Design of a User-Interpretable Math Quiz Recommender System for Japanese High School Students

Yiling Dai[1], Brendan Flanagan[1], Kyosuke Takami[1], and Hiroaki Ogata[1]

[1] *Academic Center for Computing and Media Studies, Kyoto University, Japan*

Abstract

In the context of K-12 math education, identifying the quizzes of the appropriate difficulty is essential to improve the students' understanding of math concepts. In this work, we propose a quiz recommender system which not only considers the difficulty of the quiz but also the expected learning outcome of solving that quiz. To increase the students' motivation of accepting the recommendations, our system provides interpretable information, i.e., the difficulty and expected learning gain of a recommendation, to the students. We conducted a pilot to implement this recommender system in a Japanese high school classroom. Overall, the log data showed a low rate of usage. We summarized challenges in implementing the recommender system in our specific setting, which help direct the future development of the system and its evaluation at a larger scale.

Keywords

Adaptive learning, K-12 math education, Quiz recommendation, User-interpretable

1. Introduction

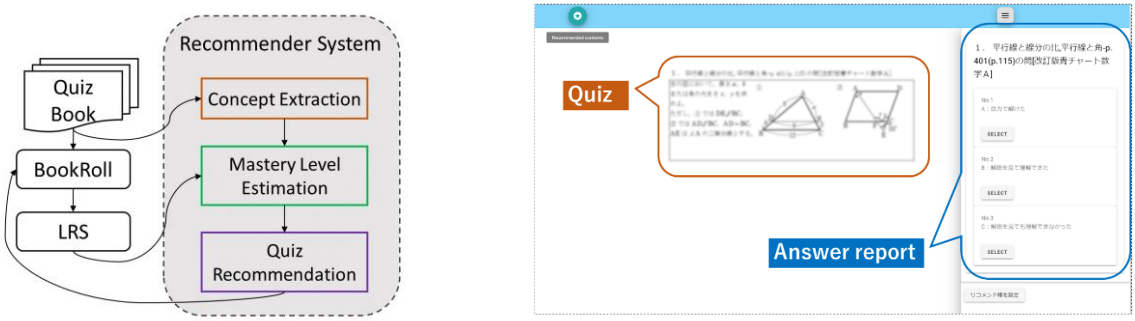
In the context of K-12 math education, identifying the quizzes of the appropriate difficulty is essential to improve the students' understanding of math concepts. Previous knowledge tracing works have focused on estimating the knowledge states of the students and predicting students' performance of the learning materials [2, 5, 7, 8, 11]. This is based on an assumption that learning happens when students attempt tasks in the zone of proximal development [13], i.e., the tasks they cannot achieve by themselves but can achieve with assistance. However, these works did not consider how knowledge states are improved and what improvement brought by the "proximal" learning materials should be prioritized. In this work, we propose a quiz recommender system which not only considers the difficulty of the quiz but also the expected learning outcome of solving that quiz. Among various learning outcomes, we focus on the average improvement of the understanding of related math concepts. By doing so, a quiz that helps the student to practice weaker concepts will be prioritized in the recommendation.

Recent adaptive learning systems have attempted to recommend learning materials based on complex methods such as deep learning methods [6] and reinforcement learning methods [12]. However, the mechanism and output of these methods are difficult to interpret, which may lead to the decrease of students' beliefs that they are able to do the task and their perceived values of completing the task, which further decreases their motivation to participate [14]. Being intuitive and simple, our recommender system is able to provide interpretable information, i.e., the difficulty and expected learning gain of a recommendation, to the students.

Some works proposed educational recommender systems with explanations for different subjects, different types of learning tasks, and different levels of education [1, 9]. Consequently, such recommender systems are context-specific and difficult to be applied in our case. We conducted a pilot to implement our proposed recommender system in a Japanese high school classroom. By analyzing the log and questionnaire data, we identified some challenges which help direct the future development of the system and its evaluation at a larger scale.

Proceedings of the 4th Workshop on Predicting Performance Based on the Analysis of Reading Behavior, March 21-22, 2022
https://sites.google.com/view/lak22datachallenge

2. System overview



(a) Recommender system and other modules. (b) Interface of BookRoll.

Figure 1: An overview of the learning system.

