World Wide Web (2023) 26:3483–3502
https://doi.org/10.1007/s11280-023-01194-8

Course map learning with graph convolutional network based on AuCM

Jianing Xia[1] · Man Li[1] · Yifu Tang[1] · Shuiqiao Yang[2]

Received: 31 March 2023 / Revised: 15 June 2023 / Accepted: 1 July 2023 /
Published online: 29 July 2023

Abstract
Concept map provides a concise structured representation of knowledge in the educational scenario. It consists of various concepts connected by prerequisite dependencies. With the abundance of educational resources available through MOOCs, encyclopedias, and electronic textbooks, extracting prerequisite dependencies and building concept maps becomes feasible. However, publicly accessible taxonomies or learning object information that can help identify prerequisites are rare. To address this, we have constructed a comprehensive dataset called the Australian Course Map data (AuCM), specifically tailored for training concept maps in the IT/CS field. The dataset comprises course descriptions from 14 different Australian universities. To identify prerequisite relationships between course concepts, we have employed an embedding-based approach that combines the Graph Convolutional Network (GCN) with pairwise features of concepts. We have evaluated the performance of our model with non-neural classifiers and neural networks for extracting these prerequisite relations.

Keywords Course map · Prerequisite relation · Word embeddings · GCN

Jianing Xia and Man Li these authors contributed equally to this work.

This article belongs to the Topical Collection: *Special Issue on Fairness-driven User Behavioral Modelling and Analysis for Online Recommendation*
Guest Editors: Jianxin Li, Guandong Xu, Xiang Ren, and Qing Li.

✉ Yifu Tang
tangyif@deakin.edu.au

Jianing Xia
xiaji@deakin.edu.au

Man Li
amanda.li@deakin.edu.au

Shuiqiao Yang
Shuiqiao.Yang@data61.csiro.au

[1] School of Information Technology, Deakin University, Melbourne, VIC 3125, Australia

[2] Data61, CSIRO, Marsfield, NSW 2122, Australia

1 Introduction

The availability and convenience of numerous online learning resources present significant opportunities for researchers to advance educational offerings. How to recommend an appropriate learning sequence to learners with different foundations and backgrounds from the massive learning resources? Prerequisite relations are crucial in this context because they outline the pedagogical dependent relations between educational units, i.e. simple and basic content should be learned first, then complex and advanced content that builds on previous content should be taught. They could assist decision making such as curriculum planning for educational designers to improve course offerings and improve the overall learning experience for students by providing a clear understanding of what is expected and how the different topics fit together. Generally, precedence dependencies are explicit within an educational institution and implicit in the Mooc and cross-university scenario. Inspired by other application based on knowledge graph [1–3], we aim to create a concept map, which serves as a general graph, encompassing normative and discriminative concepts taught across a wide range of courses. In this concept map, the nodes represent these concepts, while the edges capture the pairwise preference in the teaching sequence.

A concept map is a useful tool for educators and students as it provides an overview of the topics covered, the sequencing of those topics, and how they relate to each other, especially the prerequisite dependence. It can also help to identify any gaps or overlaps in the content and can be used to plan and track the progress of a course.

For example, as shown in the lower part of Figure 1, in order to learn the concept of "minimal spanning tree" in the course of "Data structure and algorithms", we should have prior knowledge of "graph" and "node". So the concepts of "graph" and "node" are the prerequisites of the "minimal spanning tree". If there are numerous directed links connecting concepts from one course to concepts from another, we might infer that the two courses have

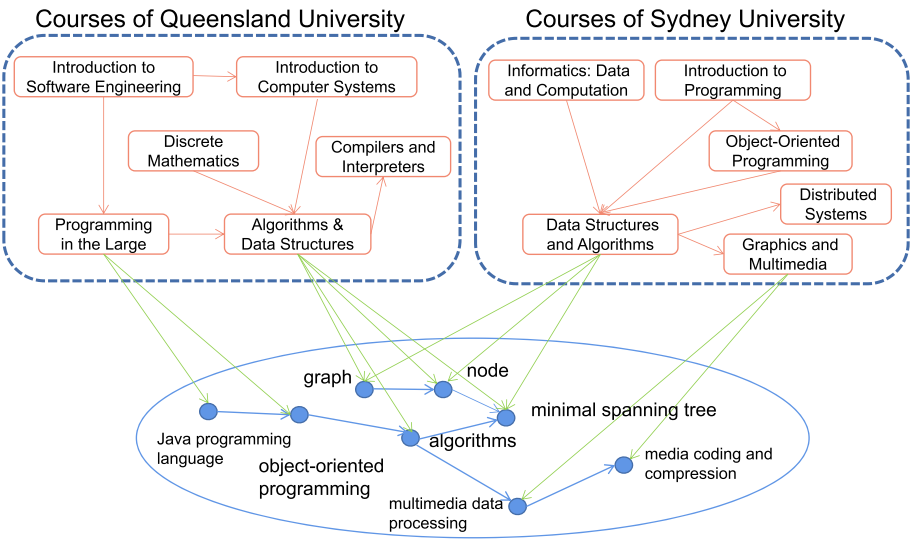

Figure 1 The framework of two-scale directed graphs: The higher-level graphs depict courses as nodes, interconnected by prerequisite dependencies shown as links. On the lower-level graph, universal concepts are represented as nodes, with pairwise preference information

a prerequisite relationship. A directed graph with a wide range of universal concepts is essential for reasoning about course content overlap and prerequisite relationships. Through the recovered explicit prerequisite relations between concepts, an optimal study sequence can be recommended to students, which could cover the online courses from different universities, such as the upper part of Figure 1. Concept maps could also advance the other educational applications such as knowledge tracing, intelligent tutoring and students' performance prediction [4–7].

There are a few efforts aiming to extract concepts and learn prerequisite relationships from educational data, such as courses description [8], MOOCs (Massive Open Online Courses) [9], textbooks [10] and so on. However, there are still many challenges to automatically mining prerequisite relations among concepts. Selecting appropriate features that effectively capture the semantic or contextual relationships between concepts can be challenging. Identifying the most relevant and discriminating features requires careful data gathering and processing, which can be time-consuming and non-trivial. Moreover, domain-dependent feature extraction strongly relies on domain expertise which can't reach the fundamental generalization, and non-optimal feature configurations could compromise classification accuracy. In this paper, we aim to extract the prerequisite dependencies between concepts based on word embedding. Our approach involves using word embeddings to map high-dimensional data to a low-dimensional latent space while preserving the semantics of the original data. Furthermore, we leverage graph convolutional networks (GCN) to incorporate information from neighboring nodes and subsequently apply neural networks to predict the concept prerequisites.

