10-2020

Educational Process Mining based on Moodle courses: a review of literature

José Costa

Ana Azevedo

Luís Rodrigues

Educational Process Mining based on Moodle courses: a review of literature

José Costa, CEOS.PP, ISCAP, P.PORTO, 2161161@iscap.ipp.pt
Ana Azevedo, CEOS.PP, ISCAP, P.PORTO, aazevedo@iscap.ipp.pt
Luís Silva Rodrigues, CEOS.PP, ISCAP, P.PORTO, lsr@iscap.ipp.pt

Abstract

With the prevalence of E-Learning, it is important to analyze how students progress in this environment. These systems collect data about the students' learning path, and Process Mining (PM) can provide a detailed model of this path. Based on the analysis of ten Educational Process Mining (EPM) case studies involving Moodle event logs, this article aims to contribute a literature review on EPM's research. Beyond a theoretical introduction to PM and its implications for educational data, the review concludes on what PM tools and techniques are used, as well as the challenges faced in practice. The technical options include software, process discovery algorithms and representation models. These results aim to create a list of available options for future EPM endeavors, in addition to a list of issues to consider in future research involving Moodle.

Keywords: *PM; EPM; moodle; algorithms; e-learning*

1. INTRODUCTION

Educational Process Mining (EPM) is the application of Process Mining (PM) techniques to educational data (Bogarín et al., 2018a), resulting in process models that represent the usage of Virtual Learning Environments (VLE) (van der Aalst, 2016) (Romero et al., 2016), also known as Learning Management Systems (LMS) (Folden, 2012). With Moodle being one of the most used LMS for E-Learning (Cole & Foster, 2008), it makes sense to study in detail the use of EPM in the context of this system.

The motivation and objective for writing this article is to provide a list of PM tools and techniques that can be used in EPM, as well as the research challenges faced, through a literature review. In other words, the present article contributes to the research of EPM by identifying software solutions, process discovery algorithms, representation models, and current challenges, based on case studies focused on the Moodle LMS.

The exclusive study of Moodle in this article is due to three reasons. The first is that this article is written in the context of an ongoing EPM Dissertation of the E-Business Master's Degree from Porto Accounting and Business School, with Moodle usage as case study. The second reason is the

ISSN 2183-489X

prevalence of Moodle in higher education (Rodrigues et al., 2017). The third reason is the open-source nature of this LMS (Figueira, 2009).

Beyond this Introduction, the remaining article is divided into a theoretical overview of (E)PM and Moodle (Background), an explanation of the research method used to develop the work presented in this article, the analysis of the case studies and respective results (Literature Review), and the Conclusion.

2. BACKGROUND

The background section includes a theoretical overview of EPM: the application of PM to educational data, including an introduction to Moodle, the LMS under study.

2.1. Educational Process Mining

The field of EPM consists in the application of PM techniques to the educational field but, more specifically, PM is applied to raw educational data in the form of event logs (Romero et al., 2016). This data is generated throughout the learner's interaction with the educational environment: clicks, chat history, learning resource usage and more (Bogarín et al., 2018a).

From the perspective of process discovery, EPM aims to create complete educational process models representative of the usage and behaviors that transpired. From the lens of the stakeholders (students, educators and educational institutions) (Romero et al., 2016), EPM should provide a compact view of the processes, so these agents can extract information from the models and act upon the findings obtained. To illustrate this point, the models can help educators to better guide their students or support the institution's decision-making process (Trcka & Pechenizkiy, 2009).

With all this information in mind, EPM can be described with the framework shown in Figure 1 (Bogarín et al., 2018a):

- Educational world: represents the (virtual) education stakeholders, namely students, educators and the respective educational institutions, as well as the learning resources involved;
- Virtual learning environments (VLE): the structure in which E-Learning takes place, which can be in LMS, Massive Open Online Courses (MOOC), Intelligent Tutoring Systems (ITS) or in Adaptive Hypermedia Systems (AHS);
- Event logs: data generated throughout the usage of the VLE to record the users' behavior and interaction with the environment; and
- Process model: representation of the VLE's usage in diagram-form, used to extract insights about the processes.

