Analysis of a short on-line course through logged data recording by a self-developed logging module

Péter Esztelecki
Faculty of Science and Informatics
University of Szeged
Szeged
epeter@inf.u-szeged.hu

Krisztina Tóth
TBA21 Hungary LTD
Szeged
ktoth@tba21.hu

Gábor Kőrösi
Faculty of Science and Informatics
University of Szeged
Szeged
korosig@inf.u-szeged.hu

Richárd Farkas
Faculty of Science and Informatics
University of Szeged
Szeged
rfarkas@inf.u-szeged.hu

***Abstract*—Online education has gained a wide popularity in today's global information boom. Prominent universities offer more and more online courses with modern audio-visual content, which have become available for almost everyone. The courses can be completed self-paced allowing for much more flexibility. Such a learning approach has already reformed higher education and seeks ways to penetrate into the realm of the secondary and primary education. Our team conducted a research in the higher classes of different primary schools in the Province of Vojvodina, Serbia. Participants of the Hungarian minority took part in a course named „Conscious and Safe Internet Use" and obtained a valuable knowledge with the help of videos and optional course materials. Student activities were all recorded with the help of a special self-developed logging software for later processing. Data were processed by statistical and data mining methods, whose findings are presented in this paper.**

Keywords—E-learning, Educational Data Logging, Log Data Analysis

I. INTRODUCTION

E-learning education supports computer learning and allows for a self-paced timing, which means participants do not need to wait for the beginning of school lessons but can acquire new knowledge self-paced in their free time. According to I. Elaine Allen and Jeff Seaman, those courses are identified as online courses whose content amount of at least 80 percent is delivered online [1]. Online courses allow for a cheaper and widely accessible education since broadband internet connection and infrastructure are given in many countries [2]. It is thus a fertile ground for the spreading of online courses.

Initial online courses contained only textual data with supporting charts and tables since video content could not be incorporated due to slow internet connection. Though, we must mention that in the field of distant learning, there were some trials with educational TV programs in the 60's. Nowadays, the majority of online courses contain audio-visual data with less textual content which aim at targeting learning through seeing and hearing. [3]. To gain a deeper overview of the entire online course and its completion success, it is not enough to observe the structure of the course material, content, acquired credits, or the time spent in on the platform but more importantly we have to shed some light on the participants` overall activities on the site. There are a couple of programs developed to suit this task to log student activities. In recent years, a relatively large body of research has been dedicated to characterizing students' behaviour on the basis of their logged activities [4]. Since, logging student activities require the storage of a huge amount of data we can safely talk about them as a form of Big Data which require statistical analysis and Data Mining approach. Big Data analysis is in close connection with data mining which connects the field of data mining, statistics, and artificial intelligence; furthermore, it aims at touching upon hidden relationships that would help gain new knowledge not only in the commercial field but also in the area of economics providing useful information. Danah Boyd and Kate Crawford created the following definition of Big Data: "It is the kind of data that encourages the practice of apophenia: seeing patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions" [5].

In this paper, we wish to present a self-developed MOOC program which saves student activities. Out of this volume, our aim is to demonstrate some information and useful findings that we came upon with after a thorough study. With the rise in number of online education platforms and the development of MOOC's, all the data gathered this way assign a completely new meaning to the course design. Big data allow for very exciting changes in the educational field that will revolutionize the way students learn and teachers teach [6].

II. BACKGROUND

According to Hershkovitz and Nachmias when we analyse logged data we tend to refer to three main parameters: the action taken, who took it and when [4]. Upon building our system we also took these three parameters into account. The identification of a participant was an easy task, since all the learning material were only available after logging into the platform, thus we could link an identification number to a

student. The time stamp was always at disposal from the server time making time logging also an important condition for a precise analysis. To record the action taken by a participant was a more difficult task, since it requires JavaScript event handlers.

Regarding courses the drop-out rate also plays a significant role. Carr states that it is widely agreed that drop out in online learning is higher, often by 10–20 percentage points, than in traditional learning [7]. Not only these value are important to be defined but also it is crucial to answer the WHY questions which require the study and analysis of numerous data [8]. Some of the following factors have impact on the drop-out rate: the structure of the course, the structure of the videos of a lesson (length, quality, etc.), the personality of the lecturer, socio-demographical issues, etc. By taking a look at the logged data with Data Mining methods, we can gain insight into new pieces of information.

