Process mining on students' web learning traces: a case study with an ethnographic analysis

Roberto Nai[1 [0000−0003−4031−5376]], Emilio Sulis[1 [0000−0003−1746−3733]], Elisa Marengo[1 [0000−0003−1879−2088]], Manuela Vinai[2 [0000−0002−1044−2764]], and Sara Capecchi[1,3 [0000−0001−6157−2932]]

[1] Department of Computer Science - University of Turin
[2] Department of Philosophy and Education Sciences - University of Turin
[3] Laboratorio CINI Informatica e Scuola

Abstract. The study of learning processes can benefit from examining the digital traces left by students while browsing educational platforms. In this work, we propose an analysis of how process mining techniques can be combined with online educational technologies. In particular, we describe a process discovery experiment based on students' movements between the web pages' paragraphs of a tutorial specifically built for teaching the programming language Python. In addition, we explore the ethnographic and conceptual genealogy of a learning project focused on a programming tutorial. Observations of the digital traces are accompanied by qualitative reflections with three objectives: human-machine interactions in a learning laboratory, the relationship between individuals and the environment, and the relationship between disciplines (alliance between computer science and anthropology) through the concept of consilience.

Keywords: Educational process mining · E-learning · Learning process discovery · Ethnography.

1 Introduction

Student pathways during learning can be monitored from multiple perspectives. The success of a learning process is influenced by several factors, including attention, duration, classroom environment, and motivation. The learning outcome can be the subject of quantitative analysis (e.g., the number of positive responses to a test) or qualitative analysis (e.g., the degree of appreciation and feedback left by students during the course). Technologies make it possible to track students behavior during learning through computer instrumentation. In this work, we exploit digital traces left by students in a web-based tutorial. The content of each page can be divided into paragraphs, so it can be monitored to understand students' movements between parts, as well as clicks or mouse movements. Timed events can then be examined using appropriate techniques to appreciate the learning process. The discipline of Process Mining (PM) perfectly fits for this purpose, and in this paper we explore process discovery techniques. As a matter

of fact, observational techniques have already shown the possibility of analysing a learning situation. This paper also introduces an exploration of the learning context with classical observational techniques, peculiar to ethnology. We consider as a case study a web tutorial of ten pages on the programming language Python[4]. The use of PM and observational techniques during the execution of the tutorial is twofold. First, to provide useful insights for improving the tutorial itself; aggregated data can be used to suggest how and where to improve the tutorial, to find bottlenecks in the sequence of activities, delays, nonlinear paths. Second, the analysis of the collected data can be used as an evaluation tool for teachers: results can provide suggestions for the improvement of student learning by identifying outliers and learning differences among students. In the following, Section 2 provides the background of our work by introducing some related work and the case study. Section 3 explores the process analysis, also with an ethnographic perspective. Finally, Section 4 presents conclusions.

2 Background and related work

Educational process mining (EPM) is a relatively new discipline which aims at studying students learning behaviours. This information is then used in various ways such as to correlate learning paths with performances or automatically extract learning preferences [2,8]. One of the key aspects contributing to the application of EPM is the availability of data which is supported by the number of courses and learning material available over the web or through Learning Management Systems (LMS) [3], among which Moodle [6]. Etinger et al. [4] and Real et al. [9] compared learning paths of different students and classify them according to their performances. We add an observation from the ethnographical perspective [11]. A LMS offers advantages for teachers and users. It supports data collection, organization, and user-friendliness. In this paper we propose an approach which goes in the direction of supporting the personalization and leaving a student the possibility of selecting the preferred learning path.

2.1 Web tutorial

We proposed a web tutorial of ten pages (in Italian) to introduce Python programming language to newcomers. Between two web-pages, a test asks for a quick answer to check learning. Each page includes three or four paragraphs (sections), up to a total of 34 paragraphs in the tutorial. Finally, the user is invited to complete a satisfaction survey. To investigate the role of the activities order in the learning process, we propose three different sequences. The tutorial has two fixed parts: first, the INTRODUCTION lessons of three web-pages, including a presentation, the meaning of a program, and variables. Second, the last page (FUNCTIONS). For the middle part of the tutorial, students can select one out of three paths, as in Figure 1. Each lesson on DATA TYPES, DATA

[4] `https://www.python.org`

STRUCTURES, and CONTROL STRUCTURES includes two web pages. To prevent order to influence students, the three topics are presented in a web page as rotating blocks, as in Figure 2.

