知识图谱构建相关

在[1]中，作者提出了一种新的知识库系统KV（Knowledge Vault）。KV同时采用从互联网网页与已有数据库（选用的是Freebase）知识抽取的方式，并将两者结合。作者比较了多种抽取方式间的准确率、召回率与AUC分数，并证明了采用融合的方式能得到更高的AUC分数（即更高的召回率与更低的假阳率）。

具体而言，作者主要将一般网页分为四类：1、文本文档（先使用NLP标注，然后分别找到每种谓语在现存数据库中的主宾语对，采用有监督学习它们同时出现的句子的句式）；2、HTML DOM树（包括需要填写表单查询的深层数据库数据。基本同1，在此基础上将主宾语在树上的距离也作为一个特征向量）；3、HTML表格（将表格两列作为主语与宾语，与现有数据库进行模式匹配，以推测两列之间的谓语关系）；4、人工标注（仅选取了一小部分，人工将网页中的谓语映射至数据库的谓语，其余同1）。四种方法中，DOM树获取的知识量最大，可信度和AUC分数也最高。

对于从已有数据库获取知识的先验概率的方式，作者主要采取了两种。1、路径排名算法（PRA，主要进行路径关系的推理）；2、多层感知机（MLP）。

将多种抽取方式合并时，对于每种不同的三元组，每个不同的抽取方法的特征向量包含两项：三元组来源的个数以及这个抽取方法对它的平均打分。最后，作者将三种方式进行比较（网页混合、数据库混合以及全部混合），全部混合拿到了最高的AUC分数。

在此基础上，作者认为还可以从三元组本身的互相约束（刚性或柔性）、三元组成立的时效性等方面进行优化。

在[18]中，作者聚焦影视领域，试图构建中英双语影视知识库。

具体而言，数据源方面，作者选择了较权威的百科网站，如豆瓣、维基百科、百度百科等；知识图谱的本体选择基本复用现有的本体，并进行一定程度上的修改与汉化；知识抽取时采用多种方法混合，如基于关键词过滤、手工构建同义词映射表、基于文本规则等；实体链接时主要采用相似度计算方法，计算多种属性相似度，并同时基于两种不同向量模型TFIDF与Word2Vec，增加准确率；实体匹配主要采用Similarity Flooding算法；跨语言实体匹配主要基于相同IMDB链接、中英文别名信息等。本文为后续的融合多数据源、特定领域的知识图谱构建做了铺垫。

知识图谱应用于计算机教育领域

在[2]中，作者以“程序设计”课程为中心，构建了基于知识图谱的可视化知识点导航系统。论文重点在于全自动地构建关于程序设计知识点的知识图谱。

具体而言，首先利用爬虫从专业教学网站获取结构化数据，并从搜索引擎中爬取非结构化数据。对于结构化数据，作者制定对应的规则，从各HTML标签中获取三元组。对于表格内容，作者将章节小标题作为主语，表格标题作为谓语，表格每一行串联在一起作为宾语形成三元组；对于代码示例，作者将章节小标题作为主语，“举例”作为宾语，所有包含在“<code>”标签中的内容作为宾语。

对于非结构化数据，作者选用依存句法解析解决“长距离依赖”问题。系统先将句子分词并标注，然后使用依存句法解析获取词与词间的关系，并对每个动词寻找其对应主语与宾语。若能成功找到，则将（主，谓，宾）作为一个三元组保存，并寻找对应的定状补进行信息补充；若找不到，则不将这个动词视为一个谓语。

最后，作者使用MVC模式构建了一个可视化的知识点导航系统。作者认为未来还可以继续在属性抽取与知识点混合方面进行优化，并基于知识图谱完成学习路径推荐、学习结果评估等功能的实现。

