

融合专题知识和科技文献的科研 知识图谱构建*

李娇¹ 孙坦^{2,3} 黄永文¹ 鲜国建^{1,3} 罗婷婷¹ 赵瑞雪^{1,3}

(1. 中国农业科学院农业信息研究所, 北京 100081; 2. 中国农业科学院, 北京 100081;

3. 农业农村部农业大数据重点实验室, 北京 100081)

摘要: 本文为解决领域科技文献与专题价值的割裂问题提出深度融合科技文献、科研活动等科研对象与领域专题数据资源的图谱构建方法。通过主题词关联设计, 构建包含期刊论文、期刊、科研机构、科研人员及专题实体类型的科研本体, 选取机器学习专题构建科研知识图谱, 并基于图数据库Neo4J进行图谱管理与查询验证。该专题科研知识图谱可以支持单实体/属性、多实体事实性问题的复杂图谱查询, 有效揭示专题、科技文献的关联关系, 具有较强的应用价值, 可以为面向文献数据的智能知识服务提供新的思路 and 方向。

关键词: 科研本体; 专题数据; 科研知识图谱; 复杂图谱查询

中图分类号: G203 **DOI:** 10.3772/j.issn.1673-2286.2021.01.001

引文格式: 李娇, 孙坦, 黄永文, 等. 融合专题知识和科技文献的科研知识图谱构建[J]. 数字图书馆论坛, 2021 (1) : 2-9.

知识图谱是采用图结构来建模和记录世界万物之间关联关系和知识的技术^[1], 可分为语义万维网领域提出的基于RDF三元组存储的语义知识图谱(关联数据)和数据库领域发展形成的基于属性图数据库的广义知识图谱, 与之对应的图谱存储方式分别是面向RDF的三元组数据库(如OpenLink Virtuoso、GraphDB等)和原生图数据库(如Neo4J、JanusGraph等)。随着三元组数据库和图数据库的相互融合发展, 知识图谱的构建、存储与管理也呈现交叉丰富之态。图情及信息管理领域, 语义知识图谱是主流的数据关联整合及语义组织方式^[2], 基于RDF三元组的表示方法是构建语义知识图谱时的常用方法, 其核心在于为数据资源嵌入机器可读的语义标签, 它提供了知识建模的灵活性, 同时也支持基于实体模型的实体、属性及实体关系的扩展, 支撑推理应用。

近年来, 知识图谱相关研究取得了极大进展, 实践

成果丰富。在以科技文献资源为语料基础的学术领域, 知识图谱通常用于支持数据发布、知识搜索、知识标引等形态的应用, 服务于专业技术人员或特定行业从业人员。其中典型的知识图谱有Springer Nature推出的SciGraph^[3-4], 通过构建本体SciGraphCore ontology并采用序列化N-Triples格式的三元组进行数据表示, 集成整合了如科研机构、资助项目、科研单位、会议、期刊论文、图书和论文附属数据等各种信息, 包含15亿~20亿条学术对象的信息(三元组), 并与Springer Nature API Portal集成在一起支持Springer Nature REST API直接查询数据^[5]; 开放学术组织(Open Academic Society)发布的亿级学术图谱——开放学术图谱(Open Academic Graph, OAG)^[6]集成了微软学术图谱(Microsoft Academic Graph, MAG)及科技情报大数据挖掘与服务系统平台AMiner^[7]的数亿篇论文, 并生成两个图之间的6 000多万链接(匹配)关系供科研使用; Elsevier

*本研究得到国家社会科学基金一般项目“科技论文全景式摘要知识图谱构建与应用研究”(编号: 19BTQ061)资助。

Knowledge Graph以Elsevier学术资源为基础语料构建科研、生命科学、医疗保健3个领域的知识图谱, 包含简介、机构、基金、论文、机构等信息。

然而, 现有基于科技文献、科研活动等科研对象的图谱实践与领域专题关联较弱, 融合深度欠缺。专题知识图谱构建语料多来自百科、垂直网站、开源数据等, 通常采用基于规则、统计模型或深度学习的知识抽取方式, 依赖大量的样本与人力标注; 涉及主题模型的学术图谱则多是基于关键词匹配且主题内词间关系不明晰, 语料及技术路径的差异性导致科技文献与专题价值呈现割裂状态, 难以支撑涉及领域及文献多实体的复杂图谱查询等应用, 科技文献与专题知识两者关联关系的构建是解决这一难题的关键。本文在现有研究基础上, 设计了涵盖数据预处理、RDF格式知识抽取、RDF2PG的科研知识图谱构建技术路径。具体来说, 基于专题数据特征设计描述专题内词间关系的数据模型, 并结合科技文献类的成熟语义描述模型, 以主题词为关联点设计构建融合专题数据模型和科技文献结构特征的科研本体模型。以机器学习(Machine Learning)专题为例, 选取Web of Science核心合集作为文献来源检索筛选出200篇英文期刊论文构建专题语义知识图谱即RDF图, 继而基于图谱映射规则、Neo4J图数据库插件及导入函数进行RDF图的属性图转化, 实现科研知识图谱管理并支持事实型客观问题查询。

