# MOOCCubeX: A Large Knowledge-centered Repository for Adaptive Learning in MOOCs

Jifan Yu[1,2], Yuquan Wang[1], Qingyang Zhong[1], Gan Luo[1], ,Yiming Mao[1],Kai Sun[1],Wenzheng Feng[1],Wei Xu[1],Shulin Cao[1],Kaisheng Zeng[1],Zijun Yao[1],Lei Hou[1],Yankai Lin[2], Peng Li[2], Jie Zhou[2], Bin Xu[1], Juanzi Li[1], Jie Tang[1], Maosong Sun[1]
[1] Department of Computer Science and Technology, Tsinghua University, China
[2] Pattern Recognition Center, WeChat AI, Tencent Inc., China
yujf18@mails.tsinghua.edu.cn

## ABSTRACT

The prosperity of massive open online courses provides fodder for plentiful research efforts on adaptive learning. However, current open-access educational datasets are still far from sufficient to meet the need for various topics of adaptive learning. Existing released datasets often cover only small-scale data, lack fine-grained knowledge concepts. They are even difficult to curate and supplement due to platform limitations. In this work, we construct MOOCCubeX, a large, knowledge-centered repository consisting of 4, 216 courses, 230, 263 videos, 358, 265 exercises, 637, 572 fine-grained concepts and over 296 million behavioral data of 3, 330, 294 students, for supporting the research topics on adaptive learning in MOOCs. Licensed by XuetangX, one of the largest MOOC websites in China, we obtain abundant and diverse course resources and student behavioral data and are permitted to make subsequent periodic updates. We propose a framework to accomplish data processing, weakly supervised fine-grained concept graph mining, and data curation to improve usability and richness. Based on the fine-grained concepts, we re-organize the data from the knowledge perspective and acquire more external learning resources from the web. Our repository is now available at https://github.com/THU-KEG/MOOCCubeX.

## CCS CONCEPTS

• **Applied computing** → **Education**; • **Information systems** → **Extraction, transformation and loading**; **Data mining**.

## KEYWORDS

Adaptive Learning, Concept Extraction, Open-Access Datasets

**ACM Reference Format:**
Jifan Yu[1,2], Yuquan Wang[1], Qingyang Zhong[1], Gan Luo[1], ,Yiming Mao[1],Kai Sun[1],Wenzheng Feng[1],Wei Xu[1],Shulin Cao[1],Kaisheng Zeng[1],Zijun Yao[1],Lei Hou[1],Yankai Lin[2], Peng Li[2], Jie Zhou[2], Bin Xu[1], Juanzi Li[1], Jie Tang[1], Maosong Sun[1]. 2021. MOOCCubeX: A Large Knowledge-centered Repository for Adaptive Learning in MOOCs. In *CIKM '21: ACM International Conference on Information and Knowledge Management, November 01–05, 2021, Online.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456


*CIKM '21, November 01–05, 2021, Online*

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

*Education is not the filling of a pail, but the lighting of a fire.* Stimulating and addressing the needs of each individual learner has always been the goal of generations of educators. Adaptive learning [18], also known as adaptive teaching, was proposed with this expectation. Different from traditional learning (e.g. courses in classrooms) that presents the same material to all learners, adaptive learning aims at providing personalized learning items tailored to individual learners. Facilitated by the prosperity of Massive Open Online Courses (MOOCs), researchers make plentiful efforts and summarize three important directions, including (1) Adaptive Content, i.e. modeling and organizing the learning materials with text mining or knowledge discovery [23, 41]; (2) Adaptive Assessment, i.e. estimating the students' mastery on skills via cognitive diagnosis or knowledge tracing [2, 29]; (3) Adaptive Sequence, i.e. recommending appropriate items or optimal learning paths to learners [1, 21].

Despite the increasing research interests, existing educational datasets are gradually failing to meet the needs of systematic research on adaptive learning, due to the following challenges:

• **Insufficient Data Coverage**: Most of the educational datasets are built for a certain task or method, e.g., Chen et al. constructs a dataset only for prerequisite-based knowledge tracing [2] and Yu et al. build datasets only for concept discovery [41]. These datasets were not originally designed to cover the entire learning cycle of students, which makes them difficult to support the exploration of new settings or to systematically analyze the associations between tasks [31]. However, the regulation of online education platforms has greatly increased the difficulty of accessing data, which not only limits the size of education datasets but also makes them almost impossible to maintain and update. Once a dataset lacks a specific type of data, it will be very hard to be replenished effectively.

• **Coarse Concept Granularity**: Adaptive learning relies on accurate knowledge modeling of students and learning resources. Yet existing datasets are far from sufficient to meet this requirement, e.g. Assistment2009 [7] only contains 101 coarse-grained concepts for knowledge tracing. Based on such data, we can only roughly estimate what subjects the student is better at, rather than his mastery of specific knowledge. However, due to the mismatch between the high cost of expert annotation [24, 41] and the explosive growth of learning resources, constructing fine-grained knowledge concept graphs for learning resources is seriously challenging.

• **Limited Data Curation**: As mentioned above, diverse teaching resources (e.g., courses, exercises, and even blogs) and rich student behaviors (e.g., video watching, assignments, exams) are in the consideration of adaptive learning. The diverse distribution of data from these heterogeneous sources imposes high demands on data fusion and organization. Consequently, existing attempts are often small in scale, e.g., TutorialBank [8] only involves less than 1,000 resources. How to organize the heterogeneous educational resources to facilitate access, use, and retrieval by researchers on different topics remains an important challenge.

Therefore, we propose MOOCCubeX, a large open-access repository for adaptive learning. Licensed by XuetangX[1], one of the largest MOOC websites in China, MOOCCubeX integrates 4,216 courses, 230,263 videos, 358,265 exercises and over 296 million diverse behavioral data of 3,330,294 students. Based on these resources, we propose a framework for constructing a fine-grained concept graph via weak supervision, which contains over 600k concepts. The concept system is then employed in the effective organization, representation and retrieval of the resources. Furthermore, we also release several toolkits for the refinement and usage of this repository, so that researchers can conveniently build high-quality datasets for diverse topics.

**Contributions.** Our contributions can be divided into three parts:

(1) A large, freely available, high-coverage education collection based on the massive open online courses;

(2) A general framework for weak supervised fine-grained knowledge concept acquisition and concept-centric data organization;

(3) A series of toolkits that can be employed to replicate, analyze, and extend the repository, and to assist with constructing new datasets for adaptive learning.

