软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):2365−2391 [doi: 10.13328/j.cnki.jos.006808]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向关系型数据与知识图谱的数据集成技术综述*

高云君[1], 葛丛丛[2], 郭宇翔[1], 陈 璐[1]

[1](浙江大学 计算机科学与技术学院, 浙江 杭州 310027)

[2](华为云计算公司 数据智能创新 Lab, 浙江 杭州 310052)

通信作者: 高云君, E-mail: gaoyj@zju.edu.cn

摘 要: 目前, 各个国家和地区均已将大数据视为重要的战略资源. 然而, 大数据时代普遍存在数据流通困难、数据监管不足等问题, 致使数据孤岛现象严重, 数据质量低下, 数据要素潜能难以释放. 这驱使研究人员探索数据集成技术, 以打破数据壁垒、实现信息共享、提升数据质量, 进而激活数据要素潜能. 关系型数据和知识图谱作为两种至关重要的数据组织与存储形式, 在现实生活中应用广泛. 为此, 聚焦关系型数据和知识图谱, 归纳总结并分析实体解析、数据融合、数据清洗 3 方面的数据集成关键技术, 最后展望未来研究方向与趋势.

关键词: 关系型数据; 知识图谱; 数据集成

中图法分类号: TP311

中文引用格式: 高云君, 葛丛丛, 郭宇翔, 陈璐. 面向关系型数据与知识图谱的数据集成技术综述. 软件学报, 2023, 34(5): 2365–2391. http://www.jos.org.cn/1000-9825/6808.htm

英文引用格式: Gao YJ, Ge CC, Guo YX, Chen L. Survey on Data Integration Technologies for Relational Data and Knowledge Graph. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2365–2391 (in Chinese). http://www.jos.org.cn/1000-9825/6808.htm

Survey on Data Integration Technologies for Relational Data and Knowledge Graph

GAO Yun-Jun[1], GE Cong-Cong[2], GUO Yu-Xiang[1], CHEN Lu[1]

[1](College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

[2](Data Intelligence Innovation Lab, Huawei Cloud Computing Technologies Co. Ltd., Hangzhou 310052, China)

Abstract: Recently, big data is considered a critical strategic resource by many countries and regions. However, difficult data circulation and insufficient data regulation commonly exist in the big data era, thereby leading to the serious phenomenon of data silos, poor data quality, and difficulty in unleashing the potential of data elements. This provokes researchers to explore data integration techniques for breaking data barriers, enabling data sharing, improving data quality, and activating the potential of data elements. Relational data and knowledge graphs, as two significant forms of data organization and storage, have been widely applied in real life. To this end, this study focuses on relational data and knowledge graphs to summarize and analyze the key technologies of data integration, including entity resolution, data fusion, and data cleaning. Finally, it prospects future research directions.

Key words: relational data; knowledge graph (KG); data integration

随着物联网、社交媒体、电子医疗等技术的高速发展, 全球数据呈现爆炸式增长的态势. 根据国际数据公司(International Data Corporation, IDC) 统计, 到 2025 年全球数据量预计将达 175 ZB, 表明人类社会已进入大数据时代[1]. 近年来, 各个国家和地区已陆续将大数据上升至战略层面. 例如, 2015 年我国在十八届五中全会上首次提出“国家大数据战略”, 同年国务院印发《促进大数据发展行动纲要》, 以推进我国大数据发展进程, 加速数据强国建设. 此外, 美国实施的《大数据研究和发展计划》、英国发布的《英国数据能力发展战略规划》以及欧盟力推的《数据价值链战略计划》等均已显示出布局大数据战略的迫切性. 可以说, 大数据正在改变全球社会的发展动力

* 基金项目: 国家重点研发计划 (2021YFC3300300, 2021YFC3300303); 国家自然科学基金 (62025206, 61972338, 62102351)
收稿时间: 2022-06-21; 修改时间: 2022-08-18; 采用时间: 2022-09-11; jos 在线出版时间: 2022-12-30
CNKI 网络首发时间: 2023-03-17

与发展方式, 重塑世界格局[2].

然而, 大数据时代普遍存在数据流通困难、数据监管不足等问题, 数据孤岛现象严重、数据质量低下, 进而导致数据要素潜能难以释放. 2020 年《中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见》指出: 要加快培育数据要素市场, 推进政府数据开放共享, 加强数据资源整合, 提高数据质量和规范性 (详见 http://www.gov.cn/zhengce/2020-04/09/content_5500622.htm). 因此, 各行业各领域对于数据集成的需求日益迫切. 数据集成的最终目标是为驻留在不同数据源中的异构数据提供统一访问渠道, 它是打破数据壁垒, 实现信息共享, 提升数据质量的重要手段. 同时, 也为下游的各类数据驱动应用提供可靠的数据基础.

数据集成的概念广泛, 包括实体解析、数据融合数据清洗、关系解析、语义消歧等技术. 专家学者们对于实体解析、数据融合以及数据清洗技术的关注度日益增加, 实体解析、数据融合以及数据清洗已成为数据集成领域的关键研究方向. 实体解析是实现数据集成的先决条件, 旨在关联不同来源中指向同一实体的数据实例. 在执行完实体解析后, 需要将已关联的不同来源的数据集成至统一的数据库中, 使得数据内容更丰富, 从而发现新的价值信息. 然而, 由于不同数据集的异构性、信息不完整、数据错误或数据过时等问题, 可能在数据集成过程中发生冲突. 因此, 需要通过数据融合以解决来自不同数据源的同一实体在集成过程中产生的冲突问题, 从而保证数据的正确性与一致性, 提升数据价值. 此外, 不同来源的数据本身以及数据集成过程中很可能产生数据质量问题. 所以, 数据清洗是贯穿整个数据集成过程的关键技术, 旨在检测并修复脏数据, 以确保数据集成的有效性.

尽管目前已有若干关于数据集成的综述性文献, 但现有的综述性文献侧重于 (1) 描述数据集成的框架概念[3,4]、发展脉络[5]; 或是 (2) 对数据集成中的某一关键技术 (譬如实体解析[6−8]、数据融合[9]、数据清洗[10,11]等) 进行综述, 尚缺乏对数据集成中各项关键技术研究现状的全面探讨与分析. 此外, 随着 5G 和物联网等技术的飞速发展, 网络数据内容呈现爆炸式增长的态势. 由于互联网内容的大规模、异质多元、组织结构松散等特点, 为人们有效地获取信息和知识提出了巨大挑战. 不同于传统的关系型数据, 知识图谱 (knowledge graph, KG)[12]以其强大的语义处理能力和开放组织能力, 已成为一种流行的数据组织形式. 近年来, 工业界和学术界都致力于构建大规模知识图谱. 然而, 尽管这些知识图谱的规模较大 (存储了真实世界中的数百万条事实), 但仍然是高度不完整的. 例如, 开源知识库 Freebase 中 71% 的人没有对应的出生地, 75% 的人没有对应的国籍信息. 此外, 对于一些不常见的事实描述可能更不完整. 因此, 数据集成所关注的数据类型已不仅局限于传统的关系型数据, 知识图谱亦是数据集成所需应对的关键数据类型. 此外, 亦有一些研究工作涉及面向半结构化数据 (JSON、XML 等)、非结构化数据 (多媒体数据) 的数据集成问题[13,14], 然而此类工作仍处于起步阶段, 尚未形成完整的体系.

鉴于此, 本文从关系型数据和知识图谱两种关键数据类型出发, 归纳总结并分析实体解析、数据融合、数据清洗 3 方面的数据集成关键技术 (如后文图 1 所示), 最后展望未来研究方向与趋势.

1 实体解析

实体解析是数据库、信息检索、机器学习、自然语言处理等领域的研究重点. 近年来, 专家学者对于实体解析的关注度日益提升, 已提出了许多面向不同数据类型 (包括知识图谱[7]、关系型数据[15,16]、文本数据[17]、图像数据[18]等) 的实体解析技术. 本节聚焦面向关系型数据与知识图谱的实体解析技术, 下面分别对这两部分工作予以阐述和分析.

1.1 面向关系型数据的实体解析

概念与定义: 在现实生活中, 大量的数据被存储为关系型数据. 然而, 这些数据通常分散在彼此孤立的数据库中, 从而导致数据孤岛, 阻碍数据的关联与共享[19]. 关系型数据实体解析长期以来是学术界和工业界所共同关注的研究热点[20,21], 其旨在识别来自两个不同来源的元组是否指向真实世界中的同一对象 (或称两者为正确匹配项), 以打破数据孤岛, 实现跨源数据之间的关联互通, 从而为数据集成奠定基础.

令 D 代表由$|D|$个元组和 m 个属性构成的关系型数据集, $A=\{A[1], A[2], \ldots, A[m]\}$为 D 的属性集合. 每个元组 $e \in D$ 均由一系列属性值组成, 记为 $V=\{e.A[1], e.A[2], \ldots, e.A[m]\}$. 这里的 $e.A[m]$ 表示元组 e 在第 m 个属性 (即 $A[m]$) 上的属性值. 给定两个数据集 D 和 D', 关系型数据实体解析任务旨在为每个元组对 $(e,e') \in D \times D'$ 赋予一个

二元标签 $y \in \{0,1\}$. 其中, $y=1$ 代表 e 和 e' 为正确匹配项; $y=0$ 则表明 e 和 e' 指向真实世界中的不同对象.

图 2 给出了一个面向关系型数据的实体解析示例. 图 2(a) 为一个关于信用卡信息的关系型数据集 D, 图 2(b) 为用户购买商品的支付记录数据集 D'. 图中能够正确匹配的元组对以“√”标识, 不能匹配的元组对以“×”标识. 在图 2 中, 元组对 (e_1, e'_1) 和 (e_2, e'_2) 为正确匹配项, 而元组 e_3 和 e'_3 指向真实世界中的不同对象.

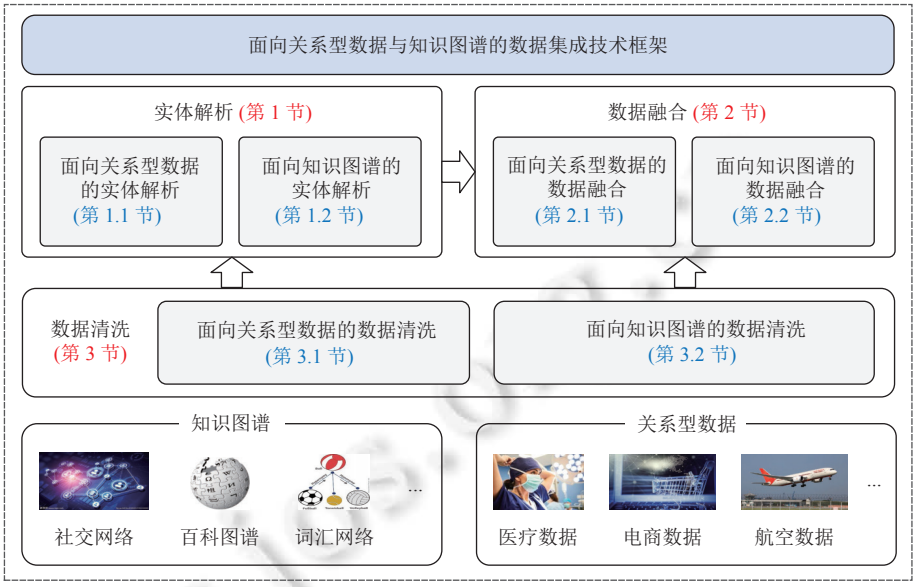


图 1 面向关系型数据与知识图谱的数据集成技术框架图

	用户识别码	姓名	性别	联系地址	电话	邮箱
e_1	10012301	张三	男	杭州市上城区华池北路 1 号	0571-1111111	zhangsan@cc.com
e_2	10012302	李四	女	杭州市西湖区浙大路 38 号	0571-2222222	lisi@dd.com
e_3	10012303	王五	男	杭州市西湖区天目山路 148 号	0571-3333333	wangwu@ee.com

(a) 关于信用卡信息的关系型数据集 D

	订单号	姓名	联系地址	联系方式	邮箱	支付金额
e_1'	001	张三	杭州市上城区华池北路 1 号	1111111	zhangsan@cc.com	180
e_2'	002	李四	杭州市西湖区浙大路 38 号	0571-2222222	lisi@dd.com	759
e_3'	003	王二	杭州市西湖区文三路 188 号	4444444	mc@gm.com	312

(b) 关于支付记录的关系型数据集 D'

图 2 关系型数据实体解析示例

评价指标: 评估关系型数据实体解析结果质量的指标主要包括精确率、召回率和 $F1$ 分数. 精确率为预测所得的真正匹配元组对占所有预测为匹配元组对的比例; 召回率为预测所得的真正匹配元组对数量占真值 (即真正匹配元组对的总数) 的比例; $F1$ 分数为精确率和召回率的调和平均值, 如公式 (1) 所示:

$$F1\text{分数}=2\times\frac{\text{精确率}\times\text{召回率}}{\text{精确率}+\text{召回率}} \tag{1}$$

研究现状及分析: 早期, 专家学者根据规则[22−26]、众包技术[27−31]以及传统机器学习模型[32−38]对关系型数据进行实体解析. 基于规则的方法依靠人工提供的声明性匹配规则[23]或程序合成的匹配规则[26]来查找匹配的元组对. 基于众包的方法需要众包工人人工识别两个元组是否指向现实世界中的同一对象. 基于传统机器学习的方法利用高斯混合模型[35]、支持向量机 (SVM)[36]、朴素贝叶斯[37]等技术进行实体解析. 随着大数据时代的来临, 数据规模不断增大且数据来源变化多样, 为前述方法带来了巨大挑战.

近年来, 随着深度学习技术在机器学习领域的快速发展, 大量基于深度学习的关系型数据实体解析方法已被提出, 并取得了可喜的性能. 根据深度学习方法对于关系型数据特征的建模方式, 可将现有的主流方法分为两类: (1) 基于序列特征的关系型数据实体解析方法; (2) 基于图结构的关系型数据实体解析方法.

基于序列特征的关系型数据实体解析方法将元组视为序列形式, 并采用 RNN、基于 Transformer 的预训练语言模型 (以下简称预训练语言模型) 等序列特征模型对元组特征进行学习. Ebraheem 等人[39]提出了 DeepER, 该方法联合长短期记忆网络 (LSTM)[40]与 GloVe[41]词嵌入向量共同训练实体解析模型. Mudgal 等人[42]提出的 DeepMatcher 为基于序列特征的深度实体解析模型定义了一个解决方案空间, 并探讨了聚合操作、RNN 模型和注意力机制 (attention) 对于实体解析训练的作用. Nie 等人[43]提出了 Seq2SeqMatcher, 该方法将元组建模为令牌级别的序列形式以捕获令牌之间的语义相关性, 并基于令牌语义相关性对实体解析关系进行预测. Fu 等人提出的 MPM[44]则关注于通过基于 RNN 的深度相似性度量方法为不同属性选择合适的度量准则, 从而提升匹配性能. Zhang 等人[45]提出了 MCA, 该方法采用多上下文注意力机制 (包括自注意力、配对注意力以及全局注意力) 来丰富元组的序列特征. Kasai 等人[46]提出了 DTAL, 其通过迁移学习和主动学习方法, 以提升序列特征模型在低资源场景下的实体解析效果. 尽管上述工作在一定程度上展示了基于序列特征的模型在关系型数据实体解析任务上的有效性, 但它们只是简单地将 RNN 与注意力机制或聚合操作相结合, 尚未充分挖掘关系型数据的内在特征. 得益于预训练语言模型的强大表征能力, 随后的研究工作倾向于将预训练语言模型融入实体解析任务中, 以进一步提升匹配性能. Zhao 等人[47]提出了一种基于迁移学习的深度实体解析方法 Auto-EM, 其通过在实体类型检测的辅助任务上对深度实体解析模型进行预训练以提升模型的匹配性能, 并减少甚至消除模型对于匹配标签的依赖. Li 等人[15]提出了 DITTO, 该方法利用预训练语言模型与多种数据增强策略的联合作用, 以实现高质量的实体解析结果. 进一步地, Li 等人[48]指出: 虽然预训练语言模型能为关系型数据实体解析带来最佳的性能, 但是其模型庞大、参数众多, 易导致计算代价随数据规模的变大而急剧增加. 鉴于此, Li 等人提出了 BertER[48], 其通过孪生网络结构加速匹配模型训练的元组交互过程, 从而提高实体解析效率.

