Behavioral Web Tracking in e-Learning: An Educational Process Mining Application

Behavioral web tracking in e-learning: an educational process mining application

Andrea Rocco Racca
Computer Science Department
University of Torino
Torino, Italy
0000-0002-8158-2561

Emilio Sulis
Computer Science Department
University of Torino
Torino, Italy
0000-0003-1746-3733

Sara Capecchi
Computer Science Department
University of Torino
Torino, Italy
0000-0001-6157-2932

***Abstract*—This paper introduces an experiment and the first results of a research on computer programming education using process mining methods. A web-based tutorial addresses the topic of agent-based modeling by introducing a guided exercise with NetLogo, a widely used tool for modeling natural and social phenomena. A goal of the project is to analyze the goodness of the learning process, also through appropriate tests placed between the pages and at the end of the tutorial. Actual data extracted on student behavior (e.g., length of time spent on different parts of each web page, movements on the page, mouse position and mouse clicks) are examined using process discovery technique. Special attention is given to the return of student learning outcomes through visualization. Our solution includes heatmaps and direct-follow graphs of the real processes. Initial results are encouraging on the possibility of improving the assessment of learning processes by relying on techniques from the discipline of process mining, as shown by the case of web-based behaviour tracking data.**

***Index Terms*—Educational process mining, User behaviour, Learning process, Online learning**

I. Introduction

Online teaching delivery is receiving increasing attention, with relevant implications for extending teaching and training opportunities to a broad spectrum of students. The delivery of online courses also had an additional impact during the recent pandemic. Web-based teachings and online tutorials make it possible to take new directions to research. In the area of learning analytics (LA), several studies plan to exploit new technologies to infer the success and goodness of learning, by proposing new courses or identifying personalized courses. The large amount of data that can be collected has therefore fostered the emergence of an area of research involving learning processes with educational data mining (EDM). In this direction, a typical possibility involves predicting behavior from large volumes of educational data. In fact, this situation makes it possible to take advantage of the new opportunities offered by web programming to improve learning and analyze student data, as in the case of tracking user behavior. In particular, the traces left by users are very useful to understand whether a course is successful and where it can be improved.

This research paper explores educational process mining (EPM) as a relatively new research discipline in which the learning process plays a central role. In this regard, we have created a web tutorial on an introductory topic for use in our experiment. The digital traces left by students during the course of the tutorial can be used to discover, analyze, and appropriately visualize the students' learning process.

The goal of collecting data during the execution of the tutorial is twofold: first, to provide useful insights for improving the tutorial itself; second, to propose suggestions for the improvement of student learning (also by identifying outliers). In fact, aggregate data can be used to suggest how and where to improve the tutorial, to find bottlenecks in the sequence of activities, delays, nonlinear paths.

Behavioral web tracking investigates where students may be having difficulty or may report where the tutorial seems lacking. The immediate return in terms of visualization, either through heatmaps or flowcharts on actions taken, further facilitates the achievement of this goal. Results about the visualization possibilities related to learning tracking are finally presented by referring to an initial administration of the tutorial to high-school students in Piedmont, a region in northern Italy.

In the remainder of the paper we first introduce related work and the use case in section II. Then we describe the methodological framework in section III. The results are detailed in section IV, as well as the visualization efforts in section V. Finally, some discussions are included in section VI, while section VII concludes the paper.

II. Background and related work

EDM and LA studies have evolved rapidly over the past decades, with methods and tools applied to educational data increasingly focusing on e-learning systems [1]–[3].

Online learning systems register logs of the digitized traces of the users in a cyberlearning system. In [4], authors tracked data of 80 students from a total of four course sections. The data collected included the amount of time that each student spent using each component of the learning system or the paths used to navigate through these components. Similarly, clustering methods have been applied to the analysis of large volumes of usage tracks from the various actions performed by a user in a digital learning environment. The events tracked can be the timed sequence of actions completed by a user/learner, including navigation through each web-page or mouse clicks [5]. Several works focused on predicting student success in a learning online environment [6], also by implementing nudges [7].

