Behavior Data Collectionin in Collaborative Virtual Learning Environment

Tianqi Wu, Juanqiong Gou, Wenxin Mu, Zhe Wang

Wu T., Gou J., Mu W., Wang Z. (2021) Behavior Data Collection in Collaborative Virtual Learning Environments. In: Camarinha-Matos L.M., Boucher X., Afsarmanesh H. (eds) Smart and Sustainable Collaborative Networks 4.0. PRO-VE 2021. IFIP Advances in Information and Communication Technology, vol 629. Springer, Cham. https://doi.org/10.1007/978-3-030-85969-5_18

Behavior Data Collection in Collaborative Virtual Learning Environments

Tianqi Wu[1], Juanqiong Gou[1], Wenxin Mu[1] and Zhe Wang[2]

[1] School of Economic and Management, Beijing Jiaotong University,Beijing,China
{20120614, jqgou, wxmu}@bjtu.edu.cn
[2] Shanghai Siye Network Technology Co., Ltd,Shanghai,China
447255981@qq.com

Abstract. Educational Data Mining has gained a variety of attention. It describes students' cognitive needs through data mining, and provides individualized knowledge support for cognitive differences. Although the application of data mining algorithms is relatively mature, the data pre-processing based on data collection still suffers from high costs. The paper focus on a research question: how to effectively collect behavior data in virtual learning environments? "Effectively" in the sense of ensuring that value-intensive behavior data on decision-making can be accurately collected which reflects the students' cognitive. Therefore, the paper presents a method to achieve the object. The method comprises six steps, including extraction, transformation, determination, design, trigger and store. Based on the fact that all behavior data generated by the interaction is objective, identifying the collection points on the trigger event matches the granularity level of behavior data. Considering the related platforms and intelligent applications, the method can be used, providing behavior data support for the research of knowledge services.

Keywords: Behavior Data Collection, Virtual Learning Environments, Educational Data Mining

1 Introduction

There are two types of virtual learning environments available: one is the learning management system represented by MOOC, which is mainly in the form of online teachers teaching knowledge to students and students internalizing it through learning. The other type is the gamified learning platform represented by serious games, where achievement is mainly addressed the results of the game. The disadvantage of the former is that traditional teaching methods don't allow for personalized teaching, while the latter's disadvantage is that students focus more on the game itself than on knowledge. Therefore, a focus on both "learning" and "practicing" is necessary to achieve knowledge collaboration in a virtual learning environment, as shown in Fig. 1.

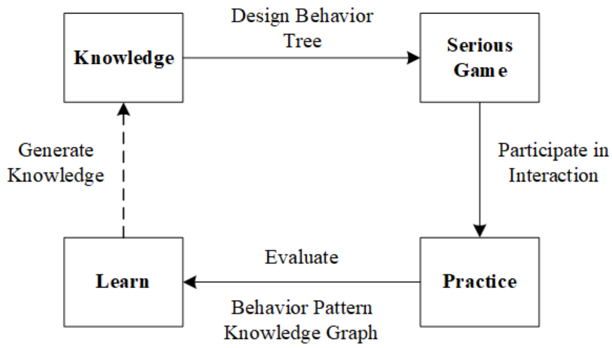

Fig. 1. Knowledge Collaboration in Virtual Learning Environments

Educational Data Mining (EDM) is a special area [1], accompanied by discovering valuable and potential information from the vast amount of data available in educational settings [2]. One of core objectives of EDM is to offer more personalized, interactive learning environments based on students' cognitive needs [1,3]. Although there has been much work on mining algorithms [4], substantial work has been done related with data collection [2]. Furthermore, data collection plays a vital role as it provides the foundation for EDM [2]. According to the background mentioned above, the research is conducted by a main question:

How to effectively collect student behavior data in virtual learning environments for behavior data mining?

Therefore, the paper needs to achieve the following objectives:

1) To reduce the cost of subsequent pre-processing by designing a structured format.
2) To ensure that value-intensive data can be collected accurately.

