Research Topics on Educational Data Mining in MOOCS

Vanessa Faria de Souza
vanessa.faria@ufrgs.br
Graduate Program of Informatics in Education
University Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil.

Tony Carlos Bignardi dos Santos
tony.santos@ufrgs.br
Graduate Program of Informatics in Education
University Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil.

Abstract

Educational Data Mining Techniques have been widely used in MOOC environments to conduct different educational analyzes. In this context, a systematic mapping was conducted in five databases in order to verify which aspects of studies are inherent to the use of Educational Data Mining in MOOCs. The search comprised the period from 2015 to 2019, and 253 searches were found, out of this total, 133 studies were selected. The results revealed that studies on performance analysis, behavior analysis, forum analysis and implementation of recommendation systems are the most frequent themes.

Keywords: MOOCs, Educational Data Mining, Systematic Mapping;

1. Introduction

Massive open courses on open lines (MOOCs) drew the attention of education scholars as a new possibility of access to learning. The proposal, in general, is to serve as a knowledge platform for anyone, anytime, anywhere, making it an emerging and powerful learning strategy, with repercussions in the technological and educational areas (Zheng et al. 2016). Zheng et al. (2016) also state that MOOCs originated with the aim of promoting educational innovation through pedagogical approaches that expand learning possibilities, in order to reach a large number of students.

For Mackness et al. (2010), MOOCs are modern means of teaching and supporting learning, with high potential, for the exponential dissemination of knowledge. They are based on connectivist theory, enabling the creation and generation of knowledge through the interaction between participants, encouraging them to use digital technologies, social networks, among others, with a focus on collaborative learning. For Comier (2010), MOOCs are a combination of actions that start in connection, go through collaboration until reaching involvement in the learning process, in which a large group of people, concerned with a certain topic, get together to discuss it in a structured way.

MOOCs can be seen as an evolution of existing online courses, offering an opportunity to rethink new open educational models. For Butcher (2014), compared to traditional online courses, MOOCs have two main characteristics that distinguish them, namely: 1) Open access: courses must be open and free,

allowing anyone to participate in the course; and 2) Scalability: they must support a large volume of participants.

In this sense, through the variety of learning resources and due to the high number of subscribers that MOOCs reach, different types of data can be generated, as the platforms record information about the interactions of students through log files, which leads to to new possibilities of understanding learning from a quantitative perspective (Lee, 2018). It is concluded, therefore, that a differential of the MOOC courses is the large amount of data generated by the interactions in the Virtual Learning Environment (VLE). In this context, the analysis of students' actions is an important activity to detect barriers to learning, especially in MOOCs that live with low completion rates, and dropout is quite frequent (He et al., 2015).

In this way, the potential for new studies was visualized and some areas of research have emerged in recent years to assist in issues such as these. For example, Educational Data Mining (EDM), which is an area of interdisciplinary research that deals with the development of methods to explore data originating in the educational context (Romero and Ventura, 2016). EDM techniques allow the extraction of valuable information from the data and learning interactions stored during the performance of a MOOC, which leads to the identification of behavioral characteristics and indicators related to learning (Lu et al., 2017).

Thus, considering that the MDE techniques work in the context of analysis of learning data obtained from online interaction records, this research aimed to carry out a systematic mapping of the literature, in order to survey studies focusing on the application of EDM in MOOCs , and to identify the thematic aspects and the general purposes of its use in courses of this nature.

2. Educational Data Mining Applied to MOOCs

MOOCs generate a large amount of data, and the analysis of large volumes of data without the aid of computational resources is impractical, as knowledge is often hidden in these bases. Therefore, it is essential to provide tools that assist people in the task of verifying, interpreting and relating this data, in order to generate useful and relevant knowledge (Goldschmidt and Bezerra, 2015). In this perspective, the MDE techniques emerged to contribute to the extraction of relevant information from the large mass of data generated in current educational contexts, such as MOOCs.