In summary, our work makes the following contributions:

- Proposal of a novel model for identifying concept prerequisite relations: We introduce a unique approach that combines concept representation, graph neural networks (GCN), and concept pairwise features. By integrating these components, our model effectively captures contextual and structural information from course descriptions to construct an initial concept graph. Through the utilization of GCN, we can update concept representations and extract crucial information from the constructed concept graph.
- Integration of contextual and structural information: Our approach leverages both the textual information in course descriptions and the structural relationships between concepts to enhance the accuracy of identifying prerequisite relations. By considering the context in which concepts are mentioned and their connections within the concept graph, our model provides a more comprehensive representation of concept relationships.
- Evaluation on the AuCM dataset: We introduce a newly collected dataset called AuCM, comprising real courses from 14 Australian universities. This dataset allows us to evaluate the performance of our method in a realistic educational setting. By conducting extensive experiments on AuCM, we demonstrate the effectiveness of our model in accurately identifying prerequisite relationships between concepts. We also verify the usability and applicability of the dataset for concept map construction tasks.
- Comparison with baseline models: To establish the superiority of our approach, we compare its performance with baseline models based on the AuCM dataset and another dataset from the USA. Through these comparisons, we provide empirical evidence showcasing the enhanced performance and robustness of our proposed model.

This paper is structured as follows: In Section 2, we conduct a review of the relevant literature concerning the exploration of prerequisite relationships. In Section 3, we introduce the construction of our dataset AuCM and the statistical and semantic analytics on it. In Section 4, we present the formulation of the problem and describe the model in detail, which

involves constructing an initial concept map and utilizing GCN and Siamese networks for concept fusion and prerequisite prediction. In Section 5, we conduct a comprehensive set of experiments to evaluate the efficacy of our proposed method. Finally, in Section 6, we summarize our research findings and outline potential avenues for future research.

2 Related work

Based on various educational resources, several efforts have been devoted to the goal of automatically discovering prerequisite relations among concepts. These methods can be classified into two categories: recovery-based methods and learning-based methods.

Recovery-based methods. Recovery-based methods assume that prerequisite strength between two courses is a cumulative effect of the prerequisite strengths of all concept pairs. These approaches aim to recover the underlying concept graph by mapping courses to concepts and identifying concept-level dependencies. Yang et al. [11] created a concept graph by mapping courses to concepts and learning concept-level dependencies based on the optimization method of adapted versions of SVM algorithms. They can predict unobserved precedence relationships among any courses using the induced concept network. Using a similar setting as that of [11], Liang et al. [12] adopted a methodology that involved representing each course using Wikipedia concepts and assigning tf-idf weights to these concepts. To optimize their approach, they employed a variant of the soft-margin Support Vector Machine (SVM) algorithm. Experiments had been done on both a synthetic data set and a real university course data set to show their superior performance. It is important to note that recovery-based approaches primarily focus on the dependencies between courses while neglecting semantic or structural details among concepts. To improve the accuracy of extracting prerequisite relations, many researchers have turned to learning-based strategies.

Learning-based methods. Learnig-based approaches leverage machine learning techniques to capture the semantic and structural aspects of the relationships between concepts. In the field of concept prerequisite relation discovery, these approaches typically involve training prerequisite classifiers using manually created or automatically generated features. One common source of information utilized by learning-based methods is Wikipedia. As the largest Internet encyclopedia, Wikipedia contains a vast amount of articles covering various knowledge concepts across different subjects. Researchers have exploited the content of Wikipedia articles and their linkage structures to extract prerequisite evidence. Liang et al. [8] first proposed an approach based on the hypothesis that the prerequisite relation is related to the degree to which two related concepts refer to each other in Wikipedia. They defined a statistical function called reference distance (RefD) with the input of item frequencies to calculate the presence of prerequisites. However, RefD is only applicable to Wikipedia concepts due to its reliance on the link structure of Wikipedia. To overcome this limitation, Pan et al. [9] proposed a method MOOC-RF by devising semantic, contextual, and structural features based not only on Wikipedia articles but also on course videos and video sentences from Massive Open Online Courses (MOOCs). They built a set of features and trained 4 binary classifiers to recognize prerequisite relationships between concepts in video transcripts. Similarly, Xiao et al. [13] utilized the frequency and position of concepts in course descriptions, as well as the learning order between courses, to construct features for recovering prerequisite dependencies. They combined these features with category and clickstream information from Wikipedia and evaluated their performance using six common classifiers. Combining the recovery-based technique from [12] and the learning-based model

from [9], an iterative prerequisite relation learning approach was proposed by [14]. They first extracted domain-specific concepts by a graph-based ranking method. Then concept pair features and dependencies among learning materials were utilized to explore the prerequisite weight matrix among concepts.

In addition to traditional binary classifiers, researchers also utilized other deep learning models for concept map generation, such as Siamese networks and MLP. In order to infer concept prerequisite relations using university courses and MOOC datasets, Roy et al. [15] proposed a new supervised learning method using the pairwise-link LDA model to create vector representations of concepts and a Siamese network to forecast unknown concept prerequisites. Using the same Siamese network, Jia et al. [16] proposed an approach for concept prerequisite relation learning (CPRL). They utilized concept pairwise features inspired from [9] and [10] and concept representation learned from a heterogeneous graph. Xiao et al. [17] proposed an approach for mining precedence relations between lecture videos in a MOOC. By automatically extracting main concepts from video captions and utilizing an LSTM-based neural network model to measure prerequisite relations among these concepts, the authors effectively identified the precedence relations between lecture videos.

While previous studies on learning-based approaches have achieved promising results in predicting concept prerequisite relations, there are still challenges that need to be addressed in this field. One major challenge lies in feature engineering and classifier selection, which can introduce biases and limitations, especially when working with proprietary synthesis data. Furthermore, in neural network settings, the complex and numerous relationships between concepts and learning resources have not been fully exploited. This presents an opportunity to enhance the accuracy and effectiveness of predicting concept prerequisite relations. To address these challenges, our study proposes a novel approach that combines handcrafted features and a Graph Convolutional Network (GCN) to improve the prediction of concept prerequisite relations. By combining handcrafted features and GCN, our proposed approach aims to overcome the limitations of traditional feature engineering and classifiers, while leveraging the inherent power of neural networks to capture the intricate relationships between concepts and learning resources.

3 AuCM dataset

The current landscape of educational concept map analysis lacks publicly available datasets to support related research. Consequently, researchers often resort to generating private datasets through concept and link extraction methods. However, this approach poses several challenges. First, the data collection process can be time-consuming, requiring significant effort and resources. Additionally, the synthesized private data may be influenced by the researchers' own analysis results, potentially introducing bias into the dataset. To overcome these limitations, we built a comprehensive dataset named AuCM (Australian Courses Map data) for learning concept maps. Specifically, We collected all possible courses for bachelor's degrees from 14 Australian universities in the area of IT and CS. For comparative analysis, these universities were selected from five states including the Australian Capital Territory (ACT) and half of the universities were from the "Group of Eight (G8)"- a prestigious group of comprehensive research universities that are recognized as world-class institutions. Our work on the AuCM dataset construction consisted of two steps: data scraping and data processing. During the data scraping phase, we collected WEB pages from 1292 undergraduate courses. Then we extracted the relevant information and organized the raw data during the

data processing phase. To the best of our knowledge, this is the first dataset containing course information from Australian universities. Besides, we analyzed the statistical attributes and semantic properties based on the words and concepts retrieved from the course descriptions and visualized them to illustrate how our dataset could be used.