1 科研知识图谱本体建模

知识图谱由模式层和数据层两个逻辑层次组成, 模式层即描述抽象知识的本体层, 是知识图谱的核心, 构建在数据层之上并用于约束数据层; 数据层则是由大量描述具体实体对象的事实和数据组成。知识图谱的建设核心是建立本体模型和实体数据库, 两者构建顺序定义的方法有“自顶向下”和“自底向上”两种。针对本次研究选用的专题数据情况及知识体系完备程度, 本文采用自顶向下的方式进行科研知识图谱构建。

本体知识建模是知识图谱内实体连通的语义基础, 需要设计概念集合和框架并对知识进行合理的组织以准确描述相关的类和属性, 即实体间的关系。知识建模需要结合数据及应用特征, 准确地描述知识本身与知识的关联, 以尽量减少数据冗余、提高应用效率为

根本原则设计科学高质量的模式。

本文选取期刊论文为主要研究对象, 并根据来源数据特征和应用场景, 继承复用RDF、SKOS等数据结构及BIBO^[8]、PRISM^[9]等成熟本体模型或元数据标准中定义的基础元属性, 建立融合专题数据模型和科技文献结构特征的科研本体, 命名空间为http://linked.scikg.com/onts/ (前缀scikg:), 包含期刊论文、期刊、主题、科研人员、科研机构5种实体类型, 并为其建立URI (见表1)。其中, scikgId为数据资源在数据库中的唯一标识符。

表1 实体对象的URI

实体对象	URI
Article (期刊论文)	http://linked.scikg.com/articles/{scikgId}
Journal (期刊)	http://linked.scikg.com/journals/{scikgId}
Subject (主题)	http://linked.scikg.com/subjects/{scikgId}
Person (科研人员)	http://linked.scikg.com/person/{scikgId}
Organization (科研机构)	http://linked.scikg.com/organizations/{scikgId}

科研本体各实体类型数据关系模型如图1所示。期刊论文类通过对象属性scikg:about、scikg:subjectOf与主题类进行链接, 实现领域主题和科技文献的关联关系构建; 通过scikg:rightHolder、scikg:hasAuthor、dcterms:isPartOf分别与科研机构类、科研人员类、期刊类进行链接。同时, 期刊论文类数据模型提供了12个数据属性, 包括类名、题名、摘要、关键词、出版日期、文献重要度(scikg:importantDegree)、期号、卷号、起始页、结束页、数字对象标识符(Digital Object Identifier, DOI)、被引频次(scikg:timeCited)。

考虑到面向终端科研用户的文献遴选等应用场景需求, 引入文献重要度^[10]这一重要指标属性, 取值为数值。文献重要度与科技文献的多个属性相关, 如被引次数、所在期刊的影响因子、作者科学价值、文献内容等。

主题类数据模型用于描述专题知识范畴内各子概念或术语的词间关系, 需在充分分析所选专题特征及知识服务需求的前提下进行设计。本文以机器学习专题为例, 通过相关调研的总结归纳确定了scikg:isKindOf、scikg:useTechnique、scikg:learningParadigm、scikg:subjectOf 4个对象属性, 以及类名、术语名、所属学科、术语简称、应用于5个数据属性。期刊数据类主要包括1个对象属性dcterms:isPartOf, 以及类名、期刊名、出版商、ISSN、期刊影响因子(Journal Impact Factor,

