Analytics 2.0 for Precision Education Driven by Knowledge Map

Ying Li[1] liying@buaa.edu.cn
JinCheng Qiu[1]
ShiJie Gui[1]
You Song[1] songyou@buaa.edu.cn Corresponding Author
Tianqi Liu[2]

[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]School of BeiJing, Beihang University

***Abstract*—Aiming to solve learning difficulties caused by information explosion in the current era of education informatization, this paper proposes a precise teaching model driven by knowledge graphs. Knowledge graphs are essentially learning tools that effectively build a correct and complete curriculum knowledge system and accurately promote personalized learning paths. The main research contents are: (1) In view of the problem that the general knowledge graphs currently in use are not applicable in the field of education, we define the knowledge graph ontology structure of special courses based on Bloom's teaching target system; (2) In view of the diversification of various teaching data sources, a simple knowledge graph representing structured data is used as a heuristic condition to extract a complete teaching sequence between knowledge points through self-expansion. (3) In view of the problem faced when trying to navigate relevant knowledge points in personalized learning, this paper exploits the internal relationship between knowledge point loopholes and ability achievements, and constructs an accurate path recommendation model based on quantitative data analysis. Taking the C language programming course as experimental data, this paper verifies the effectiveness of this model by quantitative means, which can significantly provide accurate teaching quality and realize the "multi-directional adaptation" among teachers, courses and students.**

Keywords—Course knowledge graph, Precision Education, Personalized Learning Paths, Relationship Between Knowledge Points

I. INTRODUCTION

Course knowledge graph (CKG) is used to reconstruct and reorder the knowledge points from a large number of course materials. It extracts knowledge points and represents their semantic issues depending upon the structure of nodes and edges within a knowledge graph [1]. CKG focuses on knowledge organization, which determines how to sequence the delivery of knowledge across the curriculum. With the help of CKG, massive fragmented knowledge can be effectively integrated to help learners connect the knowledge system intuitively and with ease, while supporting precision teaching such as personalized learning paths.

It has always been very popular for researchers to apply knowledge graphs and knowledge bases in the field of education. Penghe proposed a system to automatically construct a knowledge graph for education [1]; [2] provide a knowledge graph of courses via machine learning methods to study MOOC courses with ease; [3] develop a system to extract educational concepts and identify implicit relations for K-12 educational subjects. Some researchers have recommended a learning path for students based on the knowledge graph [4][5]. Knowledge graphs can also be used for micro-learning. Some researchers have even gone as far as to build a knowledge base to support the decision-making process of micro-open educational resources adaptation.

At present, practical applications of a knowledge graph in the field of education are still in their infancy, and the following problems still exist in terms of knowledge granularity, domain adaptation, and construction methods.

(1) Knowledge category is coarse-grained: in a common knowledge graph, the nodes are mostly used to represent real entities, and the granularity structure is uncertain. It is difficult to directly represent knowledge elements in the course.

(2) Domain adaptation is not better: Due to the lack of a suitable corpus in the education field, it cannot well simulate and test the individual level of students' cognitive abilities.

(3) The level of automation in construction is relatively low. Many knowledge graphs rely heavily on expert knowledge. The cognitive deviation of different experts on the same knowledge point makes it difficult to guarantee the scientificity and consistency.

Based on the above existing problems, this paper proposed a method of constructing an educational knowledge graph which uses more intelligent means such as big data and deep learning (also known as Educational Knowledge Graph 2.0). With the knowledge graph acquired, this paper hopes to achieve a precise teaching goal that abides the principle of teaching formulation based upon specified learning requirements, and teaching implementation according to students' aptitude. The specific researches are as follows:

(1) Extract knowledge point entities based on vast amounts of teaching data: using that teaching data, different knowledge graph construction methods can be adapted according to the complexity of the data structure. For structured text data, an experience-driven simple knowledge graph is established; for semi-structured data, a rule-driven enhanced knowledge graph is established. For unstructured data, a learning-driven complete knowledge graph is established.

(2) Construct a relationship between knowledge entities based on the cognitive laws: in order to simplify knowledge points that are highly abstract or have complex relationships with one another, a hybrid knowledge-point-relationship-extraction-algorithm is proposed. The algorithm extracts the explicit relationship based on syntactic structure and the implicit relationship based on semantic features. Furthermore, to ensure the consistency and fluidity between the order of textbooks, the order of teaching, and the order of learning as much as possible, Bloom's mastery learning theory is applied to extract crucial relations to help analyze logical features such as timing, causality and condition between knowledge points.