## 2 BACKGROUND

### 2.1 Adaptive Learning

Adaptive learning, also known as adaptive teaching, is the delivery of custom learning experiences that address the unique needs of an individual through just-in-time feedback, pathways, and resources, rather than providing a one-size-fits-all learning experience [18]. This topic is highly associated with educational data mining, interactive design, and learning analytics [1]. It is widely accepted that adaptive learning includes three main research directions.

**Adaptive Content.** Also known as *expert model*, this direction aims to model the learning materials, which generally employ the methods of data mining [41], knowledge extraction [8] and natural language processing [24] for resource organization and comparison.

**Adaptive Assessment.** Also know as *student model*, this direction aims to model the students. Such attempts also come from Learning analytics, e.g., cognitive modeling and knowledge tracing [2, 29], which requires the analysis of the records of students' problem-solving.

**Adaptive Sequence.** A core module for adaptive learning, also known as *instruction model*, needs to fully consider the students' historical performance, knowledge states, and candidate resources to plan the learning path or recommend the next learning item. Existing related studies associate it with sequence recommendation, knowledge structure acquisition, etc [1, 21, 39].

Researchers have also worked on building a unified theoretical system [31] for educational applications, but exploration in these directions is still in its infancy. MOOCCubeX can provide a platform and fodder for related topics.

### 2.2 Educational Datasets

With the growing research community and influence, several open-access education datasets are built for different tasks. According to the main idea of construction, these datasets can be divided into *learner-centered* and *resource-centered*.

The *learner-centered* datasets, e.g., ASSISTment [9, 28], KDD Cup [5], EdNet [3], organize the data by mainly considering the behavioral data from students so that they can better support the tasks of student modeling and cognitive modeling. The *resource-centered* datasets, e.g., PRL [23], UniCourse [20] and TutorialBank [8], regard learning resources as the subject of data organization. They can better support the modeling of learning resources, with the help of techniques such as concept extraction and relation mining.

However, these datasets are designed only for a particular task and method, and limited in scale. They not only fail to meet the exploration of more advanced tasks but even struggle to support the enhancement of existing methods if needing more types of data. To alleviate this problem, we have previously tried to build a heterogeneous MOOC knowledge base, MOOCCube [40], with an initial attempt to integrate the two dataset construction ideas. However, MOOCCube has many inherent design flaws, including the lack of data types, too coarse granularity, insufficient data size, and lack of extension tools. Therefore, we re-collect the data and design a brand new data repository, MOOCCubeX, to be released as its full version in order to provide an open and abundant data resource for relevant researchers.

## 3 DATA AND FRAMEWORK

Our data comes from XuetangX, a partner of edX. The system was launched in October 2013 and up to May 31th, 2021, it has offered over 6,000 courses (including courses from Tsinghua University, Peking University, and edX courses from MIT, Stanford, UC Berkeley, etc.) and attracted 4,500,000 registered users. XuetangX provides a wealth of learning resources, allowing users to freely enroll in courses and participate in a complete learning process including video learning, homework exercises, and discussions. These data are highly correlated and well-maintained, so we use them as an ideal basis for MOOCCubeX. In this section, we first describe how to obtain raw data from XuetangX, and then introduce the design of our technical framework, which finally completes a large-scale, high-coverage knowledge repository for adaptive learning.

### 3.1 Raw Data Collection

*3.1.1 Course Resource.* The starting point for MOOCCubeX is obtaining course data from XuetangX. After eliminating test courses and deactivating courses, we first collect detailed information on the obtained 4,216 *courses*. At this stage, the name and description of each course are stored as `text` type, and each course is assigned

[1] https://www.xuetangx.com/

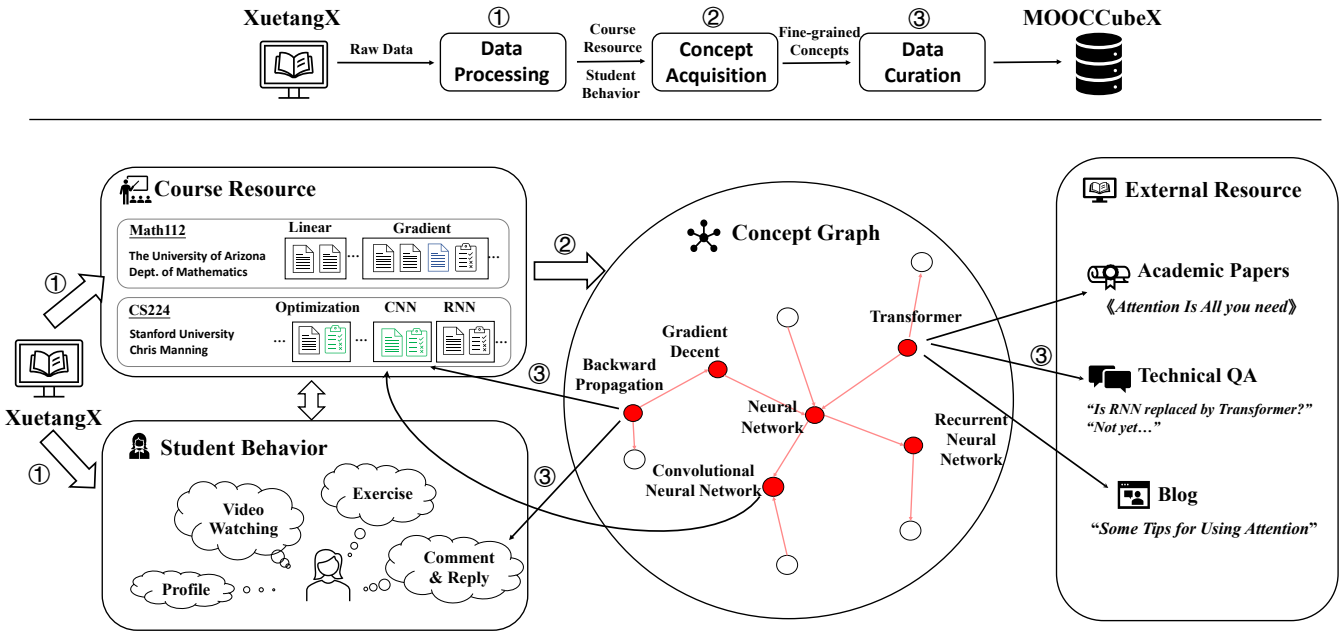



**Figure 1: The structure of the MOOCCubeX repository and the construction framework.**

a course id. As shown in Figure 1, *courses* in MOOCs are not independent. A course consists of multiple teaching chapters [26], and a chapter is usually composed of a series of *videos* and *exercises*. Such structured information is also very important, so we then focused on the collection of course-related information, including the syllabus of the course, and the list of resources it contains (including videos, exercises, and comments), saved as the `list` type. Furthermore, we preserve the *teacher* and *school* of the course, with their introduction crawled from the web. This kind of information can build associations for courses and support related tasks such as teaching style detection. After completing the collection of course information, we dive deeper into the obtaining of each course resource:

**Video.** 4,216 courses provide 2,798,892 pre-recorded videos, which are the main knowledge carriers for online learning. We record the title of each video, as well as the name of the chapter it is in, as `text` type. Moreover, for each video, we preserve video subtitles with a timeline, i.e. for each sentence in the subtitles, researchers can easily find out its start time and end time (to the millisecond level). Each video has a video id which serves as the identifier of the video (starting with `V_`).

