Educational Datamining in Virtual Learning Environments

http://dx.doi.org/10.3991/ijac.v7i1.3557

Z. Lustigova, P. Brom
Charles University in Prague, Prague, Czech Republic

***Abstract*—The present article describes the results of a medium-scale (N = 77) study, using log files from open remote laboratory at Charles University in Prague, Faculty of Mathematics and Physics, to observe students' behavior during their work in virtual environment. Simple data mining and text mining techniques were used to reveal individual user's behavioral patterns, to detect disengagement, and to compare learning outcomes and student preferences.**

***Index Terms*—datamining, educational research, log files analysis, remote and virtual laboratories,**

I. INTRODUCTION

A. *Definitions of educational data mining*

There are many different definitions of educational data mining and its main issues.

From the statistician and data miners point of view Educational data mining (EDM) is a field that exploits statistical, machine-learning, and data-mining algorithms over the different types of educational data with the main goal to analyse these types of data in order to resolve educational research issues.

Policy makers and administrators usually think that EDM is mostly about mining enrolment and students' performance data for improving the services they provide and for increasing student grades and retention.

Generally EDM is concerned with developing methods to explore the specific types of data obtained in educational settings and, using these methods, to better understand students and the settings in which they learn. On one hand, the increase in both instrumental educational software as well as state databases of student's information have created large repositories of data reflecting how students learn (Koedinger et al, 2008). On the other hand, the use of Internet in education has created a new context known as e-learning or web-based education in which large amounts of information about teaching–learning interaction are endlessly generated and ubiquitously available.

There is also a third, rediscovered, way, how to understand the process of students' learning. Simple, noninvasive, low cost measurements of neurophysiological factors like eyes blinks, galvanic skin response (GSR) or heart and breathe rate, together with screen activities and events recording are nowadays easily available. They became cheaper and more and more transferable to the „out of laboratory" conditions – into the real learning environments. This kind of data, gathered and processed in real time, has a great potential to provide the immediate and individualized reaction on the decreasing attention, increasing visual or cognitive information load, task difficulty, tension, arousal, stress and/or achievement of the learning subject. (Lustigova et al, 2010).

Educational data mining and learning analytics are more and more used to research and build models in several areas that can influence learning process itself, or at least to improve online learning systems.

B. *Description of our research problem and its "state of art"*

Our research was focused mainly on users modeling and disengagement detection and prediction within remote 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. Nowadays students have also to master their basic science concepts, to understand the role of direct observation, to distinguish between inferences based on theory and the outcomes of experiments, 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 et al. 2012). 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 personalized 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 Lustig, Lustigova 2011).

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 analyzed. 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' behavior and learning process are quite rare and often based on direct (at place) observation, results and reports discussion, or survey data (Lustigova et al, 2011)

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.

Remote laboratory at Charles University in Prague belongs to so called "open remote laboratories", which means that the local laboratory through a remote control option is available to any visitor, who is interested. 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.

Unlike many remote laboratories, laboratory at Charles University offers quite favorable conditions for high school students. The impression of the real presence is emphasized by installed web cameras that provide real time image transmission of the most interesting parts of selected experimental setup or its results. Simultaneously, different variables are measured and visualized in a form of graphs.

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 unnecessarily long wait for the event or feedback, confusing information and instructions or other problems, that cannot be easily identified with the use of traditional techniques.

We also wanted to discover behavioral and problem solving patterns with the help of user modeling technique, described above.