III. Methodology

A. Population

After managing to put the final version of the logging server together in May, 2017, we announced a course with the title: Conscious and Safe Internet Use. Primary school participants from classes 5,6,7, and 8 took part in the course with a Hungarian minority background in the northern Province of Vojvodina in Serbia. Altogether 1370 learners logged into the system when the course was live. Finally, we were left with 1076 participants since some who initially signed up to the course did not fill in every test or just logged in once during the whole course. 54% of the students were females. The birth dates of the learners vary according to the following dates: 2001: 1,02%, 2002: 15,24%, 2003: 30,67%, 2004: 26,95%, 2005: 23,23 , 2006: 2,88%.

B. Curriculum and Tests

The course named Conscious and Safe Internet Use began with a pre-entry test to assess students' knowledge, which were filled in under teacher supervision in the informatics lab of a school where the participants belonged. Learners were able to provide their gender, date of birth, name and place of their school and had to answer 10 questions in connection with the upcoming course. These questions were present in the end term tests, as well, which allowed to check how much they improved by the end of the course. The learning material consisted of three main parts: Digital Footprint, Safe Use of Internet, and Online Mobbing. Each module contained a video lesson (the professional videos were shot with the help of the green box technic and were 13-14 minutes long) and there were several references to relevant literature. The time frame to watch a video and to learn the material was one week. Students thus could use their own time management to watch a video and to prepare for the end term test. Finally, they had to fill in three tests connecting to three curriculum modules, each containing 10 questions.The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font).

C. The Logging Module

With the Software Engineering Department of the University of Szeged, we developed a Moodle based logging module, which is capable of recording all the activities taken in the system for further analysis. The client side JavaScript software recognizes student activities (for example, clicking, mouse movement, video play) and stores them in an event buffer to optimize data traffic and then they are sent and saved on the server in a Mongo Database (see Fig. 1.).

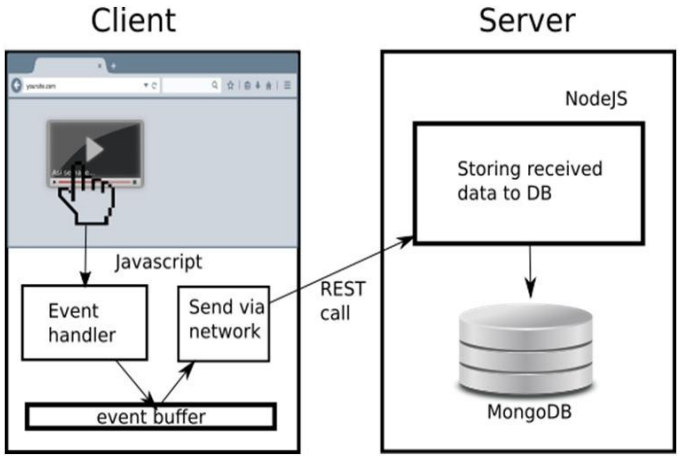

Fig. 1. Event logging

The system creates a record of the event's timestamp in every single case. It saves on which site the event occurred and which user triggered the event (before logging into the Moodle system, the users are identified as Guest users). The module records the coordinates of the mouse movement from to which the user moved the cursor, which pixel was clicked, from which position to which he scrolled, the opening and closing of a page, when was the page put into focus or was blurred, the click in an input field and characters typed in (in case of a password the characters are always hidden). Furthermore, the module records a downloaded file, a started or stopped video, a resized screen, audio volume change, and video seek. The script in a JSON file (JavaScript Object Notation) generated by the Mongo database regarding mouse movement is the following:

{ "type" : "videoPlay",
"data" : { "actualTime" : 14.878231, "videoId" : "video1", "totalTime" : 844.881269, "src" : "https://elearning.szte.hu/elearningdata/videok/tudatos_es_biztonsagos_internethasznalat/1-A_digitalis_labnyom.mp4" },
"time" : "2017.05.25. 12:29:29.855",
"page": "https://tanul.sed.hu/mod/szte/course.php?id=3...,
"user" : 2365,
"_id" : { "$oid" : "5926b20bf52e3962c1f8587c" } }

The first parameter describes the event type then come data about the actual time, id, total time and source of the video, following the timestamp, page address and the user id, and finally an entry identification code which is generated by the Mongo database system.

