Concept Extraction and Prerequisite Relation Learning from Educational Data

Weiming Lu, Yangfan Zhou, Jiale Yu, Chenhao Jia
College of Computer Science and Technology, Zhejiang University, Hangzhou, China
{luwm, 21621119, 21721115, 21821117}@zju.edu.cn

Abstract

Prerequisite relations among concepts are crucial for educational applications. However, it is difficult to automatically extract domain-specific concepts and learn the prerequisite relations among them without labeled data.

In this paper, we first extract high-quality phrases from a set of educational data, and identify the domain-specific concepts by a graph based ranking method. Then, we propose an iterative prerequisite relation learning framework, called iPRL, which combines a *learning based model* and *recovery based model* to leverage both concept pair features and dependencies among learning materials. In experiments, we evaluated our approach on two real-world datasets *Textbook Dataset* and *MOOC Dataset*, and validated that our approach can achieve better performance than existing methods. Finally, we also illustrate some examples of our approach.

Introduction

Domain-specific concepts and their prerequisite relations are the core of knowledge space and personalized learning theory (Falmagne et al. 2006). Prerequisite relations can be essentially considered as the dependency among concepts, and they are crucial for people to learn, organize, apply and generate knowledge (Margolis and Laurence 1999). For example, the prerequisite knowledge of ”Conditional Random Fields” is ”Hidden Markov Model”. Thus, organizing the knowledge in prerequisite relations can improve the educational tasks, such as curriculum planning (Liu et al. 2016) and intelligent tutoring (Wang and Liu 2016).

However, recent researches mainly focus on keyword extraction (Mihalcea and Tarau 2004; Liu, Chen, and Sun 2011; Tixier, Malliaros, and Vazirgiannis 2016) or high-quality phrase mining (Liu et al. 2015; Shang et al. 2018), and there are only a few efforts aiming to extract domain-specific concepts and learn prerequisite relations from educational data, such as courses (Yang et al. 2015; Liang et al. 2015; Liu et al. 2016; Liang et al. 2017), MOOCs (Massive Open Online Courses) (Pan et al. 2017a), textbooks (Wang et al. 2016; Liang et al. 2018) and scientific papers(Gordon et al. 2016).

The prerequisite relation learning methods can be classified into three categories: local statistical information based methods, learning based methods and recovery based methods. For the local statistical information, *reference distance* and *cross-entropy* were proposed (Liang et al. 2015; Gordon et al. 2016) to measure prerequisite relations among concepts. For the learning based methods, several features were proposed for the prerequisite classification. For example, Pan et al. (2017a) proposed contextual, structural and semantic features, and Liang et al. (2018) utilized graph-based and text-based features. In contrast, recovery based methods do not need to extract features to learn a prerequisite classifier. For example, Liang et al. (2017) can recover concept prerequisite relations from course dependencies directly.

However, there are still many challenges to automatically mine prerequisite relations among domain-specific concepts. (1) It is difficult to extract fine-grained concepts for each domain. (2) It is time-consuming to label the prerequisite relations in the learning based methods. (3) The recovery based methods only rely on the dependency among educational data, but neglect the contextual, structural and semantic features.

In order to address these challenges, we propose a domain-specific concept extraction approach, called **DsCE**, and an iterative prerequisite relation learning approach, called **iPRL**. DsCE can extract high-quality phrases from documents firstly, and then identifying domain-specific concepts with a graph based ranking method. While iPRL can utilize both the advantage of the *learning based model* and the *recovery based model*. It can learn the prerequisite relations among concepts without human labeled data, and improve the performance gradually with the interaction between two models.

We conduct extensive experiments on two real-world datasets: *Textbook* and *MOOC* datasets. The results show that our approach outperforms the state-of-the-art methods.

Our Approach

Problem Formulation

As shown in Figure 1, given a set of educational data in one domain, which is modeled as a set of sequential learning materials (denoted as LMs for short) such as textbook chapters and MOOC videos, we want to extract the domain-specific concepts and the prerequisite relations among them from these data.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

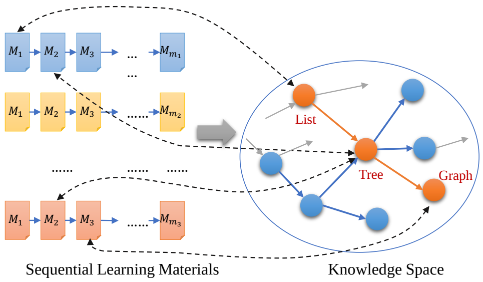


Figure 1: Illustration of Domain-specific Concept Extraction and Prerequisite Relation Learning from Sequential Learning Materials, which are denoted as $\{M_1, M_2, ..., M_{m_i}\}$.

For convenience, we list the main symbols used in this paper in Table 1.