然而, 由于不同属性之间具有序列无关性, 故将元组视为序列形式可能丢失部分数据特征[19]. 例如, 相隔较远的相关属性间的依赖关系无法得到充分的表征. 另外, 基于序列的方法将每个元组视为一个独立的序列, 因此无法捕获不同元组之间的复杂关系. 为此, 近期已有一些专家学者提出了基于图结构的关系型数据实体解析方法, 通过将关系型数据转换为图结构的形式, 以捕获不同属性之间与不同元组之间更丰富的语义关系. Cappuzzo 等人指出: 尽管预训练语言模型在一些通用领域能够发挥其强大的语义表征能力, 但难以在具有自定义词汇表的企业数据集中发挥其应有的性能[19]. 为此, Cappuzzo 等人[19]提出了 EMBDI, 该方法先为关系型数据构建一个紧凑的图结构, 而后通过随机游走模型从图结构中学习数据 (包括元组和属性) 之间的相似性, 并基于相似性度量预测元组之间的匹配关系. Li 等人[49]提出了一种基于图卷积神经网络的深度实体解析模型 GraphER, 该模型将元组转化为令牌级别的节点, 并通过连边将具有关联性的节点进行连接, 从而实现图构建. 虽然图结构能够表达关系型数据的丰富内在特征, 但是图结构往往具有错误敏感性[50], 真实关系型数据集中的脏数据容易误导基于图结构的数据特征学习, 从而导致匹配效果不佳.

此外, 尽管基于深度学习的关系型数据实体解析方法能够在大量训练标签 (或称监督信号) 的支持下获得高质量的匹配结果, 但获取实体解析标签需要付出高昂的成本, 这在现实生活中并非易事. 所以, 如何减少实体解析任务对标签的强依赖性成为当前研究通用实体解析技术的一个关键问题[47]. 虽然 Auto-EM[47]和 EMBDI[19]在无监督方面进行了一些尝试, 但是其各自存在一定缺陷. 如前所述, 基于序列特征的 Auto-EM 难以挖掘充足的数据特征, 基于图结构特征的 EMBDI 则易受脏数据的干扰, 因而两者在匹配性能上均存在一定的限制. 另外, 部分专家学者也探索了一些无监督的非深度学习实体解析方法. 例如, Li 等人[51]提出了一种新颖的无监督模糊相似连接方法 Auto-FuzzyJoin, 其能够以极少的参数量实现良好的匹配性能. Zhang 等人[52]提出的 ITER+CliqueRank 则通过构建加权二分图与迭代相似性估计算法实现无需人工参与的关系型数据实体解析. Ge 等人[53]提出了一种基于特征协同的关系型数据实体解析方法 CollaborEM. 该方法设计了一种无需人工参与的自动标注策略, 其通过自动挖掘元组之间的语义相似性以生成高质量的匹配标签. 表 1 总结了当前基于深度学习的关系型数据实体解析方法.

表 1 基于深度学习的关系型数据实体解析方法比较

深度学习方法	方法名称	有监督信号	无监督信号
序列特征模型	DeepER[39]	√	×
	DeepMatcher[42]	√	×
	Seq2SeqMatcher[43]	√	×
	MPM[43]	√	×
	MCA[45]	√	×
	DTAL[46]	√	×
	Auto-EM[47]	×	√
	DITTO[15]	√	×
	BertER[48]	√	×
图结构特征模型	EMBDI[19]	×	√
	GraphER[49]	√	×

总的来说, 尽管目前已有许多面向关系型数据的实体解析方法, 但现有工作尚未找到合适的建模方式以充分表达关系型数据的内在特征, 致使匹配精度受限. 此外, 仍需进一步降低关系型数据实体解析对人工标注训练数据的依赖, 以降低人力成本从而应对真实世界中更广泛的实体解析场景.

1.2 面向知识图谱的实体解析

概念与定义: 知识图谱由一系列三元组 (或称事实) 所构成, 其中每个三元组包含两个实体以及连接它们的关系. 知识图谱作为一种被广泛使用的知识表现形式, 能够以一种便于机器存储、识别和理解的方式对数据进行有效的组织与管理. 在现实生活中, 不同来源的知识图谱具有异构性和不完整性的特点, 因而需要关联共享来自不同来源或不同语言的异构知识, 以扩大知识规模、丰富知识内容, 从而实现知识集成. 面向知识图谱的实体解析是知识集成的先决条件, 旨在关联不同来源知识图谱中指向真实世界同一对象的等价/匹配实体. 长期以来, 专家学者一直致力于探索各类知识图谱实体解析技术.

令 $\mathcal{G}=(\mathcal{E},\mathcal{R},\mathcal{T})$ 代表一个知识图谱, 其中 $\mathcal{E}$ 表示实体集合, $\mathcal{R}$ 表示实体之间的关系集合, $\mathcal{T}$ 表示由实体与关系构成的三元组集合. 给定两个实体 e_i, $e_j\in\mathcal{E}$ 和一条由实体 e_i 指向实体 e_j 的关系 $r\in\mathcal{R}$, 其对应的三元组表示为 $t=(e_i, r, e_j)\in\mathcal{T}$. 通常, 面向知识图谱的实体解析任务旨在找到从源知识图谱 $G_s=(\mathcal{E}_s,\mathcal{R}_s,\mathcal{T}_s)$ 到目标知识图谱 $G_t=(\mathcal{E}_t,\mathcal{R}_t,\mathcal{T}_t)$ 中每个实体的一一匹配关系 (表示为矩阵形式 $\mathbf{P}$), 如公式 (2) 所示:

$$\mathbf{P}=\{0,1\}^{|\mathcal{E}_s|\times|\mathcal{E}_t|},\text{w.l.o.g. }|\mathcal{E}_s|\leqslant|\mathcal{E}_t| \tag{2}$$

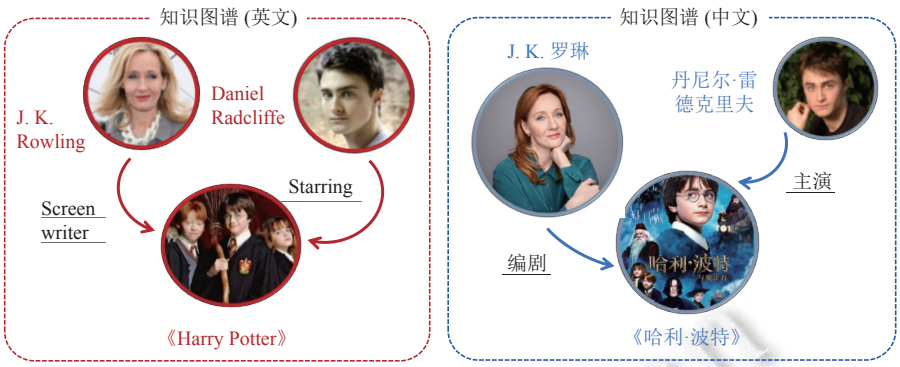
对于任意实体 $e_s^i\in\mathcal{E}_s$ 和 $e_t^j\in\mathcal{E}_t$, 当且仅当 e_s^i 和 e_t^j 匹配时 (表示为 $e_s^i\equiv e_t^j$), 匹配矩阵 $\mathbf{P}$ 中的对应元素 P_{ij}=1. 否则, P_{ij}=0, 表示 e_s^i 和 e_t^j 不匹配.

图 3 给出了一个面向知识图谱的实体解析示例. 图 3(a) 为一个英文知识图谱样例 KG_{EN}, 图 3(b) 为一个中文知识图谱样例 KG_{ZH}. KG_{EN} 和 KG_{ZH} 各自包含 3 个实体, 每个实体之间存在一一匹配关系, 分别为: (1) J. K. Rowling ≡ J. K. 罗琳; (2)《Harry Potter》≡《哈利·波特》; (3) Daniel Radcliffe ≡ 丹尼尔·雷德克里夫. 对于图 2 而言, 面向知识图谱的实体解析任务旨在找到上述一一匹配关系.

评价指标: 知识图谱实体解析的常用评估指标为 Hits@N ($N\geqslant1$) 和平均倒数秩 (mean reciprocal rank, *MRR*). Hits@1 (H@1) 表示能够正确匹配的实体占真值 (即真正匹配实体对的数量) 的比例. Hits@N ($N>1$) 表示在相似度排名前 N 的实体对中正确匹配项的比例. *MRR* 可以通过计算所有正确匹配项的相似度倒数排名的均值得到, 如公式 (3) 所示:

$$MRR=\frac{1}{N}\sum_{i=1}^{|N|}\frac{1}{rank_i} \tag{3}$$

其中, $rank_i$ 表示第 i 个正确匹配实体对的相似度排名. Hits@N 和 *MRR* 的值越高, 则实体解析的精度越高.



(a) 英文知识图谱样例 (b) 中文知识图谱样例

图 3 知识图谱实体解析示例

研究现状及分析: 早期的知识图谱实体解析主要利用人工特征提取[54]、本体等价推理[55]、众包技术[56]以及相似度计算[57]等方法, 但这些方法难以有效地匹配异构知识图谱, 限制了其实际应用范围. 近年来, 随着表示学习 (representation learning) 的迅速发展, 涌现了大量基于表示学习的知识图谱实体解析方法. 此类方法根据知识图谱中实体与关系所蕴含的语义或结构信息, 将实体与关系均表示为低维嵌入向量的形式, 以解决早期研究工作难以有效地表征异构知识图谱的问题. 最后, 此类方法通过度量这些嵌入向量之间的相似性以找到实体之间的匹配/等价关系.

根据知识图谱内在特征的建模方式, 基于表示学习的知识图谱实体解析方法主要分为 5 类, 即基于三元组表征模型 (譬如 TransE[58]和 ComplEx[59]) 的方法、基于图神经网络 (GNN) 的方法、基于概率模型的方法、基于循环神经网络 (RNN) 的方法和基于 Transformer 模型的方法.

基于三元组表征模型的方法利用知识图谱的三元组表征模型学习实体和关系的嵌入向量. 其中, Chen 等人[60]提出了 MTransE, 首次将 TransE 这一经典的三元组表征模型应用于知识图谱实体解析任务. 随后, IPTransE[61]和 BootEA[62]将迭代优化策略与三元组表征模型相结合以实现可喜的知识图谱实体解析性能. 进一步地, Pei 等人[63]指出: 知识图谱中实体的度 (degree) 差异会对具有不同度数的实体造成不同的匹配难度, 并基于此提出了一种基于度感知的知识图谱实体解析方法 SEA. 该方法通过结合三元组表征模型与对抗训练方法, 使其能够在了解度差异的情况下优化实体嵌入向量, 从而提升匹配性能.

然而, 三元组表征模型主要依赖于三元组的局部语义信息, 缺乏对知识图谱全局语义信息的有效捕获. 为此, 许多的研究人员探索了基于 GNN 的知识图谱实体解析方法, 其通过聚合实体的邻居信息来学习实体的嵌入向量, 以捕获更丰富的知识图谱结构特征. Wang 等人[64]首次将图卷积网络 (GCN) 引入知识图谱实体解析问题, 并通过大量实验验证了图卷积网络相较于平移模型的实体解析性能优势. 随后, Cao 等人[65]提出了 MuGNN, 该方法将图注意力机制 (GAT) 作为实体结构特征挖掘的核心模型. 不同于 GCN 将实体的所有邻居信息均一视同仁的做法, 图注意力机制为实体的不同邻居赋予不同的权重, 以表示各邻居对于该实体的重要程度, 从而改进匹配性能. 为进一步提升知识图谱实体解析性能并缩减模型规模, Sun 等人[66]提出了 HyperKA, 该方法首次引入了基于层次结构的双曲关系图神经网络 (hyperbolic relational GNN), 其能够以更低维的实体嵌入向量实现卓越的匹配性能. 此外, Pei 等人[67]提出了 REA, 以解决人工标注产生的噪声匹配标签问题, 使得实体解析结果具有一定的鲁棒性.

另外一些方法基于概率模型、循环神经网络或 Transformer 模型实现实体解析. 基于概率模型的方法[68,69]通过结合概率模型与知识图谱的嵌入向量以指导模型训练. 基于循环神经网络的方法[70–72]旨在利用 RNN 捕获实体之间的长期依赖关系挖掘丰富的实体特征. 基于 Transformer 模型的方法[73–75]通过 Transformer 强大的语义表征能力[76]捕获知识图谱中的实体关系、路径以及邻域上下文关系.

在现有基于表示学习的知识图谱实体解析方法中, 大量研究工作完全依靠数据的内在特征进行实体解析. 然而, 若干研究工作[77–79]指出, 知识图谱数据的内在特征不足, 难以在真实场景中产生可观的匹配结果. 鉴于此, 大批专家学者试图将辅助信息 (譬如实体属性[80]、实体描述[70]、实体名称[77]、实体图像[81]、文本语料[82]等) 与各类

基于表征学习的知识图谱实体解析方法相结合, 其旨在通过辅助信息中的额外信息增强实体特征, 以弥补图结构特征不足的问题, 从而提升匹配效果. Liu 等人[81]提出了 EVA, 该方法首次将实体的图像信息运用于实体解析任务. Chen 等人[83]提出了 JEANS, 该方法将知识图谱与文本语料共同映射至同一向量空间, 并通过迭代方法不断地调整匹配特征. Yang 等人[73]提出了 HMAN, 该方法融合 GCN 模型和基于 Transformer 的预训练语言模型 BERT[84], 其中 GCN 用于获取实体属性与名称特征, BERT 则用于挖掘实体描述中包含的语义信息. Tang 等人[74]进一步探索了 BERT 与实体解析任务的结合方式, 提出了 BERT-INT 方法. 相较于 HMAN, BERT-INT 摆脱了对于图结构特征的依赖, 并实现了更优的匹配结果. 虽然大量的研究工作验证了辅助信息对于知识图谱实体解析任务的促进作用, 但是并非所有的实体都能够找到其所对应的辅助信息. 此外, 辅助信息的质量与获取代价同样是需要进一步考虑的问题, 前者 (例如, 实体属性中存在噪声[85]) 会影响实体解析结果的正确性, 后者通常需要付出高昂的人力成本以获得充足的辅助信息, 从而确保高质量的匹配结果.

知识图谱实体解析的结果质量亦与匹配标签数量高度相关. 大多数的现有方法为有监督方法, 其需预先提供大量的匹配标签以指导模型训练. 然而, 获取充足的正确匹配标签往往需要高昂的人力成本, 这在真实场景下是不切实际的. 为此, 已有部分研究工作在如何减少人力标注成本方面进行了一定的探索, 例如, 半监督方法[61,62,70,83]、特征驱动的标签生成方法[81,86]和无监督方法[69,87].

基于半监督的知识图谱实体解析方法旨在将无标签数据与有标签数据一同送入模型进行训练, 包括自训练 (self-training)[61,62,83,88]和协同训练 (co-training)[70]. 前者根据有标签数据所蕴含的特征迭代地为无标签数据进行自动标注, 以增加匹配标签数量. 后者通过两个不同模型对不相交的实体进行特征学习, 并通过彼此之间的协同作用增强各自的匹配特征. 然而, 半监督方法仍需提供少量标签, 因而部分研究人员提出了特征驱动的标签生成方法和无监督方法, 以摆脱知识图谱实体解析任务对于已知标签的依赖. 特征驱动的标签生成方法根据知识图谱中的数据特征自动生成匹配标签. 其中, EVA[81]根据实体的图像信息自动生成匹配标签, 但获取充足的实体图像往往需要额外的人力成本, 这限制了 EVA 方法的应用范围. MRAEA[86]先利用谷歌翻译对实体名称进行翻译, 将翻译后互为最近邻 (用编辑距离度量) 的实体对视为匹配标签, 而后通过迭代优化策略进一步对标签进行扩充. 然而, MRAEA 会在迭代过程中累积错误, 从而导致错误匹配标签的生成. 为了减少错误标签生成并进一步提升匹配结果质量, Ge 等人[89]提出了一种端到端的知识图谱实体解析方法 EASY, 其通过挖掘实体的丰富名称特征与结构特征以自动生成高质量的匹配标签. PRASE[69]根据概率模型自动生成实体解析标签, 并通过迭代方式不断扩充标签量. 在无监督方法中, Mao 等人[90]指出: 虽然现有的方法侧重于通过将正确匹配标签所对应的实体嵌入向量彼此拉近的方式, 以促进实体解析模型的有效训练, 但实际上知识图谱实体解析任务受益于使大量未标记的负实体对 (两者无法正确匹配) 所对应的嵌入向量彼此远离. 进一步地, Mao 等人[90]根据此项发现提出了一种跨知识图谱的对比学习策略, 以实现无需标注的知识图谱实体解析. 此外, 不同于前述基于表示学习的方法, Mao 等人[91]提出了一种非表示学习的无监督方法 SEU, 该方法将实体解析问题视为最优分配问题, 以避免表示学习所需耗费的高昂时间成本以及人工标注成本[92], 进而实现高质量且高效的匹配性能.

