Knowledge graph construction for computer networking course group in secondary vocational school based on multi-source heterogeneous data

Gang Li, Hong Wang*, Hong Liu

School of Information Science and Engineering, Shandong Normal University

Jinan, China

e-mail: wanghong106@163.com

Abstract—Computer science major in secondary vocational schools focuses on cultivating students' ability to solve complex engineering problems. Thus both theoretical and practical teaching of professional courses are critically important. Due to the rapid update of knowledge in the computer domain, it is becoming challenging to better integrate theory and practice knowledge for teaching with advanced intelligent information technology. This paper constructs the knowledge graph for the computer networking course group in secondary vocational schools. We first propose methods for data acquisition, keyword extraction, relationship extraction, and knowledge storage. Then, we describe the complete process for constructing the knowledge graph. Finally, we visualize the relationship between knowledge points for theoretical and practical courses with the tool SmartKG. Our proposed method provides new ideas for teaching computer courses in secondary vocational schools, which helps cultivate high-quality skilled talents.

Keywords-computer networking course group, knowledge graph, SmartKG, secondary vocational schools

I. INTRODUCTION

In 2022, the Ministry of Education announced its latest work priorities, mentioning the implementation of the Education Digitalization Strategy Initiative, which is strengthening demand, deepening integration, application-driven and actively promoting "Internet+Education", accelerate the digitalization and intelligent transformation and upgrading of education [1]. As the beginning of vocational education, secondary vocational schools are responsible for basic education and teaching, and provide a solid foundation for the success and development of students. Therefore, it is crucial to explore new educational and teaching methods in secondary vocational schools. By adopting a rich supply of digital education resources and services, the foundation is laid for exploring smart classroom construction and improving classroom teaching models. At present, in order to improve the quality of teaching, enhance students' interest in learning and strengthen their abilities, it is important to diversify and enrich teaching and learning activities by integrating internet technologies into teaching and learning, using a variety of open source software and devices.

Although knowledge graph has been developed for several years, it is still at the initial stage of application in the field of education. In 2019, the "National Education Informatization Work Conference" pointed out that promoting the development of key technologies such as big data, deep learning, and knowledge graph is one of the main elements of basic science research in education [2]. This paper applies SmartKG to realize the construction and application of the knowledge graph for computer networking course group in secondary vocational schools. The approach we propose new ideas for the integration and innovation of information technology and educational teaching, as well as for the pedagogical development of computer courses.

As a fundamental course in computer science, Computer Network covers a wide range of knowledge and has a strong theoretical. It has a strong correlation with practical courses, such as Network Equipment Installation and Network Integrated Cabling. In the curriculum arrangement of secondary vocational school, these courses are arranged in different semesters, which result in the weak correlation ability of students to knowledge points and the inability to learn knowledge accurately and effectively. At the same time, due to the different basic levels of students, it is impossible to uniformly measure their mastery of knowledge. However, as a set of knowledge points, knowledge graph not only facilitates students' evocation and extension of knowledge points, but also enriches the classroom format and strengthens students' desire for learning.

II. RELATED CONCEPTS AND TECHNOLOGIES

A. Knowledge Graph

It has been more than ten years since the concept of the knowledge graph was introduced. Knowledge graph is a type of knowledge base known as semantic network, which is a knowledge base with a directed graph structure. Knowledge graph consists of vertices and edges, where each vertex corresponds to an entity and each edge corresponds to a relationship between entities [3]. The Education Knowledge Graph is an ontology construction of knowledge about a subject, which is stored and presented as a graph and can provide learning recommendation paths [4].

During the development of the Knowledge Graph, numerous more mature projects have emerged. For example, Common Knowledge Repository: Cyc, WordNet and so on; Internet Knowledge Graph: Freebase, DBpedia, Schema, Wikidata and so on; Domestic Knowledge Graph: OpenKG, CN-DBpedia, open class Chinese Encyclopedia Knowledge Graph (zhishi.me) and so on [5].

At present, knowledge graphs are widely used in various industries. For example, the knowledge graph enhances

urban management capabilities, which enables the mutual integration and sharing of data from all areas of production and life, and helps to improve the quality of urban management services and people's living standards. The public security knowledge graph, which enables the association of key people, time, place, events, objects and other entities, and helps to improve the efficiency of public security personnel.