As Figure 1a shows, our recommender system is built on an integrated learning system where an e-book reading module called BookRoll [3] works to register and present the quizzes, and a learning record store (LRS) module stores the students' answers of the quizzes. The interface of viewing and answering the quizzes is shown in Figure 1b. In this work, we only record whether the student correctly answered the quiz for each attempt. In the recommendation module, we first extract the necessary concepts for the registered quizzes. Then, we estimate the students' mastery levels on the concepts based on their answers on the quizzes, which is further utilized to recommend quizzes that complement and extend their knowledge. In the following section, we describe the mechanism of the recommender system in detail.

3. Proposed recommender system

3.1. Problem definition

Suppose we have a student s and $|Q|$ quizzes. We model the student's attempts on the quizzes as $s_t = \{(q_1, r_1), \ldots, (q_t, r_t)\}$, where r_t is the student's response to q_t at step t. r equals 1 if the answer is correct and 0 otherwise. Our goal is to select and rank a subset of the quizzes that improve the student's knowledge acquisition. We denote it as $Recommend(Q|s_t) = (q_1, q_2, \ldots, q_n)$, where n is a predefined number of quizzes to recommend. In the following sections, we describe the procedures in detail with a running example showed in Table 1.

3.2. Procedure

3.2.1. Extracting underlying concepts.

Table 1.
Examples of quizzes.

Quiz	Content
q_1	**Proof** of the following **equality** of **triangles** $(b-c)sinA + (c-b)sinB + (a-b)sinC = 0$.
q_2	What is the shape of the **triangle** if the **equality** $asinA = bsinB$ holds?
q_3	Given $\triangle ABC$ and the **radius** of its **circumcircle** is R, what is b and $cosA$ when $a = 2, c = 4cosB, cosC = -1/3$?
q_4	In $\triangle ABC$, $\angle B = 60^\circ$, $AB + BC = 1$. M is the **midpoint** of BC. What is the length of BC such that the length of **line segment** AM is minimum.

Solving a math quiz requires knowledge of related concepts. For instance, to solve q_1 in Table 1, the students should know and be able to apply the knowledge of proof, law of sines, and etc. Understanding how students master the underlying concepts of the quizzes help us estimate their knowledge states more precisely, therefore, provide better remedial strategies.

Identifying the concepts required to solve the quizzes is not an easy task. Some of the concepts can be identified in the textual information of the quiz and its standard answer while some cannot. In this work, we take an initiative attempt to utilize the noun phrases as the concepts. For example, we deem proof, equality, and triangle as necessary concepts to solve q_1. We denote the underlying concepts of a quiz set as C. For each pair of q and c, $relatedness(q, c)$ denotes the relatedness between a quiz and a concept, which indicates the degree to which the knowledge of a concept is necessary in solving the quiz. For preprocessing, we use pdftotext[2] to extract the plain texts of quizzes from PDF format. Then, we adopt Janome[3] to parse the Japanese terms in the quiz text. Last, we use a classic term weighting method TFIDF [10] to compute $relatedness(q, c)$, which is implemented in scikit-learn[4].

3.2.2. Estimating mastery level on concepts.

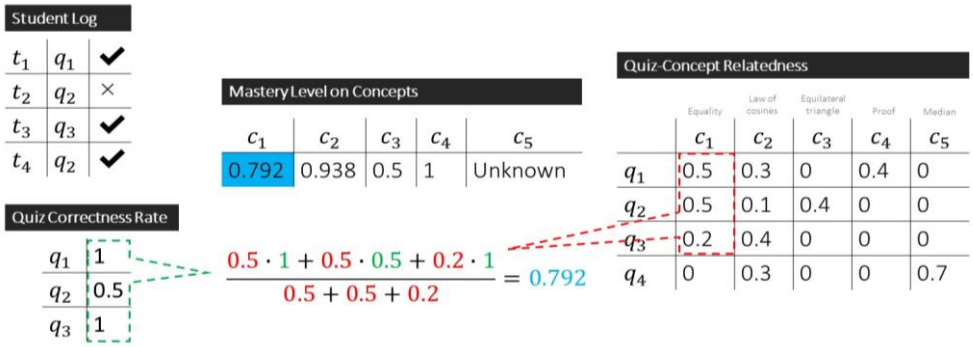


Figure 2: An illustration on computing mastery level on concepts.

