International Journal of Innovation and Research in Educational Sciences
Volume 10, Issue 2, ISSN (Online) : 2349–5219

Research on the Construction of Knowledge Graph of Education Big Data with Multi-Source Data

Zhiwen Ren, Jiashan Liu and Shanwen Zhang*
School of Electronic Information, Xijing University, Shaanxi 710123 China.
*Corresponding author email id: wjdw716@163.com

Date of publication (dd/mm/yyyy): 31/03/2023

***Abstract* – With the advent of the computer age, education has also entered the development path of informatization and intelligence, and education is a national plan. Wisdom education has attracted more and more attention. The knowledge graph in the field of education is the cornerstone of wisdom education, and the intelligence of education cannot be separated from the knowledge graph of education. Taking the construction of educational knowledge graph as the goal, this paper puts forward the construction method of educational knowledge graph with multi-source data from the technologies of data acquisition, data processing, entity extraction, entity connection and integration. The construction of educational knowledge graph is a key step of educational informatization and the basic work of educational intelligence.**

***Keywords* – Knowledge Graph, Education Big Data, Data Processing, Big Data, BILSTM.**

I. Research Status

Computers promote education to education informatization, and now the increasingly popular big data further promotes education informatization to education intelligence. The educational system of colleges and universiti--es, such as educational administration system and student-worker system, will produce a large number of educational data, which contain information about students' learning behavior and life behavior, and this information can serve education in reverse, however, few people pay attention to how to use big data to promote students' learning and growth. The premise of using big data is to have a knowledge base [1]. How to integrate a large number of educational data into a structured knowledge base has become one of the current research hotspots. This paper is a study on the construction of educational big data knowledge graph with multi-source data.

II. Research Significance

Through the data analysis of multi-source and heterogeneous educational big data, the educational knowledge graph is constructed, and the learning behavior characteristics of students are obtained, which can provide support for the adaptive learning system. In their research, Li et al. emphasized the importance of integrating educational psychology with data technology in the critical period of educational development. Combining knowledge graph with educational big data is conducive to the promotion of personalized learning resources, the provision of personalized learning paths and the personalized development of students. The development of wisdom education is a key step.

III. Research on the Construction of Knowledge Graph Model in Education Field

Firstly, a large number of data are obtained from the academic affairs office and logistics office of the school, and then information is extracted from all kinds of data through semantic technology. The relationship between student entities and other entities is mined with emphasis on student entities, and data is stored and visualized by using RDF and Secondary. The constructed knowledge graph can provide data support for educational intelligent question answering system, adaptive learning system and learning path recommendation system. Fig. 1 is the construction idea of educational knowledge graph.