相比[1]中的构建，[2]中对于非结构化数据的处理更为简单，具有可行性，但对于表格数据的处理比较粗糙，可能会出现较多知识出错与知识冗余的情况。

在[3]中，作者提出将知识图谱引入本科教学中，尤其是在“计算机组成原理”这门课程中。与[2]中类似，作者也是聚焦于某一门计算机专业相关课程，并且“计算机组成原理”这门课程的前后续课程关系复杂，也可以拆分成更小的部分（如处理器、内存）进行前后续课程知识图谱的绘制；同时，作者最后采用调查问卷进行教学结果的评估，这些点值得借鉴。

在[4]中，作者以Python程序设计为例，绘制了该课程部分教学知识点的知识图谱，并提出采用路径搜索算法进行原子知识点间的路径重构，形成知识点间的局部树状知识图谱，以达到学习路径推荐的效果。

在[5]中，作者聚焦数据库课程的线上微课堂，构建了数据库课程知识的知识图谱，并允许用户进行可视化查询，管理员进行知识管理。

具体而言，作者首先利用爬虫完成数据爬取，利用TFIDF进行知识点K-Means聚类以形成知识点间关系的参考，然后构建了与[6]类似的学科知识点本体结构，并最终将三元组存储于Neo4j中。实现系统时，作者采用前后端分离，后端Spring Boot与Neo4j交互，前端使用Ajax请求数据异步更新，并利用D3.js进行知识图谱的可视化，实现架构值得借鉴。最后，作者认为未来可以继续优化聚类算法，并利用Neo4j自身提供的算法优化知识点的推荐功能。

在[8]中，作者聚焦在线教育MOOC课堂，并以计算机网络课程为例，从课程大纲抽取课程知识图谱的基础框架，并从视频字幕中利用新提出的NNGP模型抽取知识点进行补充。这种新模型在对视频字幕进行知识点抽取的性能比现有模型更优。

具体而言，对于课程大纲采用了基于规则的命名实体识别，先进行中文分词与词性标注，然后使用正则表达式匹配得到知识点实体，并根据大纲自带的包含顺序确定知识点间的包含关系；对于视频字幕，先使用科大讯飞工具将音频转换为文字，然后利用Bert-BILSTM-CRF模型进行每个视频中候选知识点的标注，然后使用置信度传播模型计算每个知识点的置信度，并选取0.6为阈值最终确定知识点。实验证明这种新的NNGP模型在知识点的准确率与召回率方面都有显著提升，并且最终能够通过一整套方法获取关于计算机网络课程基本准确的知识图谱。

在[13]中，作者提出了一种从课程幻灯片中提取概念图的方式，并对系统进行实验，收集来自老师与学生的反馈。

具体而言，作者将课程概念的来源限定于课程幻灯片（PPT），并根据其特点制定了自然语言处理的方式（抽取关系三元组，并根据PPT中的层级关系补充缺失的主语或宾语），定义概念间层级关系（主要根据PPT层级排版）和重要性的方式（根据概念所在句子的复杂程度，字体、位置等结构性信息等）。此外，作者重用了之前研究中的方式抽取概念间的关系，并将其制成概念图。接着，作者挑选了一些大学计算机课程PPT进行实验，并收集了一些老师和学生的意见及建议。实验表明，该系统在某些课程（如软件工程和计算机体系结构）上效果很好（这些课程的PPT结构明确，逻辑清晰），但在另一些课程（如程序设计和面向对象程序设计）上效果较差（PPT中包含更多的代码片段等）。同时，大部分老师和学生认为将这种概念图作为扁平化PPT学习的补充是有利的。

在[16]中，作者提出构建数据库课程的知识图谱，并以知识卡片的形式呈现。

具体而言，在进行实体抽取时，作者主要关注“术语”和“人物”两部分。作者采用BIO标注法，利用Jieba进行分词，word2vec进行词向量转换，并使用BILSTM-CRF模型进行实体识别。属性与关系抽取时，使用基于规则的抽取方法，主要关系包括“包含”、“相关”与“研究领域”。同时，作者将构建好的知识图谱以知识卡片的形式呈现，有利于知识的结构化展现，并为老师与学生提供便利，值得借鉴。

