Journal of Artificial Intelligence Practice (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2024.070323
ISSN 2371-8412 Vol. 7 Num. 3

Development of a Knowledge Graph for Database Courses through the Integration of Multi-source Educational Data

Wen Pan

Noncommissioned Officer Academy of Pap, Hangzhou, Zhejiang, China
Mooncake9828@163.com

Keywords: Multi-source education data, Database course, Knowledge graph, Link prediction

Abstract: This paper aims to develop a comprehensive knowledge graph dataset for database curriculum based on the fusion of multi-source educational data, addressing the pressing need for efficient and systematic knowledge management in the education sector. By integrating diverse teaching data from textbooks, online courses, learning management systems, and more, advanced data fusion technologies are employed to facilitate precise extraction of knowledge points and construction of relationships. The validity of this knowledge graph is substantiated through link prediction experiments. This research not only enhances the utilization of educational data resources but also lays a robust foundation for applications such as personalized instruction, intelligent recommendations, and learning path planning—significantly contributing to the advancement of intelligent and accurate teaching methodologies in database courses.

1. Introduction

Database technology is one of the core technologies in the development of application software systems. This course is not only a required course for students majoring in computer-related fields, but also an important auxiliary in cultivating their database management and application abilities.

Knowledge graph is a graph-structured data model used to show the relationships between entities and enhance the ability to analyze problems from a "relationship" perspective. It has deepened its impact in various fields, including helping to decipher data and uncover hidden value. For example, in the public security field, knowledge graph has enhanced the practical effectiveness of operations[1]; in the field of education, the digital construction of public security subjects and criminal investigation courses[2], as well as the integration of information technology and online courses[3], have all demonstrated the role of knowledge graph in promoting modern and personalized teaching. Overall, knowledge graph in the field of education is a tool for structured presentation of teaching content, promoting the construction of subject knowledge systems, and achieving intelligent education. However, the construction of multi-source database course knowledge graph in vocational education is relatively lacking, and the corresponding knowledge graph datasets are also relatively limited.

A knowledge graph dataset based on multi-source data fusion can provide students with more diverse and comprehensive learning resources, helping them quickly build knowledge frameworks and deeply understand course content. It has significant advantages in data richness, learning efficiency, intelligent education, and knowledge sharing, which make it an important foundation for building intelligent and personalized educational platforms. It contributes to the digital transformation and high-quality development of the education industry.

This paper studies the construction of a knowledge graph dataset for database courses, aiming to comprehensively organize the teaching knowledge points covered in the course, including textbooks, online courses, and learning management systems. Building a knowledge graph for database courses is a systematic process that involves a deep understanding of the course content and its transformation into graph-based knowledge representation. After research, our findings are as follows:

● The knowledge graph of the "Database Fundamentals and Applications" course was constructed.

● The validity of the knowledge graph of the "Database Fundamentals and Applications" course was verified using a knowledge graph link prediction model.

2. Related Research

Knowledge graph is an advanced data organization form that is based on graph databases and used to represent entities (such as people, places, events, etc.) and their various complex relationships. Domain knowledge graph is the application of knowledge graph in a specific field, which abstracts and structurally represents the knowledge in that field, and its construction includes data acquisition, entity and relationship extraction, knowledge fusion, etc. Ultimately, it forms a large structured knowledge base.

With the rapid development of big data and artificial intelligence technology, more and more scholars are applying knowledge graphs to the field of education to achieve personalized and intelligent teaching. For example, Zhang Jun[4] constructed a knowledge graph for the "Introduction to Artificial Intelligence" course, reorganizing knowledge points into an orderly network; Sheng Ying and Han Tingxiang[5] created a knowledge graph platform for the "Ventilation Engineering" course, optimizing the learning route; Zhou Dongdai[6] constructed multi-level subject teaching knowledge graphs to support the digital transformation of education; Ge Jinmei[7] used knowledge graphs to analyze the hot topics of hybrid teaching in vocational education; Chen Jianhui[8] constructed a knowledge graph for "Advanced Language Programming", promoting innovation in teaching content and methods.