II. RESEARCH PROCESS AND RESULTS

Each particular record in log file, pre-processed by special SW without losing any information, contains a string, describing individual user activity, (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
```

Figure 1. The example of an individual user's activity string, derived from the log file

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.

A. *Descriptive statistics*

From the collection of 613 sessions within first half of 2011, just 155 belonged to the experimental group (April 2011) and from that number just 15 sessions finished with measurement or data downloading. The length of the connections changes from very short to very long (up to one hour). The length of the connection says nothing about the meaningfulness of the activities. Some short connections finished with data downloading, while some very long connections string descriptions contain absolutely no activity (see histogram of connection length on figure 4, notice that time axe is nonlinear). The average length of any connection was 354,7 seconds, while the average length of meaningful connection (connection finished with data download or measurement) was 756,2 seconds.

Our experimental group users connected from 43 different IP addresses. The users preferred to work in late afternoons and evenings (see Fig. 4). Notice that some of these secondary schools students worked after midnight as well.

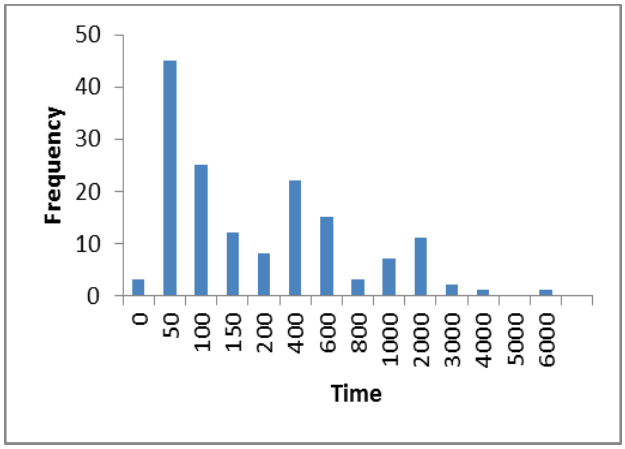

Figure 2. Time duration of an individual user connection (absolute frequencies histogram)

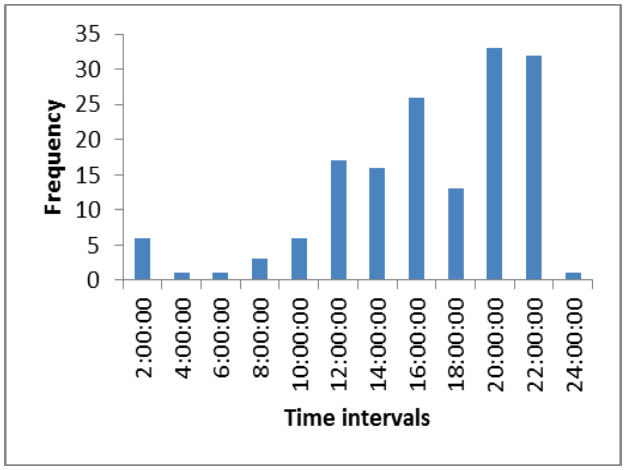

Figure 3. Daily variations of the connection time (absolute frequencies histogram)

If we define a session as a chronological series of a connections from defined IP address within the same day and setup the interconnection "no activity interval" up to 15 minutes (900 s), the number of sessions decreases to 56. Since the number of participants in our experimental group was slightly higher, it gives us evidence that some of them were not able or did not want to work within the remote lab environment.

B. Behavioural patterns

Individual user observation (selected examples):

User A (IP: 88.102...) connected to the remote experiment repeatedly and had to wait in a queue (W). Finally he/she downloaded someone else's (User B's) data (Pd). User B (IP: 81.25...) on 11/4/2011 first explored the volt-ampere characteristics of a vacuum phototube. On 17/4/2011 (see figure 1) user B had to wait in a queue, but after 88 s of waiting user B took control of the remote experimental setup, explored the interface and after a short time of playing at the beginning he started with systematic measurement activities afterwards. User B spent 1035 s (i.e. 17 min) performing the remote measurement with data acquisition.

The sessions recorded under IP address 81.25.16.87 (Fig 4) from April 11 2011 informs us about different behavioral pattern. This whole session lasted approximately 63 minutes. The user spent 2447 seconds (i.e. approximately 40 minutes) with playing all buttons and measuring. He/she started at about 8 p.m. and luckily was alone. But he did not use the occasion. After while (waiting for 2) he/she took the control and started to work. The activity record, presented by following string (figure 4), belongs to the longest ones, but surprisingly has no real output.

"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.

III. 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 revealed, that:

1. although our remote laboratory is open to individual secondary school students, the overwhelming majority of them is 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. The "play phase" seems to be very important. Just those, who played for a while, were able to setup the apparatus, to start the measurement, to finish it correctly and to save the measured data. But finally, even these students mostly preferred data download.

