An Automatic Knowledge Graph Construction System for K-12 Education

Penghe Chen
Advanced Innovation Center for Future Education
Beijing Normal University
chenpenghe@bnu.edu.cn

Yu Lu*
School of Educational Technology, Faculty of Education
Beijing Normal University
luyu@bnu.edu.cn
*Corresponding Author

Vincent W. Zheng
Advanced Digital Sciences Center
Singapore
vincent.zheng@adsc.com.sg

Xiyang Chen
Advanced Innovation Center for Future Education
Beijing Normal University
chenxiyang@bnu.edu.cn

Xiaoqing Li
Advanced Innovation Center for Future Education
Beijing Normal University
lixiaoqing8507@bnu.edu.cn

ABSTRACT

Motivated by the pressing need of educational applications with knowledge graph, we develop a system, called *K12EduKG*, to automatically construct knowledge graphs for K-12 educational subjects. Leveraging on heterogeneous domain-specific educational data, *K12EduKG* extracts educational concepts and identifies implicit relations with high educational significance. More specifically, it adopts named entity recognition (NER) techniques on educational data like curriculum standards to extract educational concepts, and employs data mining techniques to identify the cognitive prerequisite relations between educational concepts. In this paper, we present details of *K12EduKG* and demonstrate it with a knowledge graph constructed for the subject of mathematics.

Author Keywords

Knowledge Graph, Educational Concept, K-12 Education, Online Learning

INTRODUCTION

Knowledge graph is a core component of new generation online education platforms for intelligent education. Different from traditional massive open online course (MOOC) platforms focusing on learning resources provision, new generation online education platforms target on intelligent personalized educational services, such as automatic learning obstacle diagnosis, personalized learning resource recommendation, subject concept and relation visualization, and concept-level knowledge proficiency examination. Our *Smart Learning Partner (SLP)* is such a new generation education platform, developed by Beijing Normal University, that officially serves more than 15,000 local primary and middle school students in Beijing. To better support SLP services, we propose this *K12EduKG* system which can automatically build knowledge graphs for different K-12 subjects by consisting of subject concepts as nodes and concept relations as edges.

Compared to generic knowledge graphs, educational knowledge graph construction faces two major challenges: firstly, the desired educational concept entities are more abstract than real world entities like PERSON, ORGANIZATION, LOCATION. Secondly, the desired relations are more cognitive and implicit, so cannot be derived from the literal meanings of text like generic knowledge graphs. In this paper, we try to tackle these two challenges with this proposed *K12EduKG* system. More specifically, for abstract educational concept extraction, we utilize special data sources like curriculum standards from the education domain. In addition, we utilize students' performance data collected from our SLP education platform to identify relations through data mining techniques.

The proposed *K12EduKG* system endeavors to tackle the above challenges, and it principally makes the following key contributions:

- We propose a novel but practical system to construct knowledge graphs specifically for k-12 educational subjects.
- We employ named entity recognition (NER) techniques to extract educational concepts, and propose association rule mining algorithms to identify prerequisite relations among the concepts.
- We demonstrate an exemplary case with constructing a knowledge graph for the subject of mathematics

L@S 2018, June 26–28, 2018, London, United Kingdom

 ISBN 978-1-4503-5886-6/18/06…15.00

DOI: https://doi.org/10.1145/3231644.3231698

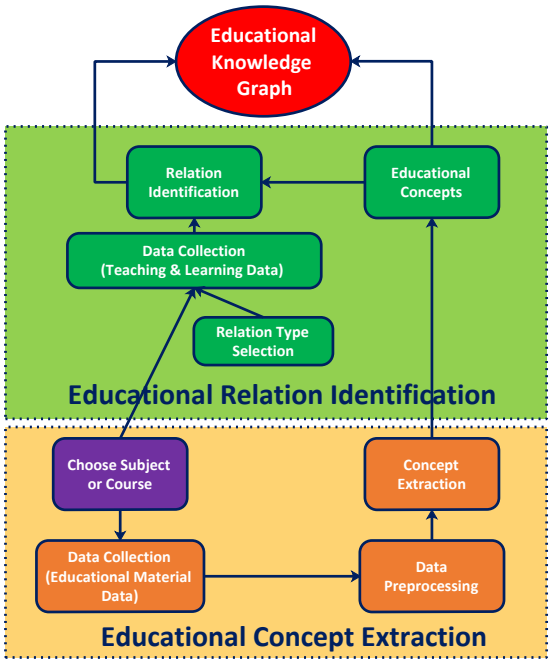

Figure 1. System Architecture of K12EduKG

RELATED WORK

In the education domain, a few studies have been conducted on knowledge graph construction recently. Chaplot and Koedinger use educational data to induce the prerequisite relations among multiple units in a course [2], and Liang et al. recovers prerequisite relations from course dependencies [4]. Another group of researchers utilize the observed relations among courses to create a directed universal concept graph [8, 5]. Different from those systems, our system tries to construct knowledge graph on subject concepts rather than courses, and the objective is to aid teaching and learning rather than finding course dependencies. In addition, we utilize students' performance data rather than course descriptions to identify relations.