3.1 Statistical analysis of AuCM

The statistics of courses from various universities of 5 states and the ACT of Australia are listed in Table 1, which covers the total number of courses and the number of courses with prerequisite requirements. Data comparison in Table 1 shows that universities belonging to G8 generally offer significantly more courses than non-G8 universities in Bachelor of IT/CS programs. Like courses of USYD and ANU are more than 100 while UTS offers 71 courses and ACU only offers an example study plan which has 25 courses in total. The average number of courses in these G8 universities is 115 which corresponds to 69 of non-G8 universities. This may be attributed to the fact that most non-G8 universities don't offer a bachelor of CS or CS as a major is included in the bachelor of IT.

Moreover, a majority of courses have at least one prerequisite course. Of the 14 universities, there are 811 unique IT&CS undergraduate courses, 511 courses with prerequisite requirements, and 1475 pairs of courses with prerequisite relations. In addition, the average number of prerequisite links per course is 1.82. Figure 2 shows the proportion of courses with prerequisites at each university. The red line is drawn by a linear regression algorithm, indicating that the proportion of prerequisite courses is around 75%.

Based on our statistical analysis, a comparison between universities belonging to the G8 group and those outside the G8 group revealed that G8 universities offer a greater number of courses for a bachelor's degree in IT/CS. We also confirmed that most courses have prerequisite requirements providing evidence for conceptual prerequisite learning.

Table 1 IT&CS courses of 14 universities in Australia

#	G8 or not	University	#courses	#courses with prerequisites
1	G8	Sydney University, USYD	126	52
2		The Australian National University, ANU	116	109
3		The University of Queensland, UQ	160	148
4		University of Western Australia, UWA	131	111
5		The University of Adelaide, ADELAIDE	85	68
6		Monash University, MONASH	119	44
7		University of Melbourne, UofMELB	70	59
8	Not G8	University of Technology Sydney, UTS	71	62
9		Australian Catholic University, ACU	25	16
10		Queensland University of Technology, QUT	85	63
11		Edith Cowan University, ECU	50	26
12		University of South Australia, UN	69	50
13		Deakin University, DEAKIN	126	81
14		Victoria University, VU	60	45

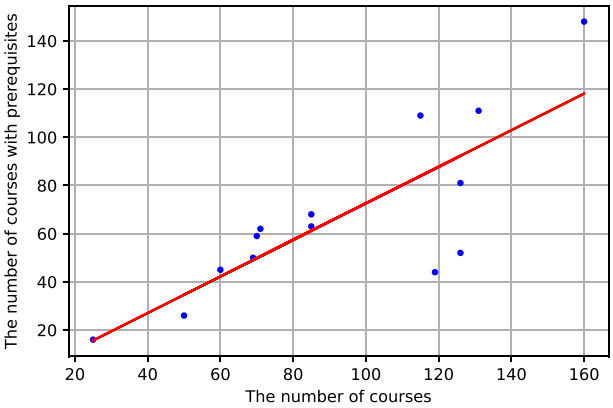

Figure 2 #Courses with prerequisites vs. #courses

3.2 Semantic feature extraction and analysis

To recover the concept-level semantic features, we first collect concepts from Wikipedia and process course descriptions by employing various traditional NLP tools, such as stop word removal, sentence segmentation, and lemmatization. Then we match the Wikipedia concept appearance with the course description. By using a pre-trained tokenizer from the BERT model, we tokenize the course description and retrieved concepts respectively in order to capture semantic characteristics. Furthermore, we employ t-SNE (t-Distributed Stochastic Neighbor Embedding) to project the high-dimensional features onto a two-dimensional map.

By drawing scatter plots for concept/course description based on the projected 2D features, we can see the difference in terms of semantic richness between G8 and non-G8 universities. The comparison of course descriptions between G8 schools and non-G8 institutions is shown in Figure 3.

The scatter plot on the left gives the distribution of words in the course descriptions, where the blue dots are the words described in the courses of G8 universities, and the red dots belong to the non-G8 school courses. It is clear that the blue dots are more scattered and span a larger

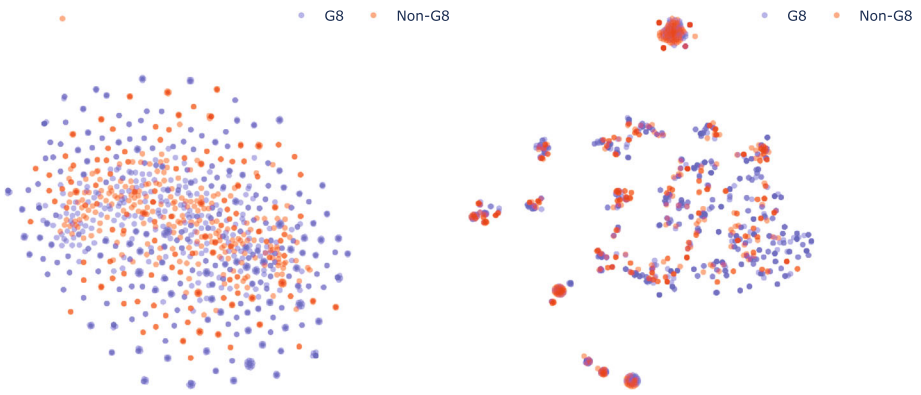

Figure 3 Scatter plots of projected features extracted from course descriptions/concepts in the courses from G8 and non-G8 universities

area, indicating that the G8 course descriptions employ more semantically rich words and cover a broader diversity. The scatter plot on the right, which displays the distribution of concepts, also demonstrates this. Compared to relatively dispersed blue points, the red points cluster in some conceptual areas, showing that non-G8 courses may pay more attention to particular concepts rather than breadth.

3.3 Concept map learning

In addition to conceptual semantic comparison among universities in Australia, our dataset is primarily utilized to create concept maps. Concept map construction intends to extract structured information from unstructured text and represent it as a graph, where concepts constitute the vertices and prerequisite dependencies make up the links. An example of the concept map is shown in Figure 4, where the vertices come from the extracted concepts, and the links, represented by blue dotted lines, indicate the relationships between concepts based on a simple assumption: the prerequisite relationship between two courses means that the concepts contained in these courses have links. It is observed that the vertex "data type" at the center of the concept graph has many connections with other vertices, indicating that the "data type" constitutes a prerequisite for many concepts. In this study, we further proposed a novel model to discover the prerequisite relationships between concepts. The proposed model incorporates multiple components, including word representation, handcrafted features, and course dependencies, within a neural network framework. The combination of these features allows for a more comprehensive and effective approach to identifying prerequisite relationships among concepts.

