JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

ANALYZING CONTEXTUAL DATA IN EDUCATIONAL CONTEXT: EDUCATIONAL DATA MINING AND LOGFILE ANALYSIS

Saleh Alrababah* and Gyöngyvér Molnár**

* Doctoral School of Education, University of Szeged, Szeged, Hungary
Email: alrababah.saleh.ahmad@edu.u-szeged.hu
**Department of Learning and Instruction, MTA–SZTE Research Group on the Development of Competencies, University of Szeged, Szeged, Hungary
Email: gymolnar@edpsy.u-szeged.hu

In recent decades,attention has been focused on analyzing contextual data in educational contextusingeducational data mining (EDM), which is a process of employing data mining techniques to transform initial data collected throughеducational systems into meaningful information. Logfile analysis entails analyzing behavioral processes, time-on-task, and the sequence of actions captured in logfiles and thusintroduces novel methods to the analysis of the instruction and learning process as well as toeducational assessment.This research paper aims to present the developmental trendsinEDM techniques and logfile analysis in educational context and their contribution to a better understanding of the contextual data, collected beyond the particular responsedata.We conducted a comparison analysis based on the Scopus database to show developmental trends by year and domain. According to the results, (1) research interestin this field has grown immensely in the last few years; that is, EDM is an emerging discipline. In addition, (2) EDM and logfile analysis examine earlier hidden informationto provide explanations ofstudents' learning and testing behaviour from a new perspective, thus broadeningour understanding of students' behavior, interests, learning processes, motivational aspects, and test results and the reason for their learning outcomes.
Keywords: data mining, educational data mining, contextual data,logfile analyses

Introduction

Improvements in the computer and data sciences have created new methods and possibilities from whicheducation can benefit. Data mining (DM) is a crucial data analysis methodology that has been used in many domains successfully (Tanimoto, 2007). Recently, the expansion of computer-based testing and training in education (Molnár & Csapó, 2019a,2019b; Mousa & Molnár, 2019) has made informationavailable (Molnár & Csapó, 2018;Tanimoto, 2007; Wu,& Molnár, 2019), which can provide important indicators of students' learning processes but were hidden by traditional assessment and training methods. Nowadays, both educationaldata mining (EDM) and logfile analysishave become a state-of-the art area of educational assessment, as theycan build the basis for new learning theories and new knowledge about how students' learn, behave, complete,and interact with the tasks and problems administered to them (see e.g. Al-Kabi, Shannaq,& Alsmadi, 2011; Chen & Chen, 2009; Guo, Deane, van Rijn, Zhang, & Bennett, 2018; Minaei,Kortemeyer,& Punch, 2004; Mylonas, Tzouveli,& Kollias, 2004; Romero,Ventura,& García,2008; Xing, Guo, Petakovic,& Goggins, 2015; Zaiane & Luo, 2001; Zhang, Bennett, Deane,& van Rijn, 2019). Based on the literature in the field of assessment, the most commonly employed techniques are time-on-task analysis (see e.g. Alzoubi, Fossati, Di Eugenio, Green & Chen, 2013;Goldhammer, Naumann, Stelter, Tóth, Rölke, & Klieme, 2014; Greiff, Niepel, Scherer,& Martin, 2016; Naumann,2019;Zoanetti & Griffin, 2017) and analysis of students'exploration and problem-solving behavior in connection with complex problems developed throughthe MicroDyn approach (seee.g.Goldhammer & Barkow, 2017; Greiff, Wüstenberg,& Avvisati, 2015; Molnár & Csapó, 2018; Tóth, Rölke,Wu,& Molnár, 2019). The possibilities are unlimited, but new methods and new models are needed to use these possibilities in research and educational practice (Molnár & Csapó, 2019b).

This paper summarizes and evaluates the main studies in this field in so as to present a theoretical framework for using contextual data in educational context. It visualizes how contextual data analyses have been utilized in different contextsover the years. The basis forthe meta-analysisconsisted ofpapers available in Scopus databases. This study is expected to provide both researchers and educators with information about the possibilities oflogfile analysis and highlight the importance of using and analyzing contextual data for a deeper understanding of learning

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

processes.Furthermore, the paper presents how logfile analysis has been used in the educational process with outstanding results, especially in assessment. It has been employed in exploring the relationship between time-on-task and students' performance, exploring students' strategies and behaviors on problem-solving tests as well.

1. Developmental trendsand challenges in EDM

1.1. Data mining

DM is a multidisciplinary field which is considered to be a branch of computer science. It is regarded as an exploratory process, but it could be used forconfirmatory investigations (Cheng, 2017). Algarni (2016) views DM as a logical step of the knowledge discovery in database operation. DM methods have links to artificial intelligence, machine learning, statistics, and computer science (Dunham, 2006). However, it is somehow different from other search and analysismethods because it is exploratory, while other analyses are more confirmatory.Hidden patterns can be uncovered with acombination of an explicit knowledge rule, domain knowledge, and advanced analytical skills. As a consequence, the detected patterns and trends can assist organizations with meaningful information and guide their decision-making(Kiron, Shockley, Kruschwitz, Finch,& Haydock, 2011).

DM is a useful artificial intelligence tool that has the potential to uncover helpful information by analyzing data from many dimensions, classifying the information, and summarizing the specified relationships in the database (Algarni, 2016).Afterward, this information assistance makes or improves decisions. In DM solutions, algorithms can be employed to achieve the required results. A case in point is clustering algorithms that recognize modes and group data into varying groups. The data in each group vary from high to less consistent. Based on this, the findings can assist in creating a better decision model, while multiple algorithms implement one solution as they can conduct separate functions. For instance, using a regression tree methodmakes it possible to obtain financial predictions or association norms to conduct a market analysis (Algarni, 2016).

Nowadays, a significant amount of data in databases surpassesthe human ability to extract and analyze the most useful information without the aid of new analysis techniques (Cheng, 2017). Knowledge discovery is the process of significant extraction of involved, unidentified, and potentially meaningful information from a big database (Kiron et al., 2011).