Symbol	Meaning
$\mathcal{D}$	a set of educational data in one domain, and $\mathcal{D} = \{d_m\}_{m=1}^{\lvert\mathcal{D}\rvert}$.
$\mathcal{C}$	a set of concepts extracted from $\mathcal{D}$, and $\mathcal{C} = \{c_i\}_{i=1}^{\lvert\mathcal{C}\rvert}$.
$\mathcal{P}$	a set of high-quality phrases extracted from $\mathcal{D}$, and $\mathcal{P} = \{p_i\}_{i=1}^{\lvert\mathcal{P}\rvert}$.
$d_m \in \mathcal{D}$	an educational data with a sequential LMs, such as a book with a sequence of chapters or a MOOC course with a sequence of videos.
$Q(p_i)$	the quality score of $p_i \in \mathcal{P}$ of being a phrase.
$conf(p_i)$	the confidence score of $p_i \in \mathcal{P}$ of being a domain-specific concept.
$\mathbf{x}_i^m \in \mathbb{R}^{\lvert\mathcal{C}\rvert}$	the vector representation of the i^{th} LM $M_i^m \in d_m$ in concept space $\mathcal{C}$.
$\mapsto$	prerequisite relation between concepts or LMs.
Ω_m	a set of prerequisite relations among LMs, and $\Omega_m = \{i \mapsto j\}_{M_i^m, M_j^m \in d_m, M_i^m \mapsto M_j^m, d_m \in \mathcal{D}}$.
$\mathbf{A}$	$\mathbf{A} = (a_{i,j}) \in [-1,1]^{\lvert\mathcal{C}\rvert\times\lvert\mathcal{C}\rvert}$ represents prerequisite relations among concepts which is obtained from a recovery based model, and $a_{i,j}$ is the weight quantifying how concept i depends on concept j.
$\mathbf{F}$	$\mathbf{F} = (f_{i,j}) \in [-1,1]^{\lvert\mathcal{C}\rvert\times\lvert\mathcal{C}\rvert}$ represents the predicted results from a learning based model $\mathcal{F}$, and $f_{i,j}$ is the predicted result for concept i and concept j.
$\mathbf{W}$	$\mathbf{W} = (w_{i,j}) \in [0,1]^{\lvert\mathcal{C}\rvert\times\lvert\mathcal{C}\rvert}$ represents the relatedness among concepts.

Table 1: Meaning of symbols used

Therefore, the problem could be formally defined as: given a set of educational data in one domain, $\mathcal{D} = \{d_m\}_{m=1}^{M}$, and each data is a sequential learning materials $d_m = \{M_i^m\}_{i=1}^{\lvert d_m\rvert}$. The goal is to extract domain-specific concepts $\mathcal{C}$ from $\mathcal{D}$, and then learn the prerequisite relation matrix $\mathbf{A}$ among these concepts by an iterative prerequisite relation learning approach.

Domain-specific Concept Extraction

Domain-specific concepts should be identified before prerequisite relation learning.

However, it is time consuming and tedious to annotate all fine-grained concepts in each domain, and predefined part-of-speech rules such as $Noun^{+}Noun$ and $[Adj|Noun]^{+}Noun$ are unsuitable for the extraction task. Because some domain-specific concepts are complex, e.g., *non-homogeneous linear differential equation*, and *linear differential equation with constant coefficients of the second order* in the *Calculus* domain. In addition, not all phrases are domain-specific, e.g., *basic theory* is a good phrase, but it is not a domain-specific concept.

We developed an unsupervised and domain-independent approach (DsCE) to extract domain-specific concepts from textbooks, which consists of two steps: (1) extracting high-quality phrases from textbooks. (2) identifying domain-specific concepts from the phrases.

Extracting high-quality phrases A phrase is defined as a sequence of words that appear consecutively in the text, forming a complete semantic unit in certain contexts of the given documents (Finch 2000). A high-quality phrase is a phrase which has a high probability to be a complete semantic unit, meeting the criteria of *Popularity*, *Concordance*, *Informativeness* and *Completeness* (Liu et al. 2015; Shang et al. 2018).

Given the set of educational data $\mathcal{D}$ in one domain, we used *AutoPhrase* (Shang et al. 2018) to extract the high-quality phrases $\mathcal{P}$ to form the candidates of domain-specific concepts, since *AutoPhrase* has the following advantages: (1) It is a robust positive-only distant training method, which can be used in many different domains with the minimal human effort. (2) It can leverage part-of-speech (POS) tags to further enhance the extraction performance.

However, the high-quality phrases extracted by *AutoPhrase* may not be domain-specific. We extracted phrases from textbooks in three domains: "Calculus"(**CAL**), "Data Structure"(**DS**), and "Physics"(**PHY**), and calculated the average ratio of the domain-specific concepts in top K phrases according to its quality score. The result is shown in Figure 2, which demonstrates that high-quality phrases may not be domain-specific, and has many noise in the top K phrases. Therefore, we should identify the domain-specific concepts from these high-quality phrases.

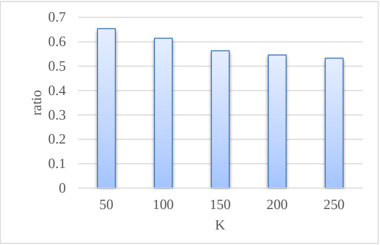


Figure 2: The ratio of the domain-specific concepts in top K phrases of the results of *AutoPhrase*.

Identifying domain-specific concepts In order to identify domain-specific concepts in the high-quality phrases, we resorted to a graph propagation based ranking method (Pan et al. 2017b).

First, a weighted undirected fully-connected graph $\mathcal{G} = (V, E)$ was constructed, where V is the vertex set of $\mathcal{G}$ and

E is the edge set of $\mathcal{G}$. Each vertex in V is a phrase $p_i \in \mathcal{P}$ with a quality score $Q(p_i)$ extracted by *AutoPhrase*. For each edge $(p_i, p_j) \in E$, its weight $w(p_i, p_j)$ is the semantic relatedness of the two phrases p_i and p_j. We trained word embedding vectors based an online encyclopedia corpus via word2vec (Mikolov et al. 2013b), and then obtained the semantic representation of each phrase via the vector addition of its individual word vectors. Then, the semantic relatedness of two phrases is defined as the cosine similarity of their vectors.