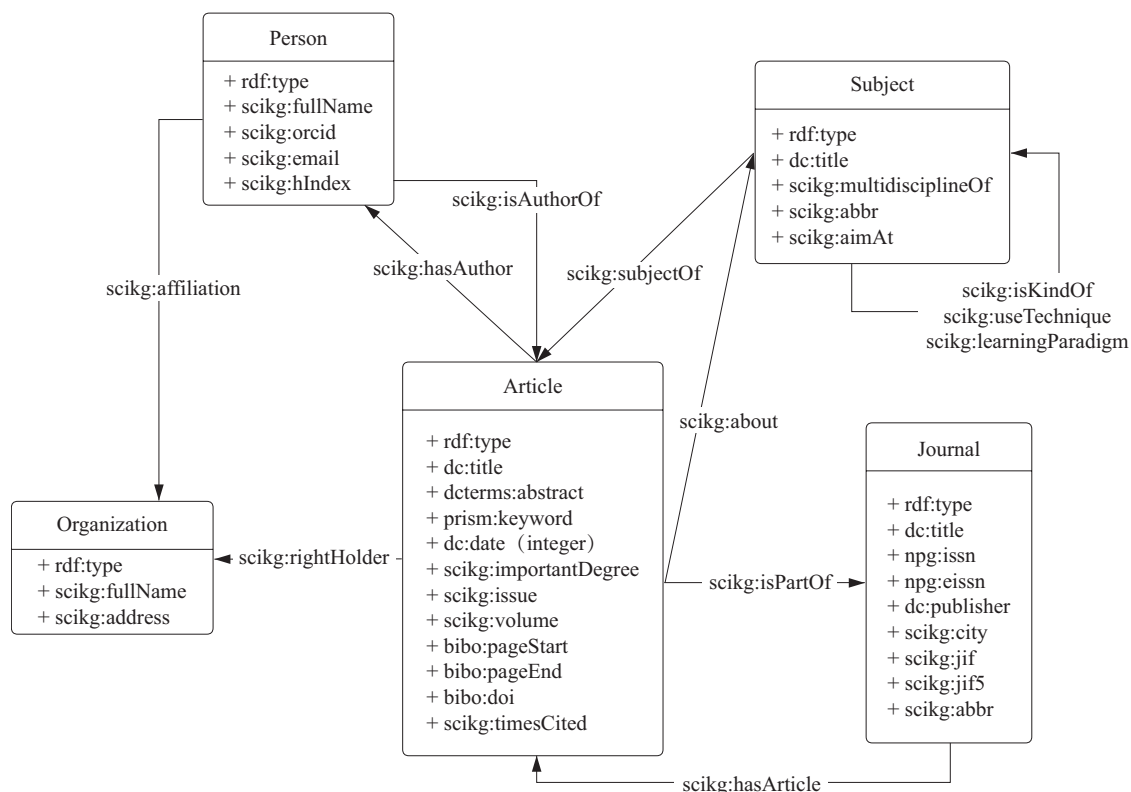


图1 科研本体数据关系模型

JIF)等8个数据属性,科研机构类和科研人员类各包括3个、7个属性。

上述为本次研究中结合可获取数据来源特征并支撑验证实验的科研本体数据关系模型,实际应用中可根据来源数据元数据复杂度进行扩充或修改,如添加更多实体类型如资助项目、会议论文、著作等,以及各类的属性。

2 科研知识图谱数据层构建

科研本体构建后表示知识图谱的模式层已经建立,本节将重点介绍知识图谱数据层(实例层)的构建。随着信息技术和大数据的发展,科技文献的结构化发布模式极大地方便了科研知识图谱的构建及语义揭示,批量文献语料的获取方式也愈加丰富,如长期积累自建的文件数据、知识服务网站(或学术信息数据库)以及开源数据下载等。由于人工标注数据集的方式成本高且规模有限,本文面向研究任务需求遴选了高质量开放数据集并辅以人工补充,主要包括专题数据和英文期刊论文数据两部分。

2.1 专题模型构建

选定的Machine Learning专题模型构建数据来源是从AMiner提取的结构化实体网络——Aminer开放数据集“Knowledge Graph for Machine Learning”^[11],通过百科及其他文献著作^[12]予以修改补充,生成实例节点100个,其中一级节点及其关系如图2所示,知识图谱构建时通过定义的若干对象属性实现URI的内部链接,包括所属学科、术语简称、应用于等。

此外,本文将所有节点的名称(术语名)、别名、单复数形式等整合构建主题词表,作为后续期刊论文主题标引的标注语料库来源。

2.2 数据获取与预处理

基于知识模型的语料数据补全及预处理是科研知识图谱构建的重要环节,主要目的是解决语料数据异构、缺省等问题,并通过主题匹配的方式在科技文献数据和专题数据之间建立相关关系。本文以Web of Science核心合集为文献来源,选取并导出200篇期刊论文全纪录作为本文知识图谱数据来源,包括标题、摘要、关键词、

