教育大数据驱动的高校学科知识图谱构建与应用研究

郭　娜

(防灾科技学院计算机科学与工程学院,河北廊坊　065201)

[摘　要]随着信息技术的飞速发展,教育大数据已成为推动高等教育变革的核心驱动力。本文立足于教育大数据时代背景,以知识体系复杂、更新迭代迅速的计算机学科为研究对象,深入探讨了学科知识图谱的构建方法与应用范式。研究首先分析了多源异构教育数据的采集与处理技术,以计算机学科为例,提出了基于深度学习与本体论的计算机学科知识图谱构建方案;其次,构建了涵盖课程规划、资源推荐、科研趋势分析及个性化学习路径规划的应用体系;最后,通过量化与质性相结合的评估方法,验证了该图谱在提升教学质量、优化科研管理及促进学生深度学习方面的显著效果。本研究旨在研究高校教育数字化转型的理论支撑与技术路径,促进计算机学科教育向智能化、精准化方向发展。

[关键词]教育大数据;学科知识图谱;计算机学科;个性化学习;知识表示

[作者简介]郭娜(1980—),女,吉林白山人,防灾科技学院计算机科学与工程学院副教授,工学硕士,研究方向:教育大数据、教育数字化。

[基金项目]本文系 2024 年廊坊市科学技术研究与发展计划自筹经费项目“教育大数据驱动的高校学科知识图谱构建与应用研究”(项目编号:2024011027)。

[DOI]https://doi.org/10.62662/kxwxz0304036　　[中图分类号]G434

[本刊网址]www.oacj.net　　[投稿邮箱]jkw1966@163.com

引言

(一)研究背景

当前,全球正处于从“信息化”向“数智化”转型的关键时期,大数据、人工智能等新兴技术正深刻重塑着社会的各个领域,教育行业亦不例外。根据教育部发布的《教育信息化中长期发展规划(2021—2035 年)》,利用数据驱动教育创新已成为高校教学改革的重中之重。在这一背景下,教育大数据(Educational Big Data)应运而生,它涵盖了从学生学习行为、教学管理过程到科研产出的全方位数据流。

计算机学科作为现代科技发展的基石,其知识体系呈现出显著的复杂性、动态性和交叉性。一方面,计算机技术更新迭代速度极快,新的编程语言、框架和算法层出不穷;另一方面,其知识结构并非线性排列,而是呈现出网状的关联特征。传统的以教材为中心、线性授课的教学模式,难以适应这种快速变化的知识体系,也难以满足学生个性化、差异化的学习需求。因此,如何利用教育大数据技术,将离散的、非结构化的学科知识转化为结构化的、可计算的知识网络,成为亟待解决的科学问题。

学科知识图谱(Discipline Knowledge Graph)作为一种以图形化方式揭示知识内在逻辑与关联的技术工具,能够有效解决上述痛点。它不仅能够将计算机学科中的概念、算法、模型、应用等实体及其相互关系进行显式表达,还能通过语义推理挖掘潜在的知识关联。因此,构建教育大数据驱动的高校计算机学科知识图谱,不仅是技术发展的必然趋势,更是提升高校计算机人才培养质量的迫切需求。

(二)研究意义

本研究通过探索教育大数据环境下学科知识图谱的构建机理,丰富了知识工程在高等教育领域的理论框架。它突破了传统知识组织的静态局限,引入了基于多源异构数据的知识表示与演化理论,为构建动态、自适应的学科知识体系提供了新的理论视角。

在实践层面,本研究具有多重价值。对于教学而言,知识图谱能够辅助教师进行课程体系的优化设计,实现从“以教为中心”向“以学为中心”的转变;对于学生而言,基于图谱的个性化学习路径推

荐能够有效缓解“信息过载”问题,提升学习效率;对于科研管理而言,图谱能够直观展示学科发展脉络与热点前沿,为科研选题和跨学科合作提供决策支持。

(三)研究问题与目标

本研究旨在解决以下核心问题:

数据融合难题:如何从海量的、非结构化的教育大数据(如 MOOC 视频、学术文献、教学大纲)中高效抽取计算机学科知识实体与关系?

动态构建机制:如何设计一套既能准确表示静态学科知识,又能动态更新以适应技术快速迭代的图谱构建方案?

应用场景落地:如何将构建好的知识图谱有效应用于教学、科研及学习评估中,并验证其有效性?