Second, a propagation based ranking method was executed on the graph $\mathcal{G}$ with the assumption that high-confidence concepts in the graph could propagate their confidence scores to their neighbor nodes that have high semantic relatedness with them, to discover other potential domain-specific concepts. Each vertex p_i has a confidence score $conf(p_i)$ of being a domain-specific concept, and $conf^k(p_i)$ is the score of p_i in the k-th iteration of the propagation. We set the initial confidence score of each vertex as $conf^0(p_i) = 1$. The propagation functions are defined as:

$$\begin{aligned} s^k(p_j, p_i) &= opf(p_i, p_j) \cdot Q(p_j) \cdot e(p_i, p_j) \cdot conf^k(p_j) \\ conf^{k+1}(p_i) &= \frac{1}{Z}\left(\frac{\sum_{p_j \in A(p_i)} s^k(p_j, p_i)}{|A(p_i)|}\right) \end{aligned}$$

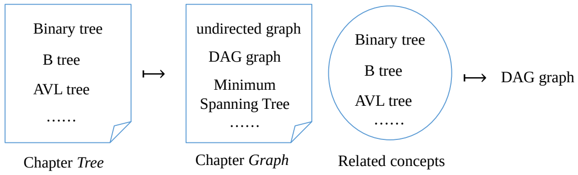
where $s^k(p_j, p_i)$ is the voting score that p_j propagates to p_i in the k-th iteration, which is determined by the semantic relatedness between p_i and p_j $e(p_i, p_j)$, the quality score of p_j $Q(p_j)$, the confidence score of p_j in the k-th iteration $conf^k(p_j)$, and the overlapping penalty between p_i and p_j $opf(p_i, p_j)$. If p_i and p_j contain one or more identical words, we say they are overlapping, and should be penalized during the score propagation. For example, *this algorithm* and *sort algorithm* are related by cosine similarity, but *sort algorithm* should not propagate its confidence score to *this algorithm*, since *this algorithm* is not a domain-specific concept. We set $opf(p_i, p_j) = 1$, if p_i and p_j are not overlapping. Otherwise, $opf(p_i, p_j) = \lambda$, $\lambda \in (0, 1)$ (We set $\lambda = 0.5$ in the experiments). $conf^{k+1}(p_i)$ is the new confidence score of p_i, which is dependent on the average voting score of vertexes in $A(p_i)$. $A(p_i)$ is the vertex set which will propagate the voting scores to p_i in each iteration. After each iteration, the confidence scores should be normalized, so Z is the normalization factor.

Finally, the phrases with the confidence score greater than $\phi = 0.6$ are selected as the domain-specific concepts $\mathcal{C}$.

Prerequisite Relation Learning

After obtained the domain-specific concepts $\mathcal{C}$, we would like to learn the prerequisite relations among these concepts.

Inspired by Liang et al. (2017), we observed that (1) The dependency among learning materials is caused by sufficient evidence provided by prerequisite relations among concepts in that materials. This means if one learning material depends on another one for knowledge learning, there must be sufficient concept pairs between them which have prerequisite relations. Taking Figure 3(a) as an example, because chapter *Tree* ↦ *Graph*, then the concepts in these chapters would like to have prerequisite relations, such as *binary tree* ↦ *undirected graph* and *B tree* ↦ *DAG graph*, where ↦ means there is a prerequisite relation between LMs or concepts.



(a) Observation 1 (b) Observation 2

Figure 3: The observations of the prerequisite relations among concepts.

(2) Related concepts would have similar prerequisite relations with other concepts. As in Figure 3(b), *binary tree* and *B tree* are highly related, so if *binary tree* ↦ *DAG graph*, then *B tree* ↦ *DAG graph*.

(3) The results of the *recovery based model* and the *learning based model* should be consistent. That is, it would be better that $\mathbf{A}$ equals to $\mathbf{F}$, where $\mathbf{A}$ is the result of the *recovery based model* and $\mathbf{F}$ is the result of the *learning based model*.

Thus, the optimization function is designed as follows.

$$\begin{aligned} \min_{\mathbf{A},\xi,\mathbf{F}} \quad & ||\xi||_2 + \lambda_1 \frac{||\mathbf{A}-\mathbf{F}||_F}{|\mathcal{C}|^2} + \\ & \lambda_2 \sum_{1 \le i \neq j \le |\mathcal{C}|} w_{i,j} \frac{||\mathbf{A}(i,:) - \mathbf{A}(j,:)||_2}{|\mathcal{C}|} \\ s.t. \quad & (\mathbf{x}_i^m)^T [\mathbf{C}_{i \mapsto j}^m \odot \mathbf{A}] \mathbf{x}_j^m \ge \theta - \xi_{i,j}^m, \\ & \forall d_m \in \mathcal{D}, (i,j) \in \Omega_m && (1) \\ & \mathbf{A}_{i,j} + \mathbf{A}_{j,i} = 0, 0 < i \neq j < |\mathcal{C}| && (2) \\ & -1 \le \mathbf{A}_{i,j} \le 1, 0 < i \neq j < |\mathcal{C}| && (3) \\ & \mathbf{A} \odot (\mathbf{1} - \mathbf{D}) = \mathbf{0} && (4) \\ & \mathbf{A}_{i,j} > 0, \forall d_m \in \mathcal{D}, pos_m(i) < pos_m(j) && (5) \end{aligned}$$

where $\xi_{i,j}^m$ is a slack parameter for a pair of LMs $M_i^m \mapsto M_j^m$ in $d_m \in \mathcal{D}$ which has a prerequisite relation. θ, λ_1 and λ_2 are positive hyper parameters, and they are set to 1.0, 0.2 and 0.2 in the experiments respectively.