The adoption of machine learning (ML) methods has grown in the last years with the design, implementation and delivery of web-based education systems [8], [9]. Learning Management Systems (LMS) increasingly store a large amount of educational data used to improve the learning, the teaching and the administration processes. ML algorithms improve their performance with experience, in many fields of research such as those contexts where students interact with learning systems leaving useful tracks.

In [10], e-learning systems register the student's interactivity and behavioral features. Therefore, the application of data mining techniques in an educational background discovered hidden knowledge and patterns to support the decision-making processes for improving the educational system. Another work predicts learner performance, engagement and potential problems by extracting useful information from web-based LMS data [11]. This analysis supports organizations to undertake actionable decisions as preventive measures for the overall teaching and learning support.

Process Mining (PM) is a recent discipline focused on extracting knowledge from the log of events recorded in the information system [12]. A recent review of the EPM state of the art [13] suggests the behavioral research, as proposed in our work, as a promising research direction, as well as collaboration, interaction and performance in the development of teaching-learning activities [14].

An ongoing educational project. Our work improves the above mentioned research direction with an EPM framework that monitors students' web behavior to detect relevant information and propose learning suggestions. We based on a previous project proposing to introduce digital competences in high schools, especially where curricula do not contain Computer Science [15]. The suggestion concerns the adoption of an interdisciplinary perspective for students and researchers, such as agent-based modeling (ABM) [16]. On the one hand, this is useful in stimulating students toward a deeper understanding of topics in disciplines already in their curricula such as mathematics, physics, and chemistry. Moreover, the approach leads students to acquire programming skills that give them the ability to modify simulations they have already developed. In particular, a web tutorial is proposed to approach programming using a well-known tool in the field of natural and social sciences, such as NetLogo [17]. The tool is suited to approach computer programming by the ease of Logo-like language, designed precisely to introduce the key concepts of programming.

III. Methodology

A. Tutorial Web

Our methodological framework is based on the construction of a student behavior tracking system within a six-page web tutorial to learn NetLogo [1]. The tutorial proposes to develop an easy example of model, which we have created for the occasion, and which can facilitate the learning of the basic constructs of the tool. During the course of the tutorial the behavior of the student in the six web pages can be analyzed. This section introduces the tutorial content.

[1]https://abmsim.di.unito.it/tutorial-netlogo/

A practical example. The tutorial describes the passages to create a simple NetLogo model. To approach the student with a nice modeling and simulation example, we designed a basic model called "Agents on the move". The model concerns agents (people) moving in an environment where there are two types of obstructions: walls and puddles in the pavement (patches) that prevent or delay movements. Agents wander randomly, until sooner or later two agents will cross paths. We keep track of these encounters by representing them with a link. The simulation ends when all agents have crossed paths. Such a simple model makes it possible to represent diffusion phenomena, such as in the case of contagion or spread news [18].

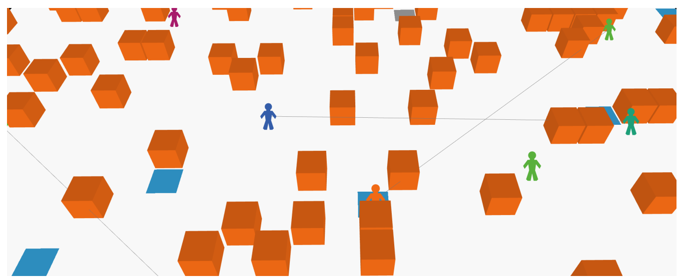

Fig. 1. A view of the NetLogo program proposed in the tutorial.

