计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(7):26−34 [doi: 10.15888/j.cnki.csa.006977]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

教育测评知识图谱的构建及其表示学习①

罗　明

(北京工业大学 信息学部, 北京 100124)
通讯作者: 罗　明, E-mail: wjlm2016@163.com

摘　要: 知识图谱旨在描述现实世界中存在的实体以及实体之间的关系. 自 2012 年谷歌提出“Google Knowledge Graph”以来, 知识图谱在学术界和工业界受到广泛关注. 针对教育领域中信息缺乏系统性组织的不足, 本文构建了面向高中的教育测评知识图谱 (Educational Assessment Knowledge Graph, EAKG), 其中 EAKG 的构建包括基于本体技术的知识图谱模式层构建和依托于模式层结构的知识图谱数据层构建. 与传统通过网页爬虫等技术手段构建的知识图谱相比, 本文构建的知识图谱优点在于逻辑结构清晰, 实体间关系的刻画遵循知识图谱模式层的定义. EAKG 为领域内知识共享, 知识推理, 知识表示学习等任务提供了良好的支撑. 在真实模考数据上的实验结果表明: 在试卷得分预测, 知识点得分预测的实体链接预测和三元组分类嵌入式表示学习任务上, 引入领域本体作为模式层构建的 EAKG 的性能优于没有领域本体模式层单纯由数据事实构成的 EAKG, 实验表明, 领域本体的引入对知识图谱的表示学习具有一定的指导意义.

关键词: 知识图谱; 教育测评; 语义网; 本体; 表示学习

引用格式: 罗明.教育测评知识图谱的构建及其表示学习.计算机系统应用,2019,28(7):26–34. http://www.c-s-a.org.cn/1003-3254/6977.html

Construction and Representation Learning of EAKG

LUO Ming

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: The Knowledge Graph is intended to describe the entities that exist in the real world and the relationships between entities. Since Google introduced the “Google Knowledge Graph” in 2012, knowledge graph have received widespread attention in academia and industry. Aiming at the lack of systematic organization in the field of education, the Educational Assessment Knowledge Graph (EAKG) for high schools is constructed. The construction of EAKG includes knowledge graph schema layer construction based on ontology technology and knowledge graph data layer construction based on schema layer structure. Compared with the traditional knowledge graph constructed by web crawling and other technical means, the knowledge graph constructed in this study has the advantages of clear logical structure and the description of the relationship between entities follows the definition of knowledge graph schema layer. EAKG provides good support for knowledge sharing, knowledge reasoning, knowledge representation learning and other tasks in the field. The experimental results on real simulated test data show that the EAKG constructed by introducing domain ontology as schema layer has better performance than EAKG constructed by data facts alone without domain ontology schema layer on the embedded representation learning tasks such as entity link prediction of test paper score prediction, knowledge point score prediction and triplet classification. Experiments show that the introduction of domain ontology has a certain guiding significance for knowledge graph representation learning.

① 基金项目: 国家重点研发计划 (2016YFB1000902); 南京普雷软件工程开发有限公司
Foundation item: National Key Research and Development Program of China (2016YFB1000902); Nanjing Pulei Software Engineering Co. Ltd.
收稿时间: 2019-01-12; 修改时间: 2019-02-03; 采用时间: 2019-02-18; csa 在线出版时间: 2019-07-01

Key words: knowledge graph; educational assessment; semantic web; ontology; representation learning

1 引言

教育测评包括两部分, 教育测量和教育评价. 教育测量是指针对学校教育影响下学生各方面的发展, 侧重从量的规定上予以确定和描述的过程, 关注点在于学校的教学效果, 反馈关于课堂教与学两方面的信息; 教学评价是指按照一定的价值标准和教育目标, 侧重对学生行为的定性描述, 利用测量和非测量的种种方法系统地收集资料信息, 对学生的发展变化及其影响学生发展变化的各种要素进行价值分析和价值判断, 并为教育决策提供依据的过程[1]. 纵观现有教育测评模式, 中学教学大多通过组织大规模的考试活动, 如多校联考, 地区统考、月考、周考等, 将学生组织在一块进行考试, 频繁的考试活动耗费师生大部分时间精力; 卷面测评覆盖内容有限和涉及知识点零散不连贯的固有弊端导致学生缺乏对知识系统性的认知[2]; 传统集中式的卷面测评很难适合各阶段学习情况的学生, 教学针对性很差; 为此, 各种测评系统不断涌现, 如全通教学质量监测平台[3]、教研测[4]等, 但是它们大多采用数学统计的方法整理分析考试结果并制作数据报表, 统计结果往往局限于学生直观的考试分数和排名, 无法对知识进行系统性地建模以帮助学生了解自己在整个知识结构上的掌握情况; 统计结果相对离散, 数据间的关联关系无法得到有效刻画, 在当今知识付费的大数据时代背景下, 无疑是对教育领域信息资源的浪费.

大数据时代的发展, 涌现了很多新技术, 其中知识图谱自谷歌公司于 2012 年提出“Google Knowledge Graph”以来受到学术界和工业界的广泛关注, 其技术特征在于描述现实世界中存在的实体以及实体之间的关系. 由于其强大的实体间关联关系刻画能力, 知识图谱被很多研究学者引入了教育领域, 并得到了应用实践, 如基于知识图谱的教育应用领域热点问题和前沿探索[5,6]、学科教学研究图谱分析[7]、领域信息可视化和知识挖掘等[8]. 本文将知识图谱引入教育领域, 提出构建教育测评知识图谱 (Educational Assessment Knowledge Graph, EAKG), 与其他领域知识图谱构建方法相比, 我们构建方法主要分为两层: 基于 Ontology[9] 的知识图谱模式层构建和依托于模式层结构的知识图谱数据层构建.