For the objective function, the first term is the empirical loss as in Liang et al. (2017). The second term is based on the third observation, which makes $\mathbf{A}$ be close to $\mathbf{F}$. The third term is based on the second observation, which means if concept i and j are related, $\mathbf{A}(i,:)$ and $\mathbf{A}(j,:)$ should be similar. Here, $w_{i,j}$ is the relatedness between concept i and j, which would be computed in advance. How to calculate $\mathbf{F}$ and $\mathbf{W}$ will be elaborated in the following section.

For the constraints, the first three constraints are learned from Liang et al. (2017). The first constraint is based on the first observation mentioned above, where $\odot$ is element-wise product, and $\mathbf{C}_{i \mapsto j}^m$ is used to remove the contribution from the common concepts between M_i^m and M_j^m, so $\mathbf{C}_{i \mapsto j}^m = (c_{s,t}^m) \in \{0,1\}^{|\mathcal{C}| \times |\mathcal{C}|}$, and $c_{s,t}^m = 0$ if $\mathbf{x}_i^m[s] > 0$ or $\mathbf{x}_j^m[t] > 0$, otherwise $c_{s,t}^m = 1$, and $\xi_{i,j}^m$ is the penalty of this constraint. The second constraint means if concept i

is a prerequisite of j, then j is not a prerequisite of i. The third constraint bounds the strength of prerequisite relation in $[-1, 1]$.

The fourth constraint assumes that there is no prerequisite relation between two concepts if one concept is not contained in other's encyclopedia article according to Talukdar and Cohen (2012). Here, $\mathbf{1} = \{1\}^{|\mathcal{C}|\times|\mathcal{C}|}$, and $\mathbf{D} \in \{0,1\}^{|\mathcal{C}|\times|\mathcal{C}|}$ is introduced to filter concept pairs. When the article of concept i (j) contains concept j (i), $\mathbf{D}_{i,j} = 1$, otherwise $\mathbf{D}_{i,j} = 0$. The last constraint specifies that if concept i always occurs before j, then i may be the prerequisite concept of j, but j must not be the prerequisite concept of i. Here, $pos_m(i)$ is the position of concept i in d_m.

The optimization problem is solved iteratively. If we don't have any labeled data, **F** is initialized randomly, otherwise a probabilistic classifier $\mathcal{F}$ is trained to predict **F**. Then, by fixing **F**, the problem becomes a quadratic optimization problem, and **A** and ξ are solved by MOSEK[1] software. When obtaining a new **A**, new training data can be selected to enhance the classifier $\mathcal{F}$. In our experiments, the performance of our model increases gradually and tends to converge.

Specifically, during the k^{th} iterations, three datasets are formed firstly: $P^k = \{(i,j)|\mathbf{A}_{i,j} > \tau\}$; $N^k = \{(i,j)||\mathbf{A}_{i,j}| < \tau\}$; $O^k = \{(i,j)|\mathbf{A}_{i,j} == 0\}$. Then, the positive samples is formed by selecting the top $(\frac{k}{k+1})^2|P^k|$ concept pairs from P^k. The negative samples also have the same number of the positive samples, but they consists of three parts: 20% data is randomly selected from $\{(j,i)|(i,j) \in P^k\}$, and the remaining 80% data is randomly selected from N^k and O^k equally. $\tau = 0.6$ in our experiments. The procedure is shown in the Algorithm 1.

```
Input: $\mathcal{D}$: a set of educational data in a domain.
$\mathcal{C}$: a set of concepts extracted from $\mathcal{D}$ by DsCE.
$\mathbf{x}$: the vector representations of all learning materials in $\mathcal{D}$.
Output: $\mathbf{A}$: the prerequisite relations among concepts in $\mathcal{C}$.
Initialize $\mathbf{F}$ randomly, or train a probabilistic classifier $\mathcal{F}$
  using labeled data to initialize $\mathbf{F}$;
Calculate the concept relatedness matrix $\mathbf{W}$ according to $\mathcal{D}$;
Obtain $\mathbf{A}$ by solving Eq. 1 with MOSEK;
while $A$ is not convergent do
    Update training data for $\mathcal{F}$ according to $\mathbf{A}$; (Update the
      Learning based Model)
    Update $\mathbf{F}$ using $\mathcal{F}$;
    Fixing $\mathbf{F}$, obtain new $\mathbf{A}$ by solving Eq. 1; (Update the
      Recovery based Model)
end
```

Algorithm 1: Iterative Prerequisite Relation Learning.

Concept Features and Relatedness

To train $\mathcal{F}$, we used several features to capture whether a concept pair has a prerequisite relation.

For MOOCs, in order to compare the method in (Pan et al. 2017a), we still used the same features, including semantic relatedness, video reference distance (refd), sentence refd, Wikipedia refd, average position distance, distributional asymmetry distance and complexity level distance.

