# Combining Wikipedia to Identify Prerequisite Relations of Concepts in MOOCs

Haoyu Wen, Xinning Zhu(✉), Moyu Zhang, Chunhong Zhang, and Changchuan Yin

School of Posts and Telecommunications, Beijing University, Beijing, China
{wenhaoyu,zhuxn,zhangmoyu,zhangch,ccyin}@bupt.edu.cn

**Abstract.** Many applications like the personalization recommendation system of online learning are based on prerequisite relations of concepts, which prompted us to automatically infer the prerequisite relations between the concepts in Massive Open Online Courses (MOOCs). The previous methods mostly use artificial features to identify the prerequisite relations from learning materials and Wikipedia. However, artificial features are complicated to deeply mine prerequisite information in MOOC videos and the Wikipedia-directed graph, resulting in poor performance. We propose a new and more effective method to identify prerequisite relations from the above two data resources. We first use a graph embedding algorithm to learn the vector representations of concepts from the created Wikipedia-directed graph and use the cosine similarity between the vectors to represent the semantic and structural relevance between the concepts. Second, we pre-train a Siamese network whose inputs are representations of course concepts learned by a variation of the LDA model to find more practical information of prerequisite relations from MOOC subtitles. Then, the concept similarities related to topic distribution can be represented by the pre-trained Siamese network's outputs. Finally, we add some excellent artificial features to expand the information of the prerequisite relations and input them together into a binary classifier to identify the prerequisite relations of the concepts in MOOCs. Our experiments on two MOOC datasets indicate that the proposed method achieves significant improvements comparing with existing methods.

**Keywords:** Prerequisite relation · Graph embedding · Siamese network

## 1 Introduction

Recently, the growth of available educational data has made a variety of emerging educational applications possible [1]. And prerequisite relations are important for describing the fundamental directed relations among concepts in knowledge structures. This paper focuses on the concept prerequisite learning problem in MOOCs, whose purpose is to predict whether a concept A is a prerequisite of a concept B given the pair (A, B) by considering different data sources comprehensively.

The data sources for discovering prerequisite relations can be divided into two categories: Wikipedia [2, 4, 5] and learning materials such as MOOCs or textbooks


T. Mantoro et al. (Eds.): ICONIP 2021, CCIS 1516, pp. 739–747, 2021.
https://doi.org/10.1007/978-3-030-92307-5_86

[3, 6, 7, 9]. A Wikipedia article, usually identifying a notable topic or concept with varying granularity levels, contains many hyperlinks which imply various relations between concepts, are exploited by many previous methods to detect prerequisite relations of concepts. However, most of these methods, only based on some handcrafted graph features, do not fully explore concepts with their relations in hyperlinks. On the other hand, students usually gain knowledge by watching instructional videos in MOOCs, so the order in which students acquire new concepts is highly consistent with the video playback sequence. When using MOOC videos as data sources to detect prerequisite relations, it is common to construct various manually defined features to excavate the information in educational materials and model it as a binary classification problem [10, 12]. However, due to the complexity of online resources, it is hard to achieve high performance with purely handcrafted features.

Based on the challenges raised above, we put forward a new model that fully excavates the information from Wikipedia and MOOC data. For Wikipedia data sources, we first construct a directed graph of concepts based on hyperlinks between Wikipedia articles. Then, use the node2vec [14] graph embedding algorithm to learn low-dimensional representations for concepts, capturing the diversity of connectivity patterns in the constructed concept graphs. For MOOC data, we make use of the Pairwise-Link-LDA [13] and Siamese network as a pre-training model to obtain the topic distribution of concepts. Instead of predicting the prerequisite relations directly as in [12], the pre-trained topic distribution of concepts is aggregated into the final classification model, with concept embeddings obtained from Wikipedia and some other important manually extracted features, further to improve the performance of identification of prerequisite relations. To evaluate the proposed method, we compare our method with the representative works of prerequisite learning on two MOOC datasets [6,12] and the experimental results show that our method achieves state-of-the-art results in the prerequisite relations discovery in MOOCs.

## 2 Problem Statement

In this section, we first give some definitions and then formulate the prerequisite identification problem.