As it can be observed in the connection between event logs and process models in Figure 1, PM techniques can be used for three scenarios: to discover new process models (process discovery), to check the models' conformance with reality (conformance checking), and to improve the models (process enhancement) (van der Aalst et al., 2012) (Bogarín et al., 2018a).

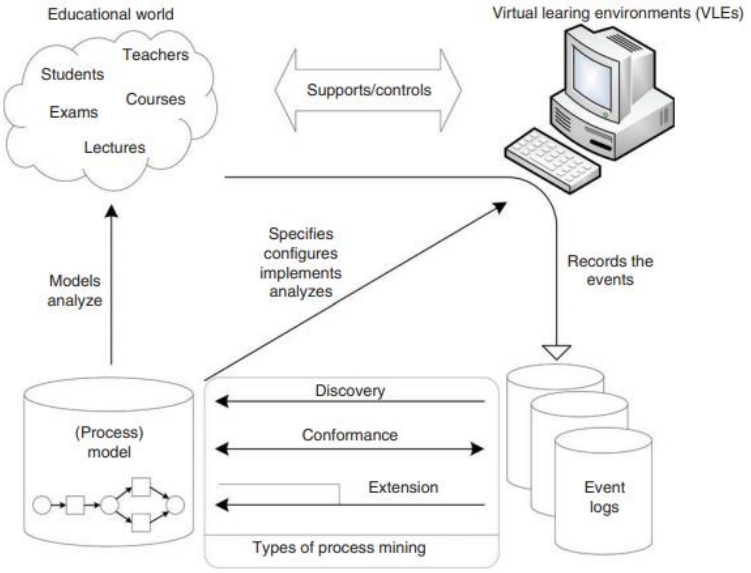

Figure 1 – EPM framework (Bogarín et al., 2018a)

2.2. Value of Process Mining

Beyond explaining what is EPM, it is necessary to highlight the value of PM and how its use in educational data can be beneficial. Much like Data Mining (DM) allows for the automatic extraction of knowledge (Cios et al., 2007), PM produces analogous results for process models via process discovery. However, and this is where the two mining fields differ, PM tightly couples event data (event logs) with process models, that is, PM provides a process-centric approach with a holistic view of the situation at hand (van der Aalst, 2016).

Through the analysis of event logs, PM can go a step further and check if a process conforms with reality, detect deviations and/or bottlenecks, and even suggest improvements (van der Aalst et al., 2012). While the two main objectives of DM are the automatic extraction of patterns and the verification of hypotheses (Fayyad et al., 1996), due to having its attention centered in data it can't provide an end-to-end perspective of processes (van der Aalst, 2016).

Regarding the data used for analysis, while DM can use almost any type of data, as long as it is significant enough for the field of application (Han & Kamber, 2012), PM requires event logs, commonly in the form of hierarchically-structured XES (eXtensible Event Stream) or MXML (Mining eXtensible Markup Language) files (van der Aalst et al., 2012). These can be generated by Information Systems, databases, data warehouses, etc. and contain at least the following fields in order to understand the processes under analysis: the activity executed, a case ID to identify the

case/set of activities to which the activity pertains, and the time of execution (timestamp) (van der Aalst, 2016).

In summary, PM acts as a bridge between the traditional model-based process analysis and data-driven analysis, namely DM, with the ability to answer questions such as what happened, why did it happen, what caused deviations in the process, what is likely to happen next, how to improve the process and more (van der Aalst, 2016).

2.3. What is Moodle?

The Modular Object Oriented Developmental Learning Environment (Moodle), is a free open-source LMS that allows for the creation and distribution of learning resources to support E-Learning and/or B-Learning (Figueira, 2009).

According to Cole & Foster (2008), Moodle manages to stand out thanks to its robustness of features and to the size of the community that supports and develops it. On top of that, as Rice (2006) puts it, Moodle uses a "social constructionist pedagogy" approach, that is, Moodle works under the philosophy that students get the most out of learning by interacting with the materials and with their peers. Among its features, it includes online tests, work submission, lessons as a set of associated learning resources with the respective assessment task, forums, text chat, wikis and more (Figueira, 2009).