Six pages and tests. Each part of the tutorial focuses on specific subjects. The six pages focus on different aspects of programming. A first page introduces ABM and the tool, while the second page describes the main ingredients of a modeling effort. The third page creates simple agents moving around in the environment. A fourth page improves the environment in which agents move by adding obstacles (walls) or slow-downs (puddles). The fifth page enhances agent behavior by adding links between agents in contact, and finally the sixth page introduces monitoring and key performance indicators to perform *what-if* analysis.

An obligatory test allows students to move from one page to the next. It has the dual purpose of collecting results on the learning outcomes, as well as returning feedback to the student, motivating him or her to continue on the learning path. Finally, a question is posed to the users about the evaluation of the tutorial, with four allowable answers (excellent, good, medium, poor).

B. Technologies

The technology used in the present work includes both an implementation part, related to web programming and data collection, and an examination part of the information collected, using process mining tools.

Web programming. On the web programming side, Angular was used for the front-end part, then HTML, CSS and typescript. To host the front-end part the Netlify service was used[2]. For the back-end part nodeJS was used, then JavaScript and a SQL database all hosted on Heroku[3]. Cookies allow to track user sessions for investigating individualized learning.

Educational process mining. To explore the event logs collect in the research we adopted typical tools for process mining, i.e. ProM[4] and python's library PM4PY[5]. Timed events recorded in the information system were processed with a python script to build a log file to be provided as input to the PM tools.

C. User behaviour indicators

The tracking of user behavior data during the execution of the tutorial can be grouped into appropriate indicators [3]. The goal is to describe student learning with respect to the following points:

- *duration* - the user's stay in the tutorial, either on a single page or on a specific part of the page; the overall time of the tutorial execution is the so-called *cycle time*;
- *movements* - the user's movements forward and backward on the page and between pages of the tutorial;
- the *positioning* of the user in the page during the course of the tutorial (eg, where the users' mouse remains stationary for a longer time);
- *clicks* made on each page and in total.

We are aware of the difference between mouse movements and actual attention or looking at the web page [19], but we are interested in exploring the usefulness of digital traces. These indicators can be computed at the individual and aggregate level (for example, by group of students, as in the case of a class).

D. Research in classroom

This tutorial was tested with two classes from a high school in northern Italy. Fourth-year classes were chosen, with boys aged 17-18 years, from a section of scientific high school, whose curricula do not include high-level courses in computer programming. Students performed the online tutorial independently and individually, after a brief introduction by the teacher, who presented the initiative and remained available during the execution of the tutorial to monitor the autonomous performance, support any technical difficulties, as well as to collect useful elements from a qualitative perspective.

E. Visualization effort

The results have been presented to the stakeholders in a way that makes learning outcomes appreciable, either of the individual student or aggregated by group to which they belong, such as a class or a portion of students with certain characteristics (e.g., students who did well versus those who responded poorly to tests). Visualization also makes it easier to redesign the teaching effort, i.e., the tutorial used in this exercise, by identifying the most problematic and the most successful parts. To present the results, we have identified two main types of visualization (see, for example, figures 3 and 4 in the Results section). First, heatmaps, capable of highlighting with warm colors (red, orange) the points with higher density in a space, while low density is presented with green/blue colors. Second, diagrams allow to show the flow between parts of the tutorial. In particular, we adopt direct-follow graph (DFGs), which are easy to understand. This visualization is frequent in PM to indicate the number of times a single activity is performed by placing the number in parentheses next to the activity name, framed by a rectangle. The flow is indicated by the arrows, which are weighted by the number of occurrences.

F. Privacy issues

This tutorial protect the confidentiality of students, as we do not ask for personal information, such as surname, e-mail, telephone number, as not useful for mining purposes. The system currently does not record sensitive data, but whether such information will be collected in the future, attention must be paid to data protection regulations, as well as we can use guidelines about ethical issues when using educational data [20]. Another aspect to note is that to ensure awareness of the purpose of the study, the initial presentation by the lecturer and a home page describe the purpose of the research.