Data mining employed in knowledge discovery has revealed patterns(Algarni, 2016). The precise discovery of patterns through DM is affected by various factors, such assample size, data fairness, and endorsement of the knowledge involved. All of them affect the degree of certainty that is suited to determining patterns (Dunham, 2006). Even though DM uncovers various patterns in databases, some of them can be interesting from the point of view of learning. It is also crucial to examine the extent of confidence in a given pattern when rating the validity of the results. Loops can take place betweenany two steps in the process, an area whichcalls for more iteration(Algarni,2016).

In using data mining techniques, the following steps are required: (1) developing an understanding of the domain under examination, building relevant prior knowledge, and defining the objective of the analysis;(2) understanding the structure of the target dataset by focusing on the subset of data/variables which are targeted in the analysis; (3) cleaning and pre-processing databases by eliminating noise, developing methods for dealing with missing data, and accounting for time sequence details and documented changes; (4) the data reduction and projectionstage, e.g. reduction of dimensionality or methods of transformation; (5) selecting the best fitting DM techniques according to what we callKnowledgeDiscoverygoals; (6) fittingDMalgorithms to the dataset to identify patterns; (7) deriving meaningful patterns from a specific form or collection of representations; and finally, before reporting, (8) explaining these mined and detected patterns and/or returning to previous steps for further iteration(Algarni,2016).

1.2.Educational data mining (EDM)

EDM has become greatly important in research interests and possibilitiesrecently (Amrieh, Hamtini, &Aljarah, 2016). It is a growing field which uses DM techniques and methodstoanalyze and extract hidden and unknown knowledge from educational data (Amrieh et al., 2016; Baker &Yacef, 2009; Dahiya, 2018; Romero et al., 2008). Silva and Fonseca (2017) defined EDM as a process of transforming initial data collected throughеducational systems into meaningful information that can be utilized to make decisionsand answer research questions.Further issues, such as time, serialization, and context, also play a significant role in analyzing educational contextual data (see e.g. time-on-task; Fatima, Fatima, & Prasad, 2015). To sum up, educational data mining is a multidisciplinary field, including and combining knowledge froma number ofdomains, such as information retrieval, recommender systems, visual data analytics,data visualization, social network analyses, cognitive psychology,and psychometrics (Cheng, 2017),machine instruction and learning (Baker & Yacef, 2009),and process mining(Juhaňák, Zounek,& Rohlíková, 2019).

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

The knowledge discovered by EDM through DM techniques and complex analysis (Amrieh et al., 2016) can assist educational researchers in developing new learning and teaching techniques, understanding learners' behavior,anddesigning more effective, motivating, and supportivelearning environments, that is, enhancing the learning process. It can assist in producing high-quality research results (Amrieh et al., 2016). Dahiya (2018) divided the objectives of EDM into threecategories: (1) educational or academic objectives (e.g. designing educational content); (2) administrative or management objectives (e.g. the maintenance of educational infrastructure), and (3) commercial or the market objectives (e.g. capturing the market in terms of enrolments).EDM also employs unique ways, methods, and techniques to address relevant educational problems and questions (Cheng, 2017), such as predicting students' performance (Miguéis, Freitas, Garcia,& Silva, 2018) and predicting students' academic failures (Arora, Singhal & Bansal, 2014; Costa, Fonseca, Santana, Araújo & Rego, 2017; Manhaes, da Cruz,& Zimbrão, 2014).

1.3.Source of the data used in EDM

A database, which can build the basis for educational data mining analysis,can be collected from different sources, such as computer-based assessments (Molnár &Csapó, 2018), web-based education systems, and online learning management systems, such as Moodle (Silva & Fonseca, 2017), educational repositories,or traditional surveys (Amrieh et al., 2016). The structuresof these databases differ greatly, that is, the "one size fits all" approach cannot be applied in these types of analyses. Every database is unique and requires "personalized" methods.

1.4. Methods

DM is a powerful technology with great capacity to map student's behavior beyond particular answer data (Silva & Fonseca, 2017) by developing methods to explore large data sets collected in educational settings in order to gain a better understanding of students' behavior, motivation, interests, and results. For educational issues and problems, EDM uses DM techniques and tools to discover unknown relationships and patterns in large-scale data. These techniques and tools involve machine learning methods, mathematical algorithms, and statistical models, such as clustering and decision trees. They also detect information within the data which queries and reports based exclusively on response data cannot effectively detect (Amrieh et al., 2016; Silva & Fonseca, 2017).

1.5.The use of educational data mining to evaluate students' behavior and actions

EDM is utilized to analyze students' behavior to detect the significantly different ways and learning behavior patterns of students inlearning management systems(Kularbphettong, 2017). Evaluating activities and understanding behavior patterns offer new developmental perspectives (Rodrigues, Isotani, & Zarate, 2018).

Romero, Ventura, and García (2008) summarized the benefits of DM techniques and methods in connection with a management system courseat the University of Cordoba, Spain. A total of 438 students took part in the research. EDM techniques assisted in classifying students and detecting the sources of any contradiction values from student activities.

Zaiane and Luo (2001) attempted to analyse 395university students' activities in Britainand Canadato identify typical behavioral patterns from access logs and activities used by the students. They extracted meaningful behavioral patterns based on weblog analysis, which assisted teachers in their students' evaluation and pointed out students' needstoward improving their learning outcomes.

Minaei et al. (2004) analyzed data extracted from an online educational systemdeveloped at Michigan State University with Computer-Assisted Personalized Approachto detect and define profiles for different students based on their problem-solving behavior and to develop and recommend decision-making strategies that best fit the different profiles. Association rules were used to detectpatterns and cluster profiles.

Chen, Chen,and Liu (2007) assessed the learning outcomesof students by analyzing data on their interactions and access to different educational media. The aim was to provide better insights for teachers into the main factors affecting students' learning outcomes. To achieve these aims, Grey's relational analyses, clustering techniques, fuzzy association rule techniques, and fuzzy inference algorithms were used. The information extracted by these algorithms aided in developing better educational strategy plans and designing more effective educational materials.

There are no definitive techniques to assess performance during the e-learning process; typically, the final outcomes are given (Mylonas et al.,2004). Chen and Chen (2009) designed an online system using statistical association analyses and fuzzy clustering methods, among other techniques. The main aim was to develop a system which uses formative assessment techniques to profile students and personalize pedagogical materials. This system allows teachers and researchers todetect factors that impact students' learning for better curriculum developent and to find, fit, and personalizethe best teaching strategy. The results confirm the importance of performing inferences in

the individual performance of studentsand to evaluate educational strategies recommended to reduce students'failing and dropping out.

