Research on Knowledge Graphs for Academic Book Topic Selection (Postprint)

Authors: Wu Nada, Li Caishan, Feels great!

Date: 2023-10-08T00:00:00+00:00

Abstract

Knowledge graphs are structured semantic knowledge bases that describe knowledge resources in a graphical manner. Through visualized knowledge graphs, the relationships among relevant information for academic book topic selection and the overall knowledge context can be clearly displayed, providing support for optimizing topic-related content and effectively analyzing and mining promising topics. This paper investigates knowledge graphs for academic book topic selection, proposes representation and construction methods for such knowledge graphs, and provides auxiliary decision-making support for academic book topic planning.

Full Text

Preamble

Research on Knowledge Graphs for Academic Book Topic Selection

Authors: Wu Nada, Li Caishan*, Ha Shuang
Affiliation: Beijing Xintong Media Co., Ltd.

Abstract: A knowledge graph is a structured semantic knowledge base that describes knowledge resources in a graphical manner. Through visualized knowledge graphs, the correlations among relevant information for academic book topic selection and the overall knowledge structure can be clearly displayed, providing support for optimizing topic-related content and effectively analyzing and mining valuable topics. This paper investigates knowledge graphs for academic book topic selection, proposing representation and construction methods to provide decision support for academic book topic planning.

Keywords: Academic book topic selection; Knowledge graph; Decision support
CLC Number: G350
Document Code: A

Article ID: 1671-0134(2019)12-078-03
DOI: 10.19483/j.cnki.11-4653/n.2019.12.022

In the publishing domain, a topic refers to a publishing project selected after comprehensive analysis and consideration of a subject [**?**]. Traditional topic planning relies heavily on editors' experience, with data processing conducted in isolated ways—typically analyzing each content item separately. Data storage methods are simplistic, retention periods are short, and fine-grained information management with multi-dimensional, multi-angle extensions cannot be achieved, resulting in insufficient mining of collected data value. In today's era of rapid mobile internet development and widespread adoption, data must be reconstructed and applied through appropriate means to rapidly obtain valuable information amid explosive content growth.

This paper investigates the representation and construction of knowledge graphs for academic book topic selection, aiming to provide effective decision support for editors planning academic book topics through constructed knowledge graphs.

1. Current Status Analysis of Academic Book Topic Selection

Generally, topics originate from certain intentions or aspirations that editors develop during information collection. These gradually form concrete proposals through thorough analysis of subjective and objective conditions and multi-party validation. Information collection represents a crucial step in topic planning, serving as the foundation for topic discovery, planning, verification, and demonstration.

1.1 Information Types Collected for Academic Book Topic Planning

Information collected for academic book topic planning generally includes five categories: social information, disciplinary information, publishing information, market (reader) information, and author information [**?**].

(1) **Social information** includes central and local government regulations, policies, white papers, and public information related to the discipline.

(2) **Disciplinary information** encompasses the scope and main content of the discipline, domestic and international development status, frontier and key research directions, major research projects, research topics, leading scholars, academic achievements, and awards.

(3) **Publishing information** comprises two aspects: first, information about similar topic books from the same publisher, including variety, quantity, authors, and sales performance; second, information about comparable

books from other publishers, including variety, quantity, authors, and sales performance.

(4) **Author information** includes the author' s academic background, position/title, research direction, primary work, applied research projects, and publication history.

(5) **Reader information** reflects personalized characteristics, including basic reader demographics (age, profession, etc.), purchasing power, actual needs, potential demands, book market distribution, and market feedback.

1.2 Problems in Information Collection

Currently, information obtained during academic book topic planning is mostly stored in documents, spreadsheets, and a few databases, presenting several significant problems.

1.2.1 Information Heterogeneity During topic information collection, data sources are diverse with inconsistent structures, creating substantial difficulties for data integration and storage.

1.2.2 Information Redundancy Combining information from different sources is challenging, with poor complementary advantages and low information integrity. Data contains substantial redundancy and noise, and information accuracy requires improvement.

1.2.3 Information Isolation **Spatial discontinuity:** The information content of interest cannot reflect connections between events in the data.
Temporal discontinuity: The information content of interest has short temporal continuity.

1.2.4 Information Stasis Existing information cannot be effectively utilized for discovery and prediction.

1.2.5 Information Visualization Difficulties Data cannot be represented in multiple forms to make it more intuitive and comprehensible.

These problems result in difficulties storing, retrieving, reusing, and sharing book topic information. In traditional information collection processes, editors manually gather information for simple storage (primarily in documents, spreadsheets, and limited databases), then subjectively analyze information about readers, authors, content, and marketing to develop topic proposals. Traditional academic book topic selection methods involve significant subjective factors, imprecise information analysis, low reliability and scientific rigor, and challenges in information retrieval, reuse, and sharing. Therefore, this paper investigates knowledge representation and construction methods for academic book topic selection knowledge graphs, aiming to reconstruct information data

for clearer, more three-dimensional understanding and to simulate the brain' s comprehensive information analysis capability through reasoning algorithms, thereby supporting effective decision-making for book topic selection.