IV. Results

A. Overview

The tutorial has been administered to two high school classes, for a total amount of 45 students. The web tutorial received a fair rate of appreciation in the students, as clearly noted during the execution of the project. This is also described by the rate of students who completed the whole tutorial, coming up with the last satisfaction test (82%). Similarly, the number of students who completed each of the six tests posed during the web tutorial sees a slight decline: it went from full participation in the first test (45), with a steady decline in the second (44), third (41), fourth (39) and fifth tests (39). We believe that this is a physiological decrease, even considering that the tutorial was proposed voluntarily and administered during school hours.

a) Test results: For students' evaluation, test results were saved anonymously and the errors per capita have been calculated (see *#Errors* columns in Table I).

Regarding the five tests provided in the tutorial among the six web pages, most students either made no errors or one error (73.3%). In one out of four cases, students made two errors (24.5%). Only one case recorded a somewhat high number of errors. Figure 2 describes the error rate of students.

B. Web tracking

a) Duration: Among the information saved for further analysis is also the average length of stay on every single page, so that the complexity of the subject and the students' concentration on each topic can be easily understood.

[2]https://www.netlify.com
[3]https://www.heroku.com
[4]https://www.promtools.org/doku.php
[5]https://pm4py.fit.fraunhofer.de/

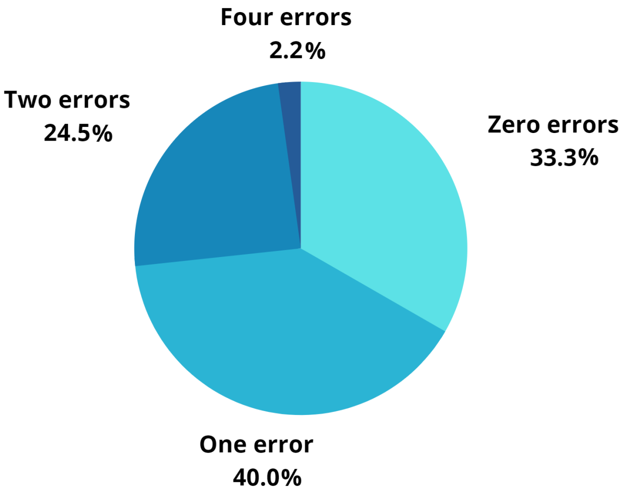

Fig. 2. Chart regarding students' errors rate.

Table I describes a sample of the main indicators concerning the information about the students tracking in the learning process including the average duration (AvgDur) of the visits for each page (from PAGE1 to PAGE6). For example, PAGE5 and PAGE6 have an average duration of more than 10 minutes, while the initial two pages last about three minutes. It is interesting to note that some pages seem more complex than others, a signal that could be a warning about the parts of the tutorial considered more difficult by the students. This is a suggestion to try to make the content more homogeneous in pages that see a high duration.

b) Movements: In order to understand how much the user focused on the topics and how much the user understood of the tutorial, information concerning the student's movement between paragraphs, i.e. how many times the student returned to topics already covered, has been saved.

In Table I, the average number of forward (*AvgFor*) and backward movements (*AvgBack*) per page are introduced in order to better understand which topics may have been more difficult for the students to understand. It is possible to appreciate the high average number of movements in last pages: PAGE5 reaches the highest value with an average of about 27 forward and 22 backward movements, compared with the lowest value recorded for PAGE2, with an average number of 10 and 7 movements, respectively.

c) Click: Furthermore, the number of clicks of every student on each single page of the tutorial was recorded. Thanks to the collection of this data, it was possible to calculate the average number of clicks on each page, and as expected, some pages revealed a higher number of clicks being those where the student was more hesitant to answer due to the difficulty of understanding the specific topic. Only the first two pages record a very low average number of clicks, a sign that perhaps the complexity of these two pages could be increased.