基于此,本研究的目标是提出一套完整的计算机学科知识图谱构建与应用体系,开发相应的原型系统,并建立科学的效果评估指标。

一、计算机学科知识图谱构建

(一)数据来源与采集

构建高质量的知识图谱,首要任务是获取全面、准确的数据。本研究设计了多维度的数据采集策略:

1. 数据源选取

(1)学术文献库:选取 CNKI、Web of Science、IEEE Xplore 等数据库,获取计算机学科的前沿研究成果,提取核心概念与技术演进脉络。

(2)在线教育平台:爬取 MOOC(慕课)、Coursera、edX 等平台上的计算机课程数据,包括教学大纲、课件、习题及视频字幕,这些数据直接反映了教学知识体系。

(3)开源社区与技术文档:GitHub、Stack Overflow、官方技术文档(如 Python 官方文档)提供了大量关于编程语言、框架和工具的实战知识。

(4)高校内部数据:在脱敏处理的前提下,整合教务系统中的课程设置数据和学生的学习行为日志。

2. 采集技术

(1)网络爬虫技术:针对网页数据,采用 Scrapy 等框架编写分布式爬虫,高效抓取网页内容。

(2)API 接口调用:对于提供 API 接口的学术数据库和在线平台,直接通过 API 获取结构化数据,确保数据的规范性。

(3)多媒体解析:针对教学视频,利用 ASR(自动语音识别)技术提取音频文本,利用 OCR 技术提取视频中的关键帧信息。

(二)数据处理与知识抽取

原始数据往往包含大量噪声,且格式不一,因此需要进行严格的数据清洗与知识抽取。

1. 数据清洗

(1)去重与降噪:利用哈希算法去除重复的网页记录,利用正则表达式去除 HTML 标签、广告等无关信息。

(2)标准化处理:统一日期格式、单位制,并对同义词进行归一化处理(例如将“AI”“人工智能”统一为“人工智能”)。

(3)基于 BERT 的数据校正:引入 BERT 模型对文本进行语义增强和拼写纠错,特别是针对学生论坛中非规范的表达进行规范化处理,提高后续抽取的准确率。

2. 知识抽取

(1)实体识别(Named Entity Recognition,NER):识别文本中的关键实体,如“Java”“区块链”“哈希表”等。本研究采用基于 BiLSTM-CRF 的深度学习模型进行实体识别,并针对计算机专业术语进行微调。

(2)关系抽取(Relation Extraction):

文本数据:利用依存句法分析(Dependency Parsing)识别句子中的主谓宾结构,从而提取实体间的关系。例如,从句子“Java 继承了 C++的语法”中提取出“Java—继承—C++”的关系。

表格与结构化数据:采用基于规则的映射方法,直接将数据库表中的字段映射为图谱中的属性。

(3)实体对齐:由于数据来源不同,同一实体可能有不同表述。本研究利用 Word2Vec 模型计算实体名称的语义相似度,并结合图算法(如 SimRank)计算结构相似度,实现跨数据源的实体对齐。

(三)知识表示与存储

知识表示决定了计算机如何“理解”知识,而存储方案则决定了知识的访问效率。

1. 知识表示模型

(1)符号表示:采用 RDF(资源描述框架)和 OWL(网络本体语言)作为形式化表示工具。定义计算机学科的顶层本体,包括“课程”“知识点”“技术”“人物”“论文”等核心类(Class),以及“先修”“包含”“应用”等属性(Property)。

(2)分布式表示(Knowledge Embedding):为了支持深度学习和知识推理,本研究采用 TransE、

TransR 等算法将图谱中的实体和关系映射到低维向量空间。这种表示方法能够捕捉知识的语义特征,支持计算实体间的相似度。

2. 知识存储架构

(1)图数据库选型:选择 Neo4j 作为底层存储引擎。相比于传统的关系型数据库(MySQL),图数据库在处理复杂关联查询时具有指数级的性能优势。例如,查询“学习深度学习需要掌握哪些前置数学知识”在图数据库中只需几毫秒,而在关系型数据库中可能需要多表连接,耗时较长。

(2)系统架构设计:设计分层架构,包括:

数据接入层:负责数据的采集与缓存(Kafka)。

知识处理层:负责数据清洗、抽取、融合及入库。

服务接口层:提供 RESTful API 供上层应用调用。

可视化层:利用 ECharts 或 Gephi 实现知识图谱的可视化展示。

二、计算机学科知识图谱应用

(一)教学应用

1. 课程规划与体系优化

传统的计算机课程设置往往依赖于专家经验,容易出现课程内容陈旧或知识点衔接不畅的问题。基于知识图谱的课程规划能够提供数据驱动的决策支持。

(1)知识点依赖分析:通过图谱分析,可以清晰地看到哪些知识点是基础(如“数据结构”),哪些是进阶(如“操作系统”)。这有助于教务管理者合理安排课程的开课学期,确保学生具备必要的先修知识。

