Dang FR, Tang JT, Pang KY *et al.* Constructing an educational knowledge graph with concepts linked to Wikipedia. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 36(5): 1200–1211 Sept. 2021. DOI 10.1007/s11390-020-0328-2

Constructing an Educational Knowledge Graph with Concepts Linked to Wikipedia

Fu-Rong Dang[1,+], Jin-Tao Tang[1,+], *Senior Member*, *CCF*, Kun-Yuan Pang[1]
Ting Wang[1,*], *Senior Member*, *CCF*, Sha-Sha Li[1], and Xiao Li[2]

[1] *College of Computer Science and Technology, National University of Defense Technology, Hunan 410073, China*
[2] *Information Center, National University of Defense Technology, Hunan 410073, China*

E-mail: 8mimixi8@163.com; {tangjintao, pangkunyuan, tingwang, shashali, xiaoli}@nudt.edu.cn

Received January 11, 2020; accepted August 18, 2020.

Abstract To use educational resources efficiently and dig out the nature of relations among MOOCs (massive open online courses), a knowledge graph was built for MOOCs on four major platforms: Coursera, EDX, XuetangX, and ICourse. This paper demonstrates the whole process of educational knowledge graph construction for reference. And this knowledge graph, the largest knowledge graph of MOOC resources at present, stores and represents five classes, 11 kinds of relations and 52 779 entities with their corresponding properties, amounting to more than 300 000 triples. Notably, 24 188 concepts are extracted from text attributes of MOOCs and linked them directly with corresponding Wikipedia entries or the closest entries calculated semantically, which provides the normalized representation of knowledge and a more precise description for MOOCs far more than enriching words with explanatory links. Besides, prerequisites discovered by direct extractions are viewed as an essential supplement to augment the connectivity in the knowledge graph. This knowledge graph could be considered as a collection of unified MOOC resources for learners and the abundant data for researchers on MOOC-related applications, such as prerequisites mining.

Keywords concept extraction, educational resource, knowledge graph, massive open online course (MOOC), prerequisite

1 Introduction

Recent years have witnessed the rapid development of MOOCs (massive open online courses)①, a totally new online educational model, which provides high-quality educational resources of top universities to global learners in a more distributed and more Internet-friendly format than the schooling.

MOOCs are characterized by openness and autonomy, which has distinguishing advantages in traditional educational tasks such as educational resources selection and learning path planning. Although it has seen significant progress, cross-platform online education resources bring information overload and learning trek to learners [1], because MOOCs which even have the same name may vary in languages, subjects, levels, teaching methods, and goals.

Studies show that in the early stages of the MOOC learning process, the dropout rate can be up to 91% [2]. For people who fail to keep pace and have a tendency to drop out, the main reason is that the course content does not comply with the learning needs or the difficulty of the course exceeds the learner's ability [3]. In the past, prerequisites among courses were manually created by experts over the decades. While free accessible MOOCs have increased steadily over the years, prerequisite relations are required for online educational applications to give effective guidance and cut the dropout

Regular Paper
Recommended by CCKS 2019
This work was supported by the National Key Research and Development Program of China under Grant No. 2018YFB1004502, and the National Natural Science Foundation of China under Grant Nos. 61532001, 61702532 and 61303190.
+Contributed equally to this work
*Corresponding Author
①https://www.bestcolleges.com/research/#reports, Sept. 2021.

rate. It serves students of varying educational backgrounds and is adaptable to any subjects.

Knowledge graph describes the concepts, entities and their relations in the objective world in a structured form, and expresses the mass knowledge of the Internet into a form closer to the human cognitive way [4]. It has recently drawn increasing interest as an effective approach to the generation of many intricate domains, which also can facilitate online educational resource organizations across platforms much more than simple information extraction techniques.

Our research not only focuses on the construction of the cross-platform knowledge graph about MOOC resources but also proposes concept extraction that could fuse the knowledge for better concept retrieving and improve learners' experience. To be specific, extracting concepts from materials in MOOCs would help students better grasp the main points as well as support research towards deeper analyses such as automatic labeling [5].

The rest of this paper is organized as follows. Section 2 describes related studies of knowledge graph construction with a focus on the MOOC domain. Section 3 presents our certain method for the knowledge graph construction, concept extraction, and prerequisite generation. Section 4 gives a detailed analysis for our knowledge graph and the comparison with a topic-similar one. Section 5 concludes this paper and discusses future work.

2 Related Work

In recent years, knowledge graph has become a main form for semantic representation of networked information to support intelligent systems. The research of knowledge graph construction has extended from the general-purpose open domain, such as DBpedia [6], to the domain-specific knowledge graph, such as financial, medical and educational [7–9].

In the field of education, knowledge graph has been extensively used for learning representation and effective evaluation. For improving adaptive learning, [10] proposes a learner preference model based on Semantic Web and mainly considers the cognition method, interests and learning ways; Chaplot *et al.* induced structures of multiple units in a course [11]; Sun *et al.* used the knowledge graph technology to express associations among contents for visualization [12]. Chen *et al.* proposed a system that could extract concepts from the educational domain and dig out their relations to build an educational knowledge graph [13].

In order to better understand the prerequisite relationship among MOOC resources, a considerable number of researchers used different algorithms to infer prerequisite dependencies of MOOCs. Alsaad *et al.* [14] proposed unsupervised methods to determine concept co-occurrence in the lecture transcripts. Yang *et al.* proposed the Concept-Graph-Learning (CGL) framework to learn the relationship between MOOCs by inferring concept prerequisite relations [15].

It can be seen that course descriptions often are too disparate structurally and semantically to compare for online learners with limited knowledge. Existing researches are concentrated on constructing knowledge graphs to show how to organize resources in an efficient way but not resource itself.

We aim to offer users with an informative educational knowledge graph with complete resource lists and prerequisite dependencies. Our study distinguishes in reorganizing educational resources and bringing implicit knowledge out by extracting concepts to achieve more satisfying MOOCs learning experience.

3 Construction of Knowledge Graph

To build a knowledge graph, the first thing is to figure out its architecture. There are two layers in the knowledge graph, the logical layer (always represented by the ontology) and the data layer (a large number of instances) [16]. The logical layer scaffolds the knowledge graph which serves refined regulations or axioms to support the data layer in a normal way. In the data layer, knowledge is organized as the triples ("entity-relation-entity" or "entity-property-value") and constitutes a large network.