The structure of the paper is as follows: Section 1 introduces the main question of the paper. Section 2 presents an overview of behavior data collection. Section 3 describes a method of behavior data collection. In order to better illustrate the method, Section 4 describes a case study. Research prospects will be presented by Section 5.

2 Related Works

2.1 Data Collection and Data Pre-processing for EDM

In many cases, data collection is categorized as the first step in data pre-processing, which is the first step in the data mining process [5]. Fig.2 shows the main pre-processing steps with educational data [6]. It is not difficult to find that most studies define the data collection phase as the collection of raw data and separate out the stages of cleaning, identifying and filtering of raw data as other stages of data pre-processing. However, the data pre-processing stage essentially consumes 60-90% of the time, resources and efforts in the whole data mining process, remains a challenge that needs to be addressed [6].

Many researchers have proposed solutions to the high cost of data pre-processing, mainly by converting manual data pre-processing to automated data pre-processing [7,8,5], while others have argued for a standardized data format starting with data collection [2,3,6,8]. In this paper we focus on the data collection itself. [2] proposed a framework for collecting educational data based on data needs, but with a wide range of data sources and no guidance for data collection in virtual learning environments. [6] raised the need to collect data from multiple types of virtual learning environments, arguing that a collected data set with its own educational benchmark eliminates the need for pre-processing. [8] believed that one of the future research directions could focus on standardizing the format of data collected in virtual learning environments in order to shorten the most time-consuming pre-processing. In general, the solution of relying on educational data collection itself to improve the efficiency of data pre-processing is still at a preliminary stage, mainly because most objective data collected comes from system log files [7]. The raw data in these log files can’t be easily made to change and therefore can only be collected indiscriminately.

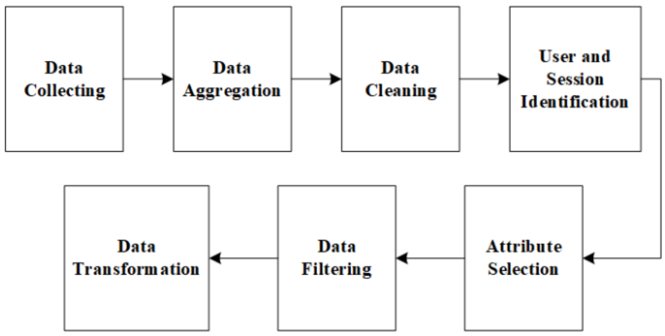

Fig. 2. The Main Pre-processing Steps with Educational Data [6]

2.3 Behavior Data Collection in Virtual Learning Environments

First, learning environments can store large amounts of data from multiple sources, such as interactions between students, teachers and virtual platforms, administrative data, statistics, student affectivity and so on [9]. So, it is important to determine the level of granularity at which behavior data is collected [10]. Fig. 3 depicts different granularity levels and their relationship to the amount of data, from the smallest (Events) to the largest (Courses) [9], which implies different collection frequencies and corresponding data set sizes [6]. The level of granularity chosen for the behavior data described in the paper is the smallest level. Therefore, the data collected can be characterized by whether event actions are captured, and the status of the event feedback [10].

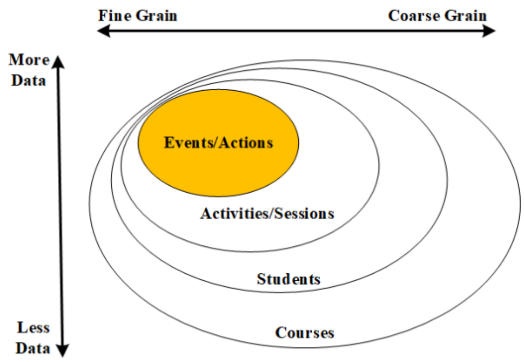

Fig. 3. Different Granularity Levels and Their Relationship to the Amount of Data [9]