After extracting the concepts and their relatedness to the quizzes, we estimate the students' mastery level on the concepts $mastery\_level(c|s_t)$ from their learning history in Equation (1).

$$mastery\_level(c|s_t) = \frac{\sum_{q \in Q_{s_t}} relatedness(q,c) \cdot correctness\_rate(q|s_t)}{\sum_{q \in Q_{s_t}} relatedness(q,c)} \quad (1a)$$

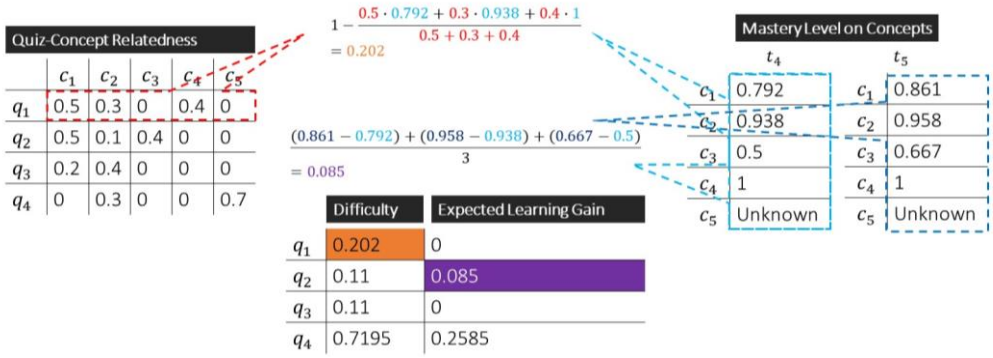
$$correctness\_rate(q|s_t) = \frac{s_t^{(q,1)}}{s_t^{(q)}} \quad (1b)$$

where:
Q_{s_t} = the set of quizzes in a student's attempts
$correctness\_rate(q|s_t)$ = the correctness rate of q in a student's attempts
$s_t^{(q,1)}$ = the correct attempts on q
$s_t^{(q)}$ = all the correct attempts on q

As illustrated in Figure 2, we compute the mastery level on each concept by looking at how the student answered the quizzes which require the knowledge of this concept. Suppose the student had attempted three quizzes q_1, q_2, and q_3, all of which require the knowledge of c_1. However, the student failed on q_2 at the first attempt and it is not counted in the computation of mastery level on c_1. Note that for the unseen quizzes, their requirement on the concepts is not considered in the computation. In other words, the mastery level only reflects the student's understanding of the concepts in the context of quizzes s/he had attempted.

[2] https://github.com/jalan/pdftotext
[3] https://mocobeta.github.io/janome/
[4] https://scikit-learn.org/stable/

3.2.3. Estimating quiz difficulty and expected learning gain.



	c_1	c_2	c_3	c_4	c_5
q_1	0.5	0.3	0	0.4	0
q_2	0.5	0.1	0.4	0	0
q_3	0.2	0.4	0	0	0
q_4	0	0.3	0	0	0.7

	t_4	t_5
c_1	0.792	0.861
c_2	0.938	0.958
c_3	0.5	0.667
c_4	1	1
c_5	Unknown	Unknown

	Difficulty	Expected Learning Gain
q_1	0.202	0
q_2	0.11	0.085
q_3	0.11	0
q_4	0.7195	0.2585

Figure 3: An illustration on computing quiz difficulty and expected learning gain.

To improve the acquisition of knowledge, the student needs to practice on quizzes that are of appropriate difficulty and provide learning gain at the same time. Therefore, we propose two criteria—quiz difficulty and expected learning gain— to decide which quizzes to recommend. Quiz difficulty reflects the probability that the student will give a wrong answer. As shown in Equation (2) and Figure 3, it is inferred from the student's mastery level on the concepts required in this quiz. Taking q_1 for example, the students' mastery level on c_1, c_2, and c_4 indicates how possible s/he is able to solve q_1. By subtracting it from 1, we obtain the probability that the student will fail to solve q_1, i.e., the difficulty of q_1. Note that for the unseen concepts, we set the mastery level as 0 to simplify the computation.

$$quiz\_difficulty(q|s_t) = 1 - \frac{\sum_{c \in C} mastery\_level(c|s_t) \cdot relatedness(q,c)}{\sum_{c \in C} relatedness(q,c)} \quad (2)$$

The expected learning gain of a quiz at step t is the average mastery level update on the concepts if the student successfully solves the quiz at the next step $t + 1$, as computed in Equation (3).

$$expected\_learning\_gain(q|s_t) = \frac{\sum_{c \in C_q} (mastery\_level(c|s_t \cup (q,1)) - mastery\_level(c|s_t))}{|C_q|} \quad (3)$$

where C_q is the set of concepts required in q. As illustrated in Figure 3, when computing the expected learning gain of q_2, we first update the student's mastery level on concepts assuming s/he successfully solved q_2. Then the total improvement on the mastery level is normalized by the number of concepts required in q_2.