The MDE is defined as the research area whose main focus is the development of methods to explore data sets collected in educational environments (Baker, Isotani and Carvalho, 2011). For Romero and Ventura (2013), the MDE area can be defined as the application of data mining techniques for a specific type of data set coming from educational environments, to answer important questions in this area.

Through MOE MDE, it is possible to understand students more clearly and adequately during the learning process, in addition to other factors that influence learning. For example, it is possible to identify in which situation a type of instructional approach (e.g. individual or collaborative learning) provides better educational benefits to the student.

It is also possible to check if the student is learning or confused, identify levels of motivation, involvement in online activities, discover elements or behavioral indicators of completion and success in a course, identify patterns of interaction, discover strategies that contribute to permanence of students (Pursel et al., 2016), elements that can help to personalize the teaching environment and methods to offer better

learning conditions (Baker, Isotani and Carvalho, 2011).

In short, the MDE is concerned with the development, research and computerized application of methods to detect patterns in large sets of educational data, which would otherwise be difficult or impossible to analyze due to its enormous volume (Romero and Ventura, 2010). The techniques most used by the MDE are those of classification (decision tree) and grouping, in addition to the discovery of associations (Romero and Ventura, 2010). In MOOCS, for example, classification techniques can be used to predict students' performance and / or to detect student behavior. The grouping technique can be used to group students based on their learning and behavior patterns (Romero and Ventura, 2013).

The various possibilities of using the techniques of DEM in MOOCs environments motivated the accomplishment of this research in order to find out which are the strands of studies related to this theme.

3. Methodology

In this research, a systematic mapping of the literature was carried out, with the objective of selecting publications that apply EDM in MOOCs. Mapping sought to find out what the main themes exist when it comes to this premise, permeating what are the main objectives of EDM applied to MOOCS. In this sense, this article was inspired by the work developed by Souza, Wives and Perry (2019) who carried out the same mapping, but with a focus on Learning Analysis in MOOCs, it should be noted that the contribution of this research was relevant to the development of this manuscript.

A mapping, as its name implies, does not seek to evaluate how the set of literature answers certain research questions, but to make an overview of a given area, presenting a panorama that allows the identification of research opportunities. As in a systematic review, moreover, one must use well-defined procedures to find, evaluate and synthesize the results of relevant research in the area under study - but it does so with more scope.

The Kitchenham and Charters (2007) guidelines were used to carry out the Mapping, which present a protocol - based on other protocols widely used in medical research based on evidence - which is the most used both in the area of computing, in general, and in work systematic survey of literature in the field of Informatics in Education. The RSL process, as presented in these guidelines, includes several activities (Figure 1), which can be grouped into three main phases: planning, conducting and reporting.