**Exercise.** We also obtain 358,265 exercises from courses in XuetangX. Exercises consist of one or more problems, which are the important resources used in the learning evaluation process. We assign the exercise id as the identifier (Starting with `Ex_`), and then we crawl the problems of this exercise. MOOCCubeX preserves the correspondence between exercise id and problem id (starting with `Pm_`). We employ 0, 1, and 2 to identify the problem types: Single choice, multiple choice, or subjective problems. Each of the choices and the standard answer of the problem are also collected. Finally, we get 2,454,397 individual problems.

*3.1.2 Student Behavior.* Besides the static course resources, the various types of student behaviors are also essential for adaptive learning research, which helps the modeling of student's learning intents [10], cognitive levels [2, 29] and social activities [22]. Therefore, we collect granular records from XuetangX, including student profiles, video watching, exercises, and discussions. These behaviors are naturally linked to the course resources. Although getting the license from the platform, we still need to carefully perform desensitization operations such as anonymization during the data processing.

**Student Profile.** Crawling from the MOOC platform, we preserve the gender, location, age, and grade level of 3,330,294 students. To prevent privacy disclosure, we remove sensitive user data (such as names, phone numbers, emails, etc.) during collection. Furthermore, before the storage of user data, we employ anonymization and static masking techniques [11] to prevent possible risks caused by query attacks.

**Video Watching Behavior.** When a user studies MOOCs, the main activity is watching the MOOC videos. We collect the *video learning record* from January 19, 2020 to November 3, 2020, covering all 4,124 courses. In this way, we obtain 154,332,174 raw data of video watching logs. These data are student video 'HeartBeat' logs at five-second intervals, i.e., every five seconds, the system records what video that student is currently watching and what position he is studying. This data is highly detailed, from which specific learning trajectories of students can be inferred, including whether they jumped video positions, watching speed, etc.

**Exercise Behavior.** The record of doing exercises is the basis for modeling the students' mastery of knowledge [29]. It is a measure of the outcome of the learning process. For each student, we record the timestamp of completing each problem, the submitted answer, the score of each student. Furthermore, for those problems that students submitted answers multiple times, we preserve a history of each submission. Finally, we obtain 133, 384, 333 problem-solving records for data preparation.

**Comment and Reply.** Students' online learning process is often accompanied by discussions and communication. These discussions not only reflect the social relationships among students but also serve as important feedback on the course design. MOOCCubeX completes the storage of discussion content by using the comments and replies obtained from the platform. Each *comment* is attached to a certain video or exercises, started by a student and replied by more students. Therefore, for each *comment*, we crawl the text of it and attach the comment id to the corresponding video id or exercise id, as well as the owner student id. For each reply, we preserve the text information, its owner student id and original comment id, and finally obtain 8, 422, 134 comment-reply records.

After the collection of these raw data, we obtained a sufficient amount of course resources as well as rich student behaviors to support the adaptive learning task. However, these data still cannot be directly utilized by researchers: (1) the data are too detailed to be brought to use and need to be reasonably aggregated; (2) these resources are still loosely organized in terms of courses and lack knowledge-level modeling; (3) the repository includes only intra-course resources and lacks a large number of heterogeneous external resources. Therefore, we propose a framework to process the raw data and build a fine-grained concept graph to complete the data curation, making MOOCCubeX a knowledge-centric and high-availability repository.

## 3.2 MOOCCubeX Construction Framework

Figure 1 shows the framework of the construction of MOOCCubeX. We first aggregate, align and label different types of course resources and student behaviors into data cube with appropriate granularity. Then, we design weakly supervised concept acquisition methods so as to construct a fine-grained concept graph for MOOCCubeX without deploying heavy expert annotation. Finally, we curate the overall data based on these fine-grained concepts, thus fusing resources from inside and outside the MOOC. Specifically, this framework consists of three main stages.

(1) **Data Processing.** Given the raw data from XuetangX, this stage mainly addresses issues such as course classification, resource deduplication, and reasonable data aggregation of student behavior, which makes the resources more accessible.

(2) **Fine-grained Concept Acquisition.** In this stage, we design weakly supervised methods to extract fine-grained concepts from the course video texts and discover the prerequisite relationships among them to build a large concept graph.

(3) **Data Curation.** Based on the fine-grained concepts, we further conduct concept annotation on other types of resources, so as to re-organize the course resources. Furthermore, we employ fine-grained concepts to correlate intra-course resources with diverse external to complete the data curation.

# 4 IMPLEMENTATION

In this section, we introduce the implementation details of our construction framework and present how to get access to the MOOCCubeX repository.

## 4.1 Data Processing

The raw data obtained from XuetangX still has many pressing issues, including low course relevance, duplication of some resources, and overly granular student behavior. We designed separate solutions to improve the data quality.

*4.1.1 Course Classification.* MOOC courses are independent of each other, but this is not conducive to data query and storage, and can significantly increase the number of annotation demands in subsequent processing. Therefore, we enrich course associations by grouping courses into subject classification systems. According to the fields of Classification and code of disciplines[2], three experts label the field which the course belongs to. To control the granularity of the classification, we labeled all courses to the second level (88 candidate disciplines), thus forming a course discipline tree.

*4.1.2 Resource Deduplication.* Many popular courses are offered repeatedly. Although they may be slightly tweaked, a large number of course resources are duplicated, which obviously leads to problems such as data sparsity and redundant labeling. Therefore, we conduct an initial filtration based on information such as course name, teacher, and school to find repeated courses. We then accurately match the video and the exercises using their content texts. For each resource (video and exercise), we label it with a ccid. Resources with the same ccid denote the same resources that recur in repeated courses. The amount of deduplicated resources was refined to 10.3% of the original size.