TABLE I
A SAMPLE OF INDICATORS FOR EACH OF THE SIX PAGE OF THE WEB TUTORIAL

Page number	**Saved information**				
	AvgDur	***AvgCli***	***AvgFor***	***AvgBack***	***#Errors***
PAGE1	3m 42s	1,93	11,71	9,53	3
PAGE2	2m 42s	4,13	10,16	7,02	2
PAGE3	5m 24s	10,50	10,14	8,05	3
PAGE4	6m 34s	13,20	14,59	12,10	6
PAGE5	10m 07s	14,46	27,46	22,72	13
PAGE6	12m 32s	12,69	12,49	10,97	17

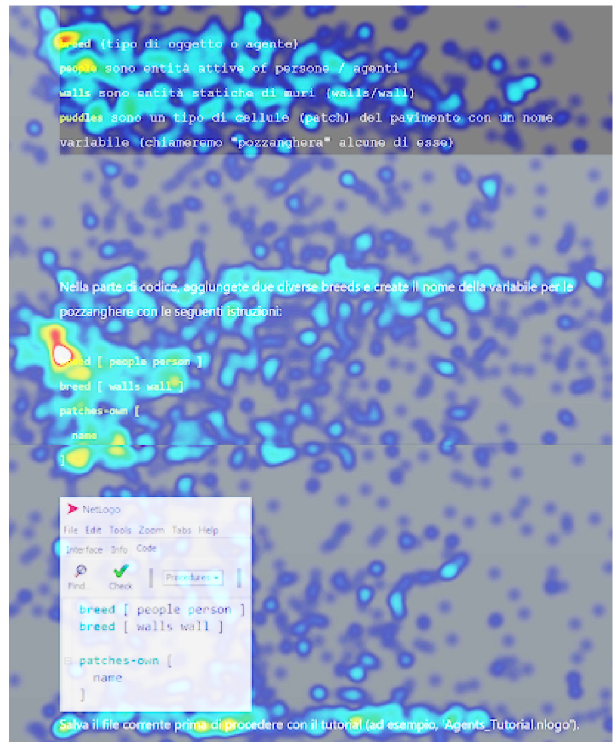

Fig. 3. An example of the distribution of mouse movements during the execution of the tutorial in a part of a web page. Red and yellow colors indicate a concentration of movements.

V. VISUALIZING RESULTS

A. *Visualize the density of movements*

Visualizing the results with heatmaps allows us to highlight the aggregate behavior of students performing the tutorial. Such maps represent well, in a very intuitive way, information about users' behavior on web pages. This kind of tool is advantageous for highlighting areas of greatest interest or criticality. Figure 3 depicts a portion of one page of the tutorial. Through such a visualization, it is possible to identify where attention was most focused during the course of the tutorial.

B. *Visualization of the flow*

The diagram containing the sequence of timed events affecting the student's process through the web page can be used to show the end user its progress. On the other hand, such a visualization allows an immediate tool for representing possible critical cases to be returned to the stakeholders involved in the analysis. A clear path of the process (eg, unidirectional, always forward path or few backward arcs), for example, can indicate a positive case of a learning pathway. Conversely, arcs going back many activities to a specific paragraph in the tutorial may indicate how the student did not understand that point and thus signal a criticality. Moreover, aggregate examination of diagrams can be done by distinguishing appropriate groups. One can think, for example, of examining and comparing the paths of two or more classes. Another analytical division can be made by distinguishing based on positive or negative evaluation of the tutorial, or by grouping those who answered text questions correctly between the tutorial pages and those who made many errors.

The DFG presented in Figure 4 describes the sequence of mouse clicks or movements over paragraphs as visited by students in their learning process across the web pages of the tutorial. Such a visualization makes it possible to appreciate the flow between the paragraphs of the 6 pages, named INTRO, INGRED, AGENTS, WORLD, BEHAV, SIMUL. Such diagrams allow immediately to identify whether some cases exit the learning process from initial or intermediate activities. Self-loops indicates the permanence over a page over 30 seconds. The number of iterations on the same paragraph indicates the most relevant ones, while backward arcs with greater weight indicate which passages were less clear and, therefore, probably required the user to go back in search for clarification.