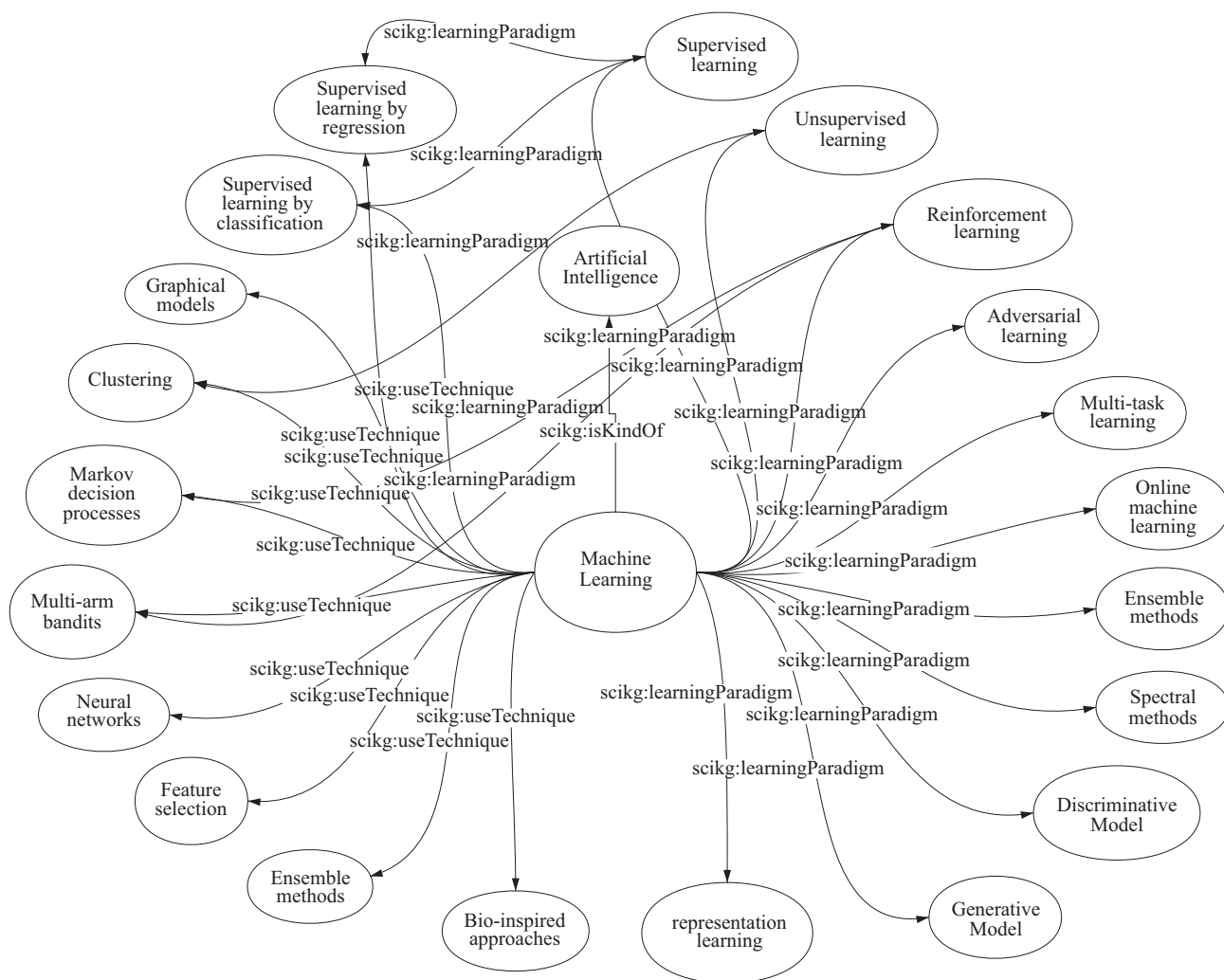


图2 Machine Learning专题模型一级节点及其关系

作者列表、地址、期刊等信息, 本次研究不涉及处理文献全文。除此之外, 通过人工方式补充机构名、作者record_id、H-index、总发文量、总被引频次、期刊影响因子(2019年)、期刊5年影响因子等字段信息, 构建以实体类型为主的关系型数据库表, 并进行相应的数据加工预处理, 如实体消歧、文献重要度计算、文献主题标引与关联。

2.2.1 实体消歧

实体歧义是影响语料数据质量的重要因素, 同一实体名称对应多个概念会直接降低基于知识图谱的信息检索、研究评估等各类应用的准确性, 因此实体消歧工作必不可少。科研本体涉及的实体类型中, 主要以科研人员、科研机构两类数据的消歧为主。