*4.1.3 Student Behavior Integration.* As mentioned in data collection, the students' video watching data is 'HeartBeat' logs at five-second intervals, i.e., every five seconds, the system records what video that student is currently watching and what position he/she is studying. Such data is not convenient to use, so we then further process the learning behavior data into watching segments, i.e., each segment indicates which segment of the video the student is watching (start and end time) and the playback speed at that time. If there is a jump or pause in the video, it is considered a new segment. Finally, we preserve 1, 852, 256 watching segments.

## 4.2 Fine-grained Concept Acquisition

Concepts refer to the knowledge concepts taught in courses, e.g., "Binary Search Tree" of Data Structure course. These concepts are a summary of learning resources from a knowledge perspective. In recent years, concept graphs have been very popular in many topics related to adaptive learning, including knowledge discovery in MOOCs, concept-based data organization, knowledge-enhanced recommendation, and cognitive modeling [8, 9, 19]. A common solution for concept acquisition is extracting concepts and their relationships from educational texts [24, 41]. The previous course information collection stage provides rich textual resources, including the subtitles of videos, the description of courses, teachers, schools,

[2]The standard of GB/T 13745-2009. It is the current official version.

and the content of exercises. However, high-quality concept acquisition from these texts is challenging: costly manual annotation is insufficient to cover large-scale MOOC resources, limiting the selection of methods.

To tackle the challenge, we propose a weakly supervised fine-grained concept extraction method, and then design an interactive co-learning mechanism for discovering the prerequisite relationships of concepts with minimal annotation. Besides, we only employ the texts of video subtitles as the source, since the texts of other resources are too short for concept acquisition.

*4.2.1 Concept Extraction.* In order to obtain high-quality concepts from texts, a general solution is to divide the process into two stages: candidate extraction (for recall) and concept ranking (for precision). Due to the costly annotation, we use three methods in the candidates' extraction stage, i.e., phrase mining, entity linking, and named entity recognition, which are mainly supported by external knowledge graphs and a small amount of annotation. As for the concept ranking stage, we employ a cluster-based unsupervised method [41] to determine the extraction quality.

**Candidate Extraction.** Concept candidates are selected as the union of the following three extracted sets.

(1) *Phrase Mining*. We select the noun-phrase titles of Chinese Wikipedia[3] as a phrase table. These phrases are crowdsourcing annotated high-quality entities. So we preserve the phrases in the phrase table for each video that appear in its subtitles as concept candidates.

(2) *Entity Linking*. Entity linking aims to discover the mentions of an external knowledge base. For each video, we perform entity linking with XLink [42] and select linked entities as concept candidates, which can help us to extract the concepts of large-scale knowledge base Xlore [16].

(3) *Named Entity Recognition*. We adopt pre-trained language models for fine-grained concept extraction. The training scheme can be regarded as Named Entity Recognition with a single category. If a phrase appearing in the video subtitle is a concept, then we annotate its span as a "named entity". For each discipline, we annotate concepts from $120-150$ videos' subtitles of $20-30$ random courses. We then fine-tune RoBERTa [4] with token classification loss by Huggingface Transformers [38]. We concatenate a video's titles before its subtitles as hints, separated by a special token [SEP]. Finally, we run prediction on all video subtitles and select phrases with confidence larger than 0.85 as concept candidates. Experimental results show that RoBERTa-NER method extracts more fine-grained concepts than phrase mining and entity linking.

**Concept Ranking.** Employing concept clusters to build a course concept space has been proven to be effective in unsupervised concept ranking. Therefore, to improve the precision of extracted concepts, we follow the idea of Yu et al. to cluster each course's concepts into 15 clusters by its BERT embedding with K-means and select those in the top 2 highest scored clusters as machine extracted concepts of the course. The score of cluster $j \in \{1..15\}$ is calculated by

$$score(j) = \min_{1 \le i \le 10} d(s_i, c_j)$$

where $s_i$ and $c_j$ are the center of $i$-th seed cluster and $j$-th candidates, $d(\cdot,\cdot)$ is cosine similarity function of BERT embedding and the labeled seed concepts of each discipline is clustered into 10 clusters.

*4.2.2 Prerequisite Discovery.* The prerequisite relation among concepts indicates whether concept A is beneficial for understanding concept B [12]. However, the sparsity of this relationship poses a challenge for the annotation and extraction [20]. In MOOCCubeX, we propose an interactive co-training method that combines a text-based method and a graph-based method for discovering such relationships. To reduce the label costs, we ensemble our text model and graph model results to automatically generate candidate pairs and manually find out the positive pairs.

**Method Building.** We produce a text-based method and a graph-based method for prerequisite relation discovery, which are two main typical methods in this task.

*Text-based Method.* We apply a simple neural network classifier to predict whether a concept pair has the prerequisite relation. A concept consists of several tokens. Unlike previous neural models that use fixed word embeddings, we employ a text encoder of BERT [6], to obtain the embedding of a concept. We take the output vector at the end position and the "[CLS]" token as the embedding vector for LSTM and BERT, respectively. Then we concatenate two embeddings of the concept pair and predict the binary score.

*Graph-based Method.* Similar to text methods, we employ graph encoders to obtain concept embeddings first, e.g. Graph Attention Networks (GAT) [36] to obtain concept embeddings through their initial representations and the graph structure between them. Then we concatenate two embeddings of the concept pair and predict the binary score. For the initial representation of a concept, we averaging the embeddings from text encoders of it. For the graph structure between concepts, we utilize each MOOC course's video order of the course. Since courses are taught in a cognitive order, we can infer that concepts in a video may have prerequisite relations with concepts in the following videos.

**Interactive Labeling.** To reduce the label costs, we ensemble our text model and graph model results to automatically generate candidate pairs and manually find out the positive pairs. Specifically, our method takes the previously labeled positive pairs and randomly sampled negative pairs as training data and ranks other pairs according to the predicted probability. Then experienced annotators label the top pairs with higher positive probability. And another experienced annotator checks the positive pairs and removes graph cycles to keep the topology. Except for the initial iteration that we manually label a small number of seed positive pairs, this generation and labeling process repeats alternately multiple times until we have enough positive pairs.

## 4.3 Data Curation

Facilitated by the construction of the fine-grained concept graph, we can enrich the association of the heterogeneous MOOC resources and integrate more types of external resources. For the heterogeneous MOOC resources, we conduct concept annotation on them to link them in the concept graph. For external resources, we integrate them into existing resources through concepts for data curation.

