Discovering students' navigation paths in Moodle

Alejandro Bogarín
Department of Computer Science
University of Cordoba, Spain
(+34) 679 30 54 86
abogarin@uco.es

Cristóbal Romero
Department of Computer Science
University of Cordoba, Spain
(+34) 653 46 28 13
cromero@uco.es

Rebeca Cerezo
Department of Psychology
University of Oviedo, Spain
(+34) 627 60 70 21
cerezorebeca@uniovi.es

ABSTRACT
In this paper, we apply clustering and process mining techniques to discover students' navigation paths or trails in Moodle. We use data from 84 undergraduate Psychology students who followed an online course. Firstly, we group students using on Moodle's usage data and the students' final grades obtained in the course. Then, we apply process mining with each cluster/group of students separately in order to obtain more specific and accurate trails than using all logs together.

Keywords
Clustering, process mining, navigation paths, trails in education.

1. INTRODUCTION
One of the current promising techniques in EDM (Educational Data Mining) is Educational Process Mining (EPM). The main goal of EPM is to extract knowledge from event logs recorded by an educational system [4]. It have been observed that students show difficulties when learn in hypermedia and Computer Based Learning Environments (CBLEs) due to these environments seems to be highly cognitive and metacognitive demanding [1]. In this sense, the models discovered by EPM could be used: to get a better understanding of the underlying educational processes, to early detect learning difficulties and generate recommendations to students, to help students with specific learning disabilities, to provide feedback to either students, teachers or researchers, to improve management of learning objects, etc. In a previous work [2], we found two problems when using EPM: 1) the model obtained could not fit well to the general students' behaviour and 2) the model obtained could be too large and complex to be useful for a student or teacher. In order to solve these problems, we proposed to use clustering to improve both the fitness and comprehensibility of the obtained models by EPM. However, in this paper we propose to use a Hypertext Probabilistic Grammar (HPG) algorithm instead of Heuristics Net [2] because it provides more informative graphs.

2. METHODOLOGY
A traditional approach would use all event log data to reveal a process model of student's behaviour . Nevertheless, in this paper, we propose an approach that uses clustering for improving EPM (see Figure 1). The proposed approach firstly applies clustering in order to group students with similar features. And then, it applies process mining for discovering more accurate models of students' navigation paths or trails. In fact, we propose to use two different grouping methods:

1) Clustering students directly by using the students' grades obtained in the final exam of the course.
2) Clustering students by using a clustering algorithm over the student's interaction with the Moodle's course.

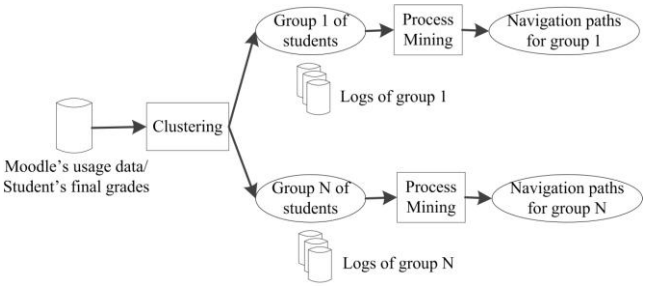

Figure 1: Proposed approach for discovering students' navigation paths.

3. DESCRIPTION OF THE DATA AND EXPERIMENTS
In this work we have used real data collected from 84 undergraduate Psychology students who followed a Moodle course. Firstly, we have divided the student's log provided by Moodle in two different ways. In a first way, we divided directly the original log file into two datasets: one that contains the 68 students who passed the course and other with the 16 students who failed. In the second way, we have used the Expectation-Maximization (EM) clustering algorithm provided by Weka [6] in order to group together students of similar behaviour when using Moodle. In this case we have obtained three clusters/datasets with the following distribution:

- **Cluster 0**: 23 students (22 pass and 1 fail).
- **Cluster 1:** 41 students (39 pass and 2 fail).
- **Cluster 2**: 20 students (13 fail and 7 pass).

After clustering, we applied EPM through HPG over the previous datasets. We have used the HPG model in order to efficiently mine trails or navigation paths [3]. HPG uses a one-to-one mapping between the sets of non-terminal and terminal symbols. Each non-terminal symbol corresponds to a link between Web pages. Moreover, there are two additional artificial states, called *S* and *F*, which represent the start and finish states of the navigation sessions respectively. The probability of a grammar string is given by the product of the probability of the productions used in its derivation. The number of times a page was requested, and the number of times it was the first and the last page (state) in a session, can easily be obtained from the collection of student navigation sessions. The number of times a sequence of two pages appears in the sessions gives the number of times the corresponding link was traversed. The aim is to identify the subset of these trails that correspond to the rules that best characterize the student's behavior when visiting the Moodle course. A trail is included only if its derivation probability is above a cut-point.

The cut-point is composed of two distinct thresholds (support and confidentiality). The support (Sup) value is for pruning out the strings whose first derivation step has low probability, corresponding to a subset of the hypertext system rarely visited. The confidence (Con) value is used to prune out strings whose derivation contains transitive productions with small probabilities. Support and confidence thresholds give the user control over the quantity and quality of the obtained trails, while α (Alp) modifies the weight of the first node in a student navigation session: when α is near 0, only those routes that start in a node which started a session are generated; when α is near 1, all weights are completely independent of the order within the session.