在[17]中，作者提出构建C语音课程的知识图谱。

具体而言，作者选取百度百科和一些领域垂直网站作为数据源，分析其网站格式，并使用jsoup爬虫爬取网站内容。实体识别时，使用开源的ICTCLAS系统，将术语添加至字典并进行标注；关系识别时，使用指示词、词条标签与章节目录进行关系的识别与补充；属性识别时，使用百度百科信息框，并处理多属性问题。最后使用Neo4j进行存储。

知识图谱应用于教育领域

在[6]中，作者认为目前的高中教学答疑中存在种种问题，并提出构建基于学科知识图谱的智能问答系统自动回答学生的问题，同时利用大数据算法分析学生的答疑行为，为教师形成反馈。

具体而言，作者先构建了学科知识点的本体结构，实体包含章节、小节与知识点，关系包括顺序、包含和关联。接着进行知识图谱的构建，存储于Neo4j中。进行智能问答时，系统利用RMM进行分词，利用CRF进行关键词（知识点）匹配，利用TFIDF计算问题与知识点间的近似程度，并最后使用Cypher查询得到结果。在大数据阶段，系统采用方差分析、相关分析、因子分析等方式进行分析，并利用预测模型等对学生的行为进行预测，最后将结果可视化地返回给老师或学生。

本文重点在于学科知识点本体结构以及问答系统的构建，并创新地加入了大数据分析系统，具有借鉴价值。作者认为未来的优化方向在于通过深度学习提高系统对于自然语言问题理解的精准度，并扩充学科知识图谱的范围。

在[7]中，作者将学生的回答分数与电商用户的评价打分进行类比，并提出现存的推荐系统虽然可以套用在课堂学习中，但仅关注知识点间的相似性，而缺乏对知识间存在的学习顺序的考虑。作者构建了KG-PKP系统：构建知识点间的知识图谱，将知识点嵌入极坐标系，并构建评估方程。最后，系统在知识点推荐的召回率、准确率等得分方面远超现存的推荐系统。

具体而言，先将知识点置入在极坐标系中，模长被用来表示不同层次的知识点，而相同模长的不同弧度被用来区分同一层次的不同知识点。在实际使用中，使用CBOW计算知识点的条件概率并进行推荐。在构建评估方程时，同时考虑到了正确率、题型、完成时间等因素，并采用0.65作为掌握与否的阈值。低于阈值时，选择推荐相似知识点进行补充学习；高于阈值时，选择推荐学习路径上的后续知识点进行更深入的学习。总而言之，这个系统同时考虑到了知识点的相似性与知识点间的递进关系，因此针对知识点推荐的各项表现都优于现有的推荐系统。

在[9]中，作者试图构建针对基础教育的知识图谱系统K12EduKG，借助命名实体识别技术识别各种来源中的学科知识点，并利用数据挖掘技术找出潜藏的知识点间的前提关系。最后，作者构建了数学的部分学科知识图谱证明系统的可行性。

具体而言，抽取学科知识点时，作者主要采用线性链CRF进行知识点标注；抽取知识点间前提关系时，作者主要采用概率关联规则挖掘技术，设定概率阈值0.7，并寻找两知识点间同时具有“会A即会B”与“不会B即不会A”两项概率都高于阈值的情况，认定B为A的前提知识点。在设定合理参数的情况下，系统的关系推断与专家认定基本一致。

对比[8]，两者在抽取知识点时思路基本一致，[8]的机器学习模型更复杂，可能效果会略有进步；同时，[8]使用其他来源的资源对抽取得到的知识图谱进行了补充，与[9]的侧重点不同。

在[10]中，作者构建课程知识图谱，明确知识点间的先后关系，并将每个知识点的难度、重要性作为客观因素，将学生对知识点的掌握程度及学习知识点的收益作为主观因素，结合计算各知识点的权重，以达到个性化推荐后续学习路径的目的。