(1) 科研人员消歧即论文作者消歧。由于部分数

据缺学术标识符(如ORCID、ResearcherID), 通过补充添加Web of Science中每篇期刊论文链接到作者详情页地址栏中的record字段作为唯一标识符record_id进行消歧。

(2) 科研机构消歧。Web of Science中导出获取的是机构地址信息, 通过人工方式补充相应的机构名称并进行消歧。

2.2.2 文献重要度计算

前文提及科技文献重要度是期刊论文数据模型中的一个至关重要的属性, 通常是指依托期刊或会议发表的科技论文影响力与学术价值的定量评估, 与论文被引次数、所在期刊或会议的影响因子、作者的权威性、发表年份及内容质量等多个因子紧密相关。目前常用的评价方法主要有依据文献被引频次的引用计数算法^[13]、

Google提出的基于PageRank算法和文献间链接关系的PR_P排名算法^[14]、Teoma中应用的链接分析算法HITS (Hyperlink - Induced Topic Search)^[15]及以此为基础的Co-Rank模型^[16]等,相关研究成果多应用于网络中的搜索引擎,图情领域的案例有Aminer溯源树MRT研究中的文献重要程度^[17]。本文在已有研究成果的基础上结合可获取数据的属性特征通过复合加权建模的方式设计文献重要度评价算法,指标说明如表2所示。

表2 文献重要度各参数指标

指标维度	指标 (I_i)	说明
内容	年均被引频次	该篇论文总被引次数与发表年限之比
载体质量	期刊影响因子	该刊前两年发表的文献在当前年的平均被引用次数
	期刊5年影响因子	该刊前5年发表文献的平均被引次数
作者科学价值	篇均被引频次	发表论文的总被引次数与其发表论文数之比
	作者影响因子	作者科学价值指数

根据数据特征,本文分别从内容、载体质量、作者科学价值3个维度遴选指标 I_i ($i \in [1, 5]$)并确定相应权重 W_i ($i \in [1, 5]$),其中作者影响因子 (Author Impact Factor, AIF) 以H指数 (H-index) 表示,且每篇论文的作者科学价值为所有作者的平均值。对于较近年份的需做人工审核,避免因文献太新导致的重要度低估。

各指标归一化计算得分如公式 (1) 所示。

$$N_i = \frac{I_i - I_{\min}}{I_{\max} - I_{\min}} \quad (1)$$

其中, $I_{\max}$ 为指标 I_i 的最大值, $I_{\min}$ 为指标 I_i 的最小值。

文献重要度得分如公式 (2) 所示。

$$\text{importantDegree} = \sum_{i=1}^5 N_i \times W_i \times 100\% \quad (2)$$

需要说明的是,文献重要度评价中各参数均为当前年份数据,此项属性计算可依据需求定期更新,或作为期刊论文类的复合属性,基于图谱数据中涉及的相关属性值进行实时计算。

2.2.3 文献主题标引与关联

主题词是论文数据模型与主题数据模型的关联枢纽,本文的论文主题标引方案主要包括如下步骤。①待标引文献分词处理与关键词提取。通过本体化改进的IK分词工具及构建的专题术语词典和若干成熟通用词

表如STKOS、CAT等对待标引文献的题录信息 (标题、摘要及作者关键词) 进行切词并提取关键词。②关键词语义匹配,即词串匹配的相符性比较。将提取的关键词与选定的标注语料库中关键词进行语义匹配。具体地,该方法在传统文本匹配 (字面匹配) 的基础上同时进行关键词词义扩展的对比分析,如单复数、缩写等的匹配。其中,标注语料库是将专题术语词典与已有自建库及专题相关领域权威语料库叠加使用。③打分与主题词确定。基于命中的关键词及其所在题录信息的位置 (如标题、摘要或者作者关键词) 进行加权计算,最高分指定为论文主题词。此过程需要涉及一定的人工审核以确保标引质量。

文献主题标引完成后,可通过在专题数据表与期刊论文数据表添加匹配上实例的资源唯一标识符的方式实现关联。如某篇期刊论文关键词为Electromyography、Supervised learning、Clustering algorithms、Electrodes,即可命中Supervised learning及Clustering两个主题词并将其添加在期刊论文数据表的主题字段中,使得后续知识图谱构建中将这篇文献通过对象属性scikg:about关联到这两个主题实例的URI。