(2)课程内容动态调整:通过监测图谱中新涌现的热点节点(如“大模型”“AIGC”),教学管理者可以及时将前沿技术引入课堂,更新教学大纲,保持课程的先进性。

2. 精准资源推荐

(1)语义化检索:在海量的在线学习资源中,学生往往面临“选择困难症”。传统的关键词搜索只能匹配字面,而基于知识图谱的搜索能够理解语义。例如,学生搜索“如何做推荐系统”,系统不仅能返回包含该关键词的视频,还能返回“协同过滤算法”“矩阵分解”等相关知识点的资源。

(2)个性化推荐算法:结合协同过滤与知识图谱路径分析,构建推荐模型。若系统发现学生 A 掌握了“Python 基础”但对“Django 框架”不熟悉,而学生 B(学习路径相似)在掌握 Python 后学习了 Django 并获得很高的评价,则系统会向学生 A 推荐 Django 相关资源。

(二)科研应用

1. 学科发展趋势分析

利用知识图谱的可视化与计量分析功能,可以全景式地展示计算机学科的发展态势。

(1)热点探测:通过分析图谱中节点的增长速度和连接密度,识别当前的科研热点。例如,近年来图谱中“边缘计算”“联邦学习”节点的连接数激增,表明这些是当前的研究热点。

(2)演化路径追踪:通过时间切片技术,观察知识图谱随时间的演变。例如,追踪“人工智能”这一概念从符号主义到连接主义,再到深度学习的演化路径,帮助科研人员把握学科发展的历史规律。

2. 跨学科研究辅助

计算机学科具有极强的工具属性,常与其他学科交叉。知识图谱能够打破学科壁垒。例如,系统可以自动发现“计算机视觉”与“医学影像诊断”之间的潜在联系,为从事医疗 AI 研究的学者提供跨学科的文献推荐和合作线索。还可以将学者、机构、项目纳入图谱,分析学者之间的合作强度与知识互补性,从而辅助组建高效的科研团队。

(三)学习应用

1. 个性化学习路径规划

每个学生的基础和兴趣都不同,千篇一律的教学进度无法满足所有人的需求。

(1)诊断式学习路径生成:系统首先通过前测构建学生的“知识画像”,识别其已掌握和未掌握的知识点。基于图谱中的拓扑结构,利用最短路径算法或遗传算法,为学生生成一条从当前状态到目标状态的最优学习路径。

(2)动态调整机制:在学习过程中,系统根据学生的实时答题情况动态调整路径。若学生在“指针”概念上反复出错,系统会自动插入相关的基础练习,或推荐替代的学习资源,体现了“最近发展区”理论。

2. 学习效果评估

传统的考试往往只能给出一个分数,缺乏过程性评价。基于知识图谱,可以将学生的知识状态映射为图中的一个子图。通过计算该子图与标准图的覆盖率、连通度等指标,可以量化学生对知识体系的掌握程度。系统不仅能指出学生哪道题做错了,还能通过反向追溯图谱,找到导致错误的根本

原因(例如是因为“循环语句”没学好,还是“数组”概念不清),从而提供针对性的改进建议。

三、应用效果评估与问题挑战

(一)应用效果评估

为了验证本研究提出的方案有效性,本研究选取作者所在高校计算机学院进行了为期一学期的实证研究。

1. 量化评估指标

(1)图谱构建质量:采用准确率(Precision)、召回率(Recall)和 F1 值评估知识抽取的效果。结果显示,实体识别的 F1 值达到 89.5%,关系抽取的 F1 值达到 85.2%。

(2)推荐系统性能:在资源推荐场景中,采用覆盖率、多样性和点击率(CTR)作为指标。实验组(使用图谱推荐)的资源点击率比对照组(传统推荐)提高了 23%。

(3)学习成效:对比实验班和普通班的期末成绩。实验班学生的平均分提高了 12%,且在解决综合性、设计性题目上的表现显著优于普通班。

2. 质性评估

(1)问卷调查:向参与实验的师生发放问卷。90% 以上的学生认为基于图谱的学习路径规划有助于理清知识脉络;85% 的教师认为图谱辅助了课程设计。

(2)访谈分析:深度访谈发现,学生普遍反馈系统推荐的资源更精准,减少了寻找资料的时间;教师则认为图谱帮助他们发现了课程体系中的一些冗余和缺失。

(二)面临的问题与挑战

尽管应用效果显著,但在实际推进过程中仍面临诸多挑战。

1. 数据隐私与安全

教育大数据包含大量学生的个人信息和行为数据。在数据采集和利用过程中,如何平衡数据价值挖掘与隐私保护是一大难题。现有的差分隐私技术虽然能保护隐私,但往往以牺牲数据精度为代价。

