Generating Learning Sequences Using Contextual Bandit Algorithms

Le Minh Duc Nguyen[1], Fuhua Lin[2][0000-0002-5876-093X], and Maiga Chang[3]

[1]Athabasca University, 1 University Dr., Athabasca, T9S 3A3, Canada
lnguyen12@learn.athabascau.ca
[2]Athabasca University, 1 University Dr., Athabasca, T9S 3A3, Canada
oscarl@athabascau.ca
[3]Athabasca University, 1 University Dr., Athabasca, T9S 3A3, Canada
magaic@athabascau.ca

Abstract. Personalized learning paths have become a promising instructional strategy in online learning, as they can cater to individual learners' needs and preferences. However, creating effective personalized learning paths is a complex task due to the high degree of variability in learners' characteristics, behaviors, and learning contexts. Existing recommendation methods do not adequately address this challenge, as they do not work effectively in dynamic environments. This paper tries to address this gap by proposing a personalized learning path recommendation system using a contextual multi-armed bandit approach to offer a student an optimal learning sequence and provide the student with a modified sequence when re-planning is required.

Keywords: Multi-Armed bandit (MAB) algorithms, knowledge components (KC), adaptive learning, exploration and exploitation, personalized learning.

1 Introduction

The effectiveness of traditional one-size-fits-all approaches to course designing has been a subject of debate due to their limited ability to address the diverse needs and interests of learners. However, the advent of personalized learning has introduced a transformative paradigm that can tailors instruction to match the unique characteristics of each learner. This approach acknowledges the inherent differences among learners, including their backgrounds, learning strategies, and preferences, emphasizing the importance of providing personalized learning paths to help them achieve their learning goals more efficiently. Personalized learning involves the customization of learning trajectories, which consist of carefully selected sequences of learning activities and resources, designed to facilitate learners in attaining their specific educational goals [1].These tailored trajectories act as individualized roadmaps, guiding learners through a set of activities that have been specifically adapted to address their distinct needs and aspirations. Achieving this level of customization in learning trajectories requires an in-depth understanding of learners' characteristics, encompassing their prior

knowledge, areas of interest, and preferred modalities of learning, in order to effectively improve the learning experience and optimize learning outcomes[2].
Our research proposes a contextual multi-armed bandit (MAB) approach for personalized learning path recommendation in the domain of online education. The contextual bandits approach, also referred to as associative Reinforcement Learning[3], is an iterative process. An agent at every time step receives a context vector generated by the environment and selects an option from the set of choices (which are referred to as "arm"). Each selected arm is associated with a stochastic reward that the environment reveals to the agent. The primary objective is for the agent to optimize its acquired rewards over the long term by leveraging the historical data of its previous actions.
We structure our paper into the following sections: Related Work (Section 2 to discuss recent related research of personalized learning), Research Problem (Section 3 to define a formalized research problem), Methodology (Section 4 to describe the methods in our research), Proposed Algorithm (Section 5 to propose algorithms), Experiment &Simulation (Section 6), and Conclusion (Section 7).

2 Related work

The field of personalized learning paths has been extensively researched in recent years. Various approaches have been proposed to personalize learning paths based on individual learners' characteristics, behaviors, and learning contexts. According to [4], those approaches can be categorized into two main types: Course Generation and Course Sequence. Course Generation (CG) approaches involve generating and recommending the entire learning path to a user in a single recommendation. In this approach, the user is presented with a complete set of learning content and activities to follow to learn a course. The evaluation of the learning path effectiveness occurs only after the completion of the entire path, rather than at each step along the way. In CG approaches, Shi et al[5] proposed a graph traversal algorithm in their paper to generate all paths considering the students' learning objectives and learning need and recommending the one with the highest score. Niknam et al. [6] proposed Ant Colony Optimization algorithm, combined with Fuzzy C-Mean Cluster algorithm to select a path for a cluster of learners based on their prior knowledge. While Course Generation methods are commonly used by researchers to generate personalized learning paths, they are associated with several limitations. One of the major drawbacks of this approach is that it often fails to account for changes that may occur during the learning process, such as a user's evolving skills, interests, or preferences. Consequently, learners may be at risk of receiving an inappropriate or unmanageable learning path, leading to inefficiencies or disengagement. Another challenge is the potential for CG methods to become slow when presented with a large amount of data, such as a high volume of learning objects or user profiles. This sluggishness can negatively impact the user experience, as it may take too long to generate a personalized learning path or respond to the learner's needs in a timely manner. In contrast to the CG approach, Course Sequence (CS) methods suggest personalized learning paths to users one step at a time, considering their current progress and performance. This method allows for ongoing evaluation and adaptation of the learning path, ensuring that users are not overwhelmed by information and can