VI. Discussion

Tracking limitations. This approach does have some tracking limitations. For example, a single user accessing at the same time from multiple browsers or devices will appear as multiple users. About cookies, a limitation is that some users can erase cookies after a session, maybe after re-opening browser. In this case, users cannot be tracked from one session to the next. In other cases, some users may have disabled tracking cookies. To overcome these limitations, user logins are suggested for future versions.

Improving representations. Heatmaps representation allows an immediate understanding of the most dense activity areas for the stakeholders involved in the project. About DFG, the diagram makes it immediately clear if there are any critical points, represented by the arcs going backwards, indicating the activities that were most critical, probably due to a flaw in the tutorial that needs to be improved or an intrinsic difficulty in the subject matter of that paragraph. Some improvements can be obtained by filtering the diagram of arcs with a few number of occurrences, in order to focus only on the most relevant paths.

Students' evaluation. The assessment of learning can be done either qualitatively or quantitatively. Qualitatively, the researchers involved in the project noted excellent engagement by the students, who were not familiar with the subject matter. There were no particular problems during the execution of the tutorial. In general, the feedback obtained at the end of the

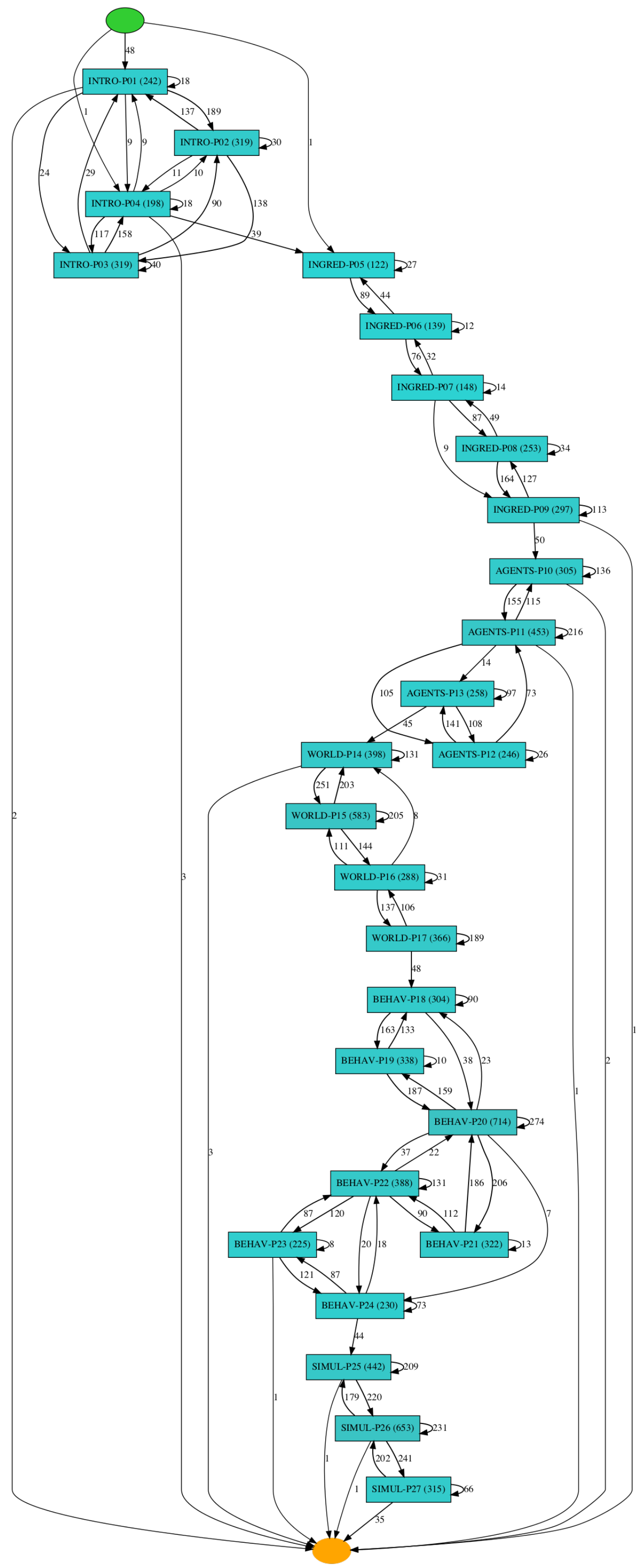

Fig. 4. A diagram of the learning process of student behavior as discovered by the heuristic miner algorithm. The weighted arcs indicate the movements or mouse clicks between the paragraphs of the tutorial in their six pages, i.e. INTRO, INGRED, AGENTS, WORLD, BEHAV, SIMUL.

tutorial, and both by observation during its execution, was satisfactory.

From a quantitative point of view, the results from the direct question asked at the end of the web tutorial, on a scale of 1 to 4, are encouraging. In particular, 52% expressed high satisfaction (the highest rate) and 43% said they were fairly satisfied. Only a small portion (5%), on the other hand, said they were dissatisfied.

VII. Conclusions and future work

This paper described an educational process mining example by using an instructional tutorial for approaching computer programming, analyzing students results and introducing aggregate analysis with a visualization effort. We demonstrate how process mining techniques, starting with the discovery of aggregate web behaviors, makes it possible to extract useful information for the improvement of the tutorial, as well as the understanding of the learning differences among students.