2. Definition of Academic Book Topic Selection Knowledge Graph

An academic book topic selection knowledge graph aims to construct a large-scale semantic network describing various entities and their relationships in the academic book topic planning process. Graph nodes represent entities, while edges represent relationships. It can also be viewed as a large-scale knowledge base providing effective storage, retrieval, and visualization of complex data involved in academic book topic selection, offering reliable and clear information and context for planning.

Currently, data sources for academic book topic selection knowledge graphs primarily derive from the five categories of information described in Section 1: social information, disciplinary information, publishing information, market (reader) information, and author information, with potential for expansion based on practical requirements. These five categories involve three main data types.

2.1 Structured Data

This primarily refers to tables in relational databases, Excel spreadsheets, and other structured data. In knowledge graph construction, these mainly originate from publisher system databases at various levels and databases provided by partners.

2.2 Unstructured Data

In academic book topic selection knowledge graph construction, this primarily refers to plain text materials, such as master' s and doctoral theses, newspapers, and image and audio data from conferences.

2.3 Semi-Structured Data

This refers to data between structured and unstructured data, typically including XML and HTML web pages. In academic book topic selection construction, these mainly derive from information obtained from various websites, such as book information (title, editor' s recommendation, author biography, etc.) and book market information (review ratings, review times, reviewer locations, etc.) extracted from XML of e-commerce sites (bookstores), and public information extracted from XML of the Ministry of Industry and Information Technology website.

3. Representation of Academic Book Topic Selection Knowledge Graph

The structure of an academic book topic selection knowledge graph consists of a node set and an edge set, formally represented by Equation (1):

$$ATS\_KG = \{< ATS\_N >, < ATS\_R >\} \tag{1}$$

where $< ATS\_N >$ represents the node set for academic book topic selection, with nodes being various entities in topic information such as authors and books; $< ATS\_R >$ represents the edge set, which can be expressed as Equation (2):

$$< ATS\_R >= \{< ATS\_T >, < ATS\_D >, < ATS\_G >\} \tag{2}$$

where $< ATS\_T >$ represents the relationship type set, such as "author-book relationship" or "book-publisher relationship" ; $< ATS\_D >$ represents the relationship direction set, such as "author->book" or "book<-publisher" ; $< ATS\_G >$ represents the triplet set expressing semantic relationships, with each triplet representing a fact, as shown in Equation (3):

$$< ATS\_G >= \{(ATS\_N1, ATS\_T1, ATS\_N2)\} \tag{3}$$

In Equation (3), ATS_N1 and ATS_N2 represent different nodes (entities), and ATS_T1 represents the semantic relationship between ATS_N1 and ATS_N2, with direction from ATS_N1 to ATS_N2. For example, the fact "Author Li Jie, Work *Industrial Big Data*" can be represented by the triplet (Li Jie, author-book relationship, *Industrial Big Data*).

4. Construction of Academic Book Topic Selection Knowledge Graph

Academic book topic selection knowledge graph construction involves two main steps: storage schema design for academic book topic databases and knowledge graph construction using graph databases, as shown in [Figure 1: see original paper].

4.1 Database Storage Schema Design

This involves organizing academic book topic-related information, formally modeling entities and their relationships, and providing clear definitions. Based on the three data types discussed in Section 1 and the representation method described in Section 3, we design the storage schema for academic book topic selection knowledge graph databases. The three data types are addressed as follows.

(1) **Structured data** primarily originates from relational databases (e.g., MySQL, SQL Server) and Excel files. The main approach involves analyzing table and field information to extract relationship patterns, design transformation rules, and establish table structures for the academic book topic selection knowledge graph database.

For example, a relational MySQL database contains a table named "Author Information," shown in .

** Author Information**

ID	Name	Age	Institution	Position	Research Direction	Published Books
1	Liu	34	University 1	Professor	Autonomous Driving, Reinforcement Learning	*A*
2	Zhang	56	Institute 1	Researcher	Computer Vision	*B*
3	Guo	48	University 2	Professor	Wireless Communication	*C*

Based on the knowledge graph representation, we can abstract node sets and relationship sets:

Name = {Liu, Zhang, Guo, ···}
Age = {34, 56, 48, ···}
Institution = {University 1, Institute 1, University 2, ···}
Position = {Professor, Researcher, ···}
Research Direction = {Autonomous Driving, Reinforcement Learning, Computer Vision, Image Processing, Wireless Communication, ···}
Published Books = {*A*, *B*, *C*, ···}

