Prediction of Students' performance in E-learning Environments based on Link Prediction in a Knowledge Graph

Antonia Ettorre, Franck Michel, Catherine Faron

Prediction of Students' performance in E-learning Environments based on Link Prediction in a Knowledge Graph

Antonia Ettorre[0000−0003−4868−2584], Franck Michel[0000−0001−9064−0463], and Catherine Faron[0000−0001−5959−5561]

Université Côte d'Azur, CNRS, Inria, I3S, Sophia Antipolis, France
{aettorre,fmichel,faron}@i3s.unice.fr

Abstract. In recent years, the growing need for easily accessible high-quality educational resources, supported by the advances in AI and Web technologies, has stimulated the development of increasingly intelligent learning environments. One of the main requirements of these smart tutoring systems is the capacity to trace the knowledge acquired by users over time, and assess their ability to face a specific Knowledge Component in the future with the final goal of presenting learners with the most suitable educational content. In this paper, we propose a model to predict students' performance based on the description of the whole learning ecosystem, in the form of a RDF Knowledge Graph. Subsequently, we reformulate the Knowledge Tracing task as a Link Prediction problem on such a Knowledge Graph and we predict students' outcome to questions by determining the most probable link between each answer and its correct or wrong realizations. Our first experiments on a real-world dataset show that the proposed approach yields promising results comparable with state-of-the-art models.

1 Introduction

The increasingly easy access to high-quality educational resources combined with the recent advances in AI and Web technologies have fostered the development of smart user-centered learning environments. The success of such systems is mainly due to their ability to assist users in their learning process, by offering real-time automated tutoring and personalized revision suggestions. One of the main requirements for these systems is the capacity to trace the knowledge acquired by users over time and assess their ability to face a specific knowledge concept in the future with the final goal of presenting learners with the pedagogical content that will most effectively improve their skills. The challenge of predicting students' outcomes when interacting with a given educational resource, also known as *Knowledge Tracing (KT)*, has been largely investigated and several approaches have been proposed throughout the years, e.g. *Bayesian Knowledge Tracing (BKT)* [3], *Additive Factors Model (AFM)* [2] and *Deep Knowledge Tracing (DKT)* [5]. Although very different, these approaches present a major

commonality: their predictions rely on very limited and simply structured information, i.e. the student approaching the question, the question being answered, and the list of the skills (or knowledge components) involved in the question. Conversely, in real-world scenarios, students' performance can be influenced by several additional factors that are miss-represented or missing in the previously mentioned models, such as type of questions, number of possible answers, assignment or test length, hierarchical organization of knowledge components, etc.

In this paper, we present an approach to represent and exploit this heterogeneous information to provide reliable predictions for students' outcomes to questions. Firstly, we rely on the expressiveness offered by Semantic Web models to represent the whole learning ecosystem as an RDF Knowledge Graph (KG). Then, we reformulate the KT task as a Link Prediction (LP) problem on such a KG and we determine the most probable links between the answers whose outcomes need to be predicted and their correct or wrong results, and we convert these predictions into binary labels for answers' correctness. We think that reformulating the problem of predicting students' outcomes to questions in such a way allows us to take advantage of a much wider amount and variety of information about the learning environment while reusing widely-known and well-established Deep Learning methods for KGs, and avoiding the burden of features engineering. To empirically confirm the validity of the proposed approach, we apply it on a real-world dataset and compare its performance with state-of-the-art KT models.

2 Link Prediction for Students' Outcomes

The approach presented in this paper is mainly based on the hypothesis that it is possible to turn the KT task into a LP problem, after modeling the learning environment as a KG. In other words, instead of predicting the probability that a given student correctly answers a specific question, we evaluate the possibility that an implicit link (i.e. a triple) exists in the KG between the expected student's answer and its correct or incorrect result. Finally, an answer is labeled as `1` (correct) if the link towards the correct result has a higher score than the link towards the incorrect one, while it is labeled as `0` (incorrect) in the opposite case. For example, to predict the positive or negative outcome of the answer given by student A to question 1 (fig. 1), we compute the score of the two triples $\langle answer1, has\_result, correct\rangle$ and $\langle answer1, has\_result, incorrect\rangle$, and we predict that A's answer will be correct if the first triple has a higher score than the second one. To empirically validate our hypothesis, we designed and developed an end-to-end pipeline depicted in fig. 2, which takes as input the traces of the students' learning history, possibly enriched with contextual knowledge, and the list of the student-question interactions whose outcomes must be predicted. The framework implements four steps:

1. **Graph Building**: create a KG representing the learning ecosystem;
2. **Graph Augmentation for Prediction**: inject into the previously created KG new nodes representing the new answers we aim to predict;

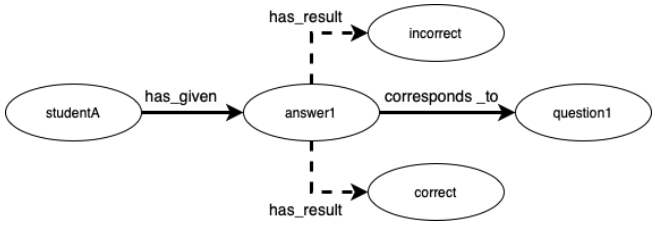

Fig. 1: Example of a KG representing the answer of a student to a question. The dashed lines represent the links we aim to predict.

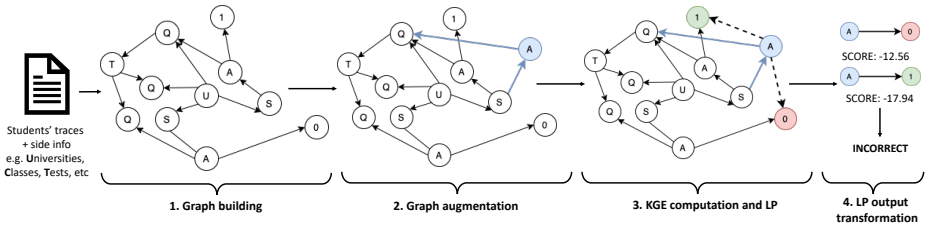

Fig. 2: Depiction of the different steps for the proposed Link Prediction-based approach.

3. **KGE Computation and Link Prediction**:compute the KGEs and use them to assess the scores of the triples to be predicted;
4. **LP Output Transformation**: convert triples' scores into the corresponding probabilities for the binary classification of students' answers (correct or incorrect).

3 Evaluation

Experimental Setup. To validate the proposed approach and to be able to compare it with state-of-the-art KT models, we decided to test our method on a widely-used benchmarking dataset: the ASSISTment 2009-2010 skill builder dataset [4] that stores the learning logs of the users of the ASSISTments platform. For each attempt of a user to a problem, it contains information such as user and problem identifiers, skills required for the problem, problem type (main or scaffolding), answer type (open answer, multiple-choice, etc.), assignment and assistment in which the problem was faced, response time, etc. The first step to apply our approach is to model this information as a KG. For the modeling of the user-problem interactions, we linked each problem to two nodes, one for the positive and one for the negative result, and users' answers are connected to such nodes based on their result. The second step is the computation of the KGEs which has been carried out using TransE [1] to obtain embeddings of dimension 100.

Results and Discussion. Table 1 shows that the newly proposed LP approach achieves an improvement of 2% in terms of bACC, when compared to DKT. We believe that the main reason for this improvement is the ability of our method to consider a greater variety of knowledge about the learning ecosystem. It is also interesting to point out that, for both models, there is a strong difference between negative and positive F1 scores, with Link Prediction achieving slightly more balanced results. This gap in the F1 values can be explained by the highly unbalanced distribution of the target values in the subject dataset, which contains about 70% of correct answers.

Table 1: Results of Link Prediction and DKT approach.

KT model	**F1 (0)**	**F1 (1)**	**bACC**
Link Prediction	0.544	0.734	**0.657**
DKT	0.512	0.755	0.640

4 Conclusion

In this work, we reformulated the KT task into a KG LP problem able to take advantage of all the information commonly available in a smart learning system. Using a well-known benchmarking dataset that we transformed into a KG based on Semantic Web models, our empirical evaluation showed that LP performs slightly better in predicting students' outcomes to questions results, when compared to DKT. Although modest, this improvement suggests that rich context information usually ignored by traditional KT approaches can help achieve better prediction. In the future, we wish to further explore this lead, notably by measuring and enriching the information encoded in KGEs.

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. Advances in neural information processing systems **26** (2013)
2. Cen, H., Koedinger, K., Junker, B.: Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. In: Intelligent Tutoring Systems. pp. 164–175. Springer (2006)
3. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction **4**(4), 253–278 (Dec 1994)
4. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User modeling and user-adapted interaction **19**(3), 243–266 (2009)
5. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep Knowledge Tracing. In: Advances in Neural Information Processing Systems 28, pp. 505–513 (2015)