知识图谱的表示方法, 是知识图谱构建与应用的基础, 表示方式的好坏直接影响着在知识图谱上的计算效率, 进而影响着知识图谱在具体应用上的表现效果或性能. 目前主流的关于知识图谱表示方法有两种: 基于符号表示和基于分布式向量表示. 其中基于符号表示的知识图谱通常借助于逻辑规则、产生式等进行知识刻画, 其特有的强逻辑关联与规则推理能力, 在常识性知识图谱领域具有广泛应用, 如语义信息检索、智能问答等; 基于分布式向量表示的知识图谱弥补了符号表示面临的数据稀疏、图算法复杂、难以适应大规模计算等困境, 通过将知识图谱中的实体和关系的语义信息嵌入到连续稠密低维向量空间中, 在简化操作与计算的同时最大程度保留了原始的网络结构, 知识的向量表示为基于连续数值空间上计算的知识应用提供了应用基础.

本文工作主要贡献概括为以下两点: 首先, 将知识图谱技术引入教育测评领域, 基于真实高中模考数据, 构建了一个面向高中的教育测评知识图谱, 其中包括基于 Ontology 的知识图谱模式层构建和依托于模式层结构定义的知识图谱数据层构建; 其次, 结合主流的嵌入式表示翻译模型对构建的 EAKG 进行嵌入式表示学习, 将符号表示的知识嵌入到连续稠密低维向量空间中; 实验表明: 加入模式层信息的 EAKG 在数值向量空间上计算的知识应用如实体链接预测、三元组分类任务上的性能要优于没有模式层结构信息的 EAKG 在该类任务上的表现.

在本文的剩余部分, 我们首先回顾相关工作研究, 在第 3 节中简要介绍知识图谱整体构建流程以及 EAKG 模式层本体构建; 第 4 节介绍 EAKG 数据层构建; 第 5 节主要阐述 EAKG 的嵌入式表示学习以及实验结果; 最后部分关于工作总结以及未来工作展望.

2 相关工作

Ontology 在计算机科学领域的核心意思是一种知识表示模型, 用于描述由对象类型、属性以及关系类型所构成的领域知识库. 斯坦福大学计算机科学家 Tom Gruber 对于计算机科学术语“Ontology”给出了审慎的定义: 一种对于某一概念体系 (概念表达或概念化过程) 的明确表述[9]. 本体是对领域知识的提炼总结, 利

用公理、规则和约束条件来规范实体、关系以及实体的类型和属性等对象之间的关系, 并逐步成为知识图谱模式构建核心[10].

知识图谱, 是结构化的知识库, 目的在于描述真实世界中存在的各种实体或概念以及之间的联系[10]. 其中, 每个实体或概念用一个全局唯一确定的 ID 来标识, 称为它们的标识符 (identifier); 关系 (relation) 用来连接两个实体, 刻画它们之间的关联, 而属性—值对 (attribute-valuepair, 又称 AVP) 则用来刻画实体的内在特性. 知识图谱技术的兴起让海量数据信息已更好的组织形式得到管理, 实现领域知识共享; 其强大的知识推理能力让智能语义搜索、深度问答、社交网络以及垂直领域内的信息挖掘成为可能. 常见的知识图谱如由专家人工创建的 WordNet、Cycorp, 由大众协作编辑创建的 Freebase、WikiData, 基于信息抽取自动创建的 Nell、YAGO、ProBase, 以及垂直领域内的谷歌大脑知识图谱、百度搜索知识图谱、阿里电商知识图谱等.

随着知识图谱研究热潮的兴起和教育智能化的发展, 人们逐渐把目光转向了两者的有机结合上. Xie[11]等人从知识表示, 知识获取和知识推理三个维度对文本和多媒体领域的知识进行阐述; Sun[12]等人通过利用自然语言处理技术提取实体和实体关系构建教育知识图谱, 利用基于事件网络和时间轴的拓扑结构, 构建视觉分析可视化平台 EduVis 来清晰地呈现知识图谱的内部结构, 从而发掘出隐藏在知识图谱中的关于教育舆论和文本语料中的主题信息; London[13,14]等人提出了四种不同的适合于公共教育的学生, 教师和学科的网络图谱表示, 并提出了一些图挖掘技术来获得关于它们的详细信息, 如文章中定义了一个有向加权学生网络, 并结合图挖掘方法获得有别于传统简单的统计分析得来的更为详细的学生成绩和排名信息; 假定两个学生在分数上的相近来刻画学生间的相似性, 通过定义一个无向加权图, 结合社区检测算法将学生进行分组, 从而从所得的学生分组中发掘出组内学生共有的重要信息; Nieto-Isidro[15]在文章中指出评估是对教学过程的质量贡献最大的教育因素之一, 不仅仅在于衡量, 而且也需要一个连续的决策过程, 它必须符合客观性、有效性、可靠性和灵活性的标准, 这些标准的应用可以确保评估过程既衡量教育质量, 又促进教育质量, 从而成为教育系统各个层面的关键要素; 蒋彦[16]等人基于本体构建的数学知识库一定程度上实现了教学领域数学学科知识的共享和基于符号规则推理的知识应用, 但是并没有对学科领域知识层次结构进行刻画, 无法让学生对自己知识掌握情况有系统性地认知, 同时符号表示的知识图谱上的应用有限, 无法实现连续数值向量空间上计算的知识应用.