For textbooks, in addition to the features mentioned above, where we replaced video refd and sentence refd with chapter refd, we also added Wikipedia abstract occurrence and content refd.

chapter refd: For a concept pair (a, b), chapter reference weight (crw), which is to qualify how b is referred by chapters of a, is defined as: $crw(a,b) = \frac{\sum_{d_m\in\mathcal{D}}\sum_{M\in d_m} f(a,M)\cdot r(M,b)}{\sum_{d_m\in\mathcal{D}}\sum_{M\in d_m} f(a,M)}$, where $f(a, M)$ indicates the term frequency of concept a in chapter M, and $r(M, b) \in \{0,1\}$ denotes whether concept b appears in chapter M. Then, *chapter refd* of (a, b) is $crw(b, a) - crw(a, b)$.

Wikipedia abstract occurrence: For a concept pair (a, b), if concept a occurs in the Wikipedia abstract of concept b, *Wikipedia abstract occurrence* of (a, b) is 1, otherwise 0.

Wikipedia content refd: For a concept pair (a, b), the *Wikipedia content refd* is defined as: $f_b(a) - f_a(b)$, where $f_b(a)$ is the term frequency of concept a in the article of b.

To define **W**, we proposed two relatedness measurements between concept i and j.

Semantic Relatedness: $\mathbf{W}_{i,j}$ is defined as the normalized cosine distance between two embedding vectors: $\mathbf{W}^s_{i,j} = \frac{1}{2}(1 + \frac{v_i\cdot v_j}{||v_i||\cdot||v_j||})$, where v_i is the word embedding vector learned by word2vec (Mikolov et al. 2013a).

Position Relatedness: if two concepts have the similar positions in textbooks or MOOC videos, they would have similar prerequisite relations with other concepts. Thus, $\mathbf{W}_{i,j}$ is defined as: $\mathbf{W}^p_{i,j} = \frac{1}{2}\sum_{d_m\in\mathcal{D}}\sum_{M\in d_m}(p^M_i + p^M_j)|p^M_i - p^M_j|$, where $p^{M\in d_m}_i = \frac{n(i,M)}{n(i,d_m)}$, $n(i, M)$ is the term frequency of concept i in the material M.

Experiments

In this section, we will evaluate the domain-specific concept extraction and prerequisite learning respectively.

Evaluation on Domain-specific Concept Extraction

Datasets In order to evaluate the domain-specific concept extraction, we collected six Chinese textbooks in each domain: "Calculus"(**CAL**), "Data Structure"(**DS**), and "Physics"(**PHY**) from our digital library, and then extracted the content of the textbooks for concept extraction.

Baseline Methods We employed the following three baseline methods to compare with our DsCE method.

TextRank (Mihalcea and Tarau 2004): An unsupervised graph-based ranking model for keyword extraction.

THUCKE (Liu, Chen, and Sun 2011)[2]: A word trigger method for keyword extraction based on word alignment in statistical machine translation.

AutoPhrase (Shang et al. 2018)[3]: An automated phrase mining method with POS-guided phrasal segmentation.

[1] http://www.mosek.com

[2] https://github.com/thunlp/THUCKE

[3] https://github.com/shangjingbo1226/AutoPhrase

Performance comparison and analysis To measure performance, we asked three students who are majoring in the corresponding domain to annotate whether the concepts extracted from each method are domain-specific or not, and then took a majority vote of the annotations to create final domain-specific concept set for each domain. *Precision* and *Recall* are used to evaluate the extraction results.

In Figure 4, we show the precision of domain-specific concept extraction in top K results of different methods in the three domains.

We find that our method (DsCE) outperforms baseline methods across all domains. Specifically, we have the following observations. First, *AutoPhrase* can extract high-quality phrases, but it does not focus on keyword extraction, so the top ranked phrases may not be domain-specific. Second, the precision in *Calculus* domain is relatively low comparing to other two domains. This is because there are many mathematical formulas in the textbooks of *Calculus*, which affect the OCR (Optical Character Recognition) quality for textbooks. Third, the graph propagation based ranking method can indeed improve the performance of *AutoPhrase* for domain-specific concept extraction.

In addition, we also calculate the average precision and recall for each method across all domains, and show the precision-recall curves in Figure 5. Obviously, we find that our method outperforms baseline methods greatly.

Figure 6 illustrates some examples of the top ranked domain-specific concepts extracted for each domain. We find that (1) *TextRank* and *THUCKE* are unable to extract complex concepts, which makes them have low recall. (2) Some top-ranked concepts extracted by *AutoPhrase* are obscure, such as *michelson interferometer* and *Nicol prism*, while the basic core concepts are not ranked in the top.

Evaluation on Prerequisite Relation Learning

Datasets We evaluated the prerequisite relation learning with the following two different datasets.

Textbook: To the best of our knowledge, these is no public dataset for mining prerequisite relation between concepts in textbooks. We also chose Chinese textbooks in three domains: *Calculus*, *Data Structure*, and *Physics* to create the dataset. The construction procedure is as follows. For each domain, we extracted concepts from textbooks by our DsCE method, and asked three students to select the most related domain concepts. Finally, the prerequisite relation among them were annotated through a majority vote strategy.

MOOC: We used MOOC data mentioned in (Pan et al. 2017a), which has been released[4].

The statistics of the datasets are listed in Table 2, which lists the number of concepts and the positive pairs in each domain.