3. RESEARCH METHOD

The research started with the literature retrieval. For the effect, the following queries were used: "educational process mining"; "educational process mining" AND "moodle"; and "process mining" AND "moodle". Concerning the databases, Google Scholar, IEEE Xplore, ScienceDirect, and SpringerLink were used. The number of documents found for each of the query and databases are presented in table 1. A total of 949 documents were found, considering all the databases.

QUERY	GOOGLE SCHOLAR	IEEE XPLORE	SCIENCEDIRECT	SPRINGERLINK
"educational process mining"	286	7	9	21
"educational process mining" AND "moodle"	120	0	3	8
"process mining" AND "moodle"	440	2	17	36

Table 1 – Literature retrieval results

To identify the suitable literature among the 949 results (duplicates included), the main criterion for inclusion was the relevance of the work. In other words, literature about (E)PM or case studies of the area, applied to Moodle. Authors such as van der Aalst, Romero and Bogarín (prominent authors in the field of PM and/or education), and literature released in or after 2015 were given preference. Specifically, for the EPM case studies, the reproducibility of results was also considered, that is, if it is possible to reproduce the study with the methodology and data described in the paper.

Besides the literature identified from this group, six additional documents were identified later, namely those to introduce Moodle and the software identified in the analysis, using the same inclusion criteria as above. This literature was known by the researchers ahead of time and so was included, in part, due to its known relevance.

With everything considered, this article cites a total of 25 articles and/or books. The analysis on its own cites 10 studies, with the remaining 15 documents being cited throughout the article. The analysis was limited to 10 studies, in part, due to the lack of quality of the studies found about Moodle.

After the literature retrieval and definition of inclusion criteria, the selected literature was then read and analyzed to write this article, that is, the theoretical background of (E)PM and the literature review of empirical studies.

4. LITERATURE REVIEW FOR EDUCATIONAL PROCESS MINING IN MOODLE

This literature review analyzes ten EPM case studies that involved Moodle event logs to identify the PM software, discovery algorithms and representation models used, as well as current research challenges.

4.1. Algorithms and Representation Models

The three techniques introduced previously, process discovery, conformance checking, and process enhancement, are the usages of EPM found in the literature but, among them, there are various technical options, without forgetting the accompanying tools (software).

Table 2 presents the results of the analysis of ten EPM studies with Moodle event logs, including the discovery algorithm, representation model, and software used in each study. The objective in all ten case studies was to find a new process model from the collected event logs.

STUDY	DISCOVERY ALGORITHM	REPRESENTATION MODEL	SOFTWARE USED
(Ariouat et al., 2016)	Heuristic Miner	Heuristic Net	ProM
(Bogarín et al., 2018b)	Inductive Miner	BPMN	ProM

(Cerezo et al., 2020)	Inductive Miner	BPMN	ProM
(Dolak, 2019)	Fuzzy Miner	Dependency Graph	Disco
(Dorrer & Dorrer, 2019)	Heuristic Miner	Causal Net	ProM
(Etinger et al., 2018)	Fuzzy Miner	Dependency Graph	R (bupaR) Disco
(Intayoad et al., 2018)	Heuristic Miner	Petri Net	ProM
(Juhaňák et al., 2019)	Fuzzy Miner	Dependency Graph	Disco
(Nafasa et al., 2019)	Alpha Miner	Petri Net	ProM
(Romero et al., 2016)	Heuristic Miner	Heuristic Net	ProM

Table 2 – State-of-the-Art usage of PM algorithms, representation models and software

Table 2 depicts Heuristic Miner as the most common discovery algorithm, being used in four of the ten studies. It makes sense to see it being used in education too since it is one of the most common algorithms for discovering process models (Bogarín et al., 2018b).