结合日趋完善的知识图谱技术以及当下提倡的精准化、智能化教育测评的时代背景, 本文提出构建面向高中教育测评知识图谱, 旨在建立领域中各种测评指标、数据以及诸如学校、学生、试卷、试题、知识点等概念对象、实体、属性等之间的关联关系, 实现教育领域知识共享与互联的同时为智能教育测评提供更为广泛的知识应用; 与其他工作不同在于, 本文将知识图谱技术应用于教育测评领域, 学科知识点层次结构的关系建模让学生对知识有了更系统性的认知, 细致全面的学生学习能力的刻画让教学更具有针对性和指导性, 知识图谱的表示学习则为教育测评的智能化提供更广泛的实现基础.

3 EAKG 模式层构建

3.1 EAKG 构建的整体框架与流程

EAKG 的构建首先从外源数据中进行信息抽取, 包括从结构化和半结构化以及非结构化数据中提取诸如学生、知识点、测评指标等概念、属性、关系等信息; 其次, 构建 EAKG 模式层本体结构, 即对抽取的概念、属性、关系等进行明确和形式化表达, 其中包括概念的定义, 属性的定义以及多元关系的定义, 该过程是对教育测评领域知识的提炼总结, 是 EAKG 的模板和核心; 然后, 构建 EAKG 数据层, 数据层的构建是依托于模式层的结构定义完成概念实例对的生成和实例属性值的生成, 本文采用 Jean 框架通过调用 OWL API 实现数据层的半自动化构建, 其中, 实例、属性值是从真实模考数据中经过数据清洗、转换获得, 数据清洗转换操作主要是针对抽取的实例、属性值进行冗余控制、格式转换解析、同步更新、规范化等操作; 接着进行知识存储, 知识存储是知识图谱构建与应用的重要基础, 目前主流的知识图谱存储方式有基于文档数据库、图数据库、关系数据库和分布式存储这几类, 本文对构建的 EAKG 采用基于文档数据库和分布式存储两种方式进行存储, 其中基于文档数据库存储形式是为了便于符号表示的知识推理, 而分布式存储形式则是为了便于在数值向量空间上的显式推理; 最后基于构建的 EAKG 进行知识推理, 知识推理是指从已有数据出发, 通过计算发掘隐含在已有知识中的新

知识, 建立实体间的新关联, 丰富与完善知识图谱的过程, 如对 EAKG 中以符号表示的知识通过调用推理机针对对象属性进行的属性推理或以分布式向量表示的知识通过数值计算得到的诸如实体相似度计算、链接预测等的显式推理; 对推理结果进行人工检查与校验, 去除冗余、矛盾、准确性不高的知识, 调整数据之间的层次、逻辑等结构后最终得到完整的教育测评知识图谱. EAKG 构建整体流程如图 1 所示.

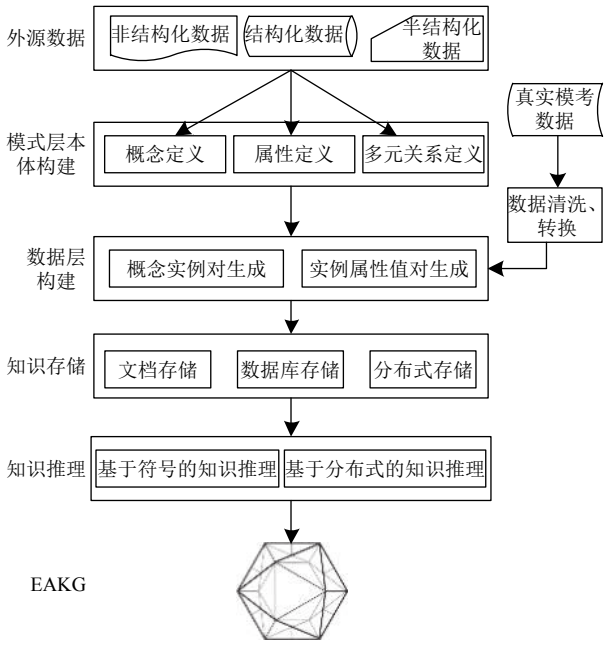


图 1 EAKG 构建整体流程

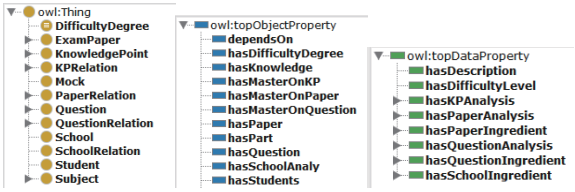
3.2 EAKG 模式层本体构建

知识图谱模式层位于数据层之上, 是知识图谱的核心, 通常采用本体库来管理知识图谱的模式层[10]. 本体是结构化知识库的概念模板, 通过本体库而形成的知识库不仅层次结构较强, 并且冗余程度较小. 本体的构建尚没有一个完全统一或通用的规范标准, 面向领域的知识图谱构建往往需要领域专家的协作或指导, 目前业内主流的构建方法有骨架法、评估法、四步法、七步法等方法[10,17–19].