Generally, there will be multiple courses in MOOC related to one subject area. Let $V = \{V_1 \dots V_k\}$ denotes the corpus of k courses, and $V_k = \{v_{k1} \dots v_{km}\}$ denotes a video sequence of $k^{th}$ course, where $v_{ki}$ is composed of concepts in the video subtitle text of the $i^{th}$ video of the $k^{th}$ course. According to the order of video playlists, a directed graph $G_V(V, E_V)$ can be constructed, in which nodes represent MOOC videos, and edges indicate the order of videos, i.e., $E_V$ contains a directed edge $e_{v_k}(i, j)$ if and only if MOOC video $v_{ki}$ plays before video $v_{kj}$. Let $C$ be the set of all concepts of interest in one subject area that is assumed to be known in advance in this study. For the $i^{th}$ video of the $k^{th}$ course, $v_{ki} = \{C \cap W_{ki}\}$, where $W_{ki}$ is the set of n-grams in $v_{ki}$, $n \in \{1, 2, 3\}$. As for Wikipedia, many concepts have their corresponding Wikipedia articles, which contain hyperlinks to related articles. We build a directed graph based on the linked information in Wikipedia, expressed as $G_W(C_W, E_W)$, where nodes set $C_W$ is a subset of $C$, that includes only the concepts having corresponding articles in Wikipedia. Edges set $E_W$

represent hyperlinks between Wikipedia articles, i.e., the directed edge $e_w(i, j)$ exists in $E_W$ if and only if the concept $c_j$'s wiki article contains a hyperlink to the concept $c_i$. Let $G_C(C, E_C)$ be a directed graph, called concept graph, where nodes represent concepts and edges represent prerequisite dependency, i.e., $E_C$ contains the directed edge $e_c(i, j)$ if and only if the concept $c_i$ is a prerequisite of concept $c_j$. For a given set of concepts $C$, we want to infer concept prerequisites $E_C$ from the set of MOOC videos $V$, the known video directed graph $G_V$ and the directed graph $G_W(C_W, E_W)$ created from Wikipedia.

## 3 Prerequisite Relations Extraction

### 3.1 An Overview of the Proposed Method

This paper proposes a novel prerequisite relation identification method that can make full use of Wikipedia and MOOC data. A framework of our proposed prerequisite identification model, called GESN, is shown in Fig. 1. For Wikipedia data, we first construct a directed graph represented as $G_W(C_W, E_W)$, and then use node2vec to obtain the vector representations of the concepts. The features about Semantic and Structural Relevance (SSR) between two concepts can be defined and used as an essential feature in the final prerequisite classifier. For MOOC data, the set of MOOC videos $V$ and the directed graph $G_V(V, E_V)$ are used to pre-train the Pairwise-Link-LDA model. The topic distribution of the concepts obtained from this model is used to pre-train a Siamese network with the known prerequisites' labels to get the features of the concepts' Similarities related to Topic Distribution (STD). Finally, the SSR, STD, and some other excellent handcrafted are used as the input of a binary classifier.

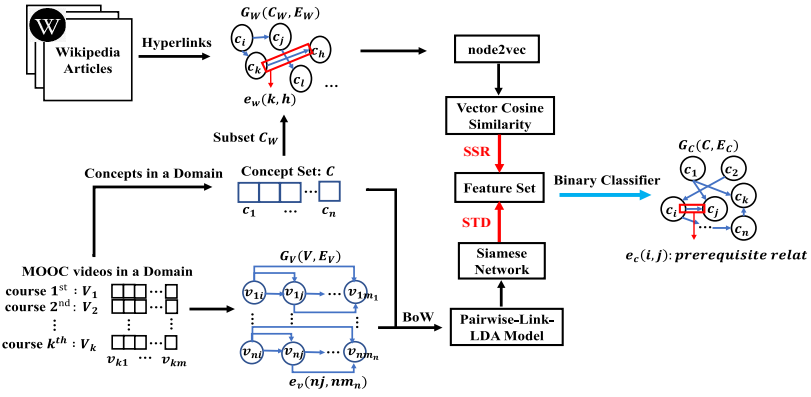


**Fig. 1.** Overall process of GESN.

### 3.2 Graph Embedding Based Wiki Information Extraction

Hypertexts in a Wikipedia article could help understand that article, reflecting the information about the prerequisite relations of concepts. It is worth noting that there may be some concepts that do not have related Wikipedia articles. We first need to find a set

of concepts $C_W$ from $C$ that have related Wikipedia articles. Then the directed graph $G_W(C_W, E_W)$ can be built based on the hyperlinks between Wikipedia articles of the concepts in $C_W$. The direction of the edge is similar to the direction of the prerequisite relations between concepts.

We thought of using a practical and scalable representation learning algorithm that can reflect network and node neighbors' characteristics to comprehensively consider the relations between all nodes in the whole digraph. The node2vec algorithm is a good choice. Node2vec improves the random walk method in DeepWalk [8], comprehensively considering the characteristics of Breadth-First Search (BFS) and Depth-First Search (DFS) due to the return parameter $p$ and in-out parameter $q$.