Textbook			
dataset	#books	#concepts	#pairs(+)
CAL	6	89	439
DS	6	90	453
PHY	6	139	630
MOOC			
dataset	#courses	#concept	#pairs(+)
ML	5	244	1,735
DSA	8	201	1,148

Table 2: Statistics of the datasets for evaluating on prerequisite relation learning.

Baseline Methods We use the following state-of-the-art methods as baselines:

RefD (reference distance) (Liang et al. 2015): A link-based metric for measuring the prerequisite relations among concepts.

CPR-Recover (Liang et al. 2017): An unsupervised method, which can recover concept prerequisite relations from course dependencies. In our datasets, we used it to recover concept prerequisite relations from book chapter or MOOC video dependencies.

PRinMOOC (Pan et al. 2017a): A supervised method for prerequisite relation learning in MOOC, which proposed contextual, structural and semantic features, and then used random forest to capture the prerequisite relations between concepts.

In the experiments, we trained the English and Chinese word vectors by word2vec (Mikolov et al. 2013a) on English Wikipedia and Baidu Baike[5] respectively, and the dimensions are both set to 100.

Performance comparison and analysis We evaluated the methods under several scenarios, including with non-labeled data, with labeled data, different concept relatedness and iterative learning.

(1) **Evaluation with non-labeled data**. We compared **iPRL** with **RefD** and **CPR-Recover** without any labeled data in both *Textbook* and *MOOC* datasets. The results are shown in Table 3.

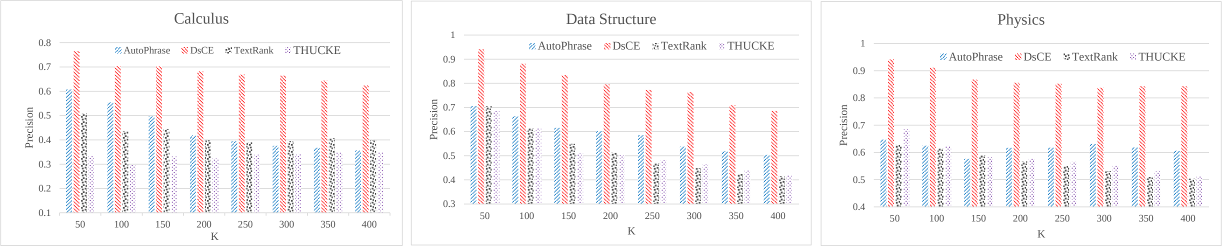
Method		*Textbook*			*MOOC*	
		CAL	**DS**	**PHY**	**ML**	**DSA**
RefD	P	35.49	34.78	27.96	71.70	72.69
	R	57.40	44.54	50.32	30.4	37.8
	F_1	43.86	39.06	35.94	42.70	49.74
CPR-R	P	44.69	31.70	29.07	57.79	55.94
	R	62.41	57.68	60.89	35.6	35.8
	F_1	52.09	40.91	39.35	44.06	43.66
iPRL	P	91.38	58.86	71.69	65.72	72.48
	R	55.58	64.80	57.85	46.4	43.2
	F_1	**69.12**	**61.69**	**64.06**	**54.40**	**54.14**

Table 3: Comparison with non-labeled data(%) on *Textbook* and *MOOC* datasets

From the tables, we can see that (i) **iPRL** outperforms other methods under the non-labeled data situation, which achieves $+16.69\%$ for **CPR-Recover** and $+18.44\%$ for **RefD** with respect to the average F_1. (ii) **CPR-Recover** can achieve a comparable *recall*, but its *precision* is too low.

[4] http://keg.cs.tsinghua.edu.cn/jietang/software/acl17-prerequisite-relation.rar

[5] https://baike.baidu.com/



(a) Calculus (b) Data Structure (c) Physics

Figure 4: The precision of domain-specific concept extraction of different methods in three domains.

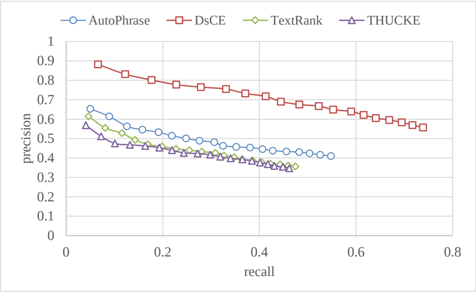


Figure 5: The precision-recall curves of domain-specific concept extraction

(2) **Evaluation with labeled data**. For the labeled data, we applied 5-fold cross validation to evaluate the performance of the proposed method,i.e., using 4 folds in methods such as **PRinMOOC** and **iPRL**, and then testing them with the remaining fold. We also selected random forest as $\mathcal{F}$ as in (Pan et al. 2017a). The comparison between **PRinMOOC** and **iPRL** is shown in Table 4.

Method	measure	**ML**	**DSA**
PRinMOOC	P	72.08	88.32
	R	59.4	48.4
	F_1	65.13	62.53
iPRL	P	65.96	70.93
	R	68.6	74.2
	F_1	**67.25**	**72.53**

Table 4: Comparison with labeled data(%) on *MOOC* dataset

From the table, we can see that: (i) **iPRL** outperforms **PRinMOOC**, which only has the *learning model*. (ii) Labeled data can promote **iPRL**. Comparing Table 4 with Table 3, the average F_1 increases by 12.85% and 18.39% on **ML** and **DSA** datasets.