本文中, 利用本体可视化工具 Protégé[20]来构建教育测评领域的知识本体, 对该领域中涉及到的基本概念、关系、函数、公理等进行明确和形式化地表达. EAKG 模式层本体构建使用的真实数据从全通教学质量监测平台获得, 该平台是在江苏省地区部分中学被实际应用于教学工作活动中, 主要用于对日常教育测评活动产生的数据进行分析统计, 其中包括对学生、学校、测评试卷、试题、知识点、各种测评指标的分析, 本文中所有概念、属性、关系等元素的抽象、定义均受实际教育测评活动的启发并在此基础上经过进一步整理加工而来, 以尽可能的贴近实际情形并服务于实际教育测评活动. EAKG 模式层本体的构建主要包括 3 部分: 概念分类结构、属性定义和多元关系定义.

3.2.1 EAKG 概念分类结构

概念类 (Class) 在本体中被定义为对该领域概念的描述, 是对象实例的集合, 包括概念的名称、与其它概念间关系的集合以及用自然语言对概念的描述. 本文中, 抽象概括的类有: Student(学生类)、Subject(科目类)、ExamPaper(试卷类)、Question(试题类)、KnowledgePoint(知识点类)、School(学校类)、DifficultyDegree(难度等级类)、PaperRelation(试卷关联类)、QuestionRelation(实体关联类)、KPRelation(知识点关联类) 等总共 585 个概念类以及 Ontology 内置的一些类, 如表示数值类型的类: Integer(整形类)、Float(浮点类), 表示文本类型的 String(字符串类) 等; 其中大部分概念类具有父子类层次关系, 如 ExamPaper 下又分为 MathExamPaper(数学试卷类)、ChineseExamPaper(语文试卷类) 等九门学科试卷类, 其部分概念类分类结构如图 2(a) 所示.



(a) 概念类分类结构 (b) 对象属性定义 (c) 数据属性定义

图 2 EAKG 模式层本体结构

3.2.2 EAKG 属性定义

属性定义包括对象属性和数据属性的定义. 对象属性用于描述概念类间关联关系, 本文中定义了如刻画试卷与试题间的包含与从属关系 hasQuestion 及其逆反关系 isQuestionOf、试题与知识点间的包含从属关系 hasKnowledge 及其逆反关系 isKnowledgeOf、学校与学生间的包含与从属关系 hasStudents 及其逆反关系 isBelongTo、知识点间整体与部分关系 hasPart 和 isPartOf、知识点间前后置依赖关系 dependsOn 等 25 种对象属性关系并进行额外的控制约束, 如限制值域定义域等, 部分对象属性关系示意图如图 2(b) 所示. 教育教学过程中, 学科知识点并不是孤立存在的, 而是

在整个教与学过程中都遵循从部分到整体以及一定的先后依赖顺序. 知识点整体部分关系主要用于刻画某个大的知识点包含若干小知识点, 反之, 某些小知识点包含于某个大知识点这种包含与被包含关系; 以高中数学学科为例, 知识点在教学大纲中以章、节、小节组织形式编排, 总共有 535 个知识点, 具体包括 33 大章、140 节、362 小节这样的知识点整体部分结构; 如图 3 所示, 大知识点“集合”包含三个小知识点, 分别为“集合的含义”、“集合的关系”、“集合的运算”, 反之, 这三个小知识点均包含于大知识点“集合”; 知识点间的整体部分关系定义有助于更精准化地测评学生在知识结构上的具体掌握情况, 因此, 在教育测评过程中, 衡量某个学生在知识点“集合”上的掌握情况可以从“集合的含义”、“集合的关系”、“集合的运算”三个小知识点上的掌握情况来进行测评, 帮助更精准化教育测评的实施.

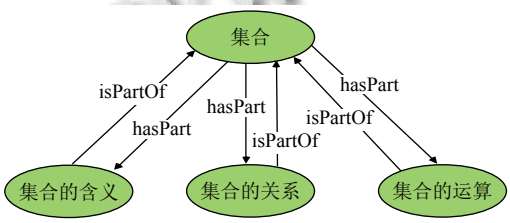


图 3 知识点“集合”整体部分关系示意图

知识点间前后置依赖关系用于描述在整个知识体系教学过程中知识点的教授先后顺序以及学生学习过程中知识点学习的先后顺序. 本文中, 以数学学科为例, 对学科中知识点间前后置依赖关系进行建模, 如图 4 所示, “集合的表示法”这个知识点依赖于“集合的概念”, 用三元组表示为 (集合的表示法,dependsOn,集合的概念), 其含义表示知识点“集合的概念”是知识点“集合的表示法”的前置知识点, 即在高中数学教学过程中教学工作者欲教授知识点“集合的表示法”, 则需先教授知识点“集合的概念”; 对于学生来说则是若欲掌握知识点“集合的表示法”, 则需提前完成知识点“集合的概念”的掌握学习, 换言之掌握知识点“集合的概念”是学好知识点“集合的表示法”的前提基础. 知识点前后置依赖关系的定义主要用于帮助学生分析其经常在某些知识点上失分而找不到具体原因的困境, 从当前知识点溯源找到前置知识点, 往往能从知识点结构上帮助学生定位具体问题所在, 进而帮助教研工作者对学生进行具有针对性的辅导教学.