In this paper, we use the top-down approach to construct the knowledge graph. Simply speaking, we build the ontology first, and then link instances to corresponding entities in the ontology [17]. The building process is generally divided into four modules: knowledge modeling, knowledge acquisition, knowledge fusion, and knowledge storage [18].

Specifically, as shown in Fig.1, the first step of our approach is knowledge modeling which builds the ontology with Protégé②. The second step is knowledge ac-

②Protégé is an open-source tool that assists in the construction and modification of ontologies with an intuitive user interface. https://protege.stanford.edu/, June 2020.

③Neo4j was selected to store this knowledge graph because of high read and write performance, user-friendly graph query language

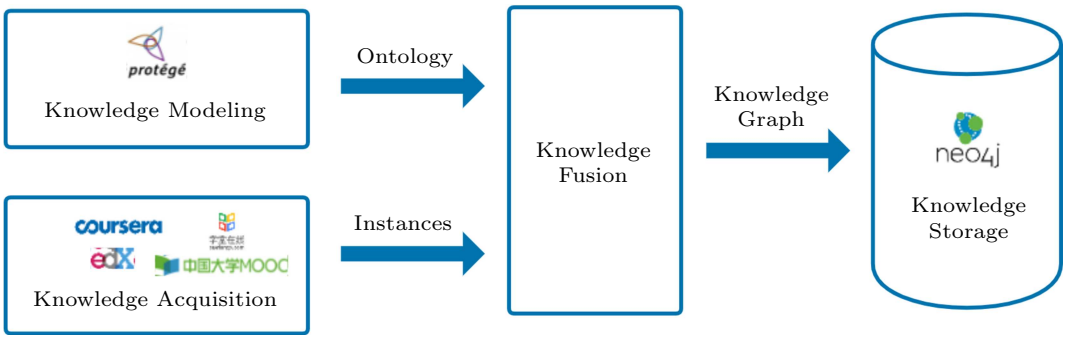


Fig.1. Construction process of the knowledge graph.

quisition which retrieves data from major MOOC platforms. After data fusion, we accomplish the knowledge graph and store it into Neo4j graph database③.

3.1 Knowledge Modeling

Knowledge modeling is to establish a conceptual model of knowledge graph for organizing and describing the associations among entities and their relations.

Here we construct an ontology for the knowledge graph while the ontology is the logical layer and the instances belong to the data layer. A knowledge graph G is composed of the ontology G_l, the data graph G_d and the corresponding relation R between them. It can be represented as $G = (G_l, G_d, R)$ and $G_l = (N_l, P_l, E_l)$. To be specific, N_l is the set of the class nodes, P_l represents properties and E_l is the set of relations that connect classes [19].

In our research, we define five interrelated classes and their relations, including platforms, courses (MOOCs), universities, instructors and concepts. Specifically, each platform cooperates with many universities to provide courses, and universities employ lots of instructors to teach various courses. Each course includes lots of concepts to represent knowledge points.

Fig.2 shows a schematic of the ontology G_l we built. Classes N_l are shown in circles with different colors. Therefore, instances can be distinguished easily in color. Properties P_l are the colored rectangles which have the same color as the corresponding class. For example, "name", "link", and "language" are partial properties of courses. E_l is the black line between classes, such as universities "EMPLOY" instructors.

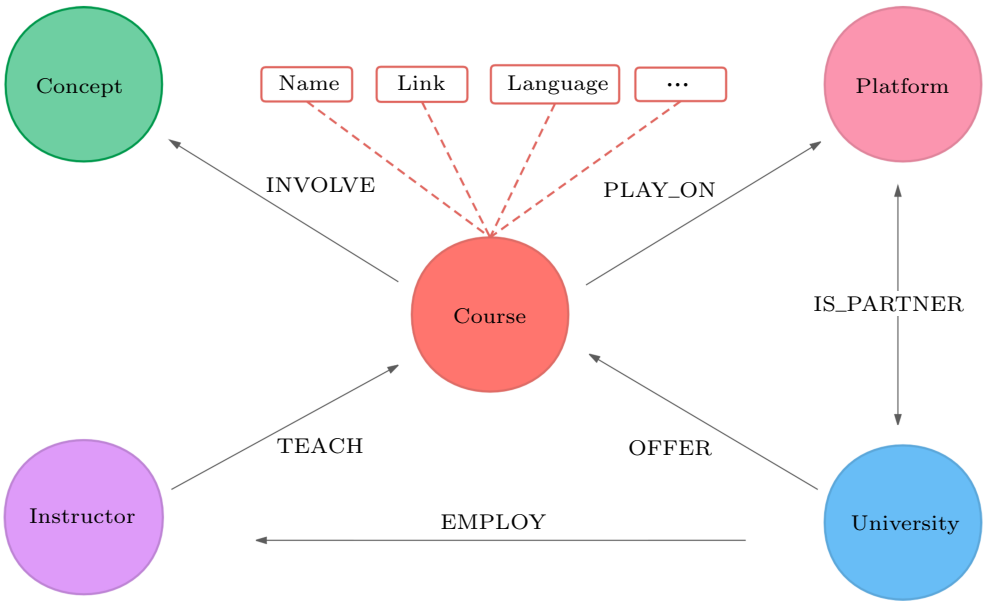


Fig.2. Ontology of the knowledge graph.

Cypher and visualization of data. https://neo4j.com/, June 2020.

3.2 Knowledge Acquisition

In this subsection, we describe the knowledge acquisition process which serves as a key part in our framework. Due to the diversity and complexity of real-world MOOCs, it is crucial to accurately and efficiently acquire the fact data from different sources for building a knowledge graph. As we mainly deal with course properties, concepts and prerequisite relations of MOOCs, we can categorize the knowledge acquisition process into three groups.

3.2.1 Knowledge Acquisition from Web Information

When students learn a MOOC, they can find the course properties (e.g., title, abstract, and language) as well as relations with universities and instructors directly from the web portal. Thus, in this category of knowledge acquisition, we aim at directly extracting course information from MOOC platforms as well as their relations with other classes. Extracting course information from online websites could be viewed as a task of knowledge acquisition with semi-structured data [20]. We compare the type of users, enrollment, profit, reputation, the number of MOOCs, partners, and the feedback of all platforms globally. As a result, Coursera, EDX, XuetangX, and ICourse are selected as the data source [21]. The former two are English-based MOOC platforms where most MOOCs come from American and British institutes. The latter two are Chinese-based platforms focusing on MOOCs from Chinese institutes.

