Problems and Prospects in the Applying Methods of Analysis Educational Data

Tarana Aliyeva[1*], Ulviyya Rzayeva[2], Jeyran Khalilova[3]
[1,3] Azerbaijan State University of Economics, Faculty of Technology and Design, Department of Digital Economy and ICT
[2] Azerbaijan State University of Economics, the Center of Digital Economy
*tarana.aliyeva@unec.edu.az

Article History: Received: 10 November 2020; Revised 12 January 2021; Accepted: 27 January 2021; Published online: 5 April 2021

ABSTRACT
At the present stage, education is the most important component of the development of a country's economic growth. Often a changing situation requires high professionalism and considerable intellectual effort to make an effective decision. The increase in information flows and the analysis of relevant information generated by the participants in the educational process plays an important role in the process of quality management of the education process. The creation of training systems and the spread of network technology has led to the accumulation of a large amount of data, and this has in turn aroused great interest in the study of Data Mining methods used to analyze the new type of educational data. In this paper, a comparison of methods for analyzing educational data (EDM) and learning analytics (LA) was made, and attention to their peculiarities was paid.
As a practical implementation, the behavior of a time series is proposed, which describes the number of studies in 4 scientific fields Azerbaijan for the period from 2014 to 2019, and a short-term forecast is presented.

Keywords
Intelligent Analysis of Educational Data; Learning Analytics; E-Learning; Time series; Predictive model

Introduction

In connection with the active use of digital technologies and the development of e-learning systems during the traditional educational process, relatively large data arrays are accumulated, and therefore, in recent years, there has been an exponential growth of data in the educational sector. This has led to the emergence of a new direction - the analysis of educational data (Educational Data Mining - EDM) (Baker & Siemens, 2012) in the field of artificial intelligence in the early 2000s. EDM is an interdisciplinary area that originated at the junction of other disciplines.

In fact, EDM can be represented as a combination of three main areas: computer science, education, and statistics. The intersection of these three areas also forms other disciplines such as computer-aided learning, DM and machine learning, and Learning analytics (LA), which are closely related to EDM (Fig.1). Of all above-mentioned areas, LA is the area most relevant to EDM and can be defined as measuring, collecting, analyzing, and presenting data about students, education objectives and optimization, and the conditions in which learning takes place. It is focused on decision making based on data generated in the learning process.

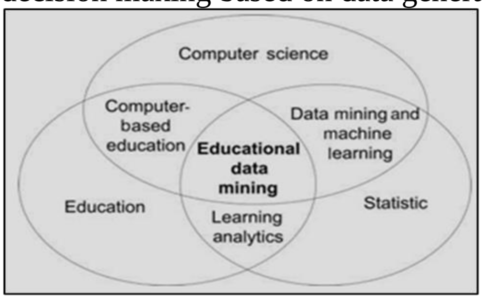

Figure1. Main areas related to Educational Data Mining (EDM)

EDM develops, improves methods for processing data generated in the learning process and extracts patterns from them. For a certain session of the electronic educational environment (EEE), some amount of data containing specific details for analysis is generated. Before EOS, the flow of information required sophisticated methods to collect, analyze and interpret student tracks to regulate and improve education. LA emerged as a knowledge discovery paradigm that provides valuable insights and helps interested parties understand the learning process and its implications.

In order to extract patterns in the educational process, EDM methods use various types of data, helping to improve the design of the educational environment and the educational process.

Despite the development of high information technologies and the e-learning method, each site visitor faces some problems due to the lack of direct human contact, i.e. there is a certain barrier in this area of application. This is due to the lack of preparedness of educational workers in the interpretation of the volume of big data and the learning process, as well as specialists in the relevant field. After examining some of the similarities and differences between the EDM and LA methods, we can conclude, “EDM focuses more on methods and methodologies, and LA focuses on applications” (Ferguson, 2012).

The purpose of this article is to set out the features of the EDM and LA methods for preventing a barrier in this area, to consider the features of these two relatively new and increasingly popular areas of research related to the collection, analysis and interpretation of educational data, and to explore problems and trends caused by the increasing enormous growth of information (Belonozhko & Silin, 2014).