B. Knowledge Graph Construction Process

The construction of knowledge graph mainly adopts two methods of top-down and bottom-up [3]. The top-down method is used to construct the domain knowledge graphs, which takes the data within the domain or the enterprise as the main source and extract ontology and pattern information from high quality data to form knowledge graphs. This method requires high data quality and the resulting knowledge graphs are extremely accurate. The bottom-up method is mainly used to construct the general knowledge graphs, which extracts data from various public data sets on the Internet with the help of certain technical means to form knowledge graphs. This method generally contains a large number of things and a variety of common knowledge information.

The process of building a knowledge graph mainly consists of data collection, knowledge extraction, knowledge integration and knowledge application. The technical architecture of the knowledge graph is shown in Figure 1.
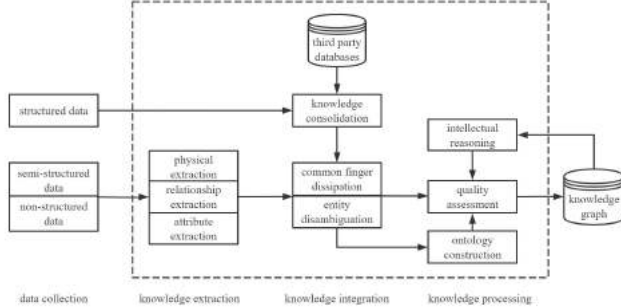

Figure 1.   The technical architecture of the knowledge graph

Knowledge acquisition refers to the adoption of various ways and methods to obtain the data, including structured data, non-structured data and semi-structured data, and ultimately to extract structured information such as entities, relationships and entity attributes from the acquired data [6]. The key techniques involved in knowledge acquisition mainly include entity extraction, relationship extraction, attribute extraction, etc.

Knowledge representation refers to the representation of knowledge and the relationships between knowledge in a way that humans can carry out to recognise and understand it. It can also be understood as the representation of various types of knowledge in the real world as a computer storable and computable structure [7]. Currently, knowledge

representation method is mainly described in the form of RDF (Resource Description Framework) triples.

Through knowledge acquisition, the results obtained may contain a large amount of redundant data, while there is a flattening of data relationships between them, lacking hierarchy and logic. In order to solve the above problems, a knowledge integration approach has been adopted. The key technologies involved mainly include entity linking and knowledge fusion.

The results obtained through knowledge acquisition and knowledge integration are not equal to knowledge. To obtain a structured and networked knowledge system, it is necessary to go through the process of knowledge processing. The key technologies involved in knowledge processing mainly include ontology construction, knowledge reasoning and quality assessment [8].

The content contained in the knowledge graph needs to be constantly updated with the changes of the time. And there are two methods of updating it. The first method is a complete update, which means building a Knowledge Graph from scratch. This method is relatively simple, but requires a lot of resources. The second method is incremental updating, which means adding new knowledge based on the existing knowledge graph. This method is more complex, but consumes less resources [6].

C. SmartKG

SmartKG is an open source knowledge graph building product developed by Microsoft. The product is mainly used in knowledge graph teaching systems and is a lightweight tool for verifying the ontology design of knowledge graphs. SmartKG helps users to quickly build and visualise knowledge graphs. At the same time, according to the user-defined knowledge graphs, the system will automatically generate CheckBot with the graph as the knowledge base for simple question and answer operations.

The user describes the knowledge graph by using the corresponding template provided by SmartKG, which is an Excel document. The document contains two tables, one is an entity table and the other is a relationship table. Each row corresponds to an entity or a relationship. The entity table contains the entity ID, entity name, entity label, lead word, attribute name and attribute value. The relationship table contains the relationship type, source entity ID, and target entity ID. The specific meaning is shown in Table 1.

TABLE I.   SmartKG TABLE PROPERTIES

Name	connotation
entity ID	ID number
entity name	Node Name
entity label	Node Type
lead word	Auxiliary as CheckBot
attribute name	Attribute name
attribute value	Instance of attribute name
relationship type	Relationship Name
source entity ID	ID of the header entity to be connected
target entity ID	Required The ID of the tail entity to be connected

SmartKG has the following features: (a) Visualization: The whole process is including template filling, data uploading and data presentation can be operated intuitively. (b) High performance: The product has a built in high performance graph database engine that allows users to set up hundreds or even thousands of entity nodes and display the relationships that exist between them. (c) Easy to extend: Users can target the entities and their relationships to be added by making changes to the table data, and timely updates can be made to display new diagrams. (d) Intelligent: Users can search and query to find a node and its corresponding relationships after generating a knowledge graph using SmartKG tools. At the same time, the SmartKG tool generates a corresponding knowledge base based on the documents provided by the user, which can be used for question and answer operations.