Fig.3 shows an example of extracting a course entity on EDX whose attributes and relations with instructors and the university are circled in red rectangles. We transform these web text into structured data to save them into the knowledge graph. Here we simulate the original screenshot and circle properties out for a higher resolution of the figure.

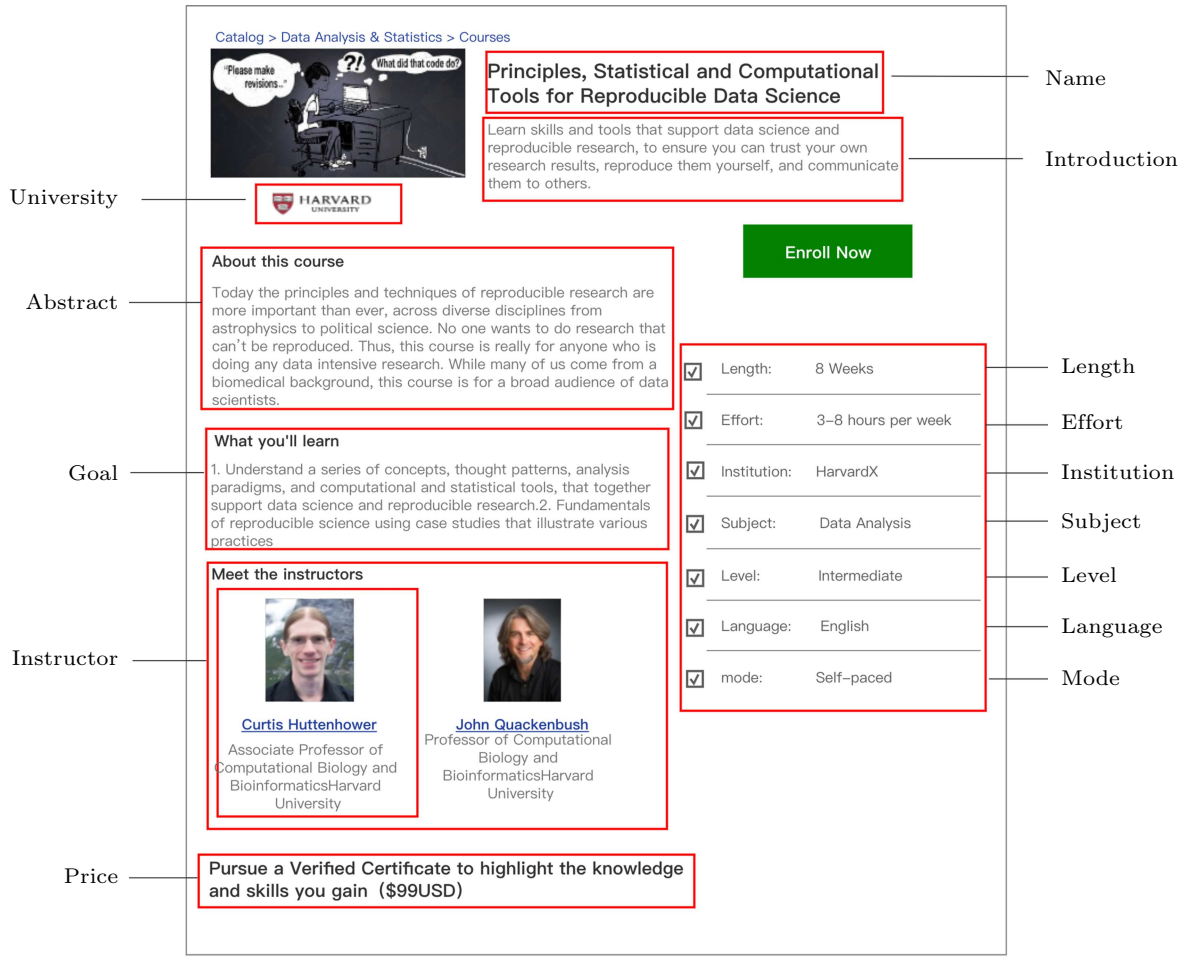


Fig.3. Information extraction of an example MOOC on EDX.

3.2.2 Knowledge Acquisition from Concepts

• *Motivation to Extract Concepts.* Compared with traditional higher educational courses that have limited numbers of students, a MOOC may draw more than 100 000 registrants with diverse backgrounds all over the world [22]. In most cases, key points presented in MOOCs are well grasped by certain students who may have the stronger understanding while they are indigestible to other students [23]. Thus identifying key knowledge points in MOOCs is very important.

Usually, concepts or knowledge points hide in the text body of course information. It requires large amounts of effort to acquire and analyze their relevance as well as importance in a MOOC [24]. From a view of text mining, concept extraction focuses on extracting important semantic key phrase of course content, which is similar to keyword acquisition in the traditional text mining task [25]. To ease the burden of manual keyword annotation, automatic keywords identification has recently become a promising research direction, such as concept map from textbooks [26], concepts extraction from video transcripts [27], searching for new concepts in MOOCs [28], automated categorization and extraction of concepts from titles of scientific articles [29]. Most of these researches are unsupervised with limited guidance.

However, unsupervised keyword extraction methods may not work well for MOOCs because of the lack of sufficient text labels. For example, in a MOOC named "Data Mining Application", keywords such as "SVM" may only occur once in a video. This can be very hard for unsupervised methods. In this paper, we use external resources such as Wikipedia to enrich text representation so that core knowledge points in MOOCs can be accurately identified.

• *How to Extract Concepts.* Among the attributes of MOOCs, the directory written by instructors could give a comprehensive list of the domain. Besides, compared with MOOCs on Coursera and EDX, nearly all MOOCs on ICourse and XuetangX are annotated with their own directories. Thus, we choose the latter ones as data sources.

In this paper, we extract concepts from directories of MOOCs and link them with Wikipedia. Generally, we need to generate a group of concept mentions (i.e., candidate concepts) first. If these concept mentions are consistent with certain Wikipedia entries, we can successfully link them. In other cases, although we cannot find exact text counterpart of concept mentions in Wikipedia, it does not mean these concept mentions are invalid. Thus, we need to find these probable relevant concepts.

To accurately evaluate the general meaning of concept mentions, we use a word embedding based method which embeds basic units of words and constitute a higher hierarchy like the phrase and named entity [30], for the measurement between Wikipedia entries and concept mentions. We give a brief description of this algorithm (see Algorithm 1).