The essence of the research is presented in the following sections of the work:

-The emergence and general objectives of EDM and LA methods;

-Features and advantages of EDM in the educational process;

-The main similarities and differences between EDM and LA methods;

-Directions of research: problems and solutions, trends;

-Conclusions.

Literature Review

Since the 80s of the 20th century, the creation of training systems and the spread of network technology has led to the accumulation of a large amount of data, and this has had in turn aroused great interest in the study of Data Mining methods used to analyze the new type of educational data. Thanks to technology similar to learning management systems such as Moodle, Sakai and ILIAS, it has become possible to obtain information about student behavior outside the traditional educational environment. At the same time, at international conferences on the use of artificial intelligence methods in education (Romero & Ventura, 2010), regular seminars dedicated to the development of methods in the educational sector were held (Belonozhko & Silin, 2014).

The evolution of learning analytics has gone through three eras (Peña-Ayala, 2017), these periods are closely related to the creation and development of the Society for Learning Analytics Research (SoLAR). The first epoch corresponded to the earliest jobs published until 2011. A job printed in 1996 examines the scholarship report declared by Boyer (Boyer, 1996), where he analyzed American higher education.

The second era began in 2011 with the goal of encouraging and supporting research, cooperation, and the spread of LA labor throughout the world and continued until 2013. Separate articles were published on journals and materials of the LAK conference (Learning Analytics & Knowledge Conference), cited during this period.

As for the third and current epoch, this one began in 2014 and was still ongoing. LA is currently receiving more attention at conferences and journals indexed by TR-JCR (Thomson Reuters - Journal Citation Reports).

The popularity of the existing two methodologies - EDM and LA is due to the following factors. First, statistical methods, methods of machine learning and data collection, and the development of predictive models or decision rules are powerful mathematical tools in the field of educational data analysis, and new technologies expand their capabilities. Secondly, there is increasing interest in using a data-driven approach to make better decisions (Daradoumis et al, 2010) in the educational field to improve the quality of the learning process.

The main essence of EDM and LA is to extract information from data related to education. Information may be targeted to several stakeholders (Daradoumis et al, 2013) - instructors, students, managers and researchers. Each of them is an executor of certain functions: the instructor develops and organizes the learning process and evaluates its effectiveness; students have the opportunity to get recommendations on resources, taking into account their performance, goals and motivation, analyze the results of the educational process, comparing them with other courses; managers, using information, more efficiently allocate human and material resources in order to improve the overall quality of their academic offerings; researchers conduct studies based on educational data.

Bienkowski's (Bienkowski et al, 2012) report about main problems of implementing and applying the methods of EDM and LA, and Peña-Ayala's work (Peña-Ayala, 2017), describing in detail the applications and methods of EDM, has a wide citation.

The next step in the development of this direction is connected with the holding of annual conferences devoted to EDM, the emergence of mass publicly available online courses (MEPs) with extensive data collection capabilities such as Khan Academy, Coursera, edX, Udacity, etc.

Methods

The methodological basis of the study was a system-evolutionary theoretical approach based on the complementarity of system principles with the principles of evolutionary development, including the concept of classical analysis.

At present, a tendency has been outlined for the development of a comprehensive science of education - the so-called educational science or educology, the main theoretical constructs of which can serve as a methodological basis for the development of the education economy.

The article considers the methodology of education informatization as a purposeful organization of the process of providing the education sector with methods, technology and practice for creating and making optimal use of scientific, pedagogical, educational and methodological and software and technological developments aimed at realizing the didactic opportunities of information and communication technologies used in comfortable and health-saving conditions.

Features and Advantages of EDM in The Educational Process

EDM features include goals, data, research methods and applications.

EDM objectives (Baker & Yacef, 2009) consist of:

1. Prediction of students' behavior in the learning process;
2. Development of new models and ways of presenting knowledge in the subject area;
3. Study of the interaction effects in the system "learning – student";

4. Development of knowledge about the phenomenon of learning and the psychology of students.

Complicated data, which educators usually do not deal with, present difficulties for analysis by traditional methods:

- the number of visits to the EEE website;
- the number of the most frequent visited pages;
- the number of views or downloads of the necessary materials for study;
- the information on browsers and frequency of visits to certain pages in time;
- the information on origin of visitors;
- the information on the number of visits and their duration for each student for a certain period of time;
- the information on the most popular keywords to search information in the system;
- the information about electronic resources downloaded, read or viewed by a student and about the amount of material for study.