(3) Build a personalized learning path based on the learner's cognitive level: determine an internal relationship between the learning process data (such as homework) and the degree of knowledge mastery,

perform a detailed scan to capture knowledge point loopholes and weak points in learning ability. Finally, accurately recommend learning resources and learning paths in a schematized form to promote personalized teaching in the context of large-scale education.

II. THE RELATIONSHIP BETWEEN KNOWLEDGE GRAPH AND PRECISION TEACHING

Precision teaching aims to personalize the curriculum for each learner and maximize learning efficiency. CKG is an effective means to achieve that precision teaching by providing an individual learning effect measurement model. CKG can link a student's learning to cognitive processes and knowledge structures.

A. To Make Teaching Objectives More Precise

As a start, we introduce a CKG driven by students' cognitive rules, in order to develop more effective teaching objectives (TO). According to Bloom's Taxonomy for cognitive processes, TO comprises the following three levels: (1) knowledge objective (KO) level corresponds to "remember" and "understand"; (2) skill objective (SO) level corresponds to "apply" and "analyze"; and quality objective (QO) level corresponds to "evaluate" and "create".

First, CKG can represent the knowledge points and their sequential relationship which should be taught in this lesson visually. Second, CKG can both accurately understand the students' current learning level and judge a student's zone of proximal development and mark every student's sweet spot. The zone of proximal development (ZPD) is the space between what a learner is capable of doing unsupported, and what they can do if supported [9] [10].

B. To Make Teaching Process More Precise

The implementation of this core state is based on precise teaching objectives. In order to make teaching activities more structured, the class teaching is divided into six links: Bridge-in, Outcomes, Pre-assessment, Participatory Learning, Post-assessment, and Summary. It is a cycle model for lesson planning called BOPPPS. With the support of CKG, BOPPPS can monitor students' learning behavior, provide feedback at each stage, and collect all kinds of fine-grained teaching data. In this case, the experience-based teaching has been optimized to rules-based or even data-driven teaching. Meanwhile, CKG can improve the traditional offline and online independent teaching, into an offline and online blending teaching. It involves a hybrid of online teaching activities and classroom teaching, integrated alongside advanced information technology. This way, students and teachers can interact both in the classroom and Internet at the same time.

III. KNOWLEDGE GRAPH IMPLEMENTATION

This study uses knowledge graph, a kind of semantic network, to extract knowledge points and to describe their relationship in the course. Moreover, a learning path is built to achieve an accurate evaluation of student achievement by carefully scanning of knowledge loopholes and weak links.

A. Model Definition and Description

This paper intends to define a new formal language for knowledge representation.

Definition 1: Course Knowledge Graph CKG = <Model, Data>. Model = <V, P, R>. V, P, R represents a set of extracted knowledge points, their properties or features, and the relationship between entities, respectively. In CKG, the model aimed to organize and integrate data according to a kind of educational ontology. Ontology often contains a subclass-based taxonomic hierarchy, which can have different classifications of the concepts.

Data is the foundation of CKG. Data is gathered from three types of text. Data = da ∪ db ∪ dc. da, db and dc refers to structured, semi-structured, and unstructured text-type data, respectively. Structured data is organized in a pre-defined format; semi-structured data is information that doesn't consist of structured data (relational database) but still has some structure to it; unstructured data is a set of text-heavy data that does not have a pre-defined data model. da = {Table, Multiple choice questionnaire,…}, db = {Syllabus, Text-survey, Database, …}, dc = {Tutorial, Courseware, Homework,…}.

In order to create a more standardized dataset, nested feature vectors were used to represent numeric or symbolic characteristics of each knowledge point. Vi = [course, name], the entity name is defined according to educational ontology; Pi = [type, objectives, difficulties]. In the elements of vector Pi, type = [knowledge, skills and abilities], type indicates the nature of the knowledge, which is divided into three categories; objectives = [remember, understand, apply, analyze, evaluate, create]. In this paper, objectives are defined as a set of six hierarchical models according to the Bloom's taxonomy. difficulties = [easy, relatively easy, normal, hard, very hard], difficulties is used to predict how difficult it is for students to learn a certain knowledge unit, Vi. Ri = [inclusion, precursor, identity, brother, cause-and-effect,…]. Not only does sequence relationship exist in educational entities, but it also has many special relations, as described in Table 1.