Here are relevant variables in Algorithm 1: a course corpus is composed of n MOOCs, denoted as $M = \{m_i\}, i = 1, ..., n$, where m_i is a MOOC and an object to deal with. We assume that d_i is the directory (containing titles of each chapter) of m_i. And several concept mentions could be extracted from d_i. The concept mention set is denoted as C. c is each mention in d_i and $c \in C$. Vector set V_t is trained with the Wikipedia text by the CBOW algorithm [31]. Our goal is to choose a proximate Wikipedia entry e for each concept mention c. Concrete steps are shown in Algorithm 1.

We conduct a user study to rate Algorithm 1. Based on the experiment, 90.7% of concept mentions exactly match the Wikipedia entries, which can be linked straightforwardly. To calculate the accuracy of links between concept mentions and semantically similar Wikipedia entries, we randomly sample 100 concept mentions not in Wikipedia entries. Two M.Sc. students and one Ph.D. student majoring in software engineering are invited to judge whether the links between concept mentions and Wikipedia entries are proper. They are asked to label "proper" or "improper" and links labeled differently would be abandoned. After evaluation, we find the accuracy of the algorithm for concept mentions which need to be calculated is 20%. The concept mentions not in Wikipedia entries account for 9.3%. Eventually, the global accuracy rate of links between concept mentions and Wikipedia entries is 92.6%.

Fig.4 illustrates how Algorithm 1 works④. The words marked in red and blue from the directory of "Principles of Management" in XuetangX both are concept mentions. Blue ones could be linked with certain Wikipedia entries directly while red ones need to be calculated for the nearest Wikipedia entries. Eventually, each concept mention is transformed into the concept with word embeddings directly or after being cal-

④We conduct the concept extraction only on the Chinese-based platforms and MOOCs. For illustrative purposes, the directory and concepts are written in English and Chinese in the figure (Fig.4).

Algorithm 1. Link Concept Mentions of the MOOC with Certain Wikipedia Entries

Input: a directory d_i, the vector training set V_t, Wikipedia entry set E and its vector set V_e

Output: Wikipedia linked entries with concept mention set C

1. Preprocess d_i, including sentence segment and stopping word removing. Concept mention set C is generated and the vector-based representation is V_c.
2. Set $n = 100, s = 0.2$, n is the number of neighbor vectors, and s is the minimum neighboring distance.
3. **for** $c \in C$ **do**

if $c \in E$ **then**
c has a totally corresponding linked entry as itself.

else

1) After word segment on c, word set P is generated. Then remove single words and stopping words in P.

2) Initialize the neighbor set $N = \emptyset$.

3) **for** $p \in P$ **do**

a) **for** $v_t \in V_t$ **do**
Calculate the cosine similarity

$$sim(v_p, v_t) = \cos(v_p, v_t)$$

b) Rank v_t in descending order of $sim(v_p, v_t)$ and put top n into N, then

$$N = \{v_t | sim(v_p, v_t) > s\} \text{ and } |N| = n$$

4) K-means algorithm splits N into $|P|$ groups denoted as $N_1, N_2, ..., N_k$. Choose the biggest group and the final vector-based presentation of c is the average vector of the group, then

$$v = average(v_t), \text{ where } v_t \in N_{\max}.$$

5) **for** $e \in E$ **do**
Calculate the cosine similarity

$$sim(v) = \cos(v, v_e).$$

6) Rank e in descending order of $sim(v)$ and choose the top 1 as the linked Wikipedia entry of c.

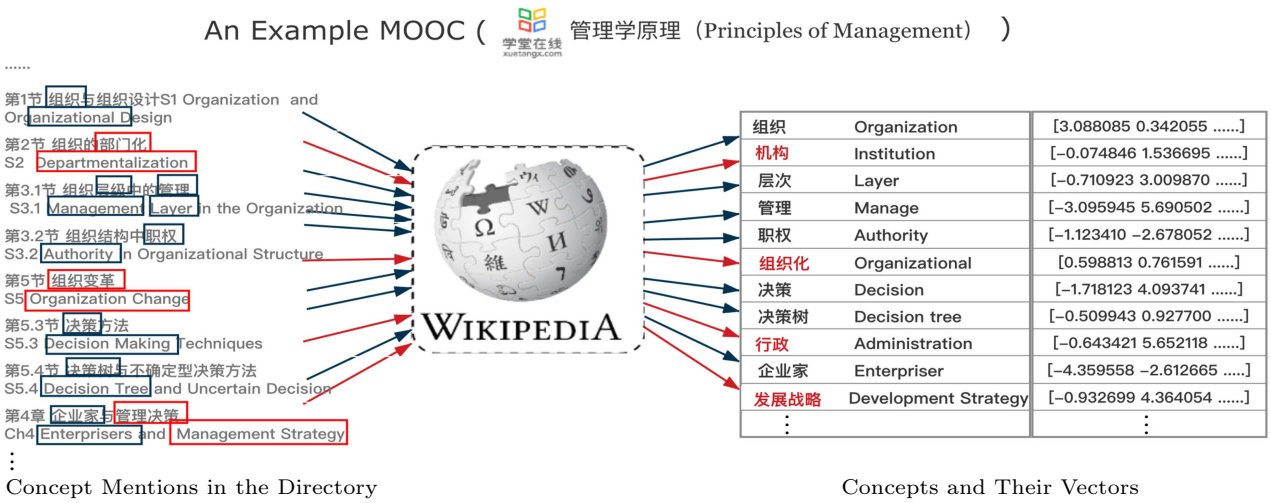


Fig.4. Example of concept extraction and links with Wikipedia entries.

culated.

For 2 197 and 1 644 MOOCs separately on ICourse and XuetangX, each directory is about hundreds of words ordinarily. After phrasing and filtering, 616 MOOCs (48 MOOCs on ICourse and 568 MOOCs on XuetangX) lost their meaningless directories. We extract and link 24 188 concept mentions from other directories with Wikipedia entries. For the knowledge graph, there are 24 188 instances and more than 200 000 relations between concepts and courses supplemented.

The concept count is shown in Table 1.

3.2.3 Knowledge Acquisition from Prerequisites

Prerequisite knowledge of MOOCs can help learners understand the teaching plan and provide professional guidance to most learners [32]. Despite being a relatively new research area, learning prerequisite relations of courses has been explored in numerous ways, especially data-driven [33, 34].

To learn prerequisite relations between MOOCs, we use a simple yet effective text mining method. Usually, in MOOCs, instructors often write course prerequisite knowledge in the introduction page. These descriptions are flat text which describes the basic requirement of mastering MOOCs. The accuracy of such prerequisites information is usually high. However, they are not structured as MOOC titles or rarely mentioned, while we assign the prerequisite MOOCs according to the description as following steps.