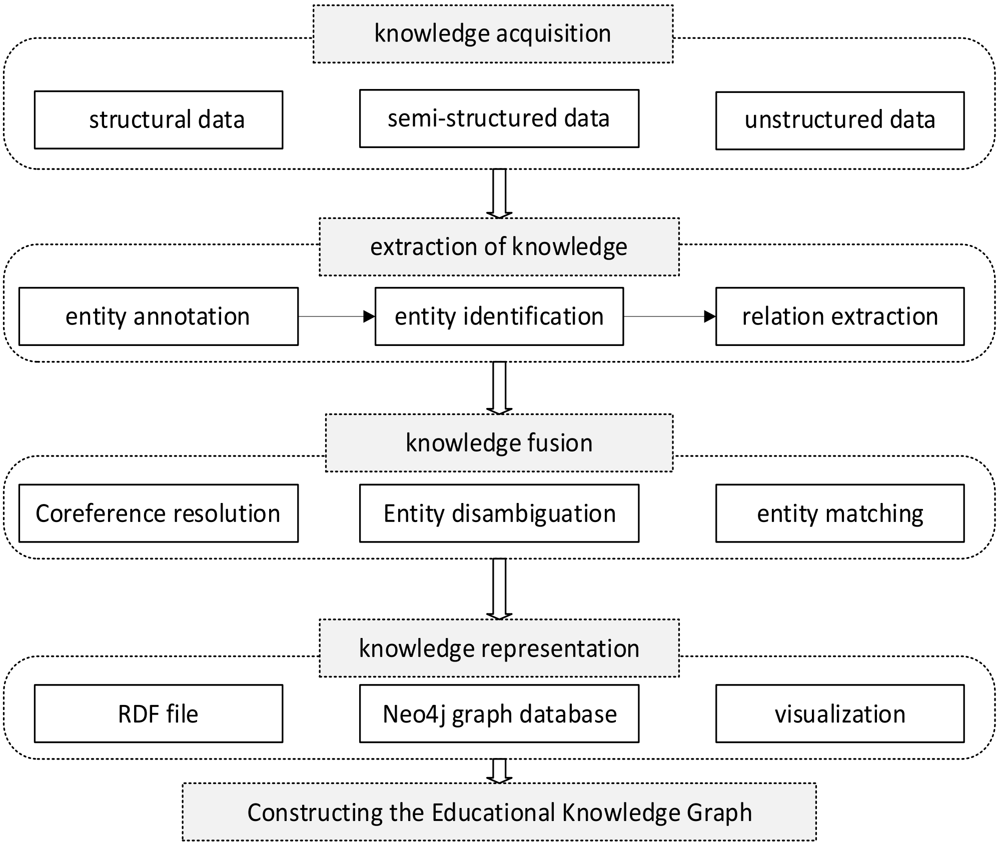


Fig. 1. Construction of educational knowledge graph based on multi-source data.

1. *Education Big Data and Processing*

Education big data is a large number in the field of education, including data generated through various educational activities. Common sources include education and teaching activities involving students and teachers, educational administration system, student-worker system and other campus management systems. Education big data has comprehensive advantages, but also has difficulties such as different data structures and diverse data forms. Education big data is mainly divided into two categories, namely structured data and unstructured data. Structured data includes traditional education data, such as academic performance, which can be transformed into corresponding entities and relationships for identification and extraction. Unstructured data include text, audio and video data contained in learning platforms such as surveillance videos and rain classes. For text type data, because it comes from different systems, its data format, storage file type and text encoding method are different.

There are three main steps to process text data. The first step is to unify the storage format, coding method and file type of data. The second step is to clean the data. In order to ensure the validity of the data, the general practice is to eliminate useless and incomplete information; The third step is to use NLP technology such as Chinese word segmentation and text labeling to process the cleaned data. The marked text is shown in the figure.

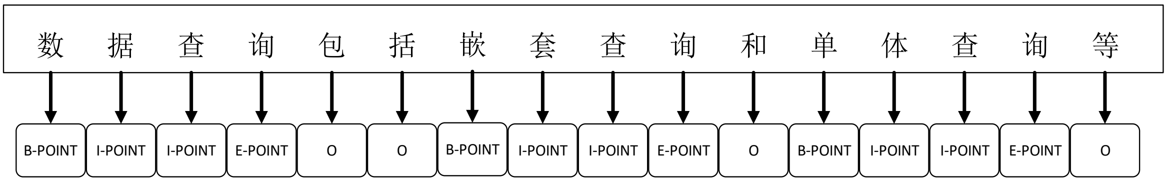


Fig. 2. Marked text.

2. *Educational Entity Identification*

The recognition of educational entities is a key step to construct the knowledge graph in the field of education. There are three main technical paths for entity recognition technology. The first is the traditional method based on rules and dictionaries, the second is the method based on machine learning, and the third is the method based on deep learning. Collins et al. found in the experiment that adding seven "seed rules" to the data can make the clustering effect better. Based on this discovery, Collins proposed two unsupervised algorithms for entity recognition. Yuan et al. proposed a parallel association rule mining approach for educational big data [3]. Huang et al. combined bidirectional long-term and short-term memory network (LSTM) with CRF, and proposed Bi-LSTM-CRF model, which significantly improved the recognition accuracy [4]. The combination of bidirectional long-term memory network and CRF has a good performance in the work of named entity recognition [5]. The overall framework of BILSTM+CRF algorithm includes the main structures such as input layer and embedded layer. The embedded layer can learn the semantic information of the text from the input text and express the text with word vectors or word vectors. In the coding layer, BILSTM deep learning framework is used to extract features of information such as word vectors and code them. The decoding layer decodes and predicts by CRF decoder; The output layer outputs the marked final result. The combination of bidirectional long-term memory network and CRF has made remarkable achievements in improving the recognition of educational entities.