3.2.4. Recommending quizzes.

We recommend quizzes based on the estimated difficulty and expected learning gain. Since both of quiz difficulty and expected learning gain are computed from students' mastery level of the concepts, they are intertwined with each other. Intuitively, a more difficult quiz may provide more learning gain if the student successfully solved it. Meanwhile, it requires the students to have more tolerance of uncertainty and to seek extra assistance in solving the quiz. Therefore, we adopt a recommending policy that the quiz is neither too difficult nor too easy to the students, yet provides learning gain as much as possible. As a result, the quiz to be recommended is not necessarily the one with the highest expected learning gain.

To do so, we select the quizzes in a proper range of difficulty and then rank them based on the expected learning gain in descendant order. Given a predefined number n of quizzes to recommend, the recommended set of quizzes and their order is $Recommend(Q|s_t) = (q_1, \ldots, q_n)$ such that for any $1 \le i \le n$, $\alpha \le quiz\_difficulty(q_i|s_t) \le \beta$, and for any $1 \le i \le j \le n$, $expected\_learning\_gain(q_i|s_t) \ge expected\_learning\_gain(q_j|s_t)$, where α and β are the

thresholds of difficulty. In the example of Figure 3, given the estimated difficulty and expected learning gain, we recommend q_2 if the acceptable difficulty is no greater than 0.3.

3.3. Strengths and limitations

Compared with previous works that recommend quizzes (or exercises in other contexts), our proposed recommender system has the following strengths:

- Our recommendations are user-interpretable.

As described in Section 3.2, every step in the recommender system is intuitive and simple, which makes it easy to provide explanations for the recommended quizzes. For example, we recommend q_2 to the student since s/he does not fully understand c_3 and by attempting q_2 s/he may deepen the understanding of c_1 and c_3. This type of information is difficult to extract in deep knowledge tracing methods [5–8] since the quizzes and knowledge states are usually embedded as vectors and processed in complex computation.

- Our model takes the future learning gain into consideration.

In a knowledge tracing model, the main goal is to model and trace the changes of knowledge states when students attempt quizzes. As a result, these models output the probability that a student can solve a quiz successfully while they do not consider whether and how the knowledge state improves by solving the quiz. In contrast, our model recommends quizzes taking the expected learning gain into account, which is an effort to optimize the learning outcomes.

- Our model does not rely on the data of other students.

Unlike data-driven methods which rely on a large dataset of student attempts on the quizzes, our model is content-based and is able to recommend quizzes merely using the data of one student. This is important when implemented in a real classroom setting that usually does not have sufficient student data on the quizzes.

- Our model is flexible and easy to modify.

Except for the two criteria we use in this work, our model is flexible to accept other criteria and modifications. For example, we can recommend a quiz which improves the understanding of the weakest concept instead of the overall understanding of the concepts. Or as in [1], we can set a target of concepts as a learning goal, then the quizzes supporting the understanding of these concepts will be prioritized.

Being intuitive and simple brings limitations too. For example, we do not consider the students' ability to apply their previous knowledge when solving a new quiz. Besides, we simply assume the students know nothing about unseen quizzes or concepts, which is not the case in real situations since the unseen concepts may be related to the known concepts. As described in Section 3.2.1, we focused on explicit noun phrases as the necessary concepts for solving a quiz. To obtain more underlying concepts, methods such as topic models can be integrated in our recommendation framework. However, interpreting the latent vectors could be another challenge. To address these limitations, considering the relationships between concepts and introducing parameters that model more complex learning behaviors are future works.

4. Pilot in high school classroom

Figure 4: The interface of the recommender system used in the pilot.

We conducted a pilot of the proposed recommender system in a real classroom setting as preparation for a well-designed comparative experiment. In the pilot, one class of students of a Japanese high school was invited to use the recommender system to solve quizzes during summer vacation from July 20th to August 23rd, 2021. The teacher had assigned 54 quizzes and asked the students to finish the quizzes and check the answers by themselves. Note that the students were highly recommended to solve the quizzes and report their answers in the learning system. However, they were not required to do so since they also can choose to solve quizzes in paper-based textbooks. In the BookRoll system, the students can access the quizzes from a book directory, a list of quiz assignment, or from the recommender system tab. While the students started to attempt the quizzes in the system, we recommended the assigned quizzes accordingly. We set the acceptable quiz difficulty in the range of [0.1, 0.6]. Figure 4 shows the interface of the recommender system where the estimated mastery level and the reasons why a quiz is recommended were shown to the students. During the summer vacation, the log data of quiz answers and recommendation clicks were collected[5]. After the summer vacation, we conducted a questionnaire survey on the students' perceptions of the recommender system.