After obtaining the vector representation of each concept, the information of the prerequisite relations between two concepts can be represented by the cosine similarity of two vectors, and its essence is also a measure of the distance between the vectors. We define this distance as the Semantic and Structural Relevance (SSR) between two concepts.

### 3.3 Information of Topic Distribution from a Pre-trained Siamese Network

Recently, many handcrafted features have been proposed to mine the information about the prerequisite relations in MOOCs, but the complexity and scale of learning resources make it hard to improve the performance further.

Inspired by the method in [12], we use the combination of the Pairwise-Link-LDA model and a Siamese network to learn the topic distribution of concepts and get the similarities of the concepts related to the topic distribution by pre-training the Siamese network. The whole process is shown in Fig. 2. The Bag-of-Words (BoW) model is used to represent the subtitles of each video. After that, the MOOC video graph $G_V$ and the BoW vectors of MOOC videos $V$ are input into the Pairwise-Link-LDA model. Explicit modeling of directed links between ordered pairs of MOOC videos $E_V$ can better capture the MOOC videos' topics and the distribution of words over topics to capture the prerequisite relations between the words themselves. Each MOOC video generation process is the same as LDA. The topic-word distribution $\beta$ describes the topic distribution of each word. Based on the Pairwise-Link-LDA model, the word distribution over topics $\beta_{K\times|V|}$ can be obtained, where $V$ is the number of n-grams, and $K$ is the chosen number of topics. Please refer to [12] for more details of the Pairwise-Link-LDA model. The learned $\beta$ will be used as the input of a pre-training Siamese Network whose weights of the sub-networks are tied. Each concept is represented as a vector of dimension K. The input of the Siamese Network is the obtained vector representation of each pair of concepts symbolized as $(x_i, x_j)$. The pair $(x_i, x_j)$ is passed through the corresponding sub-network $G_w(.)$ which include fully connected neural network layers and a rectified linear unit, yielding two corresponding outputs $(o_i, o_j)$.

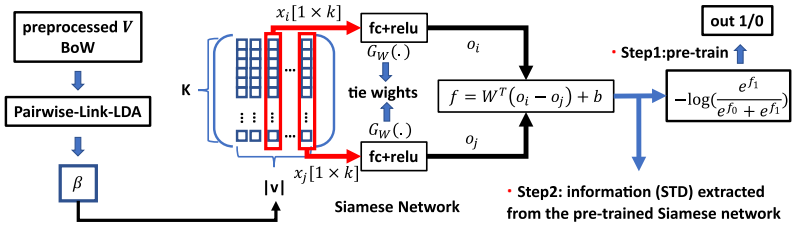


**Fig. 2.** Overall process of obtaining STD.

As we can see in Fig. 2, the whole process is divided into two steps. First, in the pre-training phase of the Siamese network, labeled pairs of such vectors from our training set are used to pre-train the Siamese network. These vectors are passed through the sub-networks $G_w(.)$. We denote $f$ as the sum of the weighted element-wise differences between the two feature vectors $o_i$ and $o_j$, and then obtain the probability $P_{x_1,x_2}$ of the first input vector $x_1$ is a prerequisite to the second vector $x_2$:

$$P_{x_1,x_2} = log\left(\frac{e^{f_1}}{e^{f_0} + e^{f_1}}\right) \tag{1}$$

where $f = W^T\left(G_w\left(\beta_{x_1}^T\right) - G_w\left(\beta_{x_2}^T\right)\right) + b$ is a two-dimensional vector and $f_n$ is the $n^{th}$element in it. Finally, the cross-entropy loss function is optimized concerning the parameter vectors controlling both the subnets through the stochastic gradient descent method using the Adam optimizer.

In step 2, for each pair of concepts, we obtain $\frac{e^{f_1}}{e^{f_0}+e^{f_1}}$ from the pre-trained Siamese network and use it as the feature of the concepts, defined as Similarities related to Topic Distribution (STD), for the final classification.

### 3.4 Classification for Prerequisite Relations

Some essential existing manually defined features include some information that cannot be obtained through SSR (#1) and STD (#2) at the same time, so we add them to the final binary classifier. First, some of these features are from [6], including Structural Features (#3) and Contextual Features that are variants of RefD (#4) and semantic similarity (#5). For more details of these features, please refer to [6]. Second, From many Wikipedia-based features proposed by [5], we use the Wikipedia-directed graph we constructed to calculate the PageRank difference of each pair of concepts (#6). Finally, all these six parts are combined as a binary classifier's input to capture the prerequisite information as much as possible.