Such data are provided in particular by the Moodle system (Kay et al, 2006; Nesbit et al, 2008).

The investigation of data generated in the learning process for possible analyzing the students' learning processes depending on their interaction with the environment (Baker et al, 2012), EDM develops and adapts various methods. Prediction, clustering, classification, search for sequential patterns, text mining and methods, search for binding rules specific to EDM - discovery with the help of models and data distillation for human judgment (Baker & Siemens, 2013) refer to traditional DM methods. Each of them is applied to solving problems of a specific nature (Romero & Ventura, 2010).

There are a number of advantages of EDM methods used by participants of the educational process - students and teachers. Students have the opportunity to adapt the course in terms of the level and assimilation of knowledge, since EEE, taking into account the duration and frequency of the visit, collects detailed information about each individual's action, processes and forms a learning model. Based on the analysis of the collected data, the EEE generates an adapted hint; the student is compensated for the loss of time for learning and is offered a new course for study. At the received hint, teachers study the situation and make adjustments in the content of the course, follow the learning process and classify students according to specific characteristics (by academic performance, activity, preliminary preparation, etc.) assess their knowledge. Because of these situations, the administration of EEE is able to evaluate the effectiveness of the course and improve its condition.

The Main Similarities and Differences Between EDM and LA Methods

EDM usually looks for new patterns in data and develops new models; LA applies known prognostic models in education systems. EDM and LA are aimed at the same goal: to improve the quality of education by analyzing a huge amount of data to extract useful information for interested parties. These methods are closely related to each other; from this point of view, they have many common characteristics, at the same time, significant differences. These differences lie in the following features (Baker & Siemens, 2013):

- EDM allows studying the components of the system and the relationships between the components; LA enables to explore the whole system.
- EDM is based on educational software, while LA is connected to a specific semantic network.
- EDM performs automated adaptation; LA informs and enhances faculty and students.
- EDM uses classification, clustering, Bayesian modeling, prediction, detection with models, and visualization methods;

• LA aims at analyzing social networks, tonality and influence, predicting student success, analyzing the concept and models for creating meaning.

In some studies (Baker & Siemens, 2013), EDM and LA are considered as separate areas that study the automation of patterns' identification in educational data and ensure the preparation of data in a suitable form for human analysis. According to the abovementioned authors, these differences are broad trends in each community and, as a result, do not define the relevant areas. A similar idea is expressed in (Baker & Inventado, 2014), which states that "the overlap and differences between communities are largely limited, evolving from the interests and values of specific researchers, rather than reflecting a deeper philosophical split".

Bienkowski (Bienkowski et al, 2012) showed that LA covers more disciplines than EDM. In addition to computer science, statistics, psychology, and the sciences of learning, LA is related to computer science and sociology.

Before the automation of the process of obtaining, storing and processing data attempts have been made to draw conclusions on a sample of experts - specialists.

At the present stage of development of technical means, researchers are provided with enormous opportunities when working with a huge amount of data and people. Technical progress, leading us to the era of big data, leads to faster and more reliable results, and in turn, solutions that are more effective. The combination of these two methodologies is a promising direction for government bodies as well.

Directions of research: problems and solutions, trends

According to the paper we can say that these two areas are relatively new areas of research, and they still have a number of unsolved problems:

1. Lack of theoretical and practical knowledge among a significant proportion of teachers and managers regarding the use of the necessary tools. To solve this problem, researchers must disseminate their results by developing a data-driven culture in an educational environment (Romero & Ventura, 2010) by collaborating with a large number of teachers and/or students to evaluate their proposals during experiments to facilitate data analysis.
2. Additional costs for storing and managing data, because different data analysis packages may not always easily integrate with each other and assistive devices.
3. Ethics and personal privacy. The ethics provided in (Greller & Drachsler, 2012) should be taken into account at all stages of data analysis, from data collection to interpretation of results and decision-making. Consideration should be given to the ownership of student data that differs from country to country.
4. The specificity of the results in the field of EDM. Since most of the research on EDM was conducted in North America and Western Europe, the results obtained in them may differ significantly from those obtained in countries with different cultural traditions (Baker & Yacef, 2009)