D. The Log File

All the events during the course, namely the 2.425.484 interpretable lines were recorded into a 674,3MB long JSON file. Most of the data processing software available is incompatible with the JSON format, while the CSV (Comma Separated Values) file types can be easily handled. Thus, we designed a PHP based software which converts logged data (see Figure 2) and the output is saved in the following way into the user-1548.csv file:

mousemove,310,1296,31,18,38.98,2017.05.29. 08:06:04.575,https://tanul.sed.hu/mod/szte/course.php?id=3...

(Remark: the last parameter of the file containing all the user information is the user id). The index page of the program allows for the choice of events to be saved into the output file using check boxes; furthermore, it is possible to choose one CSV file for every user or individual files for all the users, or the compression of the consecutive mouse movement. After hitting the Generate! button, the program reads the input file by lines and puts the entries into the output file.

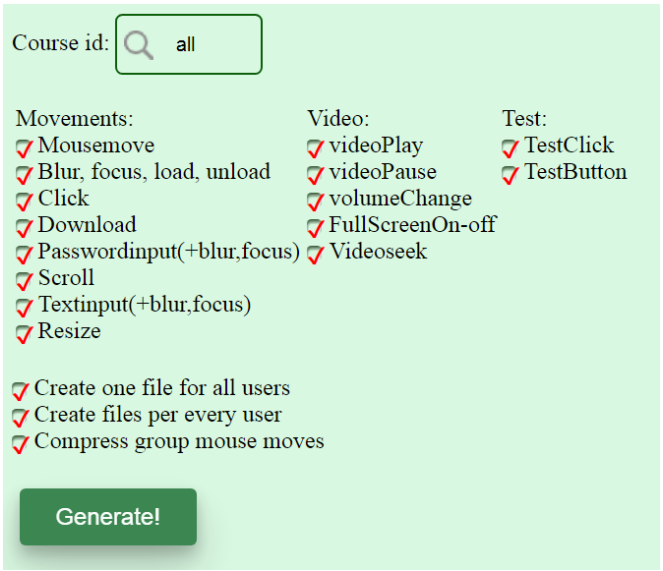

Fig. 2. Screenshot of the software for conversion

After running the code, the CSV files are closed and a summary file is also created with the number of processed data lines, events taken into account, the amount of time needed for conversion, the size of the input file and the output file respectively. Furthermore, the program attaches and saves events that belong to an individual user. If the mouse move compression was also ticked, the program creates a CSV file for every user where all the consecutive mouse movements are saved in the following format:

mousemovecompr,5,281,1278,308,1300,85.24,2017.05.29. 08:06:04.075,2017.05.29. 08:06:06.075, https://tanul.sed.hu/mod/szte/course.php?id=3...

The compression of the saved data is important because 75% of all the recorded data are mouse movements, which can be easily contracted, thus 5 times less mouse move event is saved into the output file allowing for a faster processing.

Participants had to provide some personal initial data: gender, date of birth, place of residence and the name of the school. Some false data were also recorded into the database but they were easily corrected by looking at peer data of those who filled in the paper in the same time. Regarding town names, we had to manually control and correct typical mistakes or incompletion, for example learners mistyped or were not able to write down correctly the name of their hometown, however, the most typical mistake was providing the Serbian alternative of the town names. Subsequently, we concluded that a drop-down menu would have eliminated this problem, since out of the 31 town names, the participants came up with about 100 false entries. The number of citizens has been added later to town names based on the census from 2011 and divided into four categories.

The supervising teachers had also an access to the system allowing them to watch the videos. The system, additionally, recorded the events created by the administrators. Later, these records were removed as they would lead to false conclusions. Actions taken by the Guest users were also deleted as they were rated as uninterpretable. The result files generated by the Moodle had to be also reprocessed: the points were summed up and instead of showing every answer a,b,c, or d options were recorded. Time spent on a test was not saved as a string with n minutes and m seconds but was converted into an integer with only the sum of seconds. After cleaning and converting the database, we joined the tables with SQL commands to suit further statistical and Data Mining processing. By joining the tables, we were able to assign events to individual users and to find test related data or to search for personal learner information.