(3) **Evaluation with concept relatedness**. We evaluated both *semantic relatedness* and *position relatedness* on *Textbook* dataset in **iPRL**, and the result is shown in Table 5. From the table, we can see that (i) Both *semantic relatedness* and *position relatedness* achieve a better performance. (ii) *Position relatedness* outperforms *semantic relatedness* slightly. The position of concepts may be more suitable for deducing prerequisite relations.

TextRank	THUCKE	AutoPhrase	DsCE
		Calculus	
函数 function	函数 function	最小正周期 minimal positive period	微分中值定理 differential mean value theorem
方程 function	极限 limit	常系数线性微分方程 linear differential equation with constant coefficients	柯西中值定理 Cauchy mean value theorem
级数 series	方程 function	分部积分法 integration by parts	微分公式 differential formula
		Data Structure	
算法 algorithm	数据 data	索引顺序文件 indexed sequential file	线性链表 linear linked list
节点 node	存储 storage	数组元素 array element	数组下标 array index
查找 search	算法 algorithm	空指针 null pointer	指针变量 pointer variable
		Physics	
运动 movement	运动 movement	迈克耳孙干涉仪 michelson interferometer	电场强度 electric field strength
物体 object	速度 velocity	质心运动定理 theorem of the motion for center of mass	涡旋电场 vortex electric field
速度 velocity	过程 procedure	尼科耳棱镜 Nicol prism	电场能量 field energy

Figure 6: Examples of domain-specific concepts extracted for each domain with different methods, where the English phrases under the Chinese phrases are the translations.

(4) **Evaluation with iterative learning**. Figure 7 shows the performance of **iPRL**^{p} on *Textbook* dataset varies with the iterations. The interaction between *Learning based Model* and *Recovery based Model* in **iPRL** can really make the performance better gradually.

Finally, we shows some examples of prerequisite relations among concepts extracted by **iPRL** in Figure 8.

Related Work

Domain-specific Concept Extraction

Keyword extraction is the most related task to domain-specific concept extraction. There are several approaches to extract keywords. Supervised learning methods usually extract syntactic and lexical features for keyword extraction. For example, KEA (Witten et al. 1999) proposed two features: *TFIDF* and *first occurrence*, and then used a clas-

Method	measure	**CAL**	**DS**	**PHY**
iPRL	P	91.38	58.86	71.69
	R	55.58	64.80	57.85
	F_1	69.12	61.69	64.06
iPRL^s	P	92.96	59.16	69.70
	R	54.21	66.14	60.09
	F_1	68.48	62.46	64.54
iPRL^p	P	92.48	60.16	71.12
	R	56.03	65.25	59.61
	F_1	**69.78**	**62.60**	**64.86**

Table 5: Performance of concept relatednesss(%) on *Textbook* dataset, where ^{s} and ^{p} denote the *semantic relatedness* and *position relatedness*

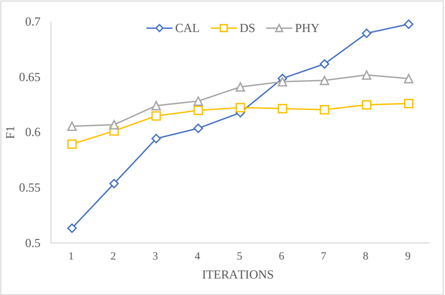


Figure 7: Performance of iPRL^p on *Textbook* varies with the iterations

sifier to determine the keywords. Recently, an encoder-decoder framework (Meng et al. 2017) was applied to generate keyphrase. Unsupervised learning methods usually apply graph-based semantic relatedness measures for keyword extraction. For example, TextRank (Mihalcea and Tarau 2004) represented a document as a term graph, and used a graph-based ranking algorithm to assign importance scores to terms. Tixier, Malliaros, and Vazirgiannis (2016) proposed a graph degeneracy-based approach for keyword extraction.

However, these researches mainly focused on the selection of words that can best describe the document, but ignore the fact that domain-specific concepts are usually complex and fine-grained.

Recently, several works (Liu et al. 2015; Shang et al. 2018) focus on how to extract high-quality phrases. SegPhrase (Liu et al. 2015) proposed a segmentation-integrated framework by integrating phrase extraction and phrasal segmentation to mutually benefit each other. AutoPhrase (Shang et al. 2018) extended SegPhrase by introducing two new technologies: robust positive-only distant training, and POS-guided phrasal segmentation.

Prerequisite Relation Learning

A few efforts aim to learn prerequisite relations from educational data, such as courses (Yang et al. 2015; Liang et al. 2015; Liu et al. 2016; Liang et al. 2017), MOOCs (Pan et al. 2017a), textbooks (Wang et al. 2016; Liang et al. 2018) and scientific papers(Gordon et al. 2016).

The statistical information such as *reference distance* and *cross-entropy* (Liang et al. 2015; Gordon et al. 2016) was proposed to measure the prerequisite relations among concepts in courses or scientific papers. While Pan et al. (2017a) proposed semantic, contextual and structural features to detect prerequisite relations among concepts in MOOCs.

Calculus
函数(function) ↦ 切线(tangent) ↦ 微分(differential)↦ 中值定理(mean value theorem)↦ 拉格朗日中值定理(Lagrange mean value theorem) 函数(function)↦导数(derivative)↦微分方程(differential equation)↦一阶线性微分方程(linear first-order differential equation)
Data Structure

数据(data)↦树(tree)↦二叉树(binary tree)↦二叉搜索树(binary search tree)↦平衡二叉树(balanced binary tree)
图(graph)↦结点(node)↦最小生成树(minimal spanning tree)↦克鲁斯卡尔算法(Kruskal's algorithm)

Physics
速度(velocity)↦动能(kinetic energy)↦机械能(mechanical energy)↦机械能守恒定律(conservation law of mechanical energy) 电荷(charge)↦电流(electricity)↦电磁感应(electromagnetic induction)↦电磁感应定律(law of electromagnetic induction)

Figure 8: Examples of prerequisite relations extracted by **iPRL**.