Trends in future research on EDM are based on the following provisions (Belonozhko et al, 2014):

1. EDM tools must be fairly convenient, simple, and integrated into EOS and provide an interface for accessing data.
2. There should be possibility for uniformly describing the models obtained by using educational data.
3. Methods of data analysis should be adapted to the application of educational data.
4. The problem of incomplete data collected when using popular social networks - Facebook, Vkontakte, etc. should be eliminated through the integration of social networks into the educational environment and the performance of part of their functions by Massive Open Online Courses (MOOCs).

The use of EDM and LA methods in network environments is determined by the generation of large educational data with dimensions that go beyond the capabilities of common software tools for capturing, storing, managing, and processing in a reasonable amount of time (Snijders et al, 2012). The main differences between big data and analytics are volume, speed, and diversity (McAfee & Brynjolfsson, 2012).
The Environments MOOCs such as Coursera, edX or Class2Go of large universities, which have been popular since 2012, allow students from all over the world to attend a variety of courses, free of charge, to narrow the gap between educational opportunities associated with economic inequality. Typically, a large number of students are trained in such courses. This creates a problem of scalability of visitors (Kay et al, 2013), very high dropout rates and very different participation models (Clow, 2013).
The maximum potential of EDM and LA in MOOCs justifies itself in the diversity of students and the extremely high level of student instructors. Different origin of participants, language skills, goals, experience, and levels of education, needs and learning styles indicates the relevance of course personalization in the automation of these systems. It is known that the existing MOOCs' platforms provide limited data storage, adaptive MOOCs (aMOOCs) appear. To improve and personalize the management of MOOCs, it is proposed to use software agents that can redesign them according to the profile of each participant. Unlike many MOOCs, described by sets of consecutive videos and quizzes, large companies such as Google or Amazon use algorithmic approaches to select searches, announcements, and purchase recommendations. Sonwalkar (Sonwalkar, 2013) describes the development of the first aMOOCs platform, which is implemented using the cloud architecture of Amazon Web Services.
Adaptive learning is very relevant today. An avalanche of information flows and overloads, a rapidly changing modern world and the need for continuous learning require the development of new learning skills. Learning should be so dynamic as to allow the formation of personal learning pathways tuned to the level of knowledge and needs of a particular student.

Study of the main development trend and forecast of the future number of scientific researches in Azerbaijan

An important direction in Learning Analytics is the study of the general development trend. Changes in the levels of time series are caused by the influence of various factors on the process under study. In general, they are heterogeneous in strength, direction and time of exposure. Many leading think tanks and universities are now engaged in forecasting science and technology.
One of the important catalysts for the spread of a new technological wave is the convergence of various fields of science and technology, the development of which can make a significant contribution to the implementation of scientific and technological responses to global challenges. A striking manifestation of this trend is the growth in the number of defended PhD thesis.
The article examines the short-term behavior of a time series describing the number of studies in 4 fields of science in Azerbaijan for the period from 2014 to 2019 and presents a forecast for 2020. Initial data are taken from the official website of the State Statistics Committee of Azerbaijan (Table 1).

Table 1. Number of researchers in the field of science (by the end of the year)

Years	Fields of science	Numbers of defended thesis
2014	technical	2621
	medical	2049
	agriculture	917
	social	1620
2015	technical	2424
	medical	2054
	agriculture	886
	social	1977
2016	technical	1960
	medical	2152
	agriculture	715
	social	2362
2017	technical	1858
	medical	2130
	agriculture	803
	social	2326
2018	technical	1773
	medical	1941
	agriculture	855
	social	2423
2019	technical	2856
	medical	1956
	agriculture	812
	social	2001

Source: the official website of the State Statistics Committee of Azerbaijan (https://www.stat.gov.az/source/education/)