For the representation models, a consensus was not found. It is true Dependency Graphs were the most used (three out of ten) but, for the remaining seven studies, Petri Nets, Heuristic Nets and BPMN models were chosen twice each. Furthermore, the software used may have played a part in the choice of the representation model. While Disco is restrained to Dependency Graphs (most likely the reason for the high frequency of this representation model) (Günther & Rozinat, 2012), in a tool like ProM it is trivial to transform, for instance, a Petri Net into a Heuristic Net (van Dongen et al., 2005). When the software offers multiple ways to visualize processes, the result becomes notation-agnostic to a certain degree (van der Aalst, 2016).

4.2. Software Tools

After analyzing the results related to the algorithms and the representation models, table 3 includes a more detailed view of the features of each software used in the case studies: ProM, Disco and the R programming language (the bupaR library).

ASPECT	PROM	DISCO	R (BUPAR)
Type	Software application	Software application	Programming language library
Source	Open-source	Proprietary	Open-source
Process Discovery	Yes	Yes	Yes
Conformance Checking	Yes	No	No
Process Enhancement	Yes	No	No

Table 3 – Detailed PM software options (van Dongen et al., 2005) (Günther & Rozinat, 2012) (Janssenswillen et al., 2019)

ProM is used in seven out of the ten case studies most likely due to its robustness and plethora of technical options included (van Dongen et al., 2005), but Disco is good for quick exploration and fast results (Günther & Rozinat, 2012). However, Etinger et al. (2018) combined the quick data

exploration of Disco with a programming language's interface (bupaR) to discover a process model. While this R library is a collection of eight complementary libraries, bupaR is considered the central one (Janssenswillen et al., 2019), hence using it as the identifier.

4.3. Challenges

Table 4 summarizes the challenges identified in the case studies, ordered by frequency.

CHALLENGES	FREQUENCY	REFERENCES
Data pre-processing	7	(Bogarín et al., 2018b) (Dolak, 2019) (Dorrer & Dorrer, 2019) (Etinger et al., 2018) (Intayoad et al., 2018) (Juhaňák et al., 2019) (Nafasa et al., 2019)
Comprehensibility of results	3	(Bogarín et al., 2018b) (Cerezo et al., 2020) (Romero et al., 2016)
Portable solutions	3	(Bogarín et al., 2018b) (Cerezo et al., 2020) (Romero et al., 2016)
Integration of different types of data	2	(Cerezo et al., 2020) (Juhaňák et al., 2019)
Semantics	2	(Ariouat et al., 2016) (Romero et al., 2016)
Decomposed process discovery	1	(Ariouat et al., 2016)
Logging limitations	1	(Juhaňák et al., 2019)

Table 4 – Challenges identified in EPM studies

Data pre-processing is an important initial step in the application of PM because it ensures the anonymity of students and, unfortunately, this is not integrated in Moodle yet (Juhaňák et al., 2019). Additionally, there is a need for noise-filtering as Moodle event logs can easily lead to unstructured process models (Etinger et al., 2018).

Another concern is that the resulting process model needs to be as easy to understand as possible (Romero et al., 2016). The notation and parameters used to represent the final model need to consider how it impacts the model's comprehensibility for the stakeholders/end-user.

Besides data pre-processing and comprehensibility, it is essential to work towards solutions that can be used in multiple scenarios, that is, EPM solutions that can be applied in multiple educational fields (Bogarín et al., 2018b) and multiple (virtual) learning environments (Cerezo et al., 2020).

Without this extended application and/or testing, the quality of results is not certain and new challenges may be detected (Juhaňák et al., 2019).

In this context of further application, it is important to consider the integration of different types of data (for example, time) and semantics. Event logs do not capture all the possible scenarios for the process model (van der Aalst, 2016), and these two factors would allow the learning paths to be analyzed in greater detail (Cerezo et al., 2020).

When clustering techniques are used in the analysis, Ariouat et al. (2016) reminded the need for a solution for the whole and not just the clusters created. It is easier to work with groups of smaller data, especially to deal with the heterogeneity of data (Romero et al., 2016), but the bigger picture cannot be forgotten.