数据属性用于刻画概念类自身特性, 本文中定义了如描述试卷类的数据属性 hasPaperAnalysis、描述试题类的数据属性 hasQuestionAnalysis、描述知识点类的数据属性 hasKPAnalysis、描述学校类的数据属性 hasSchoolIngredient 等总共 35 个数据属性; 其中大部分数据属性具有父子属性关系, 如数据属性 hasPaperAnalysis 下包含试卷难度、区分度、可信度等子属性, 数据属性 hasKPAnalysis 下包含知识点难易程度、知识点掌握程度、得分情况等以及数据属性 hasSchoolIngredient 下的学校达标率等; 部分数据属性关系示意图如图 2(c) 所示.

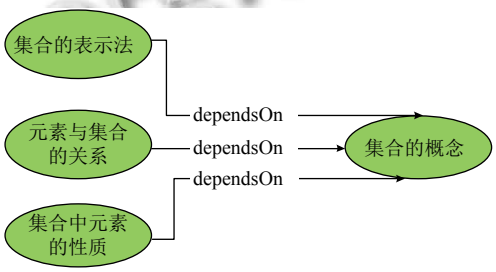


图 4 知识点前后置依赖关系结构图

3.2.3 EAKG 多元关系定义

语义 Web 语言如 RDF 或 OWL 中, 属性是二元关系, 只能用于两个实体间的关联或实体与属性值间的关联; 然而, 在某些情况下, 人们为了以更直观自然的方式表达某个概念时往往会使用关系将概念实体连接到多个概念实体或属性值, 在 W3C 官方文档介绍中称这些关系为 N 元关系 (N-ary relations)[21], 常见的如关系的属性便是一种多元关系, 实际工作中常常需要表达关系的确定性、关系的严重性、关系的强度、关系的相关性等. 针对教育测评领域数据特点, 我们定义了多种多元关系, 如学生与知识点的多元关系、学生与试卷的多元关系、学生与试题的多元关系; 其中以学生与知识点间的多元关系为例, 具体讲解 EAKG 构建过程中对于多元关系的本体模式的构建. 为了实现对学生精准测评, 需要刻画学生在知识点上的掌握情况, 如学生在某次模考中在某个知识点上的掌握程度、得分、排名等; 然而由于三元组 (h,r,t) 只能用于刻画两个实体间的关联关系, 表示范畴局限于实体 h 和实体 t 间的二元关系 r, 为了表示学生与知识点间关系 r 额外的属性, 如对知识点的掌握程度、得分、具体在哪次模考上等属性, 本文采用 W3C 官方网站推荐的处理多元关系的本体模式, 如图 5 所示, 其核心思想在于为了刻画概念类 A 与概念类 B 之间关系 R 的额外属性

(如图 5 中 C 所示), 通过引入空白节点 (blank node) 将连接两类间的二元关系 R 表示为类而不是作为属性来描述类之间的多元关系, 如图 5 中下图所示, 将连接两类间的关系 R 表示为形如 xxRelation 这样的中间类起过渡作用来刻画概念类 A 与概念类 B 之间关系 R 的额外属性 C.

本文中, 受图 5 多元关系结构模型启发, 为了刻画学生类 (Student) 与知识点类 (knowledgePoint) 间的多元关系, 如学生具体在某次模拟考试中在某个知识点上的具体掌握程度、排名等额外属性, 本文将学生类与知识点类间的原有的二元关系表示为中间类 KPRelation, 通过 KPRelation 中间类过渡将学生类 (Student) 关联到模考类 (Mock)、知识点类 (knowledgePoint) 以及整数类 (Integer) 从而实现学生在具体某次模考中, 在具体某个知识点上的具体情况 (如掌握程度、排名等) 的刻画, 其结构示意图如图 6 所示; 同理, 借助于图 5 所示的多元关系本体表示结构, 实现学生类与试卷类的多元关系以及学生类与试题类间的多元关系的刻画.

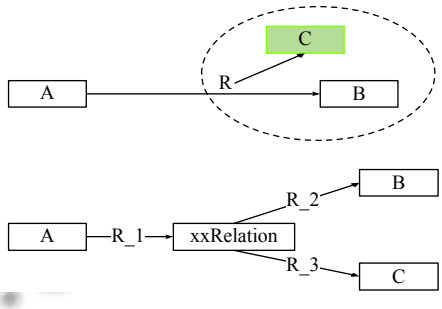


图 5 多元关系 (N-ary relations) 结构模型[21]

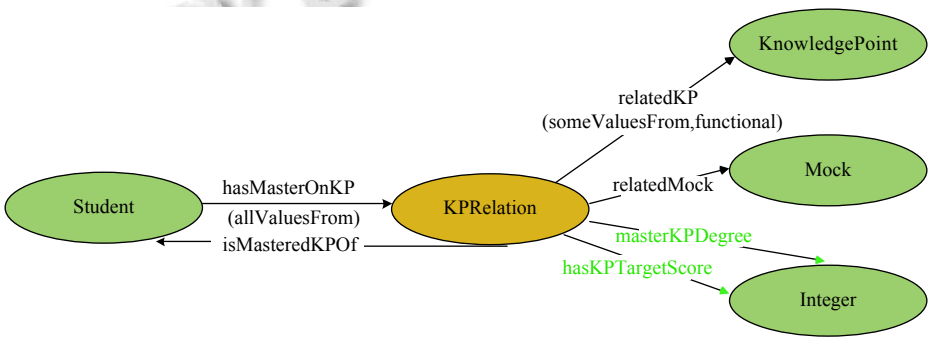


图 6 学生-知识点多元关系结构图

随着概念类、属性、关系、公理等元素的不断加入, 知识图谱模式层结构得到不断迭代更新和完善, 本文构建的 EAKG 模式层结构如图 7 所示.