2.3 语义知识图谱实现

知识图谱数据层构建过程本质上是从多源异构数据中抽取知识的过程,面向不同的数据源,知识抽取涉及的关键技术和需要解决的技术难点有所不同,需要综合考虑数据的规模、操作的稳定性和效率。语义知识图谱的构建方式主要是基于本体模型的三元组抽取,具体来说,在数据预处理过程中,将不同实体类型对应的数据文件进行分开存储,即可在实体类型已知的前提下通过工具或者算法进行实体及关系的抽取并存入知识图谱数据库中,入库后的各知识节点 (实体) 通过URI对齐。本研究预处理后的语料数据存储于关系型数据库,专题科研知识图谱生成主要是基于本地化改造的RDF ETL工具完成,该工具以开源ETL工具Kettle^[18]为运行环境,结合开源RDF数据处理框架RDF4J进行插件扩展,支持多源多类型数据通过W3C映射语言Direct Mapping的方式转换生成RDF三元组。

以此种方式构建的机器学习专题科研图谱分别按实体类型以序列化格式N-Triples的文件形式存储于本地,最终构建的机器学习专题科研知识图谱的数据记录如表3所示,共11 426个三元组。

表3 机器学习专题科研知识图谱的数据记录

数据类型	数据量/条	生成的RDF三元组数量/个
期刊论文	200	3 945
期刊	150	1 500
科研人员	747	4 139
科研机构	384	1 152
主题	254	690

3 科研知识图谱管理与查询

知识存储是支持用户查询、检索、分析等各种应用需求的必要条件,随着语义网领域的三元组库和数据库领域的图数据库的交叉融合发展,数据管理者可基于实际需求选择或定制知识图谱的存储和管理方式,本文选取原生图数据库Neo4J作为科研知识图谱存储和应用支撑工具。Neo4J是面向图的数据库,以属性图的方式存储、处理、查询和展示数据,在关系遍历和路径发现等应用中性能优越,且数据存取效率高,支持知识图谱的可视化及声明式查询语言Cypher。本文涉及的语义知识图谱的Neo4J存储过程本质上是RDF图向属性图的转化(RDF2PG),须在遵循图谱映射规则的前提下结合相应的工具及函数实现。

3.1 科研知识图谱属性映射规则

RDF图存储有两个关键特征:①以图形的形式表示、存储和查询数据;②具备语义,可以明确地描述数据的含义^[19]。语义知识图谱的Neo4J存储过程本质上是RDF图向属性图的转化,RDF图数据在“非语义”(non-semantic)图形数据库Neo4J中的存储需要解决schema信息及数据信息的存储问题,Neo4J在设计上支持OWL和RDFS这些基本的本体定义语言,进而可以支持语义词汇表。RDF图数据迁移到属性图数据库

的Neo4J首先要做的是模型映射,RDF图是主谓宾结构,主语和谓语都是资源,宾语可以是资源也可以是文本或数值等字面量(literal),若宾语是字面量这一特殊情况,则不能成为其他三元组的主语。需要说明的是,资源都由URI唯一标识。RDF图到Neo4J中的映射规则如下^[20]。

规则1(主语映射规则):RDF三元组主语节点的映射。Neo4J中表示RDF资源的节点标记为:resource,并具有带有资源URI的属性uri。

$$(S, P, O) \Rightarrow (:Resource \{uri:S\}) \dots$$

规则2a(谓语映射规则):如果三元组的宾语是字面量,那么在Neo4J中三元组的谓词将映射到节点属性。

$$(S, P, O) \ \&\& \ isLiteral(O) \Rightarrow (:Resource \{uri:S, P, O\})$$

规则2b(谓语映射规则):如果三元组的宾语是资源(resource),那么在Neo4J中三元组的谓词将映射到关系。

$$(S, P, O) \ \&\& \ !isLiteral(O) \Rightarrow \\ (:Resource \{uri:S\}) - [P] -> (:Resource \{uri:O\})$$

在遵循图谱映射规则的前提下,通过调用semantics.importRDF命令将构建的知识图谱文件导入Neo4J,即可实现可视化。

3.2 科研知识图谱存储过程

在遵循图谱映射规则前提下,Neo4J内置了支持RDF及其相关语法(如OWL、RDFS、SKOS等)的插件neosemantics(n10s)^[21],提供导入函数importRDF(),可以实现RDF图数据的无损导入。RDF图谱数据的Neo4J存储与管理流程如图3所示。

