moodle LMS during the period of one academic year, collecting all the interactions of students with the system, between students and the teacher, and between students and the digital pedagogical materials made available by the teacher in the LMS platform. As we used data from courses with more than 300 students, these data scale up to more than 75K records of interactions where we applied data mining tools.

Our strategy was to use three features to classify student's interactions in order to predict grades, but, ultimately to prevent student failures in the course. Therefore, our final goal was to predict if a student would pass or fail, by observing its behavior in respect to resources usage. This goal is an extension both in accuracy and precision from the work developed in [8].

We use Moodle logs to access information regarding students' daily interaction with Moodle, the actions performed, the resources used, and the sequence of using these resources. We retrieved the full log from February2015 until June2015. During this period, there taken more than 55K student interactions with the platform. From the logs, we understood that from the initial 332 students enrolled in the course, only 311 actually interacted with the platform.

In the next sub-sections, we describe the main three features, or criteria, that we used for classification.

A. Feature 1: number of accesses/records

Intuitively, we would expect that the more interactions the more dedication and interest on the subject would lead to better grades. On the other hand, it is also fair to expect that students with more interactions are the ones with more difficulties.

B. Feature 2: coverage of digitally provided learning material

A second feature is to know if the students have covered all the coursework, the available resources and the proposed activities in the platform. Intuitively, we would expect every student to access every lecture handout, all the provided slides, and enrolled in every proposed activity.

However, from an analysis of the logs, we discovered that this is not true. In fact, 76% is the average of accessed material in the platform. Therefore, we wanted to understand the impact that this factor would have in the final grade. To reach this goal we created a matrix where each row corresponds to a student and each column to the number of times a specific online resource/activity was accessed. Curiously, in this matrix we found many students with multiple accesses to the same activity, even if it was a single-access resource, like the handouts of one lesson (we had one student with 25 accesses to that file, and another one checked his current grades 48 times).

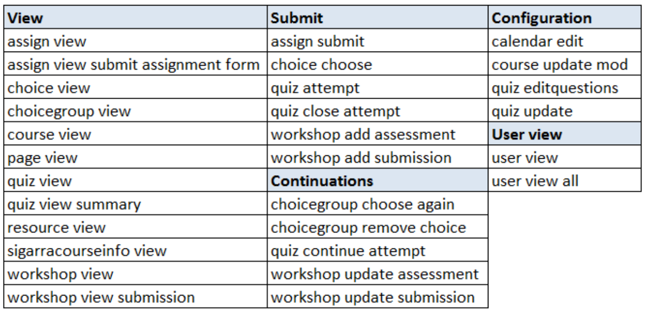
C. *Feature 3: correct sequence of resource usage*

Lastly, we grouped all course mandatory activities into groups of more general activities, as we list in Table 1. As we gathered actions into these general groups, we analyzed the interaction graph showing the number of interactions of each type along the semester. As a conclusion, we learn that the students used most of the interactions for viewing resources.

We focused on the peaks of the graph and tried to model that behavior according to the sequence in which actions were being made. For example, it's common for a student to skip seeing/reading some lecture slides until the very last moment before the test. Or, eventually, he skept some handouts, and later on had some regret for doing so and tried to recover wasted time by accessing them all in a row. All in all, and to summarize these type of behaviors, what we propose as a subject of concern is that the sequence in which activities occur is not independent from any permutation or a sub-set of this "normal" sequence. We stress that we are not forcing a specific behavior, but just detecting problematic situations.

For this purpose, we serialized all activities and resources made online available to students and marked the time in which it was profitable to use them. We mean "profitable" as having the contents, or the methodology to apply in the evaluation activities currently being undertaken.

Table 1. General grouping of actions.



View	**Submit**	**Configuration**
assign view	assign submit	calendar edit
assign view submit assignment form	choice choose	course update mod
choice view	quiz attempt	quiz editquestions
choicegroup view	quiz close attempt	quiz update
course view	workshop add assessment	**User view**
page view	workshop add submission	user view
quiz view	**Continuations**	user view all
quiz view summary	choicegroup choose again	
resource view	choicegroup remove choice	
sigarracourseinfo view	quiz continue attempt	
workshop view	workshop update assessment	
workshop view submission	workshop update submission	

We made a list of 17 different materials used in the course and did a partial order on that set. Then, we marked every interaction record for every student with its corresponding serial. Finally, we compared each student resource access sequence with our *golden standard*. For the sake of clarity, we present an example. Let us assume the golden standard is the following sequence: 1, 2, 3, 4, …, 16, 17. Now, imagine that for student X, the sequence is: 2,12,2,16,12,17,8,4,7,5,4,2. It should be clear that the student's sequence may not be either of the same length as the *golden standard* or with the same ordering.