Building a knowledge graph typically involves using either a top-down or a bottom-up approach. In this paper, we take a bottom-up approach, starting with the extraction of entities from a large amount of knowledge data, followed by data analysis to determine the relationships between entities, and finally, gradually building a database course knowledge graph.

3. Definition of the Knowledge Graph Model Layer

The knowledge graph of the Database Foundation and Applications course constructed this time has a total of 7 entity concepts, namely "Course", "Knowledge Chapter", "Knowledge Point", "Knowledge Sub-Point", "Knowledge Content", "Relevant Knowledge", and "Relevant Skills"; and 6 relationship concepts, namely "Includes", "Contains", "Is Fine-Grained As", "Contains Content", "Covers", and "Relates To".

3.1 Meaning of Entity Concept

In the field of education, a **"Course"** is a systematically planned set of learning contents and processes aimed at facilitating the all-round development of students. A **"Knowledge Chapter"** is an independent and thematically concentrated section in teaching materials, which supports the structure of the curriculum. A **"Knowledge Point"** is the smallest unit for students to master the knowledge system of a discipline. A **"Knowledge Sub-Point"** is a refined key point within a "Knowledge point", helping to organize learning. **"Knowledge Content"** is a collection of core information, theories, and skills in educational activities. **"Relevant Knowledge"** is the background knowledge that assists in understanding or supports decision-making. **"Relevant Skills"** are the knowledge, abilities, and techniques required to complete tasks, enhancing learning efficiency and the application of knowledge.

3.2 Meaning of Relation Concept

In a knowledge graph, "relationship" refers to the connection or link between entities, which is one of the basic elements that constitute a knowledge graph. This knowledge graph contains six types of relationships: "Includes" indicates the relationship between a part and a whole; "Contains" indicates a strong relationship between an entity and a part or a characteristic; "Is Fine Grained As" is the process of breaking down a whole into smaller parts; "Contains Content" refers to specific information within a course or other resources; "Covers" involves a broad range and emphasizes comprehensiveness; "Relates to" indicates aspects or factors that are related to something.

In this thesis, the relationships between various entity sets are hierarchical, with the upper-level entity types being compatible with the lower-level entity types. This structure forms a tree-like structure for the course of Database Fundamentals and Applications.

4. Extraction of the Instance Layer from the Knowledge Graph of Database Foundations and Applications Course

This thesis first performs entity extraction and relationship extraction operations on textbooks, online courses, and learning management systems (LMS), and then merges them into a global database course knowledge graph based on the hierarchical relationship between the entities from top to bottom. The core lies in extracting structured information (triple knowledge points) from unstructured/semi-structured data (text, video, images). The following will give extraction methods and examples for each of these resources:

The first one, textbook content extraction, involves identifying noun entities first, then determining the relationships between entities, and finally connecting them into triples, such as constructing triples (database concept design, includes, data) based on the content of Chapter 1 of the textbook. The second one, online course content extraction, requires OCR technology to recognize and extract text from pictures, then determines the relationships between entities to form triples, such as identifying the noun entities "data model" and "conceptual model" from a screenshot of an online course video, and then combining them with the overall knowledge content of databases to form triples (database concept design, further divided into, data model) and (conceptual design phase, covers, conceptual model). The third one is the example content extraction from a learning management system (LMS), which can also use OCR technology. For example, based on a screenshot of the LMS, we can extract noun entities such as "T-SQL statements" and "E-R diagrams", and then combine them with the overall knowledge content of databases to form triples (T-SQL statements, further divided into, data query statements) and (the steps of conceptual design, include content, design local E-R diagrams).

In practical applications, the aforementioned three extraction processes sometimes necessitate the incorporation of natural language processing (NLP) techniques, such as named entity recognition (NER), relation extraction (RE), etc., as well as customized processing logics for specific resources. For non-textual resources like videos, techniques such as image recognition and speech recognition are occasionally required to aid in the extraction.