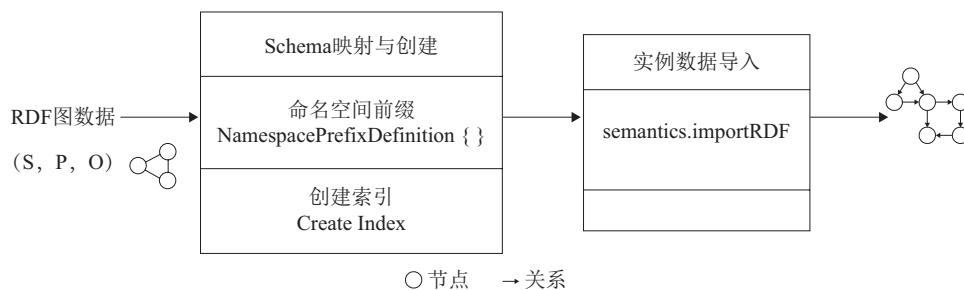


图3 RDF图谱数据的Neo4J存储与管理流程

转换并导入Neo4J的机器学习专题科研知识图谱共含2 181个节点,4 034个链接关系。期刊论文、期刊、科研人员、科研机构、主题5类实体类型在Neo4J中的节点标签分别为scikg_Article、scikg_Journal、scikg_Person、scikg_Organization、scikg_Subject,关系类型(Relationship Types)也均以“命名空间前缀_关系”的格式表示,如期刊论文实体类型和主题实体类型的“关于”关系可表示为scikg_about。

3.3 基于科研知识图谱的事实型问题查询

构建专题科研知识图谱拥有明确的主谓宾结构、语义结构清晰,可以支持事实性客观问题的查询,揭示专题、科技文献的关联关系。事实型问题在知识应用中处理频率较高,包含多种问题类型,如答案为单一对象的谓词型问题、返回一系列结果的列表型问题。具体来说,可以分为3类:命名实体基本定义、实体属性及复杂知识图谱查询。

(1) 单实体/属性查询。单实体/属性的查询是指询问命名实体的基本定义或者实体数据属性,如查询“机器学习主题最重要的五篇文献的标题及其重要度”,问题可解析为:找出期刊论文实体类型(scikg_Article)实例中,节点属性为文献重要度(scikg_importantDegree)且属性值最大的5个节点,输出其节点属性标题(dc_title),结果会返回整个科研知识专题知识图谱中文献重要度值为top5的期刊论文。

(2) 多实体联合查询。多实体联合查询则是指询问中涉及多类命名实体,如问题查询“主题为Supervised learning且影响因子大于5的文献的标题”,涉及期刊论文、主题、期刊3个实体类型,scikg_about、scikg_isPartOf 2个边属性和dc_title、scikg_jif 2个节点属性。

4 结语

知识图谱在文献资源与专题知识关联整合及复杂事实性问题查询方面发挥着重要作用。本文以机器学习专题为例,采用自顶向下的方式依次构建专题知识图谱概念层及数据层,基于“图谱映射规则+特定插件+函数”的方式实现图谱格式的转化与原生图数据库Neo4J的存储管理。构建的专题科研知识图谱语义结构清晰,支持实体列表、统计结果或者实体关系等复杂

知识图谱查询,具有一定的应用价值,可以为各类知识应用提供图谱数据及相应的技术支撑。

本文提出的专题科研知识图谱构建方法中,主题类数据模型的合理性及文献与主题模型的映射关联关系构建是关键。受限于实验数据清洗效率,未将其他类型文献如会议、会议论文、著作等实体类型纳入本次实验中,未来的研究将进一步扩展资源类型,扩充该知识图谱。此外,本文仅基于科技文献题录信息等结构特征及主题语义元素进行实验研究,未来的研究希望可以在以假设、观点、事实、结论等科学陈述为基本形式的陈述型语义特征上有所结合,丰富专题科研知识图谱形态,以期支撑图谱式研究侧写、智能问答、主题溯源等智能知识服务。

参考文献

- [1] 王昊奋,漆桂林,陈华钧.知识图谱:方法,实践与应用[M].北京:电子工业出版社,2019.
- [2] 张永娟,刘炜,于建荣,等.基于IIIF和语义知识图谱的印章资源整合与知识发现研究[J].图书情报工作,2020,64(7):127-135.
- [3] Springer Nature. SN SciGraph A Linked Open Data Platform for the scholarly domain [EB/OL]. [2021-01-10]. <https://www.springernature.com/gp/researchers/scigraph>.
- [4] 白林林,祝忠明. Springer Nature SciGraph关联开放数据分析[J]. 知识管理论坛,2018,3(1):2-11.
- [5] SN SciGraph Data Explorer 2019. Springer Nature SciGraph Explorer 2019 [EB/OL]. [2021-01-10]. <https://scigraph.springernature.com/explorer>.
- [6] Open Academic Graph [EB/OL]. [2021-01-10]. <https://www.openacademic.ai/oag/>.
- [7] TANG J, ZHANG J, YAO L, et al. ArnetMiner: Extraction and mining of academic social networks [C] // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 990-998.
- [8] Bibliographic ontology [EB/OL]. [2021-01-10]. <http://purl.org/ontology/bibo>.
- [9] PRISM METADATA [EB/OL]. [2021-01-10]. <http://www.idealliance.org/prism-metadata/>.
- [10] 张甫. 基于多信息融合的科技文献重要度评价方法的分析与研究[D]. 北京:北京工业大学,2016.
- [11] AMiner. AMiner Dataset [EB/OL]. [2021-01-10]. <https://>