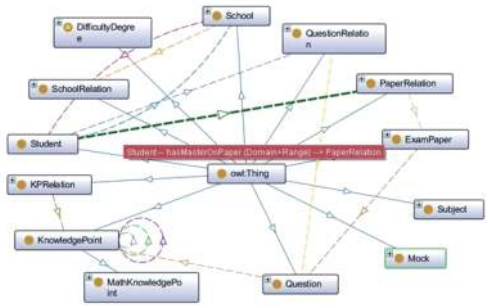


图 7 EAKG 模式层网络结构

4 EAKG 数据层构建

EAKG 数据层构建首先从外源数据 (真实模考数据) 中进行信息抽取, 提取出模式层概念类所需要的实例数据和属性值, 如从结构化数据中提取出学科知识点, 从半结构化和非结构化数据中提取出学生考试得分、排名、学校测评指标等数据信息; 经过数据清洗、转换, 即针对抽取的实例、属性值进行冗余控制、对数值型数据如得分、排名、难度值等进行格式转换解析和规范化操作, 使抽取得到的数据层数据符合模式层结构定义; 然后使用 Jena 框架调用 OWL API 实现概念实例对和实例属性值对的自动化生成, 完成依托于模式层结构定义的知识图谱数据层构建. 其中, Jena 是一个免费开源的 Java 框架, 主要用于构建语义网络和链接数据相关应用, 其强大的数据处理能力为知识图谱数据层面的构建提供了良好的技术支持, 在实体关系数达到百万级别情况下, 其实时响应速度仍然在秒级水平, 为知识图谱构建的快速实现提供了很好的保障. 完成数据层构建的 EAKG 本体度量信息如表 1 所示.

表 1 EAKG 本体度量信息

对象	数目
公理数目	2 248 818
概念数目	585
对象属性数目	25
数据属性数目	35
实例数目	418 109

基于符号表示的知识图谱不仅解决了领域知识互联与共享的难题, 其强大的知识推理能力为知识图谱的更新与完善提供了很好的支撑. 知识推理是基于已有实体、关系等知识出发, 利用推理机, 建立实体间的新关联, 发现新的知识, 丰富与完善知识图谱. 常见的知识推理如子属性关系推理、等价类、等价属性推理、逆反属性推理、属性断言推理等. 本文对于构建的 EAKG 进行知识推理, 由于实体、关系数量较大, 一次性吞吐所消耗的时间比较长, 实时性较差, 借鉴于深度学习过程中处理大数据集的做法, 采用批量处理措施, 对 EAKG 进行分批知识推理再去重整合, 解决了大吞吐量带来的高昂时间消耗问题的同时其实时性达到了分钟级别, 其可行性在基于 3000 多名高三学生真实数据的教育测评知识图谱的构建上得到了验证, 并能够满足实际应用工作需求. 对于推理结果进行人工检查与校验, 由本体开发人员和领域专家共同协作完成, 首先针对 EAKG 模式层结构的定义进行核查, 包括概念的定义、属性定义及其诸如属性的值域、定义域、函数、传递性等额外的约束性条件限制等; 其次, 借助于本体编辑可视化工具 Protégé, 打开推理机先进行一致性检测 (包括语法、语义的一致性检测), 可以直观检测出现有本体中存在的不一致性问题, 如名称 ID 唯一性冲突、类包含关系冗余、属性定义域值域冗余、属性值与类型不匹配等, 发现问题并及时修改; 最后, 对推理结果如知识点间的依赖 (dependsOn) 关系、整体部分 (hasPart-isPartOf) 关系等推理结果进行人工核查, 参考领域专家意见并对不合理结果进行回溯动态调整修改. 而对于结果的准确性, 则有两方面进行保证, 一方面有严格的模式层结构定义, 在领域专家参与指导下对概念、属性、关系以及约束性条件限制等进行严格把关, 尤其在模式层知识点结构定义部分, 一定程度上保证了结构的合理、准确性, 且模式层结构具有一定的动态调整修改的扩展性, 能根据领域专家对推理结果准确性的反馈作适当调整; 另一方面, 本文采用基于描述逻辑的 Pellet 推理机进行知识推理, 其中 Pellet 是基于 Java 使用 Tableaux 算法设计实现的 OWL-DL 推理机, 其对完备性、可判定性支持的特性从底层技术实现上确保了推理结果的准确性. 本文构建的 EAKG 经过知识推理且进行人工检查与校验, 去除冗余、矛盾、准确性不高的知识以及逻辑结构后, 整合得到完整的教育测评知识图谱, 记为 EAKG-withSchema; 另外, 按照传统网页爬虫以及信息自动抽取的方式构建了没有依托于模式层本体结构定义, 单纯由数据构成的教育测评知识图谱, 记为 EAKG-noSchema. 知识图谱主要由一系列的事实组成, 知识以事实为单位进行存储, 一般采用形如 (实体 1, 关系, 实体 2)、(实体、属性, 属性值) 这样的三元组来表达. 对于三元组的表示, 主要有定义法、图表示法和基于 XML 表示法三种方式; 其中基于图表示法的开源工具有 Neo4j、Twitter 的 FlockDB、Sones 的 GraphDB 等; 本文采用定义法对构建的 EAKG 以三元组形式进行表示, 两种 EAKG 详细信息如表 2 所示.