可扩展性问题是当前知识图谱实体解析任务的另一研究热点. 具体来说, 由最新的实验评估工作[93]可知: 绝大多数现有方法关注于小规模知识图谱的实体解析任务, 其难以处理真实场景中普遍存在的大规模数据. 当前具有高质量实体解析效果的模型 (例如, BERT-INT[74]、RREA[94]) 往往由于其复杂的模型设计而导致计算效率低下且不易扩展. 相反, 诸如 MTransE[60]、IPTransE[61]等简单模型虽然计算高效, 但其匹配效果较差. 为了在保证实体解析质量的前提下提升实体解析方法的可扩展性, Ge 等人[95]提出了一种大规模知识图谱实体解析方法 LargeEA, 其通过基于图划分的小批量 (mini-batch) 生成策略以及实体名称特征和结构特征的有效挖掘, 实现了高质量、可扩展的大规模知识图谱实体解析. 随后, LIME[96]在 LargeEA 的基础上提出了一种双向图划分策略, 以支持更大规模的知识图谱实体解析. Gao 等人[87]则提出了一种基于多轮采样的大规模实体解析方法 ClusterEA, 其通过采样的方式对知识图谱进行读取和训练, 每次仅更新采样能够装进内存的样本进行训练. 经过多轮采样后, 采样结果将覆盖整个知识图谱.

此外, 当前的知识图谱实体解析方法几乎都基于一个重要假设: 一个知识图谱中的某个实体一定能在另一知

识图谱中找到匹配的对应实体[97]. 由于在真实场景中, 待匹配的知识图谱可能在规模、来源等方面都存在较大差异, 一个知识图谱中的某个实体 (悬空实体) 可能在另一知识图谱中不存在与之匹配的实体. Sun 等人[97]首次考虑了悬空实体, 提出了一个结合悬空实体检测和实体解析的多任务学习框架, 在检测并移除悬空实体后再对知识图谱进行实体解析. 为解决真实场景中难以获取带标签的悬空实体的问题, Luo 等人[98]通过引入空实体, 将实体解析与悬空实体检测问题转化为全局最优传输问题. 这些方法都假设悬空实体的嵌入表示应该是孤立的, 并且与其他实体相距甚远, 因而无法有效检测出与可匹配实体相似的悬空实体. 表 2 通过分析知识图谱中数据的内在特征建模方式以及知识图谱与辅助信息的交互方式, 对现有基于表示学习的知识图谱实体解析方法进行了总结.

表 2 基于表示学习的知识图谱实体解析方法比较

表示学习方法	辅助信息	方法名称
三元组表征模型	无	MTransE[60], IPTransE[61], BootEA[62], SEA[63], OTEA[99], TransEdge[100], MMEA[101], AKE[102], REA[67]
	属性	JAPE[80], AttrE[71], MultiKE[77], COTSAE[85], JarKA[93]
	描述	KDCoE[70]
	名称	MultiKE[77]
	文本语料	JEANS[82]
图神经网络	无	MuGNN[65], NAEA[103], AVR-GCN[104], KECG[105], AliNet[106], SSP[107], HyperKA[67], STLT-EA[108], KE-GCN[109], RAC[110], ActiveEA[88], TEA-GNN[111]
	属性	LinkNBed[112], GCNAlign[64], Cross-Align[70], HMAN[73], CG-MuAlign[113], AttrGNN[78], EPEA[114], EVA[81]
	名称	RAGA[115], GMNN[116], RDGCN[117], HGCN[118], MRAEA[86], GM-EHD-JEA[119], CEA[120], CEAFF[121], NMN[122], RREA[94], AttrGNN[78], EPEA[114], RE-GCN[123], DINGAL[124], RNM[82], Dual-AMN[125], PSR[90], ERMC[126], LargeEA[95], LIME[96], SelfKG[127], EASY[89], ClusterEA[87], ZeroMatcher[128]
	文本语料	JEANS[82]
	类别	LinkNBed[112]
	图像	EVA[81]
概率模型	无	NTAM[68]
	属性	PRASE[69]
循环神经网络	无	RSNs[72]
	属性	COTSAE[85]
	描述	KDCoE[70], AttrE[71]
Transformer模型	无	IMEA[75]
	属性	BERT-INT[74]
	描述	HMAN[73], BERT-INT[74]
	名称	BERT-INT[74]

总的来说, 尽管专家学者们已提出了大量的知识图谱实体解析方法, 但现有工作需: (1) 进一步探讨实体解析的成本代价问题, 以增强实体解析技术在真实场景中的普适性; (2) 在可扩展性与带悬空实体的知识图谱解析方面的研究尚处于起步阶段, 需深入探索具有高精度保障的大规模知识图谱实体解析方法.

2 数据融合

由于数据不完整、数据错误和数据过时等问题, 不同数据源可能产生相互矛盾的数据, 对后续查询分析的结果产生误导作用. 例如, 人们拨打无效的电话号码可能无法联系上对方, 或是导航到错误的诊所导致错过最佳就诊时间, 从而导致严重后果. 因此, 解决不同来源数据之间的冲突问题并识别相关数据的真实/正确性显得至关重要. 近年来, 随着互联网的迅速发展以及网络监管的局限性, 互联网成为虚假信息泛滥的重灾区, 数据冲突问题变得尤为突出. 数据冲突可以被归纳为两种问题, 即不确定性和矛盾性. 不确定性由信息的不完整引起, 指非空值的数据与一个或多个空值数据之间的数据冲突. 矛盾性是指两个或多个不同的非空值之间的冲突, 这些非空值亦是对同一实体中相同属性的描述. 数据融合的目标是通过鉴别不同来源数据的真实性, 解决来自不同数据源的数据冲突

问题, 进而确保数据集成过程中的数据一致性.

2.1 面向关系型数据的数据融合

概念与定义: 互联网极大地改变了人们的生活. 近年来, 互联网上的数据量正在以惊人的速度增长, 且已渗透到生产生活的各个方面. 这些数据大多存储于底层关系型数据库中, 用户通过 Web 表单对其进行查询, 从而可视化地展现在网页上. 然而, 这些数据的质量和可信度却难以保证. 因此, 通过数据融合技术鉴别来自不同来源的关系型数据的真实/正确性, 从而解决数据冲突显得至关重要.

令集合 O 表示来源于不同数据源 S 的一系列数据对象. 给定一个数据源 $s \in S$ 和一个数据对象 $o \in O$, 由该数据源 s 所提供的关于数据对象 o 的观测值可表示为 $V_o^s = \{v_{o,1}^s, v_{o,2}^s, \ldots, v_{o,|V_o^s|}^s\}(|V_o^s| \geqslant 1)$. 数据融合的目标是为每个数据对象 o 从若干数据源得到的一系列观测值中推测出对应的真值, 表示为 $V_o^* = \{v_{o,1}^*, v_{o,2}^*, \ldots, v_{o,|V_o^*|}^*\}(|V_o^*| \geqslant 1)$.

评价指标: 面向关系型数据的数据融合常用评估指标为数据对象的真值评估准确度. 考虑到数据对象的取值可能为连续值或离散值, 下面分别给出对于连续值和离散值的真值评估准确度的计算方式.

对于连续值而言, 通常采用绝对平均误差 (mean absolute error, *MAE*) 进行计算, 如公式 (4) 所示:

$$MAE = \sum_{i=1}^{|O|} \frac{\sum_{j=1}^{|V_i|} |\hat{v}_{i,j} - v_{i,j}^*|}{|O||V_i|} \tag{4}$$

其中, $|O|$表示数据对象数量, $|V_i|$为第 i 个数据对象的数值总数, $\hat{v}_{i,j}$ 表示由数据融合方法推测所得的正确值, $v_{i,j}^*$ 为与 $\hat{v}_{i,j}$ 所对应的真值.

对于离散值而言, 数据对象的真值评估准确度 (*Accuracy*) 为由数据融合方法推测所得的真值数量占数值总数的比例, 如公式 (5) 所示:

$$Accuracy = \sum_{i=1}^{|O|} \frac{N_i}{|O||V_i|} \tag{5}$$

其中, N_i 为由数据融合方法推测所得的与第 i 个数据对象相关的真值数量.

研究现状及分析: 对于冲突的数据, 如何分辨哪个数据源中包含的信息为正确, 进而保证数据融合的质量是极其重要的. 一种朴素的方法是多数投票策略, 该策略通过计算每个信息出现的次数, 将次数最大项视为正确值. 然而, 该方法在大数据环境下容易出错, 其原因归结于许多数据来源之间会相互传播虚假消息, 或者盲目复制过时或错误的数据. 为了确保数据融合的准确性与有效性, 专家学者做了若干研究工作, 主要分为基于迭代的方法、基于优化的方法和基于概率图模型的方法, 如表 3 所示.

表 3 面向关系型数据的数据融合方法比较

假设类型	基于迭代的方法	基于优化的方法	基于概率图模型的方法
源一致性假设	[129–133]	[134,135]	[136]
源独立性假设	[129–131]	[135]	[136,137]
源依赖性假设	[132,138]	[139,140]	[132,141–144]

文献 [145] 对这 3 类方法进行了较为全面的总结与分析. 为了便于理解, 本文先依据文献 [145] 简单介绍上述 3 类方法的核心思想, 而后给出进一步分析.

基于迭代的方法认为对于数据对象的真值预测与数据来源的可信度估计之间是相辅相成的. 基于此, 该类方法通常被设计为一个迭代过程. 每轮迭代包含两个子阶段, 即: (1) 数据对象的真值预测阶段; (2) 数据来源的可信度估计阶段, 两者将被迭代计算直至收敛.

在真值预测阶段, 假定每个数据来源的可信度固定不变, 随后通过聚合不同数据源之间的可信度 (譬如投票法) 来预测每个数据对象 o 的真值. 文献 [129] 给出了一种基于加权投票法的真值估计方法, 计算公式为:

$$\Pr(v) = \left(\sum_{s \in S_v} \frac{w_s}{|V_s|} \right)^{1.2} \tag{6}$$

其中, v 代表数据对象 o 的一个观测值 (可视为真值的一个候选项), $\Pr(v)$ 表示 v 为真值的可能性, S_v 代表包含该观测值 v 的所有数据来源集合, w_s 表示数据源 s 的可信度权重, $|\mathcal{V}_s|$ 为数据源 s 中的所有取值总数. 根据公式 (5) 的计算结果, 对每个候选项进行排序, 最终被视为真值的结果往往来自具有高权重的数据源. 这是因为, 专家学者通常认为: 可靠的数据来源更有可能提供正确信息.

在数据来源的可信度估计阶段, 假定每个观测值为真值的可能性固定不变 (由公式 (6) 计算所得), 每个数据来源的可信度权重可根据当前每个观测值为真值的可能性 $\Pr(v)$ 进行计算, 如公式 (7) 所示:

$$w_s=\sum_{v\in\mathcal{V}_s}\left(\Pr(v)\times\frac{\dfrac{w_s}{|\mathcal{V}_s|}}{\displaystyle\sum_{s'\in S_v}\dfrac{w_{s'}}{|\mathcal{V}_{s'}|}}\right) \tag{7}$$

结合公式 (6) 和公式 (7) 可知, $\Pr(v)$ 的值越高, 其对应的观测值 v 对数据源权重估计的贡献亦越大.

在基于迭代的方法中, Galland 等人[130]受信息检索中相似性度量算法启发, 引入余弦相似度、2-Estimates 方法和 3-Estimates 方法, 迭代地估计源数据的可信度. Pasternack 等人[129]在迭代模型中引入用户的领域知识与一般常识 (统称为先验知识), 让系统根据先验知识而不是多数投票来确定真相. 然而, 此类方法容易在迭代过程中累积错误, 从而对结果产生误导, 进而限制了此类方法的数据融合精度.

基于优化的方法通常根据已知信息 (或称标签数据) 对每个数据对象的观测值进行优化, 如公式 (8) 所示:

$$\arg\min_{\{w_s\},\{v_o^*\}}\sum_{o\in\mathcal{O}}\sum_{s\in\mathcal{S}}w_s\cdot dis(v_o^s,v_o^*),\ \text{s.t. }\delta(W)=1 \tag{8}$$

其中, $dis(\cdot)$ 表示一个距离度量函数. 例如, 对于离散值而言, 可以采用具有 0-1 分布的损失函数进行距离度量; 对于连续值而言, 则可以采用 L^2-norm 进行距离计算. W 代表所有数据来源的权重集合, $W=\{w_1, w_2, \ldots, w_{|W|}\}$, $\delta(W)=1$ 表示对所有数据来源的权重进行归一化处理. 在 $\delta(W)=1$ 的权重归一化约束条件下, 通过令公式 (8) 得到的结果最小化, 能够尽可能地满足: (1) 可靠的数据来源所提供的观测值与其对应的真值距离接近; (2) 不可靠的数据来源所提供的观测值与其对应的真值距离较远. 与基于迭代的方法相类似, 基于优化的方法以迭代方式对 $dis(v_o^s,v_o^*)$ 与 w_s 进行计算, 直至收敛. 此外, 为了便于描述, 对于任一数据对象 o, 公式 (8) 假定每个数据源有且仅有一个与 o 相关的观测值 v_o^s, 其亦可拓展至具有多个观测值/真值的情况.

在基于优化的方法中, CATD[134]指出了数据源存在的长尾现象 (long-tail phenomenon)——大多数来源只提供少量的信息, 只有少数来源能够提供大量的信息. 例如, 很少有像维基百科这样覆盖信息广泛的网站. 这些只包含少量信息的源数据可信度的波动范围可能很大. CATD 被用于自动检测来自具有长尾现象的冲突数据的正确性, 并用优化方法估计数据可信度的置信区间. SLiMFast[139]利用判别模型减少对数据源分布情况的假设条件, 并设计了一种能够自动学习 SLiMFast 所需参数的优化器, 相关实验证明该优化器能够准确地给出参数, 从而产生最佳的数据融合结果.

基于概率图模型的方法先根据数据源中已捕获的数据构建概率图模型 (probabilistic graphical models, PGMs), 而后预测每条源数据真实性的概率 (或称可信度), 常规计算公式为:

$$\prod_{s\in\mathcal{S}}p(w_s\mid\beta)\prod_{o\in\mathcal{O}}\left(p(v_o^*\mid\alpha)\prod_{s\in\mathcal{S}}p(v_o^s\mid v_o^*,w_s)\right) \tag{9}$$

在基于概率图模型的方法中, TruthFinder[131]是基于概率和迭代模型的混合框架, 并结合语句的相似性对源数据的可信度进行估计. AccuPr[132]通过贝叶斯分析发现数据源之间存在的依赖关系来预估源数据的可信度. LCA 方法通过提取网页数据源中存在的各种特征作为参数, 并根据最大后验概率 (MAP) 来分析数据可信度.

现有面向关系型数据的数据融合方法通常遵循以下 3 种源依赖关系假设, 分别为源一致性假设、源独立性假设以及源依赖性假设. 早期, 大多数工作遵循源一致性假设[129,131–136]. 此类假设认为来自同一个数据源的不同数据为真的可能性相同. 该假设是估计源可靠性的最重要假设之一, 其在一些实际应用中具有一定的合理性, 因为可靠的来源往往能够提供可靠的数据. 随后, 一些研究工作[129,130,133,135–137]提出了源独立性假设的概念——不同数据源

之间是相互独立的. 在该假设下, 由不同来源所提供的真实/正确数据更相似, 而虚假信息则大相径庭. 另外一些研究工作[132,138−144]则认为不同数据来源之间存在依赖关系 (即源依赖性假设). 例如, 若一个数据来源所提供关于动作电影的信息具有较高的可靠性, 则该数据来源也可能提供较可靠的冒险电影相关信息. 其原因在于, 动作电影通常与冒险电影在题材上存在一定程度的相关性. 总的来说, 尽管现有方法提出了多种不同的源依赖关系假设, , 但每种假设在适用范围上均存在一定的局限性, 难以有效地应对复杂多样的真实数据融合场景.

2.2 面向知识图谱的数据融合

概念与定义: 随着互联网的发展, 知识图谱以其强大的语义处理能力和开放组织能力为“知识互联”时代的到来奠定了基础. 传统的知识图谱构建方法主要依赖于人工知识的输入, 这些方法产生的知识量较有限. 因此, 专家学者致力于通过自动化构建的方法进一步扩充知识图谱, 实现知识集成. 例如, 从网络中自动提取三元组信息, 以补充从人工输入和现有知识图谱中未能包含的知识. 然而, 这类方法通常会提取到嘈杂且不可靠的三元组. 为了避免噪声数据对集成高质量的知识图谱造成不利影响, 面向知识图谱的数据融合 (以下简称知识融合) 概念被引入, 其旨在验证知识图谱中三元组的正确性.