SYSTEM OVERVIEW

Figure 1 illustrates the architecture of *K12EduKG* system, which consists of two modules: *Concept Extraction Module* and *Relation Identification Module*. General descriptions of these two modules are given as follows.

- *Educational Concept Extraction Module*: the main objective of this module is to extract educational concepts for any given subjects. Different from generic knowledge graph construction, the data sources are mainly from education domain, typically including the curriculum standards, textbooks and course manuals. Such data may be in the format of printed text, so a step of converting data into machine-encoded text is usually needed. With the produced data, a number of information extraction techniques can be employed to recognize educational concepts, during which the unique characteristics and properties of educational data will be leveraged. The key outputs of this module would be the extracted educational concepts.

Table 1. Sequence labeling example

y	O	O	B-CP	I-CP	O	...
x	Apply	the	Pythagorean	theorem	in	...

- *Implicit Relation Identification Module*: the main objective of this module is to identify the implicit relations between educational concepts. Different from traditional relations derived mainly based on the literal meanings of text, we focus on the implicit relations with educational significance that can directly aid the cognitive process of learning. To achieve this objective, data of learners can be utilized, based on which latest data mining and machine learning techniques can be employed to conduct the algorithms and models. The key outputs of this module are implicit relations that interlink educational concepts.

Finally, the extracted educational concepts and identified relations would formulate an educational knowledge graph that can be used for a variety of applications for both learners and teachers. In the following two sections, we will elaborate our current design for these two modules respectively.

CONCEPT EXTRACTION

Educational concept extraction on text data can be regarded as a word sequence labeling problem. Specifically, given a word sequence such as "*Apply the Pythagorean theorem in* ...", the main objective is to annotate each word with a label specifying whether the word belongs to an educational concept. We define three labels: 1) *B-CP*, which means "beginning of a concept"; 2) *I-CP*, which means "inside a concept"; 3) *O*, which means "outside a concept". By denoting an input word sequence of length T as $x = \{x_1, x_2, \ldots, x_T\}$, and the sequence's labels as $y = \{y_1, y_2, \ldots, y_T\}$, we can illustrate the ideal concept extraction results for a given word sequence with Table 1.

Given the input word sequence x, Conditional Random Field (CRF) [7] is an undirected graphical model to find the best output sequence y that labels each word in x. Among different CRF models, we choose the one with a linear-chain structure because our input and output word sequences have similar chain structure in nature. This linear-chain CRF model with its parameter $\Lambda = \{\lambda_1, \lambda_2, \ldots\}$ can be defined as follows:

$$P_\Lambda(y|x) = \frac{1}{Z_x} \exp\Big(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, x, t)\Big), \qquad (1)$$

where t represents the position in word sequence, and $f_k()$ is the feature function that compute the value for the k^{th} feature of each word. In a nutshell, those feature functions mainly capture the state transition from y_{t-1} to y_t and the observation of sequence x at position t. Parameters λ_k determine the weights of different features, and Z_x is the normalization constant.

Using the above model defined by (1), for any given word sequence x, we can predict its label sequence y^* by solving the following optimization problem defined by (2). In addition, model inference and learning are solved by the methods

mentioned in [7].

$$y^* = \arg\max_y P_\Lambda(y|x), \tag{2}$$

RELATION IDENTIFICATION

As mentioned earlier, the main task of this module is to identify the implicit relations of educational concepts. In the education domain, there are a number of relations critical to both teachers and learners, such as causal relation, progressive relation and inclusion relation. In this work, we mainly focus on the prerequisite relation which is an important one for teaching and learning. It is also a typical implicit relation that hard to be derived from the literal meanings of educational text data.