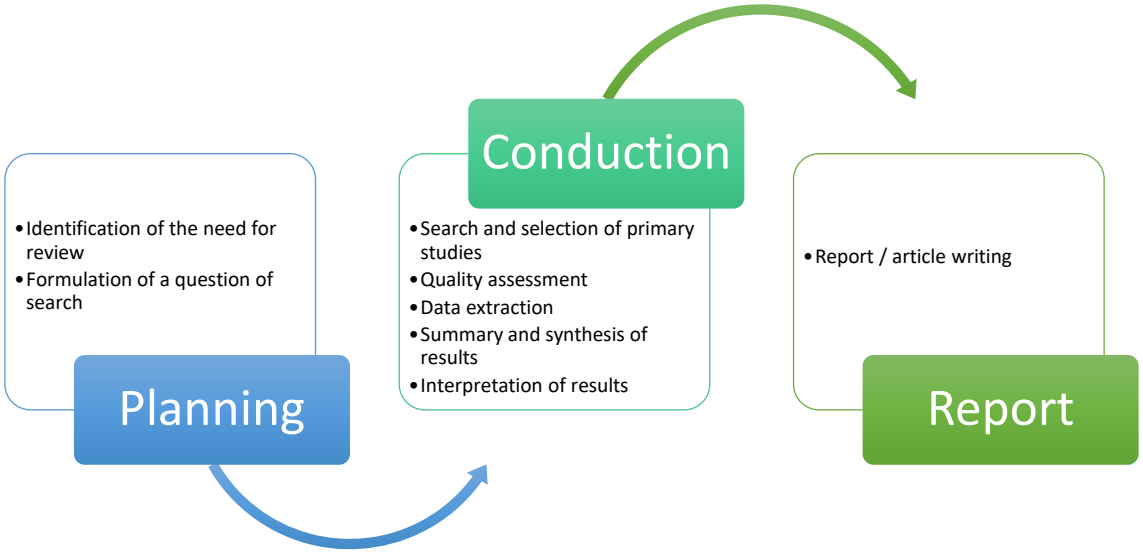

Figure 1. Phases and activities of the systematic literature review and mapping process.
Source: Based on Kitchenham and Charters (2007).

The Mapping then started planning by asking 4 research questions:

1) Which databases have the most publications in the Area?
2) Which years have there been more publications?
3) What topics are most addressed when it comes to EDM in MOOCs?

The conduct of this mapping was formed by the stages of formatting the search string, definition of the databases, establishment of the inclusion and exclusion criteria.

Then the Search String was defined and formulated as follows: ((“educational data mining” OR “EDM”) AND (“massive open online courses” OR “MOOCs”)).

Next, the search databases were defined, which were: (1) IEEEXplore Digital Library; (2) ACM Digital Library; (3) ERIC - Education Resources Information Center; (4) Science @ Direct and (5) SciElo.

The inclusion and exclusion criteria were as follows: Inclusion: 1) Full or summary articles, 2) Articles published between 2015 and 2019, 3) Articles written in English, and 4) Articles that describe the application of EDM in MOOCs. Exclusion: 1) Duplicate articles, 2) Use of EDM in contexts other than MOOCs, 3) Application of different techniques than EDM on MOOCs, 4) Literature reviews or mappings, and 5) Articles in languages other than English.

After applying the search string on the bases and returning the articles, it was then possible to select those that met the inclusion criteria or exclude those that met the exclusion criteria. For this, at first, only the abstracts of the articles were read, for a first screening, then the articles that were selected were analyzed in more detail.

For the third phase of the mapping, generating the report, Excel was used to extract and analyze the data, corresponding to the last stage of the mapping. The records were formed by 8 attributes[1]: title, authors, place of publication, year, thematic / objective summary, criteria (whether the inclusion / exclusion criteria

[1] The articles returned in the survey are available at this link: https://drive.google.com/file/d/1I2v83vJI4HXUvX2FW-gZQ2hlwml1dWwC/view?ths=true

were met or not) and situation (included or excluded). The results were analyzed taking into account the number of publications per year, place of publication and the research topic. These items are described in section 4 that presents the results.

4. Results

The organization of this section was inspired by the work of Souza, Wives and Perry (2019), in this way the systematic mapping selected articles from 2015 to 2018, from the cited databases, it is noteworthy that articles that did not use EDM in MOOCs were not selected, nor were articles that were literature reviews or mappings. Table 1 lists the number of articles returned from each database.

Table 1. Totals of papers found and selected

DATA BASE	RETURNED ITEMS	SELECTED
IEEE	98	67
ACM	22	11
ERIC	47	43
SCIENCE@DIRECT	64	12
SCIELO	22	0
TOTAIS	**253**	**133**

As shown in Table 1, of the total of 253 studies found, 133 studies met the inclusion criteria. As can be seen in the SciElo database, no article was selected, since all those dealing with the EAW in the context of MOOCs were texts in Spanish, therefore, by the exclusion criterion, none can be selected. Below, each of the research questions listed in section 3 is answered.

4.1 Which databases have the most publications in the Area?

The IEEE database presented the largest number of publications according to the inclusion criteria, with 67 selected searches, followed by ERIC with 43 selected studies. Science @ Direct, despite presenting a significant number of returned works, however, most were not within the scope of this mapping. The 133 works selected and retrieved by the databases come from conferences and journals, published in 71 different locations, 60 of which are conferences and 11 journals. Of this total, those who returned at least 3 publications on the topic were classified, as shown in Table 2. Table 2 shows that the International Conference on Educational Data Mining (EDM) is the event that has the largest number of surveys that report the use of Educational Data Mining in MOOCs.