focus on mastering one concept at a time. By dynamically adjusting the learning path as a user progresses, CS methods are better able to accommodate changes in users' performance and adapt to their unique learning needs. Xu et al. [7] proposed Naïve Bayes algorithm, combined with KNN to recommend an optimized learning objective to a student. Cai et al. [8] proposed a Reinforcement Learning based method, combined with Neural Network and Knowledge Tracing Model (KTM) to recommend the most suitable learning path based on the specific knowledge points and the individual learner's needs throughout the entire learning journey. The study of Cai et al. (2019), however, posts some limitations [8]. It does not address dependencies and constraints among learning modules. In addition, the proposed method is not efficient in online learning, as the reinforcement-learning model can only be trained after the training of the knowledge tracing model is completed. Lastly, there is no handling of altering or adjusting the recommended learning path once required.

Although the field of learning path personalization has seen significant attention from researchers, a number of challenges and limitations still persist. First and foremost, it is essential to consider the learners' time constraints when designing personalized learning paths. Time is a valuable resource for learners, and an effective learning path personalization method should consider learners' schedules to optimize their learning outcomes. Another challenge in learning path personalization is scalability. Designing methods that can handle large-scale datasets is a complex problem that has only been addressed by a few studies in the literature. Scalability is critical to the success of personalized learning, as the method must be able to efficiently process and respond to large amounts of data to maintain learners' engagement. In addition, learners' profiles should be regularly updated to reflect changes in their responses and learning progress. This requires the learning path personalization method to adapt to the learners' changing needs and provide them with the most appropriate learning content. Evaluation is also a significant challenge in learning path personalization methods. The lack of a general evaluation framework makes it difficult to compare different methods, and a reliable evaluation framework should include guidelines for data sources and principles to ensure a consistent and accurate evaluation. The ability to update the learning sequence actively and dynamically, when one or more student’s surrounding factors alter is another challenge. Finally, recommender systems play a crucial role in adaptive learning by predicting student preferences. However, they face the exploration-exploitation dilemma when making recommendations. They must balance exploiting their knowledge about the content chosen by previous students with exploring new materials that may be better suited to the current student's needs.

Reinforcement learning (RL) has been utilized to create effective and adaptable pedagogical policies. Recently, there has been a growing interest in the use of MAB algorithms for adaptive learning. MAB algorithms fall under the broader category of RL and are named after the problem faced by a gambler who must decide which arm of a K-slot machine to pull in order to maximize their total reward in a series of trials. These algorithms can navigate exploration-exploitation trade-offs and make sequential decisions under uncertain conditions. They have been employed in real-world applications

to solve optimization problems, such as experimental design and website optimization. As MAB algorithms actively select which data to receive and analyze in real-time, they lend themselves naturally to the problem of eliciting adaptive sequences of content and assessment in adaptive learning environments[9]. The multi-armed bandit framework has the potential to address the challenge faced by many applications when no prior information is available, especially for large-scale recommender systems. Their continuous exploration approach can also help address the cold start problem in recommender systems.