2. 知识更新的时效性

计算机学科发展日新月异,新的知识(如新的编程语言特性)每天都在产生。目前的图谱更新多为批量处理,存在时间滞后。如何实现知识的流式处理和实时更新,是未来需要攻克的技术难点。

3. 技术门槛与成本

构建和维护一个高质量的知识图谱需要高昂的成本,包括高性能计算资源、专业的 NLP 工程师团队等。这对于许多普通高校而言是一笔不小的负担。

四、结论与展望

(一)研究总结

本研究以教育大数据为驱动,深入探讨了高校计算机学科知识图谱的构建与应用。主要结论如下:

1. 构建了完善的图谱体系

提出了一套从多源数据采集、清洗、抽取到存储的完整技术路线,成功构建了覆盖计算机核心课程与前沿技术的知识图谱。

2. 验证了应用价值

通过在教学、科研和学习评估中的应用实践,证明了知识图谱能够有效提升教学资源的匹配效率,优化科研管理决策,并显著提高学生的学业成绩。

3. 提供了评估依据

建立了量化与质性相结合的评估体系,为后续相关研究提供了参考范式。

(二)未来展望

未来的研究将从以下几个方向展开:

1. 隐私计算技术的融合

探索联邦学习(Federated Learning)在教育大数据中的应用,实现“数据不动模型动”,在保护隐私的前提下进行知识图谱的联合构建。

2. 智能化的动态更新

研究基于增量学习的知识图谱更新算法,结合大语言模型(LLM)的自动摘要和推理能力,实现图谱的自动化、实时化维护。

3. 全息化教育场景

结合 VR/AR 技术,将知识图谱从二维屏幕延伸至三维空间,为学生提供沉浸式的知识探索体验。

参考文献:

[1]许惠,马雪赟,刘春蕾. 基于 CiteSpace 的国内计算机学科发展历程及前沿剖析[J]. 内蒙古科技与经济,2023(19):133-138.

[2]李家瑞,李华昱,闫阳. 面向多源异质数据源的学科知识图谱构建方法[J]. 计算机系统应用,2021,30(10):59-67.

[3]程格平,谷琼,宁彬,等. 基于学科知识图谱的个性化学习模型构建研究[J]. 科教文汇,2024(9):58-62.

[4]李华昱,刘烨宸,李家瑞,等. 基于异质数据源的计算机学科知识图谱构建[J]. 计算机系统应用,2022,31(6):10-18.

[5]高茂,张丽萍.融合多模态资源的教育知识图谱的内涵、技术与应用研究[J].计算机应用研究,2022,39(8):2257-2267.

[6]张勇,杨进才.基于学科知识图谱的高校教学模式研究[J].计算机教育,2021(6):141-144.

Research on the Construction and Application of University Disciplinary Knowledge Graph Driven by Educational Big Data

GUO Na

(School of Computer Science and Engineering, Institute of Disaster Prevention, Langfang Hebei 065201, China)

Abstract: With the rapid development of information technology, educational big data has emerged as a core driving force propelling the transformation of higher education. Against the backdrop of the educational big data era, this paper takes computer science—a discipline characterized by a complex knowledge system and rapid updates—as the research subject, and delves into the construction methods and application paradigms of disciplinary knowledge graphs. Initially, the study analyzes the collection and processing techniques for multi-source heterogeneous educational data, take computer science as an example, proposing a scheme for constructing a computer science knowledge graph based on deep learning and ontology. Subsequently, an application system is established, encompassing course planning, resource recommendation, analysis of research trends, and personalized learning path planning. Finally, through a combination of quantitative and qualitative evaluation methods, the study verifies the significant effects of the knowledge graph in enhancing teaching quality, optimizing research management, and facilitating students' deep learning. This research aims to explore the theoretical underpinnings and technological pathways for the digital transformation of university education, fostering the development of computer science education towards intelligence and precision.

Key words: educational big data; disciplinary knowledge graph; computer science; personalized learning; knowledge representation