[3] https://zh.wikipedia.org

| CourseId | CName | Field | ChapterId | VideoId | VideoText | ExerciseId | ProblemId | ProblemText | PType |
|---|---|---|---|---|---|---|---|---|---|
| *C_1729* | *Artificial Intelligence* | *CS* | *L_4522* | *V_59697* | *AI is the intelli -gence exhibited by machines or software.* | *Ex_7552* | *Pm_14512* | *Which of the follow -ing are the research fields of AI? **A**:...* | 0 |
| | | | | | | *Ex_7554* | *Pm_14520* | *What is the AI method to represent information by symbols and their relationship?* | 2 |

**Table 1: An example of course resource in MOOCCubeX. The example shows the meta information, teacher, school, a video and its corresponding two problems in course *Artificial Intelligence*. Phrases in blue are concepts.**

| UserId | CourseId | CName | VideoId | ExerciseId | ProblemId | Answer | Comment | Reply | Time |
|---|---|---|---|---|---|---|---|---|---|
| *U_112* | *C_1729* | *Artificial Intelligence* | *V_59697* | \ | \ | \ | *Is the Turing test complete?* | \ | 2020-04-20 16:57:50 |
| | | | \ | *Ex_7552* | *Pm_14512* | A | \ | \ | 2020-04-21 10:14:13 |
| | | | \ | *Ex_7554* | *Pm_14520* | *Symbolicism* | \ | \ | 2020-04-23 14:29:06 |
| | | | *V_59703* | \ | \ | \ | \ | *I think under -standing and intelligence are two things...* | 2020-04-26 21:35:45 |

**Table 2: An example of student behavior in MOOCCubeX. The example shows the video watching behavior, exercise behavior and comment and reply behavior of student *U_112* in course *Artificial Intelligence*. Phrases in blue are concepts.**

*4.3.1 Multi-resource Concept Annotation.* Existing MOOC resources retain only very sparse connections through the course structure. By linking various resources to the fine-grained concept graph, we can enrich their knowledge-level connections. As the concepts are extracted from video subtitles, the *videos* are naturally annotated with fine-grained concepts. However, other types of resources are still lack of concept annotation. For each *course*, its concepts are the union of its video concepts. For the *exercises*, *comments* and *replies*, we first select the concepts of the videos in the same chapter as candidates and then employ their BERT embedding to match top $1-3$ concepts, according to the threshold 0.8 of cosine similarity. As a result, each MOOC resource is linked to the concept graph. Measured in terms of 10 common concepts, each course is correlated with 26 other courses. And respectively measured in 3 and 1 common concepts, each video has 422 related videos and each exercise has 364 correlations, showing the richness of knowledge-based resource connections.

*4.3.2 External Resource Curation.* Based on concepts, we further search for diverse external resources as a supplement. *Academic Papers*: We crawl 10 corresponding papers from ArnetMiner [34] for each concept. *Blogs and Technical QA*: Employing concepts as keywords, we crawl related blogs from CSDN[4] and technical QA content from Zhihu[5]. This type of data is also saved to MOOCCubeX and associated with in-class resources through concepts.

[4]https://www.csdn.net/. One of the largest computer technical blog sites in China.
[5]https://www.zhihu.com/. A famous Chinese question-and-answer website.

## 4.4 Availability

Our repository is now publicly available on https://github.com/THU-KEG/MOOCCubeX. All the collected courses, videos, exercises, student behaviors, and external resources are presented in corresponding subpages and are free to download. Table 1 and 2 shows the data example after the concept-based curation. Specially, we provide researchers with two toolkits for an easy and quick start with our repository: 1) *MOOCCube Dataset Builder* toolkit; 2) *MOOCCube Concept Helper* toolkit.

• **MOOCCube Dataset Builder.** This toolkit is built for easily constructing individual datasets for diverse adaptive learning-related tasks. We package 10 common operations for querying different types of data in MOOCCubeX. These operations can also be combined for building more complex datasets.

• **MOOCCube Concept Helper.** Researchers can utilize our published methods as introduced in Section 4.2, for concept extraction and prerequisite discovery. Following the guidance of these methods, researchers can build fine-grained prerequisite concept graphs with a small number of annotations.

We will continue to update the content of MOOCCubeX, including the ongoing addition of more prerequisite relations of some specific disciplines, and regular updates to the student behaviors. The user privacy protection will also follow the anonymization process as in Section 3.1.

# 5 EXPERIMENT

In this section, we reveal the characteristics of MOOCCubeX by presenting statistics and comparing them with other educational

| Dataset | Course | Video | Exercise | Chapter | Concept | Prerequisite | Student | Prof. | Prob. | Disc. | Behavior | External |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PRL | 20 | 1.3k | ✗ | ✗ | <1k | 3.5k | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| UniCourse | 654 | ✗ | ✗ | ✗ | <1k | 1.1k | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| NPTEL | 38 | <1k | ✗ | ✗ | <1k | 1.5k | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| ASSISTment | ✗ | ✗ | 26.6k | ✗ | <1k | ✗ | 4.2k | ✗ | ✓ | ✗ | 347k | ✗ |
| KDD Cup 2015 | 39 | ✗ | ✗ | ✗ | ✗ | ✗ | 112k | ✗ | ✗ | ✗ | 1319k | ✗ |
| EdNet | 1,021 | ✗ | 13.2k | ✓ | <1k | ✗ | 298k | ✗ | ✓ | ✗ | 89270k | ✗ |
| TutorialBank | ✗ | ✗ | <1k | ✗ | ✗ | <1k | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| LectureBank | 60 | ✗ | ✗ | ✓ | <1k | <1k | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| MOOCCube | 706 | 38.1k | ✗ | ✗ | 106k | 17.6k | 199k | ✗ | ✗ | ✗ | 4874k | ✓ |
| Our | 4,216 | 230k | 358k | ✓ | 637k | 126.7k | 3330k | ✓ | ✓ | ✓ | 296138k | ✓ |

**Table 3: Statistics of existing Educational datasets. *Prof.*, *Prob.* and *Disc.* are the abbreviations of user profiles, exercising records, and discussions. *Behavior* is the sum of raw behavior records.**