Secondly, it is essential to understand how behavior data is currently collected in virtual learning environments. In the beginning, many case studies focused on questionnaires and continued to quantify their indicators in the hope of increasing accuracy [11,12]. However, as questionnaires are strongly subjective in nature [11], it was gradually replaced by log files represented by Moodle [13,14,8,7]. Log files are based entirely on student interactions with the platform [8]. The granularity of behavior data is also at the event level [6], but as mentioned in the previous section, it is difficult for researchers to make changes to log files in order to ensure the authenticity of raw data. So, the cost of subsequent data pre-processing is very high [5]and the existing solution is to automate pre-processing process [8,5,7]. In addition to these, some researchers have suggested that publicly available data sets are easier than collecting own behavior data [9], but only if these public data sets don't involve data ethics and privacy [6,9].

In summary, the "effective" collection of behavior data proposed in this paper involves two objects: *(I) Ensuring the granularity and objectivity of the behavior data itself. (II) Attempting to structure behavior data, reducing the cost of subsequent data pre-processing*.

3 Method of Behavior Data Collection

As shown in Fig.4, the method comprises six stages: extraction, transformation, determination, design, trigger and store, which are further explained below.

(Ⅰ) Extraction

The behavior of students interacting with the platform depends on the rules of the platform [15]. Therefore, the first stage is to extract the logic rules of the platform applied by the configuration platform. These rules are the core basis for the subsequent construction of behavior logic model. There are two methods of extraction, one is based directly on the documents provided by platform developers and the other is based on the whole interactions with the platform, generating rule documents.

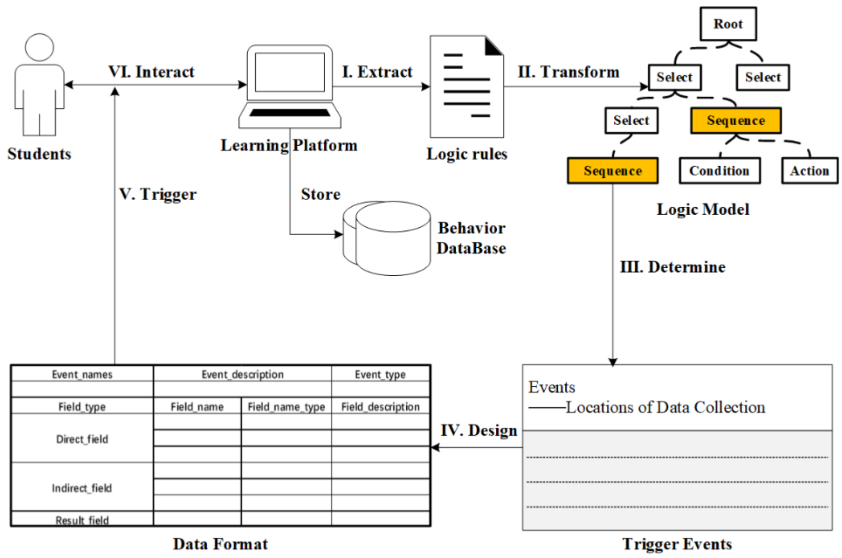

Fig. 4. A Method of Behavior Data Collection in Virtual Learning Environments

(Ⅱ) Transformation

Once we have the required logic rules, we need to transform them into the behavior logic model. In order to achieve the goal of logic clarity and visualization, the paper refers to Behavior Tree theory. Behavior Tree (BT) is a formal modeling language. Its concept was developed by Dromey in 2001 [16]. Behavior Tree has been widely applied on behavior decision-making of intelligent object NPC (Nonplayer Character) in interactive simulation applications [17]. The basic idea of Behavior Tree is to decompose individual behaviors into multiple levels. The nodes of logic model in Fig. 4. can be divided into the following types according to BT: **(1)** Select node (Select, SEL), a node that describes the interactive selection of students on the virtual learning platform. Its parent node is the root or a select node and its child nodes are select or sequential nodes. **(2)** Sequence node (Sequence, SEQ), a node that describes decision-making, usually containing a set of decision conditions and actions. Its parent node is a select node and its child nodes are a few condition nodes and action nodes. **(3)** Condition node (Condition, CON), is the leaf node to determine whether to execute or jump out of the sequence node. Its parent node is a sequence node. **(4)** Action node (Action, ACT), is a leaf node that represents the student's operational behavior. Its parent node is a sequence node.