III. KNOWLEDGE GRAPH CONSTRUCTION FOR COMPUTER NETWORK COURSE CLUSTERS

In this paper, the bottom-up method is used for knowledge graph construction. Firstly, the data is acquired from data sources such as e-textbooks, Zhihu and w3cschool. Then, entity identification, relationship extraction and attribute extraction are performed on the acquired data. Finally, the processed structured data is collated and imported into SmartKG to complete visualisation and other tasks.

A. Data Acquisition

Due to regional differences and teaching resources, secondary vocational schools choose different textbooks, but their teaching contents are generally the same. Therefore, this paper selects Jingdong Bookstore, w3cschool, Baidu Encyclopedia and Zhihu as the network data sources and crawls the data through python.

In Jingdong Bookstore, input "computer network" and select the keyword "secondary vocational school and technical materials" to get hundreds of related books. Each book contains information such as "Content Introduction", "Author Introduction", "Contents", "Foreword", and so on. This paper focuses on obtaining the "Contents" information. In w3cschool, by searching the keyword "Computer Network", and crawl the data from the two parts of the computer network basic knowledge summary and TCP / IP tutorial. In the blog park, input the keyword "Computer Network Knowledge" to crawl the data of the computer network basics summary.

In addition, this experiment manually processes the content of the theoretical course "Fundamentals of Computer Network Technology" and the practical course "Installation and Commissioning of Network Equipment" for the secondary vocational school information security majors, to prepare for the subsequent keyword extraction, relationship extraction and entity attribute.

B. Keyword Extraction

A key step in building a knowledge graph is through keyword extraction of the acquired data. There are many algorithms for keyword extraction, and this paper mainly applies the LDA document topic generation model, the statistical-based TF-IDF algorithm, and the network graph-based TextRank algorithm for keyword extraction.

LDA (Latent Dirichlet Allocation), also known as the three-layer Bayesian probabilistic model, is to discover the subject information contained in documents through unsupervised learning. It consists of a three-layer structure of words, topics and documents. By using the co-occurrence relationship of words in the document to cluster the words, two probability distributions of "document topic" and "topic word" are obtained [9]. The relationships are shown in Figure 2.
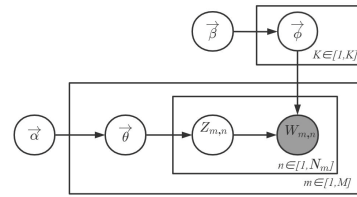

Figure 2. LDA model

TF-IDF is a statistical method that measures the importance of a word to a document. The keywords of the document are extracted by using the statistical information of the words in the document. Its advantages are simplicity, ease of implementation and generalisability. TF-IDF not only considers the frequency of occurrence of a word TF, but also considers the inverse frequency IDF that the word does not appear in other documents, which gives an excellent indication of the differentiation of feature words and is a widely used retrieval method in the field of information retrieval [10]. Its formula is as follows:

$$tfidf_{i,j} = tf_{i,j} * idf_i \qquad (1)$$

The tf represents word frequency. If a word occurs a total of i times in the document and there are N words in the entire document, the value of tf is:

$$tf = \frac{i}{N} \qquad (2)$$

The idf is the inverse document frequency. If the whole document has n articles and a word appears in k articles, the value of idf is:

$$idf = log\frac{n}{k} \qquad (3)$$

The TextRank algorithm is a text ranking algorithm based on network graphs. One aspect of the network graph-based algorithm that differs from the above algorithms is that both the topic-based and statistical analysis methods require an off-the-shelf corpus, whereas the TextRank algorithm can be used to extract keywords from a single document out of the context of a corpus [11]. Its core formula is as follows:

$$WS(V_i) = (1 - d) + d * \sum_{V_i \in (V_i)} \frac{W_{ji}}{\sum_{V_{k \in out(v_j)}} W_{jk}} \qquad (4)$$

Through the above three algorithms, this paper extracts keywords from the crawled data in turn. At the same time, stop words were added to filter out irrelevant words during the experiment. The final ranking was done by keyword weight, and Table 2 shows the top 10 keywords for each algorithm.