Miguéiset al.(2018) utilized EDM techniques to predict graduating students' overall academic performance. Barros, Neto, Plácido, Silva,and Guedes(2019) also used EDM techniques (e.g. decision tree, neural networks, and Balanced Bagging) to predict dropout ratesin higher education in Brazil.

Process miningis among the basic EDM techniques. It is attracting increasing research attention (Juhaňák, Zounek,& Rohlíková, 2019; Reimann & Yacef,2013). Reimann, Markauskaite, and Bannert (2014) focused on process mining in data-intensive research methods from the perspective of methodological challenges.

Schoor and Bannert (2012) analyzed empirical studies using process mining techniques in the context of computer-supported collaborative learning to map social organizational processes.They concluded that these methods are helpfulto gain insights into the process of learning and recommendedthem for further analyses. Bannert, Reimann, and Sonnenberg (2014) used a process miningtechnique in self-regulated learning to analyze qualitative data collected through a think-aloud protocol. Sedrakyan, De Weerdt,and Snoeck (2016) employedprocess mining methods in connection with complex problem-solving data to detect students' learning behavior patterns and to link identified profiles to students' learning outcomes. They concluded that the use of such methods was highly beneficial to monitor and analyze cognitive learning processes.In the virtual learning environment,Vidal,Azquez-Barreiros, Lama, and Mucientes(2016) made use ofthesetechniques to analyze log event recordingsof students' and teachers' behavior.

Romero, Cerezo, Bogarín, and Anchez-Santill (2016) applied process mining techniques in the analysis of data collected via the Learning Management System (LMS) Moodle, while Papamitsiou and Economides (2016) focused on quiz-taking behavior through a process mining approach by identifying patterns of guessing behavior during testing. They suggested that process mining can provide new opportunities in modelingand detecting various types of students' quiz-taking and guessing behavior in online learning and test environments.

The application and development of DM in educationare limited and fairly late compared toother fields, e.g. the life sciencesandbusiness administration. Indeed, EDM faces many challenges. One of these arises from the specific attributes of data (Silva & Fonseca, 2017). However,growth can be observed in the amount of educational data, thus opening doors for EDM analysis.Nowadays, EDM has becomea significant aspect of analysis, the basis of further developments in educational research and practice (Dahiya, 2018).

2. Developmental trends and challenges in logfile analysis

The emergence of computers in the psychological and educational context has enabled us to analyze the behavioral processes and sequence of actions through the information stored in logfiles (Greiff et al., 2015). Analysis of such data as the interaction between the user and the computer system dates back to 1967 (Al-Kabi et al., 2011).

In the context of computer-based assessment, student interactions duringtasks are recorded easily to produce logfiles. These records typically describe keystrokes and mouse events (cursor movement, clicking, dropping, typing, dragging, etc.). Therefore, each separate processis typically logged using a timestamp or a similar occurrence time (Zoanetti & Griffin, 2017). The logfilescan be used as a direct source of information containing data on each behavioral action for each student (Wu & Molnar, 2019). The challenge here is how to make sense of this massive amount of data (Zoanetti & Griffin, 2017). Contextual data introduces new possibilities and raises new research questions for a better understanding of the educational phenomenon under examination (Molnár & Csapó, 2019).

There are many challenges in working with logfiles across a number ofproblem domains, such as event distribution and data size.A wide range of logfile formats, and different transport and storage methods exist. Usually, it is not possible to manually scan all of the logfile data to identify patterns, anomalies, or different tendencies in the data (Green,2015).

2.1.A review of empirical studies usingtime-on-task analysis

Time-on-task can be defined as the amount of time spent on task completion. It involves the time spent identifying the task, the time spent completing it, and the time spent entering the solution. Ontasks that require participants to take multiple steps and/orto interact with the problem environment, we can define time-on-task in smaller steps, that is, the time between each click.

There are two different approaches to modeling time-on-task. Time-on-task can be taken as an indicator of the latent construct or it can be employed in explanatorymodels as a predictor to explain differences in performance (Goldhammer et al., 2014).

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

Nowadays, most studies focus on the second approach and test the effect of time-on-task on the academic performance of students in a computer-based learning environment (Lee, 2018). Lustria (2007) argued that interactivity can significantly affect comprehension as well as attitudes towards health Web sites, that isundergraduate students who spent more time using interactive websites on health-related information achieved significantly higheron a related achievement test. Louw, Muller, and Tredoux (2008) analyzed the predictive power of differentvariables, such as previously existing mathematics knowledge, computer access outside of school, time spent on computers both outside and inside of school, confidence in using information technology, computer literacy, degree of enjoyment in learning mathematics, intention to study after school, motivation towardmathematics, parental encouragement, language used at home, and time spent with the computer-based tutoring system employed in the study. According to their research findings, time spent in the computer-based tutoring environment was the most influential predictive factor for academic successamong participating students.

Krause, Stark, and Mandl (2009) studied the learning behavior of 137 undergraduate students engaged instudying statistics coursein a computer-based learning environment. They found that time-on-task significantly correlated with the students' learning outcomes. Macfadyen and Dawson (2010) analyzed log data for a learning management system. They concluded that the number of log-ins and time spent in the learning management system explain more than 30% of the variance of the final grade earned bythe participating undergraduate students. Cho and Shen (2013) confirmed this result and reported that time-on-task logged in a learning management system, along with effort regulation, predicted students' academic achievement.

Landers and Landers (2015) focused on the predictive power of time-on-task in a game-based learning environmentforacademic achievement. They reported that undergraduate students engaged in learning in a game-based learning environment spent much more time on the school material than their peers working in a non-game-based learning environment. The additional time improved their academic performance significantly. These studies unanimously confirmed that time-on-task in a computer-based learning environment is a significant predictor for academic success.