1) Extract prerequisite description *pre* of each MOOC m.

2) Add consolidated course names to facilitate word parsing on *pre*. After stopping words removing and word segmentation, the keywords in *pre* are extracted as $K = \{k_1, k_2, ..., k_j\}$.

3) Compare k_i with all the MOOC names with the Sorensen distance and Jaccard distance, and then select MOOCs whose distances are smaller than the threshold value as the alternative MOOC set $M_i' = \{m_1^{i'}, m_2^{i'}, m_3^{i'}, ...\}$.

4) Rank M_i' by the enrollment number and select the top 1 as the corresponding prerequisite course m_i' of k_i.

5) By analogy, the prerequisite MOOCs of m could be assigned as $M' = \{m_1, m_2, ..., m_j\}$.

Then we find the most prerequisite relations mentioned in descriptions. Table 2 gives an example of mining prerequisite courses. "Marketing" is one of prerequisites to "Brand Management" and "Advanced College English" is prerequisite to "Marketing English".

To study the effect of this method, we select ICourse as the experimental bed. Eventually, for 2 197 MOOCs from ICourse, we find 2 979 prerequisite relations among 1 282 MOOCs. We randomly sample 300 prerequisites relations and the accuracy is 80%. We insist this simple and effective method is suitable for MOOCs with abundant properties.

Another side effect of this method is that we can study the pedagogy of MOOCs. An appropriate learning path could be presented in a multi-way tree structure where the higher the MOOCs, the more the prerequisites needed. For example, Fig.5 derives a prerequisite graph from Table 2. We find a clear path to master the course "Brand Management" and we need to study "Marketing", "Marketing English" and "Advanced College English" first. This is very helpful for junior students who are not familiar with a domain. We hope that this work serves as a step toward developing a data-driven model of learning path design.

Table 1. Concept Statistics

Platform	MOOCs with Directory	Number of Direct Links	Number of Calculated Links	Concept
ICourse	2 149 (97.8%)	134 376	14 378	21 412
XuetangX	1 076 (65.5%)	62 604	6 054	13 308
Total	3 225 (83.9%)	196 980	20 432	24 188

Note: "Direct Link" means the concept mentions are involved in Wikipedia entries while "Calculated Link" means the concept mentions are linked to the most similar entry by Algorithm 1. The statistics in the second column mean the numbers and percentages of "MOOCs with Directory" on three platforms.

Table 2. Example of Extraction from Prerequisite Description

No.	MOOC Name	Keywords in Prerequisite Description	Prerequisite MOOCs No.
151	Brand Management	Marketing, Consumer, Behavioral Science	826, 1719, 1823
477	Marketing English	CET6, English major	1596
826	Consumer Protection Act	/	/
1596	Advanced College English	/	/
1719	Marketing	Marketing, English	477
1823	Organizational Behavioral Science	/	/

Note: MOOCs in ICourse are described in Chinese. For illustrative purposes, we translate MOOC names and keywords into English. Besides, CET6 (College English Test Band 6) is an acknowledged English proficiency test in China.

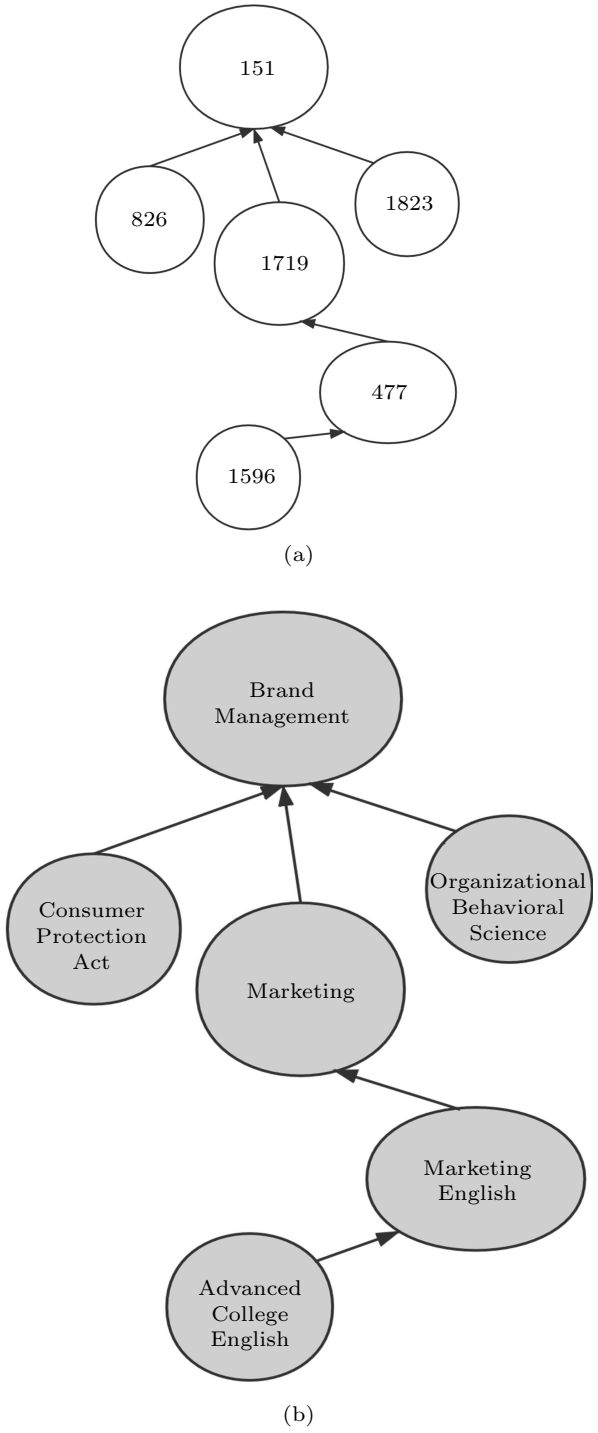


Fig.5. Example of learning path generated by prerequisite description. (a) Learning path of course numbers. (b) Learning path of course names.

3.3 Knowledge Fusion

The third step of our framework is knowledge fusion. To build a clean and condense knowledge graph for MOOCs, we need to carefully align different entities with similar names because different MOOCs platforms have a large portion of overlap in universities, instructors and courses. The interrelations among entities are also complicated, resulting in serious name repetition, normalization, incompleteness or inconsistencies [35]. In addition, aligning a class will affect other related classes.