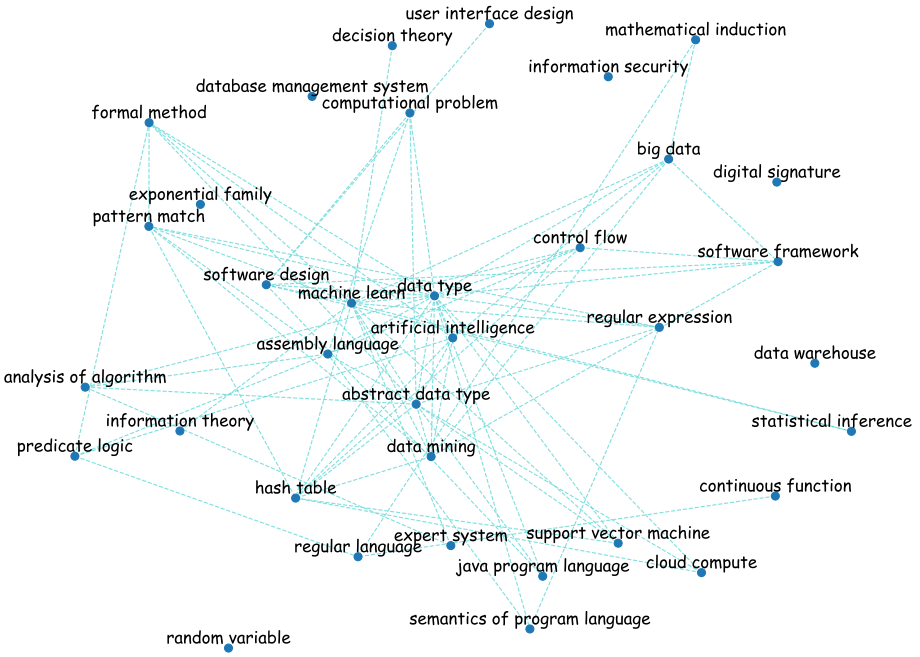

Figure 4 An example of concept map

4 Concept map learning with graph convolutional network (CML-GCN)

Our model, which we have named "Concept Map Learning with Graph Convolutional Network," will be referred to as CML-GCN for brevity. The primary objective of the CML-GCN approach is to enhance the performance of concept prerequisite learning by leveraging the strengths of both graph convolutional networks and Siamese networks. Concept maps, which serve to distill and represent the main concepts and their interactions within the input corpus, are utilized in this approach. In this section, we will formally present our problem statement and provide a detailed description of our proposed method.

4.1 Problem formulation

We formulate the generated concept map $G(C, E)$ as a unified local graph whose nodes represent concepts and edges represent prerequisite dependencies. Course concepts refer to the subject knowledge taught in a particular course. These concepts are not just mentioned but are actively discussed and taught throughout the course. Let us denote the course concept set of D as $C = \{c_1, c_2, ..., c_m\}$ where c_a is the course concept in the concept set C. E contains the directed edge (c_a, c_b) if and only if concept c_a is a prerequisite of concept c_b .

Creating a concept map G to visualize the dependencies between concepts in a course can be challenging without an automated approach. To address this, we first extract concept C from course description D using information from Wikipedia. Then, we infer concept prerequisites E using the course description and other course-related information, such as course dependencies.

The identification of prerequisites relations can be formalized as given a course description corpus D, its corresponding course concepts C, and the course precedence, the objective of the study is to learn a function $P : C^2 \rightarrow \{0, 1\}$ that takes a concept pair (c_a, c_b) as input, where c_a and c_b belong to the set of concepts C. The function P aims to map this concept pair to a binary class, indicating whether c_a is a prerequisite concept of c_b. In other words, the function P predicts whether there exists a prerequisite relationship between concept c_a and concept c_b, with the output being either 0 (indicating no prerequisite relationship) or 1 (indicating the presence of a prerequisite relationship). For convenience, we list the main symbols used in this paper in Table 2.

Table 2 Meaning of symbols used

Symbol	Meaning
D	a set of course text data, and $D = \{d_1, d_2, ..., d_n\}$
X	an n-by-n matrix where each cell is the binary indicator of the prerequisite relation between two courses, i.e., $x_{ij} = 1$ means that course j is a prerequisite of course, i, and $x_{ij} = -1$ otherwise
$G(C, E)$	a unified directed graph, called concept map
C	a set of concepts extracted from D, i.e.$C = \{c_1, c_2, ..., c_m\}$
E	a m-by-m matrix, represent prerequisite relations among concepts
$Set(c_a)$	all the courses which include concept c_a

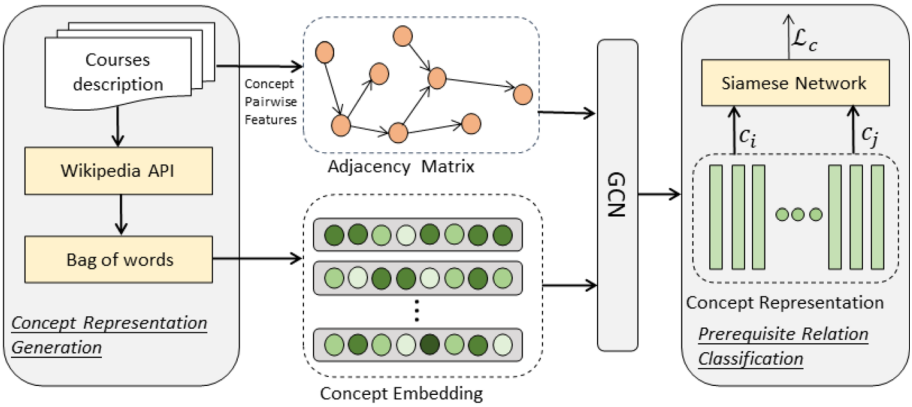

Figure 5 The framework of proposed method CML-GCN

4.2 Framework of CML-GCN

Based on the observation that most universities include an overview of a course's topics in its description, the proposed method aims to extract course concepts from these descriptions. The architecture of the proposed method is depicted in Figure 5. First, all Wikipedia concepts are extracted from the course descriptions using the Wikipedia API. An initial concept graph is generated using Bag-of-words (BoW) for concept representation and handcrafted features to construct an adjacent matrix. The concept graph is then processed using a Graph Convolutional Network (GCN) [18] to update the concept representations. The updated concept representations are then input into a Siamese network to predict prerequisite relations between concepts.

4.3 Initial concept map generation

In order to construct the initial concept graph, we first extract concepts from course descriptions using Wikipedia API and then represent the concepts in our corpus utilizing a straightforward method of Bag-of-words (BoW). To construct the adjacency matrix A, pairwise features for concepts are extracted according to textual and structural information from course descriptions which provide important clues to infer prerequisite relations among concepts. The edges of matrix A are generated through the combination of various features using a weighted average.