令 $\mathcal{G}=(\mathcal{E},\mathcal{R},\mathcal{T})$ 代表一个知识图谱, 其亦可被视为若干三元组的集合, 其中 $\mathcal{E}$ 表示实体集合, $\mathcal{R}$ 代表实体之间的关系集合, $\mathcal{T}$ 表示由实体与关系构成的三元组. 三元组的形式化定义为 $t=(e_i, r, e_j)\in\mathcal{T}$. 关于知识图谱的详细定义已在第 1.2 节给出, 因而此处不再赘述. 假定 $f=(t, s)$ 为一个事实对象, 其代表三元组 t 是由数据来源 s 所提供的. 令 $C=\{c_1, c_2, \ldots, c_{|C|}\}$ 为从不同来源提取得到的所有事实对象的集合, 知识融合的目的在于从不同来源提取得到的事实对象集合中识别出正确的子集, 表示为 $C'\subseteq C$. 随后, 这些被识别为正确的事实对象将被添加到已知的知识图谱中, 以扩充知识图谱内容, 实现知识集成.

评价指标: 知识融合的评价指标同面向关系型数据的数据融合方法一致 (详见第 2.1 节), 此处不再赘述.

研究现状及分析: 通常来说, 面向关系型数据的数据融合方法也可以运用于知识融合, 其原因在于这些方法并未将关系型数据的独有特性纳入考量, 弱化了关系型数据与知识图谱在数据结构方面的差异性. 为了进一步提升知识融合质量, 一些研究工作探讨了知识图谱的内在特性, 代表性工作有 OKELE[146]和 TKGC[147].

Cao 等人[146]指出长尾现象大量存在于知识图谱中, 这里的长尾现象是指大部分数据源仅提供关于特定三元组的信息, 仅有少数三元组能够从多个不同来源中获取. 知识融合方法的一种常见假设是: 若与特定三元组相关的来源数量越多, 则该三元组为真/正确的可能性越大. 然而, 由于长尾现象的普遍存在, 依靠这种假设难以有效地评估仅由少量数据源所提供的三元组的正确性. 为此, Cao 等人[146]提出了 OKELE, 其通过一种新颖的概率图模型对具有长尾现象的知识图谱进行有效建模, 从而准确判别每个三元组的正确性, 图 4 为该概率图模型的示例.

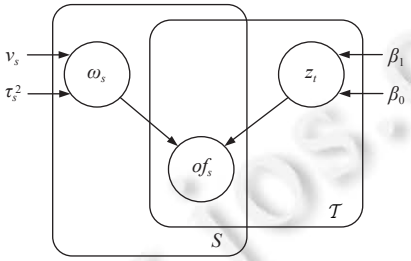


图 4 OKELE 的概率图模型示例

对于每个三元组 t, OKELE 将其为真的概率建模为一个在 [0, 1] 区间内的连续概率分布 (或称 beta 分布), 用隐变量 z_t 表示, $z_t\sim(\beta_1,\beta_0)$, β_1 和 β_0 为控制该分布的超参数. 在 OKELE 中, $\beta=(\beta_1,\beta_0)$ 代表决定 t 为真的先验概率分布, 其中 β_1 表示与 t 相关的先验知识中认为其为真的来源数量; 反之, β_0 表示与 t 相关的先验知识中认为其为假的来源数量. 给定 z_t, t 是否为真可以通过阈值 θ 进行推测——若 $z_t\geqslant\theta$, 则 t 为真; 反之, 若 $z_t<\theta$, 则 t 为假. 对于每个数据源 s, OKELE 通过缩放逆卡方分布对其误差方差 ω_s 建模, 表示为 $\omega_s\sim\text{Scale-inv-}\chi^2(v_s,\tau_s^2)$, 该分布由超参数 v_s 和 τ_s^2 控制. 此外, OKELE 假定每个事实对象 $f=(t, s)$ 为真的可能性 o_{t_s} 均遵从正态分布, 表示为 $o_{t_s}\sim N(z_t,\omega_s)$, 其

中 z_t 和 ω_s 分别可被视为控制正态分布的均值和方差. OKELE 认为事实对象为真的可能性以 (未知的) 真值为中心, 其受数据源的质量影响逐渐偏离中心. 若数据源不可靠, 则由该数据源所提供的事实对象将偏离中心真值.

Huang 等人[147]指出不同类型的值 (包括离散值、连续值、字符串) 在知识融合时发挥着不同的作用, 应对其进行有效的区分, 并基于此提出了一种基于多源噪声数据的可信知识图谱补全技术 TKGC. TKGC 包含一种基于混淆概率 (confusion probability) 的半监督知识融合方法, 其能够为知识图谱补充新的可靠三元组信息, 从而实现知识集成. 图 5 给出了 TKGC 的技术流程框图.

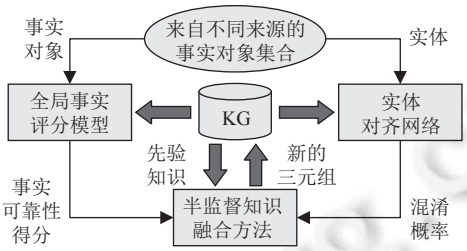
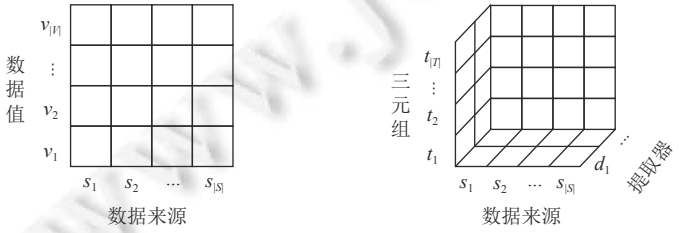


图 5 TKGC 技术流程框图

由图 5 可知, TKGC 主要由 3 个模块构成, 即全局事实评分模型 (holistic fact scoring)、半监督知识融合方法 (semi-supervised truth inference) 以及实体对齐网络 (value alignment networks). 给定来自不同来源的事实对象集合, 全局事实评分模型通过学习相应的嵌入向量来表征知识图谱中的实体和关系, 以衡量从不同来源中提取所得的事实对象的可靠性. 实体对齐网络旨在解决同一实体之间的异质性 (譬如“UK”与“United Kindom”指向真实世界中的同一实体). 此外, 根据实体对齐网络可以计算得到实体之间的混淆概率. 这里的混淆概率可以被视为观测值偏离真值的距离, 其旨在模拟数据源所引入的错误情况——由于数据源所提供的事实有误, 可能造成实体取值 (或称观测值) 偏离真值. 值得一提的是, 实体对齐网络还可以通过实体对齐的方式发现以前未见过的新的实体及其相关的事实对象. 随后, 根据事实可靠性得分 (由全局事实评分模型计算所得) 和混淆概率 (由实体对齐网络计算所得), 半监督知识融合方法可以通过知识图谱中的 (小部分) 先验知识从不同来源的事实对象集合中推断出可靠的三元组. 最后, 这些可靠三元组将被加入至知识图谱, 以实现知识扩充.

然而, 上述方法仍然不足以得到理想的知识融合效果, 其主要归因于知识图谱与关系型数据在数据提取/构建时的内在差异——关系型数据具有高度的结构化特性, 且在数据录入/修改时受数据库规范限制, 因而对于关系型数据的数据融合可以着重于考虑数据本身的正确与否. 相比之下, 知识图谱往往是从来源各异的数据源 (譬如网页表格、文本、XML、JSON 等) 中自动提取得到, 其不仅受数据本身存在的错误影响, 而且受制于知识提取器 (knowledge extractions) 的有效性[148]. 图 6 展示了传统的数据融合与知识融合的区别. 基于此, Dong 等人[148]提出了一种概率知识融合方法 KnowledgeVault. 给定一个三元组, 该方法将已知的高质量知识图谱作为先验知识, 结合由知识提取器得到的该三元组的特征向量, 以合理估计该三元组为真的可能性. KnowledgeVault 为如何将知识提取器的信息纳入知识融合技术提供了一种直观的解决方案, 但其尚未深入挖掘知识提取器的内在特征, 亦尚未深入考虑不同提取错误/误差对于知识融合效果的影响大小. 综上所述, 知识融合技术仍处于起步阶段, 尚需专家学者进行深入探索.



(a) 面向关系型数据的数据融合 (b) 面向知识图谱的数据融合 (知识融合)

图 6 面向关系型数据的数据融合与知识融合的区别示意图

3 数据清洗

在现实生活中, 脏数据无处不在, 各组织或机构的研究调查报告了惊人的脏数据比例: 在全球顶尖的企业中, 超过 25% 的关键数据都存在一定的数据缺陷[149]. 脏数据的存在, 不仅会导致错误的决定和不可靠的分析, 还可能对企业经济造成巨额损失. 例如, IBM 的报告指出: 脏数据造成美国每年损失近 3 万亿美元[150]. 近几十年来, 数据清洗已成为学术界和工业界的研究重点, 其旨在检测与修复脏数据 (包括属性值域错误、错别字、缺失值、数据冗余等), 从而提升数据质量[151]. 本节围绕关系型数据和知识图谱的数据清洗技术展开研究, 下面将分别对这两部分的相关工作进行阐述和分析.

3.1 面向关系型数据的数据清洗

概念与定义: 关系型数据广泛存在于现实生活中的各个领域, 譬如医疗信息系统、政府数据管理、企业业务管理等. 高质量的数据是支撑数据分析与决策的重要基石, 而无处不在的脏数据严重影响了分析决策的可靠性. 因此, 关系型数据清洗技术在数据管理过程中的重要性日渐凸显.

通常, 面向关系型数据的数据清洗任务假设数据集中的脏数据是由不准确的赋值所导致. 给定一个关系型数据集 D 及其相应的属性集合 $A=\{A[1], A[2], \ldots, A[m]\}$, D 可被视为一系列元组的集合, 其中每个元组 $e \in D$ 由一系列属性值 $V=\{e.A[1], e.A[2], \ldots, e.A[m]\}$ 组成. 对于元组 e 的第 i 个属性值 $e.A[i]$, 令 $e.A[i]^*$ 代表与之对应的真值. 基于此, 脏数据可被表示为 $e.A[i] \neq e.A[i]^*$. 面向关系型数据的数据清洗任务旨在检测给定数据集 D 中的脏数据, 并为每个脏数据推测出其对应的修复值 $\hat{v}$; 若 $\hat{v} = e.A[i]^*$, 则该脏数据已被正确修复.

表 4 给出了一个面向关系型数据的数据清洗示例, 其数据来源为美国 1994 年成人人口普查数据的部分记录[152]. 该数据集具有 7 项属性, 其中 CLS 指代工作情况; ED 表示受教育程度; MR 代表婚姻状况; REL 表示家庭关系; GEN 为性别; SAL 代表薪资情况. 为了便于描述, 表 4 中的脏数据均已加粗显示, 例如, $e_2.A[6]$="AR", $e_8.A[1]$="private". 面向关系型数据的数据清洗任务旨在检测上述脏数据, 并对其进行修复, 以提升数据质量. 例如, $e_2.A[6]$="AR"应被修复为 $e_2.A[6]^*$="AL", $e_8.A[1]$="private"应被修复为 $e_8.A[1]^*$="Self-emp".

表 4 美国 1994 年成人人口普查数据样例[152]

元组	CLS	ED	MR	REL	City	State	SAL
e_1	private	Bachelors	Married	Husband	DOTHAN	AL	<50 k
e_2	private	Bachelors	Married	Husband	DOTHAN	**AR**	>50 k
e_3	Self-emp	Masters	Married	Wife	DOTHAN	AL	>50 k
e_4	Local-gov	HS-grad	Married	Husband	BOAZ	AL	>50 k
e_5	Self-emp	Masters	Married	Wife	NOME	AK	>50 k
e_6	Never-worked	7th-8th	Divorced	Not-in-family	NOME	AK	<50 k
e_7	Self-emp	HS-grad	Never	Own-child	NOME	AK	<50 k
e_8	**private**	Some-college	Never	Own-child	BOAZ	AL	<50 k
e_9	State-gov	Masters	Never	Own-child	BOAZ	AL	>50 k
e_{10}	Local-gov	Bachelors	Divorced	Not-in-family	BOAZ	**AK**	<50 k
e_{11}	Self-emp	Some-college	Never	Own-child	BOAZ	AL	>50 k

研究现状及分析: 目前, 专家学者已提出了大量的关系型数据清洗技术, 主要可以分为: (1) 基于规则的关系型数据清洗方法; (2) 基于机器学习的关系型数据清洗方法; (3) 人机协同的关系型数据清洗方法. 接下来分别对其展开介绍与分析.

基于规则的关系型数据清洗方法旨在检测或修复违反规则约束的脏数据. 关系型数据的规则主要包括函数依赖、条件函数依赖以及否定约束等. 这些规则往往需要资深的领域专家进行人工定义. 然而, 雇佣资深专家需耗费高昂的人力成本且效率低下. 为此, 研究人员提出了许多数据驱动的自动规则挖掘方法, 譬如函数依赖自动挖掘方法[79−158]、条件函数依赖自动挖掘方法[159,160]和否定约束自动挖掘方法[161−165]. 基于规则的关系型数据清洗方法可

根据给定的规则执行数据清洗. 早期的方法只能针对单一类型的规则进行清洗[166−172]. Chu 等人[173]提出了一种全局清洗方法, 该方法首次将多种规则融合于统一的清洗框架, 以扩大规则覆盖面, 从而提升数据清洗效果. 随后, 若干研究工作[174−180]对支持多种规则的数据清洗方法进行了深入的探讨. 另外, Giannakopoulou 等人[181]设计的 CleanM 为用户提供了一种新的声明式查询语言, 以支持用户表达个性化的清洗需求. 进一步地, Giannakopoulou 等人[182]研发了 Daisy 系统, 该系统将数据清洗与用户查询需求进行无缝集成, 使得用户能够按需修复违反否定约束的脏数据. 尽管目前已有大量基于规则的关系型数据清洗方法, 但此类方法受制于给定规则的正确性与充分性, 不正确或不充分的规则不可避免地会导致修复效果不佳. 所以, 专家学者仍将持续探索基于自动规则生成的关系型数据清洗方法.

基于机器学习的关系型数据清洗方法旨在通过数据的概率分布对脏数据进行检测或修复. Yakout 等人[183]提出了一种基于统计概率的数据修复方法 SCARE, 其消除了数据修复对于预先定义规则约束的必要性. 给定已知的脏数据集合, SCARE 通过最大似然估计法为每个脏数据推测其最有可能的修复方案. ERACER[184]和 ActiveClean[185]探讨了迭代式数据清洗方法, 其能够通过机器学习模型与每次迭代过程中标记的干净/脏数据之间的交互作用, 不断提升数据清洗效果. Rekatsinas 等人[175]提出了一种概率推理驱动的全局数据修复框架 HoloClean. 给定已知的脏数据集合, HoloClean 先根据该数据集需遵循的否定约束规则为其生成概率图模型, 而后通过概率图模型计算每个脏数据的最大后验概率, 以推测其最有可能的修复方案. Mahdavi 等人[186]提出了 Baran, 该方法通过捕获每个脏数据的上下文信息以增强数据修复的可靠性. 前述基于机器学习的方法关注于如何有效地修复关系型数据, 而前置的错误检测阶段则往往仍依赖于基于规则的方法. 为此, HoloDetect[187]和 Raha[188]将关系型数据的错误检测问题抽象为机器学习中的二分类任务. 给定一部分标签样本数据, 此类方法旨在通过训练二分类模型, 将包含脏数据的数据集分为干净子集与错误子集两类, 其中错误子集即为检测到的脏数据集合. 虽然机器学习为关系型数据清洗提供了一种行之有效的解决方案, 但是基于机器学习的关系型数据清洗方法对于标签有着一定的依赖性, 同时标注可靠且充足的标签通常需要高昂的人力成本. 此外, 相较于基于规则的清洗方法, 机器学习方法可能损失数据清洗结果的可解释性, 即难以给出错误识别以及错误修复的依据, 不利于用户对清洗结果的有效分析. 鉴于此, 如何在保障数据清洗性能的前提下, 降低人力成本并提高清洗结果的可解释性还有待深入探索.

基于人机协同的关系型数据清洗方法旨在通过将人类智慧用于指导数据清洗过程, 从而确保清洗结果的有效性. Yakout 等人[171]提出了一种引导式的数据修复方法 GDR, 其能够在数据清洗过程中结合用户反馈, 以增强和加速现有的错误修复技术, 并最大限度地减少用户参与来降低人力成本. Chu 等人[189]提出了 KATARA, 该方法将高质量的知识库与众包技术作为识别脏数据的依据, 并给出可能的修复建议. Luo 等人[190]提出了一个可视化的交互式清洗系统 VisClean, 该系统允许用户通过可视化的界面轻松地回答数据清洗问题, 并通过多次迭代的人机交互模式提升清洗结果质量. 尽管目前已有的人机协同方法提供了多种利用人类智慧赋能数据清洗的有效途径, 但现实生活中的脏数据类型复杂多样, 仍需继续深入探索用户友好的数据清洗方式, 以低成本、易交互的方式使得人类智慧在数据清洗任务上发挥其最大限度的作用.