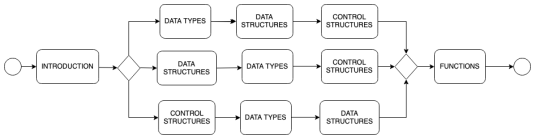

Fig. 1. The three different learning paths; on the top, the standard learning flow. Full size image available at `https://github.com/roberto-nai/ectel2023`

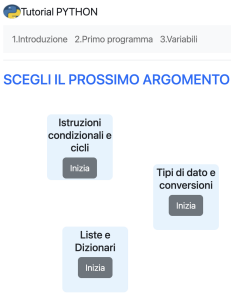

Fig. 2. The rotating blocks allowing students to select next lessons (web-page in Italian). Full size image available at `https://github.com/roberto-nai/ectel2023`

2.2 Case study

The tutorial has been proposed to students of an Italian university course, not expert in computer science. During the tutorial, an observational study was conducted by an ethnologist. The data collected via the web tutorial are: entering a page, scrolling through a paragraph (section) of the page, exiting the page, accessing the quiz related to the page, and as last activity accessing the satisfaction survey.

2.3 Process mining

The PM discipline is oriented toward timed events in which the sequence of activities is important [1]. Process discovery algorithms are of particular interest to this case study since they allow us to reconstruct the students learning paths. The data described in Section 2.2, used for process discovery, were collected through the web tutorial and saved in a SQL relational database. Each

event record (*event*) in an event log includes at least three basic features [1]: 1) the identifier of the process instance it belongs to; 2) the name of the activity which generated the event; 3) the execution timestamp. The sequence of events generated by the same process forms a *trace* or *case* [1]. To analyse the event logs we used academically licensed Apromore [7][5].

2.4 Ethnography and learning

The idea of adding traditional observations to digital observations, applies first and directly to the classroom context. Direct observation is used as a source of ethnographic information, enables contextual data collection and leads to a descriptive type of restitution of the setting (in which project activities take place). Here *ethnographic research* is proposed, alongside the proposition of the tutorial by the computer scientists involved. At the theoretical level, an effort has been made to study how science is produced. We introduce concepts that can help contextualize the case study. In particular, qualitative reflections have three objectives: human-machine interactions in a learning laboratory, the relationship between individuals and the environment, and the relationship between disciplines through the concept of consilience.

3 Results of the observational study

3.1 Event log

Following Section 2.3, in our case the session ID (an alphanumeric string generated for each http connection to the web tutorial) is the process instance identifier, while the different activities involved in the process are 11, starting from the INTRODUCTION up to FUNCTIONS and then the survey page (Section 2.1). Each event includes the timestamp at the granularity level of day, hour, minute and second on which the event occurred. Two event logs were analysed: one at the page level and one at the level of individual paragraphs within pages. In both cases, we focused on complete traces, i.e. cases reaching FUNCTIONS or survey page. We have 70 valid cases, with an average duration of 37.7 minutes, a median of 31.6, and a standard deviation of 24.6. Variability of duration is high, in a range from 3 minutes up to one hour and a half. Sampling data is available[6].

3.2 Observation from digital traces

Process discovery outputs described in Figure 3, show the case frequency: the paths by users after the third page (VARIABLES) are fairly evenly distributed (36% for route A, 27% for route B, and 37% for route C). We can explore meaningful research directions, i.e. by considering students reaching good or bad

[5] `https://apromore.com`
[6] `https://github.com/roberto-nai/ectel2023`

results in quizzes. We considered the correct answers to obtain two balanced subsets of cases, the one with a percentage of correct quizzes below 70% (i.e. 31 cases of *low performance*) and above 70% (i.e. 39 cases of *good performance*). The aggregate learning processes appear immediately different: Figure 4.a shows a quite linear learning process for *low performance* cases, where the mean case duration is 28.4 minutes. Figure 4.b suggests how best students have a more complicated learning process, having a mean duration of 43.6 minutes. The comparison suggests that who spent more time on the tutorial pages and visited some pages more than once, also answered the quizzes better.