According to the above process, we first removed duplicate items from the extracted dataset, and then built three small-scale knowledge graphs based on the data from textbooks, online courses, and learning management systems. We then linked these three sub-graphs through the transitive relationships between the entity sets at each level, thus obtaining the overall database course knowledge graph. The extraction algorithm is shown in the following Table 1:

Table 1: Extraction algorithm of instance layer of knowledge graph of database foundation and application course.

Input	Textbook on Database Fundamentals and Applications(Text_DatabaseBook), Textual Material for Online Database Course (Text_DatabaseOLC), Textual Material for Database Learning Management System (Text_DatabaseLMS)
Output	Database Course Knowledge Graph (DBCKG)
1	Text_DatabaseBook_Filter = TextProcessing(Text_DatabaseBook|Text_DatabaseOLC|Text_ DatabaseLMS) //Process the original text materials of database textbooks, online courses, and learning management systems respectively, and retain the core content.
2	def getPaddleEntities(model, text) // Define the function getPaddleEntities, which represents using PaddleNLP to extract entities and their labels.
3	ner_model=load_paddlenlp_ner_model()//Load the PaddleNLP model for named entity recognition.
4	Entities_Book=getPaddleEntities[(ner_model, Text_DatabaseBook_Filter)|OCR(ner_model, Text_DatabaseOLC_Filter)|OCR(ner_model, Text_DatabaseLMS_Filter)]//According to the PaddleNLP tool, relevant entities and their labels are extracted from the preprocessed corpus in sequence, including those from textbooks, online courses, and learning management systems. OCR() denotes a function for recognizing text in images.
5	def getPaddleTriple(model, entity, text)//Define the function getPaddleTriple, which represents extracting relevant triples (head entity - relation - tail entity) for each entity.
6	rel_model = load_paddlenlp_rel_model()//Loading PaddleNLP models for relation extraction
7	Triples = []
8	for entity in Entities:

9	tripleBook=getPaddleTriple(rel_model, entity, Text_DatabaseBook _Filter) //Using the PaddleNLP tools, we obtain the triples (relations) for the entities in the current curriculum.
10	tripleOLC = getPaddleTriple(rel_model, entity, Text_DatabaseOLC _Filter) //Using the PaddleNLP tool, we obtain triples, i.e., relationships, for the current online course entities.
11	tripleLMS = getPaddleTriple(rel_model, entity, Text_DatabaseLMS _Filter) //Using the PaddleNLP tool, we obtain triples, i.e., relationships, for the entities in the current learning management system.
12	Triples.append(triplesBook,tripleOLC,tripleLMS) //Add the current entity's triple to the triple list.
13	DBCKG = connect(neo4j(tripleBook) , neo4j(tripleOLC),neo4j(tripleLMS)) //Construct a comprehensive database course knowledge graph based on the triad of textbooks, online courses, and learning management systems.

Through the above code, we create an example of a knowledge graph for the database course, as shown in Figure 1:

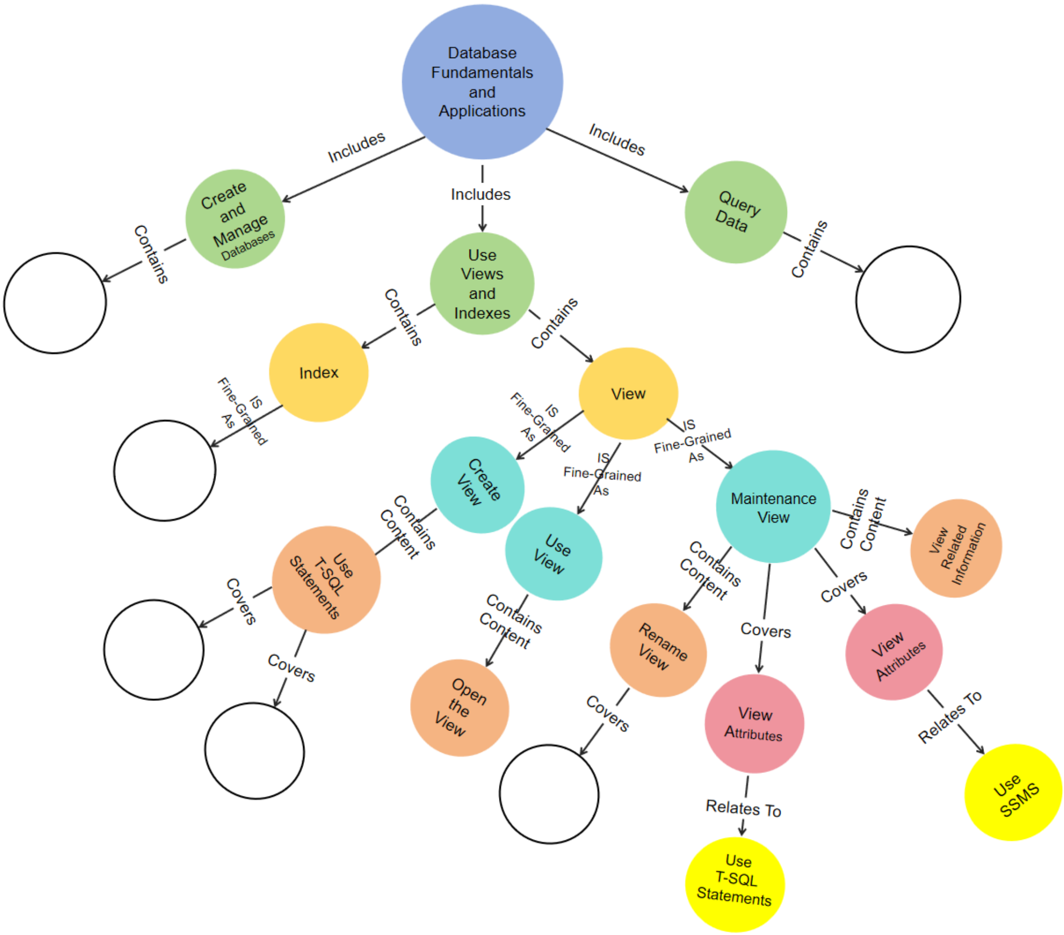

Figure 1: Knowledge graph diagram of database foundation and application course.

We found that the database course knowledge graph presents a tree-like knowledge graph structure, as shown in the above figure, which corresponds to the characteristic of gradually deepening subject content. It reveals the intrinsic relationships between knowledge points and, combined with big data analysis, can provide evaluation and optimization suggestions to educational administrators, promoting the rational allocation of educational resources and improvement of quality.

5. Dataset Analysis

5.1 Scale Analysis

After instance layer extraction, the DBCKG dataset for database course knowledge graph constructed in this paper contains a total of 1,697 entity nodes, 6 types of relationships, and forms 1,696 triples. The distribution of the number of entities is shown in the following Table 2:

From the number of entity nodes in Table 2, the entity nodes of knowledge content have the most, followed by related knowledge entity nodes, and the entity nodes of "Knowledge Chapters" (except for "Courses") have the least; from the number of relationship types, the number is stable at only 6 types; from the number of triples, it is almost the same as the number of entity nodes.

According to the calculation of the possible relationships between any two nodes in the database course knowledge graph, the knowledge graph theoretically contains nearly 1.03×1013 (1×10×43×327×580×462×274) relationships in total. The above analysis suggests that the current database course knowledge graph is relatively sparse and there are still many relationships to be mined, which is also one of our research directions in the future.

Table 2: DBCKG dataset size.