In the field of educational assessment until the 21st century,the main focus was almost exclusively on the realization of reliable and valid tests and test-level achievement. With traditional assessment, it was not necessaryto collect contextual data, which could be important indicators ofbehavioral processes that lead to students'final performance. Contextual data can be recorded via technology-based assessment systems, which can help the researcher to extract descriptors of theoretical importance to the task completion process (Goldhammer et al., 2014) and time spent on the task (or part of the task) during testing (Zoanetti & Griffin, 2017). Previous researchhas indicated that it is a challenging task to define the predictive power of time-on-task to task performance (Naumann, 2019). Spending too much time on complex problem-solving tasks was associated with poor performance (Greiff et al., 2016). In contrast, Alzoubi et al. (2013) observedthat if students spent more time on ataskon a problem-solving test, theirperformance became significantly better. Therefore, we would expect that a longer time spent on a problem-solving task results in higher achievement, meaning a longer time allows for longer planning and better planned solutions. Notably, according to the research results, for weak problem-solvers, spending more time on a task may be helpful in compensating for the lack in reading or computer usage (Goldhammer et al., 2014). That is, the keyquestion is ifstudents' attitudes affect their time in completing a task, which may affect their performance and skills in the target domain (Naumann,2019).

Naumann (2019) examined three variables in digital reading tasks, comprehension skills, reading enjoyment, and reading strategies,whichcan predict students' performance. It also focused on the time students spend on digital reading tasks with varying difficulty. It analyzed computer-based assessment data retrieved in PISA 2009. Two indicators were built by the researcher in connection with time-on-task: the exact time spent by students on a task and the average time students spent on each task. The study found positive correlations between enjoyment of reading, comprehension skills, and reading strategies knowledge, and concluded that students with high comprehension skills, enjoyment of reading, and reading strategies were better at dealing with task difficulty with respect to their time usage than thosewho were lessskilled in theseareas.

Goldhammer et al. (2014) analyzed the relations between time-on-task and students' achievement in reading and problem-solving. They used the (N=1020) International Assessment of Adult Competencies (PIAAC) computerized reading and problem-solving test data collected in Germany. The confirmed resultsforitem difficulty, namely, time-on-task, increased with item difficulty.In problem-solving, time-on-task correlated positively with item difficulty, while with reading items, which called for more routine processing, time spent on tasks negatively correlated with achievement. They concluded that there is no common interpretation between time-on-task and achievement.

Scherer, Greiff, and Hautamäki (2015) distinguished CPS time-on-task and CPS ability with a positive correlation. According to their findings, cognitive and motivational factors have different predictive power on CPS

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

time-on-task and CPS ability. As the developmental level of CPS skills isnot only influenced by that of thinking skills (Goldhammer et al., 2014), the development of both cognitive and affective factors can be seen as a significant educational goal. "Understanding CPS time-on-task and CPS ability is therefore crucial for making progress in the conceptualization and assessment of CPS within and beyond educational context" (Scherer, Greiff,& Hautamäki, 2015, p. 47).

2.2. Analyzing learning processes

A unique challenge is to understand the hundreds of pieces of information that students may produce when participating in complex assessments. Questions such as (1) how to definethese different pieces of information, (2) what makes sense, and (3) how to combine them intoanevidence-based approach may have meaning in connection with auto-judging, fluency, or motivation, generally speaking, in educational context (Csapó, Ainley, Bennett, Latour & Lawet, 2012). Using logfile analysis can lend these kinds of data meaning and importance, and makes it possible to interpret them (Molnár & Csapó,2018). One of the advantages of this broad shift is that a large amount of the observable behavior is stored in logfiles created by the computer and can be accessed to provide additional information about students' behaviour, how they completed the tasks, and how they behavedduring testing, thusofferingmore detailed information beyond the actual test scores. Since computers are available for assessment, the potential for this almost unlimited amount of information has been widely praised(Eichmann, Goldhammer, Greiff, Brandhuber,& Naumann, 2020).

Assessment and learning can be integrated through direct feedback and customized interventions from an educational science perspective. The emergence of different generations of computerized tests has finally led to intelligent measurement, a description of the overall integration of behavioral processes to assess students' skills while students engage in learning using computers. This vision is consistent with a description of logfile-driven integration of learning and computer-based assessment (Greiff et al., 2016). Log data analysis is considered a new approach for the assessment of learning to learn from a self-regulative and motivational perspective(Vainikainen, 2019).

Logfile analysis has introducednumerous modern methods for analyzing learning processes (Molnar & Csapó, 2018). One of them was used by Zhang et al. (2019). They recorded data onstudents' keystroke logs extracted from writing processes to monitor gender-level differences. The results indicated a female advantage in writing essays and highlighted the gender-level differences in the writing process. In the same line, Guo et al. (2018) also used keystroke logs to distinguishprocesses used by students in essay composition.

3.Growingresearch interest in logfile analysis and educational data mining: analyses based on papers indexed by Scopus

The sample for these analyses consisted ofpaperspublished between 1966 and 2019 and indexed by Scopus. We used the following keywords during the filtering process: contextual data, log file analysis, data mining, and educational data mining. To detect the significance of contextual data analysis in educational research, we investigated the history and main topics in logfile analysis and educational data mining that have been tackled.

We met 167,563 records, distributed as follows according to the keywords*contextual data*(1249), *log file* or *log-file* or *logfile analysis*(405) (with different spelling),*data mining*(163,496), and *educational data mining*(2008).The huge number of studies confirmed that there is great research interest in working with contextual data.

Figure 1 shows the almost linear trend of how the number of data mining papers in Scopus has risenin the last twenty-five years. In 1995,there were one hundredsuch publications, while that number was 15,466 in 2019 (please note that this data only refers tothe Scopus databases). It must also be noted that the very first study was published in 1966.