The network structure model of BILSTM+CRF is show in Fig 3:

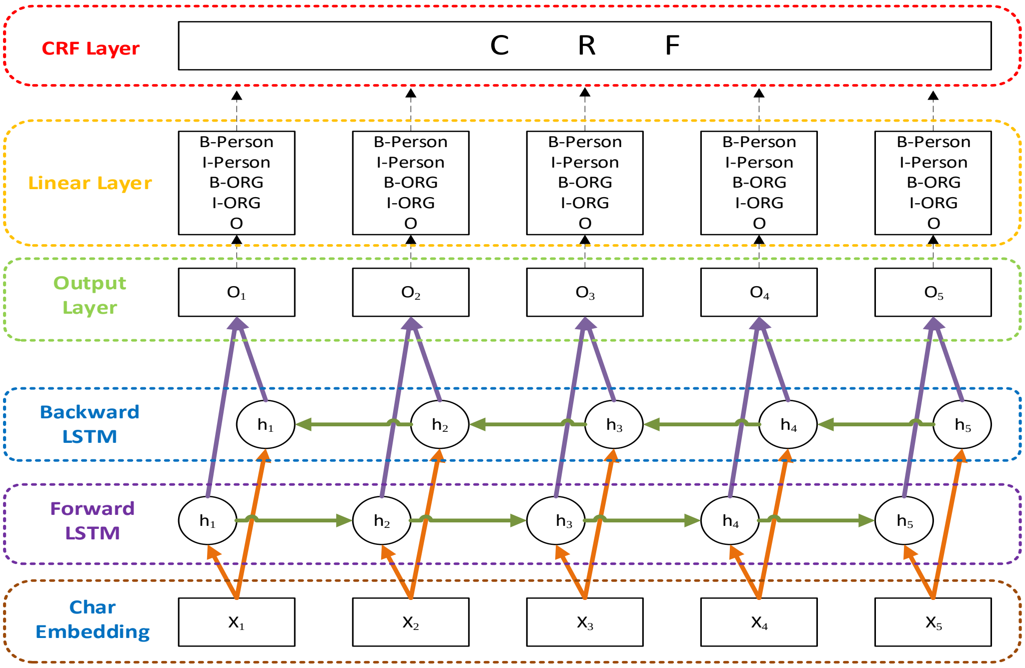


Fig. 3. Structure of BILSTM+CRF.

Through BILSTM+CRF algorithm, the entities in the field of education are identified, and the entities are extracted from the preprocessed data and distributed into training set, test set and verification set at the ratio of 6:2:2. After 10 epochs of training, the ideal effect is achieved. Then throw the unprocessed data into the trained model and output the labeling results, so as to complete entity extraction.

3. *Relationship Extraction*

In the field of education, the relationship between the two entities mainly includes the relationship between learning priority and learning dependence. By unsupervised relationship extraction, ALSAADF et al. proposed Bridge Ensemble Measure and the Global Direction Measure to infer the existence and directionality of dependencies between entities, and discover graphs by using the time sequence between entities [6]. By supervised relationship extraction, Yosef et al. studied the learning order relationship between entities extracted from textbooks. Chen et al. use data mining technology to identify and extract the cognitive premise relationship between educational entities [7]. For educational data from different sources, features are designed to distinguish the sequential relationship between entities, and then classified by machine learning classification model [8].

