A Review of Automated-Feedback Generation in Intelligent Tutoring Systems

Abstract

Educational technologies facilitate students' learning in both formal and informal contexts within various disciplines. **Intelligent tutoring systems (ITS)** have been developed to support students in improving their performance and practicing their skills. However, research on implementing and validating automated-feedback generation systems is lagging. This study reviews 11 recent articles on feedback in ITS environments and identifies current research trends. Results show that most studies: 1) were based on programming tasks, with a few exceptions using mathematical tasks; 2) used supervised-learning techniques, with only 3 studies employing reinforcement learning techniques to track students' learning status and generate hints or feedback; 3) detected students' knowledge mastery levels and provided useful feedback to improve students' performance through automated-feedback generation systems. This study systematically summarizes and compares the feedback-generating tools based on their characteristics, subject domains, and effectiveness. Findings from the review could inform future directions in designing and implementing automated-feedback generation systems.

Keywords: Intelligent Tutoring Systems (ITS); Feedback; Machine Learning Algorithms

Purpose

Background. Over the last decades, there has been a growing interest in the field of education to develop online learning platforms such as Massive Open Online Courses (MOOCs) and Intelligent Tutoring Systems (ITS) to make quality educational resources more accessible to students in and outside the classroom. The technology-enhanced learning platforms can effectively deliver asynchronous materials and administer both formative (i.e., assessment *for* learning) and summative (i.e., assessments *of* learning) assessments (Harlen & James, 1997).

Challenges. However, one of the main issues that hinders the popularization of the e-learning platforms is that they often fail to detect students' learning status (i.e., whether students' successfully mastered a skill; Lin & Chi, 2017; Mao et al., 2019) and make efficient interactions with learners or provide concrete feedback with respect to learners' performance on the tasks (Baker et al., 2012). Few studies have been conducted to implement and validate systems that can automatically generate feedback within the ITS. Although it is crucial in the process of knowledge acquisition, feedback has been understudied in relation to ITSs (Shute, 2008). Even less research has attempted to summarize and categorize the related empirical studies. Therefore, more research is needed on this topic.

Research Questions. The purpose of this literature review is to examine the current empirical research published within the last decade that applied machine learning algorithms to detect students' learning status and automated-feedback generation in ITSs. Therefore, this review is guided by the following research questions:

1. What are the target population, discipline, and platform used in the studies?
2. What machine learning algorithms are used in the studies to detect learning status and provide feedback?
3. Does feedback effectively improve participants' performance in the studies examined?

Theoretical Framework

Feedback. In education, feedback is defined as the information provided by an agent (e.g., teacher, peer, book, parent, self, or experience) regarding aspects of one's performance or understanding (Hattie & Timperley, 2007). A model of feedback was proposed (Hattie & Timperley, 2007) that conceptualized feedback as a tool that reduces discrepancies between current and desired performance and addresses four levels of feedback, including task, process, self-regulation, and self. The task level and the process level refer to how well the task is being performed; the self-regulation level describes how students monitor, direct, and regulate actions toward the learning goal; and the self level presents the affective evaluations about the student, which are usually not related to the task itself (Brophy, 1981).

Methods

Search Process. Since the topic of interest is relatively new, only a few studies have been published in the last decade. We adopted a snowballing approach that starts with the most recent related studies (Noy 2008). Both journal articles and conference proceedings were included in the search process.

Selecting Studies. Search results were refined by the following criteria for excluding articles for the following reasons: (1) not empirical and (2) not related to feedback generation.

Results

1. Eight out of eleven studies were conducted in the U.S., one in Brazil, one in the U.K., one in the Netherlands, and one in Australia. In most studies (n = 10), participants were undergraduate students from computing science courses. Only one study recruited crowd workers using Figure-Eight (a machine learning and artificial intelligence company based in San Francisco) in Australia. In terms of disciplines, nine studies provided feedback for programming tasks in platforms including iSnap (e.g., Mao et al., 2019; Marwan et al., 2019; Price et al., 2019), iList (Fossati et al., 2015), and Sleuth (Katan et al., 2020). Three studies provided feedback for mathematics in Pyrenees ITS (Lin & Chi, 2017) and Deep thought ITS (Ausin et al., 2019).