## 4 Experiment

### 4.1 Data

We conducted related experiments on two datasets from different domains. First, we use a published MOOC dataset, named NPTEL, which belongs to the domain of computer

science. This dataset, from [12], is based on video playlists from a MOOC corpus. There are 382 videos from 38 different courses, and from 345 concepts, 1008 pairs of concept pairs with prerequisite relation were manually labeled. Out of 345 concepts, two concepts do not have related Wikipedia articles.

Second, we use a MOOC dataset in the domain of machine learning proposed by Pan et al. in 2017, named W-ML [6]. Then the W-ML dataset extracts 120 concepts from 244 concepts in the ML dataset and contains 486 pairs of annotated concept pairs with prerequisite relations. Besides, a total of 548 videos are included in the five courses. We regard the video under each main course's first-level heading as a short course, dividing the five main courses into 49 miniature courses.

Dataset statistics are detailed in Table 1.

**Table 1.** Dataset statistics.

| Dataset | \| \| | \| \| | \| \| | ***Pairs +*** | \| \| | \| \| |
|---|---|---|---|---|---|---|
| NPTEL | 382 | 38 | 1445 | 1008 | 345 | 343 |
| W-ML | 548 | 49 | 1171 | 486 | 120 | 120 |

### 4.2 Parameter Settings

First, for the Pairwise-Link-LDA model, the parameters we chose are the same as those used in [12]. In particular, we choose the number of topics $K = 100$ and a fixed Dirichlet parameter $\alpha = 0.01$. The Siamese network is pre-trained with a learning rate of 0.0001 and a batch size of 128. For the parameters in node2vec, we have done the same experiment as in [14] to analyze the parameter sensitivity. Except for the parameter being tested, all other parameters assume default values. The best in-out and return hyperparameters were learned using 10-fold cross-validation on 10% labeled data with a grid search over $p, q \in \{0.25, 0.50, 1, 2, 4\}$. Finally, we set dimensions $d = 128$, return parameter $p = 0.25$, in-out parameter $q = 4$, walk length $l = 80$, walks per node $r = 10$ and context size $k = 10$. For the NPTEL dataset, it is worth noting that two concepts do not have related wiki pages, and the calculation result of the cosine similarity of the above two concepts involved is set to 0.

Finally, we choose Random Forest (RF) as the binary classifier.

### 4.3 Results

The baselines we use are: RefD [2], iPRL [11], PREREQ [12], and MOOC-RF [6]. We compare our method with these baselines using precision (P), recall (R), and F1-score (F1). We summarize the comparing results of different methods across the two datasets in Table 2. We find that our method outperforms baseline methods across both two datasets. For example, the F1-score of our method on NPTEL outperforms PREREQ and MOOC-RF by 18.7% and 8.5%, respectively. Specifically, we have the following observations.

First, RefD achieves relatively high precision but the lowest recall. The reason may be that the interpretation of Wikipedia concepts is different from the teacher's interpretation of knowledge points. Second, the features extracted by MOOC-RF do not fully reflect the information about the prerequisite relations of concepts. Third, PREREQ achieves relatively high recall but with the lowest precision, which tends to identify more concept pairs that have prerequisite relations and will identify more negative samples as positive samples. Finally, even if iPRL considers the information in Wikipedia, its performance is not good due to the incomplete mining of prerequisite relation information and lack of annotated data.

**Table 2.** Comparison with baselines.

| Methods | NPTEL | | | W-ML | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| RefD | 68.3 | 34.7 | 46.0 | 72.6 | 36.4 | 48.4 |
| iPRL | 65.7 | 46.5 | 54.4 | 64.6 | 46.0 | 53.7 |
| PREREQ | 53.2 | 70.7 | 60.7 | 56.3 | 71.6 | 63.0 |
| MOOC-RF | 67.9 | 74.2 | 70.9 | 71.7 | 70.1 | 70.9 |
| **GESN** | **76.5** | **82.6** | **79.4** | **75.2** | **83.3** | **79.0** |