TABLE II.   Top 10 keywords

LDA_model		Textrank		TF-IDF	
Keywords	Weight	Keywords	Weight	Keywords	Weight
Protocol	0.0351	Network	0.0351	Practical	0.1643
Network	0.0205	Address	0.0182	IP	0.1485
IP	0.0195	Protocol	0.0135	Address	0.1304
TCP	0.0175	Data	0.0134	TCP	0.1124
Data	0.0152	LAN	0.0119	Network	0.1056
Computer	0.0136	Network Card	0.0103	Protocol	0.1047
Servers	0.0127	Computer	0.0094	IP Address	0.0923
LAN	0.0127	Task	0.0094	Host	0.0882
Technology	0.0118	IP Address	0.0093	Task	0.0861
Internet	0.0109	Servers	0.0088	Servers	0.0704

C.   Relationship Extraction

Relationship extraction connects entities to each other and it is the key step to form knowledge graph. Commonly used methods for relationship extraction include rule-based methods, supervised learning methods, semi-supervised learning methods and unsupervised learning methods. Due to the small size of this experimental dataset, a rule-based approach combined with manual assistance is used for relationship extraction in order to improve the accuracy of the results.

Most rule-based methods rely on manual work, with more accurate results and no need for training data. As far as possible in the dataset, various hidden relationships such as dependencies and inclusion relationships between entities are found. In this paper, partial rules are first defined, and then relationship extraction is performed based on these rules. Some of the rules are shown in Table 3.

TABLE III.   Relationship extraction rules

Rules	Contents
Rule 1	X contains/is covered by Y
Rule 2	X relates to/is related to Y
Rule3	X is divided into Y
......	......

Based on the above table, the relationships between entities (X,contains,Y), (X,associated,Y) and other relationships can be extracted when the relevant content is matched. Finally, a total of 521 relationships were extracted in this paper.

D.   Knowledge Storage

Through data crawling, keyword extraction and relationship extraction and other related work, our last step is to store and visualize knowledge. After comparing and measuring the extracted keywords with those obtained by manual construction, 239 entities were finally selected. At the same time, the attributes contained in each entity are completed manually. This paper uses SmartKG for knowledge storage and the results are shown in Figure 3.
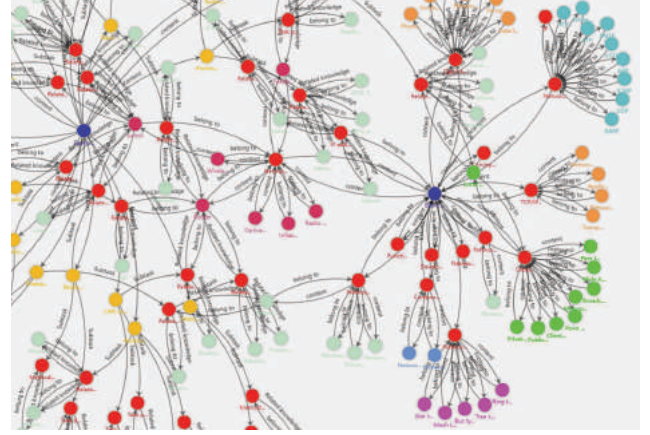

Figure 3.   Computer network knowledge graph

This paper focuses on the keywords "computer network basics" and "experimental tasks" in order to derive relevant knowledge. In total of 239 entities and 521 relationships. In the diagram, each node has a different colour, which means that it belongs to a different entity label. In this experiment, the acquired entities were classified into "Classification", "Knowledge Point", "Network structure level", "Concept", "Task" and other labels. The relationships involved are "Contains", "Belongs to", "Associated Knowledge", "Subtask" and so on.

SmartKG can search for knowledge by querying, as shown in Figure 4. By inputting the keyword "Router", and then the relevant knowledge points will appear. Take the knowledge point "Router" as an example, it is both a key knowledge point in the theory class and a equipment that is required everywhere in the practical class. As the teaching requires students to have a high level of knowledge about routers and the ability to apply this knowledge, it can be displayed in the form of knowledge graph, which can enable students to master the required knowledge content in different learning periods. As shown in Figure 5, the router knowledge graph contains the basic attributes "definition", "english name", "function", "performance indicator" and "port". Moreover, it is one of the network transmission media. At the same time, the knowledge point "Router" involves many experimental tasks, such as "Basic configuration of a router" "LAN users accessing the Internet through a router" "Accessing the Internet through a wireless routing relationship", and so on.
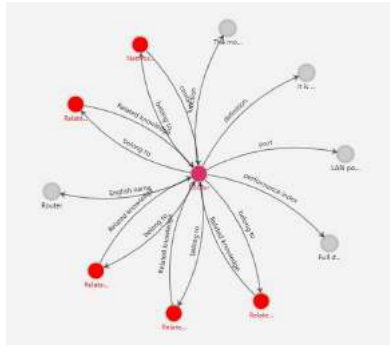
102

Figure 4.　Query function

Figure 5.　Router knowledge points

The knowledge graph and research method discussed in this paper have been applied to the teaching practice. This study selected two classes and set them as control group and experimental group respectively. In the experimental group, the constructed knowledge graph were applied in the teaching practice process; in the control group, the lessons were taught according to the conventional teaching methods. By analyzing the multi-dimensional data such as students' interest in and enjoyment of the course, students' participation in classes, and students' academic achievements. The results of the two classes are compared as shown in Table 4. The research shows that the application of knowledge graph in the teaching process can improve students' academic performance and promote the improvement of teaching quality.