We then created a function which computes the differences between the two sequences. This function is based on the Hamilton distance, as it counts the number of changes that have to be made in the student's sequence in order to obtain the golden standard. We call this value the "Ordering Distance", which we use as the third feature of the analysis.

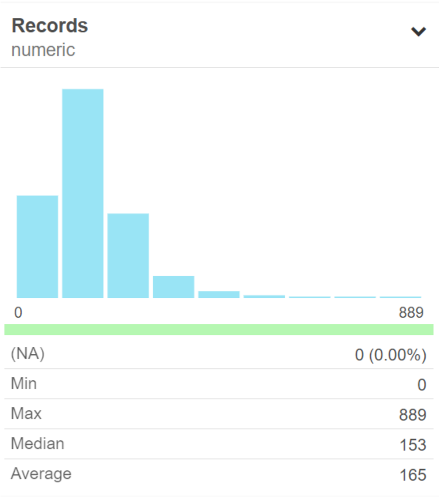
V. Data Analysis

In this section, we present our preliminary analysis of each feature, independently from each other.

A. *Number of accesses to the platform*

We computed the correlation between the number of times one student accesses online material and the final grade he had. We found that considering all student we didn't find a clear correlation (R^2 = 0.27, which is low). We also considered only the approvals, we got R^2 = 0.23 (which is also low). However, when we consider only the failures we got R^2 = 0.56 which is a meaningful correlation. Values of correlation increase to more than 0.65 as we switch from linear regression to polynomial regression of order higher than 2.

In Fig. 1 we present an exploratory data analysis of this feature regarding its distribution, min/max, and centrality values. The analysis was undertaken using the "Exploratory.io" system.



Records
numeric

(NA)	0 (0.00%)
Min	0
Max	889
Median	153
Average	165

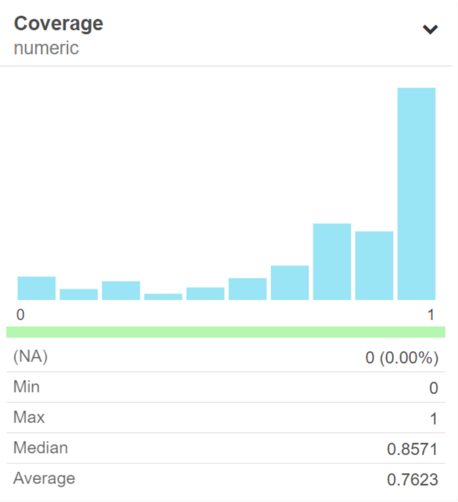
Fig. 1. Distribution of accesses to the platform.

B. *Coverage of the provided online material*

From the 14 items that each student, in theory, would need to access a single time, only two were accessed on average only once (which would be enough for any of these 14 items). However, if we count the number of times each student does not access any of these 14 items, and we compute the average of all these students, we got the number 3.32 which means that, on average, each student does not access more than 3 of needed items to successfully complete the course. Nonetheless, the average of covered material is 76%, and 60 out of 311 students had accessed all materials.In Fig. 2 we present the distribution of the covered material among students.

C. *Ordering distance*

Whilst it is fair to understand that the bigger this number, the more distant is a student sequence, to the golden standard (ie, to the natural correct sequence) this metric has an inherent problem: the shorter the sequence, the smaller would be the changes to convert it to the correct sequence.