(Ⅲ) Determination

The logic model enables modularity and independence of individual behaviors through transparent logic encapsulation. It can not only help researchers to identify the location of behavior data collection points, but also help platform operators to control the interaction logic more clearly. The decision points are a part of the student's behavior that occurs when interacting with the platform--"mouse click" [18].

From the perspective of decision making in collaborative networks, the logic model we constructed is in fact a decision model of students in collaborative virtual learning environments, as shown in Fig. 5. In Decision Model, the output data is determined from the input data, other sub-decisions and pre-defined business logic rules [19]. In the logic rule-based transformed logic model, we decouple students' decision points into multiple possible sub-decision points that correspond to different necessary conditions and actions resulting from the decision (Detail examples are shown in Table 2). A simple decision model diagram is shown in Fig.6, with each sub-decision point implying business knowledge.

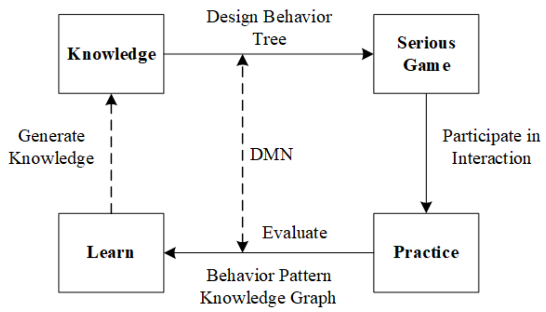

Fig. 5. Decision Model and Notation in Collaborative Virtual Learning Environments

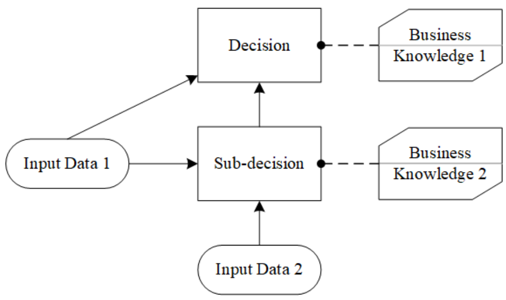

Fig. 6. Decision Model and Notation in Collaborative Networks [19]

Therefore, those sub-decision points are the part of behaviors that contains the most valuable behavior data, including information about students' cognitive and behavior patterns to be mined later. This precise data collection improves the efficiency of data pre-processing and ensures the quality of data compared to the full collection of raw data such as log files. So how to collect these behavior data? The first issue is to determine the location of behavior data collection points. In the logic model, the point at which a student makes a decision is represented by a sequential node. In addition, the leaf nodes in the logic model - the condition node and the action node - describe the trigger conditions and the corresponding actions that result when the event is triggered.

Therefore, all sequence nodes in the logic model are the behavior data collection points we need to determine. They need to be extracted and stored as trigger events.

(Ⅳ) **Design**

Once the behavior data collection points have been identified, the collection can usually begin. However, before this can be done, a structured data format needs to be developed for the behavior data in order to make it more "effective". This paper designs a structured data format based on event triggering, including event name, event description, event type, field type, field name, field data type and field description. Of these, the field types are divided into three categories: direct fields, indirect fields and result fields. We design direct fields as behavior data that can be captured directly based on the event trigger. Indirect fields can be collected indirectly through pre-defined procedures. Based on the above two types of fields, the system can output judgement results through the result fields and trigger corresponding actions.