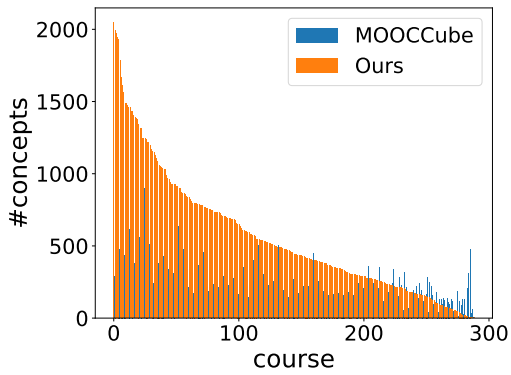



**Figure 2: Comparison with MOOCCube. The Figure shows the difference of Concept Distribution.**

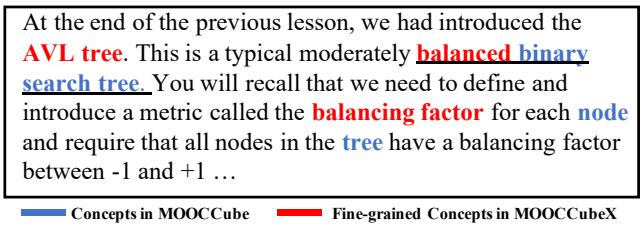



**Figure 3: An example from the "Data Structures and Algorithms" course, comparing the concept richness between MOOCCube and MOOCCubeX. The underline indicates that the concept is complemented by MOOCCubeX.**

datasets. Furthermore, we conduct an adaptive learning recommendation task upon MOOCCubeX as an usage example.

## 5.1 Characteristic

We select several famous education datasets for comparison, which can be divided into three categories: Educational knowledge discovery datasets, i.e., PRL [23], UniCourse [20] and NPTEL [30]; Learning analytics datasets, i.e., ASSISTment [9, 28], KDD Cup [5], EdNet [3]; Integrated educational resources dataset, i.e. TutorialBank [8], LectureBank [19], MOOCCube [40].

**Coverage.** Our dataset contains rich concept-based background knowledge (as in the knowledge discovery datasets), fine-grained records of a large number of student behaviors (as in the learning analytics datasets, such as Assessment, KDD Cup). Meanwhile, MOOCCubeX contains more diverse data resources than the existing integrated datasets (such as TutorialBank, LectureBank), including exercises, course structures, multiple student behaviors, etc. Compared with the other open-access education datasets, MOOCCubeX is excellent for both data size and data richness.

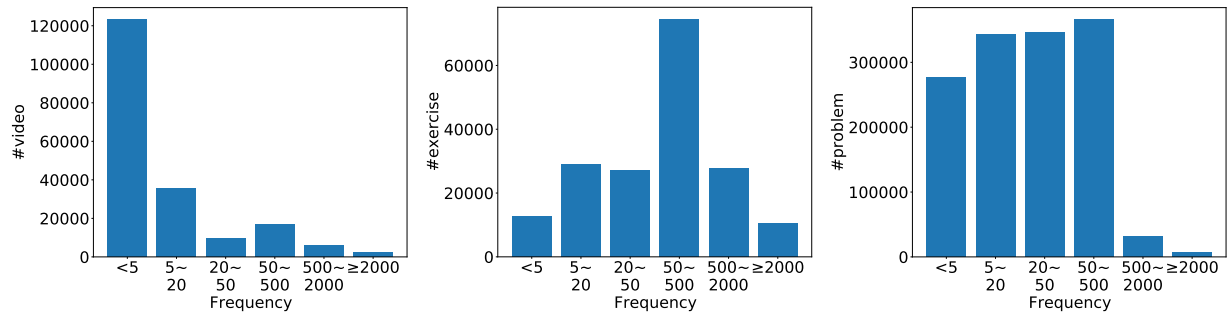
**Supported Tasks.** The limited scale or lack of data can directly lead to some adaptive learning tasks not being supported. In Table 4, we compare the MOOCCubeX's supporting tasks with existing representative datasets (UniCourse for *educational Knowledge Discovery*, EdNet for *Learning Analytic* and MOOCCube for *Integrated datasets.*). It is not difficult to observe that the limited scale or lack of data can directly lead to some adaptive learning tasks not being supported. While the existing dataset mainly serves the corresponding domain, MOOCCubeX can adequately support various adaptive learning-related tasks.

**Concept Quality.** To evaluate the quality of extracted concepts and prerequisite relationships, we randomly select 200 videos from 6 courses in different disciplines for expert annotation and evaluation. Three experts of the corresponding domain need to annotate: (1) the concepts of the videos (2) the relationships among the given concepts. The annotation results are double-checked. Finally, the F1-score of concept extraction on the evaluation set is 0.861, and the F1-score of prerequisite discovery is 0.905, reflecting the high quality of our concept acquisition.

**Concept Richness.** As a concept-based knowledge repository, we focus on comparing the richness of the concept part with the preliminary version of MOOCCube. We have 6 times the number of concepts than MOOCCube [40]. After matching our course to its, we compare the concept distribution among courses. As Figure 2 shows, 91% of the matched 302 unique courses in our repository contains more concepts than MOOCCube. Figure 3 shows an example of the concepts in the "Data Structures and Algorithms" course. From which we can see that the coarse-grained concept suffers from

| Direction | Task | UniCourse | EdNet | MOOCCube | MOOCCubeX |
|---|---|---|---|---|---|
| Adaptive Content | Concept Extraction | ✗ | ✗ | ✓ | ✓ |
| | Prerequisite Discovery | ✓ | ✗ | ✓ | ✓ |
| | Resource Evaluation | ✓ | ✗ | ✓ | ✓ |
| Adaptive Assessment | Knowledge Tracing | ✗ | ✓ | ✗ | ✓ |
| | Cognitive Diagnosis | ✗ | ✓ | ✗ | ✓ |
| | Dropout Prediction | ✗ | ✓ | ✓ | ✓ |
| | Social Network Analysis | ✗ | ✗ | ✗ | ✓ |
| Adaptive Sequence | Educational Recommendation | ✗ | ✓ | ✓ | ✓ |
| | Learning Path Planning | ✗ | ✗ | ✓ | ✓ |

**Table 4: The Adaptive Learning tasks supported by different data repositories.**




(a) Video's watching (b) Exercising (c) Problem Solving

**Figure 4: The distribution of student behavioral data.**

incompleteness (*binary search tree*) and ambiguity (*tree*). With the utilization of fine-grained concepts, concepts such as *AVL tree* and *balancing factor* in the course are discovered, resulting in more accurate modeling of the course content.