Fig. 3. Discovered process of paths among pages and quizzes. Full size image available at `https://github.com/roberto-nai/ectel2023`

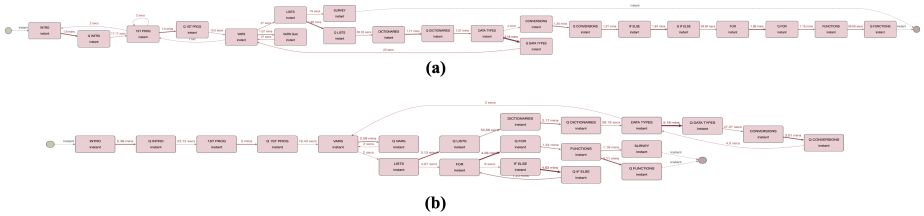

Fig. 4. Processes of low performing students (a) and good performing ones (b). Full size image available at `https://github.com/roberto-nai/ectel2023`

.

3.3 An ethnographic insight

The ethnographic exploration of the study project is well represented by *truthfulness function* identified by Cassin [5], which investigates the stages of data production. The observation of the classroom's context in which the tutorial was administered highlights the variables of interest, namely, the numerosity of students, the distribution in space, and the possibilities for comparison. The presence of the lecturer providing support for students in case of doubts facilitated the conduct of the tutorial. The proximity among students fostered a positive climate for discussions aiming at overcoming minor difficulties. Accordingly to the concept of consilience [10], the observation suggests to take into account the twofold direction between the need for informatics to consider the technical construction of the tutorial, and the anthropological one to focus on the context, i.e.

the type of classroom, the noisiness or quietness of the environment, the time of performance, and finally the relationality aspects.

4 Conclusions

This study combines digital traces and ethnographic observation to investigate students' behaviors during learning. The method is useful for discovering navigation processes on web-tutorial to identify meaningful patterns. Event logs can be used to study user experience, e.g. when students stop navigating after a series of events. PM also helps understand students' learning processes by separating students based on their performance. The timing of learning processes lead to higher achievement and recommending them to other students. Future work plans include increasing the number of students involved, verifying learners' conformance to typical direct paths, and restitution through ethnographic narrative style, including interviews and focus groups among students.

Acknowledgements This research has been partially carried out within the projects "AI-LEAP" and "Circular Health for Industry", funded by the "Compagnia di San Paolo".

References

1. van der Aalst, W.: Process mining: data science in action, vol. 2. Springer (2016)
2. Bogarín, A., Cerezo, R., Romero, C.: A survey on educational process mining. WIREs DM Knowl. Discov. **8**(1) (2018)
3. Cavus, N.: Distance learning and learning management systems. Procedia-Social and Behavioral Sciences **191**, 872–877 (2015)
4. Etinger, D.: Discovering and Mapping LMS Course Usage Patterns to Learning Outcomes. In: IHSI. pp. 486–491. Springer (2020)
5. Fassin, D.: If truth be told: the politics of public ethnography. Duke University Press (2017)
6. Gogan, M.L., Sirbu, R., Draghici, A.: Aspects concerning the use of the moodle platform–case study. Procedia Technology **19**, 1142–1148 (2015)
7. La Rosa, M., Reijers, H.A., van der Aalst, W., Dijkman, R.M., Mendling, J., Dumas, M., García-Bañuelos, L.: Apromore: An advanced process model repository. Expert Systems with Applications **38**(6), 7029–7040 (2011)
8. Racca, A.R., Sulis, E., Capecchi, S.: Behavioral web tracking in e-learning: An educational process mining application. In: et al., E.B. (ed.) 26th International Conference Information Visualisation, IV 2022, Vienna, Austria, July 19-22, 2022. pp. 269–274. IEEE (2022). https://doi.org/10.1109/IV56949.2022.00053
9. Real, E., Pimentel, E., Oliveira, L., Braga, J., Stiubiener, I.: Educational Process Mining for Verifying Student Learning Paths in an Introductory Programming Course. pp. 1–9 (10 2020)
10. Slingerland, E., Collard, M.: Creating consilience: Integrating the sciences and the humanities. Oxford University Press (2011)
11. Wilson, S.: The use of ethnographic techniques in educational research. Review of educational research **47**(2), 245–265 (1977)