In the construction of the knowledge graph, data cleansing techniques, such as grammatical normality and data normalization, are to be finished before the fusion. To reduce the computational cost, instances can be classified before matching. Obviously, XuetangX and ICourse are concentrated on MOOC resources in Chinese provided by Chinese universities and instructors. Edx and Cousrea are dominated by most of the Top 100 universities and mainly showed in English while few in minority languages. Therefore, we divide instances into Chinese-dominated and English-dominated blocks, and the alignment is performed according to the three classes, *Course*, *Teacher*, and *University*.

Then on account of that key attributes are in the form of short text, we use the semantic similarity of the respective attribute text to judge whether two similar entities match. Specifically, the editing distance would be calculated to estimate the text similarity as it can reduce the negative impact caused by the recording errors [36]. Through a test on examples, we choose the Sorensen distance for the fusion of instructor entities with the threshold of 0.5 and the Jaro distance for university entities with the threshold of 0.6. Finally, we keep all the attributes of metadata for the merged one to avoid information loss.

3.4 Knowledge Storage

So far, an educational knowledge graph which covers MOOCs on main platforms has been constructed with an integrated ontology and abundant instances. We store the whole knowledge graph into the Neo4j graph database. To prove the visual management and easy-to-query of Neo4j, we show the search results of MOOCs provided by Princeton University with related instructors and platform in Neo4j in Fig.6. And the blue node represents “Princeton” (university), the pink one, red ones and purple ones are “Coursera” (platform), MOOCs (courses) and instructors in turn. Here we map entities out to replace the screenshot for the higher resolution of Fig.6. And lines between courses and the university which means “the university offers courses” are hidden to make the figure clear.

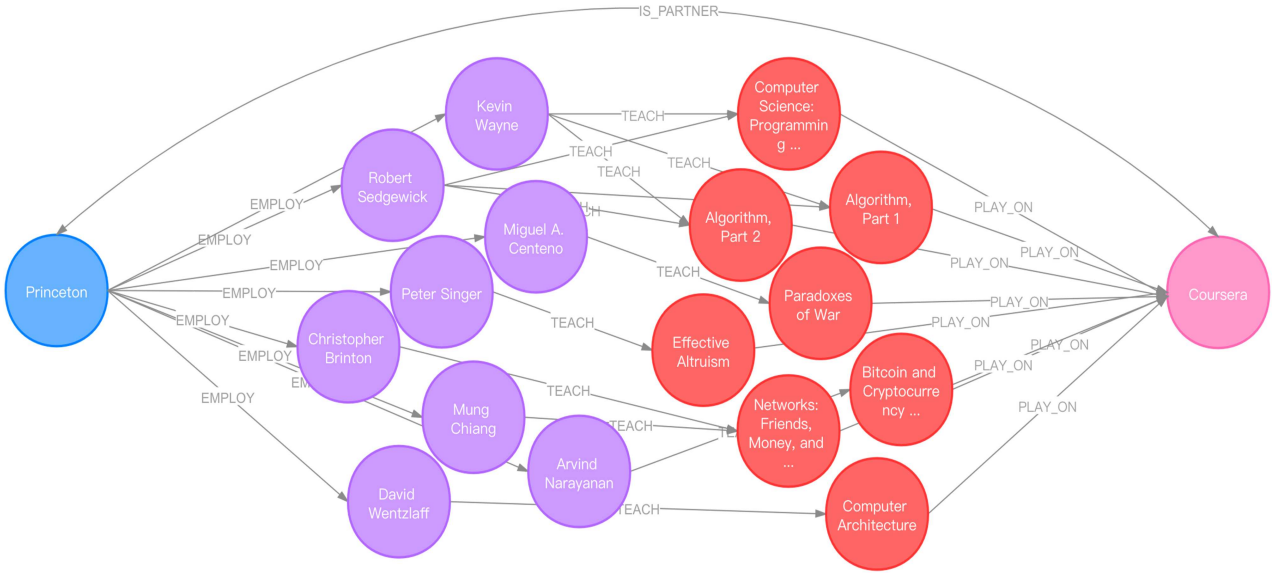


Fig.6. Search result of MOOCs provided by Princeton University with related instructors and platform in Neo4j.

For further utilization and updating, we have released the knowledge graph constructed by us on Github, the largest open development platform globally, and tend to add new resources periodically.

4 Analysis and Evaluation

The knowledge graph of MOOCs reorganizes cross-platform educational resources and uncovers implicit knowledge between heterogeneous data, which would facilitate further downstream applications, such as recommender systems, search engines and question-answer systems. With the knowledge graph of MOOCs, it is believed that we could dig insight into the characteristics of course entities and provide users with smart planning to support their decision making.

4.1 Overview of Knowledge Graph

This knowledge graph covers mainstream MOOC platforms globally with numerous MOOCs and properties. As we can see in Tables 3–5, there are five classes with 52 779 entities, including platforms, universities, instructors, concepts, and courses (MOOCs). Each entity (except the concept) has rich properties and relations with others so that more than 300 000 triples are contained in the Neo4j graph database. To the best of our knowledge, this is the largest knowledge graph of MOOC resources especially in the amount and scale to date.

Table 3. Numbers of Entities in Knowledge Graph

Platform	Course	University	Instructor	Concept
Coursera	3 664	179	3 375	-
EDX	1 718	108	3 139	-
XuetangX	1 733	258	3 149	21 412
ICourse	2 197	213	11 276	13 308
Total	9 312	604	18 671	24 188

Table 4. Numbers of Property Types in Knowledge Graph

Platform	Course	University	Instructor
Coursera	13	4	5
EDX	15	4	7
XuetangX	15	5	4
ICourse	15	4	9

Table 5. Numbers of Relations in Knowledge Graph

Entity Pair	Relation Type	Number
Course-Course	PRE_OF	2 979
Course-Platform	PLAY_ON	9 224
University-Course	OFFER	9 224
Instructor-Course	TEACH	22 912
Course-Concept	INVOLVE	217 412
University-Instructor	EMPLOY	18 671
University-Platform	IS_PARTNER	758

4.2 Comprehensive Comparison with Existing Knowledge Graph

In this subsection, we compare our knowledge graph with an existing knowledge graph of MOOCs, HEKG [37]. Even though both knowledge graphs fo-

cus on MOOCs with related entities, our study distinguishes in following aspects.

• *Emphasis of Research.* Our research is concentrated on reorganization and representation of MOOC resources while HEKG emphasizes more on the correlation between MOOCs and they generate three learning scenes, course groups for similar courses, contrastive relations as well as sequence relations.