4. RESULTS

We have carried out several experiments with the HPG algorithm to test several configurations of number of Nodes, Links, Routes, and average route length (Avg). Results obtained when using different datasets and parameters are displayed in Table 1.

Table 1. Results with different datasets and configurations.

Dataset	Alp	Sup	Con	Nodes	Links	Routes	Avg
Fail	0,2	0,05	0,5	8	7	12	3,85
Pass	0,2	0,05	0,5	12	11	20	3,81
Cluster0	0,2	0,05	0,5	8	6	12	4,16
Cluster1	0,2	0,05	0,5	9	7	14	4,14
Cluster2	0,2	0,05	0,5	5	4	6	2,75
Fail	0,4	0,06	0,3	15	15	27	3,8
Pass	0,4	0,06	0,3	25	27	47	3,96
Cluster0	0,4	0,06	0,3	13	12	21	3,66
Cluster1	0,4	0,06	0,3	15	17	29	4,11
Cluster2	0,4	0,06	0,3	12	9	18	3,66
Fail	0,5	0,06	0,3	20	19	36	4
Pass	0,5	0,06	0,3	37	41	72	4,07
Cluster0	0,5	0,06	0,3	19	17	31	3,7
Cluster1	0,5	0,06	0,3	20	21	38	4,19
Cluster2	0,5	0,06	0,3	12	9	18	3,66

Table 1 show that the smaller and more comprehensible models were obtained using logs from students who failed (Fail dataset) and students of Cluster 2. On the other hand, the models obtained with the other datasets were much bigger and complex. We think that this may be due to:

- Both dataset Fail and Cluster 2 contain mainly information about bad students who failed the course. This type of students has a low interaction with Moodle and so, they show only some frequent navigation paths.
- Datasets Pass, Cluster0 and Cluster1 contain mainly information about good students who pass the course. This type of students has a high interaction with Moodle and so, they show more frequent navigation paths.

Finally, we show an example of obtained model when using the Cluster2 dataset. In Figure 2, each node represents a Moodle's Web page, and the directed edges (arrows) indicate how the students have moved between them. These paths can be stochastically modeled as Markov chains [5] on the graph, where the probability of moving from one node to another is determined by which Web page the student is currently visiting. Edge thickness varies according to edge weight; this allows the learning designer to quickly focus on the most important edges, ignoring those that have very low weights. In addition, line widths and numerical weights are also available.

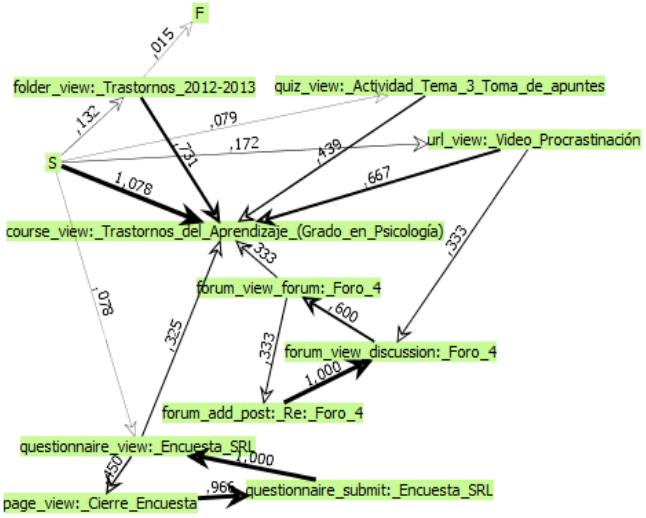

Figure 2: Navigation paths of Cluster 2 students.

Starting from Figure 2 we can see and detect what are the most frequent actions (view forum X, view questionnaire Y, view quiz Z, etc.) and in which order (navigation paths or trails) were done/followed by Cluster 2 students (normally fail students).

5. REFERENCES

[1] Azevedo, R., Behnagh, R., Duffy, M., Harley, J. M., & Trevors G. J. Metacognition and self-regulated learning in student-centered learning environments. *Theoretical foundations of student-center learning environments*. Erlbaum, Mahwah, NJ, 2nd edition, 216–260, 2012.

[2] Bogarin, A. Romero, C., Cerezo, R., Sanchez, M. Clustering for improving Educational Process Mining. *Learning Analytics and Knowledge Conference*, Indianapolis, 11-14.

[3] Borges, J., Levene, M. Data Mining of user navigation patterns. Proc. of Workshop Web Usage Analysis and User Profiling. San Diego, 2000. pp. 31-36.

[4] Pechenizkiy, M., Trcka, N., Vasilyeva, E., van der Aalst, W.M., & De Bra, P. 2009. Process Mining Online Assessment. Data. *Educational Dada Mining Conference*, Cordoba, Spain, 279-288.

[5] Kemeny, J.G., Snell, J.L. Finite Markov chains. Princeton: Van Nostrand. 1960

[6] Witten, I.H., Eibe, F., Hall, M.A. Data Mining, Practical Machine Learning Tools and Techniques. Third Edition. Morgan Kaufman Publishers, 2001.