另外, 国外已有若干成熟的典型数据清洗系统落地使用, 例如, Oracle Enterprie Data Quality、Google Refine、IBM QualityStage 和 SAP BusinessObjects 等. 然而, 尚未见国内的相关成熟系统落地. 此外, 这些现有的系统主要利用传统的 ETL 规则进行数据清洗, 难以应对复杂多样的脏数据类型与清洗需求.

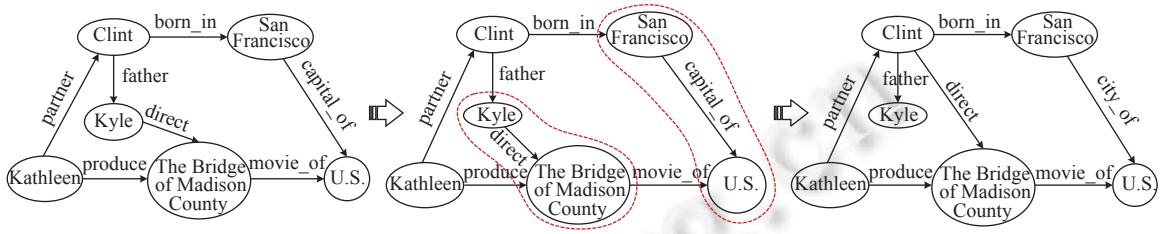
3.2 面向知识图谱的数据清洗

概念与定义: 随着知识互联时代的逐步迈进, 知识图谱已成为一项不可或缺的重要数据战略资源. 现实生活中知识图谱大多由程序自动构建, 而在构建过程中容易引入脏数据[191], 致使各类知识赋能应用的可靠性降低. 因此, 亟需探讨面向知识图谱的数据清洗问题.

令 $\mathcal{G}_d=\{\mathcal{E}_d,\mathcal{R}_d,\mathcal{T}_d\}$ 代表一个包含脏数据的知识图谱, 其中 $\mathcal{E}_d$ 为 $\mathcal{G}_d$ 的实体集合, $\mathcal{R}_d$ 为 $\mathcal{G}_d$ 的关系集合, $\mathcal{T}_d$ 为 $\mathcal{G}_d$ 中所包含的所有三元组. 面向知识图谱的数据清洗任务旨在检测并修复知识图谱中的脏数据. 脏数据检测的目的在于将给定脏知识图谱 $\mathcal{G}_d$ 分为 $\mathcal{G}_d^n$ 和 $\mathcal{G}_d^c$ 两个子集, 其中 $\mathcal{G}_d^n$ 为所有包含脏数据的三元组 (简称脏三元组) 集合,

$\mathcal{G}_d^c$ 为所有干净三元组的集合. 脏数据修复旨在根据脏数据检测结果对 $\mathcal{G}_d^n$ 中的每个脏三元组 $t=(e_i, r, e_j)\in\mathcal{T}_d^n$ 进行修复, 得到每个三元组对应的修复值, 表示为 $\hat{t}$. 假定 t^* 为 t 的真值, 若 $\hat{t}=t^*$, 则 t 已被正确修复.

图 7 给出了一个面向知识图谱的数据清洗示例. 图 7(a) 为一个包含脏数据的原始知识图谱. 图 7(b) 给出了经由错误检测得到的脏三元组, 即 (Kyle, direct, The Bridge of Madison County) 和 (San Francisco, capital_of, U.S.), 以虚线框标识. 图 7(c) 展示了修复后的知识图谱, 其中 (Kyle, direct, The Bridge of Madison County) 被修复为 (Clint, direct, The Bridge of Madison County), (San Francisco, capital_of, U.S.) 被修复为 (San Francisco, city_of, U.S.).



(a) 脏知识图谱 (b) 错误检测 (c) 错误修复

图 7 知识图谱数据清洗示例

研究现状及分析: 现有的知识图谱数据清洗方法主要依赖于图数据规则. 此类方法旨在检测或修复图数据中违反指定图数据依赖规则的脏数据. 图数据依赖规则指代正确/干净数据应保持的语义关系. 目前, 专家学者已探索了许多的图数据依赖规则[191−201], 以检测由脏数据导致的不一致问题. 随后, 各类数据修复方法被提出以修复由图数据依赖规则捕获到的脏数据. Cheng 等人[202]提出了一种图自动修复语义方法 GRR. 在该方法中, 由于 GRR 修复脏数据涉及子图同构问题 (被证明是 NP 完全问题), 故为了有效地规避子图同构所带来的复杂计算, GRR 利用了一种启发式的分解和连接策略以提高计算效率. Fan 等人[203]提出了 GFix, 该方法根据图数据依赖规则和迭代累积的真值数据推测出特定的图修复方案. Song 等人[204]提出了一种基于邻域约束的图数据修复方法, 该方法旨在检测与修复错误的顶点标签及其邻居节点. Lin 等人[205]提出了 StarFDs, 该方法定义了一种新颖的星形函数依赖, 并确保数据能够以最小的变化量进行修复, 使得修复后的数据满足星形函数依赖. Song 等人[206]设计了一种基于约束的缺失数据解释与修复方法, 该方法联合图数据依赖规则与所提出的双向算法, 以在保证信息量和高效性的基础上填补缺失数据. 尽管目前已提出了若干基于规则的知识图谱数据清洗方法, 但此类方法通常涉及 NP 难或 NP 完全问题, 随之带来的是高昂的计算代价, 因而仍需深入地研究更高效的算法, 实现快速且有效的知识图谱数据清洗效果. 此外, 虽然目前已提出了大量的图数据依赖规则挖掘方法, 但是自动生成的图数据依赖规则难以完全保证其正确性[207], 而不正确的规则不可避免地会导致错误的修复结果. 所以, 如何确保自动规则生成方法的正确性一直以来是研究人员所关注的关键问题, 并将持续深入地探索.

另外, 专家学者也已探讨了许多与知识图谱数据清洗任务高度相关的工作, 譬如知识图谱嵌入 (knowledge graph embedding), 其能够在一定程度上为知识图谱数据清洗提供技术基础或研究思路. 知识图谱嵌入旨在将知识图谱的实体与关系嵌入至连续的向量空间中[208], 以支持各种基于知识图谱的下游任务. 例如, 知识图谱补全[58]、实体解析[60]、事实检查[209]等. 其中, 事实检查可被视为知识图谱错误检测任务, 其目的在于发现知识图谱中每项三元组成立的可能性. 通常, 事实检查可以通过任何基于知识图谱嵌入的三元组分类任务进行实现: 根据知识图谱中实体与关系的嵌入向量计算每个三元组成立的可能性 (以三元组的评分函数为依据), 若可能性大于给定的阈值, 则该三元组成立, 反之不成立[210]. 根据评分函数的类别作为分类依据, 当前主流的知识图谱嵌入模型可分为基于翻译的知识图谱嵌入模型和基于语义匹配的知识图谱嵌入模型两类. 翻译模型[58,211−215]利用基于距离的评分函数进行模型训练, 此类方法将三元组成立的可能性通过两个实体之间的距离进行衡量. 语义匹配模型[59,79,212,216]则采用基于相似性的评分函数, 此类方法通过匹配实体与关系的潜在语义相似性来衡量三元组成立的合理性. 文献 [208] 综述了这两类方法的技术细节与优缺点. 尽管基于知识图谱嵌入的事实检查任务能够发现知识图谱中包含脏数据的三元组, 但其无法完成知识图谱数据清洗任务. 具体而言, 知识图谱数据清洗任务关注于检测脏数据在

三元组中的具体位置以及如何修复脏数据, 这超出了事实检查的问题范畴. 因此, 亟需进一步探索基于知识图谱嵌入的知识图谱数据清洗方法, 通过知识图谱嵌入的强大语义表征能力以有效地识别并修复知识图谱中的脏数据.

4 研究展望与趋势

面向关系型数据与知识图谱的数据集成技术是一个充满挑战性的研究课题, 目前仍有大量的研究工作亟待进一步深入探讨. 本节先依次对本文所聚焦的 3 类数据集成关键技术 (即实体解析、数据融合以及数据清洗) 的未来研究方向进行展望. 而后, 从数据类型角度出发, 探讨未来数据集成技术在不同类型数据上的研究方向. 最后, 进一步探究数据集成与数据治理之间的依存关系, 指出未来研究趋势.

(1) 放松知识图谱实体解析技术研究的前提假设. 随着知识互联时代的日渐趋近, 基于知识图谱的相关研究工作越来越受到专家学者的广泛重视. 目前现有的知识图谱实体解析方法主要建立在以下两大前提假设上: 1) 不同来源实体之间存在一一匹配关系; 2) 已知一部分既定的实体解析标签. 然而, 现实生活中不同来源的知识图谱通常具有不同规模且高度不完整, 因而无法为每个实体找到其对应的匹配项. 为此, 探讨非一一匹配约束下的知识图谱实体解析技术更具现实意义. 另外, 在现实生活中, 由于实体解析在各行各业的需求日益增大, 故在少标签甚至无标签情况下探索有效的知识图谱实体解析技术, 以降低成本、提高普适性, 这对于实际应用而言至关重要. 尽管本文已提出了若干基于实体名称信息的匹配标签生成策略, 为降低知识图谱实体解析的成本提供了行之有效的方法. 然而, 现实生活中存在一类情况——由于隐私保护或数据编码等问题, 导致实体名称差异过大, 因而难以利用名称信息实现有效的标签自动生成策略. 鉴于此, 如何利用实体本身所蕴含的内在特征探索更为通用的标签生成策略或探讨有效的无监督方法, 亦是值得进一步研究的关键问题.

(2) 支持大规模的关系型数据实体解析技术. 现有的关系型数据实体解析技术主要为单机算法, 由于单机系统在存储容量、计算资源等方面的限制, 其难以有效地处理大规模数据. 此外, 深度学习已在关系型数据实体解析任务上展现出强大的能力, 但其复杂且庞大的模型架构与模型参数对大规模的关系型数据实体解析任务提出了进一步的挑战. 因此, 在大数据环境下, 还需解决分布式存储、分布式模型训练等问题, 以确保大规模关系型数据实体解析技术的可扩展性. 另外, 计算效率也是大规模关系型数据实体解析任务所亟待解决的关键问题, 需深入地探索面向大规模关系型数据实体解析任务的数据分块、索引、剪枝等优化策略, 以进一步提高计算效率.

(3) 支持来源广泛、类型异构的复杂数据融合技术. 大数据时代, 数据来源复杂、类型异构、规模庞大, 如何高效地融合多源异构数据, 确保数据的正确性与一致性至关重要. 然而, 现有的方法大多关注于单一的数据类型, 难以有效地衡量具有异构特征但对应于同一实体的不同数据实例之间的正误性. 另外, 现有的数据融合方法耗时长且扩展性不高, 难以支持大规模的数据融合, 因而如何提升大数据环境下复杂数据融合效率仍有待进一步深入的研究.

(4) 低时延、高质量、易交互的数据清洗新模式. 随着大数据的蓬勃发展, 数据规模不断增大. 大多数现有的数据清洗方法侧重于全局数据清洗, 其在应对大规模数据时易导致效率与可扩展性问题, 难以满足现实生活中各种具有高时效性要求的数据查询与分析需求. 为此, 在线数据清洗技术应运而生. 此类技术以用户查询或分析需求为主导, 仅需清洗用户所需的数据集合, 因而大大缩小了数据清洗的范围, 能够在一定程度上提高清洗效率. 然而, 如何在包含脏数据的数据集中精准定位用户所需的数据范围, 并以低时延、高质量、易交互的方式将清洗后的干净数据及时返回给用户, 仍是需要深入探索的关键问题.

(5) 跨类型的数据集成新技术. 目前现有的数据集成技术大多关注于单一的数据类型, 而较少关注于不同类型数据之间的交互. 然而, 大数据类型多样, 半结构化数据 (JSON、XML 等)、非结构化数据 (多媒体数据) 等层出不穷, 且随着万物互联时代的日益趋近, 对于跨类型的数据集成需求日益迫切. 尽管目前已有一些专家学者进行了初步尝试 (譬如, Fan 等人[13]提出了支持关系型数据与图结构数据互通的异构实体解析技术), 但现有方法在数据种类数量、数据规模、效率等方面离满足实际的跨模态数据集成需求存在较大差距.

(6) 深入探索复杂数据治理技术. 实体解析、数据融合以及数据清洗不仅是数据集成的核心技术, 更是数据治理中的关键步骤, 但仅凭这些技术不足以解决当前复杂而多样的数据共享、共融、共用问题. 所以, 还需进一步深

入地探索复杂数据融合、元数据管理、数据风险监测与预警等技术, 从而为数据治理提供更丰富且有效的技术支撑.

5 总 结

大数据时代普遍存在数据流通困难、监管不足等问题, 导致数据共享薄弱、质量低下. 这驱使研究人员探索数据治理技术, 以实现数据共享、提升数据质量, 从而激活数据要素潜能. 数据集成作为数据治理的关键技术, 长期以来受到专家学者的重点关注. 数据集成旨在通过实体解析、数据融合和数据清洗等技术, 打破数据壁垒、实现信息共享、提升数据质量, 进而激活数据要素潜能. 本文聚焦关系型数据和知识图谱, 归纳总结并分析了实体解析、数据融合、数据清洗 3 方面的现有数据集成关键技术, 并展望了未来的研究方向与趋势, 以供相关的研究工作人员参考.

References:

[1] Reinsel D, Gantz J, Rydning J. The digitization of the world from edge to core. 2022. http://cloudcode.me/media/1014/idc.pdf

[2] China Academy of Information and Communications Technology. Big Data White Paper. 2021 (in Chinese). http://www.caict.ac.cn/kxyj/qwfb/bps/202112/t20211220_394300.htm

[3] Chen YG, Wang JC. A review of data integration. Computer Science, 2004, 31(5): 48–51 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2004.05.015]

[4] Yang XD, Peng ZY, Liu JQ, Li XH. An overview of information integration. Computer Science, 2006, 33(7): 55–59, 80 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2006.07.015]

[5] Wang S, Peng YW, Lan H, Luo QW, Peng ZY. Survey and prospect: Data integration methodologies. Ruan Jian Xue Bao/Journal of Software, 2020, 31(3): 893−908 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5911.htm [doi: 10.13328/j.cnki.jos.005911]

[6] Getoor L, Machanavajjhala A. Entity resolution: Theory, practice & open challenges. Proc. of the VLDB Endowment, 2012, 5(12): 2018–2019. [doi: 10.14778/2367502.2367564]

[7] Sun ZQ, Zhang QH, Hu W, Wang CM, Chen MH, Akrami F, Li CK. A benchmarking study of embedding-based entity alignment for knowledge graphs. Proc. of the VLDB Endowment, 2020, 13(12): 2326–2340. [doi: 10.14778/3407790.3407828]

[8] Zhuang Y, Li GL, Feng JH. A survey on entity alignment of knowledge base. Journal of Computer Research and Development, 2016, 53(1): 165–192 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20150661]

[9] Meng XF, Du ZJ. Research on the big data fusion: Issues and challenges. Journal of Computer Research and Development, 2016, 53(2): 231–246 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20150874]

[10] Guo ZM, Zhou AY. Research on data quality and data cleaning: A survey. Ruan Jian Xue Bao/Journal of Software, 2002, 13(11): 2076−2082 (in Chinese with English abstract). http://www.jos.org.cn/jos/article/abstract/20021103?st=search [doi: 10.13328/j.cnki.jos.2002.11.003]

[11] Hao S, Li GL, Feng JH, Wang N. Survey of structured data cleaning methods. Journal of Tsinghua University (Science and Technology), 2018, 58(12): 1037–1050 (in Chinese with English abstract). [doi: 10.16511/j.cnki.qhdxxb.2018.22.053]

[12] Wang X, Zou L, Wang CK, Peng P, Feng ZY. Research on knowledge graph data management: A survey. Ruan Jian Xue Bao/Journal of Software, 2019, 30(7): 2139−2174 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5841.htm [doi: 10.13328/j.cnki.jos.005841]

[13] Fan WF, Geng L, Jin RC, Lu P, Tugay R, Yu WY. Linking entities across relations and graphs. In: Proc. of the 38th IEEE Int'l Conf. on Data Engineering. Kuala Lumpur: IEEE, 2022. 634–647. [doi: 10.1109/ICDE53745.2022.00052]

[14] Ahmadi N, Sand H, Papotti P. Unsupervised matching of data and text. In: Proc. of the 38th IEEE Int'l Conf. on Data Engineering. Kuala Lumpur: IEEE, 2022. 1058–1070. [doi: 10.1109/ICDE53745.2022.00084]

[15] Li YL, Li JF, Suhara Y, Doan AH, Tan WC. Deep entity matching with pre-trained language models. Proc. of the VLDB Endowment, 2020, 14(1): 50–60. [doi: 10.14778/3421424.3421431]

[16] Azzalini F, Jin SL, Renzi M, Tanca L. Blocking techniques for entity linkage: A semantics-based approach. Data Science and Engineering, 2021, 6(1): 20–38. [doi: 10.1007/s41019-020-00146-w]

[17] Joshi M, Levy O, Weld DS, Zettlemoyer L. BERT for coreference resolution: Baselines and analysis. arXiv:1908.09091, 2019.