4.1. Students' reactions

Table 2.
Statistics of the usage of the recommender system.

	Number of Students
Total students	38
Accessed the learning system	22
Clicked the recommended quiz	2
Answered the recommended quiz	1

Table 2 shows the statistics of the system usage during the pilot period. 57.9% of the students accessed the learning system and only two students ever clicked the recommended quizzes. To further investigate the reason of the usage of the recommender system, we analyzed the questionnaire survey. The questionnaire contains 42 questions regarding the students' perceptions of the recommender system and attitudes towards math learning. Among the questions, 8 are descriptive and 34 are 5-likert scale questions, which were found to have good reliability ($\alpha = 0.836$). 30 students answered the questionnaire and 3 incomplete answers were excluded in the following analysis. We separated the students into two groups based on whether they reported that they ever used the recommender system. The students whose answers were equal to or greater than 3 are treated as self-reported users ($n = 6$), while the rest of the students are treated as self-reported non-users ($n = 21$). Note that the following analysis is based on an assumption that we "trust" the self-reported results. We then conducted 2-tailed t-test on the answers of the other questions between these two groups. Table 3 lists the questions that had a significant difference ($p < 0.05$) between the two groups. We found that the self-reported users

[5] The log data of quiz answers during August 12th and August 18th was not recorded due to a system bug.

demonstrated more positive attitudes towards the usefulness of the recommender system, more trust towards the explanations and the recommender system, and more motivation to do math quizzes. However, we are aware of the limitations such as the small sample size and the low reliability of self-reported results.

Table 3.
Analysis on the questionnaire survey.

	Self-Reported User ($n = 6$)		Self-Reported Non-User ($n = 21$)		
	mean	sd	mean	sd	t
Perception of the recommender system					
I used the recommender system because it was easy to use.	3.33	0.82	2	0.95	3.40***
The recommender system was useful.	3.33	0.82	2.38	0.80	2.53*
I wanted to use the recommender system more.	3.17	0.41	2.33	0.80	3.46***
I trusted the recommender system.	3.83	0.75	2.71	0.85	3.12**
I used the recommender system because the quizzes fitted to me.	3.67	1.03	2	1	3.51***
I understood why I need to do the recommended quizzes.	4	0.63	2	1.05	5.80***
I trusted the explanations of why the quizzes were recommended.	3.67	0.52	2.29	1.01	4.53***
Attitude towards math learning					
I became better at math because of the recommender system.	3	0.89	1.76	0.70	3.12**
I enjoyed learning math more because of the recommender system.	3.33	1.21	1.95	1.02	2.55*
The system pushed me to do math quizzes.	3.67	1.03	2.14	1.11	3.13**

4.2. Challenges

Based on the results of the pilot, we found the following challenges to implement a recommender system in our specific classroom setting:

- The inertia of the students to do quizzes in paper-based methods.

Doing math quizzes and self-checking the answers require a large visual space to write down the answer and then compare it with the standard answer. The current e-book reading system and the device could not fully support this usage, which leads to extra work for the students to report their answers in the system after solving them in the paper version. As some of the students reported in the questionnaire, it was annoying to find the same quiz in digital-version and the paper-version. To encourage more usage in the digital version, we need to improve the convenience of the e-book reading system so that the students can easily switch between both versions.

- The compatibility with the standard teaching schedule.

As a request from the teacher, we limited the recommendations to an assigned set of quizzes in the pilot. This reduces the meaning to use the recommender system as the students had to finish all the quizzes at the end. However, there does exist a concern that if the recommended quizzes for each student are distributed over diverse topics, the teacher may fail to keep a standard teaching schedule

and to give feedback instantly. In the future, we need to balance the standard teaching schedule and the personalized learning.

- The definition of the usage of the recommender system.