	Name	**Number**	**Total**
Entities	Course	1	1697
	Knowledge Chapter	10	
	Knowledge Point	43	
	Knowledge Sub-Point	327	
	Knowledge Content	580	
	Relavant Knowledge	462	
	Relavant Skills	274	
Relation	Species	6	6
Triples	Includes	10	1696
	Contains	43	
	Is Fine-Grained As	327	
	Contains Content	580	
	Covers	462	
	Relates To	274	

5.2 Analysis of Quality Index

Among the data shown in Table 3, the minimum query time refers to the fastest time of the query node, the maximum query time refers to the slowest time of the query node, the average query time refers to the average value of the query time of all nodes in the computing graph, which is used to reflect the distribution of the graph's response efficiency. The longest inference time refers to the inference time required to find the longest path in the graph, which is used to reflect the topological complexity of the graph schema.

Table 3: Evaluation index results.

Evaluation Indicators	Index value
Minimum query time	0.0
Maximum query time	0.001
Average query time	6.10 x 10^{-6}
Maximum reasoning time	0.001

Analyzing the above evaluation results, we found that the minimum query time is 10-6, which may be due to the presence of a large number of dispersed end nodes in the knowledge graph. However, the other three indicators are at a good level, indicating that if this knowledge graph is applied to query and reasoning scenarios, its response performance is excellent.

6. Experimental Results and Dataset Analysis

We conducted relationship prediction experiments on the DBCKG dataset we built, using three baseline models: the first is a translation-based model called TransE, the second is a semantic matching model called DisMult, and the third is a neural network-based model called ConvE. The specific information of the baseline models is as follows:

(1) TransE[9]. The main idea is to convert triples in the graph into vectors, adjusting the vector positions so that the head entity + relation ≈ the tail entity. The triples are evaluated for rationality using a scoring function, and correct facts conform to this pattern. TransE optimizes the vector representation by expanding the error sample score gap and improving the accuracy of graph triplets. (2) DistMult[10]. This model uses a diagonal matrix to represent relationships to solve complex situations, and uses vectors to represent entities. It uses a scoring function to show that DisMult has a deep interaction and expression ability between entities and relationships through matrix decomposition. (3) ConvE[11]. This model introduces a CNN into the graph completion task, concatenates the head entity vector with the relation vector into a two-dimensional matrix, and further extracts features through a CNN. Then, it multiplies the tail entity vector to obtain the fact score.

We selected the relatively smaller FB15k-237 and WN18R datasets as our benchmark datasets for comparison. The following evaluation metrics will be used to assess the performance of the algorithms, as shown in Table 4:

(1) MR (mean rank): This refers to the average ranking of positive samples in the candidate edge score sequence. The lower the mean rank value is, the better the graph dataset is built. The formula is as follows (M represents the number of nodes in the knowledge graph, and N represents the ranking of candidate predictions according to their scores, with the first correct prediction ranked first):

$$MR = \frac{1}{M}(\sum_{i=1}^{M} rank_i)$$

(2) Hits@k: used to evaluate the proportion of correct predictions of links being ranked in the top k in the dataset. The formula is as follows (M represents the number of nodes in the knowledge graph, and right represents the number of nodes ranked in the top k in the predicted result):

$$Hit@k = \frac{right}{M \cdot k\%}$$

Table 4: Relationship predicted experimental results.

	DBCKG(ours)			FB15k-237			WN18RR		
	MR	H@1	H@10	MR	H@1	H@10	MR	H@1	H@10
TransE	1175	8.1	20.59	209	21.72	49.65	3936	2.79	49.52
DistMult	**541**	10.9	8.82	199	22.44	49.01	5913	39.68	50.22
ConvE	1543	**21.61**	**36.82**	281	21.90	47.62	4944	38.99	50.75

* Note: Values for data sets FB15k-237 and WN18RR on TransE, DistMult, and ConvE models are taken from the review Knowledge Graph Embedding for Link Prediction: A Comparative Analysis.

