A framework for data collection, analysis and evaluation of the relationship between students' computer interaction and course grades in laboratory courses

Mustafa Coşku, ***Meltem Özturan***, *Bogazici University, Turkey*

ABSTRACT. Big data is an emerging topic, with huge investments in IT and education worlds. Together with awareness of the knowledge discovery and education improvement progresses, big data concept has come with a growing consciousness. There are several educational data mining analysis methods created, evaluated and presented in the literature in this manner. But, how the educational big data for finding "intelligently" valuable results to benefit students, teachers and administrators will be collected, filtered and handled? In this study, a unique data collection, filtering and evaluation framework for educational data mining was designed, evaluated and presented. In this perspective, an IT infrastructure and process monitoring software for gathering client computers' data of 3 laboratories in a public high school was developed. Time series big data was collected during 5 months from 62 computers with this back stage working software. After filtering this data with eligible methods, resultant data was evaluated with Pearson's correlation analysis between students' rate of interactions with computers and their exam grades on laboratory courses. Results showed that students' computer interaction and their success in the courses are highly correlated.

KEYWORDS: *Educational Big Data Analysis, Educational Data Mining, Instructional Technology*

Introduction

The field of human-computer interaction (HCI) emerged in the early nineties partly as a response to changing conditions of the digital world. "HCI is an area of applied cognitive science and engineering design. It is concerned both with understanding how people make use of devices and systems that incorporate computation, and with designing new devices and systems that enhance human performance and experience" (Carol, 2009). In the education world, the term HCI can be divided into two main concepts: "teacher-computer interaction" (TCI) and "student-computer interaction" (SCI) while there are other actors of education such as parents, administrators etc. Since "learning by doing" is accepted as the most effective way of education progress, and since we are living in a digital world in which computers are indispensable, SCI is very crucial for instructional institutions. On the other hand, the current decade, second half of 2010s, is not same as nineties or the first ten of twenties for instructional technologist.

When educational scientists intend to analyze SCI today, they should cope with big data because of the new structure of digital classrooms or laboratories. When we examine the computer facilities of education institutions for today and last decade, the results might show us the big difference. For instance, a school could have one computer laboratory for Information Technology (IT) course including 10 to 20 computers in last decade. Whereas today, there are at least 2-3 laboratories for not only IT courses but also for others. Additionally, there are computers and interactive boards in the classrooms. Also, students have computers and high speed internet connection at home. These cause an important issue that SCI researches will come up with big data to be analyzed softly.

Big data is an emerging topic, with huge investments in IT and education worlds. Together with awareness of the knowledge discovery and education improvement progresses, big data concept has come with a growing consciousness. There are several educational data mining analysis methods created, evaluated and presented in the literature in this manner (Baker, Yacef, 2009). But, how the educational big data for finding "intelligently" valuable results to benefit students, teachers and administrators will be collected, filtered and handled? In this study, a unique data collection, filtering and evaluation framework for educational data mining was designed, evaluated and presented.

In the following sections, literature about SCI and big data analysis studies and related works are reviewed for drawing the main goals and hypothesis. Then, research methodology and analysis progress are stated consequently. Afterwards, the research model is deployed and findings are discussed. Lastly, conclusions and discussions of the study are stated.

Literature review & background

The historical development of instructional technology has been as fast as the technology's and digital world's speed, at all. In this perspective several developed countries started their own (adapted or unique) strategies and future plans on usage of technology in education concept. But the main problem in this progress is that technology innovations are more about hardware not process. On the other hand, successful technology integration depends on overcoming issues with staff development (Akbaba-Altun, 2006; Holland, 2001; Cooley, 2001; Swan et. al. 2002; MacNeil, Delafield, 1998), investment in hardware and software (Casey, 1995; MacNeil, Delafield, 1998), leadership (MacNeil, Delafield, 1998; Todd, 1999; Leigh, 2000; Turan, 2002; Akbaba-Altun, 2004;), curriculum (Hakkarainen, 2000; Schuttloffel, 1995), teachers' and principals' attitudes (Casey, 1995; Swan et. al. 2002), and teacher commitment (Schuttloffel, 1995).