TABLE IV.　Score analysis

	Class	Number	Average value	Standard deviation	Standard Error of Mean
score	Class 1	25	78.52	7.397	1.679
	Class 2	25	69.92	9.142	1.828
Question type 1	Class 1	25	22.64	2.123	.425
	Class 2	25	21.44	2.752	.550
Question type 2	Class 1	25	23.20	3.266	.653
	Class 2	25	20.48	3.525	.705
Question type 3	Class 1	25	32.68	4.388	.898
	Class 2	25	28.00	4.555	.911

IV.　CONCLUSION

This paper takes the teaching content of the computer network course group as the core, involving theory and experimental operation. Firstly, we introduce the overview of knowledge graph, the construction process and the use of SmartKG application software; then we elaborate the methods applied in this paper and the experimental results in turn from four aspects: data acquisition, keyword extraction, relationship extraction and knowledge storage. The key to teaching computer networks is the common development of theory and practice, not only to cultivate students' mastery of theoretical knowledge, but also to focus on students' practical operation ability. In the process of rapid construction and development of a skill-based society, it is the key to vocational education to let students enjoy and apply their skills.

REFERENCES

[1] "Key points of work of the Ministry of education in 2022: implementing the digital education strategy," Modern education technology, vol. 32, Feb. 2022, pp. 1.

[2] Ministry of Education: "2019 national education informatization work conference," Basic education courses, vol. 09, May. 2019, pp. 4.

[3] Liu Qiao, Li Yang, Duan Hong, Liu Yao, Qin Zhiguang. "Overview of knowledge graph construction technology," Computer research and development, vol. 53, Mar. 2016, pp. 582-600, doi:10.7544/issn1000-1239.2016.20148228.

[4] Li Zhen, Zhou Dongdai, Wang Yong. "Educational knowledge graph under the vision of "Artificial Intelligence +": connotation, technical framework and application research," Journal of distance education, vol. 37, Jul. 2019, pp. 42-53, doi:10.15881/j.cnki.cn33-1304/g4.2019.04.006.

[5] Gao Mao, Zhang Liping. "Research on the connotation, technology and application of educational knowledge graph integrating multi-modal resources," Computer application research, vol. 12, Aug. 2022, pp. 1-12, doi:10.19734/j.issn.1001-3695.2021.12.0686.

[6] Hang Tingting, Feng Jun, Lu Jiamin. "Knowledge graph construction technology: classification, investigation and future direction," Computer science, vol. 48, Feb. 2021, pp. 175-189, doi:10.11896/jsjkx.200700010.

[7] Wang Zhe, Li Yaqi, Feng Xiaohui, Wang Cuilin. "Development trend and thinking prospect of artificial intelligence in Education,"]. Artificial intelligence, vol. 03, Jun. 2019, pp. 15-21, doi:10.16453/j.cnki.issn2096-5036.2019.03.003.

[8] Liu yechen, Li Huayu. "Overview of domain knowledge atlas,". Computer system application, vol. 29, Jun. 2020, pp. 1-12, doi:10.15888/j.cnki.csa.007431.

[9] Yang Mengmeng, Huang Hao, Cheng LuHong, Ma Ping, Bao WuJie. "Short text classification based on LDA theme model," Computer engineering and design, vol. 37, Dec. 2016, pp. 3371-3377, doi:10.16208/j.issn1000-7024.2016.12.044.

[10] Wang Gensheng, Huang Xuejian. "Convolutional neural network text classification model based on word2vec and improved TF-IDF," Small microcomputer system, vol. 40, May. 2019, pp. 1120-1126, doi:10.3969/j.issn.1000-1220.2019.05.036.

[11] Liu Xiaojian, Xie Fei, Wu Xindong. "Keyword extraction algorithm based on graph and LDA topic model," Journal of information science, vol. 35, Jun. 2016, pp. 664-672, doi:10.3772/j.issn.1000-0135.2016.006.010.