• *Instance Amount and Scale.* We extract 52 779 instances which include 9 312 MOOCs from Coursera, EDX, XuetangX, and ICourse. HEKG collects 1 225 MOOCs from Coursera, Open163, XuetangX, and ICourse. We insist that EDX is the second biggest MOOC platform globally which should contain significant MOOC resources.

• *Number of Properties.* Our research collects as many as possible properties as shown in Table 4 and saves more than 100 000 property triples. While HEKG does not give a precise number of properties or relations among MOOCs.

• *Way of Concept Extraction.* We process the directory of each MOOC to acquire concept mentions and links with Wikipedia to redefine concepts, which expresses knowledge in a more union and precise way. HEKG uses the TF-IDF algorithm to process the description and the content of courses to get concepts, which we think is a typical but not very successful way because of the text sparsity.

5 Conclusions

To organize online educational resources effectively and unfold the internal characteristics of MOOCs, we built a knowledge graph to represent and store detailed information about MOOCs from Coursera, EDX, XuetangX, and ICourse. We built an ontology to model relations of MOOC entities and used data mining methods to retrieve key information. Entity disambiguation was utilized to improve precision and organization. We extracted a great number of concepts to enrich the semantic representation of MOOCs which gives the pedagogical coverage in a certain way. Moreover, all concepts were linked to Wikipedia entries. Prerequisite relations were inferred by MOOC description to provide effective guidance for general learners.

Our knowledge graph contains 9 312 courses, 604 universities, 18 671 instructors, 24 188 concepts and four platforms. It is worth noticing that this has been the largest knowledge graph of MOOC resources so far，which offers the most unified-format and qualified MOOC resources as metadata for the other researches on the knowledge graph.

Future research directions would be to improve the accuracy of links with Wikipedia entries and utilize extracted concepts to estimate prerequisite relations. For concept mentions not in Wikipedia entries, enlarging Wikipedia entry set available or optimization on the algorithm would be taken into account.

References

[1] Almeda M V, Zuech J, Utz C, Higgins G, Reynolds R, Baker R S. Comparing the factors that predict completion and grades among for-credit and open/MOOC students in online learning. *Journal of Interactive Online Learning*, 2018, 22(1): 1-18. DOI: 10.24059/olj.v22i1.1060.

[2] Onah D F O, Sinclair J, Boyatt R. Dropout rates of massive open online courses: Behavioural patterns. In *Proc. the 6th International Conference on Education and New Learning Technologies*, July 2014, pp.5825-5834. DOI: 10.13140/RG.2.1.2402.0009.

[3] Gardner J, Brooks C. Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, 2018, 28(2): 127-203. DOI: 10.1007/s11257-018-9203-z.

[4] Lin Y, Liu Z, Sun M, Liu Y, Zhu X. Learning entity and relation embeddings for knowledge graph completion. In *Proc. the 29th AAAI Conference on Artificial Intelligence*, January 2015, pp.2181-2187.

[5] Wang X, Feng W, Tang J, Zhong Q. Course concept extraction in MOOC via explicit/implicit representation. In *Proc. the 3rd IEEE International Conference on Data Science in Cyberspace*, June 2018, pp.339-345. DOI: 10.1109/DSC.2018.00055.

[6] Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes P N, Hellmann S, Morsey M, van Kleef P, Auer S, Bizer C. DBpedia-A largescale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 2015, 6(2): 167-195. DOI: 10.3233/SW-140134.

[7] Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. *Scientific Reports*, 2017, 7(1): Article No. 5994. DOI: 10.1038/s41598-017-05778-z.

[8] Miao Q, Meng Y, Zhang B. Chinese enterprise knowledge graph construction based on linked data. In *Proc. the 9th IEEE International Conference on Semantic Computing*, February 2015, pp.153-154. DOI: 10.1109/ICOSC.2015.7050795.

[9] Szekely P, Knoblock C A, Slepicka J *et al.* Building and using a knowledge graph to combat human trafficking. In *Proc. the 14th International Semantic Web Conference*, October 2015, pp.205-221. DOI: 10.1007/978-3-319-25010-6_12.

[10] Qiu B S, Zhao W. Student model in adaptive learning system based on semantic web. In *Proc. the 1st International Workshop on Education Technology and Computer Science*, March 2019, pp.909-913. DOI: 10.1109/ETCS.2009.466.

[11] Chaplot D S, Yang Y, Carbonell J, Koedinger K R. Data-driven automated induction of prerequisite structure graphs. In *Proc. the 9th International Conference on Educational Data Mining*, June 2016, pp.318-323.

[12] Sun K, Liu Y, Guo Z, Wang C. EduVis: Visualization for education knowledge graph based on web data. In *Proc. the 9th International Symposium on Visual Information Communication and Interactions*, September 2016, pp.138-139. DOI: 10.1145/2968220.2968227.

[13] Chen P, Lu Y, Zheng V W, Chen X, Yang B. KnowEdu: A system to construct knowledge graph for education. *IEEE Access*, 2018, 6: 31553-31563. DOI: 10.1109/ACCESS.2018.2839607.

[14] Alsaad F, Boughoula A, Geigle C, Sundaram H, Zhai C. Mining MOOC lecture transcripts to construct concept dependency graphs. In *Proc. the 11th International Conference on Educational Data Mining*, July 2018.

[15] Yang Y, Liu H, Carbonell J, Ma W. Concept graph learning from educational data. In *Proc. the 8th ACM International Conference on Web Search & Data Mining*, February 2015, pp.159-168. DOI: 10.1145/2684822.2685292.

[16] Duan Y, Shao L, Hu G, Zhou Z, Zou Q, Lin Z. Specifying architecture of knowledge graph with data graph, information graph, knowledge graph and wisdom graph. In *Proc. the 15th IEEE International Conference on Software Engineering Research, Management and Applications*, June 2017, pp.327-332. DOI: 10.1109/SERA.2017.7965747.

[17] Francesconi E, Montemagni S, Peters W, Tiscornia D. Integrating a bottom-up and top-down methodology for building semantic resources for the multilingual legal domain. In *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, Francesconi E, Montemagni S, Peters W, Tiscornia D (eds.), Springer, 2010, pp.95-121. DOI: 10.1007/978-3-642-12837-0_6.