4. *Links to Educational Entities*

Linking educational entities is to link the entities in the data with those in the knowledge base through relationships, which can express the semantic information of educational entities more clearly, and can greatly solve the problems of multi-word synonyms and polysemy, so as to facilitate the later knowledge fusion and then apply it to specific fields. Entity connection is divided into two aspects, namely, entity disambiguation and coreference resolution. Entity disambiguation is to solve the ambiguity of entities with the same name, and coreference resolution is to solve the problem that a word refers to multiple entities. The entity disambiguation method combining BERT model and LSTM based on deep learning will achieve better results in comprehensive evaluation indicators. The algorithm based on capsule network can realize the aggregation of some features through dynamic routing, and achieve relatively good performance in Onto Notes English data set. By linking educational entities and linking educational data from different sources, the problem of incomplete coverage of a single educational data resource can be effectively solved, and multi-source data can be integrated, which can be more effectively used in knowledge quiz and optimal path planning.

5. *Storage and Display of Educational KG*

After data processing, information from different sources has been integrated into structured knowledge. The knowledge graph is represented by graph structure, and the relationship, node and attribute of graph relational database just correspond to the entity-relationship-entity structure of the knowledge graph, so the data in the educational knowledge graph is stored according to the graph data. Using the secondary graphic database, the educational knowledge graph is stored in the form of a graph because the secondary graphic database has the advantages of high reliability, high expansibility and high convenience. Part of the educational knowledge graph is shown in the following figure:

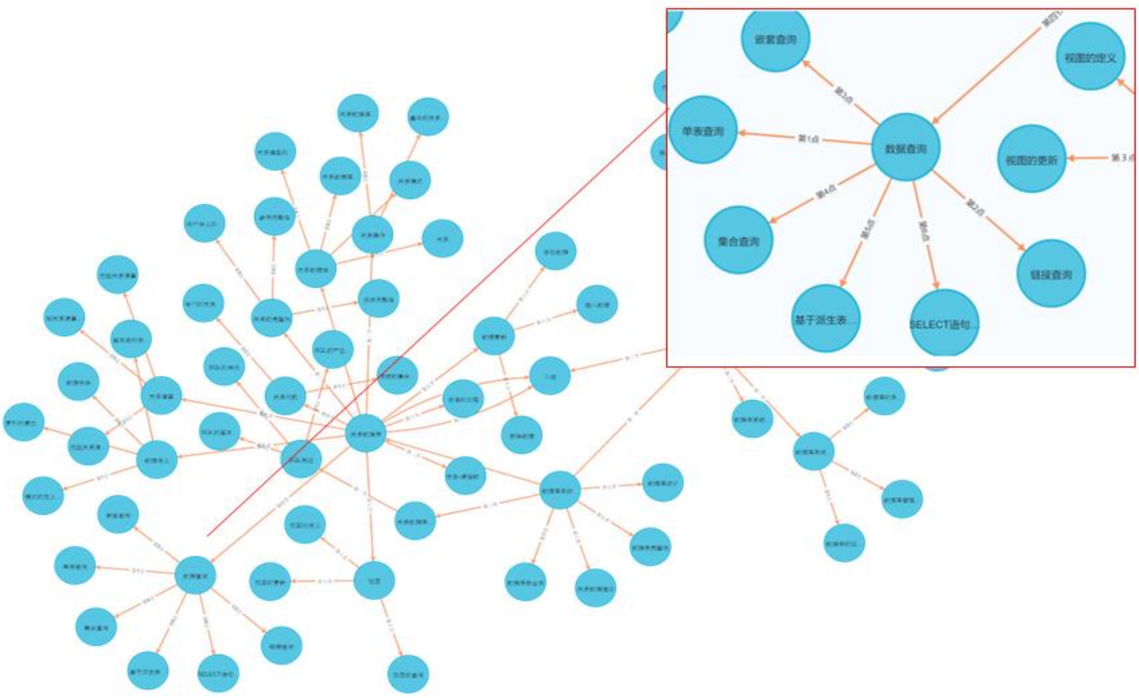


Fig. 4. Part of the educational knowledge graph display.