Other works utilized the dependency among courses or textbooks. For example, (Yang et al. 2015; Liu et al. 2016) proposed a concept graph learning framework for within- and cross-level inference of prerequisite relations at the course-level and the concept-level directed graphs. (Liang et al. 2017) addressed the problem of recovering concept prerequisite relations from university course dependencies. (Wang et al. 2016) proposed a joint optimization model for concept map extraction from textbooks that utilizes the mutual interdependency between concept extraction and prerequisite relation learning. They further applied active learning to the concept prerequisite learning (Liang et al. 2018).

Conclusion

In this paper, we propose a domain-specific concept extraction approach and an iterative prerequisite relation learning approach. This approach can extract domain-specific concepts and learn the prerequisite relations without human labeled data. In experiments, we evaluated our approach on two real-world datasets *Textbook Dataset* and *MOOC Dataset*, and validated that our approach can achieve better performance than existing methods. In addition, we also illustrated some examples of the results of our approach. In future, we plan to use deep learning and active learning technologies to further improve the performance.

Acknowledgments

This work is supported by the Zhejiang Provincial Natural Science Foundation of China (No. LY17F020015), the Fundamental Research Funds for the Central Universities (No. 2017FZA5016), CKCEST, and MOE-Engineering Research Center of Digital Library.

References

Falmagne, J.-C.; Cosyn, E.; Doignon, J.-P.; and Thiéry, N. 2006. The assessment of knowledge, in theory and in practice. In Missaoui, R., and Schmidt, J., eds., *Formal Concept Analysis*, 61–79. Berlin, Heidelberg: Springer Berlin Heidelberg.

Finch, G. 2000. *Linguistic Terms and Concepts*. How to study. Macmillan.

Gordon, J.; Zhu, L.; Galstyan, A.; Natarajan, P.; and Burns, G. 2016. Modeling concept dependencies in a scientific corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 866–875.

Liang, C.; Wu, Z.; Huang, W.; and Giles, C. L. 2015. Measuring prerequisite relations among concepts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1668–1674.

Liang, C.; Ye, J.; Wu, Z.; Pursel, B.; and Giles, C. L. 2017. Recovering concept prerequisite relations from university course dependencies. In *AAAI*, 4786–4791.

Liang, C.; Ye, J.; Wang, S.; Pursel, B.; and Giles, C. L. 2018. Investigating active learning for concept prerequisite learning. *Proc. EAAI*.

Liu, J.; Shang, J.; Wang, C.; Ren, X.; and Han, J. 2015. Mining quality phrases from massive text corpora. *Proceedings. ACM-Sigmod International Conference on Management of Data* 2015:1729–1744.

Liu, H.; Ma, W.; Yang, Y.; and Carbonell, J. 2016. Learning concept graphs from online educational data. *Journal of Artificial Intelligence Research* 55:1059–1090.

Liu, Z.; Chen, X.; and Sun, M. 2011. A simple word trigger method for social tag suggestion. In *EMNLP*.

Margolis, E., and Laurence, S. 1999. *Concepts: core readings*. Mit Press.

Meng, R.; Zhao, S.; Han, S.; He, D.; Brusilovsky, P.; and Chi, Y. 2017. Deep keyphrase generation. In *ACL*.

Mihalcea, R., and Tarau, P. 2004. Textrank: Bringing order into texts.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Pan, L.; Li, C.; Li, J.; and Tang, J. 2017a. Prerequisite relation learning for concepts in moocs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1447–1456.

Pan, L.; Wang, X.; Li, C.; Li, J.-Z.; and Tang, J. 2017b. Course concept extraction in moocs via embedding-based graph propagation. In *IJCNLP*.

Shang, J.; Liu, J.; Jiang, M.; Ren, X.; Voss, C. R.; and Han, J. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering* 1–1.

Talukdar, P. P., and Cohen, W. W. 2012. Crowdsourced comprehension: predicting prerequisite structure in wikipedia. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 307–315. Association for Computational Linguistics.

Tixier, A. J.-P.; Malliaros, F. D.; and Vazirgiannis, M. 2016. A graph degeneracy-based approach to keyword extraction. In *EMNLP*.

Wang, S., and Liu, L. 2016. Prerequisite concept maps extraction for automatic assessment. In *Proceedings of the 25th International Conference Companion on World Wide Web*, 519–521. International World Wide Web Conferences Steering Committee.

Wang, S.; Ororbia, A.; Wu, Z.; Williams, K.; Liang, C.; Pursel, B.; and Giles, C. L. 2016. Using prerequisites to extract concept maps from textbooks. In *Proceedings of the 25th acm international on conference on information and knowledge management*, 317–326. ACM.

Witten, I. H.; Paynter, G. W.; Frank, E.; Gutwin, C.; and Nevill-Manning, C. G. 1999. Kea: Practical automatic keyphrase extraction. In *ACM DL*.

Yang, Y.; Liu, H.; Carbonell, J.; and Ma, W. 2015. Concept graph learning from educational data. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 159–168. ACM.