Table 2. Total papers by publication basis

DATA BASE	# PAPERS
International Conference on Educational Data Mining (EDM)	36
IEEE Access	5
Computers in Human Behavior	4
Computers & Education	4
International Conference on Educational Innovation through Technology (EITT)	3
International Conference on Computer Science and Education (ICCSE)	3
International Conference on Learning Analytics & Knowledge Pages	3
Conference on Learning @ Scale Pages	3

4.2 Which years have there been more publications?

The second question investigated is the identification of the number of publications per year. As the mapping was carried out from the period of publications from 2015 to 2019, the following figures were identified:

- ✓ 2015 - 16 posts
- ✓ 2016 - 36 posts
- ✓ 2017 - 33 posts
- ✓ 2018 - 35 posts
- ✓ 2019 - 13 posts

This figure can be seen in Figure 1.

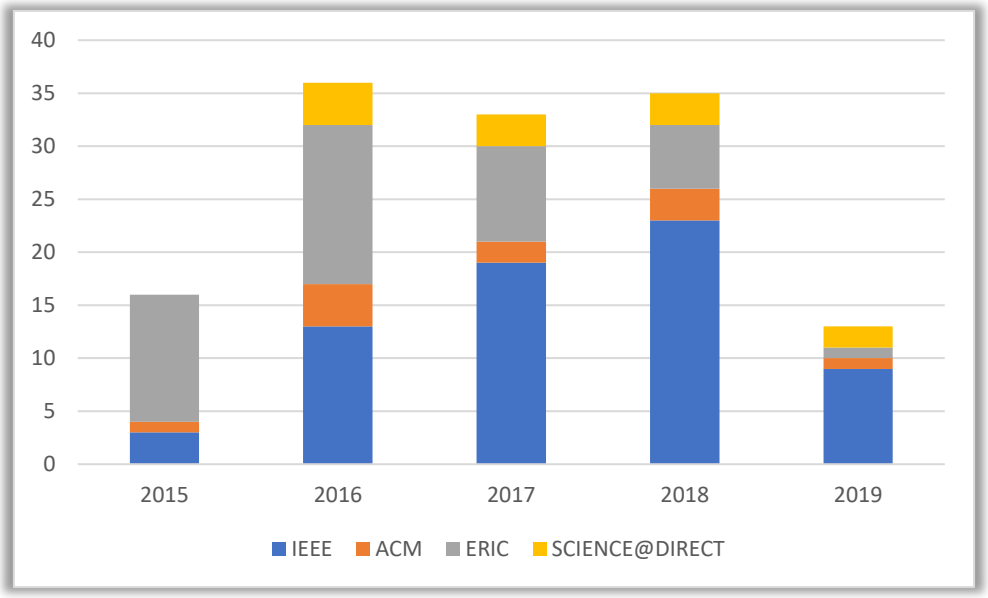

Figure 2. Number of publications per year.

In Figure 2, the number of publications per year in each database can be better viewed. Analyzing by base, it can be seen in Figure 2 that in IEEE Xplore there is an increase in the number of publications in 2018. ACM (with 4 searches), ERIC (15 searches) and Science @ Direct (4 research) presented more publications in 2016. Due to the mapping being carried out until October 2019, it is possible that many studies this year have not yet been published and made available in these databases.