Based on our analysis, we have noticed variations in the depth and complexity of concepts and vocabulary used in course descriptions between G8 and non-G8 universities. Moreover, we have identified that both core units and course pre-chains play a significant role in uncovering prerequisite connections. As a result, we have put forward three specific characteristics that utilize the influence of core units, courses from prestigious institutions, and course pre-chains on the conceptual prerequisites. These features can also be applied to other datasets that possess similar attributes.

Core Course Indicator. Core courses are typically foundational courses that are required for a particular degree or program, and often serve as a basis for more advanced coursework. Follow-up courses are professional-related courses or elective courses, usually depending

on these core units. Based on this observation, we use $Core(c_a) = \frac{\sum_{d_i \in Set(c_a)} Core(d_i)}{|Set(c_a)|}$ to denote the extent to which the concept c_a belongs to the core course, where $Core(d_i)$ is a binary indicator which indicates whether d_i belongs to core units. So the feature is defined as: $Coref(c_a, c_b) = Core(c_b) - Core(c_a)$ When the $Core(c_b)$ is bigger which means most courses containing concept c_b are core units, while $Core(c_a)$ is smaller represents less courses containing concept c_a are core units, so concept c_a may depend on core-like concept c_b.

Elite School Course Indicator. This feature captures the influence of courses offered by elite or highly-regarded schools on the prerequisite relationship between other courses. Elite schools are often associated with higher-quality instruction, resources, and prestige. In addition, we proved that course concepts from G8 universities are semantically richer and tend to be more advanced than non-G8 ones based on our AuCM dataset, and as such, courses offered by these schools may depend on more basic units in the conceptual prerequisite relationship. We use $Fam(d_i)$ to indicate whether the course is from a famous university, which refers to G8 university in AuCM. $Fam(c_a)$ is used to represent the probability of c_a belonging to courses from famous universities as follows:

$$Fam(c_a) = \frac{\sum_{d_i \in Set(c_a)} Fam(d_i)}{|Set(c_a)|}. \tag{1}$$

Therefore the feature is defined as:

$$Famf(c_a, c_b) = Fam(c_a) - Fam(c_b). \tag{2}$$

Prerequisite Depth Distance. Pre-chains are often used to ensure that students have the necessary foundational knowledge before moving on to more advanced coursework which has a significant influence on the conceptual prerequisite relationship. This feature aims to measure the depth of the prerequisite chain of a course, which is intended to reflect the level of complexity of the course. Concepts from high-complexity courses tend to be more advanced and usually depend on low-complexity course concepts. We use $Dep(d_i)$ to represent the prerequisite depth of course d_i, so the prerequisite depth of a concept can be defined as the average of the prerequisite depth of all courses that contain the concept.

$$Dep(c_a) = \frac{\sum_{d_i \in Set(c_a)} Dep(d_i)}{|Set(c_a)|}. \tag{3}$$

The feature is denoted as the "Prerequisite Depth distance" between two concepts, c_a and c_b, and is formally defined as follows:

$$Pdf(c_a, c_b) = Dep(c_a) - Dep(c_b). \tag{4}$$

This feature could capture the complexity of concepts through a common prerequisite transition pattern of multi-hop prerequisite relationships.

Contextual information contained in course descriptions also plays a crucial role in deducing prerequisite relationships among concepts. This can be achieved through the analysis of various factors such as the frequency of co-occurrence of concepts, and their relative order of appearance [13]. By analyzing such information, two features have been proposed as follows.

Concept Co-occurrence. This feature is based on the occurrence frequency of two concepts in one course which can be defined as $Ca(c_a, c_b)$. $Set(c_a)$ means all the courses which include concept c_a. $Ca(c_a, c_b) = \frac{|Set(c_a) \cap Set(c_b)|}{|Set(c_a)|}$ $Ca(c_a, c_b)$ represents the probability that c_b appears in the course description where c_a appears. When a new concept is introduced, the background information will also be mentioned. So when $Ca(c_a, c_b)$ is bigger and $Ca(c_b, c_a)$

is smaller, it indicates that the concept c_a is commonly found in courses that also contain the concept c_b. On the other hand, $Ca(c_b, c_a)$ being smaller suggests that there are relatively fewer course descriptions that include concept c_a while covering concept c_b.This implies that concept c_a is more likely to rely on concept c_b, as it is frequently encountered in courses that discuss or teach concept c_b.

The feature is defined as the difference of Ca between two concepts, as follows:

$$Caf(c_a, c_b) = Ca(c_a, c_b) - Ca(c_b, c_a). \tag{5}$$

Concept Order. This feature depends on the order in which concepts are presented in the course description. $order(c_a \leftarrow c_b)$ represents all the courses in which the position of concept c_b first appears before the first position of concept c_a. Therefore, $Co(c_a, c_b)$ is defined to measure how many courses with concept c_b precedes concept c_a: $Co(c_a, c_b) = \frac{|order(c_a \leftarrow c_b)|}{|Set(c_a) \cap Set(c_b)|}$. So the feature is defined as follows:

$$Cof(c_a, c_b) = Co(c_a, c_b) - Co(c_b, c_a). \tag{6}$$

These features illustrate the prerequisite relationships of concepts within the course using the frequency of concept co-occurrence and the order of concepts in the course description.

4.4 Concept fusion module

Graph Convolutional Network (GCN) is a type of neural network that can be used for graph-based data, such as a concept graph. Prior research has confirmed that GCN can effectively model complex transition patterns among nodes [18–21]. In this paper, we utilize GCN to capture the sequence of prerequisite transitions occurring between concepts. It applies a convolutional operation to the node feature vectors in the graph which combines information from neighboring nodes to update the feature vector of each node. After the convolutional operation, the GCN aggregates information from all the nodes in the graph to generate a new feature vector for each concept. This aggregation step enables the GCN to consider the global structure of the graph when updating the feature vectors. In the context of Graph Convolutional Networks (GCN), the goal is to learn the node representations $H = \{h_1, h_2, ..h_n\}$ in the hidden layers, given the node representation X and the adjacency matrix A. These updated feature vectors can be used as improved representations of the concepts, capturing both their local and global structure in the graph. The updated node representations produced by a GCN can be used as features in downstream tasks, such as predicting prerequisite relations between concepts. By incorporating information about the relationships between concepts, the GCN can produce more accurate representations of each node, which in turn can lead to more accurate predictions of prerequisite relations.