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

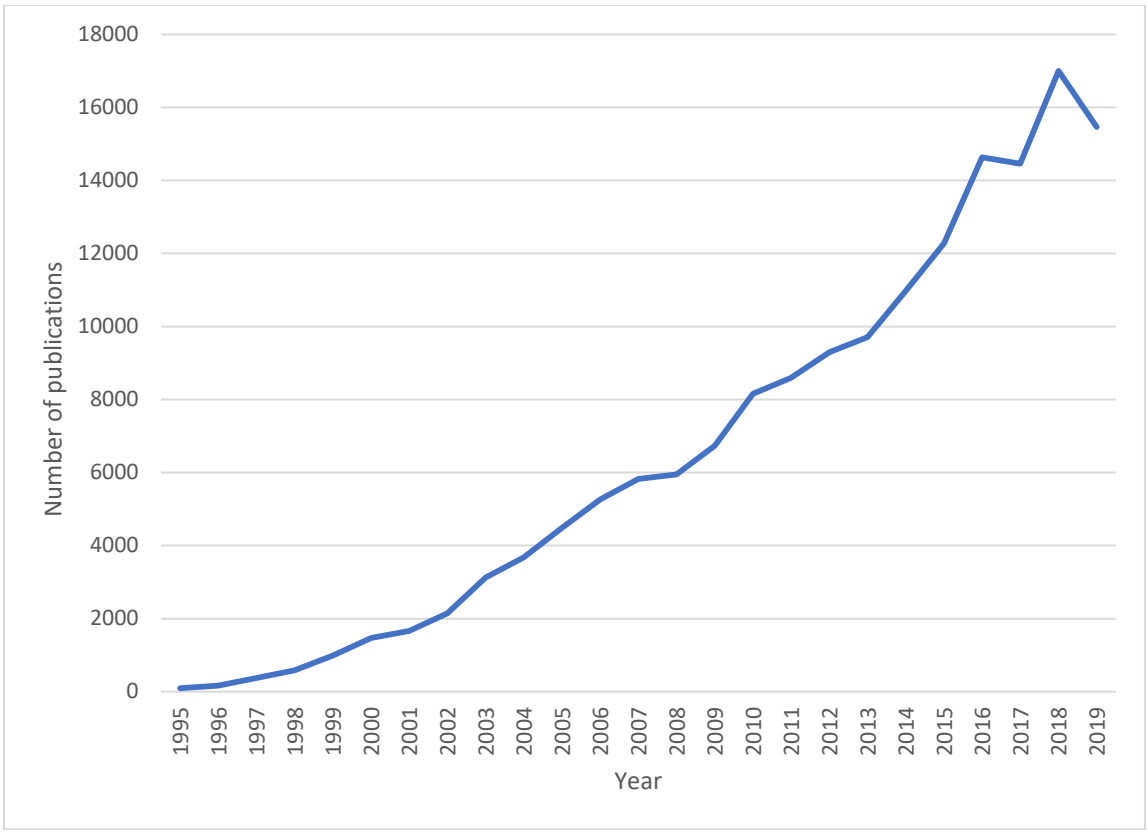

Figure 1. Number of data mining papers published annually between 1995 and 2019

As a next step, we conducted research filtered for the different subject areas defined in Scopus. Figure 2 shows the frequency distribution for dataminingpapers, but in the different fields. As we expected,computer science proved to be the top field in datamining research (113,045), followed by engineering(46,501)and mathematics(35,978).The social sciences, including education,are among the top tendomains; however, the cumulative numberof papers published beyond these three domains does not reach that published in computer science alone. Clearly, it is still rare for these techniques to be applied, as they are basically under development.