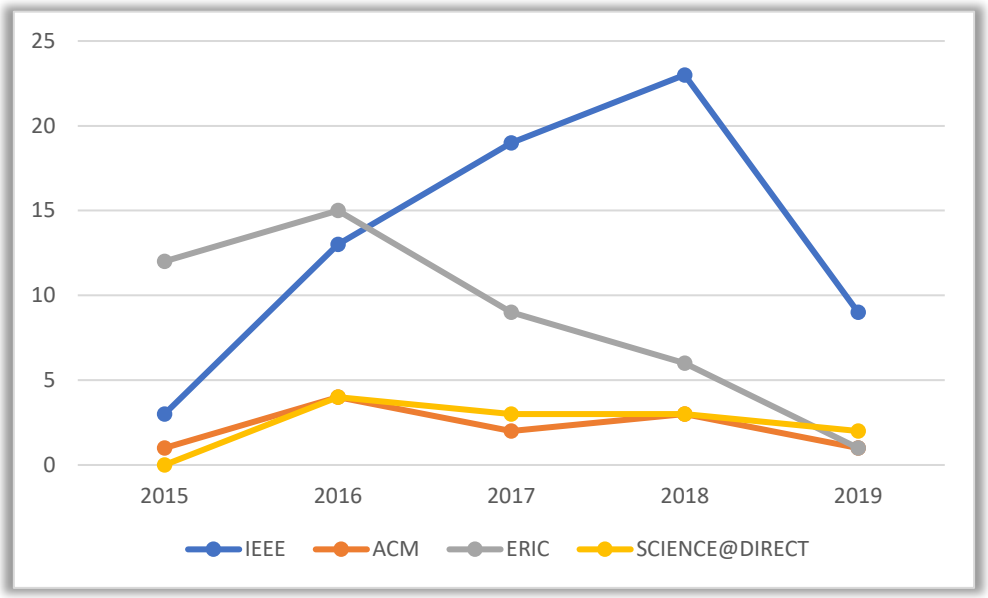

Figure 3. Number of publications per base per year.

4.3 What topics are most addressed when it comes to EDM in MOOCs?

From reading the articles, it was possible to determine the theme the article was about, and then list the main research topics on EDM in MOOCS, these can be listed in 18 main themes: 1) Behavior Analysis, 2) Behavior Analysis in Prediction / Detection of Performance (subcategory), 3) Behavior Analysis in Prediction / Detection of Abandonment (Subcategory), 4) Behavior Analysis in Conclusion of Prediction / Detection (Subcategory), 5) Analysis of Behavior in engagement prediction / detection (subcategory), 6) forum analysis, 7) recommendation systems, 8) MOOC management platforms, 9) text mining, 10) video analysis, 11) cheating identification, 12) analysis of feelings, 13) Analysis of the MOOC curricula, 14) Self-regulated learning, 15) Gamification, 16) Peer review, 17) Effort / capacity model, 18) Simulation of artificial students. The Table 3 shows the themes highlighted in quantities per year of publication.

Table 3. Total papers by publication basis

CATEGORY	Publication Year / # de Papers					TOTAL
	2015	2016	2017	2018	2019	
Performance	1	6	6	6	3	**22**
Behavior Analysis	2	5	5	7	1	**20**
Forum Analysis	6	2	2	5	2	**17**
Recommendation Systems	0	2	8	5	1	**16**
Abandonment	1	6	1	3	2	**13**
Conclusion	2	1	0	3	0	**6**
MOOC Management Platforms	0	1	2	2	1	**6**
Text Mining	1	2	2	1	0	**6**
Video Analysis	1	2	1	0	1	**5**
Cheating Identification	0	2	3	0	0	**5**
Engagement	1	2	1	0	0	**4**
Sentiment Analysis	0	3	0	0	0	**3**
Analysis of MOOC Curricula	1	0	0	1	0	**2**
Self-Regulated Learning	0	0	1	0	1	**2**
Gamification	0	0	1	1	0	**2**
Peer Review	0	1	0	1	0	**2**
Effort / Capacity Model	0	0	0	0	1	**1**
Simulation of Artificial Students	0	1	0	0	0	**1**

Table 3 shows that the student's performance forecast; investigations with an emphasis on analyzing student behavior in different contexts; analysis of discussion forums; systems of recommendation and analysis of patterns of behavior to predict and / or detect abandonment are the most frequent themes in research using EDM in MOOCs. It is worth mentioning that some thematic categories have many associated themes, for example, behavior analysis with many aspects, such as profile analysis, motivation, engagement, abandonment prediction / detection, conclusion prediction / detection and performance prediction / detection, among others.