4.5 Prerequisite relation learning

After acquiring of final concept representations, a Siamese network is employed to determine whether there is a prerequisite relation between concept c_i and concept c_j. As seen in Figure 6, concept representations are input into two feed-forward networks with shared weights. The outputs are then joined together for classification. The general process can be described as follows:

$$c_i = ReLU(W_s \cdot h_{c_i}^L + b_s) \tag{7}$$

$$p(c_i, c_j) = \sigma(W^T[\mathbf{c_i}; \mathbf{c_j}; \mathbf{c_i} - \mathbf{c_j}; \mathbf{c_i} \otimes \mathbf{c_j}] + b), \tag{8}$$

where $h_{c_i}^L$ is the output of the GCN for concept c_i in the L-th layer, σ is the sigmoid function,

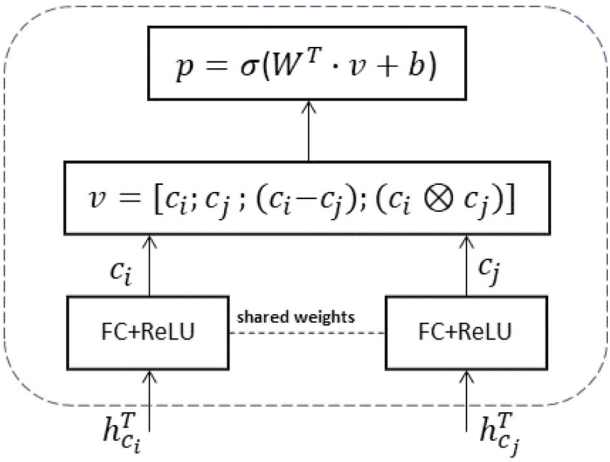

Figure 6 The Siamese network

$\otimes$ and - are the element-wise multiplication and subtraction operators, and $[\cdot;\cdot]$ represents the concatenation of vectors. Finally, the cross entropy function is used:

$$\mathcal{L}_c = \frac{1}{|T|} \sum_{(c_i,c_j)\in T} -[y_{ij} \log(p(c_i, c_j)) + (1 - y_{ij}) \log(1 - p(c_i, c_j))] \tag{9}$$

5 Experiments

To showcase the effectiveness and superiority of the proposed method, the study conducted extensive experiments and comparative analysis using two datasets: AuCM (Australian Course Map Data) and a university dataset from the USA.

5.1 Dataset

We used our dataset AuCM including course information and the corresponding course prerequisite dependencies. Unlike other datasets, most of which are established according to the curriculum of American universities, our dataset AuCM included course information from 14 Australian universities in the fields of Information Technology (IT) and Computer Science (CS). To the best of our knowledge, this is the first dataset consisting of Australian educational data for concept map generation. After removing duplicated courses and courses with identical codes, AuCM comprises a total of 738 distinct courses. Additionally, there are 1267 pairs of courses within AuCM that are related through precedence relationships. A total of 224 concepts were extracted using Wikipedia API, and 504 pairs of concepts were annotated as having prerequisite relations based on the extension of concept prerequisites in the dataset introduced in [12]. The dataset from [12] was also included in our analysis for comparative purposes. This dataset consists of 654 courses from 11 American universities, referred to as the "USA courses dataset", with a total of 861 pairs of courses that have precedence relationships. After removing duplicate concepts, we obtained a total of 290 unique concepts. Among these concepts, 681 pairs have a prerequisite relation.

5.2 Baselines

In the results of our experiment, we conducted an analysis and comparison of our proposed method with several typical concept prerequisite relations prediction baselines.

Binary Classifiers In the approach described in [9], binary classifiers were employed for the task at hand. Specifically, classifiers such as Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) were utilized. Among these classifiers, SVM and RF were selected based on their superior performance compared to the other two. To evaluate the effectiveness of the proposed model, a comparison was conducted between the model and these binary classifiers (SVM and RF).

MLP Siamese We also compared our CML-GCN model with typical neural network settings. We implemented a baseline Siamese neural network architecture to compare against our approach. The baseline network consisted of two identical neural networks, each with a single hidden layer. We used our proposed 5 features as the input feature vector of each concept for Siamese Network. The Siamese layers were followed by a final fully connected layer. We used a rectified linear unit (ReLU) activation function in each layer of the network, and the output is a binary similarity score as the prerequisite probability of concepts.

CPRL Concept prerequisite relation learning (CPRL) is a method under weakly supervision that achieves advanced outcomes. It utilized R-GCN to update the node representation of a heterogeneous graph and pairwise features for optimization. In accordance with numerous methodologies, we predominantly employed F-score (F1) to assess the performance of CML-GCN in comparison to all the baseline methods. Additionally, we compared precision (P) and recall (R) against other techniques.

5.3 Experimental settings and performance

We divided the concept prerequisite pairs into training and testing groups to evaluate our method's performance. The proportions were established as 80% and 20%.

To tackle the issue of imbalance, we employed an oversampling technique that increased the number of positive examples threefold, in order to acquire more training examples. In addition, we generated an equivalent number of negative instances by randomly selecting pairs of concepts from the list of concepts that did not belong to the original positive pairs. By doing so, we aimed to obtain more training examples and ensure a balanced training set that adequately represents both positive and negative classes. Regarding GCN, we established the number of GCN layers as L=2, with the embedding size for the first convolution layer set to 128 and for the second convolution layer set to 64. We conducted experiments with alternative configurations and discovered that minor adjustments did not significantly affect the outcomes.

Table 3 provides a comparison of the outcomes of various methods on AuCM and USA course datasets. The table shows the overall performance comparison of different methods in terms of precision, recall, and F1 score with a primary focus on the F1 score as it provides a balanced assessment of precision and recall. Our findings demonstrate that the proposed method almost outperforms all the other baseline methods on both datasets consistently. Our model, CML-GCN, consistently outperforms all other baseline methods on both datasets, as indicated by the F1 score. Specifically, we have observed a significant improvement in the F1 score of CML-GCN on the AuCM dataset compared to the best-performing baseline, surpassing the MLP Siamese model by 20.45%. Although the recall is similar to the CPRL model, the precision of CML-GCN demonstrates an improvement of 8.14% compared to

Table 3 Overall performance comparison in terms of Precision, Recall, and F-scores

Dataset	Metric	SVM	RF	MLP Siamese	CPRL	CML-GCN
AuCM	Precision	0.566	0.713	0.774	0.558	0.837
	Recall	0.615	0.692	0.679	0.898	0.895
	F1	0.590	0.702	0.714	0.689	0.860
USA Courses	Precision	0.723	0.774	0.853	0.689	0.765
	Recall	0.656	0.702	0.548	0.760	0.849
	F1	0.687	0.736	0.662	0.723	0.803

the MLP Siamese method. These compelling findings are visually represented in Figure 7, showcasing the superiority of CML-GCN in detecting prerequisite relations.

Furthermore, when considering the USA courses dataset, our proposed method achieves an F1 score and recall that outperform the best baselines, namely CPRL and RF methods, by 9.1% and 11.7%, respectively. These results further reinforce the exceptional performance of our approach in effectively identifying prerequisite relations in both the AuCM and USA datasets.

The remarkable performance of CML-GCN can be attributed to its ability to leverage the power of graph convolutional networks (GCN) in capturing the complex relationships between concepts. By integrating concept representation, GCN, and concept pairwise features, our model effectively exploits the contextual and structural information present in the course descriptions, resulting in superior performance compared to other methods.