**Student Behavior.** (1) *Video watching*. Although we employ video segments to process the raw data, the student behavior in real MOOCs is still sparse. The popularity of videos varies a lot. As shown in Figure 4, the frequency distribution of video watching is long-tailed. Among all students, 28.5% of the total have watched the most popular video. However, 39.2% of the total videos have only been watched once, which shows the extreme unbalance of the data distribution. (2) *Exercising and Problem Solving*. On the contrary, the exercising behavior is more balanced and appears like a normal distribution. When we investigate deeper into the finishing situation of a specific problem, we can find that each frequency interval is also very balanced. One potential reason of this phenomenon is that final grades and certificate issuance in MOOCs are highly correlated with participation in homework and exercising. Such data characteristics can be further analyzed by pedagogical researchers in order to develop instructional models that improve course engagement.

## 5.2 Application

One of the major aspects of adaptive learning is Adaptive Sequence, e.g. the sequential recommendation of learning resources [27]. We conduct this task on MOOCCubeX to evaluate the capability of this dataset and discuss how it can be used to conduct related studies.

Following recent efforts [21], we define this task as recommending the next video to the student based on the student's historical video learning sequence. This task not only requires good modeling of student behaviors but also considering the role of the knowledge embedded in the videos, the structure of the courses, etc.

*5.2.1 Dataset Construction.* We select all of the students' video watching segments from January 19, 2020 to November 3, 2020, covering 230,263 videos from 4,126 courses. Filtering the records shorter than 20, we have the video watching sequences of 33,650 students. Then we utilize the `concept_finder` and `course_info_finder` from *MOOCCubeX Dataset Builder* toolkit to get the concepts and course information corresponding to the videos.

*5.2.2 Baselines.* We reproduce several baselines from sequential and educational recommendations. The models of these baselines include CNN, RNN, GNN, Bayesian Networks and even pre-training.

- **KSS** [21], a simple baseline for learning decision, which ranks the videos according to the course syllabus and video order;
- **GRU4Rec** [14], a widely-used session-based recommendation model based on GRU architecture;
- **CASER** [33], which employs CNN in both horizontal and vertical way to model high-order MCs for sequential recommendation;
- **TrueLearn** [1], an educational recommendation method that utilizes a family of Bayesian Networks.

• **KGAT** [37], a GNN-based method that employs the background knowledge graph for a better recommendation. Such method is reproduced on the concept co-occurrence network.

• **BERT4Rec** [32], a typical pre-trained recommendation model based on bidirectional Transformer architecture like BERT.

*5.2.3 Evaluation Metrics.* To evaluate these models, we adopted the *leave-one-out* evaluation, which has been widely used in [13, 17, 32, 33]. For each user, we hold out the last item of the behavior sequence as the test data, treat the penult as the validation set, and utilize the remaining items for training. The models are evaluated in the *Random Negative Sampling* setting: rank each ground truth item in the test set with 100 randomly sampled negative videos that the user has not watched. We employ two typical evaluation metrics, including Normalized Discounted Cumulative Gain (NDCG) and Recall. We report the top-k NDCG and Recall (k = 1, 5, 10). For all these metrics, the higher the value, the better the performance is. Note that NDCG@1 is equivalent to Recall@1 since we only have one ground truth item for each user.

*5.2.4 Result Analysis.* The experimental results are shown in Table 5. From the analysis of these results, we can understand the characteristics of the MOOCCubeX from a practical use perspective. Moreover, we provide some insights for subsequent research based on this investigation.

**Convenience.** MOOCCubeX conveniently accommodates the varying data needs of baselines. Methods with diverse base models can be adequately tested on it. And the process of dataset generation is smooth and easeful, which meets the expectations of our design.

**Non-sequential Behaviors.** KSS heavily underperforms our estimates. At the same time, the performance of the sequential recommendation models (GRU4Rec, CASER) is still acceptable. This phenomenon indicates that students' actual video watching behaviors are not following the preset order. Therefore, adaptive recommendations for online education are an urgent need to guide students on a more appropriate learning path. In addition, exploring and mining the information implied by students' learning sequences may be a promising direction for research.

**Technique Selection.** For one thing, as evidenced by KGAT's impressive performance at the @5 level, it is promising to focus on the knowledge behind educational resources in future research due to the high correlation between education and knowledge. For another thing, the BERT4Rec achieves surprising results. As the pre-training methods have made a big splash in NLP field, the field of adaptive learning, supported by sufficient and complete data, may be completely dominated by such new methods in the future.

## 6 IMPACT

In this section, we first highlight the impact of MOOCCubeX on relevant research directions and then analyze which research groups will benefit from this repository.

**Impact on Educational Technology Promotion.** Applying advanced technologies into adaptive learning is a major trend, e.g., Self-attention in knowledge tracing [25], heterogeneous graph convolution in prerequisite discovery [15], etc. These attempts require large-scale, high-quality data, and as much side information as possible. The wealth of information covered by MOOCCubeX can provide ample sustenance for these new technologies and tasks.

| Model | Metric@1 | Metric@5 | | Metric@10 | |
|---|---|---|---|---|---|
| | NDCG | NDCG | Recall | NDCG | Recall |
| KSS | 0.98 | 2.90 | 4.94 | 4.50 | 9.94 |
| GRU4REC | 45.14 | 53.47 | 61.55 | 61.79 | 65.81 |
| CASER | 33.57 | 46.51 | 48.87 | 59.46 | 66.84 |
| TureLearn | 37.61 | 40.25 | 44.99 | 52.86 | 59.84 |
| KGAT | 42.30 | 50.72 | 63.87 | **67.95** | **74.45** |
| BERT4Rec | **60.34** | **65.13** | **69.34** | 66.12 | 72.40 |

**Table 5: Results of learning recommendation task. Results in the table are percentage numbers with '%' omitted.**

**Impact on Educational Theoretical Exploration.** In recent years, researchers in education science have diligently pursued the exploration of a unified theoretical framework for adaptive learning. Shaffer et al. propose *Epistemic Frame Theory* [31], which analyzes the network constructed from multiple learning factors. Meanwhile, Tsai et al. propose the idea that connecting the analysis results among individual tasks is incredibly valuable for mining the inner logic of learning analytics [35]. As a repository containing most of the data types required for related tasks, MOOCCubeX will play an indispensable role in theoretical exploration.

**Beneficiary Groups.** The beneficiary groups of MOOCCubeX include the researchers of education data mining, learning analytics, adaptive learning, knowledge discovery, as well as the researchers from pedagogy who devote themselves to the theory exploration. MOOCCubeX also welcomes industry developers from AI-powered education to validate next-generation educational products.

