网络课程知识图谱的分治模式设计及构建方法

管铮懿[1] 游兰[1] 吕顺营[1*] 何渡[2]

[1](湖北大学计算机与信息工程学院 湖北 武汉 430062;)

[2](湖北省科技信息研究院 湖北 武汉 430071)

摘 要 针对网络课程平台多源、缺乏统一课程知识图谱的问题，提出了网络课程知识图谱的分治模式设计与构建方法。首先，针对知识图谱的局部模式层特性，通过在核心本体基础上持续扩展的“分治法”构建模式层；其次，针对多源异构数据实体的抽取难题，提出结合构建词典库、TF-IDF 函数与 BiLSTM-CRF 神经网络的方式抽取实体；通过建立课程主题实体拓宽知识图谱深度，挖掘隐藏实体；最后结合本体推理与随机游走算法补全实体关系，丰富语义。实验以主流的三个网络课程平台为基础，构建了包含 1108 个隐藏实体，共计 20219 个实体、75102 对关系的多源网络课程知识图谱，验证了本文方法的准确性和有效性。

关键词 知识图谱 图数据库 实体抽取 本体构建 多源异构 教育大数据 关系补全

中图分类号 TP391.1 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2018.01.001

Divide-and-conquer Model Design and Construction Methods for Multi-source-based Online Course Knowledge Graph

GUAN Zhengyi[1] YOU Lan[1] LV Shunying[1*] HE Du[2]

[1](Hubei University, School of Computer Science and Information Engineering, Wuhan 430062, Hubei, China)

[2](Hubei Institute of Science and Technology Information, Wuhan 430071, Hubei, China)

Abstract To solve the difficult problems in constructing a multi-source course platform-based knowledge graph, a division-and-conquering method and a construction method have been proposed. Firstly, because of the local pattern layer characteristics of the knowledge graph, the overall pattern layer of the knowledge graph is constructed through the division-and-conquering method which is continuously extended based on the core ontology. Secondly, to solve the problem of multi-source heterogeneous data entity extraction, an entity extraction method using the Term Frequency–Inverse Document Frequency (TF-IDF) function and Bi-Directional Long Short-Term Memory Conditional Random Fields (BiLSTM-CRF) neural network is proposed to construct a dictionary database. Third, to enrich entity relationships, a subject concept to widen the depth and width of the knowledge graph using text data was built. Finally, combine ontology reasoning and random walk algorithm to complete the relationship of the knowledge graph and enrich the semantics of the knowledge graph. Based on the data of three main online course platforms, a multi-source online course knowledge graph including 1108 hidden entities, 20219 entities, and 75102 pairs of entities was constructed to verify the accuracy and effectiveness of the proposed method.

Keywords Knowledge graph Graph database Entity extraction Ontology construction Multi-source heterogeneous data Education big data Relationship completion

0 引 言

开放式网络课程是未来教学与学习模式发展的关键驱动力和趋势[1]。然而由于网络课程的大规模与开放性，使海量的课程内容分布在不同的网络平台上，呈现出去中心化、自组织化等特点，给缺乏相关背景素养的学习者带来使用困难。知识图谱是经过结构化的命名实体[2]，为知识的有效组织与语义表达提供了良好的辅助作用[3-4]。建立统一的大规模网络课程知识图谱，可以提供直观有效的知识聚合信息，并从深度与广度上有效地揭示课程资源的内在价值，为学习者提供标准统一且具有专业指导意义的网络课程数据体系。

然而，如何从复杂分散的真实网络课程数据中提取

收稿日期: 20yy-mm-dd。湖北省教育厅科学技术研究计划重点项目（D20201006）；湖北省自然科学基金面上项目（2019CFB757）。
管铮懿，主研领域：知识图谱与轨迹数据挖掘；游兰，主研领域：时空大数据与城市智能计算；吕顺营，主研领域：数据挖掘与机器学习；何渡，主研领域：高技术服务业与制造业信息化。

出可靠的知识模式与实体，从而科学构建网络课程知识图谱尚无完整成熟的方案。本文以三个中文网络课程平台作为信息源，根据大规模开放课程的数据特点，设计了针对小规模模式层构建的“分治法”，有效减少模式构建中对专业知识的依赖，提高小规模的模式层设计效率。针对数据集中不同类型的数据域与隐藏实体，结合构建字典库、建立课程主题实体、使用 BiLSTM-CRF 神经网络进行抽取来优化多源异构数据抽取实体的精度，并实现对隐藏实体的抽取。针对知识图谱中的隐藏关系，使用本体推理与随机游走算法补全关系，扩展了知识图谱语义。最终构建了一个包含 20219 个实体、75102 对实体关系的多源网络课程知识图谱。本文的构建流程如图 1。

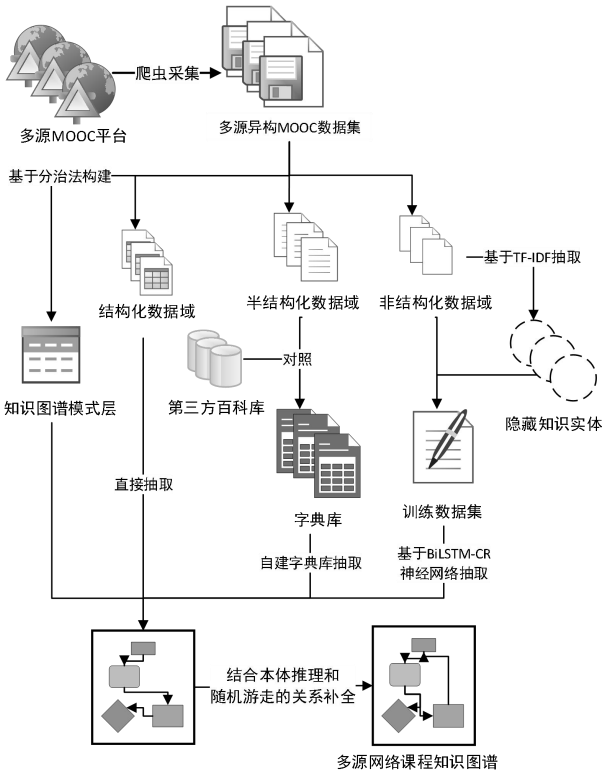

图 1 基于多源数据的网络课程知识图谱构建流程

.....

5 结 语

对于网络课程知识图谱构建过程中，多源异构数据带来的构建模式层困难、实体抽取精度不高、难以利用文本数据抽取隐藏实体的构建难题，本文详细阐述了针对多源网络课程知识图谱的分治模式设计及实体与关系抽取方法，并基于三个主流网络课程平台数据构建了网络课程知识图谱。实践表明本文“分治法”模式构建方式减少了小规模模式层的构建中对于专业知识与专业数据库的依赖，提高了构建效率；通过针对异构数据结构特性的实体抽取方式在挖掘隐藏实体、提高抽取准确率上有着明显优势；通过关系补充提高了知识图谱的信息深度与可解释性。下一步工作将针对网络口语文本优化实体抽取方法，从而适应更灵活多变的网络语境。