In this work, we mainly adopt the technique of probabilistic association rule mining to identify relations. Association rule mining is a simple, yet effective data mining technique for discovering interesting relations hidden in large databases. With two key measures, *support* and *confidence*, by defining two key parameters, *minsupp* and *minconf*, association rules can be defined using equation (3). More details can be found in [1].

$$supp(X \Rightarrow Y) \geq minsupp \text{ AND } conf(X \Rightarrow Y) \geq minconf. \tag{3}$$

From the perspective of prerequisite relation, if concept s_i is a prerequisite of concept s_j, learners who do not master s_i very likely do not master s_j, and learners who master s_j most likely master s_i. From the perspective of association rule mining, such a relation between concept s_i and concept s_j is deemed to exist when the following pair of association rules can be identified simultaneously:

$$S_j \Rightarrow S_i \text{ AND } \overline{S_i} \Rightarrow \overline{S_j}, \tag{4}$$

where $\overline{S_i}$ and $\overline{S_j}$ means learners do not master concepts s_i and s_j respectively, and S_i and S_j means learners have mastered concepts s_i and s_j respectively. Hence, we can determine the prerequisite relations between concepts based on learners' concept mastery through estimating the interestingness of those corresponding association rules.

In order to infer learners' mastery on concepts, one natural and feasible approach is to derive from learners' academic performances of related quizzes or exams. However, since academic performance may be affected by uncertain events like slipping (i.e., making an error despite mastering the concept) and guessing (i.e., giving a right answer despite not knowing the concept), learners' mastery on concepts is actually probabilistic. Hence, we adopt a modified version of association rule mining method to identify relations, namely the probabilistic association rule mining techniques [6, 3] which is an extension of the association rule mining for handling the data uncertainties. Probabilistic association rule is defined as:

$$P(S_j \Rightarrow S_i) = P(supp(S_j \Rightarrow S_i) \geq minsupp \; AND \; conf(S_j \Rightarrow S_i) \geq minconf). \tag{5}$$

Substituting the deterministic value in equation 4 with this probabilistic one, with defining a new parameter *minprob*, we can have the requirement as equation 6 and use it to identify the prerequisite relations.

$$P(S_j \Rightarrow S_i) \geq minprob \; AND \; P(\overline{S_i} \Rightarrow \overline{S_j}) \geq minprob. \tag{6}$$

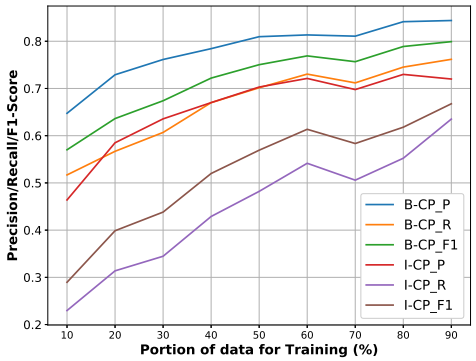

Figure 2. Concept Extraction Results

EXEMPLARY CASE AND SYSTEM EVALUATION

To demonstrate the designed system and algorithms, we build a specific knowledge graph for the subject of mathematics. Specifically, the math concept extraction is firstly conducted, and then the prerequisite relations are identified.

Concept Extraction

Dataset

To conduct the concept extraction for mathematics, we choose the Chinese curriculum standards of mathematics published by the ministry of education as the main data source, which is an official and authorized description on all the key information of this subject. The entire curriculum standards consist of 135 pages and 58,473 Chinese words in total.

Evaluation

As defined in section 4, *B-CP*, *I-CP* and *O* are used to label the word sequences, and we adopt precision, recall and F1-score as the metrics to evaluate. The ground truth is manually labeled by two domain experts. Given a data set consisting of 245 sentences containing 5827 words, we randomly split them into training examples and test examples. To evaluate the model performance with a varying number of training examples, we increase the number of training examples from 10% to 90%. All experiments are repeated 20 times to compute the average performance.

Due to space limitation, we illustrate experimental results of both B-CP and I-CP in Figure 2. From the figure, we can see that the F1-Score of B-CP and I-CP achieves 0.77 and 0.6 respectively when 60% data are used for training. Such results show a higher difficulty in *I-CP* extraction, and it is partly because the correct *I-CP* labeling requires the correct *B-CP* labeling in advance.

Relation Identification

Dataset

To conduct the prerequisite relation identification task, we utilize students' performance data collected by our SLP platform. In this platform, each core concept is associated with a unit test containing multiple questions. Thus, the score rate

Table 2. AUC with Different Parameter Pairs

AUC		*minsupp*					
		400	**600**	**800**	**1000**	**1200**	**1400**
minconf	**0.3**	0.623	0.690	0.645	0.483	0.510	0.477
	0.4	0.689	0.756	0.722	0.518	0.525	0.478
	0.5	0.868	0.874	0.838	0.623	0.559	0.493
	0.6	0.953	**0.953**	**0.954**	0.803	0.692	0.546
	0.7	0.836	0.836	0.836	0.840	0.840	0.688
	0.8	0.850	0.850	0.850	0.853	0.858	0.756
	0.9	0.735	0.735	0.735	0.735	0.747	0.682

of a student on each concept can be easily obtained and used as a measure of student performance. To conduct this experiment, we randomly select 9 mathematical concepts with their corresponding test results as a study case to investigate the prerequisite relations. The corresponding unit tests are answered by 4,488 students from 31 middle schools in Beijing.