5. CONCLUSION

After an introduction to the research of PM applied to educational data, from both the theoretical and the practical lenses, it is clear that this field can create value for the E-Learning stakeholders: students, educators, and educational institutions. However, there are challenges to be solved, namely the scarcity of EPM studies dedicated to Moodle.

On one hand, EPM creates value by providing information to educators on how to lead students in better learning paths, and by supporting the decision-making process of the educational institutions. On the other hand, in accordance with the challenges identified in the literature review, data pre-processing, comprehensibility of the results and portability of solutions are the three leading challenges.

Beyond the challenges identified, the review also highlighted ProM, Disco and bupaR (a R programming language library) as three software options for EPM, with differing levels of application and features. ProM is used as a complex yet robust tool, Disco as a simpler tool for quick results and data exploration, and bupaR as a PM extension to the R programming language.

The review did not find a single definitive answer for what discovery algorithm to use, rather, it found a set of algorithms that should be considered by researchers when undertaking a project of this nature, at least in the context of E-Learning. With that said, the literature review highlighted the Heuristic Miner and the Fuzzy Miner algorithms as the most used for process discovery, with none of the ten analyzed case studies having included conformance checking or process enhancement techniques, formally.

Regarding the representation models used to visualize the obtained educational process models, several options were found: Dependency Graphs, Petri Nets, Heuristic Nets and BPMN models. These were common choices in the studies reviewed.

While the field of EPM is by all means a promising one for the improvement of E-Learning, there are still challenges in its application, especially the portability of solutions across different fields of study and/or learning environments, and comprehensibility of the results for the end-users. For instance, even if a model was fit for Engineering students, the methodology and techniques may need to be changed for students of other fieds. Plus, there is a concern with the mined process being too complex for the end-user to understand it (for instance, an educator). Specifically, in the context of Moodle, while pre-processing event logs is not a big technical hurdle, it was still the most reported issue, in large part because this pre-processing is not integrated in the system.

Concerning the limitations, this study was limited by the scarcity of EPM literature, especially of case studies to support the literature review. Consequently, the number of studies analyzed was limited to ten by choice of the researchers.

Thus, the three prominent EPM challenges to consider in future research are (the integration of) data pre-processing in Moodle, creation of comprehensible results for the end-users and solutions that can be applied across different areas of study and learning environments.

ACKNOWLEDGEMENTS

This work is financed by Portuguese national funds through FCT - Fundação para a Ciência e Tecnologia, under the project UIDB/05422/2020.

REFERENCES

Ariouat, H., Cairns, A. H., Barkaoui, K., Akoka, J., & Khelifa, N. (2016). A Two-Step Clustering Approach for Improving Educational Process Model Discovery. 2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 38–43. https://doi.org/10.1109/WETICE.2016.18

Bogarín, A., Cerezo, R., & Romero, C. (2018a). A survey on educational process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(1), e1230. https://doi.org/10.1002/widm.1230

Bogarín, A., Cerezo, R., & Romero, C. (2018b). Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). Psicothema, 30.3, 322–329. https://doi.org/10.7334/psicothema2018.116

Cerezo, R., Bogarín, A., Esteban, M., & Romero, C. (2020). Process mining for self-regulated learning assessment in e-learning. Journal of Computing in Higher Education, 32(1), 74–88. https://doi.org/10.1007/s12528-019-09225-y

Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (Eds.). (2007). Data mining: A knowledge discovery approach. Springer. https://link.springer.com/chapter/10.1007/978-0-387-36795-8_2

Cole, J. R., & Foster, H. (2008). Using Moodle: Teaching with the popular open source course management system (2nd ed). O'Reilly Community Press.