In the forum analysis category, the most diverse approaches were observed, such as detecting student errors, thematic relevance, student engagement, posts that need the attention of teachers, among others. Along with this theme, there is the category of text mining in which the most different sources were explored, by email, texts on social networks and interactions between students on the MOOC platform, different categories were considered, because not all analyzes of the forums were by text mining, and this technique was used in different discussion forum scenarios.

Another frequent topic concerns the implementation of recommendation systems for MOOCs, which can be developed in the most diverse contexts, to recommend contacts for the exchange of information, to recommend content and also courses in this format, based on the characteristics extracted from each person. In the scope of management platforms, a survey was identified that exposed new types of AVAS for the purposes of data analysis and mining. The other identified approaches turned to video analysis, sentiment analysis, MOOC curriculum analysis, which seems to be a very promising topic, as well as peer analysis

and effort / capacity model, although with few published studies.

In addition to the most frequent thematic categories mentioned above, there were also some recent trends, such as Cheating Identification, in which students use fake accounts, called Cameo, by the authors, to get the answers in the course and put them in their real accounts. In addition, analysis of self-regulation of learning, application of EDM in MOOCS to improve Gamification, and the most innovative ones identified to perform simulations of artificial students, created through artificial intelligence of artificial students to train MDE algorithms.

4. Conclusion

An important milestone in the evolutionary process of education was the emergence of MOOCs, to meet the demands arising from a new global technological scenario. The emergence of this type of course contributed to strengthen the changes in the existing educational paradigms, in addition to meeting the process of democratization of education and the yearnings for a new student profile in the digital age, increasingly present in educational institutions. These courses are being considered the next step in distance education in the world (Wulf et al., 2014).

MOOCs are one of the greatest examples of changes in the educational standard, constituting a form of distance learning, based on the open and free offer of online courses for a large number of geographically dispersed people (Vázquez et al., 2013). Its philosophy is centered on the democratization of knowledge, making it available to people, regardless of their geographical location or financial conditions (Barak et al., 2016). In addition to all these characteristics, one of the biggest differentials of MOOCs is the amount and diversity of data generated by students on the offer platforms, a fact that made it possible to explore this mass of data, discover new knowledge about how individuals' study, learn, interact .

In this scenario, Educational Data Mining (MDE), an area of interdisciplinary research that deals with the development of methods to explore data originating in the educational field (Romero & Ventura, 2016) has gained prominence. The MDE techniques aim to extract information from the data recorded by the platforms during the course of a MOOC, and which can lead to the identification of behavioral characteristics and indicators related to learning, the main contributions of the MDE can be summarized in: (1) the creation of models to better understand the learning processes; and (2) the development of more effective methods to support learning when the student studies using educational software or Virtual Teaching Learning Environment (VLEs).

Based on these claims, this research carried out a systematic mapping of the literature that aimed to verify which research topics employ Educational Data Mining in MOOCs, EDM is one of the main mechanisms for obtaining new knowledge in large volumes of educational data. The results showed that data from MOOCs can be used in many research topics, and the best way to conduct such research on these data is EDM.

The Mapping carried out pointed out that the performance analysis, behavior analysis, forum analysis and implementation of recommendation and prediction abandonment detection systems are the main lines of research, indicating frequent occurrences in the use of MOOC data for the purpose of identification and predicting events. In fact, very new topics have been identified that have the potential to be extensively

studied, such as cheating analysis, peer review, gamification, effort/capacity model and also data production in artificial students.

Acknowledgment

I thank the Federal Institute of Science and Technology Education of Rio Grande do Sul (IFRS) for the training opportunity that gave rise to this research.

I also thank my doctoral mentor Gabriela Trindade Perry.