Although some initial, isolated, or purely theoretical research has been conducted on using MAB to elicit sequences for adaptive learning, there are still many questions that need to be answered. For example, due to the complexity of adaptive learning, standard MAB models cannot be directly applied. When an MAB-based adaptive engine makes sequential decisions to optimize learning, how does it define rewards? Which metrics should be optimized? Which algorithm of the MAB family in what parameter settings, would be best for a particular sequencing problem in adaptive online learning?

To our knowledge, none of the proposed algorithms have adequately addressed the challenges associated with the learning sequence recommendation problem. Another consideration is the contextual bandit algorithm, which is an extension of the MAB approach used to discover which actions are effective in specific contexts. Xu et al. [10] used a contextual bandit approach to recommend entire sequences of courses within a program, rather than sequencing knowledge components in an online course. In their work, courses were planning elements with fixed completion times. The work did not address knowledge component sequencing within individual courses. Additionally, after recommending a whole sequence to complete a degree, the sequence could not be altered. We recognize that one of the drawbacks of recommending an entire KC learning path to a learner is that it ignores the learner's actualized learning performance and the context changes that occur during the learning process. As a result, the learner may waste time by receiving a path that may not be optimal.

In the next section, we will formally define the research problem, followed by our proposed methodology and algorithm.

3 Research Problems

MAB algorithm is a subset of RL algorithm, and contextual MAB (CMAB) is an extension of the MAB approach where environment contexts are factored in. MAB (or CMAB) approach is different from the general RL approach in that MAB is for solving a stateless Markov Decision Process (MDP) problem, where all the states (observations) are independent, and the agent gets a reward immediately after choosing an action. In order to fit the adaptive sequence learning problem into CMAB approach, we have the option to follow the CG approach (i.e... generate entire learning path). However, due to the disadvantage of CG approach, we give favors to the CS approach. It is a challenge to creatively model our problem to fit CMAB in the CS approach. Another challenge is that even when our problem is well modelled, the standard CMAB

approach does not well support re-planning (i.e., the current recommended learning module failed to fit a student's current competency, and a new different learning module should be recommended). For this challenge, Wacharanwan et al. in their research [11] proposed applying correlation analysis in CMAB approach. It recommends the closest learning path having the highest rank of correlation measurement. However, the approach is not effective as the learning problem is non-stationary [11]. Before we discuss further, we first consider how we are going to model our learning problem.

There are many ways to model a modular content hierarchy. Duval and Hodgins [12] introduced the content hierarchy consisting of 5 levels: Course, Lesson, Learning Object, Raw Content, and Information Object. Nabizadeh et al in their research paper [4] suggested another level named Topic. Referencing these articles, we modeled our knowledge domain into two levels: Course and Knowledge Component (KC). A course is composed of a few learning units, each learning unit covers a concept. A KC is a learning unit, and a Course is composed by few KCs. Some KCs are the prerequisites in order to learn other KCs. Some KCs can be exchangeable or other KCs.

4 Methodologies

We can model a course domain as a KC AND-OR graph where each node in the graph presents a knowledge component. A KC AND-OR graph represents the search space for solving the problem in "Goal". We can define the root node "Goal" as the initial problem and every other node is a sub-problem.

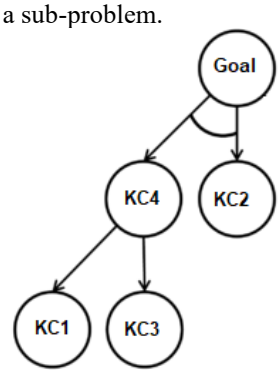

Fig. 1A KC AND-OR graph sample.

In Figure 1, we use goal-reduction methods to break down the graph:

$$\boldsymbol{GOAL} if \boldsymbol{KC4} and \boldsymbol{KC2}$$
$$\boldsymbol{KC4} if \boldsymbol{KC1} or \boldsymbol{KC3}$$

The KCs are modeled as a knowledge structure in Knowledge Space Theory (KST) [13]. To this end, the domain experts first identify the set of KCs for a course. After that, the domain experts specify the prerequisite relations among the KCs.