[18] Datta R, Joshi D, Li J, Wang JZ. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, 2008, 40(2): 5.

[doi: 10.1145/1348246.1348248]

[19] Cappuzzo R, Papotti P, Thirumuruganathan S. Creating embeddings of heterogeneous relational datasets for data integration tasks. In: Proc. of the 2020 ACM SIGMOD Int'l Conf. on Management of Data. Portland: ACM, 2020. 1335−1349. [doi: 10.1145/3318464.3389742]

[20] Konda P, Das S, Paul Suganthan GC, Doan AH, Ardalan A, Ballard JR, Li H, Panahi F, Zhang HJ, Naughton J, Prasad S, Krishnan G, Deep R, Raghavendra V. Magellan: Toward building entity matching management systems over data science stacks. Proc. of the VLDB Endowment, 2016, 9(13): 1581–1584. [doi: 10.14778/3007263.3007314]

[21] Chen C, Golshan B, Halevy AY, Tan WC, Doan AH. BigGorilla: An open-source ecosystem for data preparation and integration. IEEE Data Engineering Bulletin, 2018, 41(2): 10−22.

[22] Arasu A, Ré C, Suciu D. Large-scale deduplication with constraints using dedupalog. In: Proc. of the 25th IEEE Int'l Conf. on Data Engineering. Shanghai: IEEE, 2009. 952−963. [doi: 10.1109/ICDE.2009.43]

[23] Fan WF, Jia XB, Li JZ, Ma S. Reasoning about record matching rules. Proc. of the VLDB Endowment, 2009, 2(1): 407–418. [doi: 10.14778/1687627.1687674]

[24] Hernández MA, Stolfo SJ. The merge/purge problem for large databases. ACM SIGMOD Record, 1995, 24(2): 127–138. [doi: 10.1145/568271.223807]

[25] Singh R, Meduri V, Elmagarmid A, Madden S, Papotti P, Quiané-Ruiz JA, Solar-Lezama A, Tang N. Generating concise entity matching rules. In: Proc. of the 2017 ACM SIGMOD Int'l Conf. on Management of Data. Chicago: ACM, 2017. 1635−1638. [doi: 10.1145/3035918.3058739]

[26] Singh R, Meduri VV, Elmagarmid A, Madden S, Papotti P, Quiané-Ruiz JA, Solar-Lezama A, Tang N. Synthesizing entity matching rules by examples. Proc. of the VLDB Endowment, 2017, 11(2): 189–202. [doi: 10.14778/3149193.3149199]

[27] Marcus A, Wu E, Karger D, Madden S, Miller R. Human-powered sorts and joins. Proc. of the VLDB Endowment, 2011, 5(1): 13–24. [doi: 10.14778/2047485.2047487]

[28] Wang JN, Kraska T, Franklin MJ, Feng JH. CrowdER: Crowdsourcing entity resolution. Proc. of the VLDB Endowment, 2012, 5(11): 1483–1494. [doi: 10.14778/2350229.2350263]

[29] Gokhale C, Das S, Doan AH, Naughton JF, Rampalli N, Shavlik J, Zhu XJ. Corleone: Hands-off crowdsourcing for entity matching. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. Snowbird: ACM, 2014. 601−612. [doi: 10.1145/2588555.2588576]

[30] Chai CL, Li GL, Li J, Deng D, Feng JH. Cost-effective crowdsourced entity resolution: A partial-order approach. In: Proc. of the 2016 ACM SIGMOD Int'l Conf. on Management of Data. San Francisco: ACM, 2016. 969−984. [doi: 10.1145/2882903.2915252]

[31] Vesdapunt N, Bellare K, Dalvi N. Crowdsourcing algorithms for entity resolution. Proc. of the VLDB Endowment, 2014, 7(12): 1071–1082. [doi: 10.14778/2732977.2732982]

[32] Bilenko M, Mooney RJ. Adaptive duplicate detection using learnable string similarity measures. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2003. 39−48. [doi: 10.1145/956750.956759]

[33] Cohen WW, Richman J. Learning to match and cluster large high-dimensional data sets for data integration. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Edmonton: ACM, 2002. 475−480. [doi: 10.1145/775047.775116]

[34] Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Edmonton: ACM, 2002. 269−278. [doi: 10.1145/775047.775087]

[35] Wu RZ, Chaba S, Sawlani S, Chu X, Thirumuruganathan S. ZeroER: Entity resolution using zero labeled examples. In: Proc. of the 2020 ACM SIGMOD Int'l Conf. on Management of Data. Portland: ACM, 2020. 1149−1164. [doi: 10.1145/3318464.3389743]

[36] Collobert R, Bengio S. SVMTorch: Support vector machines for large-scale regression problems. Journal of Machine Learning Research, 2001, 1: 143–160. [doi: 10.1162/15324430152733142]

[37] Rish I. An empirical study of the naive bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. 2001, 3(22): 41–46.

[38] Reynolds D. Gaussian mixture models. In: Li SZ, Jain AK, eds. Encyclopedia of Biometrics. Boston, MA: Springer, 2015. 827−832. [doi: 10.1007/978-1-4899-7488-4_196]

[39] Ebraheem M, Thirumuruganathan S, Joty S, Ouzzani M, Tang N. Distributed representations of tuples for entity resolution. Proc. of the VLDB Endowment, 2018, 11(11): 1454–1467. [doi: 10.14778/3236187.3236198]

[40] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

[41] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods

in Natural Language Processing. Doha: ACL, 2014. 1532−1543. [doi: 10.3115/v1/D14-1162]

[42] Mudgal S, Li H, Rekatsinas T, Doan AH, Park Y, Krishnan G, Deep R, Arcaute E, Raghavendra V. Deep learning for entity matching: A design space exploration. In: Proc. of the 2020 ACM SIGMOD Int'l Conf. on Management of Data. Houston: ACM, 2018. 19−34. [doi: 10.1145/3183713.3196926]

[43] Nie H, Han XP, He B, Sun L, Chen B, Zhang W, Wu SH, Kong H. Deep sequence-to-sequence entity matching for heterogeneous entity resolution. In: Proc. of the 28th ACM Int'l Conf. on Information and Knowledge Management. Beijing: ACM, 2019. 629−638. [doi: 10.1145/3357384.3358018]

[44] Fu C, Han XP, Sun L, Chen B, Zhang W, Wu SH, Kong H. End-to-end multi-perspective matching for entity resolution. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 4961−4967. [doi: 10.24963/ijcai.2019/689]

[45] Zhang DX, Nie YY, Wu S, Shen YY, Tan KL. Multi-context attention for entity matching. In: Proc. of the 2020 Web Conf. Taipei: ACM, 2020. 2634−2640. [doi: 10.1145/3366423.3380017]

[46] Kasai J, Qian K, Gurajada S, Li YY, Popa L. Low-resource deep entity resolution with transfer and active learning. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 5851−5861. [doi: 10.18653/v1/P19-1586]

[47] Zhao C, He YY. Auto-EM: End-to-end fuzzy entity-matching using pre-trained deep models and transfer learning. In: Proc. of the 2019 World Wide Web Conf. San Francisco: ACM, 2019. 2413−2424. [doi: 10.1145/3308558.3313578]

[48] Li B, Miao YK, Wang YS, Sun YF, Wang W. Improving the efficiency and effectiveness for BERT-based entity resolution. Proc. of the AAAI Conf. on Artificial Intelligence, 2021, 35(15): 13226–13233. [doi: 10.1609/aaai.v35i15.17562]

[49] Li B, Wang W, Sun YF, Zhang LH, Ali MA, Wang Y. GraphER: Token-centric entity resolution with graph convolutional neural networks. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(5): 8172–8179. [doi: 10.1609/aaai.v34i05.6330]

[50] Zügner D, Akbarnejad A, Günnemann S. Adversarial attacks on neural networks for graph data. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. London: ACM, 2018. 2847–2856. [doi: 10.1145/3219819.3220078]

[51] Li P, Cheng X, Chu X, He YY, Chaudhuri S. Auto-FuzzyJoin: Auto-program fuzzy similarity joins without labeled examples. In: Proc. of the 2021 Int'l Conf. on Management of Data. ACM, 2021. 1064−1076. [doi: 10.1145/3448016.3452824]

[52] Zhang DX, Li DS, Guo L, Tan KL. Unsupervised entity resolution with blocking and graph algorithms. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(3): 1501–1515. [doi: 10.1109/tkde.2020.2991063]

[53] Ge CC, Wang PF, Chen L, Liu XZ, Zheng BH, Gao YJ. CollaborEM: A self-supervised entity matching framework using multi-features collaboration. IEEE Trans. on Knowledge and Data Engineering, 2021. [doi: 10.1109/TKDE.2021.3134806]

[54] Mahdisoltani F, Biega J, Suchanek FM. YAGO3: A knowledge base from multilingual wikipedias. In: Proc. of the 7th Biennial Conf. on Innovative Data Systems Research. Asilomar: CIDR, 2015. 1−11.

[55] Jiménez-Ruiz E, Cuenca Grau B. LogMap: Logic-based and scalable ontology matching. In: Proc. of the 10th Int'l Semantic Web Conf. Bonn: Springer, 2011. 273−288. [doi: 10.1007/978-3-642-25073-6_18]

[56] Zhuang Y, Li GL, Zhong ZJ, Feng JH. Hike: A hybrid human-machine method for entity alignment in large-scale knowledge bases. In: Proc. of the 2017 ACM on Conf. on Information and Knowledge Management. Singapore: ACM, 2017. 1917−1926. [doi: 10.1145/3132847.3132912]

[57] Suchanek FM, Abiteboul S, Senellart P. PARIS: Probabilistic alignment of relations, instances, and schema. Proc. of the VLDB Endowment, 2011, 5(3): 157–168. [doi: 10.14778/2078331.2078332]

[58] Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proc. of the 26th Annual Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 2787−2795.

[59] Trouillon T, Welbl J, Riedel S, Gaussier É, Bouchard G. Complex embeddings for simple link prediction. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 2071−2080.

[60] Chen MH, Tian YT, Yang MH, Zaniolo C. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Melbourne: AAAI Press, 2017. 1151−1517.

[61] Zhu H, Xie RB, Liu ZY, Sun SM. Iterative entity alignment via joint knowledge embeddings. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. Melbourne: AAAI Press, 2017. 4258−4264.

[62] Sun ZQ, Hu W, Zhang QH, Qu YZ. Bootstrapping entity alignment with knowledge graph embedding. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 4396−4402.

[63] Pei SC, Yu L, Hoehndorf R, Zhang XL. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In: Proc. of the 2019 World Wide Web Conf. San Francisco: ACM, 2019. 3130–3136. [doi: 10.1145/3308558.3313646]

[64] Wang ZC, Lv QS, Lan XH, Zhang Y. Cross-lingual knowledge graph alignment via graph convolutional networks. In: Proc. of the 2018

Conf. on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 349–357. [doi: 10.18653/v1/D18-1032]

[65] Cao YX, Liu ZY, Li CJ, Liu ZY, Li JZ, Chua TS. Multi-channel graph neural network for entity alignment. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 1452–1461. [doi: 10.18653/v1/P19-1140]

[66] Sun ZQ, Chen MH, Hu W, Wang CM, Dai J, Zhang W. Knowledge association with hyperbolic knowledge graph embeddings. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 5704–5716. [doi: 10.18653/v1/2020.emnlp-main.460]

[67] Pei SC, Yu L, Yu GX, Zhang XL. REA: Robust cross-lingual entity alignment between knowledge graphs. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. CA: ACM, 2020. 2175–2184. [doi: 10.1145/3394486.3403268]

[68] Li SN, Li X, Ye R, Wang MZ, Su HP, Ou YZ. Non-translational alignment for multi-relational networks. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 4180−4186.

[69] Qi ZY, Zhang ZH, Chen JY, Chn X, Xiang YJ, Zhang NY, Zheng YF. Unsupervised knowledge graph alignment by probabilistic reasoning and semantic embedding. In: Proc. of the 30th Int'l Joint Conf. on Artificial Intelligence. Montreal: IJCAI.org, 2021. 2019–2025. [doi: 10.24963/ijcai.2021/278]

[70] Chen MH, Tian YT, Chang K W, Skiena S, Zaniolo C. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 3998−4004.

[71] Trisedya BD, Qi JZ, Zhang R. Entity alignment between knowledge graphs using attribute embeddings. Proc. of the AAAI Conf. on Artificial Intelligence, 2019, 33(1): 297–304. [doi: 10.1609/aaai.v33i01.3301297]

[72] Guo LB, Sun ZQ, Hu W. Learning to exploit long-term relational dependencies in knowledge graphs. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 2505–2514.

[73] Yang HW, Zou YY, Shi P, Lu W, Lin J, Sun X. Aligning cross-lingual entities with multi-aspect information. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: ACL, 2019. 4431–4441. [doi: 10.18653/v1/D19-1451]

[74] Tang XB, Zhang J, Chen B, Yang Y, Chen H, Li CP. BERT-INT: A BERT-based interaction model for knowledge graph alignment. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: IJCAI.org, 2020. 3174–3180. [doi: 10.24963/ijcai.2020/439]

[75] Xin KX, Sun ZQ, Hua W, Hu W, Zhou XF. Informed multi-context entity alignment. In: Proc. of the 15th ACM Int'l Conf. on Web Search and Data Mining. Tempe: ACM, 2022. 1197–1205. [doi: 10.1145/3488560.3498523]

[76] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[77] Zhang QH, Sun ZQ, Hu W, Chen MH, Guo LB, Qu YZ. Multi-view knowledge graph embedding for entity alignment. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 5429–5435. [doi: 10.24963/ijcai.2019/754]

[78] Liu ZY, Cao YX, Pan LM, Li JZ, Liu ZY, Chua TS. Exploring and evaluating attributes, values, and structures for entity alignment. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 6355–6364. [doi: 10.18653/v1/2020.emnlp-main.515]

[79] Flach PA, Savnik I. Database dependency discovery: A machine learning approach. AI Communications, 1999, 12(3): 139–160.

[80] Sun ZQ, Hu W, Li CK. Cross-lingual entity alignment via joint attribute-preserving embedding. In: Proc. of the 16th Int'l Semantic Web Conf. Vienna: Springer, 2017. 628−644. [doi: 10.1007/978-3-319-68288-4_37]

[81] Liu FY, Chen MH, Roth D, Collier N. Visual pivoting for (unsupervised) entity alignment. Proc. of the AAAI Conf. on Artificial Intelligence, 2021, 35(5): 4257–4266. [doi: 10.1609/aaai.v35i5.16550]

[82] Zhu Y, Liu HZ, Wu ZH, Du YP. Relation-aware neighborhood matching model for entity alignment. Proc. of the AAAI Conf. on Artificial Intelligence, 2021, 35(5): 4749–4756. [doi: 10.1609/aaai.v35i5.16606]

[83] Chen MH, Shi WJ, Zhou B, Roth D. Cross-lingual entity alignment with incidental supervision. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics. ACL, 2021. 645–658. [doi: 10.18653/v1/2021.eacl-main.53]

[84] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2019.

