Detecting Disengagement of Online Students through Log Files Analysis

Zdena Lustigova[1], Veronika Novotna[2]
[1]Faculty of Mathematics and Physics, Charles University in Prague (Czech Republic), [2]Faculty of Medicine Plzen, Charles University in Prague (Czech Republic)
lustigo@plk.mff.cuni.cz, Veronika.novotna@lfp.cuni.cz

Abstract

To achieve effective learning, motivational aspects like engagement play a very important role. Within online learning applications the disengagement detection and prediction based on real data (not always in real time) is becoming more and more popular among educational specialists. Many E-learning systems, and virtual or remote learning environments, could be improved by tracking students' disengagement that, in turn, would allow personalized interventions at appropriate times in order to re-engage students.

The present article describes the results of a medium-scale (N = 56) study, using log files from Open Remote Laboratory at Charles University in Prague, Faculty of Mathematics and Physics, to observe secondary school students' behaviour during their work in virtual environment. Simple data mining and text mining techniques were used to reveal individual user's behavioural patterns and to detect disengagement. The results will be used mainly to improve the systems' adaptability to students' requirements and to prevent their disengagement.

1. Introduction

1.1 Origin of educational data mining (EDM)

At the end of last century, E-learning and online learning (also referred to as web-based education) started to generate large amounts of information describing the continuum of the teaching-learning interactions. Up to nowadays this information is endlessly generated, easily and ubiquitously available but rarely processed. "At the beginning the variety and amount of data from learning and teaching process was often seen as a blessing: plenty of information readily available just a click away. Equally it could be seen as an exponentially growing nightmare, in which unstructured information chokes the educational system without providing any articulate knowledge to its actors." [1]

Educational Data Mining was born to deal with problems like this. As a field of research, it is almost contemporary to e-learning. It is, though, rather difficult to define. Not because of its intrinsic complexity, but because "data mining has most of its roots in the ever-shifting world of business. At its most detailed, it can be understood not just as a collection of data analysis methods, but as a data analysis process that encompasses anything from data understanding, pre-processing and modelling to process evaluation and implementation." [2] [6]

1.2 Disengagement detection and prediction models

Within online learning applications the disengagement detection and prediction based on real data (not always in real time) is more and more popular.

In effective learning, motivational aspects like engagement play a very important role. E-learning systems could be improved by tracking students' disengagement that, in turn, would allow personalized interventions at appropriate times in order to reengage students.

There are a couple of studies based on Item Response Theory – like an engagement tracing or a model combining a hidden Markov model with Item Response Theory (e.g. [4]). This dynamic mixture models take into account student proficiency, motivation, evidence of motivation, and a student's response to a problem.

Another approach based on the ARCS Model (Attention, Relevance, Confidence, and Satisfaction, Keller [5]) infers three aspects of motivation: confidence, confusion, and effort, from the learner's focus of attention and inputs related to learners' actions: the time to perform the task, the time to read the paragraph related to the task, the time for the learner to decide how to perform the task, the time when the learner starts/finishes the task, the number of tasks the learner has finished with respect to the current plan (progress), the number of unexpected tasks performed by the learner which are not included in the current learning plan, and the number of questions asking for help. [1]

2. Description of our research problem and its "state of art"

Our research was focused mainly on users modelling and disengagement detection within remote and open laboratory activities.

Remote laboratories represent one of the three mostly used nowadays laboratory landscapes, together with so called virtual labs (also known under the name simulated labs) and computer-mediated, hands-on labs.

Remote labs enable experimenting and lab work in virtual conditions and with the use of remote access. Although this work is often done in environments and conditions for recent generations of students unimaginable, the main goals of laboratory work are still the same: to master students' basic concepts, to help them to understand the role of direct observation, to train them to distinguish between inferences based on theory and the outcomes of experiments, to teach them to cooperate and to develop collaborative learning skills. But they have to do all this being exposed to uncertain and not exactly defined situations, since the whole virtual and remotely controlled working environment is more complicated and thus more unpredictable. (Lustig, Lustigova in [7],[8],[9]). This brings also more and more unpredictable to the teacher (or online supervisor) and also places greater demands on the analyst and remote lab developers, who themselves have often grown up and learned in different conditions.

Also educational research within remote labs conditions has to deal with higher fuzziness and unpredictability. While in e-learning or online learning environment researchers have to their disposal plenty of structured and unstructured textual information, including discussion threads, all kind of communication between teacher and student, student-student, student-team of students, student – learning material (in form of personalised comments, reviews, etc.), in remote labs the situation is different. The remote lab communication tools are very limited and the whole work is usually task oriented: to setup the experimental environment, to gather data and to process them. If there is a team work and the negotiation connected, it is observable directly, at place (see [8]).

Remote laboratory environments offers communication tools like chats, discussion clubs or cafés, whether synchronous or asynchronous, very rarely. This means, that there is virtually no textual information available and the researchers often have to work just with log files and information hidden in there.

Within the latest "state of art" literature review focused on remote laboratories, we did not find any study based on log files analysis. It follows that log file data from remote laboratories is more often collected than analysed. Most of research papers in the field are focused on remote experiments development, online access improvement and other technical and engineering aspects of the problem. Studies of users' behaviour and learning process are quite rare and often based on direct (at place) observation, results and reports discussion, or survey data [8].

Within our research we processed data from log files, collected in spring and summer 2012 at remote laboratory belonging to Charles University in Prague, Faculty of Mathematics and Physics. In spring and early summer 2012 the most engaged were students of 5 secondary schools, who were asked to measure and process their data and report their results of photo effect experiment.