[18] Li J, Jun Z, Chen H, Liu Z, Sun L, Hou L, Xu B, Peng P. Knowledge mapping development report. Technical Report, Chinese Information Society Language and Knowledge Computing Committee, 2018. http://cips-upload.bj.bcebos.com/KGDevReport2018.pdf, June 2020. (in Chinese)

[19] Xie G. Review of knowledge graph refinement. *Application of Electronic Technique*, 2018, 44(9): 29-33. DOI: 10.16157/j.issn.0258-7998.180696. (in Chinese)

[20] Adelberg B. NoDoSE—A tool for semi-automatically extracting structured and semi-structured data from text documents. In *Proc. the 1998 ACM SIGMOD International Conference on Management of Data*, June 1998, pp.283-294. DOI: 10.1145/276304.276330.

[21] Ching N. The Best MOOC Platforms. Reviews.com. 2018. https://www.reviews.com/mooc-platforms/, October 2019.

[22] Seaton D T, Bergner Y, Chuang I, Mitros P, Pritchard D E. Who does what in a massive open online course? *Communications of the ACM*, 2013, 57(4): 58-65. DOI: 10.1145/2500876.

[23] Mesbah S, Chen G, Torre M V, Bozzon A, Lofi C, Houben G J. Concept focus: Semantic meta-data for describing MOOC content. In *Proc. the 13th European Conference on Technology Enhanced Learning*, September 2018, pp.467-481. DOI: 10.1007/978-3-319-98572-5_36.

[24] Effland T D. Focused mining of university course descriptions from highly variable sources. In *Proc. the 46th ACM Technical Symposium on Computer Science Education*, March 2015, Article No. 716. DOI: 10.1145/2676723.2693630.

[25] Atapattu T, Falkner K, Falkner N. A comprehensive text analysis of lecture slides to generate concept maps. *Computers & Education*, 2017, 115: 96-113. DOI: 10.1016/j.compedu.2017.08.001.

[26] Wang S, Ororbia A, Wu Z, Williams K, Liang C, Pursel B, Giles C L. Using prerequisites to extract concept maps from textbooks. In *Proc. the 25th ACM International Conference on Information and Knowledge Management*, October 2016, pp.317-326. DOI: 10.1145/2983323.2983725.

[27] Pan L, Wang X, Li C, Li J, Tang J. Course concept extraction in MOOCs via embedding-based graph propagation. In *Proc. the 8th International Joint Conference on Natural Language Processing*, November 27-December 1, 2017, pp.875-884.

[28] Yu J, Wang C, Luo G, Hou L, Li J, Liu Z, Tang J. Course concept expansion in MOOCs with external knowledge and interactive game. arXiv:1909.07739, 2019. https://arxiv.org/abs/1909.07739, June 2020.

[29] Krishnan A, Sankar A, Zhi S, Han J. Unsupervised concept categorization and extraction from scientific document titles. In *Proc. the 2017 ACM Conference on Information and Knowledge Management*, November 2017, pp.1339-1348. DOI: 10.1145/3132847.3133023.

[30] Pang K, Tang J, Wang T. Which embedding level is better for semantic representation? An empirical research on Chinese phrases. In *Proc. the 7th CCF International Conference on Natural Language Processing and Chinese Computing*, August 2018, pp.54-66. DOI: 10.1007/978-3-319-99501-4_5.

[31] Liang C, Wu Z, Huang W, Giles C L. Measuring prerequisite relations among concepts. In *Proc. the 2015 Conference on Empirical Methods in Natural Language Processing*, September 2015, pp.1668-1674. DOI: 10.18653/v1/D15-1193.

[32] Mikolov T, Chen K, Corrado G S, Dean J. Distributed representations of words and phrases and their compositionality. In *Proc. the 27th Annual Conference on Neural Information Processing Systems*, December 2013, pp.3111-3119.

[33] Gresch H, Bögeholz S. Identifying non-sustainable courses of action: A prerequisite for decision-making in education for sustainable development. *Research in Science Education*, 2013, 43(2): 733-754. DOI: 10.1007/s11165-012-9287-0.

[34] Vuong A, Nixon T, Towle B. A method for finding prerequisites within a curriculum. In *Proc. the 4th International Conference on Educational Data Mining*, July 2011, pp.211-216.

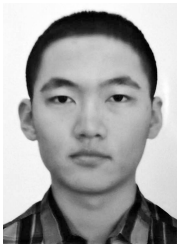
[35] Dong X L, Gabrilovich E, Heitz G, Horn W, Murphy K, Sun S, Zhang W. From data fusion to knowledge fusion. *Proceedings of the VLDB Endowment*, 2014, 7(10): 881-892. DOI: 10.14778/2732951.2732962.

[36] Zhong H, Zhang J, Wang Z, Wan H, Chen Z. Aligning knowledge and text embeddings by entity descriptions. In *Proc. the 2015 Conference on Empirical Methods in Natural Language Processing*, September 2015, pp.267-272. DOI: 10.18653/v1/D15-1031.

[37] Zheng Y, Liu R, Hou J. The construction of high educational knowledge graph based on MOOC. In *Proc. the 2nd IEEE Information Technology, Networking, Electronic and Automation Control Conference*, December 2017, pp.260-263. DOI: 10.1109/ITNEC.2017.8284984.

Fu-Rong Dang received her M.S. degree in software engineering from National University of Defense Technology, Changsha, in 2019. Her research interests include natural language processing and knowledge graph.

Jin-Tao Tang received his Ph.D. degree in computer science from National University of Defense Technology (NUDT), Changsha, in 2011. He is an associate professor in the College of Computer Science and Technology in NUDT, Changsha. His research work mainly focuses on natural language processing, open information extraction, and construction of social network.



Kun-Yuan Pang received his M.S. degree in software engineering from National University of Defense Technology (NUDT), Changsha, in 2016. Since 2017, he has been an Ph.D. candidate with the College of Computer Science and Technology, NUDT, Changsha. His research interests include machine learning, natural language processing and information extraction.

Ting Wang received his Ph.D. degree in computer software from National University of Defense Technology (NUDT), Changsha, in 1997. He is a professor and Ph.D. supervisor in the College of Computer Science and Technology in NUDT, Changsha. His research work mainly focuses on natural language processing, information retrieval, and semantic web.

Sha-Sha Li received her Ph.D. degree in computer science from National University of Defense Technology (NUDT), Changsha. She has been an assistant professor in NUDT, Changsha, since 2012. She currently focuses on natural language processing and automatic construction of knowledge graph.

Xiao Li received his Ph.D. degree in computer science from the University of Western Ontario, London, in 2013. He is currently an associate professor in National University of Defense Technology, Changsha. His research interests include recommender system, information retrieval and its application in distributed system and online education.