[85] Yang K, Liu SQ, Zhao JF, Wang YS, Xie B. COTSAE: Co-training of structure and attribute embeddings for entity alignment. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(3): 3025–3032. [doi: 10.1609/aaai.v34i03.5696]

[86] Mao X, Wang WT, Xu HM, Lan M, Wu YB. MRAEA: An efficient and robust entity alignment approach for cross-lingual knowledge graph. In: Proc. of the 13th Int'l Conf. on Web Search and Data Mining. Houston: ACM, 2020. 420–428. [doi: 10.1145/3336191.3371804]

[87] Gao YJ, Liu XZ, Wu JY, Li TY, Wang PF, Chen L. ClusterEA: Scalable entity alignment with stochastic training and normalized mini-

batch similarities. In: Proc. of the 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2022. 421–431. [doi: 10.1145/3534678.3539331]

[88] Liu B, Scells H, Zuccon G, Hua W, Zhao GH. ActiveEA: Active learning for neural entity alignment. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 3364–3374. [doi: 10.18653/v1/2021.emnlp-main.270]

[89] Ge CC, Liu XZ, Chen L, Zheng BH, Gao YJ. Make it easy: An effective end-to-end entity alignment framework. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2021. 777–786. [doi: 10.1145/3404835.3462870]

[90] Mao X, Wang WT, Wu YB, Lan M. Are negative samples necessary in entity alignment?: An approach with high performance, scalability and robustness. In: Proc. of the 30th ACM Int'l Conf. on Information and Knowledge Management. Queensland: ACM, 2021. 1263–1273. [doi: 10.1145/3459637.3482232]

[91] Mao X, Wang WT, Wu YB, Lan M. From alignment to assignment: Frustratingly simple unsupervised entity alignment. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 2843–2853. [doi: 10.18653/v1/2021.emnlp-main.226]

[92] Zhao X, Zeng WX, Tang JY, Wang W, Suchanek FM. An experimental study of state-of-the-art entity alignment approaches. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(6): 2610–2625. [doi: 10.1109/TKDE.2020.3018741]

[93] Chen B, Zhang J, Tang XB, Chen H, Li CP. JarKA: Modeling attribute interactions for cross-lingual knowledge alignment. In: Proc. of the 24th Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Singapore: Springer. 2020. 845–856. [doi: 10.1007/978-3-030-47426-3_65]

[94] Mao X, Wang WT, Xu HM, Wu YB, Lan M. Relational reflection entity alignment. In: Proc. of the 29th ACM Int'l Conf. on Information and Knowledge Management. Ireland: ACM, 2020. 1095–1104. [doi: 10.1145/3340531.3412001]

[95] Ge CC, Liu XZ, Chen L, Gao YJ, Zheng BH. LargeEA: Aligning entities for large-scale knowledge graphs. Proc. of the VLDB Endowment, 2021, 15(2): 237–245. [doi: 10.14778/3489496.3489504]

[96] Zeng WX, Zhao X, Li XY, Tang JY, Wang W. On entity alignment at scale. The VLDB Journal, 2022, 31(5): 1009–1033. [doi: 10.1007/s00778-021-00703-3]

[97] Sun ZQ, Chen MH, and Hu W. Knowing the no-match: Entity alignment with dangling cases. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. ACL, 2021. 3582–3593. [doi: 10.18653/v1/2021.acl-long.278]

[98] Luo SX, Yu S. An accurate unsupervised method for joint entity alignment and dangling entity detection. In: Proc. of the 2022 Findings of the Association for Computational Linguistics. Dublin: ACL, 2022. 2330–2339. [doi: 10.18653/v1/2022.findings-acl.183]

[99] Pei SC, Yu L, Zhang XL. Improving cross-lingual entity alignment via optimal transport. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 3231–3237. [doi: 10.24963/ijcai.2019/448]

[100] Sun ZQ, Huang JC, Hu W, Chen MH, Guo LB, Qu YZ. TransEdge: Translating relation-contextualized embeddings for knowledge graphs. In: Proc. of the 18th Int'l Semantic Web Conf. Auckland: Springer, 2019. 612–629. [doi: 10.1007/978-3-030-30793-6_35]

[101] Shi XF, Xiao YH. Modeling multi-mapping relations for precise cross-lingual entity alignment. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: ACL, 2019. 813–822. [doi: 10.18653/v1/D19-1075]

[102] Lin XX, Yang H, Wu J, Zhou C, Wang B. Guiding cross-lingual entity alignment via adversarial knowledge embedding. In: Proc. of the 2019 IEEE Int'l Conf. on Data Mining. Beijing: IEEE, 2019. 429–438. [doi: 10.1109/ICDM.2019.00053]

[103] Zhu QN, Zhou XF, Wu J, Tan JL, Guo L. Neighborhood-aware attentional representation for multilingual knowledge graphs. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 1943–1949. [doi: 10.24963/ijcai.2019/269]

[104] Ye R, Li X, Fang YJ, Zang HY, Wang MZ. A vectorized relational graph convolutional network for multi-relational network alignment. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 4135–4141. [doi: 10.24963/ijcai.2019/574]

[105] Li CJ, Cao YX, Hou L, Shi JX, Li JZ, Chua TS. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: ACL, 2019. 2723–2732. [doi: 10.18653/v1/D19-1274]

[106] Sun ZQ, Wang CM, Hu W, Chen MH, Dai J, Zhang W, Qu YZ. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(1): 222–229. [doi: 10.1609/aaai.v34i01.5354]

[107] Nie H, Han XP, Sun L, Wong CM, Chen Q, Wu SH, Zhang W. Global structure and local semantics-preserved embeddings for entity alignment. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: IJCAI.org, 2020. 3658–3664. [doi: 10.24963/ijcai.2020/502]

[108] Chen J, Li ZX, Zhao PP, Liu A, Zhao L, Chen ZG, Zhang XL. Learning short-term differences and long-term dependencies for entity alignment. In: Proc. of the 19th Int'l Semantic Web Conf. Athens: Springer, 2020. 92–109. [doi: 10.1007/978-3-030-62419-4_6]

[109] Yu DH, Yang YM, Zhang RH, Wu YX. Knowledge embedding based graph convolutional network. In: Proc. of the 2021 Web Conf. Ljubljana: ACM, 2021. 1619–1628. [doi: 10.1145/3442381.3449925]

[110] Zeng WX, Zhao X, Tang JY, Fan CJ. Reinforced active entity alignment. In: Proc. of the 30th ACM Int'l Conf. on Information and Knowledge Management. Queensland: ACM, 2021. 2477–2486. [doi: 10.1145/3459637.3482472]

[111] Xu CJ, Su FL, Lehmann J. Time-aware graph neural network for entity alignment between temporal knowledge graphs. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 8999–9010. [doi: 10.18653/v1/2021.emnlp-main.709]

[112] Trivedi R, Sisman B, Dong XL, Faloutsos C, Ma J, Zha HY. LinkNBed: Multi-graph representation learning with entity linkage. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 252–262. [doi: 10.18653/v1/P18-1024]

[113] Zhu Q, Wei H, Sisman B, Zheng D, Faloutsos C, Dong XL, Han JW. Collective multi-type entity alignment between knowledge graphs. In: Proc. of the 2020 Web Conf. Taipei: ACM, 2020. 2241–2252. [doi: 10.1145/3366423.3380289]

[114] Wang ZC, Yang JJ, Ye XJ. Knowledge graph alignment with entity-pair embedding. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. ACL, 2020. 1672–1680. [doi: 10.18653/v1/2020.emnlp-main.130]

[115] Zhu RB, Ma M, Wang P. RAGA: Relation-aware graph attention networks for global entity alignment. In: Proc. of the 25th Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Switzerland: Springer, 2021. 501–513. [doi: 10.1007/978-3-030-75762-5_40]

[116] Xu K, Wang LW, Yu M, Feng YS, Song Y, Wang ZG, Yu D. Cross-lingual knowledge graph alignment via graph matching neural network. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 3156–3161. [doi: 10.18653/v1/P19-1304]

[117] Wu YT, Liu X, Feng YS, Wang Z, Yan R, Zhao DY. Relation-aware entity alignment for heterogeneous knowledge graphs. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 5278–5284. [doi: 10.24963/ijcai.2019/733]

[118] Wu YT, Liu X, Feng YS, Wang Z, Zhao DY. Jointly learning entity and relation representations for entity alignment. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: ACL, 2019. 240–249. [doi: 10.18653/v1/D19-1023]

[119] Xu K, Song LF, Feng YS, Song Y, Yu D. Coordinated reasoning for cross-lingual knowledge graph alignment. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(5): 9354–9361. [doi: 10.1609/aaai.v34i05.6476]

[120] Zeng WX, Zhao X, Tang JY, Lin XM. Collective entity alignment via adaptive features. In: Proc. of the 36th IEEE Int'l Conf. on Data Engineering. Dallas: IEEE, 2020. 1870–1873. [doi: 10.1109/ICDE48307.2020.00191]

[121] Zeng WX, Zhao X, Tang JY, Lin XM, Groth P. Reinforcement learning-based collective entity alignment with adaptive features. ACM Trans. on Information Systems, 2021, 39(3): 26. [doi: 10.1145/3446428]

[122] Wu YT, Liu X, Feng YS, Wang Z, Zhao DY. Neighborhood matching network for entity alignment. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 6477–6487. [doi: 10.18653/v1/2020.acl-main.578]

[123] Yang JZ, Zhou W, Wei LW, Lin JY, Han JZ, Hu SL. RE-GCN: Relation enhanced graph convolutional network for entity alignment in heterogeneous knowledge graphs. In: Proc. of the 25th Int'l Conf. on Database Systems for Advanced Applications. Jeju: Springer, 2020. 432–447. [doi: 10.1007/978-3-030-59416-9_26]

[124] Yan YC, Liu LH, Ban YK, Jing BY, Tong HH. Dynamic knowledge graph alignment. Proc. of the AAAI Conf. on Artificial Intelligence, 2021, 35(5): 4564–4572. [doi: 10.1609/aaai.v35i5.16585]

[125] Mao X, Wang WT, Wu YB, Lan M. Boosting the speed of entity alignment 10×: Dual attention matching network with normalized hard sample mining. In: Proc. of the 2021 Web Conf. Ljubljana: ACM, 2021. 821–832. [doi: 10.1145/3442381.3449897]

[126] Yang JZ, Wang D, Zhou W, Qian WH, Wang X, Han JZ, Hu SL. Entity and relation matching consensus for entity alignment. In: Proc. of the 30th ACM Int'l Conf. on Information and Knowledge Management. Queensland: ACM, 2021. 2331–2341. [doi: 10.1145/3459637.3482338]

[127] Liu X, Hong HY, Wang XH, Chen ZY, Kharlamov E, Dong YX, Tang J. SelfKG: Self-supervised entity alignment in knowledge graphs. In: Proc. of the 2022 ACM Web Conf. Lyon: ACM, 2022. 860–870. [doi: 10.1145/3485447.3511945]

[128] Ge CC, Zeng XC, Chen L, Gao YJ. ZeroMatcher: A cost-off entity matching system. In: Proc. of the 45th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Madrid: ACM, 2022, 3262–3266. [doi: 10.1145/3477495.3531661]

[129] Pasternack J, Roth D. Knowing what to believe (when you already know something). In: Proc. of the 23rd Int'l Conf. on Computational Linguistics. Beijing: ACL, 2010. 877–885.

[130] Galland A, Abiteboul S, Marian A, Senellart P. Corroborating information from disagreeing views. In: Proc. of the 3rd ACM Int'l Conf. on Web Search and Data Mining. New York: ACM, 2010. 131–140. [doi: 10.1145/1718487.1718504]

[131] Yin XX, Han JW, Yu PS. Truth discovery with multiple conflicting information providers on the Web. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Jose: ACM, 2007. 1048–1052. [doi: 10.1145/1281192.1281309]

[132] Dong XL, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence. Proc. of the VLDB Endowment, 2009, 2(1): 550–561. [doi: 10.14778/1687627.1687690]

[133] Yin XX, Tan WZ. Semi-supervised truth discovery. In: Proc. of the 20th Int'l Conf. on World Wide Web. Hyderabad: ACM, 2011. 217–226. [doi: 10.1145/1963405.1963439]

[134] Li Q, Li YL, Gao J, Su L, Zhao B, Demirbas M, Fan W, Han JW. A confidence-aware approach for truth discovery on long-tail data. Proc. of the VLDB Endowment, 2014, 8(4): 425–436. [doi: 10.14778/2735496.2735505]

[135] Li Q, Li YL, Gao J, Zhao B, Fan W, Han JW. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. Snowbird: ACM, 2014. 1187–1198. [doi: 10.1145/2588555.2610509]

[136] Zhao B, Han JW. A probabilistic model for estimating real-valued truth from conflicting sources. In: Proc. of the 10th Int'l Workshop on Quality in Databases. Istanbul, 2012. 1817.

[137] Pasternack J, Roth D. Latent credibility analysis. In: Proc. of the 22nd Int'l Conf. on World Wide Web. Rio de Janeiro: ACM, 2013. 1009–1020. [doi: 10.1145/2488388.2488476]

[138] Dong XL, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world. Proc. of the VLDB Endowment, 2009, 2(1): 562–573. [doi: 10.14778/1687627.1687691]

[139] Rekatsinas T, Joglekar M, Garcia-Molina H, Parameswaran A, Ré C. SLiMFast: Guaranteed results for data fusion and source reliability. In: Proc. of the 2017 ACM Int'l Conf. on Management of Data. Chicago: ACM, 2017. 1399–1414. [doi: 10.1145/3035918.3035951]

[140] Li YL, Li Q, Gao J, Su L, Zhao B, Fan W, Han JW. On the discovery of evolving truth. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Sydney: ACM, 2015. 675–684. [doi: 10.1145/2783258.2783277]

[141] Pochampally R, Das Sarma A, Dong XL, Meliou A, Srivastava D. Fusing data with correlations. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. Snowbird: ACM, 2014. 433–444. [doi: 10.1145/2588555.2593674]

[142] Qi GJ, Aggarwal CC, Han JW, Huang T. Mining collective intelligence in diverse groups. In: Proc. of the 22nd Int'l Conf. on World Wide Web. Rio de Janeiro: ACM, 2013. 1041–1052. [doi: 10.1145/2488388.2488479]

[143] Sarma AD, Dong XL, Halevy A. Data integration with dependent sources. In: Proc. of the 14th Int'l Conf. on Extending Database Technology. Uppsala: ACM, 2011. 401–412. [doi: 10.1145/1951365.1951414]

[144] Zhao B, Rubinstein BIP, Gemmell J, Han JW. A Bayesian approach to discovering truth from conflicting sources for data integration. Proc. of the VLDB Endowment, 2012, 5(6): 550–561. [doi: 10.14778/2168651.2168656]

[145] Li YL, Gao J, Meng CS, Li Q, Su L, Zhao B, Fan W, Han JW. A survey on truth discovery. ACM SIGKDD Explorations Newsletter, 2016, 17(2): 1–16. [doi: 10.1145/2897350.2897352]

[146] Cao EM, Wang DF, Huang JC, Hu W. Open knowledge enrichment for long-tail entities. In: Proc. of the 2020 Web Conf. Taipei: ACM, 2020. 384–394. [doi: 10.1145/3366423.3380123]

[147] Huang JC, Zhao Y, Hu W, Ning Z, Chen QJ, Qiu XX, Huo CF, Ren WJ. Trustworthy knowledge graph completion based on multi-sourced noisy data. In: Proc. of the 2022 ACM Web Conf. Lyon: ACM, 2022. 956–965. [doi: 10.1145/3485447.3511938]

[148] Dong XL, Gabrilovich E, Heitz G, Horn W, Murphy K. From data fusion to knowledge fusion. Proc. of the VLDB Endowment, 2014, 7(10): 881–892. [doi: 10.14778/2732951.2732962]

[149] Swartz N. Gartner warns firms of 'dirty data'. The Information Management Journal, 2007, 41(3): 6–7.

[150] Redman TC. Bad data costs the U.S. $3 trillion per year. Harvard Business Review, 2016, 22: 11–18.