A learning path is also called a policy or a learning sequence. Formally, the set of KCs in a course is denoted as $list(KC) = \{\mathrm{KC}_1, \mathrm{KC}_2, \ldots, \mathrm{KC}_n\}$, where n is the total number

of the KCs of the course. For example in Figure 1, the possible learning sequences are $\{\{KC_1 \rightarrow KC_4 \rightarrow KC_2\}, \{KC_2 \rightarrow KC_1 \rightarrow KC_4\}, \{KC_3 \rightarrow KC_4 \rightarrow KC_2\}, \{KC_2 \rightarrow KC_3 \rightarrow KC_4\}\}$. The contextual multi-armed bandit algorithm is an extension of the classical multi-armed bandit algorithm[14] that considers the context in which the decisions are made. In the CMAB problem, besides the observed rewards, the agent also considers the additional information received about the context to make the decision. The context can be thought of as a set of features that describe the state of the system, and the objective is to learn a policy that maps the context to the best action to take.

5. The Proposed Algorithms

We propose a contextual multi-armed bandit algorithm for recommending the next KC for learners to learn based on two types of context information as follows.

- The similar profile, characteristics, and backgrounds of past learners
- The current situation of learners in the course. Current situation of learners can be seen as the current completion status of the KCs

Let's denote f_g^k to be k global features that are included in the first type of context. $f_g^k \in [0, 1]^k$ is a binary feature matrix of fixed dimensionality k.
Let's denote f_a^l to be l per-arm (per selection) features that are included in the second type of context. $f_a^l \in [0, 1]^l$ is a binary feature matrix of fixed dimensionality l, where $l = 2n$.
Each KC is presented by two binary features: 1- If KC is taken and 2- If KC is completed (i.e., passed/failed)
$1 \leq i \leq n | KC_i = [f_{gi}^k; f_{ai}^l]$ - KC feature matrix presenting both global features and per-arm features for each of KC (a horizontal concatenating matrix)
There are a fixed number of arms n in CMAB which are equal to the number of KCs (each KC presents an arm). At each time step t, there are a number of $n_t \leq n$ possible arms to select. The n_t possible arms are dependent on two factors: 1- the remaining KCs to complete and 2 – possible KCs from the AND-OR graph to take. If the selected arm is completed (i.e., the student passes the selected KC), the agent receives a binary reward of 1; otherwise, 0. Let's denote $r_i^t \in \{0, 1\}$ as the binary reward at time step t for selecting arm i. The KC feature matrix is then updated before the next time step.
The framework that we propose consists of offline and online learning, where offline learning stores the structure of domain model, student database and policy base [2]. On the other hand, online learning is the agent using CMAB algorithm to select the next KC for a learner. Every time a learner puts his/her attempt on a recommended KC, either he/she can complete or cannot complete the KC, the binary reward and the KC features are recorded into the student database. We use the binary classification algorithm proposed by David Cortes [15] as black-box oracles for finding the best policy on observed context and rewards. The contextual Multi-Armed Bandit algorithms have the dilemma of balancing exploration and exploitation. Some methods have been proposed such as Epsilon-Greedy, UCB, or Thompson Sampling. In this research, we compare three methods by simulating CMAB using logistic regression as black-box oracles. However, we need to establish baselines for each method first.

Algorithm 1: Epsilon-Greedy

Input probability $p \in (0,1]$, decay rate $d \in (0,1]$, oracle $\hat{f}_{1:n}$

1: For each successive round t with context x^t do
2: With probability (1 – p):
3: Select action a = $argmax_n \hat{f}_{1:n}(x^t)$
4: Otherwise:
5: Select action a uniformly at random from 1 to k
6: Update $p := pxd$
7: Obtain rewardr_a^t, Add observation $\{x^t, r_a^t\}$ to history for arm a
8: Update oracle $\hat{f}_a$ with its news history