Ministry of National Education of Turkey seriously has taken into account the education technology since 1930s (Akkoyunlu, 2002). When educational (or instructional) technology is considered, the main definition made by Commission on Instructional Technology of Turkey (1970) is: "Instructional technology means the media born of the communications revolution which can be used for instructional purposes". This old definition explains the main aim of educational technology which is employing human and nonhuman resources to bring about more effective instructions.

In addition to the important issues that have been put forward by Akbaba and Altun that affect technology integration progress, the psychological problems faced during the adoption time as three main pedagogical problems in Turkish education system related to technology usage; "lack of different materials and visuals", "failing to create necessary connections with daily life", "use of presentation method within teacher-centered settings instead of using different methods and techniques". Besides those, teachers also emphasized on the problems related to "lack of different materials and visuals" that they were facing with (Yılmaz et. al., 2016). According to them, technology integration did not seem to meet their expectations in Turkey. In other words, these did not lead to active and appropriate environments as predicted.

These problems of educational technology were recognized in Turkey as they had been in other OECD countries. Thus, two pilot educational technology problems were chosen to be solved by Ministry of National Education in vocational schools through two projects: "Empowering Vocational and Technical Education Program (MEGEP)" and "Developing the Quality of Vocational and Technical Education (METEK) Program". Those projects changed not only the curriculum and educational bureaucracy progresses but also allowed to be more and more flexible on education and technology integration period. For instance, schools had one type of financial budget and could have spent it for all type of their needs in a year period before this educational revolution. But, after those projects were started, vocational and technical institutions were allowed to spend (even earn and spend) their own compensation for different type of educational technology needs. Therefore, this flexible financial process provides a changing environment for the vocational and technical schools to be unique for achieving their institutional vision and mission. Since the financial and educational flexibility were allowed, schools could earn and spend more efficiently and effectively, thus, more and more educational technology development was done on those schools.

At the same time, Ministry of National Education started "Movement to Increase Opportunities and Technology" FATIH project for improving educational technology in all schools on November 22, 2010 (Fatih Project, 2016). In this project, firstly, at least one computer laboratory was constructed in all schools with internet connection. Also, interactive boards and tablets were started to be distributed to teachers, classrooms and students.

However, did all these technological operations and projects provide a development in education? Or can they provide? This is a big dilemma in education science. It is believed that this problem cannot be solved easily. On the other hand, today, students, teachers and classrooms are equipped with technological devices and the most considerable data can be collected, stored and analyzed to improve education progress. In the progress of educational technology development, the main aim should be analyzing and evaluating the advantages, gains and requirements. Therefore, more specifically, the main research question should be examined through vocational and technical schools (the most technologically equipped schools right now).

Research question and hypothesis

In this study the main research question is: "how the educational big data for finding *intelligently* valuable results to benefit students, teachers and administrators will be collected, filtered and handled?"
This research question brings the following sub-questions for this study:

- Can the data of SCI in courses be collected with a background software in an effective and efficient way?
- How can this data be filtered, cleaned and prepared for analysis?
- Is there a relationship between SCI and their success in laboratory courses?

Methodology

Data Collection Software

There are several software in education to detect user-computer interaction in the market such as eye-tracking, mouse-tracking, keyboard-tracking programs. But, the weakest point of these software is that they consume computer hardware resources (eg. CPU, RAM, HDD) to a great extent. Therefore, using these kind of software to gather data for analysis during education progress would not be efficient in this study. By this perspective, a simple and unique tray software has been developed by the authors of this paper to collect SCI data (Figure 1). The main applications of this software are as follows:

1. Detect and store the opening and closing time of students' computer;
2. Detect and store the opening and closing time of software on students' computer,
3. Send the data simultaneously to the remote server while executing on tray of operating system.