3. Research process and results

Our main goal during processing log files data from this students' activity was to reveal disengagement, to prevent such a situation and to improve the users' motivation within the online

learning and measuring environment. We researched mainly to avoid objective causes of disengagement, such as 1/unnecessarily long wait for the event or feedback, 2/confusing information and instructions or other problems, that cannot be easily identified with the use of traditional techniques.
Each particular record contains a string, describing individual user activity, without losing any information (see an example of an individual user activity recorded in a form of a string below).

81.25.16.87 17.4.2011 18:37:29 1035 s ID(4)

0:W(1){88}*Sv1{23}*Sv1{10}*Sr(100){71}*Sl1{1}*Sl0{4}*Sl1{7}*Mv(-12.16){0}*Mv(-445.85){0}
*Mv(-477.93){0}*Mv(-1000.00){1}*Mv(-1000.00){4}*Ma0{160}*Sf(0){1}*Sf(1){3}*Sf(0){10}*Ma1{46}
*Pr(1){9}*Ma1{43}*Sf(1){3}*Ma1{43}*Sf(2){3}*Ma1{44}*Sf(3){3}*Ma1{42}*Sf(4){3}*Ma1{43}*Sf(5){8}
*Sl0{5}*Ma1{44}*Ps(1){0}*Pd(1){12}*Pd(1)*D

Fig. 1: The example of an individual user's activity string, derived from the log file (adopted from Lustigova and Brom [10]).

While the first line in the figure above identifies the user's computer IP address, the date and time he started to measure, the whole time in seconds his activities lasted and the original ID in log file under which we can find original data, the second long line contains the full description of user activities. For the legend ask authors.

3.1 Descriptive statistics

From the collection of 613 sessions within first half of 2012, just 155 belonged to the experimental group and from that number just 15 sessions finished with meaningful measurement or data downloading. The length of the connections changes from very short to very long (up to one hour), but it says nothing about the meaningfulness of the activities. Our experimental group users connected from 43 different IP addresses. The users preferred to work in late afternoons and evenings, some of these secondary schools students worked after midnight as well.

3.2 Disengagement detection

The number of connections, where the user was alone, apparatus ready to measure, but he/she disconnected after a while for unknown reason, is surprisingly high (108). It is even higher than the number of connections finished because of necessity to wait (28). All connections that have finished with meaningful activities were the "wait" connections: waiting for one user (3 minutes) 9 connections, waiting for two users (6 minutes) 6 connections.
Our null hypothesis that users are disengaged because of waiting in queue has to be rejected.
On contrary to Corter, Nickerson, at al study (2007), where they refused real-time interaction and forced fixed scheduling on the students, we decided to offer both, real time measurement and pre-measured data.
"Early birds" students, who followed recommended time schedule, preferred real time measurement (app ¼ within each group), while those "last minute" students, cueing to operate remotely lab devices, frequently used pre-measured data, often without checking their quality and reliability.
Although the remote lab offers up to 200 stored data sets, the users in experimental group usually selected among last 3 offers without using the preview and checking their reliability and quality.

4. Conclusions

Although the students from experimental group presented nicely processed reports, the reality hidden in log files was different. On the base of educational data mining techniques, we found the following reasons for disengagement:
1/Lack of training:

Although our remote laboratory is open to individual secondary school students, the overwhelming majority of them are not able to practice in the laboratory without meaningful training. If they are forced to do so, they leave the environment without any meaningful activity or they play for a while, but then also prefer data withdrawal to the real measurement.
2/Lack of self-confidence:
Students do not trust to their own results. It might be associated with the learning and teaching paradigm change in general.
3/Lack of supervision and/or increased uncertainty in the virtual environment:
Students are not used to the "researchers' freedom" offered by remote laboratories. They are missing step by step guides and lab sheets.

References

[1] Castro, F., Vellido, A., Nebot, A. and Mugica, F. (2007) Applying Data Mining Techniques to e-Learning Problems. Springerhttp://sci2s.ugr.es/keel/pdf/specific/capitulo/ApplyingDataMiningTechniques.pdf
[2] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000) CRIPS-DM 1.0 Step by Step Data Mining Guide. CRISP-DM Consortium.
[3] Corter, J. E., Nickerson, J. V., Esche, S. K. Chassapis, C. Im, S., and Ma, J. "Constructing Reality: A study of remote, hands-on and simulated laboratories", ACM Transactions on Computer Human Interaction (14), 2, Article 7, 1-27
[4] Johns,J. and Woolf,B. (2006). A Dynamic Mixture Model to Detect Student Motivation and Proficiency. Proceedings of 21st National Conf. Artificial Intelligence (AAAI-06).
[5] Keller, J.M.: Development and Use of the ARCS Model of Instructional Design. Journal of Instructional Development, vol. 10, no. 3, pp. 2-10, 1987
[6] Koedinger, K., Cunningham, K., Skogsholm, A., and Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. In Proc. 1st Int. Conf. Educ. Data Mining, Montreal, QC, Canada, pp. 157–166.
[7] Lustigova, Z., Lustig, F. (2008). New e-learning environments for teaching and learning science In: Learning to live in the knowledge society. Springer. Milan. Italy.
[8] Lustigova, Z., Lustig, F., Mechlova, E.& Malcik, M. (2009). A New E-learning Strategy for Cognition of the Real World. In: The New Educational Review, Vol. 17. No. 1. 305-317.
[9] Lustigova, Z., Lustig, F.(2009). Remote and Open Laboratory in Science Education - Technological, Educational and Psychological Issues. Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces. Cavtat. Croatia.
[10] Lustigova, Z., Brom, P.:The log files analysis in remote laboratories as a way of educational data mining. In Proceedings of ICTE 2013. Roznov pod Radhostem.2013. Czech Republic.pp.74-80.