表 2 两种 EAKG 度量对比

EAKG	实体	关系	三元组
EAKG-noSchema	418 198	34	1 824 966
EAKG-withSchema	419 010	60	3 658 949

从表 2 中可以看出, 拥有模式层本体结构信息并进行知识推理后的教育测评知识图谱 EAKG-withSchema 无论在实体关系数还是在网络复杂度上均要高于没有模式层结构支撑单纯由数据事实构成的教育测评知识图谱 EAKG-noSchema.

5 EAKG 的表示学习及应用

知识的表示形式是知识图谱应用的基础, 目前主流的表示形式有基于符号的知识表示和基于分布式的知识表示; 其中基于符号的知识表示有如基于一阶谓词逻辑、基于语义网络等方式, 其特点在于知识推理具有可解释性, 属于隐式推理, 但难以适应大规模知识图谱且存在语义鸿沟, 而基于分布式的知识表示有如基于张量分解、翻译模型和神经网络三种表示方式, 其特点在于知识推理具有可学习、可计算, 适合大规模知识图谱且属于显式推理. 结合构建的 EAKG 规模比较大且具体知识应用往往涉及在连续数值向量空间上的计算如学生试卷成绩预测、学生在知识点上得分预测等的情况, 本文采用基于翻译模型的分布式表示

方式对 EAKG 进行表示学习.

基于翻译模型的分布式表示学习核心思想是将知识图谱中符号化知识嵌入到连续稠密低维向量空间的过程中, 将关系解释为对实体进行操作的翻译, 在保留原始图谱知识结构的同时尽可能减少了语义的丢失并简化了操作, 其代表算法便是由 Bordes 等人首次提出的 TransE[22]算法. 由于其算法的简单有效, 不少研究学者在此算法基础上进行改良优化, 提出了一系列的 Trans 系列算法, 如将关系解释为超平面上的转换操作以保持 1-N/N-1/N-N 关系映射特性的 TransH[23]、加入类别信息进行语义平滑操作的信 SSE[24]以及处理多语义表达问题的 TransG[25]等.

基于最优的嵌入式知识表示学习结果, 指导在连续数值向量空间上计算的知识应用. 诸如可以通过实体链路预测任务来实现学生在知识点上的得分预测, 模考成绩预测, 利用实体的向量化表示来完成实体相似性计算、知识补全、关系挖掘等知识应用.

本文对比单纯由数据没有模式层结构信息构建的 EAKG-noSchema 和依托于模式层结构信息构建并完成知识推理后的 EAKG-withSchema 在多次重复实验当 TransE、TransH 模型训练结果达到最优状态下, 在诸如学生模考成绩预测、知识点分数预测等实体链接预测[22]、三元组分类[26]任务上的表现, 实验结果如下表 3、4 所示, 其中数据集如表 5 所示.

表 3 实体链接预测结果

Metric	EAKG-noSchema						EAKG-withSchema					
	Hits@10(%)		Hits@3(%)		Hits@1(%)		Hits@10(%)		Hits@3(%)		Hits@1(%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
transE(Bordes et al. 2013)	0.50	0.58	0.35	0.42	0.27	0.32	0.69	**0.77**	0.57	**0.70**	0.55	**0.40**
transH(Wang et al.2014)	0.52	0.69	0.44	0.62	0.31	0.48	0.70	**0.80**	0.50	**0.69**	0.31	**0.54**

表 4 三元组分类结果

Metric	EAKG-noSchema	EAKG-withSchema
	Accuracy(%)	Accuracy(%)
TransE(Bordes et al. 2013)	0.743	**0.855**
TransH(Wang et al.2014)	0.783	**0.907**

表 5 实验数据集

EAKG	训练集	验证集	测试集	实体	关系
EAKG-noSchema	1 464 966	180 000	180 000	418 198	34
EAKG-withSchema	3 118 949	360 000	360 000	419 010	60

其中, 表 3 中, Hits@m[27]表示正确三元组排名位于 top m 的数量占所有测试三元组数量的比例, 值越高越好; Raw、Filter[23]表示统计 Hits@m 采取的不同措施, 其中 Filter 表示统计排名前删除出现在训练集、验证集或测试集中的所有其他三元组, Raw 则表示不进行 Filter 中的处理; 表 4 中, Accuracy 指代在三元组分类任务中 (二分类问题), 对于封闭世界假设中任意给定一个三元组 (h,r,t), 对其正确分类的准确率; 表 5 中训练集、验证集、测试集比例为 8:1:1, 其数据相较于嵌入表示学习领域中公开的数据集如 WN18、FB15K 等在量级上均是其好几十倍, 覆盖面比较广, 因此实验结果具有一定的可信度.

从表 3、表 4 实验结果可以看出, EAKG 的嵌入表示学习中, 依托于模式层本体结构定义构建的 EAKG-withSchema 在实体链接预测、三元组分类任务上性能均要大幅优于没有模式层结构信息支持单纯由数据构建的 EAKG-noSchema 在上述任务中的表现, 实验表明: 领域本体结构信息的定义一定程度上提高了知识图谱的嵌入表示学习性能, 在基于连续数值向量空间上计算的知识应用取得了更好的表现, 为进一步的智能教育测评提供了更好的帮助.