The temporal development of the process generally consists of several components - the demand component, the cyclical component, the irregular (random) component and the trend.
The demand component is the changes occurring in connection with various events, mandatory orders, the impact of which is limited to a certain period. Changes in demand are sometimes so strong that they violate the main line of development of the phenomenon. The cyclical component consists of successive ups and downs that are not repeated every year. The short-term irregular (random) component represents a residual variation that cannot be explained. This is the result of random fluctuations. It manifests the action of those one-time events that occur over time by chance, and not systematically. The trend indicates the actual long-term behavior of a time series, usually as a straight line or exponential curve.
So, in the general case the temporal development of the process consists of four components. We are mainly interested in the trend, since it allows us to judge the dynamics of the studied process and gives an opportunity to look into the future. However, other factors "crowd" around it, which confuse the general picture and the role of the trend may be more blurred. That is why it is important to be able to highlight:
a) evaluate the influence of each of the factors discussed;

b) mark their "weight" contribution;
c) evaluate the role of the trend.
These four basic components of a time series (trend, demand, cyclical and irregular components) can be estimated in different ways. The most convenient method is called the sliding average relationship. Sliding average is used to eliminate irregular effects by averaging over the whole year, to reduce irregularity and obtain a combination of trend and cyclical component. Dividing the original series by the smoothed sliding average series gives us a ratio to the sliding average that includes both regular and irregular values. Performing grouping by the corresponding area, and then averaging in the obtained groups, we find the demand index for each scientific area. After that, dividing each value of the series by the corresponding index of demand for the corresponding scientific field, we find the values adjusted for demand.
Time-adjusted regression of the series is used to estimate the trend as a straight line. This trend (tendency) does not reflect fluctuations and makes it possible to obtain a forecast adjusted for demand.
Forecasting can be done using a trend. By deriving the predicted values (trend) for future periods from the regression equation and then multiplying them by the appropriate demand index, we obtain predictions that reflect both short-term trend and irregular behavior.
We build a graph (Fig. 1) based on Table 1 and analyze the data.

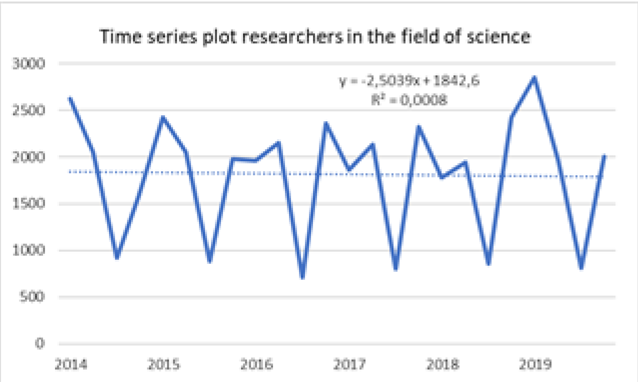

Figure 1. Initial data and trend line with regression equation

Take a close look at our experimental dependence. Based on the visual analysis, the following features can be noted:
- there are obvious fluctuations in the demand for scientific research in a particular area (there is a decline in agriculture);
- there is a certain short-term trend, namely: a general decline in defended theses (the straight line slowly creeps down);
- there is some irregularity in behavior;
- the approximating coefficient (coefficient of determination) R2 is 0.0008. This indicates the existence of a weak correlation between the studied populations.

Our goal is to isolate four basic components of a time series. Decomposition of the initial time series into these components allows us to get a clear picture of the influence of each component. Let's start by averaging the data over the year
a) to get rid of the demand component
b) to reduce the random (irregular) component.
The moving average is a new series, obtained by averaging the time series observations and moving to the next time period. As a result, we have got a smoother series (Table 2).

Table 2. Smoothing the time series using the sliding average method

Numbers of defended thesis	Before the sliding average	After the sliding average	Centered sliding average
2621			
2049			
917	1801,75	1752,5	1777,125
1620	1752,5	1753,75	1753,125
2424	1753,75	1746	1749,875
2054	1746	1835,25	1790,625
886	1835,25	1719,25	1777,25
1977	1719,25	1743,75	1731,5
1960	1743,75	1701	1722,375
2152	1701	1797,25	1749,125
715	1797,25	1771,75	1784,5
2362	1771,75	1766,25	1769
1858	1766,25	1788,25	1777,25
2130	1788,25	1779,25	1783,75
803			
2326			
1773			
1941			
855			
2423			
2856			
1956			
812			
2001			

Source: Authors' calculations

To highlight the demand, first of all, you need to get the ratio of the initial values to the sliding average. The result will include a demand component and a random component, since the sliding average excludes the trend and cyclical component from the data (Table 3).