(Ⅴ) **Trigger**

The location of behavior data collection points has been determined and the structured data format has been designed. The next stage is to trigger the event and collect the corresponding behavior data. Unlike the data collection of log files, the paper uses a front-end event monitoring mechanism, which is explained as follows: when an interaction decision is made between a student and the platform, the front-end event monitoring mechanism collects the relevant fields according to the structured data format. The collection is real-time. The mechanism will feed the trigger results (result fields, actions, etc.) back to the students.

(Ⅵ) **Store**

Finally, the platform stores collected data in a dedicated behavior database to facilitate subsequent data mining research. In addition, social data of students outside the platform, for example, can be correlated with the behavior data to analyze the impact of the environment, social and other factors on students' behaviors.

We try to make the data collection achieve two goals mentioned in Section 2. Based on the fact that all behavior data generated by the interaction is objective, identifying the collection points on the trigger event matches the granularity level of behavior data.

4 Case Study

A case study was conducted on a New Retail Business Simulation Platform to better illustrate the application of the method described in Section 3.

The New Retail Enterprise Simulation Platform is based on business simulation, game theory and other technologies, aiming at the needs of enterprise operation and management courses. Students are exposed to various situations such as purchasing, marketing and warehousing, they implement enterprise activities based on financial management, inventory management and some other multidimensional knowledge. Therefore, behavior data generated by interaction between students and the platform contains intensive individualized cognitive needs, and the main form of which is individual interaction decision-making. The purpose of the case study is to acquire these interactive behavior data for the related research of data mining such as individualized decision-making difference analysis.

1. Extraction

According to the documents provided by the platform developer, a comprehensive logic rule for the New Retail Enterprise Simulation Platform is shown in Fig. 7. This is only an overview, there are also specific logic rules for each part. These extracted logic rules are the core basis for the next stage in constructing a behavior logic model.