Evaluation

The ground truth of the prerequisite relations between selected 9 concepts are annotated manually by two domain experts. If a prerequisite relation exists from concept A to concept B, we call it a positive relation, and if no relation exists, we call it a negative relation. A positive relation is determined only if both experts annotate it. Finally, we obtained 13 positive relations and 59 negative relations between the selected 9 concepts.

Consistent to the normal ways used in educational data mining [3], score rate of concepts are treated as the probability of learners' mastery on the concepts. Hence, for each educational concept, estimated probabilistic knowledge states are derived using the performance data of 4,488 students. The data can be regarded as 4,488 transactions in association rule mining. For each prerequisite relation candidate, we adopt equation (5) to calculate its positive probability and equation (6) to identify the relations, which is mainly affected by the three parameters *minsupp*, *minconf* and *minprob*.

To evaluate the performance of this probabilistic association rule mining algorithm, we use Area Under ROC curve (AUC) as the main metrics. In this experiment, we set *minprob* as 0.7, and Tables 2 summarize the AUC values for different pairs of *minsupp* and *minconf*. Based on the experimental results, we can observe that the *minconf* and *minsupp* pair (0.6, 600) has a significant higher AUC value.

By integrating identified relations with extracted concepts, we can construct the knowledge graph of a given subject. In Figure 3, we present a simplified, yet demonstrative knowledge graph with identified concepts and relations on mathematics subject.

CONCLUSIONS

To automatically construct the domain-specific knowledge graph for education, we developed this novel system *K12EduKG*. More specifically, we elaborated how NER and data mining techniques were employed on heterogeneous educational data to conduct educational concept extraction and relation identification. In addition, we also utilized the subject of mathematics as an exemplary case to demonstrate *K12EduKG* in detail. Furthermore, *K12EduKG* will be deployed on our SLP platform soon to better serve the 15,000 local primary and middle school students as well as the teachers in Beijing.

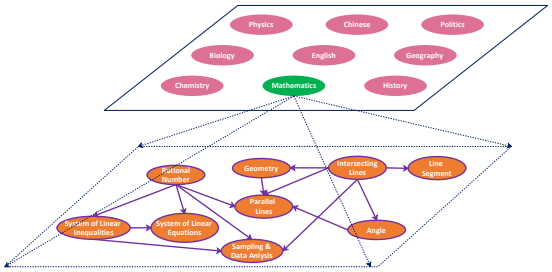

Figure 3. Demonstrative Knowledge Graph of Math

Acknowledgement

This research is partially supported by the National Natural Science Foundation of China (No. 61702039), and the Humanities and Social Sciences Foundation of the Ministry of Education of China (No. 17YJCZH116).

REFERENCES

1. Rakesh Agrawal, Ramakrishnan Srikant, and others. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. 487–499.
2. Devendra Chaplot and Kenneth R Koedinger. 2016. Data-driven Automated Induction of Prerequisite Structure Graphs. In *Proceedings of the Educational Data Mining (EDM)*.
3. Yang Chen, Pierre-Henri Wuillemin, and Jean-Marc Labat. 2015. Discovering Prerequisite Structure of Skills through Probabilistic Association Rules Mining. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM '15)*. 117–124.
4. Chen Liang, Jianbo Ye, Zhaohui Wu, Bart Pursel, and C Lee Giles. 2017. Recovering Concept Prerequisite Relations from University Course Dependencies. In *AAAI Conference on Artificial Intelligence*.
5. Hanxiao Liu, Wanli Ma, Yiming Yang, and Jaime Carbonell. 2016. Learning concept graphs from online educational data. *Journal of Artificial Intelligence Research* 55 (2016), 1059–1090.
6. Liwen Sun, Reynold Cheng, David W Cheung, and Jiefeng Cheng. 2010. Mining uncertain data with probabilistic guarantees. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 273–282.
7. Charles Sutton, Andrew McCallum, and others. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning* 4, 4 (2012), 267–373.
8. Yiming Yang, Hanxiao Liu, Jaime Carbonell, and Wanli Ma. 2015. Concept graph learning from educational data. In *ACM International Conference on Web Search and Data Mining*. ACM, 159–168.