Table 3. Behavior of the dynamic series, taking into account the demand for scientific research in a particular area

Centered sliding average	Ratio	Average ratio	Normalized
		1,261602	1,2646070
		1,188707	1,1915382
1777,125	0,516001	0,507262	0,5084704
1753,125	0,924064	1,032924	1,0353841
1749,875	1,385241	3,990497	4

1790,625	1,147085
1777,25	0,498523
1731,5	1,141784
1722,375	1,137963
1749,125	1,230329
1784,5	0,400672
1769	1,335217
1777,25	1,045435
1783,75	1,194113

Source: Authors' calculations

We complete our calculations by calculating the demand index, trend, corresponding periods and forecast. When a time series shows a short-term linear trend towards an increase or decrease, regression analysis can be used to assess this trend and predict the future (Table 4).

Table 3. Final calculations with a forecast for 2020

Years	Fields of science	No of thesis	Demand index	Trend	Periods	Forecast
2014	tech.	2621	1,26	2072,5	1	
	med.	2049	1,19	1719,6	1	
	agric.	917	0,51	1803,4	1	
	social	1620	1,03	1564,6	1	
2015	tech.	2424	1,26	1916,8	2	
	med.	2054	1,19	1723,8	2	
	agric.	886	0,50	1742,4	2	
	social	1977	1,03	1909,4	2	
2016	tech.	1960	1,26	1549,8	3	
	med.	2152	1,19	1806,0	3	
	agric.	715	0,50	1406,1	3	
	social	2362	1,03	2281,2	3	
2017	tech.	1858	1,26	1469,2	4	
	med.	2130	1,19	1787,6	4	
	agric.	803	0,50	1579,2	4	
	social	2326	1,03	2246,5	4	
2018	tech.	1773	1,26	1402,0	5	
	med.	1941	1,19	1628,9	5	
	agric.	855	0,50	1681,5	5	
	social	2423	1,034	2340,1	5	
2019	tech.	2856	1,26	2258,4	6	
	med.	1956	1,19	1641,5	6	
	agric.	812	0,50	1596,9	6	
	social	2001	1,03	1932,6	6	
2020	tech.		1,26	1832,1	7	2317
	med.		1,19	1835,2	7	2188
	agric.		0,50	1838,2	7	935
	social		1,03	1841,2	7	1906

Source: Authors' calculations

The result, adjusted for demand, turned out to be less than the actual number of defended theses in two areas of science. The fact is that the number of defended dissertations in the field of agriculture, as a rule, is lower in comparison with other fields of science. Dividing by the demand index cancels out the influence of this expected fluctuation. As a result, the number of theses defended is brought in line with the typical annual indicators of the year.

Discussions and conclusion

Being relatively new and promising areas of research and improving educational experience, these two methods - EDM and LA are aimed at enhancement the educational process and help participants in this process - students, teachers and researchers to make more effective decisions using data. By increasing the capabilities of modern technical means of information processing and the availability of DM, statistical and machine learning methods, the growth of educational data has been increased.

One of the applications is the Internet environment, in which data is constantly generated with various formats and levels of hierarchy. Unlike traditional courses, dropout rates for online courses are higher. EDM and LA are mainly used to monitor students and groups and adapt learning experiences. The methods for analyzing educational data have many similarities and at the same time several differences. Despite their current improvements, there are some barriers to the use of EDM and LA educational environments.

The application of the analysis of educational data provides a number of advantages to the participants - students, teachers and administrators of the educational process. Using EDM allows students to tailor the course for fitting their abilities. In the system, according to the accumulated information about the student depending on the duration and frequency of viewing it, a learning model is formed.

Students are offered shortened paths for completion the course, considering their interest in passing tests and homework assignments. At the prompts of EEE, problems revealed by students' errors in tests and homework are identified, they are recommended additional materials for studying the course.

Obtaining information on the course of the educational process, teachers have the opportunity to improve the content of the materials, based on data on the frequency and distribution of errors; the performing students determine the causes of these errors and eliminate them.