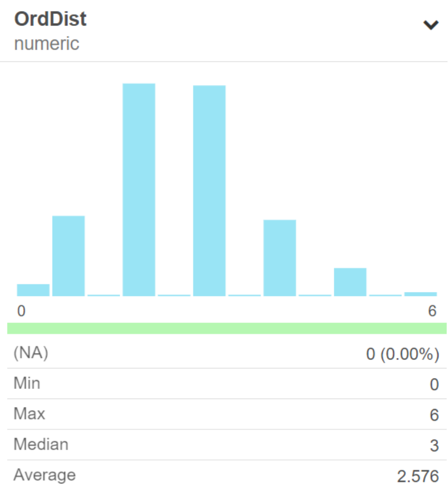


(NA)	0 (0.00%)
Min	0
Max	1
Median	0.8571
Average	0.7623

Fig. 2. Distribution of content coverage.

Therefore, this metric can be erroneous for students with short sequences. That is, students that do not access many online resources. Nevertheless, in this case study, as the resources are regularly accessed more than once, the former cases are clearly outliers.

This situation is illustrated in Fig. 3, where we see a normal distribution of the ordering distance.



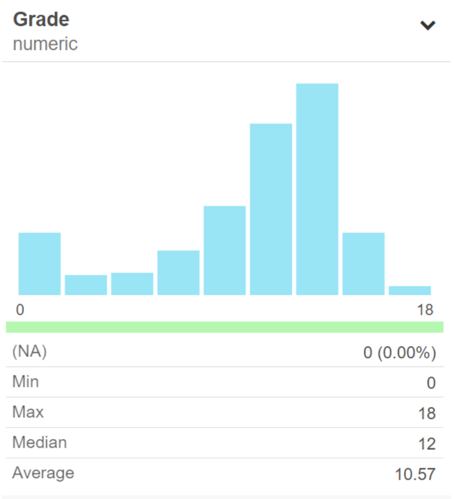
(NA)	0 (0.00%)
Min	0
Max	6
Median	3
Average	2.576

Fig. 3. Distribution of ordering distance value.

The final grades for the whole class were normally distributed, having a slightly skew at the right. This situation can be confirmed in Fig. 4. We stress that the final grades are in a scale from 0 to 20.

VI. Data Modelling

Equipped with these data set and set of features we wanted to model the student behavior and try to find a system capable of predicting if a student will pass or fail. This kind of prediction can be made with decision trees from which a set of "rules" is devised from the available data. A recent work using decision trees was able to predict grades using CART (Classification and Regression Trees) [8].



(NA)	0 (0.00%)
Min	0
Max	18
Median	12
Average	10.57

Fig. 4. Final grade's distribution.

Although it is possible to use a CART to predict, it is important to know how the tree fits to the training and test data (if available). The accuracy of a good predictive tree is never 100% otherwise, this would mean that the tree was totally fitted to the training data, therefore not capable of generalizing, nor predicting capabilities.

In this work, we do not try to predict the precise grade of a student in a scale 0-20, but only: i) if the student will pass or fail, and; ii) to measure the accuracy of such system. For this task, we use "Fast and Frugal Trees" (FFTs). A FFTree is a set of rules for making decisions based on very little information (usually 5 or fewer criteria). In our case, we use the three features described in section V to decide whether a student will pass or fail.

FFTrees are simple, transparent decision strategies that use minimal information to make decisions [9][10]. They are frequently preferable to more complex decision strategies (such as Logistic Regression) because they rarely over-fit data [11] and are easy to interpret and implement in real-world decision tasks. We implemented FFTs using the FFTrees package for the R programming language, which returns several FFTrees that attempt to classify training cases into criterion classes.

After evaluation against our prepared data, the package returned four trees, and picked the best one (tree #2), as described in the listing bellow showing its performance.

```
[1] "An FFTrees object containing 4 trees using 3 predictors
{Records,OrdDist,Coverage}"
[1] "FFTrees AUC: (Train = 0.78, Test = 0.79)"
[1] "My favorite training tree is #2, here is how it performed:"
                        train  test
n                      248.00 63.00
p(Correct)               0.81  0.81
Hit Rate (HR)            0.90  0.86
False Alarm Rate (FAR)   0.41  0.32
d-prime                  1.49  1.58
```

The dataset was split into 248 cases for training and 63 cases for testing, achieving very high correctness and hit rate, while maintaining a low false alarm rate, particularly in the test set. Tree #2 is described in Fig. 5.