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

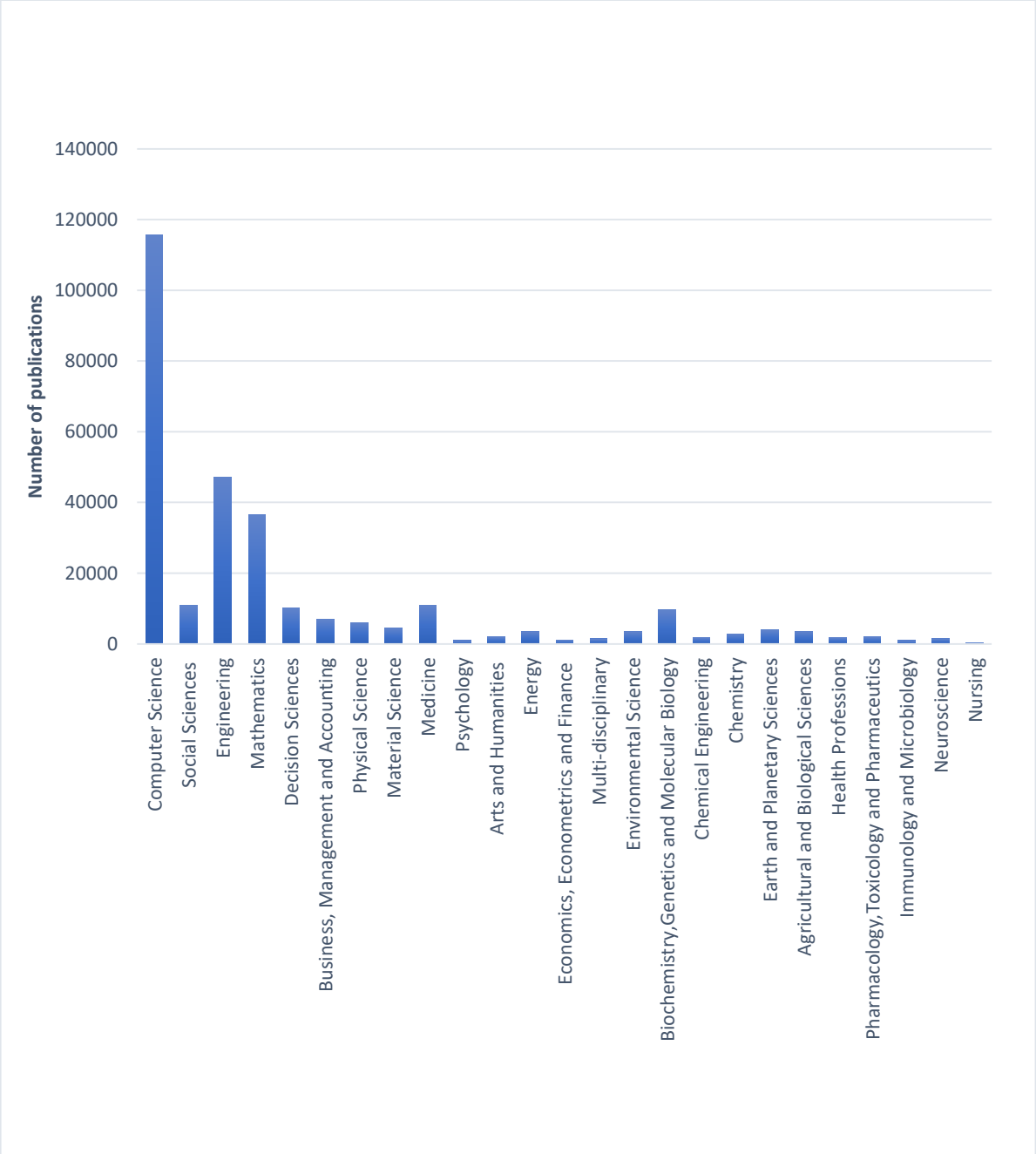

Figure 2. The frequency distribution of data mining studies in different domains

Figure3 shows the result of the third analysis. It presents how keywords, and thus research interest, have changed in datamining in the last 25 years. The terms*contextual data* (139 records) and*logfile analysis*(29 records) received less, but still increasing attentionin Scopuscompared to*educational data mining* (314 records). In contrast, the term*datamining* still tops the list. In 2019 alone, there were 15,466data mining publications in Scopus.

The most rapidly growingareaunder the umbrella term *datamining*was educational datamining. Significant changes happened in 2005, 2008, and 2012. Between 1995 and 2005, the number of publications wasalmost the same year by year, but they started to increasein 2005. Until 2008, the term *logfile* was used more often than *educational data mining*. This maybe caused by the effect of the large-scale assessment programs, which became computer-based and made great use of recorded logfiles. Another important change occurred in 2012, when*educational data mining* became the leading term.

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

To sum up, datamining became an important method in education, since technology is broadly used in educational assessment (Molnár &Csapó 2019) and learning (especially in the time of COVID 19), which areamong the main sources of the contextual data.

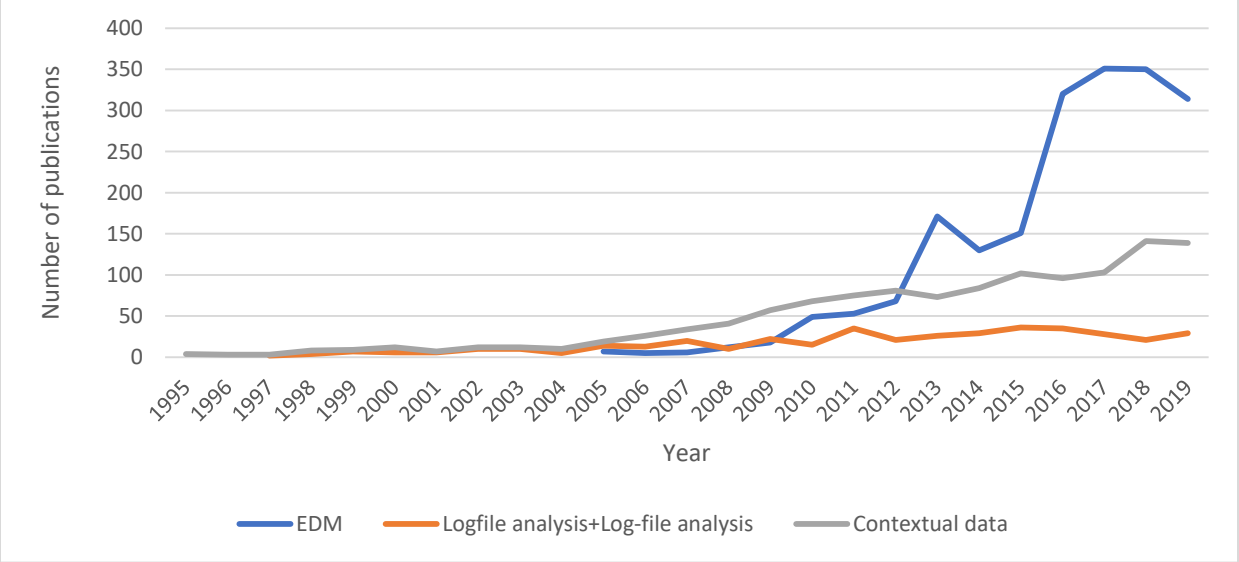

Figure 3. Theever growing research interest in EDM,logfile analysis, and contextual data

Figure 4 shows the result of the fourth analysis of the application level of educational datamining techniques, contextual data, or logfile data analysis in the different domains. The term *contextual data*was first usedin 1981, *logfile analysis*dates back to 1995, and the latest term has proved to be *educational data mining*(2000). Interestingly, this most recent term (1700 records) more often occursin papers published in computer science (641times) than in the social sciences (279). This confirms our previous statement that educational data mining techniques are still under development and that their real application is still ahead of us. This trendwas also observedforthe other two keywords (*logfile analysis*: computer science: 249; social sciences: 93; *contextual data*: computer science: 641; social sciences: 279).

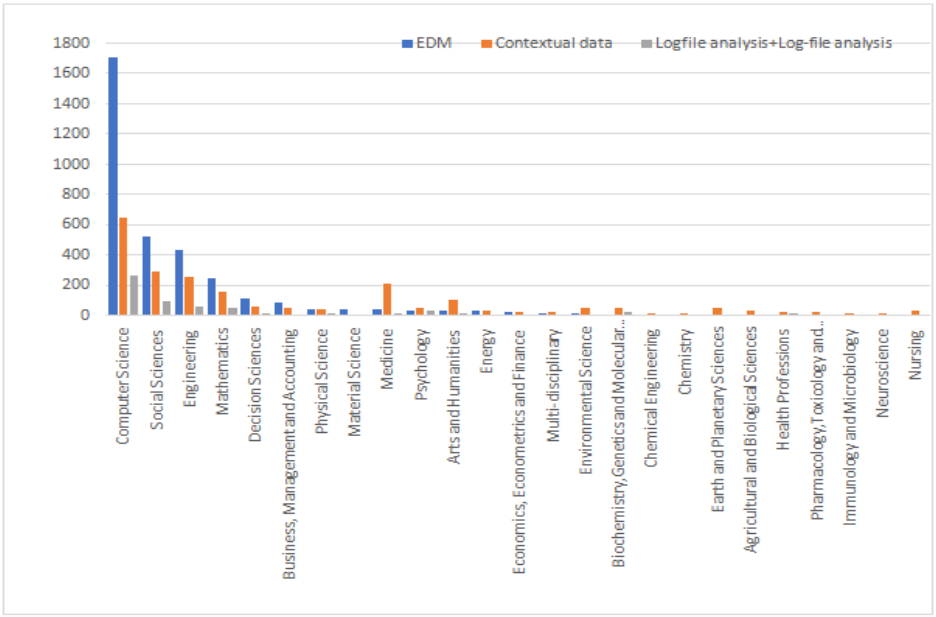

Figure 4. The frequency distribution of studies involving EDM, contextual data, or logfile analysis in different domains