IV. Results

The conversion and processing of the log file were followed by statistical and Data Mining processing.

Out of 1076participants, 632 learners filled in all the 4 tests (i.e. one pre-test and three final tests). Log data clearly show that those participants who did not fill in every test were less active during the course (see Table 1.).

TABLE I. Student Activity (N=1076)

	mousemove	bflu	click	scroll	vidplay	vidseek
4 tests completed (n=632)	1560.96	21.29	55.61	281.22	2.49	0.96
<4 tests completed (n=444)	870.6	14.82	31.02	151.69	1	0.38

The 632 learners who completed the pre-test and the three final tests had twice as more mouse movement, blur, focus, load, and unload (bflu) events. The same results were achieved with click and scroll events, while there is even a higher gap in the frequency of video play and video seek options.

Female participants had better scores at every test. The difference is in average 0.57 point (see Table 2.). The t0avg column shows the average pre-test and the t1-t2-t3avg indicates the final three tests.

TABLE II. TEST RESULTS (N=1076)

	t0avg	t1avg	t2avg	t3avg
Male (n=583)	5.05	5.03	5.44	5.12
Feale (n=493)	5.65	5.25	6.2	5.83

To pass the course, the participants had to score at least 5 points. The pre-test did not have such a prerequisite since only the level of knowledge was assessed. The course was successfully completed by 282 learners (175 girls, 107 boys) whose video play scores were 3.67 while those users who did not reach the minimum 5 points, the value was only 1.55.

By studying the correlation factors, the relationship is even more significant (at the level 0.01) when the number of clicking and mouse movement (0.475), clicking and video play (0.217), mouse movement and video play (0.317) are compared. With the analysis of the correlation factor of the number of clicking and mouse movement, their value remain under 0.1. The correlation factor between video play and the tests remain also significantly low: t0: 0.80, t1: 0.76, t2: 0.108 és t3: 0.90. However, if the same factor is taken into account between the tests (see Table 3.), we may conclude that the output test results can be more accurately predicted in comparison with the pre-test. The correlation factor between the output tests shows a high value.

TABLE III. CORRELATIONS

		t0	t1	t2	t3
t0	Pearson Correlation	1	.422**	.451**	.488**
	N	1076	730	717	674
t1	Pearson Correlation	.422**	1	.479**	.519**
	N	730	730	689	643
t2	Pearson Correlation	.451**	.479**	1	.531**
	N	717	689	717	654
t3	Pearson Correlation	.488**	.519**	.531**	1
	N	674	643	654	674

We can draw a conclusion, that those participants who scored well at the pre-test stage had expectedly a better achievement during the final testing stage. As a proof, we can look at the correlation factor between the pre-test and the end tests which is 0.545. The positive values of the correlation factor reflect this idea that higher user activities (mouse movement, video play) can be related to better test results, though these correlations are considerably weaker by comparing them to the pre-tests.

If we compare the average outcome of the three output tests with the pre-tests, the R Square statistical value is 0.297 which shows a strong correlation.

We created 4 categories from the cities where the participants come from. The correlation coefficients reveal that there is no connection between the size of a town and the number of mouse clicks, video play, or mouse movement, however, as the size of a town grows the achievement results are better at all four tests (t0: 0.104. t1: 0.94, t2: 0.163 és t3: 0.188).

Learners' age and points scored at the tests show a correlation coefficient of 0.279 which demonstrates that age plays a significant role at the test results, namely older participants had better results and are more proficient regarding course topics, as revealed during pre-testing (correlation coefficient: 0.291).

If we create three clusters taking into account mouse movements, clicking, and test results, then the algorithm sorts those cases into the first cluster (n=279) who had few activities and had worse end term results (see Fig 3..). Remark: the values were normalized because of the clustering so the table shows negative results as well. The second cluster (n=102) contains those participants who were active but had an average end term results. Finally, the third cluster (n=251) demonstrates those learners who despite the fact that had less mouse movements or clickings achieved a good result.