## 7 CONCLUSION AND FUTURE WORK

We present MOOCCubeX, a large, knowledge-centered repository consisting of 4, 216 courses, 230, 263 videos, 358, 265 exercises, 637, 572 fine-grained concepts and over 296 million behavioral data of 3, 330, 294 students, for supporting the research topics on adaptive learning in MOOCs. With a high-coverage data collection of diverse teaching resources and the behavioral data of the entire learning process, we extracted a large, fine-grained concept graph via weak supervision, and employ it to complete the heterogeneous data curation. The repository is now publicly available with our provided toolkits. We also conduct adaptive learning application to illustrate the usage of MOOCCubeX, and discuss some insights of future directions based on the experimental results.

There are several promising paths for future work. For Data Mining, Artificial Intelligence researchers, one of the most urgent concerns is to apply advanced but data-intensive techniques to existing adaptive learning tasks. Simultaneously, mining the patterns of different student behavior data is probably instructive for a large number of subsequent researches. For researchers in educational fields such as learning analytics, MOOCCubeX provides data that can support the investigation of associations between different learning analytics tasks, which can validate the hypotheses of relevant theories. We hope that the establishment of MOOCCubeX will call for more efforts in the related topics of adaptive learning.

# REFERENCES


[1] Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. 2020. Truelearn: A family of bayesian algorithms to match lifelong learners to open educational resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 565–573.

[2] Penghe Chen, Yu Lu, Vincent W Zheng, and Yang Pian. 2018. Prerequisite-driven deep knowledge tracing. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 39–48.

[3] Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. 2020. Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education*. Springer, 69–73.

[4] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101* (2019).

[5] KDD Cup. 2015. KDD Cup 2015: Predicting dropouts in MOOC.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[7] Tanya Elias. 2011. Learning analytics. *Learning* (2011), 1–22.

[8] Alexander Richard Fabbri, Irene Li, Prawat Trairatvorakul, Yijiao He, Weitai Ting, Robert Tung, Caitlin Westerfield, and Dragomir Radev. 2018. TutorialBank: A Manually-Collected Corpus for Prerequisite Chains, Survey Extraction and Resource Recommendation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 611–620.

[9] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction* 19, 3 (2009), 243–266.

[10] Wenzheng Feng, Jie Tang, and Tracy Xiao Liu. 2019. Understanding Dropouts in MOOCs. In *Proceedings of the AAAI Conference on Artificial Intelligence, (Vol 33 No 01: AAAI-19, IAAI-19, EAAI-20)*.

[11] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. 2007. Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*. 758–769.

[12] Jonathan Gordon, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully Burns. 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)*. 866–875.

[13] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[14] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[15] Chenghao Jia, Yongliang Shen, Yechun Tang, Lu Sun, and Weiming Lu. 2021. 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*. 2036–2047.

[16] Hailong Jin, Chengjiang Li, Jing Zhang, Lei Hou, Juanzi Li, and Peng Zhang. 2018. XLORE2: Large-scale cross-lingual knowledge graph construction and application. *Data Intelligence* 1, 1 (2018), 77–98.

[17] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *Proceedings of 2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[18] Andreas Kaplan. 2021. *Higher Education at the Crossroads of Disruption: The University of the 21st Century*. Emerald Group Publishing.

[19] Irene Li, Alexander R Fabbri, Robert R Tung, and Dragomir R Radev. 2019. What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6674–6681.

[20] Chen Liang, Jianbo Ye, Zhaohui Wu, Bart Pursel, and C Giles. 2017. Recovering concept prerequisite relations from university course dependencies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[21] Qi Liu, Shiwei Tong, Chuanren Liu, Hongke Zhao, Enhong Chen, Haiping Ma, and Shijin Wang. 2019. Exploiting cognitive structure for adaptive learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 627–635.

[22] Helmi Norman, Norazah Nordin, Melor Md Yunus, and Mohamed Ally. 2018. Instructional Design of Blended Learning with MOOCs and Social Network Analysis. *Advanced Science Letters* 24, 11 (2018), 7952–7955.

[23] Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. 2017. Prerequisite relation learning for concepts in moocs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1447–1456.

[24] Liangming Pan, Xiaochen Wang, Chengjiang Li, Juanzi Li, and Jie Tang. 2017. Course Concept Extraction in MOOCs via Embedding-Based Graph Propagation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 875–884.

[25] Shalini Pandey and George Karypis. 2019. A self-attentive model for knowledge tracing. *arXiv preprint arXiv:1907.06837* (2019).

[26] Laura Pappano. 2012. The Year of the MOOC. *The New York Times* 2, 12 (2012), 2012.

[27] Alexandros Paramythis and Susanne Loidl-Reisinger. 2003. Adaptive learning environments and e-learning standards. In *Second european conference on e-learning*, Vol. 1. 369–379.

[28] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. 2014. Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics* 1, 1 (2014), 107–128.

[29] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Advances in neural information processing systems* 28 (2015), 505–513.

[30] Sudeshna Roy, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. 2019. Inferring concept prerequisite relations from online educational resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9589–9594.

[31] David Williamson Shaffer, Wesley Collier, and Andrew R Ruis. 2016. A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics* 3, 3 (2016), 9–45.

[32] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[33] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 565–573.

[34] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 990–998.

[35] Yi-Shan Tsai, Vitomir Kovanović, and Dragan Gašević. 2021. Connecting the dots: An exploratory study on learning analytics adoption factors, experience, and priorities. *The Internet and Higher Education* 50 (2021), 100794.

[36] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[37] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 950–958.

[38] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.

[39] Susen Yang, Yong Liu, Lei Chenyi, Guoxin Wang, Haihong Tang, Juyong Zhang, and Chunyan Miao. 2020. A Pre-training Strategy for Recommendation.

[40] Jifan Yu, Gan Luo, Tong Xiao, Qingyang Zhong, Yuquan Wang, Wenzheng Feng, Junyi Luo, Chenyu Wang, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jie Tang. 2020. MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3135–3142. https://doi.org/10.18653/v1/2020.acl-main.285

[41] Jifan Yu, Chenyu Wang, Gan Luo, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jie Tang. 2019. Course Concept Expansion in MOOCs with External Knowledge and Interactive Game. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*. 4292–4302.

[42] Jing Zhang, Yixin Cao, Lei Hou, Juanzi Li, and Hai-Tao Zheng. 2017. XLink: An Unsupervised Bilingual Entity Linking System. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Maosong Sun, Xiaojie Wang, Baobao Chang, and Deyi Xiong (Eds.). Springer International Publishing, Cham, 172–183.