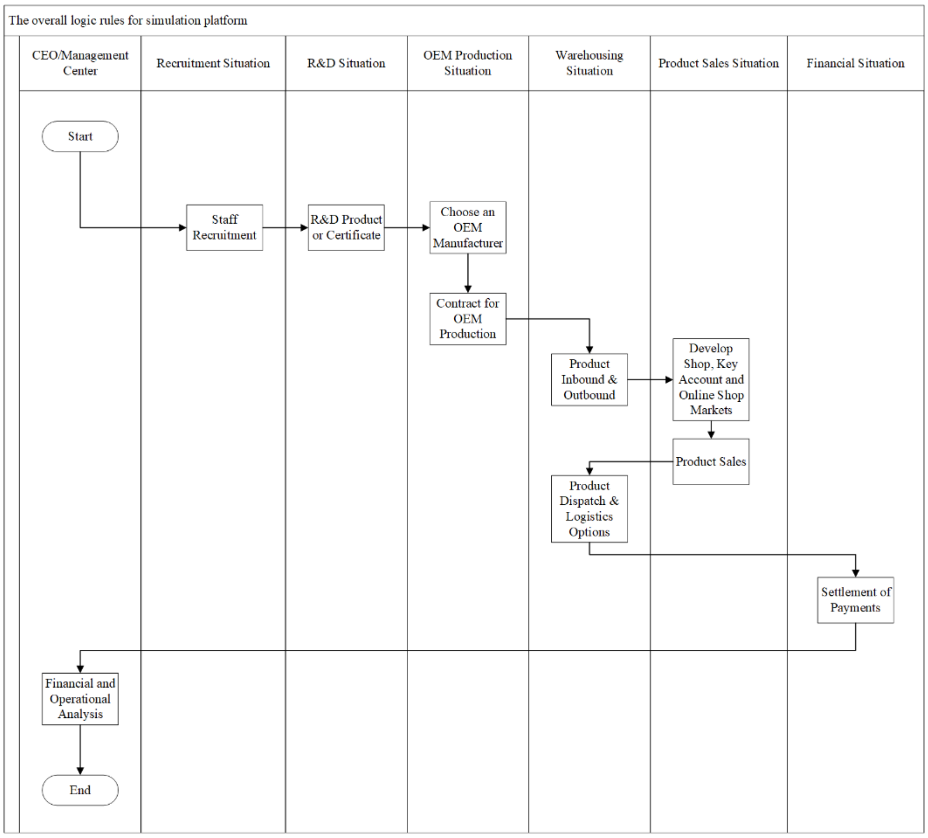

Fig. 7. The Overall Logic Rules for New Retail Enterprise Simulation Platform

2. Transformation

The logic rules are transformed into a behavior logic model as shown in Table 1 referring to Behavior Tree theory. Due to space limitations, only two of these situations are shown here: key account market and warehouse. The logic model consists of five types of nodes: select node, sequence node, behavior node, and condition node. The logic model provides a clear view of all decision points, and these are trigger events we will determine in the next stage.

Table 1. Part of the logic model for New Retail Enterprise Simulation Platform.

Key Account Market	SEL: market operation	SEQ: develop a new market	CON: development costs<=corporate funds	
			CON: have the certificate required: YES	
			ACT: develop a new market	
		SEL: R&D qualification certificate	SEQ: research and develop the certificate	CON: development costs<=corporate funds
				CON: have the certificate required: NO
				CON: number of R&D personnel >=2
				ACT: research and develop a certificate
			SEQ: recruit R&D personnel	CON: development costs<=corporate funds
				CON: have the certificate required: NO
				CON: number of R&D personnel<2
				ACT: recruit R&D personnel
	SEL: order operation	SEQ: win an order	ACT: select a developed market	
			CON: number of marketing specialists available>0	
			ACT: choose the order	
		SEQ: accept an order	ACT: obtain the information of the order	
			ACT: accept the order	
		SEL: wrap up an order	SEQ: complete the order	CON: complete the delivery: YES
				ACT: confirm completed
			SEQ: break a contract	CON: complete the delivery: NO
				ACT: break the contract
Warehouse	SEL: put in storage	SEQ: raw material inventory	ACT: select the warehousing order	
			CON: inventory required< remaining stock capacity	
			ACT: put the raw materials in storage	
		SEQ: commodity warehousing	ACT: select the incoming batch	
			CON: inventory required> remaining stock capacity	
			ACT: put merchandise in storage	
	SEQ: put out storage	ACT: select the outbound order		
		CON: procurement costs < corporate funds		
		ACT: expel merchandise from warehouse		

3. **Determination**

At this stage, we extract all sequence nodes from the logic model, some of which are shown in Table 2. This stage actually determines the location of behavior data collection points and they are also triggering events. These events contain the conditions that must be met for the trigger and all associated trigger actions. After storing the trigger events, we can design the corresponding data format for each event.

4. **Design, Trigger and Store**

The determination stage focuses on reducing the cost of data pre-processing through accurate the location of valuable behavior data. The design stage, on the other hand, is about designing a structured data format to improve the efficiency of collection and pre-processing. When we start data collection, we mainly use front-end monitoring mechanisms to collect behavior data in real time based on students' interactions with the platform. The platform stores the collected data to a back-end behavior database for behavior data mining. Taking the sequence node 'win an order' described in Table 2 as an example. Table 3 shows the structured format design of the data acquisition for the 'win an order' event. When a student makes a decision to win an order, the event 'Market Win Order' in Table 3 will be triggered to collect the relevant behavioral data. Among them, 'Direct Field' can be collected directly according to the order and the status of the enterprise. 'Indirect Field' can be obtained through the preset program of the system. Finally, according to 'Direct_Field' and 'Indirect Field', the system can calculate and output the judgement result through 'Result Field'.

Table 2. Part of The Locations of Data Collection Determined.