Table 1 Special relations in Course Knowledge Graph

Types	Descriptions
Inclusion	A contains B
Precursor	A must be learned before learning B
Identity	A and B are different descriptions of the same knowledge
Brother	A and B have the same parent C, but there is no sequence between them
Correlation	A and B do not conform to the previous relationships, although they are still relevant
Inheritance	A is parent and B is his son
cause-and-effect	A produces an event or condition; B is the results from A

From above, the model of CKG has a hierarchical structure, taking example, <Vi, Pi, Ri>. = < [course.C programming, *name*.function], [[type.skills, type.abilities], [objectives.apply, objectives.analyze], [difficulties. normal]], [Precursor: {function-definition, function- call, module programming }, Brother:{ local variable, global variable}]>.

B. Build Course Knowledge Graph

Converting raw data to a high-quality Course Knowledge Graph mainly contains three steps: extracting knowledge, extracting relationships, and building personalized learning paths.

1) Extracting Knowledge

Knowledge extraction is the process of extracting the concepts, definitions, theorems, properties, formulas, and other domain terms from teaching materials. Educational data are often heterogeneous because they stem from distinct sources and are independent, disparate activities. Data can be classified as structured, semi-structured and unstructured text data. In order to increase the accuracy of data extraction, different methods are used for different structures of data depending on the complexity of data.

First, for structured text data, a simple experience-driven knowledge graph was built. Through eliciting prior knowledge from domain experts and teachers, it can directly extract explicit knowledge from forms-questionnaires or data sheets. Some effective methods, like iterative feedback tuning and manual proofreading, were used.

Second, for semi-structured text data, an enhanced rules-driven knowledge graph was built. A combination of both manual analysis and association rule learning is likely to achieve the best results. The sentences with high semantic complexity are converted into standard key-value pairs to realize the transformation from semi-structured text data to structured text data. In order to get strong semantic expression, the self-expanding method of bootstrapping is used to continuously expand the nodes and edges in the simple knowledge graph.

Third, for unstructured text data, a completed data-driven knowledge graph was built. As we know, teaching materials follow the principles of hierarchy and correlation. It is relatively easy to extract lexical and grammatical features. Therefore, we proposed a hybrid deep learning model (CKG-DNN), which integrates the course knowledge graph and deep learning models for Natural Language Processing (NLP). CKG-DNN took a simple labeled knowledge graph as a heuristic (call seed) to label complex text data such as textbooks and courseware. Then, these historical labeled data were trained to build a predictive model to predict the outcomes for new data points. Through iterative learning, CKG-DNN can extract new entities in large amounts of unstructured text data, effectively expanding the seeds. Finally, an entity annotated corpus of the educational domain is built, as shown in Figure 1.

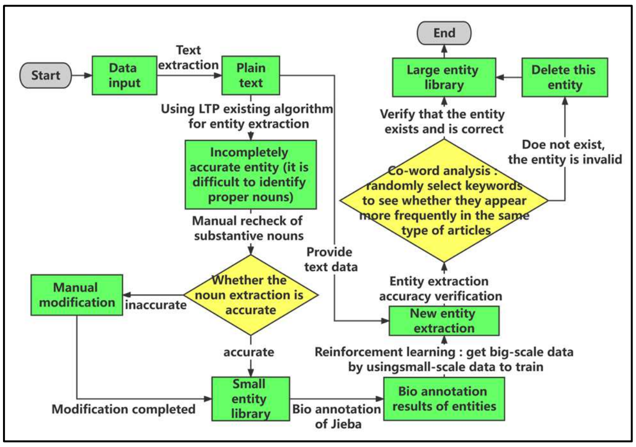

Figure 1 Extract Knowledge from unstructured text data

1) Extracting Relationships

Relation extraction aims to solve the problem of entity semantic linking. The predecessor-successor relationship is the most important relationship between knowledge, and it can generate better semantic sequences. Mastery learning states that students must achieve a level of mastery (e.g., 90% on a knowledge test) in prerequisite knowledge before moving forward to learn subsequent information. Mastery learning is an instructional strategy and educational philosophy, first formally proposed by Benjamin Bloom in 1968. E.g. when teaching Function, parameters and arguments (named knowledge A) are the prior knowledge of function (named knowledge B). According to Mastery Learning Theory, if the learner understood knowledge B, its predecessor knowledge A must also be mastered. We defined the knowledge points with sequential relationships as knowledge pairs. A large number of knowledge graph cases show that a knowledge node often has a strong and similar semantic relationship with its adjacent nodes. Therefore, we extract the relationship between knowledge pairs by measuring their similarity (pairwise similarity) [12]. In this paper, we applied a cosine similarity approach to estimate the degree of similarity between knowledge pairs, as shown in Formula 1. The closer the cosine value to 1, the smaller the angle, and the more relevant two vectors become. We interpret the given Course Knowledge Graph, CKG, as a weighted, undirected graph. The weight of the edge in CKG is given by sim(Vi, Vj), where the nodes Vi, Vj correspond to the knowledge pairs, and they are connected by an edge if sim(Vi, Vj) $\neq$ 0. The matrix Sn = {sim(Vi, Vj)}, i,j=1,2,…n, containing the pairwise similarities between the knowledge points is called similarity matrix (or affinity matrix).