IV. Conclusion

The knowledge graph in the field of education can assist teachers in preparing lessons and teaching and research; Can accurately find the weaknesses of students in learning, teachers prepare lessons according to students' learning situation; You can automatically generate the question bank and use the questions for exercises, homework and exams; The educational knowledge graph can also be used to build an adaptive learning system, and it can also carry out intelligent path planning for students. According to the dominant subjects of students in the map, the content that students should learn in the next stage can be pushed by using the correlation between knowledge points, so as to create the most suitable learning path for different students and optimize the learning effect. Wang et al. attach importance to the important role of educational big data for students, teachers and other teaching participants, and build a visual educational information platform. The construction of knowledge graph can effectively help to build a visual educational information platform, thus promoting the construction of educational informatization [9]. The construction of educational knowledge graph can effectively promote the construction of educational informatization.

Acknowledgement

This work was supported by Shaanxi undergraduate and higher continuing education teaching reform research project (21BG051) and Shaanxi Province Education Science "14th Five-Year Plan" 2021 annual project (SGH21Y0279).

References

[1] Xing Wanli Wang Xianhui. Understanding students' effective use of data in the age of big data in higher education [J]. Behaviour & Information Technology, 2022, 41(12): 2560-2577.

[2] Li Jia,Jiang Yuhong. The research trend of big data in education and the Impact of teacher psychology on educational development during COVID-19: A systematic review and future perspective [J]. Frontiers in Psychology, 2021, 12.

[3] Collins M, Singer Y. Unsupervised models for named entity classification [C]// Empirical methods in natural language processing. 1999.)

[4] He Yuan,Yi Liqiong,Xu Si. Intelligent construction and mechanism of educational big data information for resource sharing [J]. Wireless Communications and Mobile Computing, 2022, 2022.

[5] Huang Wenzhi, Qian Tao, Lyu Chen, Zhang Junchi, Jin Guonian, Li Yongkui, Xu Yongrui. A multitask learning approach for named entity recognition by exploiting sentence-level semantics globally [J]. Electronics, 2022, 11(19).

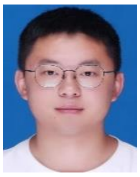
[6] Alsaad F, Boughoula A, Sundaram H, et al. Mining MOOC lecture transcripts to construct concept dependency graphs[C]// Educational Data Mining. 2019.

[7] Yosef M A, Hoffart J, Bordino I, et al. AIDA: An online tool for accurate disambiguation of named entities in text and tables [C]// DBLP. DBLP, 2011:1450-1453.

[8] Chen P, Yu L, Zheng V W, et al. An automatic knowledge graph construction system for K-12 education[C]// Proceedings of the fifth annual ACM Conference on Learning at Scale. ACM, 2018.

[9] Wang Ping, Zhao Pengfei, Li Yingji. Design of education information platform on education big data visualization [J]. Wireless communications and mobile computing, 2022, 2022.

AUTHOR'S PROFILE



Frist Author

Zhiwen Ren, was born in Hebei, China in. He received his B.S. degree from Tangshan University in 2021. He is currently pursuing the M.S. degree in Computer Technology with Xijing University. His research interests include image processing and educational knowledge graph. Facility: (School of Electronic Information, Xijing University, Xi'an, 710123, China.

email id: renzhiwen0314@163.com

Second Author

Jia Shan Liu, received the B.S. degrees from Dalian Jiaotong University in 2017. He is currently pursuing the M.S. degree in computer science and technology with Xijing University. His research interests include computer vision, image processing and deep learning. Facility: (School of Electronic Information, Xijing University, Xi'an, 710123, China.

email id: 2108544406040@stu.xijing.edu.cn

Third Author

Shanwen Zhang, was born in Shaanxi Province, China. He received his B.S. degree in mathematics from Northwest University, China, in 1988. He received M.S. degree in applied mathematics from Northwest Polytechnic University, China, in 1995. He received Ph.D. degree in electromagnetic field and microwave from Air Force Engineering University, China, in 2001. He is a professor in the Xijing University, a visiting scholar in Department of Computer Science at Virginia Tech. His research area is machine learning and its application in data mining, including machine learning, leaf image processing, data reduction, data mining, feature selection, wavelet transforms, and their application in the plant disease recognition. Facility: School of Information Engineering, Xijing University, Xi'an, 710123, China.