In our recorded log data, only two students clicked the recommended quizzes. However, 6 students reported they ever used the recommender system. One of the two students who clicked the recommended quizzes reported s/he never used the recommender system. Obviously, there is a gap between our and the students' perceptions of "usage". Besides, the students may provide unreal answers both intentionally and unintentionally. As Fredricks et.al [4] suggest, engagement is a multifaceted construct that relates to behavior, emotion, and cognition. In future work, we plan to record more students' engagement with the system in a more objective manner.

5. Conclusions and future works

In this work, we proposed a quiz recommender system which not only considers the difficulty of the quiz but also the expected learning outcome of solving that quiz. To increase the students' motivation of accepting the recommendations, our system provided interpretable information, i.e., the difficulty and expected learning gain of a recommendation, to the students. Our system has advantages over existing methods as a) the recommendations are user-interpretable, b) the future learning gain is considered, c) it does not rely on large sets of student data, and d) it is flexible to modifications.

We also conducted a pilot to implement this recommender system in a Japanese high school classroom. The log data demonstrated a low usage of the system during the pilot. However, we did find some positive signals between the attitudes toward the system and the self-reported usage of it from the limited questionnaire data. More importantly, we identified some challenges such as the inconvenience of the current e-book reading system, the incompatibility with the standard teaching schedule, and the gap between the intended usage of the system and the students' perceptions of it. In future work, we plan to redesign the evaluation experiment with more clear instructions and more thorough and objective ways to record students' engagement and learning performance improvement.

6. Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, JSPS Grant-in-Aid for Scientific Research (S) 16H06304 and NEDO JPNP20006 and JPNP18013.

7. References

[1] Jordan Barria-Pineda, Kamil Akhuseyinoglu, Stefan Želem-Ćelap, Peter Brusilovsky, Aleksandra Klasnja Milicevic, and Mirjana Ivanovic. 2021. Explainable Recommendations in a Personalized Programming Practice System. In Artificial Intelligence in Education. 64–76.

[2] Albert T. Corbett and John R. Anderson. 1994. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction 4, 4 (1994), 253–278.

[3] Brendan Flanagan and Hiroaki Ogata. 2018. Learning Analytics Platform in Higher Education in Japan. Knowledge Management & E-Learning 10, 4 (2018), 469–484.

[4] Jennifer A. Fredricks, Phyllis C. Blumenfeld, and Alison H. Paris. 2004. School Engagement: Potential of the Concept, State of the Evidence. Review of Educational Research 74, 1 (2004), 59–109.

[5] Tao Huang, Mengyi Liang, Huali Yang, Zhi Li, Tao Yu, and Shengze Hu. 2021. Context-Aware Knowledge Tracing Integrated With the Exercise Representation and Association in Mathematics. In Proceedings of the 14th International Conference on Educational Data Mining.

[6] Zhenya Huang, Qi Liu, Chengxiang Zhai, Yu Yin, Enhong Chen, Weibo Gao, and Guoping Hu. 2019. Exploring Multi-Objective Exercise Recommendations in Online Education Systems. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management.1261–1270.

[7] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2021. EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction. IEEE Transactions on Knowledge & Data Engineering 33, 01 (2021), 100–115.
[8] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Graph-Based Knowledge Tracing: Modeling Student Proficiency Using Graph Neural Network. In Proceedings of 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI). 156–163.
[9] Behnam Rahdari, Peter Brusilovsky, Khushboo Thaker, and Jordan Barria-Pineda. 2020. Using Knowledge Graph for Explainable Recommendation of External Content in Electronic Textbooks. In Proceedings of the Second International Workshop on Intelligent Textbooks 2020 co-located with 21st International Conference on Artificial Intelligence in Education (AIED 2020), Vol. 2674. 50–61.
[10] Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management 24, 5 (1988), 513–523.
[11] Xia Sun, Xu Zhao, Bo Li, Yuan Ma, Richard Sutcliffe, and Jun Feng. 2021. Dynamic Key-Value Memory Networks With Rich Features for Knowledge Tracing. IEEE Transactions on Cybernetics (2021), 1–7.
[12] Xueying Tang, Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. 2019. A Reinforcement Learning Approach to Personalized Learning Recommendation Systems. Brit. J. Math. Statist. Psych. 72, 1 (2019), 108–135.
[13] Lev S. Vygotsky. 1978. Interaction between Learning and Development. 79–91.
[14] Allan Wigfield and Jacquelynne S. Eccles. 2000. Expectancy-Value Theory of Achievement Motivation. Contemporary Educational Psychology 25, 1 (2000), 68–81.