Trigger Events—Locations of Data Collection Points
SEQ: Develop a New Market
SEQ: Research and Develop the Certificate
SEQ: Recruit R&D Personnel
SEQ: Win an Order
SEQ: Accept an Order
SEQ: Complete the Order
SEQ: Break a Contract
SEQ: View Supplier Information
SEQ: Sign Purchase Order
SEQ: Cancel the Signed Order
SEQ: Raw Material Inventory
SEQ: Commodity Warehousing
SEQ: Put out Storage

Table 3. Structured Data Format Triggered by 'Win an order' Event.

Event_names	**Event_description**		**Event_type**
Market_WinOrder	The marketing specialist wins an order		Decision trigger
Field_type	**Field_name**	**Field_name_type**	**Field_description**
Direct_field	Order_MarketType	int	The ID of the market category in which the order belongs
	Order_Avil_Person	int	The number of key account specialists currently available
	Account	int	The cash in hand
	Time	int	The current number of months from the start of the game
	GoodinStock_All	int	Total current stock
	Order_GoodType	int	The cargo type ID of the order
	GoodinStock_Use	int	Current inventory occupancy
Indirect_field	Order_Have_Amount_1	int	The quantity of Order_GoodType to be delivered in the next month
	Order_Have_Amount_2	int	The quantity of Order_GoodType to be delivered in the next two months
Result_field	Expected_Performance	tinyint	1 is expected to perform, 0 is expected to default

This case describes the process of behavior data collection. Extraction is to extract logic rules of the New Retail Enterprise Simulation Platform. Transformation is to transform complex logic rules into a clear and transparent logic model. Determination is to initially reduce the cost of pre-processing through accurate collection points. Design is to further reduce the cost by designing a structured data format. Finally, students interact to trigger events for data collection. The data stored in behavior database can be exported in .TXT or .XLS file format for subsequent behavior data mining by researchers. At the same time, whether students have access to their behavior data source is a matter for discussion, but they can certainly review their behavior patterns through behavior data visualization.

5 Conclusion

To collect value-intensive behavior data that reflects cognitive characteristics of students, the paper proposed a method for data collection. The presented method consists of six processes: extraction, transformation, determination, design, trigger and store. Achieving knowledge collaboration in virtual learning environments is a core goal of behavior data mining. Future works will focus on evaluating behavior patterns of students through behavior data mining, enabling students to move from "practicing" to "learning" based on personalized knowledge support.

There are some noted limitations in the research. Due to complex logic rules, the transformation of behavior logic rules still requires artificial participation. Therefore, future research will focus on the automatic transformation of behavior logic models. Another noted limitation is that empirical studies on more types of simulation learning platforms are also needed to support the presented method. The situation in which data collected for the case study was in the individual mode. However, the students' individual behavior patterns will be subjected to decisions of others when they collaborate in team mode. Differences in behavior patterns arising from different collaboration modes are also the focus of future research.

Acknowledgments. The presented research works have been supported by "the National Natural Science Foundation of China" (61972029).

References

1. Aldowah, H., Al-Samarraie, H., Fauzy, W.M.: Educational data mining and learning analytics for 21st century higher education: A review and synthesis. Telematics and Informatics. 37, 13–49 (2019). https://doi.org/10.1016/j.tele.2019.01.007.
2. Mahanama, B., Mendis, W., Jayasooriya, A., Malaka, V., Thayasivam, U., Thayasivam, U.: Educational Data Mining: A Review on Data Collection Process. Presented at the September 1 (2018). https://doi.org/10.1109/ICTER.2018.8615532.
3. Romero, C., Ventura, S.: Educational Data Mining: A Review of the State of the Art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 40, 601–618 (2010). https://doi.org/10.1109/TSMCC.2010.2053532.
4. Romero, C., Ventura, S.: Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications. 33, 135–146 (2007). https://doi.org/10.1016/j.eswa.2006.04.005.
5. Feldman-Maggor, Y., Barhoom, S., Blonder, R., Tuvi-Arad, I.: Behind the scenes of educational data mining. Educ Inf Technol. 26, 1455–1470 (2021). https://doi.org/10.1007/s10639-020-10309-x.