As a future work, we plan to add better analysis of behaviors investigating heat maps regarding click distribution or user attention. These visualizations can also provide useful insights into the strengths of the tutorial and improve future implementations of the tutorial accordingly.

In addition, the evaluation by users has been partly investigated but could be further explored. For instance, evaluation can be strengthened with a focus-group with experts and professors, or an online questionnaire. We also plan to add a focus-group with educators to collect meaningful insight about the proposed approach. Moreover, real data can be leveraged by platform developers to improve user experience accordingly. For example, if a certain amount of students leave the tutorial after a certain series of events, it indicates a pattern of dissatisfaction. Similarly, explicit negative comments can be explored correlating the different behaviour observed in the traces.

Another PM technique we aim to explore is the so called *conformance checking*, i.e. the comparison between the expected learning process and the real one from event logs. Similarly, several metrics allow for comparison groups of students, e.g. two different classrooms, as well as groups of students with similar outcomes. For instance, we expect to find different measures of fitness by separately examining the event log of the subgroup of students who answered the questions posed during the tutorial correctly or incorrectly, or among those who gave a positive or negative evaluation.

Acknowledgment

This work has been partially supported by *Laboratorio Informatica e Scuola* of the CINI Consortium.

References

[1] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 3, 2020.

[2] F. Sciarrone and M. Temperini, "Learning analytics models: A brief review," in *2019 23rd International Conference Information Visualisation (IV)*, 2019, pp. 287–291.

[3] M. W. Rodrigues, S. Isotani, and L. E. Zarate, "Educational data mining: A review of evaluation process in the e-learning," *Telematics Informatics*, vol. 35, no. 6, pp. 1701–1717, 2018. [Online]. Available: https://doi.org/10.1016/j.tele.2018.04.015

[4] D. S. Brogan, D. Basu, and V. K. Lohani, "Insights gained from tracking users' movements through a cyberlearning system's mediation interface," in *Online Engineering & Internet of Things - Proceedings of the 14th International Conference on Remote Engineering and Virtual Instrumentation REV 2017, held 15-17 March 2017, Columbia University, New York, USA*, M. E. Auer and D. G. Zutin, Eds., vol. 22. Springer, 2017, pp. 652–659. [Online]. Available: https://doi.org/10.1007/978-3-319-64352-6_61

[5] L. Omar and E. Zakaria, "Clustering methods applied to the tracking of user traces interacting with an e-learning system," in *Proceedings of World Academy of Science, Engineering and Technology*, vol. 68. Citeseer, 2012, pp. 894–894.