www.aminer.cn/data/#AMiner-Knowledge-Graph.

- [12] 雷明. 机器学习: 原理、算法与应用 [M]. 北京: 清华大学出版社, 2019.
- [13] BEEL J, GIPP B. Google scholar's ranking algorithm: The impact of articles' age (an empirical study) [C] // Proceedings of the 2009 3rd International Conference on Research Challenges in Information Science, RCIS 2009: 160-164.
- [14] FIALA D, ROUSSELOT F, JEZEK K. PageRank for bibliographic networks [J]. Scientometrics, 2008, 76 (1): 135-158.
- [15] HITS algorithm [EB/OL]. [2020-12-31]. https://en.wikipedia.org/wiki/HITS_algorithm.
- [16] ZHOU D, ORSHANSKIY S A, ZHA H, et al. Co-ranking Authors and Documents in a Heterogeneous Network [C] // Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007). Omaha, NE, USA: IEEE, 2007: 739-744.
- [17] 溯源树 前沿科技探源 [EB/OL]. [2021-01-10]. <https://www.aminer.cn/mrt>.
- [18] Pentaho Data Integration-Kettle ETL tool [EB/OL]. [2021-01-10]. <https://etl-tools.info/en/pentaho/kettle-etl.htm>.
- [19] BARRASA J. Building a semantic graph in Neo4j [EB/OL]. [2021-01-10]. <https://jbarrasa.com/2016/04/06/building-a-semantic-graph-in-neo4j/>.
- [20] Importing RDF data into Neo4j [EB/OL]. [2021-01-10]. <https://jbarrasa.com/2016/06/07/importing-rdf-data-into-neo4j/>.
- [21] neo4j-labs/neosemantics [EB/OL]. [2021-01-10]. <https://github.com/neo4j-labs/neosemantics>.

作者简介

李娇, 女, 1989年生, 博士, 助理研究员, 研究方向: 关联数据、知识图谱。

孙坦, 男, 1970年生, 博士, 研究员, 研究方向: 数字信息描述与组织。

黄永文, 女, 1975年生, 博士, 副研究馆员, 研究方向: 知识组织与知识服务。

鲜国建, 男, 1982年生, 博士, 研究员, 研究方向: 知识组织与知识服务。

罗婷婷, 女, 1985年生, 硕士, 助理研究员, 研究方向: 数据融合与管理。

赵瑞雪, 女, 1968年生, 博士, 研究员, 通信作者, 研究方向: 农业信息管理, E-mail: zhaoruixue@caas.cn。

Construction of Scientific Knowledge Graph by Integrating Thematic Knowledge and Scientific Literature

LI Jiao¹ SUN Tan^{2,3} HUANG YongWen¹ XIAN GuoJian^{1,3} LUO TingTing¹ ZHAO RuiXue^{1,3}

(1. Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing 100081, China; 2. Chinese Academic of Agricultural Sciences, Beijing 100081, China; 3. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China)

Abstract: In order to solve the problem of the separation of scientific literature and subject value, this paper proposes a method to construct knowledge graph that deeply integrates scientific literature, scientific activities and other scientific research objects and thematic data resources, and constructs including. A scientific ontology with journal article, journal, scientific research institution, and researcher as the main scientific research subjects, is developed through the subject association, and the subject of "machine learning" is selected to construct the thematic scientific knowledge graph, then the graph management and query verification are performed based on the graph database Neo4J. The thematic scientific knowledge graph constructed in this study can support complex graph query of factual issues with single-entity/property or multi-entity, and effectively reveal the relationship between thematic and scientific literature, which shows certain application value and can provide new ideas and directions for intelligent knowledge service.

Keywords: Scientific Ontology; Thematic Data; Scientific Knowledge Graph; Complex Graph Query

(收稿日期: 2021-01-12)