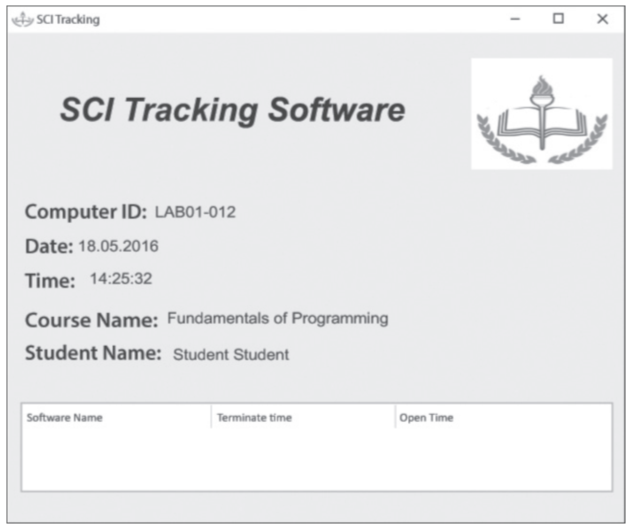

Figure 1. A Screenshot of Data Collection Software

When this software is compared with standard user tracking software, the advantages can be listed as:

- It is working on tray (hidden) which does not interrupt or annoy students;
- Its size is (362 KB) slight small compared to others;
- Since the tray programs would probably start before internet connection is established by operating system during the opening of personal computer; this program is designed to detect internet connection before trying to send data;
- It is applicable for all common versions of used operating system (Windows XP, Windows Vista, Windows 7 and Windows 10).

However, the possible disadvantages of this software are:

- Though it has simple infrastructure, it still consumes a considerable RAM storage;
- As it is mentioned by Frankel et. al. (1993), a data collection bias might probably occur in designing and evaluating research in education, if the students are allowed to unintentional behaviors or expectations on the process of data collection. And, this bias can be avoided by hidden research atmosphere. But, this hidden structure is an ethical gap for the program. Thus, at the end of the data collection progress, students are informed about the research.

Data storage remote server and structure of database

As it is mentioned in the previous section, the developed software sends all of the students' computer interactions to a remote server as time series data. This data contains computer name, time, software's name and operation information. Since there is a proxy and security license in Ministry of National Education internet service in schools, developed software cannot connect and insert data to the remote database directly. This is a common case for all school internet connections. The software has been developed to execute an API that inserts data to the remote server via an allowed internet site. This API executes a secure web site to send data by staying in background.
The structure of remote database is shown in Figure 2. As it is seen in the figure, all the students, computers and schedules are stored in master tables. Since students are not allowed to sit on different computers during the courses and all personal computers are designed same in terms of hardware and software in the laboratories; those master data have been ahead in the database before the data collection progress.

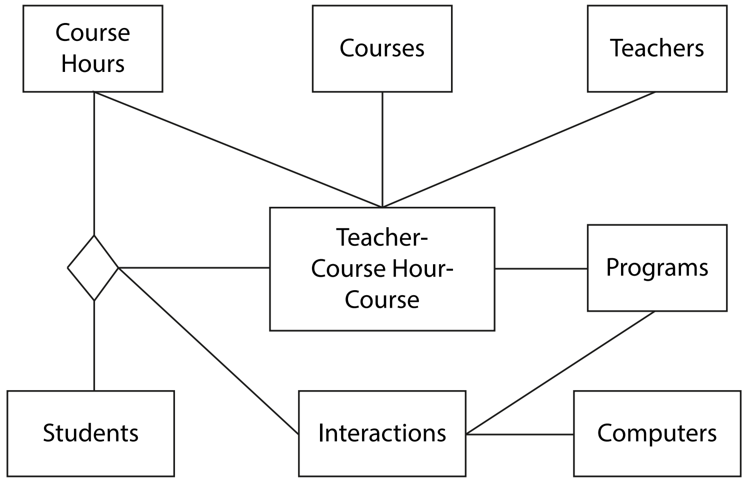

Figure 2. Structure of Database

Data collection and filtering progresses

As it is explained in the previous section, an IT infrastructure and process monitoring software has been developed for this study in order to gather client computers' data. The time series big data is collected during 5 months from 62 computers from a vocational school in Turkey using this back stage working software.
As it is seen in Figure 2, there is a transactional table (operations) designed to store students' computer operation data in the remote server database. The structure of this table can be summarized as:

- Computer ID (Foreign key of Computers table);
- Operation Type ID (Foreign key of Operations table: Start, Terminate etc.);
- Program ID (Foreign key of Programs table: Windows, ETA, Word, Excel, Visual Studio, Dreamweaver etc.);
- Time (In Unix time format for simplifying MYSQL time operations).