Author-Institution Relationship = {<author-institution relationship, author->institution, (Liu, author-institution relationship, University 1)>, <author-institution relationship, author->institution, (Zhang, author-institution relationship, Institute 1)>, <author-institution relationship, author->institution, (Guo, author-institution relationship, University 2)>, ···}

Author-Research Direction Relationship = {<author-research direction relationship, author->research direction, (Liu, author-research direction relationship, Autonomous Driving)>, <author-research direction relationship, author->research direction, (Liu, author-research direction relationship, Reinforcement Learning)>, <author-research direction relationship, author->research direction, (Zhang, author-research direction relationship, Computer Vision)>, <author-research direction relationship, author->research direction, (Guo, author-research direction relationship, Wireless Communication)>, ···}

Due to space limitations, not all node and edge sets are enumerated here. Through node sets and relationship sets, we design the storage schema for academic book topic selection knowledge graphs. Node sets map to node table groups in the graph database storage schema, while edge sets map to relationship table groups. The graph database storage schema is represented as actual stored node table groups and relationship table groups.

Node table groups effectively preserve hierarchical structures while enabling efficient entity data expansion. Relationship table groups, represented by triplets, store facts and effectively preserve hierarchical relationships of semantic relations while enabling efficient storage and expansion of other semantic relationships between entities.

Using the author-research direction relationship as an example, the relevant node sets and edge sets map to corresponding table groups. The author node table is shown in , the research direction node table in , and the author-research relationship table in .

** Author Node Table**
** Research Direction Node Table**
** Author-Research Relationship Table**

(2) **Unstructured data** is processed according to semantic relationship hierarchies and structures. For example, consider the text: "The 2016 policy hotspot keywords include big data, IoT, cloud computing, AI, 5G technology, etc. Big data disciplinary directions can be divided into big data collection and preprocessing, big data storage and management, big data computing models, big data analysis and mining, big data visualization, big data security, big data systems and memory, big data processing platforms, big data networks, etc. In the industrial big data direction book market, there is *Industrial Big Data* by author Li Jie, published by Machinery Industry Press."

Based on the knowledge graph representation, we can abstract node sets and relationship sets:

Policy Hotspot Node Set = {Big Data, IoT, Cloud Computing, AI, 5G Technology, ···}
Big Data Disciplinary Direction Node Set = {Big Data Collection and Preprocessing, Big Data Storage and Management, Big Data Computing Models, Big Data Analysis and Mining, Big Data Visualization, Big Data Security, Big Data Systems and Memory, Big Data Processing Platforms, Big Data Networks, Data Algorithm Fundamentals, Industrial Big Data, ···}

Hotspot-Discipline Edge Set = {<hotspot-discipline relationship, hotspot->discipline, (Big Data, hotspot-discipline relationship, Industrial Big Data)>, <hotspot-discipline relationship, hotspot->discipline, (Big Data, hotspot-discipline relationship, Big Data Storage and Management)>, ···}

Various relationships are defined through logic and practical application, though

not enumerated here. The method for designing graph database storage schemas from node sets and relationship sets follows the same approach as for structured data.

(3) **Semi-structured data** lies between structured and unstructured data. Hierarchical relationships are typically obtained from XML files. Based on actual conditions, combining processing methods for structured and unstructured data yields the academic book topic selection knowledge graph database schema.

4.2 Building Knowledge Graphs Using Graph Databases

Based on the designed graph database storage schema, we construct the academic book topic selection knowledge graph using a graph database. Recent years have witnessed rapid development of graph-based storage across the database field, with Neo4j holding the leading position. Neo4j is a high-performance graph database usable on Java and Python platforms. This paper uses the Java API to write academic book topic selection data from node table groups and relationship table groups into the Neo4j graph database. For building knowledge graphs with Neo4j, refer to *The Neo4j Graph Platform*; the primary method involves using Cypher statements to write designed node table groups and relationship table groups into the Neo4j graph database to obtain the knowledge graph.

Using the People's Posts and Telecommunications Press "National Heavyweight" book series as an example, [Figure 2: see original paper] shows a partial knowledge graph visualization. The figure partially illustrates relevant information about this book series.

Conclusion

This paper analyzes problems in current academic book topic planning, investigating representation and construction methods for academic book topic selection knowledge graphs to address these issues and provide effective decision support. Currently, due to data volume limitations, the knowledge graph scale remains small. Future work will focus on enriching the academic book topic selection knowledge graph and exploring knowledge reasoning methods to recommend and support academic book topic planning.

References

[1] National Publishing Professional Qualification Examination Office. *Fundamentals of Publishing* [M]. Shanghai: Shanghai Dictionary Publishing House, 2004.
[2] Zhang Boxi. Information Collection in Early Book Topic Planning [J]. *China Media Technology*, 2013(4): 226-227.

Note: Figure translations are in progress. See original paper for figures.