Dolak, R. (2019). Using Process Mining Techniques to Discover Student's Activities, Navigation Paths, and Behavior in LMS Moodle. In L. Rønningsbakk, T.-T. Wu, F. E. Sandnes, & Y.-M. Huang (Eds.), Innovative Technologies and Learning (Vol. 11937, pp. 129–138). Springer International Publishing. https://doi.org/10.1007/978-3-030-35343-8_14

Dorrer, M., & Dorrer, A. (2019). Generation of agent simulation models by using process mining methods on the example of E-learning process. Journal of Physics: Conference Series, 1399, 033077. https://doi.org/10.1088/1742-6596/1399/3/033077

Etinger, D., Orehovački, T., & Babić, S. (2018). Applying Process Mining Techniques to Learning Management Systems for Educational Process Model Discovery and Analysis. In W. Karwowski & T. Ahram (Eds.), Intelligent Human Systems Integration (Vol. 722, pp. 420–425). Springer International Publishing. https://doi.org/10.1007/978-3-319-73888-8_65

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37–37. https://doi.org/10.1609/aimag.v17i3.1230

Folden, R. W. (2012). General Perspective in Learning Management Systems. In Babo, R., & Azevedo, A. (Eds.), Higher Education Institutions and Learning Management Systems: Adoption and Standardization (pp. 1-27). IGI Global. http://doi:10.4018/978-1-60960-884-2.ch001

Figueira, Á. (2009). Moodle—Criação e Gestão de cursos online. https://hdl.handle.net/10216/84783

Günther, C. W., & Rozinat, A. (2012). Disco: Discover Your Processes. BPM. BPM 2012 Demonstration Track, Estonia.

Han, J., & Kamber, M. (2012). Data Mining: Concepts and Techniques (3rd ed). Elsevier.

Intayoad, W., Kamyod, C., & Temdee, P. (2018). Process mining application for discovering student learning paths. 2018 International Conference on Digital Arts, Media and Technology (ICDAMT), 220–224. https://doi.org/10.1109/ICDAMT.2018.8376527

Janssenswillen, G., Depaire, B., Swennen, M., Jans, M., & Vanhoof, K. (2019). bupaR: Enabling reproducible business process analysis. Knowledge-Based Systems, 163, 927–930. https://doi.org/10.1016/j.knosys.2018.10.018

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. https://doi.org/10.1016/j.chb.2017.12.015

Nafasa, P., Waspada, I., Bahtiar, N., & Wibowo, A. (2019). Implementation of Alpha Miner Algorithm in Process Mining Application Development for Online Learning Activities Based on MOODLE Event Log Data. 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), 1–6. https://doi.org/10.1109/ICICoS48119.2019.8982384

Rice, W. H. (2006). Moodle: E-learning course development: a complete guide to successful learning using Moodle (1. publ). Packt Publ.

Rodrigues, S., Rocha, A., & Abreu, A. (2017). The use of Moodle in Higher Education: Evolution of teacher's practices over time. 2017 12th Iberian Conference on Information Systems and Technologies (CISTI), 1–4. https://doi.org/10.23919/CISTI.2017.7975702

Romero, C., Cerezo, R., Bogarín, A., & Sánchez-Santillán, 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 (pp. 1–28). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118998205.ch1

Trcka, N., & Pechenizkiy, M. (2009). From Local Patterns to Global Models: Towards Domain Driven Educational Process Mining. 2009 Ninth International Conference on Intelligent Systems Design and Applications, 1114–1119. https://doi.org/10.1109/ISDA.2009.159

van der Aalst, W., Adriansyah, A., de Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., van den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., de Leoni, M., … Wynn, M. (2012). Process Mining Manifesto. In F. Daniel, K. Barkaoui, & S. Dustdar (Eds.), Business Process Management Workshops (Vol. 99, pp. 169–194). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28108-2_19

van der Aalst, W. (2016). Process Mining: Data Science in Action. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4

van Dongen, B. F., de Medeiros, A. K. A., Verbeek, H. M. W., Weijters, A. J. M. M., & van der Aalst, W. M. P. (2005). The ProM Framework: A New Era in Process Mining Tool Support. In G. Ciardo & P. Darondeau (Eds.), Applications and Theory of Petri Nets 2005 (Vol. 3536, pp. 444–454). Springer Berlin Heidelberg. https://doi.org/10.1007/11494744_25