7. References

[1] ZHENG, SAIJING; WISNIEWSKI, PAMELA; ROSSON, MARY BETH; CARROLL, JOHN M. Ask the Instructors: Motivations and Challenges of Teaching Massive Open Online Courses. Proceeding CSCW '16. Proceedings of the 19th ACM Conference on ComputerSupported Cooperative Work & Social Computing. Page 206-221. 2016.

[2] BUTCHER, N. 2014. Technologies in Higher Education: mapping the terrain. New York: Unesco, 2014. Disponível em: iite.unesco.org/pics/publications/en/files/3214737.pdf. Access 2020/06/16.

[3] MACKNESS, Jenny, Sui Mak, and Roy Williams. 2010. "The Ideals and Reality of Participating in a MOOC." In Proceedings of the 7th International Conference on Networked Learning, edited by Lone Dirckinck-Holmfeld, Vivien Hodgson, Chris Jones, Maarten de Laat, David McConnell, and Thomas Ryberg, 266–275. Lancaster: University of Lancaster, 2010.

[4] COMIER, D., Stewart, B., Siemens G. and MacAuley A. What is a MOOC? http://www.youtube.com/ watch?v=eW3gMGqcZQc, 2010.

[5] LEE, Y. Using Self-Organizing Map and Clustering to Investigate Problem Solving Patterns in the Massive Open Online Course: An Exploratory Study. Journal of Educational Computing, p. 1–20, 2018.

[6] ROMERO, C., VENTURA, S. "Educational data science in massive open online courses," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no.1, September 2016.

[7] HE, J.; BAILEY, J.; RUBINSTEIN, B. I. P.; ZHANG, R. Identifying At-Risk Students in Massive Open Online Courses. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, p. 1749–1755, 2015.

[8] LU, X.; WANG, S.; HUANG, J.; CHEN, W.; YAN, Z. What Decides the Dropout in MOOCs? In: DATABASE Systems for Advanced Applications. Cham: Springer International Publishing, 2017. p. 316–327.

[9] PURSEL, B.; ZHANG, L.; JABLOKOW, K.; CHOI, G.; VELEGOL, D. Understanding MOOC students: motivations and behaviours indicative of MOOC completion. Journal of Computer Assisted Learning, v. 32, n. 3, p. 202–217, 2016.

[10] ROMERO, C.; VENTURA, S. Educational Data Mining: A Review of the state of the art. IEEE Transactions Systems, Man, and Cybernetics, Part C: Applications and Reviews. v. 40, n. 6, p. 601-618, 2010.

[11] ROMERO, C.; VENTURA, S. Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 3, n. 1, p. 12-27, 2013.

[12] GOLDSCHMIDT, R.; PASSOS, E; BEZERRA, E. Data mining: conceitos, técnicas, algoritmos, orientações e aplicações. Rio de Janeiro: Elsevier, 2015.

[13] BAKER, R. S. J.; ISOTANI, S.; DE CARVALHO, A. M. J. B. Mineração de dados educacionais: oportunidades para o Brasil. Revista Brasileira de Informática na Educação, v.19, n. 2, p. 1-12, 2011.

[14] WULF, J.; Blohm, I.; Leimeister, J. M.; Brenner, W.; OTHERS. Massive open online courses. Business & Information Systems Engineering, 6(2):111–114, 2014.

[15] BARAK, M., Watted, A., & Haick, H. (2016). Motivation to learn in massive open online courses: Examining aspects of language and social engagement. Computers & Education, 94, 49–60.

[16] VÁZQUEZ C., E., López Meneses, E., & Sarasola, J. L. (2013). La expansión del conocimiento en abierto: Los MOOCs [The expansión of open knowledge: the MOOCs]. Barcelona: Octaedro.

[17] KITCHENHAM, B.; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, 2007.

[18] SOUZA, N. S.; WIVES, L. K.; PERRY, G. T. Tendências de pesquisas que utilizam Learning Analytics em MOOCS: um mapeamento sistemático. RENOTE, Revista Novas Tecnologias na Educação, v. 17, n. 1, p. 82-90, 2019.