At the end of the fifth month, this table contains more than 20 thousand rows (operations). In order to analyze the SCI, a non-normalized table (like OLAP table) is created with the following attributes:

- Student ID: This data was gathered from operations, students, computers and course hours tables by matching operation time with course hours and students with computers;
- Operation Duration: This data was calculated by the difference (seconds) between related operation types (open and terminate);
- Software ID.

After this first operation, the data are filtered using the developed software for students and courses only. In other words, this transactional data do not constitute all software operation durations of students. Therefore to examine the relationship between SCI and their success, only the interactions of related programs with the courses are included, and the others are dropped.

Third operation for preparing the data to the analysis is on calculating the total students' computer interaction in the courses. There are several methods and units for this preparation data. Consequently, the SCI data is calculated by finding the percentage ratio of duration of students' course related software usage to total course time. For instance, weekly course hour for "Office Applications" course is 4800 seconds (2*40 minutes). There are 20 weeks in 5 months period (data collection period). Then, total course hours are calculated as 20*4800=96000 seconds. In this course, students use Word, Excel and Power Point software. As a result, to calculate SCI ratio, total duration of those software usage of related students' computer over total course duration (96000 seconds) are calculated and converted to percentage for each student and for each course. Finally, a resultant dataset is found with those variables:

- Student ID;
- Course ID;
- SCI (Percentage);
- Student Grade for that Course (over 100).

Results and findings

Demographic findings

Demographic findings of the study can be seen in Table 1 where the two departments of the chosen vocational high school are Information Technologies (IT) and Accounting & Finance (AF). There are 11 courses given in 3 laboratories in these two departments.

Course Name	Department	Level	Weekly Course Hours
Open Source Operating Systems	IT	11	2
Fundamentals of Programming	IT	10	4
Web Design and Programming	IT	11	10
Fundamentals of Information Technologies	IT	10	4
Package Programming	IT	10	2
Internet Programming	IT	12	2
Graphical Animation	IT	11	4
Database Organization	IT	11	2
Office Programs	AF	10	4
Accounting with Computer	AF	11	6
Keyboard Usage in Computer	AF	10	2

Table 1. Departments, courses, weekly course hours

To calculate total number of course hours given in laboratories, the numbers of classes for those courses (Table 2) and their weekly course hours (Table 1) are considered. Then, total number of IT course hours given in laboratories in a week is found to be 80 hours (20 hours 10th level, 54 hours 11th level and 6 hours 12th level) and total number of AF course hours given in laboratories in a week is found to be 84 hours (42 hours 10th level and 42 hours 11th level).
The number of students in terms of departments and levels can be listed as in Table 2. As it can be seen in Table 2, total number of students in the sample is 641.

Department	Level	Classes	Number of Students
Information Technologies (IT)	10	2 (A, B)	72
Information Technologies (IT)	11	3 (A, B, C)	70
Information Technologies (IT)	12	2 (A, B, C)	73
Accounting and Finance (AF)	10	7 (A.......G)	214
Accounting and Finance (AF)	11	7 (A.......G)	212
		TOTAL	**641**

Table 2. Number of Students

The average number of operations per student and students' average interactive duration are listed in Table 3.