具体而言，知识点难度为做对该知识点习题的学生比例；知识点重要性与学生观看视频次数、习题数量与视频时挂钩；对知识点的掌握程度通过学习前后完成习题的正确率计算；学习知识点的收益则通过LSTM模型分析学生的学习习惯（行为顺序）得到。每个知识点的权重都可以通过以上几点计算得到，而系统将按权重从高到低推荐接下来可学习的知识点（已完成所有先决知识点学习的知识点，以及学习过但未掌握的知识点）。最终，实验证明以上参数的使用的确可以做到个性化学习路径推荐。

在[11]中，作者提出了建立以本体为基础的电子教育平台的必要性，并提出了构建电子信息教育本体的一些方法。本文可看作后续一些聚焦教育领域，使用知识图谱作为手段的论文的基础。

在[12]中，作者制作了用于连接几大MOOC网站资源的知识图谱MOOC-KG，并根据其中数据分析了MOOC网站相关特征。

具体而言，作者认为目前MOOC学习存在资源分散，未结构化等问题，给学习者造成困难，因此构建一个连接MOOC课程资源的知识图谱非常有必要。其次，作者详细描述了构建MOOC-KG的过程：知识建模（重用并调整了现有的本体CCSO.owl）；知识抽取（挑选了四个中/英MOOC网站，利用命名实体识别等技术进行实体、属性与关系的抽取）；知识融合（使用传统概率模型计算实体间的相似性，根据不同实体制定不同的融合规则，辅以人工审核）；知识存储（Neo4j）；知识更新（定期更新，众包）。利用构建好的MOOC知识图谱，作者进一步分析了四个MOOC网站间的异同，如课程的提供者、主要学科、主流语言、课程评分等情况。作者对KG系统构建的描述非常详细，之后对于知识图谱的分析应用也很具有现实意义，非常具有参考价值。

在[14]中，作者提出了为教育领域自动构建知识图谱的系统KnowEdu，并以数学学科为例构建了相应的知识图谱进行实践，与领域专家给出的结果进行对比，得出系统具有较高可行性的结论。

具体而言，系统主要分为两部分。在教育概念抽取部分，系统使用神经序列标注算法（一种循环神经网络）解决标注问题，其GRU（门控循环单元）的设计可以看作LSTM（长短期记忆网络）的简化版。同时，作者也给出了使用GRU、LSTM与CRF（条件随机场）三种不同方法进行概念抽取的效果：GRU与LSTM接近，并略高于CRF。在教育关系识别部分，主要采用概率关联规则挖掘技术，基本与[9]一致。最终，作者使用该系统对初中数学进行了部分知识图谱构建，并证明其中的实体和关系与领域专家给出的高度吻合。

在[15]中，作者对知识图谱及其在教育领域的应用做了简单综述，包括对知识图谱的定义，知识图谱发展遇到的需解决的问题，知识图谱嵌入技术及一个知识图谱应用于教育领域的例子。最后，作者提出，知识图谱作为一种工具，需要与实际应用或网站结合来服务于人。

参考文献

[1]Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14). Association for Computing Machinery, New York, NY, USA, 601–610. DOI:https://doi.org/10.1145/2623330.2623623

[2]Peng Zhu, Wei Zhong, and Xianming Yao. Auto-Construction of Course Knowledge Graph based on Course Knowledge [J]. Int J Performability Eng, 2019, 15(8): 2228-2236.

[3]范玉雷,雷艳静,韩姗姗,等. 基于知识图谱的计算机组成原理课程植入与导入式教学[J]. 计算机教育,2020(5):113-117. DOI:10.3969/j.issn.1672-5913.2020.05.029.

[4]鹏 刘, 畅 刘, 隽 尹. 基于知识图谱的在线授课组织研究——以"计算机程序设计语言(Python)"课程为例[J]. 教育研究, 2020, 3(5).