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad \text{Formula 1}$$

C. Create Personalized Learning Paths

According to the above analysis, we proposed a Course Knowledge Graph-based Knowledge Tracing (CKGBKT) method to create personalized learning paths. CKGBKT is a self-supervised deep learning model that finds the relationship between assignment and knowledge through training, represented by the mapping function Y = f(X), where X represents the assignment data, and Y represents the predicted knowledge points measured by X. In teaching feedback, student assignment completion is an important metric, which can well reflect students' mastery of the course knowledge points.

CKGBKT method transforms the assessment of students' mastery of a certain knowledge point into a binary classification problem of whether the knowledge points can be predicted from the assignment. The specific method is as follows:

First, CKG is divided into two parts, a single node, named PVi (represents a knowledge) and the node-removed sub-graph, named $\overline{\text{CPKG}}_\text{i}$ (not including node PVi), CKG = PVi ∪ $\overline{\text{CPKG}}_\text{i}$. Here, we present an ingenious method to generate labeled training samples by artificially constructing two tuples (PVi, $\overline{\text{CPKG}}_\text{i}$), where PVi is the predicted label, $\overline{CPKG_i}$ is the training samples. Then, given a CKG, there exists some target function y = f(x), where $\{x_i, y_i\}$ = {(PVi, $\overline{CPKG_i}$), kp_i}.

Second, in the training stage, assume that the set of correct answers for each assignment $HW^{best} = concat(HW_1^{best}...HW_n^{best})$ (n is the number of assignment) covers all the knowledge points in CKG. In theory, for any knowledge point in CKG, the set of assignment joins sub-graph data can predict the knowledge PVi. The conclusion is that, $\forall \text{PV}_\text{i} \in \text{V}$, the probability of PVi = $f(\text{HW}^{\text{best}}, \overline{\text{CPKG}}_\text{i})$ is 1. Therefore, PVi can be regarded as the label of $(\text{HW}^{\text{best}}, \overline{\text{CPKG}}_\text{i})$.

Third, in the test stage, we input both a student's submitted assignments HW^{stu} and the course knowledge graph data into the trained model CKGBKT. CKKBKT should predict whether he has mastered the knowledge points required in the assignment. If the probability of PVi' = $f(HW^{stu}, \overline{\text{CPKG}}_\text{i})$ is 1, then the submitted assignments HW^{stu} is considered the same as the correct answers. It proves that the student had mastered the knowledge PVi.

IV. EXPERIMENT

The data used in the experiment came from two of the top online learning platforms in China (https://www.educg.net/ , https://www.educoder.net/). It took the C programming course as an example and downloaded a total of more than 6000 teaching documents from cloud platform. 100 students were randomly selected as the research object.

A. Data Collection & Processing

Clean data is the foundation of knowledge graph construction. Data for C Programming Couse Knowledge Graph may be available in multiple sources. There are two types of data, digital teaching materials and learning behavior data. Digital teaching material is used to build course knowledge graph and it mainly includes text resources such as electronic textbooks, syllabuses, courseware and tests. Learning behavior data is used to evaluate students' learning status and effectiveness. It mainly includes online learning behavior (total study time, interactions with content, peers and instructor, number of submitted assignments).

Before training, we applied Jieba (a Chinese word segmentation tool) to clean and format the text data, making it adaptable to the Chinese BERT input method. Finally, 6028 Chinese characters are separated from 27042 valid sentences in the teaching materials.

B. Extracting Knowledge from CKG

After preprocessing, we implement the BERT-BiLSTM-CRF neural network model to perform a BIOES (where B stands for Beginning, I for Inside, O for Out, E for End, and S for Single) annotation on 6028 Chinese words that had undergone word segmentation, completing the task of extracting knowledge point entities. For training, we first pre-train the model using BERT to produce word vectors. Then, we input the word vectors into BiLSTM for further training. Finally, we predict the best label sequence of BIOES through CRF decoding, as shown in the Figure 2 (the nodes in the Figure 2 represent the extracted knowledge points entity).