The tree reads as: if the number of records is less than 93.58 than the student will fail; otherwise, if the ordering distance is bigger than 2, he will pass; otherwise, if the coverage is less than 57%, he will fail; otherwise, he will pass.

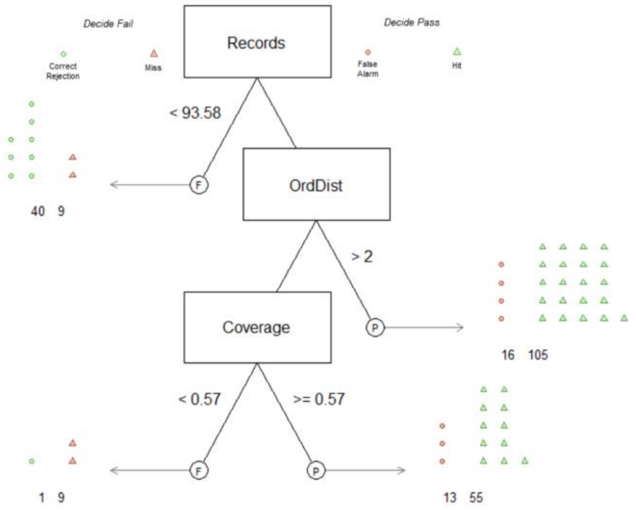
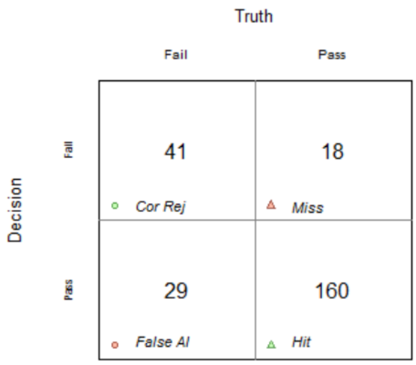


Fig. 5. The computed best FF Tree.

VII. Performance Fitting

Having the tree and an annotated data set (the final grades) it is possible to assess the tree's performance according to predicting a "fail" (grade < 10) or a "pass" (grade ≥ 10).

The classification table on Fig. 6 shows the relationship between tree decisions and the truth. CR (Correct Rejection) and H (Hit) are correct decisions. MI (Miss) and FA (False-alarm) are incorrect decisions.



		Truth	
		Fail	Pass
Decision	Fail	41 ○ Cor Rej	18 △ Miss
	Pass	29 ○ False Al	160 △ Hit

Fig. 6. Confusion matrix.

In Fig. 7 we present levels that show the cumulative tree performance in terms of Specificity, Hit Rate, D-prime, and AUC (area under the curve).

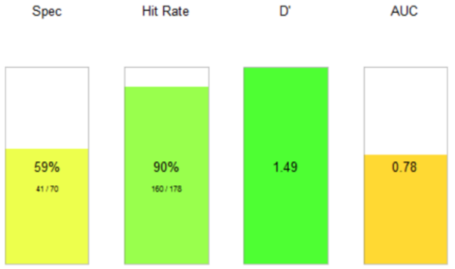


Fig. 7. Accuracy metrics.

As can be seen by the levels and correspondent colors, the performance is quite high, particularly in the hit rate (160 hits out of 178) and sensitivity (D-prime). Moreover, the tree's specificity got 41 cases out of 70, which is totally acceptable.

Finally, the plot depicted in Fig. 8 shows a ROC curve comparing the performance of all trees in the returned FFTrees object. Additionally, the performance of Logistic Regression (blue) and CART (red) are shown. The four trees created by the algorithm are plotted in green creating a line that connects each.

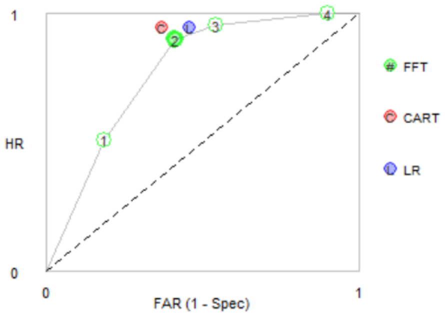


Fig. 8. Receiver operating characteristic curve (ROC).

Tree #2 is plotted in solid green and the other three trees are still all of them well above the dividing line. In Fig. 9 we present a comparative cue map for the three criteria and it is easy to see that each of them is positioned in the HR area.