To get an insight into the importance of different six parts in our method, we perform a contribution analysis. Here, we run our approach seven times on the NPTEL MOOC Dataset. In each of the seven times, one or two parts are removed. We focus on the decrease of the F1-score for each setting. Table 3 lists the evaluation results after ignoring different parts. According to the decrement of F1 scores, we find that all the proposed parts help predict prerequisite relations. Primarily, we observe that SSR, decreasing our best F1-score by 4.5%, plays the most crucial role. On the contrary, with a 1.8% decrease, variants of RefD and the structural features are relatively less important, the cause might be that STD can capture structural information better, and the variants of RefD are sensitive to the length and number of the MOOC videos. We experience a decrease of 2.4% when we do not consider STD, which can more effectively dig out the prerequisite information hidden in MOOCs than the above artificial features. Semantic similarity can also help identify the prerequisite relations between concepts because it provides semantic information not contained in other parts. As for the PageRank difference, with a decrease of 3.7%, it is the same as mentioned in [5] that PageRank difference is an excellent feature. Finally, if we ignore SSR and STD simultaneously, the performance will be significantly affected, reflecting the importance of the proposed two parts.

**Table 3.** Contribution analysis

| Ignored part | Precision | Recall | F1 score |
|---|---|---|---|
| SSR | 72.7 | 77.1 | 74.9 (-4.5) |
| STD | 73.8 | 80.7 | 77.0 (-2.4) |
| SSR+STD | **72.6** | **75.0** | **73.8 (-5.6)** |
| Semantic Similarity | 74.2 | 79.8 | 77.0 (-2.4) |
| PageRank Difference | 72.7 | 79.0 | 75.7 (-3.7) |
| Variants of RefD | 74.7 | 80.8 | 77.6 (-1.8) |
| Structural Features | 75.3 | 80.0 | 77.6 (-1.8) |

## 5 Conclusion

We develop GESN, a supervised learning method, to learn concept prerequisites from MOOCs and Wikipedia. GESN first uses node2vec to capture the semantic and structural relevance between concepts from the Wikipedia-directed graph we built. Second, GESN obtains latent representations of concepts through the Pairwise-Link-LDA model, which are then used to pre-train a Siamese network. The pre-trained Siamese network can output the similarities between the concepts related to topic distribution. Finally, we combine some excellent features and input them together into a binary classifier to identify the prerequisite relations. GESN outperforms state-of-the-art methods on the dataset NPTEL and W-ML in two different domains.

## References

1. Hu, C., Xiao, K., Wang, Z., Wang, S., Li, Q.: Extracting prerequisite relations among wikipedia concepts using the clickstream data. In: Qiu, H., Zhang, C., Fei, Z., Qiu, M., Kung, S.-Y. (eds.) KSEM 2021. LNCS (LNAI), vol. 12815, pp. 13–26. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82136-4_2
2. Liang, C., Wu, Z., Huang, W., Giles, C.L.: Measuring prerequisite relations among concepts. In: EMNLP, pp. 1668–1674 (2015)
3. Gasparetti, F.: Discovering prerequisite relations from educational documents through word embeddings. Futur. Gener. Comput. Syst. **127**, 31–41 (2021)
4. Zhou, Y., Xiao, K., Zhang, Y.: An ensemble learning approach for extracting concept prerequisite relations from Wikipedia. In:International Conference on Mobility, Sensing and Networking (2020)
5. Liang, C., Ye, J., Wang, S., Pursel, B., Lee Giles, C.: Investigating active learning for concept prerequisite learning. In: AAAI (2018)
6. 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, vol. 1, pp. 1447–1456 (2017)
7. Adorni, G., Alzetta, C., Koceva, F., Passalacqua, S., Torre, I.: Towards the identification of propaedeutic relations in textbooks. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 1–13. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_1

8. Perozzi, B., Al-Rfou, R., Skiena, S.: DeepWalk: online learning of social representations. In: KDD, pp. 701–710 (2014)
9. Alzetta, C., Miaschi, A., Adorni, G., Dell'Orletta, F., Koceva, F., Torre, I.: Prerequisite or not prerequisite? That's the problem! an NLP-based approach for concept prerequisite learning. In: CLiC-it (2019)
10. Liang, C., Ye, J., Wu, Z., Pursel, B., Giles, C.: Recovering concept prerequisite relations from university course dependencies. In: AAAI (2017)
11. Lu, W., Zhou, Y., Yu, J., Jia, C.: Concept extraction and prerequisite relation learning from educational data. In: AAAI (2019)
12. Roy, S., Madhyastha, M., Lawrence, S., Rajan, V.: Inferring concept prerequisite relations from online educational resources. In: AAAI (2019)
13. Nallapati, R.M., Ahmed, A., Xing, E P., Cohen, W W.: Joint latent topic models for text and citations. In: KDD, pp. 542–550 (2008)
14. Grover, A., Leskovec, J.: Node2vec: scalable feature learning for networks. In: KDD, pp. 855–864 (2016)