2. Seven studies used machine learning algorithms to analyze the data generated in the ITS to detect students' learning status and provide formative feedback when participants are using it. Among the seven studies, a variety of algorithms have been applied and compared including the Procedural Knowledge Model (PKM; Fossati et al., 2015), Support Vector Machine (SVM; Yaghoub-Zadeh-Fard et al., 2019), KNN (Mao et al., 2019;), deep learning models such as RNN and LSTM (Ausin et al., 2019; Lin & Chi, 2017; Mao et al., 2019), and reinforcement learning (Ausin et al., 2019; Lin & Chi, 2017). Three studies used a pre- and post-test quasi-experimental design to examine the effectiveness of the feedback provided by the ITS (e.g., Katan et al., 2020; Marwan et al., 2019). One study simply reported the usefulness of the feedback provision system (Keuning et al., 2014).

3. Among the eleven studies reviewed, eight studies reported that the automated-feedback generation systems could effectively track students' learning status, provide useful hints or feedback, and help improve students' performance on the tasks within the ITS. Three studies found their systems failed to detect students' errors and generate pertinent feedback. Of the studies that employed machine learning approach, the three studies that adopted reinforcement learning (RL) revealed that RL outperformed supervised-learning approaches such as KNN and RNN on detecting students' skill mastery profiles and generating feedback.

Conclusion and Educational Implications

This study reviewed the current trends and practices of automated feedback generation in ITS and identified several gaps in the literature. Most reviewed articles: 1) were based on programming tasks; 2) focused on ITS; 3) used supervised-learning techniques; and 4) revealed that the automated feedback generated could effectively improve participants' performance within the ITS. Practitioners may refer to the results of the present review when implementing automated-feedback generation systems.

References

Ausin, M. S., Azizsoltani, H., Barnes, T., & Chi, M. (2019, January). Leveraging deep reinforcement learning for pedagogical policy induction in an intelligent tutoring system. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (eds.). *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 168-177).

Brophy, J. (1981). On praising effectively. *The Elementary School Journal*, 81(5), 269-278.

Fossati, D., Di Eugenio, B., Ohlsson, S., Brown, C., & Chen, L. (2015). Data driven automatic feedback generation in the iList intelligent tutoring system. *Technology, Instruction, Cognition and Learning, 10*(1), 5-26.

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365-379.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Katan, S., & Anstead, E. (2020, April). Work in progress: Sleuth, a programming environment for testing gamification. In *IEEE Global Engineering Education Conference (EDUCON)*. Porto, Portugal [Conference or Workshop Item].

Keuning, H., Heeren, B., & Jeuring, J. (2014, November). Strategy-based feedback in a programming tutor. In *Proceedings of the Computer Science Education Research Conference* (pp. 43-54).

Lin, C., & Chi, M. (2017, June). A comparison of BKT, RNN, and LSTM for learning gain prediction. In *International Conference on Artificial Intelligence in Education* (pp. 536-539). Springer, Cham.

Mao, Y., Zhi, R., Khoshnevisan, F., Price, T. W., Barnes, T., & Chi, M. (2019, January). One minute is enough: Early prediction of student success and event-level difficulty during novice programming tasks. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (eds.). *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 119-128).

Marwan, S., Lytle, N., Williams, J. J., & Price, T. (2019, July). The impact of adding textual explanations to next-step hints in a novice programming environment. In *Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education* (pp. 520-526).

Noy, C. (2008). Sampling knowledge: The hermeneutics of snowball sampling in qualitative research. *International Journal of social research methodology*, *11*(4), 327-344.

Price, T. W., Dong, Y., Zhi, R., Paassen, B., Lytle, N., Cateté, V., & Barnes, T. (2019). A comparison of the quality of data-driven programming hint generation algorithms. *International Journal of Artificial Intelligence in Education, 29*(3), 368-395.

Yaghoub-Zadeh-Fard, M. A., Benatallah, B., Barukh, M. C., & Zamanirad, S. (2019, June). A study of incorrect paraphrases in crowdsourced user utterances. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 295-306).