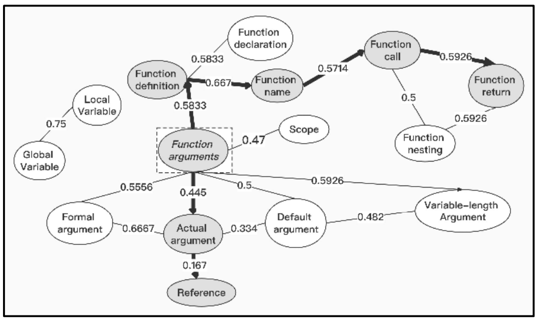

Figure 2 Extracting relationships

The paper measures the quality of the knowledge point entities extracted by the BERT-BiLSTM-CRF model from three indicators, namely: precision (P, P = TP/FN), recall (R, R = TP / FP) and F1-score (F1, F1 = 2*P*R/(P+R)). It can be seen from the Table 2 that more data or structured data will almost always increase the accuracy of a model.

Table 2 The performance of the model CKG-DNN

Data Source	Amount (Bytes)	P (%)	R (%)	F1 (%)
T1	17543KB	0.83%	0.89%	0.81%
T2	26714KB	0.75%	0.77%	0.76%
T3	2024KB	0.61	0.63%	0.60%

(Note: TP and TN stands for true positive and negative, respectively; FP and FN stands for false positive and negative, respectively; T1, T2, T3 stands for structured, semi-structured, unstructured text, respectively)

C. Extracting Relationships from CKG

Relationship extraction is a key but difficult part. The relationship between knowledge points, shown in Table 1(Part A in Section III), is different from common entity relationship. First, according to the Formula 1, we established links between the extracted knowledge points in CKG through calculating the similarity of all nodes. Second, a semantic bootstrapping algorithm was applied to automatically generate semantic lexicons. The input to bootstrapping is an unannotated text corpus and a few manually defined seed words for each semantic category. Then it can learn new words for each category through a specific candidate mechanism, measuring how strongly those contexts are associated with words that belong to the semantic category. Here, we just show a small portion of the relationships about knowledge Function, as shown in Figure 2.

D. Creating personalized learning paths based-on CKG

In this paper, cognitive model is used to measure the completion of a learner's homework and predict the mastery of a certain knowledge point based on the understanding of his cognitive processes. With it, we can tailor a recommended learning path for that specific learner. The knowledge tracking model is a special Markov model used to describe the transition process between the explicit state (the score of an assignment) and the implicit state (the mastery of the knowledge points contained in the assignment). Each knowledge point in CKG involves 5 parameters (P(T), P(L0), P(S), P(G) and P(F)), as shown in the Figure 3. As an example, if a learner's learning performance on the first n-1 assignments is known, then the knowledge state of the n^{th} time can be predicated using the following formula: P(Ln) = P(Ln-1|n-1^{th} performance) + (1-P(Ln-1|n-1^{th} performance))P(T).

The basis of building personalized learning path is to calculate the learner's mastery of a certain knowledge, and then guide the learner to learn accordingly. The specific process is as follows:

(1) Based on the extracted knowledge point entities, design a homework that corresponds to the knowledge point;

(2) Based on the cognitive model combined with the learner's homework, it predicts the learner's mastery of knowledge points. Then, according to the level of mastery of knowledge points, the corresponding index values are indicated on a radar chart (the larger the value, the higher the mastery degree);

(3) Based on the prediction results from (2) combined with the course knowledge graph, we can build a recommendation of precursor knowledge points based on the current knowledge points. In this way, learners can strengthen the fundamentals of the current knowledge points that they have not yet to master, laying a good foundation for learning the current knowledge points;

(4) Based on the prediction results from (2) combined with the course knowledge graph, we can build a recommendation of the follow-up knowledge point based on the current knowledge point. This can help guide learners that have successfully mastered the current knowledge point into learning the follow-up knowledge point more efficiently, thus improve learning efficiency.