In conclusion, our extensive experimentation and evaluation demonstrate that our proposed method, CML-GCN, outperforms the baseline methods on both the AuCM and USA datasets, providing accurate and reliable detection of prerequisite relations. These findings validate the effectiveness and superiority of our approach in concept map construction and showcase its potential for various educational applications.

5.4 Feature contribution analysis

To gain insight into the significance of each feature in our approach, an ablation study was conducted on the performance of the model based on AuCM. Specifically, each feature was

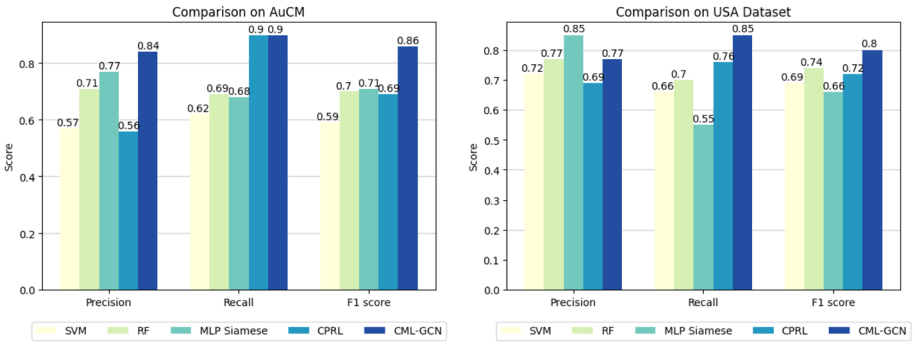

Figure 7 Overall performance comparison on AuCM and USA datasets

selectively removed from the model, and the resulting effect on the model's performance was evaluated. The goal was to assess how much each feature contributes to the effectiveness of the model. The evaluation results, as shown in Table 4, indicate the impact of ignoring each feature on the precision, recall, and F1-score of the model.

Among the features, Feature Coref, which compares the extent to which a concept belongs to a core course, was found to be particularly important in detecting prerequisite relations. Ignoring this feature resulted in a 5.4% decrease in the F1-score. This implies that considering the coreness of a concept in relation to the course it belongs to is crucial for the accurate prediction of prerequisites. Feature Cof, which captures the appearance orders of concepts, had the most significant impact on precision, leading to a 4.0% decrease in the F1-score when ignored. This highlights the importance of identifying the order in which concepts appear when determining prerequisite relationships. The model heavily relies on this feature to achieve higher precision in predicting prerequisites.

On the other hand, Feature Caf, which deals with the co-occurrence of concepts within the same course, had a relatively lower importance level compared to the other features. Ignoring this feature resulted in a 0.8% decrease in the F1 score. This suggests that while co-occurrence can be informative in some cases, there are many instances where two concepts may appear together in a course without having a prerequisite relationship. Hence, this feature has a lesser impact on the overall performance of the model.

Overall, the ablation study demonstrates that all the proposed features are useful in predicting prerequisite relations. Feature Coref and Feature Cof are particularly important, while Feature Caf has a relatively lower significance level. These findings provide valuable insights into the contribution of each feature and help in understanding the effectiveness of the model in predicting prerequisite relationships.

5.5 Effect of data size

In this section, we analyze the performance metrics of CML-GCN using various data sizes to check the stability and resilience of our model. Specifically, we randomly select 20% to 80% of the original AuCM dataset and USA dataset as our new datasets to train our model.

CML-GCN exhibits a degradation in performance with a decrease in training data size on both datasets, as indicated in Figure 8. In the case of the AuCM dataset, reducing the training data size from 80% to 40% results in only a marginal decline in performance. This decline is characterized by minimal fluctuation, indicating a stable and consistent decrease in performance. This indicates that AuCM dataset may contain sufficient and representative information, allowing the model to generalize well even with a smaller training subset. For the USA dataset, the performance of CML-GCN gradually decreases with a little fluctuation

Table 4 Feature contribution analysis

Feature	Precision	Recall	F1
CML-GCN	0.837	0.895	0.860
-Coref	0.798	0.812	0.806 (-5.4%)
-Famf	0.836	0.882	0.845 (-1.5%)
-Pdf	0.814	0.895	0.847 (-1.3%)
-Caf	0.830	0.867	0.852 (-0.8%)
-Cof	0.797	0.831	0.811 (-4.9%)

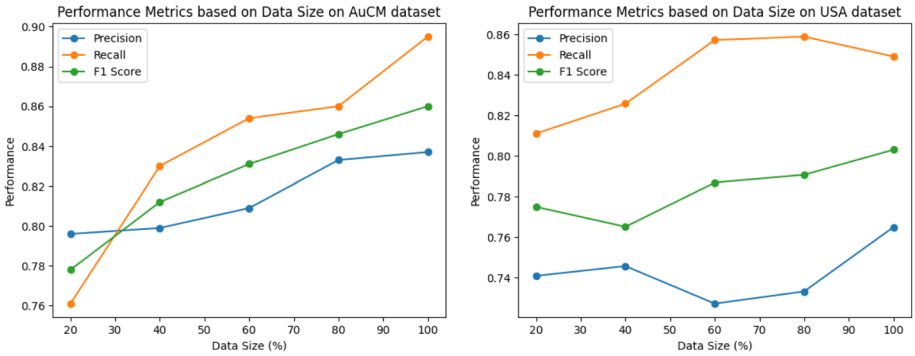

Figure 8 Performance w.r.t data size on AuCM dataset and USA dataset

as the data size drops from 80% to 20%. Despite the decrease in performance with a smaller training data size, CML-GCN still outperforms the best baseline MLP Siamese in terms of Precision and F1 Score on AuCM dataset which achieves a 2.84% higher Precision and an 8.96% higher F1 Score. This is possibly due to the fact that our model makes use of an initial concept map and a GCN network, which effectively employ information gathered from related concepts and the concept map's structure to classify prerequisite relationships across concepts more efficiently.

5.6 Case study

In order to further assess the effectiveness of our CML-GCN model, we performed a case study where we compared the concept prerequisite relations obtained from the ground truth with those learned by our CML-GCN model and a baseline model called MLP Siamese, which removes the GCN component. Specifically, we randomly selected five positive and five negative instances from the ground truth labels in the AuCM dataset. Subsequently, we obtained the corresponding concept prerequisite relations learned by our CML-GCN model and the MLP Siamese baseline. Table 5 displays the results of the comparison between concept prerequisite relations of ground truth (GT) and learned from both models, where a value of 1 indicates that concept a is a prerequisite of concept b.

The comparison in Table 5 demonstrates that both the CML-GCN and MLP Siamese models are able to learn the prerequisite relations between concepts in the majority of instances. The results obtained from the MLP Siamese model indicate that it performs less effectively compared to the CML-GCN model as it can accurately learn the concept prerequisite relation labels for three out of five positive concept pairs, while the CML-GCN model correctly learns four pairs. This proves that the GCN part in the CML model plays a crucial role in integrating semantic information from other concepts and the structural information of the initial concept map, which also can be seen from the negative instances. This integration of information significantly contributes to the enhancement of concept prerequisite relation classification accuracy. Hence, our proposed CML-GCN model successfully classifies concept prerequisite relations.