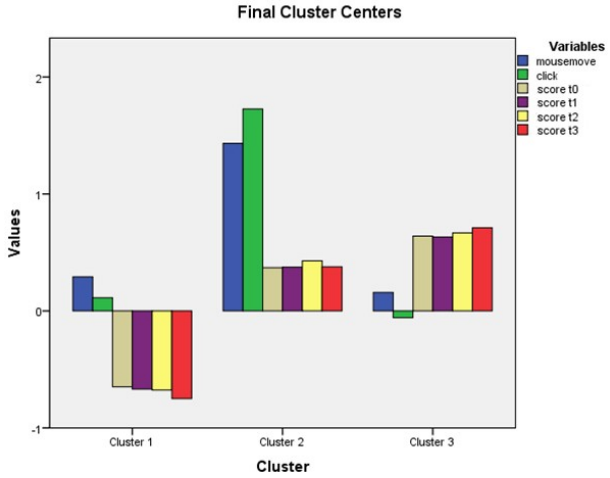

Fig. 3. Clustering

To have a better visual overview, we show here the clusters in a scatter diagram (see Fig. 4.).

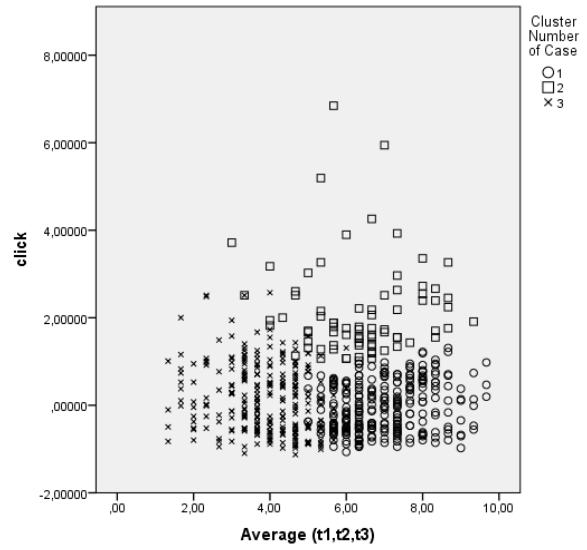

Fig. 4. Clustering

REFERENCES

[1] Allen, I. E., & Seaman, J. (2008). Staying the course: Online education in the United States 2008. Needham, MA: The Sloan Consortium. URL: https://www.onlinelearningsurvey.com/reports/staying-the-course.pdf (2017.11.10.)

[2] T. Butler, M. Haldeman, E. Laurans: Creating Sound Policy for Digital Learning. [online] Washington: Thomas B. Fordham Institute, 2012.01.11.

URL:http://www.edexcellencemedia.net/publications/2012/20120110-the-costs-of-online-learning/20120110-the-costs-of-online-learning.pdf (2018.01.20)

[3] P. Esztelecki - G. Kőrösi (2015): Idegennyelv-tanulás megvalósítása online eszközökkel. 6. Báthory-Brassai nemzetközi konferencia.URL: http://www.bbk.alfanet.eu/userspace/6bbk2015_minden/6BBK2015_Tanulmany_kotetek/6BBK_konyv-2.pdf (2018.02.01)

[4] A. Hershkovitz , R. Nachmias (2011). Online persistence in higher education web-supported courses. Internet and Higher Education 14 (2011) 98–106.

[5] Boyd, D., & Crawford, K. (2012). Critical questions for Big Data. Information, Communication & Society, 15(5), 662—679. DOI: 10.1080/1369118X.2012.678878

[6] [12] Rijmenam van M. (2013). Big Data Will Revolutionize Learning. Smart Data Collective.

URL:http://www.smartdatacollective.com/bigdatastartups/121261/big-data-will-revolutionize-learning (2018.02.01.)

[7] Carr, S. (2000). As distance education comes of age, the challenge is keeping the students. The Chronicle of Higher Education, 46(23), 39−41.

[8] J. Park (2007). Factors related to learner dropout in online learning. In: Nafukho, F. M., Chermack, T. H., & Graham, C. M. (szerk.): Proceedings of the Academy of Human Resource Development. Indianapolis: Annual Conference, 2007. 1–8 .p