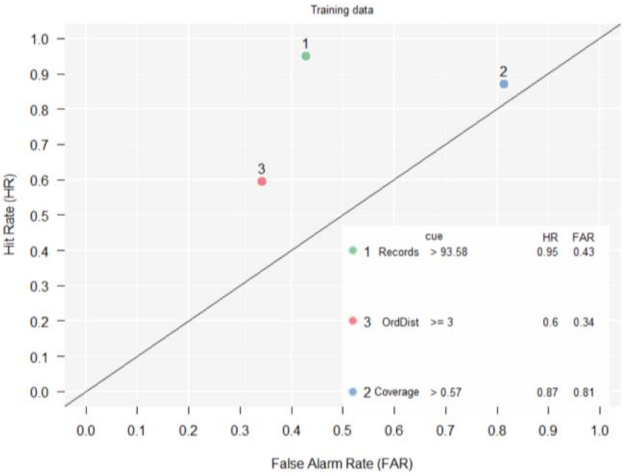


Fig. 9. Grade predictors' accuracy.

If we compare the performance of our trees with the performance of Logistic Regression and of CART (see listing bellow) we obtain the table below, once more confirming that FFTrees are a good choice for this kind of predictions.

```
        FFTrees        lr      cart
train 0.7769262 0.7433387 0.7861958
test  0.7948565 0.8080144 0.7852871
```

VIII. Conclusions

The obtained results show that our hypotheses are confirmed by the experiment. The graphical representation of sequences of communication between students is also helpful to understand the other graphic representation of resource usage, which in turn is a valuable tool to predict final results (pass or fail).

The analysis of the sequential access to pedagogical content can be complemented with an analysis of the extension of the pedagogical material that was actually covered by each student. The latter indicator provides a decisive feature in the analysis.

Moreover, we used the log data to also apply Social Network Analysis, which lead us to be able to characterize the class using simple and comparable metrics. This analysis provided us with a base structure to frame the context of the class allowing us to draw generalizable conclusions about causal factors for grade prediction.

Finally, in this study we experimented a new approach to predict passing/failing in order to prevent students from failures at early stages. The model we described can be implemented within Moodle, but it can also be generalized to any LMS with a reasonable logging system.

In this article, we discussed three types of features, that can be extracted from the logs to characterize the interaction behavior with the platform.

Acknowledgments

This work is supported by the ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within project «Reminds/ UTAP-ICDT/EEI-CTP/0022/2014»

References

[1] Hoppe, U. et al. “Building bridges within learning communities through ontologies and thematic objects”. Proceedings of the International Conference on Computer Supported Collaborative Learning. pp. 211–220 (2005).

[2] Perera, D. et al. Clustering and sequential pattern mining of online collaborative learning data. IEEE Trans. Knowl. Data Eng. **21**(6), 759–772 (2009).

[3] Niemann, K., et al. Clustering by usage: higher order co-occurrences of learning objects. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. pp. 238–247. ACM (2012)

[4] Zeibarth, S., Chounta, I., Hoppe, H. “Resource Access Patterns in Exam Preparation Activities”. Proceedings of the EC-TEL’15, LNCS, pp.497-502, 2015.

[5] André Silva and Álvaro Figueira, "Visual Analysis of Online Interactions through Social Network Patterns". In Proceedings of the IEEE ICALT conference. Rome, Italy. July, 2012

[6] Mödritscher, F., Andergassen, M., Neumann, G. “Dependencies between e-learning usage patterns and learning results”. Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies. ACM (2013).

[7] Figueira, A. “Predicting Results from Interaction Patterns During Online Group Work”. Proceedings of the EC-TEL’15, LNCS, pp.414-419, 2015.

[8] Figueira, A. “Predicting Grades by Principal Component Analysis: A Data Mining Approach to Learning Analyics”. Proceedings of ICALT’16. Austin, 2016.

[9] Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In Simple heuristics that make us smart (pp. 3–34). Oxford University Press.

[10] Gigerenzer, G., Czerlinski, J., & Martignon, L. (1999). How good are fast and frugal heuristics? In Decision science and technology (pp. 81–103). Springer.

[11] Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. Topics in Cognitive Science, **1**(1), 107–143.