	IT (215)	AF (426)	All (641)
Average number of operations per student in a course	3.22	1.16	2.48
Average student interactivity duration of computers (per 40 min. course hour)	31' 22"	18' 14"	24' 37"
Average student interactivity duration of computers (percentage over 40 min.)	78%	45%	61%
Average student grades on laboratory courses (over 100)	68	79	76

Table 3. Average Numbers of SCI

Relation between student success and student computer interaction

Pearson Product Moment Correlation (PPMC or simply Pearson's Correlation) is used for analyzing the linear relationship between two continuous dataset. In other words, this statistical analysis method is used for answering the question "is there a significant relationship between these two variables?" The resultant coefficient, which is calculated by dividing the covariance of two variables by the product of their standard deviations, gives the idea for the size of relationship.
In this part of the study Pearson's correlation coefficient of two variables "Student Success" and "Student Computer Interaction" is calculated by SPSS Statistics (Version 21). Results of the analysis are summarized in Table 4 where "N" shows the number of students in the department used for analysis, "r" shows correlation value and "p" shows level of significant value from Pearson's correlation 2-tailed test.

Department	N	r	p
IT	215	0.821	0.001
AF	426	0.637	0.001
Both	641	0.697	0.001

Table 4. Results of Pearson's Correlation Analysis

Chan (2003) states the power of linear correlation results as in Table 5.

Coefficient Interval	Power
0.80 and above	Very Strong
0.60 – 0.79	Moderately Strong
0.30 – 0.59	Fair
0.29 and below	Poor

Table 5. Levels of Pearson's Correlation Coefficients

When the results of the analysis between students grades of the laboratory courses and their computer interaction percentages are investigated, it can be seen that:

- In both departmental courses (and together), there are significant positive correlations between students' grades and students' computer interaction rates. This result means that the higher the student's computer interaction (in laboratory courses) the higher the grade of him/her;
- The correlation coefficient of the students in IT department has "very strong" power while the correlation coefficient of the ones' in AF department has "moderately strong" level. This result shows that the importance of computer interaction in IT department for students is higher than the ones in AF department for their success.

Conclusions and discussion

This study is designed for analyzing the possible relationship between SCI and their grades in computer laboratory courses. To this respect, data collection and filtering sequences of the analysis are designed and proposed for an educational data mining. Therefore a main research question is proposed as "how the educational big data for finding *intelligently* valuable results to benefit students, teachers and administrators will be collected, filtered and handled?"

In order to collect students' computer interaction data for the correlation analysis, an unique tracking software is developed. This software is installed in 62 computers in 3 laboratories of a vocational high school. By this software all computer operations of 641 students during 11 different laboratory based courses are collected to a database stored on a remote server. After data collection, by filtering and eliminating noisy data, SCI times are calculated. Lastly, the resultant times for every course and every student are converted to the percentage by finding the ratio of SCI time to actual course duration.

The relationship between students' computer interaction percentages and course grades (over 100) is examined by Pearson's Correlation analysis. The results showed that there are significant correlations between these two variables in 99% confidence level. The highest correlation is detected in Information Technology department courses (0.821), while the correlation coefficient is 0.637 in Accounting and Finance department courses.

The main contribution of this study is that a strong positive correlation is detected between SCI and course grade. Thus, true design of computer interaction (with truly essential software for the course) increases the course grades of students. But, this does not mean students should be allowed for free computer usage in the course time, because only the related software usage duration is used for the analysis. In other words, not all the program usage duration of the students positively (and significantly) affects the student grades, but only the related program usage duration does.

The sample of the study might be the main limitation of the study. In the future studies, it is advisable to enlarge the sample and this designed tracking software can be used to collect data from different types of schools (e.g. not only vocational ones) and laboratory courses.

References

Akbaba-Altun Sadegül (2005), *Bilgi teknolojisi sınıflarında denetim*, "Eurosian Journal of Educational Research", V. 5, n. 18, pp. 1-16

Akbaba-Altun Sadegül (2006), *Complexity of integrating computer technologies into education in Turkey*, "Educational Technology & Society", V. 9, n. 1, pp. 176-187

Akkoyunlu Buket (2002), *Educational technology in Turkey: Past, present and future*, "Educational Media International", V. 39, n. 2, pp. 165-174

Baker Ryan S., Yacef Kalina (2009), *The state of educational data mining in 2009: A review and future visions*, "JEDM-Journal of Educational Data Mining", V. 1, n. 1, pp. 3-17

Carroll John M. (2009), *Human–computer interaction*, "The Encyclopedia of Human Computer Interaction", Denmark, Interaction Design Foundation