Algorithm 2: Bootstrapped-UCB

Input number of re-samples$m$, percentile$p$, oracle $\hat{f}_{1:n,1:m}$

1: For each successive round t with context x^t do
2: For arm q in 1 ton do
3: Set $\hat{r}_q^{ucb} = Percentile_q\{\hat{f}_{q,1}(x^t), \dots, \hat{f}_{q,m}(x^t)\}$
4: Select action a = $argmax_q \hat{r}_q^{ucb}$
5: Obtain rewardr_a^t, Add observation $\{x^t, r_a^t\}$ to history for arm a
6: For re-sample s in 1 tom do
7: Take bootstrapped re-sample X_s, r_s from X_α, r_α
8: Refit $\hat{f}_{a,s}$ to this re-sample

Algorithm 3: Bootstrapped-TS

Input number of re-samples$m$, percentile$p$, oracle $\hat{f}_{1:n,1:m}$

1: For each successive round t with context x^t do
2: For arm q in 1 ton do
3: Select re-sample s uniformly at random from 1 tom
3: Set $\hat{r}_q^{ts} = \hat{f}_{q,s}(x^t)$
5: Select action a = $argmax_q \hat{r}_q^{ts}$
6: Obtain rewardr_a^t, Add observation $\{x^t, r_a^t\}$ to history for arm a
6: For re-sample s in 1 tom do
7: Take bootstrapped re-sample X_s, r_s from X_α, r_α
8: Refit $\hat{f}_{a,s}$ to this re-sample

Re-planning in the CMAB: Given the systematic capture of the learner's performance within the framework of per-arm features, the management of failures assumes significance, particularly in the context of re-planning. In scenarios necessitating re-

evaluation and adjustment, the agent's role extends to the discerning recommendation of the next Knowledge Component, incorporating a nuanced consideration of the failures attributed to the previously suggested KC. This approach not only ensures a meticulous handling of learning dynamics but also underscores the agent's capacity for informed decision-making and adaptive planning within the broader educational framework. The integration of failure-aware considerations into the re-planning process emerges as a pivotal aspect, contributing to the resilience and effectiveness of the learning algorithm.

6. Experiments and Simulations

In our research, we plan to conduct the incorporation of experiments and simulations. in assessing the efficacy and performance of the proposed Contextual Multi-Armed Bandit approach. The experimental design encompasses the deployment of the CMAB algorithm in diverse educational contexts, involving real-time interactions with learners. Both experiments and simulations are carried out in the environment context of course COMP 272 (Data Structures and Algorithms) from Athabasca University. We especially chose Unit 7 (sorting algorithm) of the course to model the KC spaces.

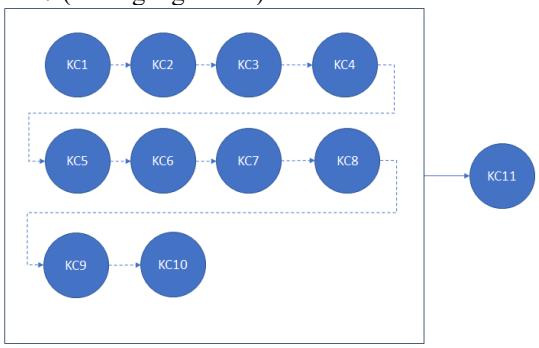

Fig. 2KC graph of sorting algorithm Unit.

Fig 2 shows the KC graph of 11 sorting algorithms in the unit of the course.
In the experiment, participants will engage with a learning system designed to recommend a KC for their learning at each step, based on their individual background and performance, and the system will adapt to their progress. On the other hand, the simulation will produce student data and emulate their learning progression within the learning system. Our primary objective is to assess and compare the performance of the adaptive learning system across various learning curves using four foundational algorithms: Bootstrapped UCB, Epsilon-Greedy, Bootstrapped TS, and Random Action Selection.