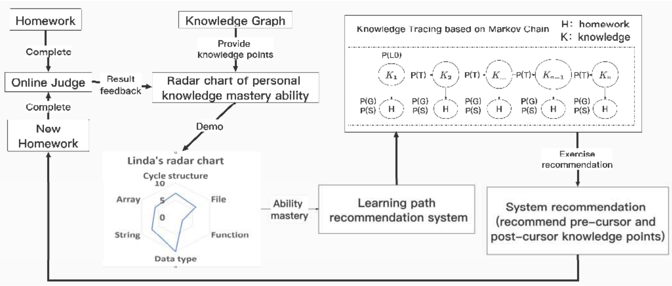

Figure 3 The process of creating personalized learning paths

To verify the validity of the personalized learning path generation, a group of 50 students in C programming course were randomly selected to participate in the study. By applying our CKGBKT model, the mean accuracy is greater than 70%. We can predict a student's own learning path according to his submitted assignment in one semester, as shown in Figure 3. Through reviewing his homework, it was found that he did not fully understand the "Function Arguments" knowledge. Hence, a personalized learning path containing the recommended pre-cursor and post-cursor knowledge was created, as shown in Figure 2. It has been verified that this student's learning path is effective. Validation of personalized path validity for a large number of students is the focus of future work.

V. CONCLUSION

This paper proposes a course knowledge map to building a personalized learning path for the implementation of precision education. In this course knowledge graph, a large number of multi-sourced and heterogeneous teaching data is extracted to cultivate knowledge closure and key knowledge paths needed in developing professional comprehensive ability. The innovations are as follows:

(1) Designed an entity extraction method based on deep learning knowledge graph, which can effectively solve the problem of arranging high-complexity heterogeneous teaching data. Construct an algorithm with the capabilities of automatic BIO entities marking, breaking the existing bottleneck of lack of entity corpus in the field of education.

(2) Designed a hybrid knowledge-point-relationship-extraction-algorithm that integrates both syntax and semantics. It adopts the extraction of explicit entity relationships based on syntactic structure and the extraction of implicit entity relationships based on semantic features. Based on traditional algorithms, key relationship information in between knowledge points is uncovered with Bloom Taxonomy as the heuristic condition. This effectively compresses the original extraction scope and improves extraction efficiency.

(3) Designed a personalized learning path recommendation model based on the importance of each knowledge point and the degree of path achievement. By uncovering the implicit relationship between homework and knowledge mastery, we can extract knowledge points that have a greater influence (either positive or negative) on students' learning outcomes and predicts a learning path that matches students' cognitive level according to the Markov chain method.

However, due to the research time being relatively short, we only focused on how to build knowledge graphs and personalized learning paths, but have not apply the research on a larger scale compared to regular teaching. This will be the key study in future.

ACKNOWLEDGMENTS

This work was supported in part by NSFC (No.62107002) and National Key R&D Program of China (2021ZD0110703).

REFERENCES

[1] Chen P, Lu Y, Zheng V W, et al. Knowedu: A system to construct knowledge graph for education[J]. Ieee Access, 2018, 6: 31553-31563.

[2] Dalipi F, Imran A S, Kastrati Z. MOOC dropout prediction using machine learning techniques: Review and research challenges[C]. 2018 IEEE global engineering education conference (EDUCON). IEEE, 2018: 1007-1014.

[3] Chen P, Lu Y, Zheng V W, et al. An automatic knowledge graph construction system for K-12 education[C]//Proceedings of the fifth annual ACM conference on learning at scale. 2018: 1-4.

[4] Shi D, Wang T, Xing H, et al. A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning[J]. Knowledge-Based Systems, 2020, 195: 105618.

[5] Durand G, Belacel N, LaPlante F. Graph theory based model for learning path recommendation[J]. Information Sciences, 2013, 251: 10-21.

[6] Ji S, Pan S, Cambria E, et al. A survey on knowledge graphs: Representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[7] Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs[J]. Proceedings of the IEEE, 2015, 104(1): 11-33.

[8] Lin J, Zhao Y, Huang W, et al. Domain knowledge graph-based research progress of knowledge representation[J]. Neural Computing and Applications, 2021, 33(2): 681-690.

[9] Shabani K, Khatib M, Ebadi S. Vygotsky's zone of proximal development: Instructional implications and teachers' professional development[J]. English language teaching, 2010, 3(4): 237-248.

[10] Zou X, Ma W, Ma Z, et al. Towards helping teachers select optimal content for students[C]//International conference on artificial intelligence in education. Springer, Cham, 2019: 413-417.

[11] Driscoll M P, Burner K J. Psychology of learning for instruction[J]. 2005.

[12] Zhu G, Iglesias C A. Computing semantic similarity of concepts in knowledge graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(1): 72-85.