Through the results of the relationship prediction experiment in Table 4, we found that: 1) The DBCKG dataset performed poorly on TransE and ConvE, but well on DistMult, showing that it has rich multi-to-multi relationships and strong extensibility, making it suitable for semantic matching and neural network knowledge graph completion training; 2) DBCKG's Hits@1 and Hits@10 scores on ConvE were higher than those on TransE and DistMult, indicating that it contains diverse entities and complex relationships, and using models such as ConvE that utilize graph convolutions for prediction can yield better prediction results; 3) The DBCKG dataset has complex relationships and obvious graph features, making it suitable for research experiments on semantic matching or neural network relationship prediction models.

7. Summary and Future Directions

At present, there are few examples of database course knowledge graph datasets based on multi-source educational data fusion. In this paper, we propose and construct a knowledge graph dataset DBCKG that integrates teaching knowledge points from textbooks, online courses, and learning management systems for database courses. We first define seven entity concepts (courses, chapters, points, subpoints, content, related knowledge, and skills) and six relationships. Then, we extract these entities and relationships from integrated course content. Finally, we construct a database course knowledge graph based on the hierarchical relationships between these entities.

The database course dataset we have built includes 1697 entity nodes, 6 relationships, and 1696 triples, and presents a topological structure of a tree-shaped knowledge graph. We conduct relationship prediction experiments on the dataset using three baseline models, and the results show that the DBCKG dataset has a high degree of relationship complexity. This feature indicates that the DBCKG dataset is more suitable for research experiments based on semantic matching or neural network-based relationship prediction models. In addition, based on the evaluation results of the knowledge graph dataset we have built, we find that the end nodes of the knowledge graph are scattered, and the overall structure is sparse. This suggests that we can use knowledge graph reasoning techniques in the next stage to fill in the graph.

References

[1] Fan Shu. Research on the Construction and Application of Disciplinary Knowledge Graph in Public Security Science [J]. Journal of Guizhou Police College, 2024(2):68-77.

[2] Zhang Wei. Research on the Construction of Criminal Investigation Course Based on Knowledge Graph [J]. Journal

of Shanxi Police College, 2024, 3(2):116-123.

[3] Wan Haipeng, Cheng Lingna, Cheng Yuemei. Research on Online Course Design Based on Subject Knowledge Graph in Information Technology [J]. China Education Informatization, 2023, 29(08): 121-128.

[4] Xia Li. Zhu Wenxuan. Research on the Reform of English Writing Teaching Evaluation under the Background of Educational Informationization [J]. Education and Teaching Forum, 2024(5):101-104.

[5] Song Ling-qing, Xu Lin. The Logical Starting Point and Boundary of Artificial Intelligence Education Application - Taking Knowledge Learning as an Example [J]. Intelligent Leadership and Wise Education, 2019(389):14-20.

[6] Xue Jian-li, Tuo Wan-liang, Qiu Xin. Intelligent Application of Non-heritage Knowledge Graph - Taking Ethnic Cultural Education as an Angle [J]. Social Scientist, 2023(11):65-70.

[7] Zhang Jun, Yuan Zhan-jiang, Yang Zhong-ming, Li Zi-jian, Liu Hao-ran, Deng Yue, Lin Ze-kai. Research on the Construction and Application of Knowledge Graph of "Introduction to Artificial Intelligence Application"[J]. Network Security Technology and Application, 2023(6):96-99.

[8] Sheng Ying, Han Ting-xiang. Construction and Application of Knowledge Graph Teaching Platform for "Ventilation Engineering" Course [J]. Intelligent Computer and Application, 2023, 13(8):135-139.

[9] Zhou Dongdai, Dong Xiaoxiao, Gu Hengnian. New Trends in Knowledge Graph Research in the Field of Education: Subject Teaching Knowledge Graph. Journal of Educational Television and Multimedia Research, 2024, 45 (2): 91-97+120.

[10] Kadlec R, Bajgar O, Kleindienst J. Knowledge Base Completion: Baselines Strike Back [J]. ACL 2017: 69.

[11] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2d knowledge graph embeddings[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018. 1811-1818.