7. Conclusion

In conclusion, this research highlights the potential of personalized learning paths as a promising instructional strategy in the realm of online learning. By catering to the unique needs and preferences of individual learners, personalized learning paths have the potential to enhance the learning experience and outcomes. However, the complexity of creating effective personalized learning paths arises from the significant variability in learners' characteristics, behaviors, and learning context. As is known, existing recommendation methods have limitations in dealing with the dynamic nature of learning environments, leading to suboptimal recommendations. To address this gap, our study proposes a novel approach: a personalized learning path recommendation system based on a contextual multi-armed bandit framework. This approach aims to overcome the challenges posed by dynamic learning contexts and adaptively offer students optimal learning sequences. By utilizing the contextual multi-armed bandit approach, the system can dynamically adjust learning paths based on real-time feedback, ensuring that students receive the most relevant and suitable content as they progress. This adaptability allows the system to respond to changes in learners' preferences and needs, providing them with a modified sequence when re-planning is required. Through this research, we hope to contribute to the advancement of personalized learning in online education and provide educators and learners with a more effective, and responsive learning path recommendation system. The findings and insights from this study have the potential to inform the design of future personalized learning platforms, enhancing the overall learning experience and promoting better learning outcomes for diverse learners in various educational settings.

ACKNOWLEDGEMENTS

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and Alberta Innovates.

References

1. S. Graf, F. Lin, Kinshuk and R. McGreal, Intelligent and Adaptive Learning Systems: Technology Enhanced Support for Learners and Teachers, Information Science Reference, 2011.
2. F. Lin, L. Howard, and H. Yan, "Learning Optimal and Personalized Knowledge Component Sequencing Policies," AIED (2), pp. 338-342, 2022.
3. P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," Journal of Machine Learning Research, 3 (Nov), pp. 397-422, 2002.
4. A. H. Nabizadeh, J. P. Leal and H. N. Rafsanjani, "Learning path personalization and recommendation methods: A survey of the state-of-the- art," Expert Systems with Applications, 159, 113596., vol. 113596, no. https://doi.org/10.1016/j.eswa.2020.113596, p. 159, 2020.
5. D. Shi, T. Wang, H. Xing, and H. Xu, "A learning path recommendation model based on a multidimensional knowledge graph framework for e- learning," Knowledge-Based System, 105618, 2020.
6. M. Niknam and P. Thulasiraman, "A bio-inspired intelligent learning path recommendation system based on meaningful learning theory," Education and Information Technologies, pp. 1-23, 2020.

7. D. Xu, Z. Wang, K. Chen, and W. Huang, "Personalized learning path recommender based on user profile using social tags," 2012 Fifth international symposium on computational intelligence and design (ISCID), pp. Vol.1, pp.511-514, 2012.
8. D. Cai, Y. Zhang, and B. Dai, "Learning path recommendation based on knowledge tracing model and reinforcement learning," IEEE 5th international conference on computer and communications (ICCC), pp. 1881-1885, 2019.
9. J. Mui, F. Lin, and M. A. A. Dewan, "Multi-armed Bandit Algorithms for Adaptive Learning: A Survey," AIED (2), pp. 273-278, 2021.
10. J. Xu, T. Xing, and M. v. d. Schaar, "Personalized Course Sequence Recommendations," Personal IEEE Transactions on Signal Processing, vol. 64, no. 20, p. 5340–5352, 2016.
11. I. Wachanrawan, K. Chayapol and T. Punnarumol, "Reinforcement Learning Based on Contextual Bandits for Personalized Online Learning Recommendation Systems," Springer Nature, 2020.
12. E. Duval and W. Hodgins, "A LOM research agenda," in the twelfth international word wide web conference (WWW2003), Budapest, Hungary, 2003.
13. J.-C. Falmagne and J.-P. Doignon, Learning Spaces: Interdisciplinary Applied Mathematics., Springer-Verlag. https://doi.org/10.1007/978-3-642- 01039-2, 2011.
14. B. Giuseppe, L. Jason, and L. Ramon, "A survey of online experiment design with the stochastic multi-armed bandit," arXiv preprint arXiv:1510.00757, 2015.
15. C. David, "Adapting multi-armed bandits policies to contextual bandits scenarios," arXiv:1811.04383v2 [cs.LG], 2019.