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

4. Discussion and conclusion

The present research paper has aimed to summarize and evaluate the different definitions and approaches in educational data mining and logfile analysis. It has revealed thateducational data mining is still strongly related to computer science in its use of data mining techniques, which are considereda useful artificial intelligence tool. Logfile analysis isalso strongly tiedto that field;however,since the emergence and spread of computers in psychological and educational research,working with structured databases has madeit possible to analyze the behavioral processes captured during datacollection and stored in logfiles.EDM researchhas extracted and analyzed hidden data to evaluate students'behaviours and actions with a focus on the following issues:

- Classifying students and detecting sources of any incongruous values from student activities(Romero, Ventura,& García, 2008).
- Assisting teachers in evaluating their students so as to improve theirperformance (Zaiane & Luo,2001).
- Detectingchanges in students' performance during the teaching and learning process (Schoor & Bannert, 2012).
- Helping teachers to design and develop good educational strategy plans and digital school materials (Chen & Chen, 2009).
- Predicting the likelihood that students willfail ordropout(Neto, Plácido, Silva,& Guedes, 2019).
- Analyzing event logs based on students' and teachers' behaviors in digital learning environments (Lama & Mucientes, 2016).
- Detecting students' quiz-taking behavior (Papamitsiou & Economides, 2016).

Molnár and Csapó (2019) stated that essentialeducational assessmentswill be administered in the near future via a technological environment, providing the option to use the numerous advantages of technology-based assessment, including the methods and tools developed from educational data mining and logfile analysis.This study aims to offer researchers and educators information about developmental trends in analyzing contextual data in educational context through an evaluation of the number and distribution of publicationsin one of the most prominent publication databases, Scopus,in the field of datamining, more specifically, educational datamining, contextual data analysis,and logfile analysis.

The interest in contextual data, especially in educational context,could already be observed in the last century; however, use of the term *data mining*dates back to 1966. In the 1980s,*contextual data*was the relevant term, while *logfile analysis* as a keyword started to beusedin the 1990s. The latest term is*educational data mining*, which has been common since 2000.Despite its educational context, itis still often found in computer sciencepublications, indicating that the methods and algorithms are still under development and not yet at a stage of widespread application.

This study confirms that providing additional indicators during the educational process, especially in assessment, draws researchers' interest to the new domain,thus supporting the educational process with an abundance of indicators. Indeed, using technology in learning has come into its own as a research field.

The acknowledgment :

Preparation of this article was funded by OTKA K135727.

References

Algarni, A. (2016). Data mining in education. *International Journal of Advanced Computer Science and Applications, 7*(6), 456–461.

Al-Kabi, M., Shannaq, M., & Alsmadi, I. (2011). A comparative study of Web usage and searches at Yarmouk University and Jordan. *Abhath Al-Yarmouk: Basic Science & Engineering, 20,* 1–21.

Alzoubi, O., Fossati, D., Di Eugenio, B., Green, N., & Chen, L. (2013). Predicting students' performance and problem solving behavior from iList log data. In *ICCE 2013, 21st International Conference on Computers in Education*.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application, 9*(8), 119–136.

Arora, Y., Singhal, A., & Bansal, A. (2014). Prediction & Warning: a method to improve student's performance. *ACM SIGSOFT Software Engineering Notes, 39*(1), 1–5.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analyzing patterns and strategies in students' self-regulated learning. *Metacognition and learning, 9*(2), 161–185.

Barros, T. M., Neto, S., Plácido, A., Silva, I., & Guedes, L. A. (2019). Predictive Models for Imbalanced Data: A School Dropout Perspective. *Education Sciences, 9*(4), 275.

Chen, C. M., & Chen, M. C. (2009). Mobile formative assessment tool based on data mining techniques for supporting web-based learning. *Computers & Education, 52*(1), 256–273.

Chen, C. M., Chen, Y. Y., & Liu, C. Y. (2007). Learning performance assessment approach using web-based learning portfolios for e-learning systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 37*(6), 1349–1359.

Cheng, J. (2017). Data-mining research in education. *arXiv preprint arXiv:1703.10117*. Retrived from https://arxiv.org/pdf/1703.10117.pdf.

Cho, M.-H., & Shen, D. (2013). Self-regulation in online learning. *Distance Education, 34*(3), 290–301.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247–256.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st Century skills* (pp. 143–230). New York: Springer.

Dahiya, V. (2018). A survey on educational data mining.*International Journal of Research in Humanities, Arts and Literature, 6*(5), 23–30.

Dunham, M. H. (2006). *Data mining: Introductory and advanced topics*. Pearson Education India.

Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2020). Using process data to explain group differences in complex problem solving. *Journal of Educational Psychology*. Advance online publication

Fatima, D., Fatima, S., & Prasad, A. K. (2015). A survey on research work in educational data mining. *IOSR Journal of Computer Engineering (IOSR-JCE), 17*(2), 43–49.

Funke, J. (2010). Complex problem solving: A case for complex cognition?. *Cognitive processing, 11*(2), 133–142.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626.

Green, B. (2015). Data mining log file streams for the detection of anomalies (Doctoral dissertation, Auckland University of Technology).

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36–46.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers and Education, 91*, 92–105.

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement, 55*(2), 194–216.

Juhaňák, L., Zounek, J., & Rohlíková, L. (2019). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. *Computers in Human Behavior, 92*, 496–506.

Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2011). Analytics: The widening divide. *MIT Sloan Management Review, 53*(3), 1–22.

Krause, U. M., Stark, R., & Mandl, H. (2009). The effects of cooperative learning and feedback on e-learning in statistics. *Learning and instruction, 19*(2), 158–170.

Kularbphettong, K. (2017). Analysis of students' behavior based on educational data mining. In *Proceedings of the Computational Methods in Systems and Software* (pp. 167–172). Springer, Cham.