Casey Patrick J. (1995), *Presenting teachers with a model for technological innovation*, "Technology and Teacher Education Annual", pp. 855-858

Chan Y. H. (2003), *Biostatistics 104: correlational analysis*, "Singapore Medical Journal", V. 44, n. 12, pp. 614-9

Commission on Instructional Technology (1970), *To improve learning: A report to the President and the Congress of the United States*, Washington, DC, USA, U.S. Government Printing Office

Cooley Van E. (2001), *Implementing technology using the teachers as trainers staff development model*, "Journal of Technology and Teacher Education", V. 9, n. 2, pp. 269-284

Earle Rodley S. (2002), *The integration of instructional technology into public education: Promises and challenges*, "Educational Technology", V. 42, n. 1, pp. 5-13

Fraenkel Jack, Wallen Norman, Hyun Helen (1993), *How to design and evaluate research in education*. Vol. 7. New York, NY, USA, McGraw-Hill

Hakkarainen Kai (2000), *Students' skills and practices of using ICT: Results of a national assessment in Finland*, "Computers and Education", V. 34, n. 2, pp. 103-117

Holland Patricia E. (2001), *Professional development in technology: Catalyst for school reform*, "Journal of Technology and Teacher Education", V. 9, n. 2, pp. 245-267

Leigh Gabrielle (2000), *Key markers in Victoria's information technology journey into the knowledge age*, "Australian Educational Computing", V. 15, n. 1, pp. 7-12

MacNeil Angus, Delafield Dennis (1998), *Principal leadership for successful school technology implementations*. Proceedings of the 9th International Conference of Society for Information Technology and Teacher Education, 10-14 March, Washington, DC, USA

Schuttloffel M. J. (1995), *A handbook for technology implementation*, "Technology and Teacher Education Annual", pp. 859-862

Swan Karen, Holmes Aliya, Vargas Juan D., Jenning Sybillyn, Meier Ellen, Rubenfeld Laster (2002), *Situated professional development and technology integration: The capital area technology and inquiry in education (CATIE) mentoring program*, "Journal of Technology and Teacher Education", V. 10, n. 2, pp. 169-190

Todd Ross J. (1999), *Transformational leadership and transformational learning: Information literacy and the World Wide Web*, "NASSP Bulletin", V. 83, n. 605, pp. 4-12

Turan Selahattin (2002), *Teknolojinin okul yönetiminde etkin kullanımında okul yöneticisinin rolü*, "Eğitim Yönetimi", V. 30, pp. 271-281

Yılmaz Rezan, Şendurur Emine, Şendurur Polat (2016), *Transition from theory to practice: a comparison of math and computer education and instructional technology teachers during school experience course*, "Kastamonu Eğitim Dergisi", V. 24, n. 3, pp. 1163-1178

Sintesi

Nella nostra attuale società il progresso tecnologico è molto spesso direttamente correlato con l'evoluzione informatica e infrastrutturale che viviamo. Proprio in tal senso, sempre più spesso, sentiamo parlare di Internet delle Cose, IoT, e Big Data proprio per indicare la possibilità di far dialogare tra loro tutte le "cose" attraverso la rete scambiandosi, istante per istante, una grande mole di dati. Quali vantaggi potremmo trarne? Sistemi pensanti, ambienti che si adattano alle persone, infrastrutture mobili che offrono servizi personalizzati e diversi. In realtà la lista sarebbe infinita ed è ragionevole credere che ancora oggi non riusciamo a cogliere tutte le possibilità offerte da questa innovazione.

L'articolo che andiamo ora a presentare, si occupa proprio di analizzare un caso studio su un'applicazione di Big Data. In particolare, i dati relativi all'utilizzo del PC da parte degli studenti sono stati raccolti nelle aule di tre laboratori didattici cercando di mettere in relazione queste informazioni con i risultati dei discenti. Dopo un'accurata e formale analisi sono emersi dati interessanti che meritano una riflessione profonda non solo sulla mera applicazione ma, come accennato, sulle enormi potenzialità che questi sistemi possono avere per creare una "didattica incentrata sullo studente".