References

[1] Ferguson, R. (2012). The State Of Learning Analytics in 2012: A Review and Future Challenges. Technical Report KMI-12-01, Knowledge Media Institute, The Open University, UK. http://kmi.open.ac.uk/publications/techreport/kmi-12-01

[2] Belonozhko, M.L., & Silin, A.N. (2014). Sociological research in the mechanism of decision-making. Tyumen: TSOG

[3] Romero C.R., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, *40*(6), 601-618. doi:http://dx.doi.org/10.1109/TSMCC.2010.2053532

[4] Peña-Ayala, A. (2017). Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning. Springer International Publishing. Consulté à l'adresse https://books.google.fr/books?id=x8omDgAAQBAJ

[5] Boyer, E. (1996). The scholarship of engagement. *Bulletin of the American Academy of Arts and Sciences*, *49*(7), 18–33.

[6] Daradoumis, T., Juan, A., Lera-López, F., & Faulin, J. (2010). Using Collaboration Strategies to Support the Monitoring of Online Collaborative Learning Activity. In M. Lytras, P. O. D. Pablos, D. Avison, J. sipior, Q. Jin, W. Leal, D. Horner (Eds.), Technology Enhanced Learning. *Quality of*

Teaching and Educational Reform, 271–277. Springer Berlin Heidelberg. doi: http://dx.doi.org/10.1007/978-3-642-13166-0_39

[7] Daradoumis, T., Bassi, R., Xhafa, F., & Caballé, S. (2013). A review on massive e-learning (MOOC) design, delivery and assessment. *Proceedings of the 8th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 208–213. Compiegne, France. doi: http://dx.doi.org/10.1109/3pgcic.2013.37

[8] Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief. Retrieved from http://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

[9] Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. JEDM | *Journal of Educational Data Mining*, *1*(1), 3-17. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8

[10] Kay J., Maisonneuve N., Yacef K., Zaiane O.R. (2006). Mining Patterns of Events in Students' Teamwork Data. *Proceedings of Educational Data Mining Workshop*. Taiwan.URL:http://www.educationaldatamining.org/ITS2006EDM/Kay_Yacef.pdf

[11] Nesbit, J.C., Xu, Y., Winne, P.H., & Zhou, M. (2008). Sequential pattern analysis software for educational event data. *6th International Conference on Methods and Techniques of Behavioral Research "Measuring Behaviour"*, 1-5, Maastricht, Netherlands.

[12] Baker, R. S. J. D., Costa, E., Amorim, L., Magalhães, J., & Marinho, T. (2012). Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. *Jornada de Atualização em Informática na Educação*, *1*, 1–29.

[13] Baker, R., & Siemens, G. (2013). Educational data mining and learning analytics, in Cambridge handbook of the learning sciences (2nd edition), R. K. Sawyer, Ed., Cambridge, UK: Cambridge University Press, (in press) http://www.columbia.edu/~rsb2162/ BakerSiemensHandbook2013.pdf

[14] Baker, R. S. J. D., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In J. A. Larusson, & B. White (Eds.), Learning Analytics: from Research to Practice (pp. 61–75). New York, NY: Springer.

[15] Greller, W., & Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Educational Technology & Society*, *15*(3), 42–57.

[16] Snijders, C., Matzat, U., & Reips, U.-D. (2012). 'Big Data': Big gaps of knowledge in the field of Internet science. *International Journal of Internet Science*, *7*, 1-5.

[17] McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*, *90*(10), 60–66.

[18] Kay, J., Reimann, P., Diebold, E., & Kummerfeld, B. (2013). MOOCs: So Many Learners, So Much Potential… *IEEE Intelligent Systems*, *28*(3), 70–77. doi: http://dx.doi.org/10.1109/MIS.2013.66

[19] Clow, D. (2013). MOOCs and the funnel of participation. In D. Suthers, K. Verbert, E. Duval, & X. Ochoa (Eds.), *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*, 185–189. doi: http://dx.doi.org/10.1145/2460296.2460332

[20] Sonwalkar, N. (2013). The First Adaptive MOOC: A Case Study on Pedagogy Framework and Scalable Cloud Architecture — Part I. *MOOCs Forum*, *1*, 22–29. doi: 10.1089/mooc.2013.0007.