Table 5 The comparison of concept prerequisite relations from annotated labels and learned by the CML-GCN and MLP Siamese models on AuCM dataset

$Concept_a$	$Concept_b$	GT	CML-GCN	MLP Siamese
computer program	logic program	1	1	1
algorithm	computer graphic	1	1	1
discrete mathematics	graph theory	1	1	1
geometry	computer vision	1	1	0
asymptotic analysis	analysis of algorithm	1	0	0
information theory	computer science	0	0	0
synchronization	recursion	0	0	0
system program	pushdown automaton	0	0	1
python program language	memory management	0	0	0
boundary value problem	computer animation	0	0	0

6 Conclusion

We conducted a study to automatically determine prerequisite relationships between concepts extracted from courses description. We clearly defined the issue and proposed a number of practical features, including contextual, structural, and semantic features to construct the initial concept graph. To further enhance the quality of concept representations, Graph Convolutional Network (GCN) was used to aggregate neighbor information as it allowed for the incorporation of contextual and structural information into the representations, resulting in more accurate and informative semantic concept representations. We utilized a Siamese network to make predictions about prerequisite relationships between concepts. The approach was able to effectively identify relationships between concepts and gain insights into the hierarchical structure of courses. The effectiveness of the suggested model has been validated by experimental results on our AuCM dataset from IT/CS domains of Australian universities and USA courses dataset.

Promising potential directions include looking into how deep learning models can be used to automatically learn useful features for better prerequisite learning and how to use these prerequisite relations on other applications, such as students' performance prediction.

Author Contributions Xia and Li wrote the main manuscript text, Tang and Yang built the main framework of the program. All authors participated in programming and reviewed the manuscript

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data Availability The AuCM dataset in this paper can be obtained from the corresponding author upon a reasonable request

Declarations

Conflicts of interest The authors declare that they have no conflict of interest

Ethical Approval Not applicable

References

1. Jia, Y., Gu, Z., Jiang, Z., Gao, C., Yang, J.: Persistent graph stream summarization for real-time graph analytics. World Wide Web (2023)
2. Qi, Y., Gu, Z., Li, A., Zhang, X., Shafiq, M., Mei, Y., Lin, K.: Cybersecurity knowledge graph enabled attack chain detection for cyber-physical systems. Comput. Electrical Eng. **108**, 108660 (2023). https://doi.org/10.1016/j.compeleceng.2023.108660
3. Bi, X., Nie, H., Zhang, G., Hu, L., Ma, Y., Zhao, X., Yuan, Y., Wang, G.: Boosting question answering over knowledge graph with reward integration and policy evaluation under weak supervision. Inf. Process. Manag. **60**(2), 103242 (2023)
4. Song, X., Li, J., Cai, T., Yang, S., Yang, T., Liu, C.: A survey on deep learning based knowledge tracing. Knowl. Based Syst. **258**, 110036 (2022)
5. Xu, C., Guan, Z., Zhao, W., Wu, H., Niu, Y., Ling, B.: Adversarial incomplete multi-view clustering. In: Kraus, S. (ed.) Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, August 10-16, 2019, Macao, China, pp. 3933–3939 (2019)
6. Xu, C., Zhao, W., Zhao, J., Guan, Z., Song, X., Li, J.: Uncertainty-aware multiview deep learning for internet of things applications. IEEE Trans. Industrial Inf. **19**(2), 1456–1466 (2023). https://doi.org/10.1109/TII.2022.3206343
7. Yang, S., Verma, S., Cai, B., Jiang, J., Yu, K., Chen, F., Yu, S.: Variational coembedding learning for attributed network clustering. CoRR arXiv:2104.07295. (2021)
8. Liang, C.,Wu, Z., Huang, W., Giles, C.L.: Measuring prerequisite relations among concepts. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, September 17-21, 2015, pp. 1668–1674. The Association for Computational Linguistics, Lisbon, Portugal (2015). https://doi.org/10.18653/v1/d15-1193
9. Pan, L., Li, C., Li, J., Tang, J.: Prerequisite relation learning for concepts in moocs. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, July 30 - August 4, Volume 1: Long Papers, pp. 1447–1456. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1133
10. Wang, S., Ororbia, A.G., Wu, Z., Williams, K., Giles, C.L.: Using prerequisites to extract concept maps from textbooks. ACM (2016)
11. Yang, Y., Liu, H., Carbonell, J.G., Ma, W.: Concept graph learning from educational data. In: Proceedings of the Eighth ACM International Conference on WebSearch and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015, pp. 159–168 (2015)
12. Liang, C., Ye, J., Wu, Z., Pursel, B., Giles, C.L.: Recovering concept prerequisite relations from university course dependencies. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, pp. 4786–4791 (2017)
13. Xiao, K., Bai, Y., Wang, Z.: Extracting prerequisite relations among concepts from the course descriptions (SEKEEO-RN). Int. J. Softw. Eng. Knowl. Eng. **32**(4), 503–523 (2022). https://doi.org/10.1142/S0218194022400034
14. Lu, W., Zhou, Y., Yu, J., Jia, C.: Concept extraction and prerequisite relation learning from educational data. In: The Thirty-Third AAAI Conference on Artificial Intelligence, pp. 9678–9685. AAAI Press, Honolulu, Hawaii,USA (2019)
15. Roy, S., Madhyastha, M., Lawrence, S., Rajan, V.: Inferring concept prerequisite relations from online educational resources. In: The Thirty-Third AAAI Conference on Artificial Intelligence, pp. 9589–9594. AAAI Press, Honolulu, Hawaii,USA (2019). https://doi.org/10.1609/aaai.v33i01.33019589
16. Jia, C., Shen, Y., Tang, Y., Sun, L., Lu, W.: Heterogeneous graph neural networks for concept prerequisite relation learning in educational data. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2021)

17. Xiao, K., Bai, Y., Wang, S.: Mining precedence relations among lecture videos in moocs via concept prerequisite learning. MATHEMATICAL PROBLEMS IN ENGINEERING 2021 (2021)
18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings
19. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: The Thirty-Third AAAI Conference on Artificial Intelligence, pp. 7370–7377. AAAI Press, Honolulu, Hawaii, USA (2019). https://doi.org/10.1609/aaai.v33i01.33017370
20. Liu, J., Chen, Y., Huang, X., Li, J., Min, G.: Gnn-based long and short term preference modeling for next-location prediction. Inf. Sci. **629**, 1–14 (2023). https://doi.org/10.1016/j.ins.2023.01.131
21. Yang, S., Cai, B., Cai, T., Song, X., Jiang, J., Li, B., Li, J.: Robust cross-network node classification via constrained graph mutual information. Knowl. Based Syst. **257**, 109852 (2022)