[5]Junhua Bai and Lei Che. 2021. Construction and Application of Database Micro-course Knowledge Graph Based on Neo4j. In The 2nd International Conference on Computing and Data Science (CONF-CDS 2021). Association for Computing Machinery, New York, NY, USA, Article 68, 1–5. DOI:https://doi.org/10.1145/3448734.3450798

[6]Yang, Z., Wang, Y., Gan, J. et al. Design and Research of Intelligent Question-Answering(Q&A) System Based on High School Course Knowledge Graph. Mobile Netw Appl 26, 1884–1890 (2021). <https://doi.org/10.1007/s11036-020-01726-w>

[7]Yakun Lang and Guozhong Wang 2020 J. Personalized knowledge point recommendation system based on course knowledge graph. Phys.: Conf. Ser. 1634 012073

[8]Chao Huang, Quanlong Li, Yuanlong Chen, and Dechen Zhan. 2021. An Effective Method for Constructing Knowledge Graph of Online Course. In 2021 4th International Conference on Big Data and Education (ICBDE 2021). Association for Computing Machinery, New York, NY, USA, 12–18. DOI:https://doi.org/10.1145/3451400.3451403

[9] Penghe Chen, Yu Lu, Vincent W. Zheng, Xiyang Chen, and Xiaoqing Li. 2018. An automatic knowledge graph construction system for K-12 education. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale (L@S '18). Association for Computing Machinery, New York, NY, USA, Article 40, 1–4. DOI:https://doi.org/10.1145/3231644.3231698

[10]Shaoyi Wang, Yan Xu, Quanlong Li, and Yuanlong Chen. 2021. Learning Path Planning Algorithm Based on Learner Behavior Analysis. In 2021 4th International Conference on Big Data and Education (ICBDE 2021). Association for Computing Machinery, New York, NY, USA, 26–33. DOI:https://doi.org/10.1145/3451400.3451405

[11]M. Bucos, B. Dragulescu and M. Veltan, "Designing a semantic web ontology for E-learning in higher education," 2010 9th International Symposium on Electronics and Telecommunications, 2010, pp. 415-418, doi: 10.1109/ISETC.2010.5679298.

[12]F. Dang, J. Tang and S. Li, "MOOC-KG: A MOOC Knowledge Graph for Cross-Platform Online Learning Resources," 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2019, pp. 1-8, doi: 10.1109/ICEIEC.2019.8784572.

[13]Thushari Atapattu, Katrina Falkner, Nickolas Falkner, A comprehensive text analysis of lecture slides to generate concept maps, Computers & Education, Volume 115, 2017, Pages 96-113, ISSN 0360-1315, <https://doi.org/10.1016/j.compedu.2017.08.001>.

[14]P. Chen, Y. Lu, V. W. Zheng, X. Chen and B. Yang, "KnowEdu: A System to Construct Knowledge Graph for Education," in IEEE Access, vol. 6, pp. 31553-31563, 2018, doi: 10.1109/ACCESS.2018.2839607.

[15]Rizun, M. (2019). Knowledge Graph Application in Education: a Literature Review. Acta Universitatis Lodziensis. Folia Oeconomica, 3(342), 7–19. <https://doi.org/10.18778/0208-6018.342.01>

[16]Yuehua Qin. et al. 2020 J. Research and Application of Knowledge Graph in Teaching: Take the database course as an example. Phys.: Conf. Ser. 1607 012127

[17]贾丙静,葛华,李德胜. "慕课"时代C语言知识图谱的构建[J]. 绵阳师范学院学报,2018,37(11):100-103. DOI:10.16276/j.cnki.cn51-1670/g.2018.11.019.

[18]王巍巍,王志刚,潘亮铭,等. 双语影视知识图谱的构建研究[J]. 北京大学学报（自然科学版）,2016,52(1):25-34. DOI:10.13209/j.0479-8023.2016.022.