6. Romero, C., Romero, J.R., Ventura, S.: A Survey on Pre-Processing Educational Data. In: Studies in Computational Intelligence. pp. 29–64 (2014). https://doi.org/10.1007/978-3-319-02738-8_2.
7. Kolekar, S.V., Pai, R.M., Manohara Pai, M.M.: XML Based Pre-processing and Analysis of Log Data in Adaptive E-Learning System: An Algorithmic Approach. In: Vincenti, G., Bucciero, A., and Vaz de Carvalho, C. (eds.) E-Learning, E-Education, and Online Training. pp. 135–143. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-28883-3_17.
8. Munk, M., Drlík, M., Benko, L., Reichel, J.: Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques. IEEE Access. 5, 8989–9004 (2017). https://doi.org/10.1109/ACCESS.2017.2706302.
9. Romero, C., Ventura, S.: Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery. 10, e1355 (2020). https://doi.org/10.1002/widm.1355.
10. Ihantola, P., Hellas, A., Butler, M., Börstler, J., Edwards, S., Isohanni, E., Korhonen, A., Petersen, A., Rivers, K., Rubio, M., Sheard, J., Skupas, B., Spacco, J., Szabo, C., Toll, D.: Educational Data Mining and Learning Analytics in Programming: Literature Review and Case Studies. Presented at the July 4 (2015). https://doi.org/10.1145/2858796.2858798.
11. Angeli, C., Howard, S.K., Ma, J., Yang, J., Kirschner, P.A.: Data mining in educational technology classroom research: Can it make a contribution? Comput. Educ. 113, 226–242 (2017). https://doi.org/10.1016/j.compedu.2017.05.021.
12. Munoz-Organero, M., Munoz-Merino, P.J., Kloos, C.D.: Student Behavior and Interaction Patterns With an LMS as Motivation Predictors in E-Learning Settings. IEEE Transactions on Education. 53, 463–470 (2010). https://doi.org/10.1109/TE.2009.2027433.
13. Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M.P., Núñez, J.C.: Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. Computers & Education. 96, 42–54 (2016). https://doi.org/10.1016/j.compedu.2016.02.006.
14. McCuaig, J., Baldwin, J.: Identifying Successful Learners from Interaction Behaviour. International Educational Data Mining Society (2012).
15. Domagk, S., Schwartz, R.N., Plass, J.L.: Interactivity in multimedia learning: An integrated model. Computers in Human Behavior. 26, 1024–1033 (2010). https://doi.org/10.1016/j.chb.2010.03.003.
16. Dromey, R.G.: From requirements to design: Formalizing the key steps. In: Cerone, A. and Lindsay, P. (eds.) First International Conference on Software Engineering and Formal Methods, Proceedings. pp. 2–11 (2003). https://doi.org/10.1109/SEFM.2003.1236202.
17. Subagyo, W.P., Nugroho, S.M.S., Sumpeno, S.: Simulation multi behavior NPCs in fire evacuation using emotional behavior tree. In: 2016 International Seminar on Application for Technology of Information and Communication (ISemantic). pp. 184–190 (2016). https://doi.org/10.1109/ISEMANTIC.2016.7873835.
18. Papamitsiou, Z., Economides, A.: Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. Educational Technology & Society. 17, 49–64 (2014).
19. Biard, T., Le Mauff, A., Bigand, M., Bourey, J.-P.: Separation of Decision Modeling from Business Process Modeling Using New "Decision Model and Notation" (DMN) for Automating Operational Decision-Making. In: Camarinha-Matos, L.M., Bénaben, F., and Picard, W. (eds.) Risks and Resilience of Collaborative Networks. pp. 489–496. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-24141-8_45.