1:W(2){1749}-Sl1{2}-Sl0{4}-Sl1{3}-Sr(100){9}-Mv(-450.160156){0}-Mv(-650.160156){1}-Mv(-759.062500){0}-Mv(-775.125000){0}-Mv(-1000.000000){1}-Mv(-1000.000000){2}-Ma1{1}-Sf(0){26}-Sf(1){8}-Sf(2){4}-Sf(2){4}-Sf(1){1}-Ma1{3}-Ma0{7}-Sf(0){11}-Sl1{9}-Sf(0){4}-Ma1{9}-Mv(-566.316406){0}-Mv(-542.222656){0}-Mv(-293.253906){0}-Mv(-285.222656){1}-Mv(260.894531){0}-Mv(260.894531){1}-Mv(268.925781){0}-Mv(276.957031){0}-Mv(903.390625){0}-Mv(919.453125){2}-Mv(967.640625){0}-Mv(983.703125){0}-Mv(999.765625){0}-Mv(999.765625){6}-Mv(-237.039063){0}-Mv(501.832031){2}-Ma0{0}-Mv(999.765625){0}-Mv(999.765625){3}-Mv(991.734375){0}-Mv(991.734375){1}-Mv(999.765625){0}-Mv(999.765625){2}-Ma0{2}-Ma0{1}-Mv(-700.000000){1}-Mv(-200.000000){0}-Mv(200.000000){1}-Mv(349.238281){0}-Mv(397.425781){0}-Mv(758.828125){0}-Mv(782.921875){1}-Mv(999.765625){6}-Mv(999.765625){6}-Sf(1){4}-Sf(2){1}-Pr(28){2}-Sf(0){2}-Sf(5){3}-Sl0{3}-Sl1{36}-Sf(1){84}-Sf(1){1}-Sf(0){3}-Sf(1){23}-Sf(0){49}-Sf(0){1}-Sf(0){2}-Sv0{4}-Sl0{5}-Sv1{1}-Sf(1){31}-Sr(10){2}-Sr(100){9}-Sf(0){1}-Sf(3){4}-Sv0{0}-Sv1{81}-Pd(22){1}-Pd(26){2}-Pd(28){4}-Pd(21){6}-Pd(23){7}-Pd(13){3}-Pd(14){5}-Pd(14){28}-Pd(14){37}-Pd(22){4}-Pd(21){5}-Pd(19){3}-Pd(17){6}-Pd(18){19}-Sf(1){11}-Sl1{3}-Sl0{4}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{15}-Sl1{1}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{1}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{1}-Sl0{5}-Sl1{2}-Sl0{2}-Sl1{7}-Sl0{2}-Sl1{3}-Sl0{2}-Sl1{1}-Ps(28){0}-Pd(28){4}-Sv1{9}-Sf(1){6}-Sl0{1}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{1}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{2}-Sl0{2}-Sl1{3}-Sl1{1}-Sl0{2}-Sl1{89}-Sl0{1}-Sl1{2}-Sl0{3}-Sl1{10}-Sl0{1}-Sl1{2}-Sl0{2}-Sl1{4}-Sl0{1}-Sl1{2}-Sl0{1}-Sl1{31}-Sl0{1}-Sl1{2}-Sl0{2}-Sl1{3}-Sl0{3}-Sl1{16}-Sf(1){1}-Sf(4){1}-Sf(5){5}-Sf(4){19}-Pd(13){10}-Sf(4)-D

Figure 4. Example of an activity record when the user might have been confused by the user interface or unsure with the assignment itself. He/she just played with all control elements.

3. The credibility of pre-measured data (doesn't matter how they look like and who is their author) is very high.
4. Students do not trust to their own results. It might be associated with the learning and teaching paradigm change in general (teamwork x individual work), lack of supervision; they are not used to, and/or increased uncertainty in the virtual environment.

REFERENCES

[1] 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 http://dx.doi.org/10.1007/BF02905780

[2] 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.

[3] Lustigova, Z., Novotna,V. (2012). The role of virtual and remote labs in promoting conceptual understanding of students. In *Proc. International conference on Interactive Mobile and Computer aided Learning (IMCL),* November 6-8. Amman, Jordan, pp. 42-47.

[4] Lustigova, Z., Zelenda, S. (1996). Remote laboratory for science education. *Proceedings of the GIREP - ICPE international conference on New Ways of Teaching Physics*. June 19-22. Ljubljana. Slovenia. 260-262.

[5] Lustigova, Z., Zelenda, S. (1996) Remote Laboratory for Distance Education of Science Teachers. *Proceedings of IFIP WG 3.6 Working Conference on Collaborative Learning and Working with Telematics*. 78-85. Dec 16-18. Vienna. Austria.

[6] 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. http://dx.doi.org/10.1007/978-0-387-09729-9_56

[7] 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.

[8] 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. http://dx.doi.org/10.1109/ITI.2009.5196119

[9] Lustigova, Z. Dufresne A, Courtemanche F, et al. (2010). Acquiring Physiological Data for Automated Educational Feedback in Virtual Learning Environments. New Educational Review Vol. 21 Issue: 2 Pp. 97-109 .

[10] 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

[11] 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).

[12] 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 http://dx.doi.org/10.1007/BF02905780

AUTHORS

Z.Lustigova works as the head of the Laboratory of Online learning at Charles University in Prague, Faculty of Mathematics and Physics, V Holesovickach 4, Prague 8, Czech Republic (zdena.lustigova@.mff.cuni.cz).

P. Brom is a PhD student at Charles University in Prague, Faculty of Mathematics and Physics, Ke karlovu 3, Prague 2, Czech Republic (pavel.brom@mff.cuni.cz).

Manuscript received 11 February 2014. Published as re-submitted by the authors 04 April 2014.