[151] Chu X, Morcos J, Ilyas IF, Ouzzani M, Papotti P, Tang N, Ye Y. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. Victoria: ACM, 2015. 1247–1261. [doi: 10.1145/2723372.2749431]

[152] Asuncion A, Newman DJ. UCI machine learning repository. 2007. https://archive.ics.uci.edu/ml/index.php

[153] Huhtala Y, Kärkkäinen J, Porkka P, Toivonen H. TANE: An efficient algorithm for discovering functional and approximate dependencies. The Computer Journal, 1999, 42(2): 100–111. [doi: 10.1093/comjnl/42.2.100]

[154] Liu JX, Li JY, Liu CF, Chen YF. Discover dependencies from data—A review. IEEE Trans. on Knowledge and Data Engineering, 2012, 24(2): 251–264. [doi: 10.1109/TKDE.2010.197]

[155] Lopes S, Petit JM, Lakhal L. Efficient discovery of functional dependencies and armstrong relations. In: Proc. of the 7th Int'l Conf. on Extending Database Technology: Advances in Database Technology. Konstanz: Springer, 2000. 350–364. [doi: 10.1007/3-540-46439-5_24]

[156] Novelli N, Cicchetti R. FUN: An efficient algorithm for mining functional and embedded dependencies. In: Proc. of the 8th Int'l Conf. on Database Theory. London: Springer, 2001. 189–203. [doi: 10.1007/3-540-44503-X_13]

[157] Papenbrock T, Ehrlich J, Marten J, Neubert T. Functional dependency discovery: An experimental evaluation of seven algorithms. Proc. of the VLDB Endowment, 2015, 8(10): 1082–1093. [doi: 10.14778/2794367.2794377]

[158] Chiang F, Miller RJ. Discovering data quality rules. Proc. of the VLDB Endowment, 2008, 1(1): 1166–1177. [doi: 10.14778/1453856.1453980]

[159] Fan WF, Geerts F, Li JZ, Xiong M. Discovering conditional functional dependencies. IEEE Trans. on Knowledge and Data Engineering, 2011, 23(5): 683–698. [doi: 10.1109/TKDE.2010.154]

[160] Rammelaere J, Geerts F. Revisiting conditional functional dependency discovery: Splitting the "C" from the "FD". In: Proc. of the 2018 European Conf. on Machine Learning and Knowledge Discovery in Databases. Dublin: Springer, 2018. 552–568. [doi: 10.1007/978-3-030-10928-8_33]

[161] Bleifuß T, Kruse S, Naumann F. Efficient denial constraint discovery with hydra. Proc. of the VLDB Endowment, 2017, 11(3): 311–323. [doi: 10.14778/3157794.3157800]

[162] Chu X, Ilyas IF, Krishnan S, Wang JN. Data cleaning: Overview and emerging challenges. In: Proc. of the 2016 Int'l Conf. on Management of Data. San Francisco: ACM, 2016. 2201–2206. [doi: 10.1145/2882903.2912574]

[163] Pena EHM, de Almeida EC. BFASTDC: A bitwise algorithm for mining denial constraints. In: Proc. of the 29th Int'l Conf. on Database and Expert Systems Applications. Regensburg: Springer, 2018. 53–68. [doi: 10.1007/978-3-319-98809-2_4]

[164] Pena EHM, de Almeida EC, Naumann F. Discovery of approximate (and exact) denial constraints. Proc. of the VLDB Endowment, 2019, 13(3): 266–278. [doi: 10.14778/3368289.3368293]

[165] Livshits E, Heidari A, Ilyas IF, Kimelfeld B. Approximate denial constraints. Proc. of the VLDB Endowment, 2020, 13(10): 1682–1695. [doi: 10.14778/3401960.3401966]

[166] Bertossi L, Bravo L, Franconi E, Lopatenko A. The complexity and approximation of fixing numerical attributes in databases under integrity constraints. Information Systems, 2008, 33(4–5): 407–434. [doi: 10.1016/j.is.2008.01.005]

[167] Beskales G, Ilyas IF, Golab L. Sampling the repairs of functional dependency violations under hard constraints. Proc. of the VLDB Endowment, 2010, 3(1–2): 197–207. [doi: 10.14778/1920841.1920870]

[168] Fan WF, Ma S, Tang N, Yu WY. Interaction between record matching and data repairing. Journal of Data and Information Quality, 2014, 4(4): 16. [doi: 10.1145/2567657]

[169] Cong G, Fan WF, Geerts F, Jia XB, Ma S. Improving data quality: Consistency and accuracy. In: Proc. of the 33rd Int'l Conf. on Very Large Data Bases. Vienna: VLDB Endowment, 2007. 315–326.

[170] Fan WF, Geerts F, Jia XB, Kementsietsidis A. Conditional functional dependencies for capturing data inconsistencies. ACM Trans. on Database Systems, 2008, 33(2): 6. [doi: 10.1145/1366102.1366103]

[171] Yakout M, Elmagarmid AK, Neville J, Ouzzani M, Ilyas IF. Guided data repair. Proc. of the VLDB Endowment, 2011, 4(5): 279–289. [doi: 10.14778/1952376.1952378]

[172] Kolahi S, Lakshmanan LVS. On approximating optimum repairs for functional dependency violations. In: Proc. of the 12th Int'l Conf. on Database Theory. St. Petersburg: ACM, 2009. 53–62. [doi: 10.1145/1514894.1514901]

[173] Chu X, Ilyas IF, Papotti P. Holistic data cleaning: Putting violations into context. In: Proc. of the 29th IEEE Int'l Conf. on Data Engineering. Brisbane: IEEE, 2013. 458–469. [doi: 10.1109/ICDE.2013.6544847]

[174] Khayyat Z, Ilyas IF, Jindal A, Madden S, Ouzzani M, Papotti P, Quiané-Ruiz JA, Tang N, Yin S. BigDansing: A system for big data cleansing. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. Victoria: ACM, 2015. 1215–1230. [doi: 10.1145/2723372.2747646]

[175] Rekatsinas T, Chu X, Ilyas IF, Ré C. HoloClean: Holistic data repairs with probabilistic inference. Proc. of the VLDB Endowment, 2017, 10(11): 1190–1201. [doi: 10.14778/3137628.3137631]

[176] Geerts F, Mecca G, Papotti P, Santoro D. The LLUNATIC data-cleaning framework. Proc. of the VLDB Endowment, 2013, 6(9): 625–636. [doi: 10.14778/2536360.2536363]

[177] Arocena PC, Glavic B, Mecca G, Miller RJ. Messing up with BART: Error generation for evaluating data-cleaning algorithms. Proc. of the VLDB Endowment, 2015, 9(2): 36–47. [doi: 10.14778/2850578.2850579]

[178] Dallachiesa M, Ebaid A, Eldawy A, Elmagarmid A, Ilyas IF, Ouzzani M, Tang N. NADEEF: A commodity data cleaning system. In:

Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2013. 541–552. [doi: 10.1145/2463676.2465327]

[179] Ge CC, Gao YJ, Miao XY, Yao B, Wang HB. A hybrid data cleaning framework using Markov logic networks. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(5): 2048–2062. [doi: 10.1109/TKDE.2020.3012472]

[180] Ge CC, Gao YJ, Miao XY, Chen L. IHCS: An integrated hybrid cleaning system. Proc. of the VLDB Endowment, 2019, 12(12): 1874–1877. [doi: 10.14778/3352063.3352088]

[181] Giannakopoulou SA, Karpathiotakis M, Gaidioz B, Ailamaki A. CleanM: An optimizable query language for unified scale-out data cleaning. Proc. of the VLDB Endowment, 2017, 10(11): 1466–1477. [doi: 10.14778/3137628.3137654]

[182] Giannakopoulou S, Karpathiotakis M, Ailamaki A. Cleaning denial constraint violations through relaxation. In: Proc. of the 2020 ACM SIGMOD Int'l Conf. on Management of Data. Portland: ACM, 2020. 805–815. [doi: 10.1145/3318464.3389775]

[183] Yakout M, Berti-Équille L, Elmagarmid AK. Don't be scared: Use scalable automatic repairing with maximal likelihood and bounded changes. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2013. 553–564. [doi: 10.1145/2463676.2463706]

[184] Mayfield C, Neville J, Prabhakar S. ERACER: A database approach for statistical inference and data cleaning. In: Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data. Indianapolis: ACM, 2010. 75–86. [doi: 10.1145/1807167.1807178]

[185] Krishnan S, Wang JN, Wu E, Franklin MJ, Goldberg K. ActiveClean: Interactive data cleaning for statistical modeling. Proc. of the VLDB Endowment, 2016, 9(12): 948–959. [doi: 10.14778/2994509.2994514]

[186] Mahdavi M, Abedjan Z. Baran: Effective error correction via a unified context representation and transfer learning. Proc. of the VLDB Endowment, 2020, 13(12): 1948–1961. [doi: 10.14778/3407790.3407801]

[187] Heidari A, McGrath J, Ilyas IF, Rekatsinas T. HoloDetect: Few-shot learning for error detection. In: Proc. of the 2019 Int'l Conf. on Management of Data. Amsterdam: ACM, 2019. 829–846. [doi: 10.1145/3299869.3319888]

[188] Mahdavi M, Abedjan Z, Fernandez RC, Madden S, Ouzzani M, Stonebraker M, Tang N. Raha: A configuration-free error detection system. In: Proc. of the 2019 Int'l Conf. on Management of Data. Amsterdam: ACM, 2019. 865–882. [doi: 10.1145/3299869.3324956]

[189] Chu X, Ilyas IF, Papotti P. Discovering denial constraints. Proc. of the VLDB Endowment, 2013, 6(13): 1498–1509. [doi: 10.14778/2536258.2536262]

[190] Luo YY, Chai CL, Qin XD, Tang N, Li GL. VisClean: Interactive cleaning for progressive visualization. Proc. of the VLDB Endowment, 2020, 13(12): 2821–2824. [doi: 10.14778/3415478.3415484]

[191] Jia SB, Xiang Y, Chen XJ, Wang K, Shijia. Triple trustworthiness measurement for knowledge graph. In: Proc. of the 2019 World Wide Web Conf. San Francisco: ACM, 2019. 2865–2871. [doi: 10.1145/3308558.3313586]

[192] Fan WF, Fu WZ, Jin RC, Lu P, Tian C. Discovering association rules from big graphs. Proc. of the VLDB Endowment. 2022, 15(7): 1479–1492.

[193] Calvanese D, Fischl W, Pichler R, Sallinger E, Šimkus M. Capturing relational schemas and functional dependencies in RDFS. In: Proc. of the 28th Conf. on Artificial Intelligence. Québec: AAAI, 2014. 1003–1011.

[194] Lausen G, Meier M, Schmidt M. SPARQLing constraints for RDF. In: Proc. of the 11th Int'l Conf. on Extending Database Technology: Advances in Database Technology. Nantes: ACM, 2008. 499–509. [doi: 10.1145/1353343.1353404]

[195] Cortés-Calabuig A, Paredaens J. Semantics of constraints in RDFS. In: Proc. of the 6th Alberto Mendelzon Int'l Workshop on Foundations of Data Management. Ouro Preto: CEUR-WS.org, 2012. 75–90.

[196] Akhtar W, Cortés-Calabuig Á, Paredaens J. Constraints in RDF. In: Proc. of the 4th Int'l Workshop on Semantics in Data and Knowledge Bases. Bordeaux: Springer, 2010. 23–39. [doi: 10.1007/978-3-642-23441-5_2]

[197] Arioua A, Bonifati A. User-guided repairing of inconsistent knowledge bases. In: Proc. of the 21st Int'l Conf. on Extending Database Technology. Vienna: ACM, 2018. 133–144. [doi: 10.5441/002/edbt.2018.13]

[198] Yu Y, Heflin J. Extending functional dependency to detect abnormal data in RDF graphs. In: Proc. of the 10th Int'l Semantic Web Conf. Bonn: Springer, 2011. 794–809. [doi: 10.1007/978-3-642-25073-6_50]

[199] He BB, Zou L, Zhao DY. Using conditional functional dependency to discover abnormal data in RDF graphs. In: Proc. of the 2014 Semantic Web Information Management. Snowbird: ACM, 2014. 1–7. [doi: 10.1145/2630602.2630605]

[200] Fan WF, Wu YH, Xu JB. Functional dependencies for graphs. In: Proc. of the 2016 Int'l Conf. on Management of Data. San Francisco: ACM, 2016. 1843–1857. [doi: 10.1145/2882903.2915232]

[201] Fan WF, Fan Z, Tian C, Dong XL. Keys for graphs. Proc. of the VLDB Endowment, 2015, 8(12): 1590–1601. [doi: 10.14778/2824032.2824056]

[202] Cheng YR, Chen L, Yuan Y, Wang GR. Rule-based graph repairing: Semantic and efficient repairing methods. In: Proc. of the 34th

IEEE Int'l Conf. on Data Engineering. Paris: IEEE, 2018. 773–784. [doi: 10.1109/ICDE.2018.00075]

[203] Fan WF, Lu P, Tian C, Zhou JR. Deducing certain fixes to graphs. Proc. of the VLDB Endowment, 2019, 12(7): 752–765. [doi: 10.14778/3317315.3317318]

[204] Song SX, Liu BG, Cheng H, Yu JX, Chen L. Graph repairing under neighborhood constraints. The VLDB Journal, 2017, 26(5): 611–635. [doi: 10.1007/s00778-017-0466-5]

[205] Lin P, Song Q, Wu YH, Pi JX. Repairing entities using star constraints in multirelational graphs. In: Proc. of the 36th IEEE Int'l Conf. on Data Engineering. Dallas: IEEE, 2020. 229–240. [doi: 10.1109/ICDE48307.2020.00027]

[206] Song Q, Lin P, Ma HC, Wu YH. Explaining missing data in graphs: A constraint-based approach. In: Proc. of the 37th IEEE Int'l Conf. on Data Engineering. Chania: IEEE, 2021. 1476–1487. [doi: 10.1109/ICDE51399.2021.00131]

[207] Loster M, Mottin D, Papotti P, Ehmüller J, Feldmann B, Naumann F. Few-shot knowledge validation using rules. In: Proc. of the 2021 Web Conf. Ljubljana: ACM, 2021. 3314–3324. [doi: 10.1145/3442381.3450040]

[208] Wang Q, Mao ZD, Wang B, Guo L. Knowledge graph embedding: A survey of approaches and applications. IEEE Trans. on Knowledge and Data Engineering, 2017, 29(12): 2724–2743. [doi: 10.1109/TKDE.2017.2754499]

[209] Hassan N, Arslan F, Li CK, Tremayne M. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Halifax: ACM, 2017. 1803–1812. [doi: 10.1145/3097983.3098131]

[210] Socher R, Chen DQ, Manning CD, Ng AY. Reasoning with neural tensor networks for knowledge base completion. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 926–934.

[211] Wang Z, Zhang JW, Feng JL, Chen Z. Knowledge graph embedding by translating on hyperplanes. In: Proc. of the 28th AAAI Conf. on Artificial Intelligence. Québec: AAAI, 2014. 1112–1119. [doi: 10.1609/aaai.v28i1.8870]

[212] Nickel M, Tresp V, Kriegel HP. A three-way model for collective learning on multi-relational data. In: Proc. of the 28th Int'l Conf. on Machine Learning. Bellevue: Omnipress. 2011. 809–816.

[213] Xie RB, Liu ZY, Lin F, Lin LY. Does william shakespeare really write hamlet? Knowledge representation learning with confidence. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence and the 30th Innovative Applications of Artificial Intelligence Conf. and the 8th AAAI Symp. on Educational Advances in Artificial Intelligence. New Orleans: AAAI, 2018. 4954–4961. [doi: 10.1609/aaai.v32i1.11924]

[214] Bougiatiotis K, Fasoulis R, Aisopos F, Nentidis A, Paliouras G. Guiding graph embeddings using path-ranking methods for error detection innoisy knowledge graphs. arXiv:2002.08762, 2020.

[215] Vashishth S, Sanyal S, Nitin V, Talukdar P. Composition-based multi-relational graph convolutional networks. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[216] Dong X, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, Strohmann T, Sun SH, Zhang W. Knowledge vault: A Web-scale approach to probabilistic knowledge fusion. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2014. 601–610. [doi: 10.1145/2623330.2623623]

附中文参考文献:

[2] 中国信息通信研究院. 大数据白皮书. 2021. http://www.caict.ac.cn/kxyj/qwfb/bps/202112/t20211220_394300.htm

[3] 陈跃国, 王京春. 数据集成综述. 计算机科学, 2004, 31(5): 48–51. [doi: 10.3969/j.issn.1002-137X.2004.05.015]

[4] 杨先娣, 彭智勇, 刘君强, 李旭辉. 信息集成研究综述. 计算机科学, 2006, 33(7): 55–59, 80. [doi: 10.3969/j.issn.1002-137X.2006.07.015]

[5] 王淞, 彭煜玮, 兰海, 罗倩雯, 彭智勇. 数据集成方法发展与展望. 软件学报, 2020, 31(3): 893−908. http://www.jos.org.cn/1000-9825/5911.htm [doi: 10.13328/j.cnki.jos.005911]

[8] 庄严, 李国良, 冯建华. 知识库实体对齐技术综述. 计算机研究与发展, 2016, 53(1): 165–192. [doi: 10.7544/issn1000-1239.2016.20150661]

[9] 孟小峰, 杜治娟. 大数据融合研究: 问题与挑战. 计算机研究与发展, 2016, 53(2): 231–246. [doi: 10.7544/issn1000-1239.2016.20150874]

[10] 郭志懋, 周傲英. 数据质量和数据清洗研究综述. 软件学报, 2002, 13(11): 2076−2082. http://www.jos.org.cn/jos/article/abstract/20021103?st=search [doi: 10.13328/j.cnki.jos.2002.11.003]

[11] 郝爽, 李国良, 冯建华, 王宁. 结构化数据清洗技术综述. 清华大学学报(自然科学版), 2018, 58(12): 1037–1050. [doi: 10.16511/j.cnki.qhdxxb.2018.22.053]

[12] 王鑫, 邹磊, 王朝坤, 彭鹏, 冯志勇. 知识图谱数据管理研究综述. 软件学报, 2019, 30(7): 2139−2174. http://www.jos.org.cn/1000-9825/5841.htm [doi: 10.13328/j.cnki.jos.005841]

高云君(1977－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为数据库, 大数据管理与分析, DB 与 AI 融合.

郭宇翔(1998－), 男, 博士生, 主要研究领域为数据集成, 数据准备.

葛丛丛(1995－), 女, 博士, 主要研究领域为数据集成, 数据治理.

陈璐(1989－), 女, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为数据库, 大数据处理, 度量空间数据管理.