6 结论与展望

本文介绍了教育测评知识图谱构建方法, 逻辑上分为基于 Ontology 的知识图谱模式层构建和依托于模式层结构定义的知识图谱数据层构建; 文中以高中数学学科为例讲解了知识点间的层次关系、整体部分关系、前后置依赖关系的定义以及诸如学生与知识点间的多元关系的定义, 构建了层次结构多元、实体对象类型丰富的教育测评知识图谱; 另一方面, EAKG 的嵌入表示学习, 将传统符号表示的知识嵌入到连续稠密低维向量空间去, 实验结果表明: 加入领域本体结构信息构建的知识图谱的分布式表示学习, 在实体链接预测、三元组分类任务上的性能要优于没有模式层信息支撑仅由数据事实构建的知识图谱在上述任务上的表现, 更好地支撑基于连续数值向量空间上计算的知识应用.

本文构建的 EAKG 尽管拥有几十万的节点和几百

万的边, 但依然存在关系不够丰富、数据稀疏等问题, 以及随着实体关系逐渐增多在知识推理、知识表示学习等在时间上的开销所带来的性能问题依然是未来亟待解决的问题.

参考文献

1 黄光扬. 教育测量与评价. 2 版. 上海: 华东师范大学出版社, 2012.

2 林汇波. 考试与评价的学习性分析及问题对策. 教学与管理, 2017, (22): 74–76.

3 全通教学质量监测平台. http://yj.hysbz.com:5005/index.aspx, 2007.

4 教研测. http://www.jiaoxuece.com/, 2016.

5 蔡建东, 马婧. Web2.0 教育应用领域知识图谱研究. 远程教育杂志, 2012, 30(2): 57–62. [doi: 10.3969/j.issn.1672-0008.2012.02.007]

6 赵露, 王林. 利用可视化应用软件进行的中学物理实验教学前沿分析. 教学与管理, 2016, (24): 25–27.

7 汪嘉慧, 王长江. 近 10 年中学物理规律教学疑难问题调查研究——基于知识图谱. 湖南中学物理, 2018, 33(9): 7–11, 35.

8 Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition, 1993, 5(2): 199–220. [doi: 10.1006/knac.1993.1008]

9 Arp R, Smith B, Spear AD. What is an ontology? Arp R, Smith B, Spear AD. Building Ontologies with Basic Formal Ontology. Cambridge: MIT Press, 2015. 248.

10 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述. 计算机研究与发展, 2016, 53(3): 582–600.

11 Xie LX, Wang HX. Learning knowledge bases for text and multimedia. Proceedings of the 22nd ACM international conference on Multimedia. Orlando, FL, USA. 2014. 1235–1236.

12 Sun K, Liu YH, Guo ZC, *et al*. EduVis: Visualization for education knowledge graph based on web data. Proceedings of the 9th International Symposium on Visual Information Communication and Interaction. Dallas, TX, USA. 2016. 138–139.

13 London A, Pelyhe Á, Holló C, *et al*. Applying graph-based data mining concepts to the educational sphere. Proceedings of the 16th International Conference on Computer Systems and Technologies. Dublin, Ireland. 2015. 358–365.

14 London A, Németh T. Student evaluation by graph based data mining of administrational systems of education. Proceedings of the 15th International Conference on Computer Systems and Technologies. Ruse, Bulgaria. 2014. 363–369.

15 Rodríguez-Conde MJ, Olmos-Migueláñez S, Nieto-Isidro S. Evaluation in education and guidance: A perspective from 2016. Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality. Salamanca, Spain. 2016. 135–138.

16 蒋彦. 基于本体的数学知识库的构建及其应用[硕士学位论文]. 成都: 电子科技大学, 2011.

17 栗斌. 面向地理事件的地理本体构建研究[博士学位论文]. 武汉: 武汉大学, 2014.

18 Xu BF, Luo XG, Peng CL, *et al*. Based on ontology: Construction and application of medical knowledge base. Proceedings of 2007 IEEE/ICME International Conference on Complex Medical Engineering. Beijing, China. 2007. 1586–1589.

19 杨玉基, 许斌, 胡家威, 等. 一种准确而高效的领域知识图谱构建方法. 软件学报, 2018, 29(10): 2931–2947.

20 Horridge M, Knublauch H, Rector A, *et al*. A practical guide to building OWL ontologies using the protégé-OWL plugin and CO-ODE tools edition 1.0. Manchester: The University of Manchester, 2004.

21 Noy N, Rector A. Defining N-ary relations on the semantic web. https://www.w3.org/TR/2006/NOTE-swbp-n-aryRelations-20060412/, [2006-04-12].

22 Bordes A, Usunier N, Weston J, *et al*. Translating embeddings for modeling multi-relational data. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 2787–2795.

23 Wang Z, Zhang JW, Feng JL, *et al*. Knowledge graph embedding by translating on hyperplanes. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Quebec City, Canada. 2014. 1112–1119.

24 Guo S, Wang Q, Wang B, *et al*. SSE: Semantically Smooth Embedding for Knowledge Graphs. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(4): 884–897. [doi: 10.1109/TKDE.2016.2638425]

25 Xiao H, Huang ML, Zhu XY. TransG: A generative model for knowledge graph embedding. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. 2016. 2316–2325.

26 Socher R, Chen DQ, Manning CD, *et al*. Reasoning with neural tensor networks for knowledge base completion. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 926–934.

27 Trouillon T, Welbl J, Riedel S, *et al*. Complex embeddings for simple link prediction. Proceedings of the 33rd International Conference on Machine Learning. New York, NY, USA. 2016. 2071–2080.