Landers, R. N., & Landers, A. K. (2015). An empirical test of the theory of gamified learning: The effect of leaderboards on time-on-task and academic performance. *Simulation & Gaming, 45*(6), 769–785.

Lee, Y. (2018). Effect of uninterrupted time-on-task on students' success in Massive Open Online Courses (MOOCs). *Computers in Human Behavior, 86*, 174–180.

Louw, J., Muller, J., & Tredoux, C. (2008). Time-on-task, technology and mathematics achievement. *Evaluation and Program Planning, 31*(1), 41–50.

Lustria, M. L. A. (2007). Can interactivity make a difference? Effects of interactivityon the comprehension of and attitudes toward online health content. *Journal of the American Society for Information Science and Technology, 58*(6), 766–776.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers &Education, 54*(2), 588–599.

Manhães, L. M. B., da Cruz, S. M. S., & Zimbrão, G. (2014). WAVE: an architecture for predicting dropout in undergraduate courses using EDM. In *Proceedings of the 29th annual acm symposium on applied computing* (pp. 243–247).

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36–51.

Minaei-Bidgoli, B., Kortemeyer, G., & Punch, W. (2004). Association analysis for an online education system. In *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, 2004. IRI 2004* (pp. 504–509). IEEE.

Molnár, G., & Csapó, B. (2018). The efficacy and development of students' problem-solving strategies during compulsory schooling: Log-file analyses. *Frontiers in Psychology, 9*, 302.

Molnár, G., & Csapó, B. (2019). How to Make Learning Visible through Technology: The eDia-Online Diagnostic Assessment System. In *Proceedings of the 11th International Conference on Computer Supported Education* (CSEDU 2019) - Volume 2, pages 122–131. ISBN: 978-989-758-367-4.

Molnár, G., & Csapó, B. (2019). Technology-Based Diagnostic Assessments for Identifying Early Mathematical Learning Difficulties. In *International Handbook of Mathematical Learning Difficulties* (pp. 683–707). Springer, Cham.

Mousa, M., & Molnár, G. (2019). Applying Computer-based Testing in Palestine: Assessing Fourth and Fifth Graders Inductive Reasoning. *Journal of Studies in Education, 9*(3), 1–13.

Mylonas, P., Tzouveli, P., & Kollias, S. (2004). Towards a personalized e-learning scheme for teachers. In IEEE *International Conference on Advanced Learning Technologies* Proceedings (pp. 560–564). IEEE.

Naumann, J. (2019). The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment. *Frontiers in Psychology, 10*, 1429.

Papamitsiou, Z., & Economides, A. A. (2016, July). Process mining of interactions during computer-based testing for detecting and modelling guessing behavior. In *International Conference on Learning and Collaboration Technologies* (pp. 437–449). Springer, Cham.

Reimann, P., & Yacef, K. (2013). Using process mining for understanding learning. In R. Luckin, S. Puntambekar, P. Goodyear, B. Grabowski, J. Underwood, & N. Winters (Eds.), *Handbook of design in educational technology* (pp. 472–481). New York: Routledge.

Reimann, P., Markauskaite, L., & Bannert, M. (2014). e-Research and learning theory: What do sequence and process mining methods contribute. *British Journal of Educational Technology, 45*(3), 528–540.

Rodrigues, M. W., Isotani, S., & Zarate, L. E. (2018). Educational Data Mining: A review of evaluation process in the e-learning. *Telematics and Informatics*, *35*(6), 1701–1717.

Romero, C., Cerezo, R., Bogarín, A., & Sanchez-Santill_an, M. (2016). Educational process mining: A tutorial and case study using moodle data sets. In S. ElAtia, D. Ipperciel, &O. R. Zaïane (Eds.), *Data mining and learning analytics: Applications in educational research* (pp. 3–28). Hoboken, NJ: John Wiley & Sons.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368–384.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence, 48*, 37–50.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence, 48*, 37–50.

Schoor, C., & Bannert, M. (2012). Exploring regulatory processes during a computer supported collaborative learning task using process mining. *Computers in Human Behavior, 28*(4), 1321–1331.

Sedrakyan, G., DeWeerdt, J., & Snoeck, M. (2016). Process-mining enabled feedback: "Tell me what I did wrong" vs. "tell me how to do it right". *Computers in Human Behavior, 57*, 352–376.

Silva, C., & Fonseca, J. (2017). Educational data mining: a literature review. In *Europe and MENA Cooperation advances in information and communication technologies* (pp. 87–94). Springer, Cham.

Tanimoto, S. L. (2007). Improving the prospects for educational data mining. *In Proceedings of the complete online proceedings of the workshop on data mining for user modelling, 11th International conference on user modeling* (UM 2007) (pp. 106–110).

ISSN- 2394-5125 VOL 8, ISSUE 01, 2021

Tóth, K., Rölke, H., Goldhammer, F., & Barkow, I. (2017). Educational process mining: New possibilities for understanding students' problem-solving skills.In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire* 21st century learning (pp. 193–210). Paris: OECD.

Vainikainen, M. P. (2019). Assessing learning to learn in the 2020S – Theory, methods, and practice. Paper presented at 17th Conference on Educational Assessment, Szeged, Hungary.

Vidal, J. C., Vazquez-Barreiros, B., Lama, M., & Mucientes, M. (2016). Recompiling learning processes from event logs. *Knowledge-Based Systems, 100*, 160–174.

Wu, H., & Molnár, G. (2019). Cross-national Differences in Students'Exploration Strategies in a Computer-Simulated Interactive Problem-SolvingEnvironment: Log-file Analyses. Paper presented at Symposium at 17thConference on Educational Assessment. Szeged, Hungary.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior, 47*, 168–181.

Zaiane, O. R., & Luo, J. (2001). Towards evaluating learners' behaviour in a web-based distance learning environment. In *Proceedings IEEE International Conference on Advanced Learning Technologies* (pp. 357–360). IEEE.

Zhang, M., Bennett, R. E., Deane, P., & van Rijn, P. W. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice*,*38*(2), 1–13.

Zoanetti, N., & Griffin, P. (2017). Log-file data as indicators for problem-solving processes. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 177–191). Paris: OECD.