[6] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 205–240, 2021.

[7] E. Jensen, T. Umada, N. C. Hunkins, S. Hutt, A. C. Huggins-Manley, and S. K. D'Mello, "What you do predicts how you do: Prospectively modeling student quiz performance using activity features in an online learning environment," in *LAK'21: 11th International Learning Analytics and Knowledge Conference, Irvine, CA, USA, April 12-16, 2021*, M. Scheffel, N. Dowell, S. Joksimovic, and G. Siemens, Eds. ACM, 2021, pp. 121–131.

[8] F. Sciarrone, "Machine learning and learning analytics: Integrating data with learning," in *2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET)*. IEEE, 2018, pp. 1–5.

[9] L. Bioglio, S. Capecchi, F. Peiretti, D. Sayed, A. Torasso, and R. G. Pensa, "A social network simulation game to raise awareness of privacy among school children," *IEEE Trans. Learn. Technol.*, vol. 12, no. 4, pp. 456–469, 2019.

[10] P. Kumari, P. K. Jain, and R. Pamula, "An efficient use of ensemble methods to predict students academic performance," in *2018 4th International Conference on Recent Advances in Information Technology (RAIT)*. IEEE, 2018, pp. 1–6.

[11] S. Islam and H. Mahmud, "Integration of learning analytics into learner management system using machine learning," in *Proceedings of the 2020 2nd International Conference on Modern Educational Technology*, 2020, pp. 1–4.

[12] W. Van Der Aalst, *Process mining: data science in action*. Springer, 2016, vol. 2.

[13] M. A. Jayanthi and I. E. Shanthi, "Role of educational data mining in student learning processes with sentiment analysis: A survey," *Int. J. Knowl. Syst. Sci.*, vol. 11, no. 4, pp. 31–44, 2020. [Online]. Available: https://doi.org/10.4018/IJKSS.2020100103

[14] A. Bogarín, R. Cerezo, and C. Romero, "A survey on educational process mining," *WIREs Data Mining Knowl. Discov.*, vol. 8, no. 1, 2018. [Online]. Available: https://doi.org/10.1002/widm.1230

[15] E. Sulis, R. Pidello, and G. B. Demo, "Designing agent-based simulations in high schools: a project with netlogo," *DIDAMATiCA 2020*, vol. 17, p. 110, 2020.

[16] S. F. Railsback and V. Grimm, *Agent-based and individual-based modeling: a practical introduction*. Princeton university press, 2019.

[17] U. Wilensky and W. Rand, *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*. Mit Press, 2015.

[18] E. Sulis and M. Tambuscio, "Simulation of misinformation spreading processes in social networks: an application with netlogo," in *7th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2020, Sydney, Australia, October 6-9, 2020*, G. I. Webb, Z. Zhang, V. S. Tseng, G. Williams, M. Vlachos, and L. Cao, Eds. IEEE, 2020, pp. 614–618. [Online]. Available: https://doi.org/10.1109/DSAA49011.2020.00086

[19] J. Huang, R. White, and G. Buscher, "User see, user point: gaze and cursor alignment in web search," in *CHI Conference on Human Factors in Computing Systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012*, J. A. Konstan, E. H. Chi, and K. Höök, Eds. ACM, 2012, pp. 1341–1350.

[20] A. Pardo and G. Siemens, "Ethical and privacy principles for learning analytics," *Br. J. Educ. Technol.*, vol. 45, no. 3, pp. 438–450, 2014. [Online]. Available: https://doi.org/10.1111/bjet.12152