Analysis of students clustering results based on Moodle log data

Angela Bovo, Stephane Sanchez, Olivier Héguy, Yves Duthen

To cite this version:

Angela Bovo, Stephane Sanchez, Olivier Héguy, Yves Duthen. Analysis of students clustering results based on Moodle log data. 6th International Conference on Educational Data Mining - EDM 2013, Jul 2013, Memphis, Tennessee, United States. pp. 306-307. hal-01146267

HAL Id: hal-01146267
https://hal.archives-ouvertes.fr/hal-01146267

Submitted on 28 Apr 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Open Archive TOULOUSE Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: http://oatao.univ-toulouse.fr/
Eprints ID: 13222

To link to this article:
URL: http://www.educationaldatamining.org/IEDMS/EDM2013

To cite this version: Bovo, Angela and Sanchez, Stephane and Heguy, Olivier and Duthen, Yves Analysis of students clustering results based on Moodle log data. (2013) In: 6th International Conference on Educational Data Mining - EDM 2013, 6 July 2013 - 9 July 2013 (Memphis, Tennessee, United States).

Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr
Analysis of students clustering results based on Moodle log data

Angela Bovo  
Andil  
Université Toulouse 1  
IRIT  
angela.bovo@andil.fr

Stéphane Sanchez  
Université Toulouse 1  
IRIT  
stephane.sanchez@irit.fr

Olivier Héguy  
Andil  
olivier.heguy@andil.fr

Yves Duthen  
Université Toulouse 1  
IRIT  
yves.duthen@irit.fr

ABSTRACT
This paper describes a proposal of relevant clustering features and the results of experiments using them in the context of determining students' learning behaviours by mining Moodle log data. Our clustering experiments tried to show whether there is an overall ideal number of clusters and whether the clusters show mostly qualitative or quantitative differences. They were carried out using real data obtained from various courses dispensed by a partner institute using a Moodle platform. We have compared several classic clustering algorithms on several group of students using our defined features and analysed the meaning of the clusters they produced.

Keywords
clustering, Moodle, analysis, prediction

1. INTRODUCTION
1.1 Context of the project
Our project aims to monitor students by storing educational data during their e-learning curriculum and then mining it. The reasons for this monitoring are that we want to keep students from falling behind their peers and giving up.

This project is a research partnership between a firm and an university. The partner firm connects our research with its past and current e-learning courses, hence providing us with real data from varied trainings.

All available data comes from a Moodle [5] platform where the courses are located. Moodle’s logging system keeps track of what materials students have accessed and when. We then mine through such logs.

1.2 Clustering as a means of analysis
Clustering is the unsupervised grouping of objects into classes of similar objects. In e-learning, clustering can be used for finding clusters of students with similar behaviour patterns. In the example of forums, a student can be active or a lurker [1, 7]. These patterns may in turn reflect a difference in learning characteristics, which may be used to give them differentiated guiding [2] or to predict a student’s chance of success [3]. They may also reflect a degree of involvement with the course, which, if too low, can hinder learning. The data contained in Moodle logs lends itself readily to clustering, after a first collecting and pre-processing step [6].

Our aim with this analysis will be to determine if there is an overall ideal number of clusters and whether the clusters show mostly qualitative or quantitative differences. We chose clustering, which is unsupervised, in order to better reflect the natural structure of our data. Because of this choice, the outcome of our experiments will not be directly relevant to the success of the students, but will rather reflect the differences in their usage of the LMS.

2. FEATURES CHOSEN TO AGGREGATE THE DATA
We have tried to aggregate the Moodle log data into a list of features that could capture most aspects of a student’s online activity. The features we have selected are: the login frequency, the date of last login, the time spent online, the number of lessons read, the number of lessons downloaded as a PDF to read later, the number of resources attached to a lesson consulted, the number of quizzes, crosswords, assignments, etc. done, the average grade obtained in graded activities, the average last grade obtained, the average best grade obtained, the number of forum topics read, the number of forum topics created, and the number of answers to existing forum topics. For every “number of x” feature, we actually used a formula that would reflect both the distinct and total number of times that this action had been done. All of our features are normalized, with the best student for each grade obtaining the grade of 10, and others being proportionally rescaled.

3. EXPERIMENTAL METHOD
4. CLUSTERING RESULTS

4.1 Best number of clusters

The following figure shows the results of the four algorithms used on each of our three datasets. The first shows the frequency at which the X-Means algorithm proposed a given number of clusters. The other three graphs show the error for a given number of clusters for K-Means, Hierarchical clustering and Expectation Maximisation. We can see that all algorithms generally agree on at most 2 or 3 clusters.

4.2 Meaning of the clusters

To our surprise, the clusters observed for all three trainings did not show anything more relevant than a simple distinction between active and less active students, with variations according to the chosen number of clusters. We did not, for instance, notice any group that would differ from another simply by their activity on the forum.

To explain this, we offer the following possible reasons. Firstly, we have a relatively small number of students in each training (between 15 and 56), which may mean less variety in behaviour. Secondly, this training may be targeted towards a relatively homogeneous audience in terms of age, professional training, and habitual use of IT. Thirdly, a vicious circle effect can happen of the forum, because if few people use it, other students have less incentive for using it.

Hence, in about all observed clusters, the students were only quantitatively differentiated by a global activity level. It is also to be noticed that when the number of clusters was too large, clusters containing only one student, the most or least active of his training, tended to form. This phenomenon might be a good indicator that the number of clusters is too high without the help of a comprehensive study.

However, the fact that all differences were proportional also means that the student’s activity level was also correlated to the grades they obtained in graded activities (which were not evaluative). This seems to indicate that in our trainings, using a quantity of activity is sufficient to help identify students in trouble, which is our global aim.

5. CONCLUSIONS AND FUTURE WORK

This paper proposes comprehensive and generic features that can be used for mining data obtained from Moodle courses. These features are then used to conduct a clustering of the data, using several algorithms, followed by an analysis, which seems to show very little qualitative difference in behaviour between students. It seems that a single feature, a kind of index of their global activity, would be almost sufficient to describe our data. This is also shown by the very little (2 to 3) number of clusters that is sufficient for describing our data. We propose several explanations for this surprising result, such as the small dataset, the homogeneity of our students and a vicious circle effect. However, the results mean that using our features or computing a quantity of activity could be enough to monitor students and notice which ones run a risk of failure.

6. REFERENCES


