doi:10.11887/j.issn.1001-2486.26020019 http://journal.nudt.edu.cn

数据工程智能化方法研究综述

张海粟[1,2],王 龙[1,2],颜登程[3*],王晓乐[3],李向朋[1,2]
(1. 信息支援部队工程大学,湖北 武汉 430010;2. 数据智能湖北省重点实验室,湖北 武汉 430010;
3. 安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:在数据工程中,通过应用人工智能方法,提高对现实世界中多源、异构、高噪声的大规模原始数据的处理效果,已经成为当前研究热点。基于数据工程的总体研究框架,按照数据清洗、数据连接、数据发现三个关键环节的设计,系统梳理了数据工程智能化方法的最新研究进展,详细分析了每个关键环节智能化方法的思路和效果。进一步地,结合智能技术发展趋势对其在数据工程领域的未来研究给出了展望。

关键词:数据工程;智能化方法;数据清洗;数据连接;数据发现
中图分类号:TP311.13 **文献标志码**:A **文章编号**:1001-2486(2026)03-211-17

Review of intelligent methods for data engineering

ZHANG Haisu[1,2], *WANG Long*[1,2], *YAN Dengcheng*[3*], *WANG Xiaole*[3], *LI Xiangpeng*[1,2]
(1. Information Support Force Engineering University, Wuhan 430010, China;
2. Hubei Provincial Key Laboratory of Data Intelligence, Wuhan 430010, China;
3. School of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: In data engineering, the enhancement of processing effectiveness for real-world massive raw data, characterized by being multi-source, heterogeneous, and high-noise, via the application of artificial intelligence methods is currently regarded as a research hotspot. Based on the general research framework of data engineering, the latest research progress in intelligent data engineering methods was systematically reviewed in accordance with the design of three key stages: data cleaning, data linking, and data discovery. Additionally, the principles and effectiveness of the methods related to each key stage were analyzed in detail. Furthermore, an outlook on future research in data engineering is provided in combination with the development trends of intelligence.

Keywords: data engineering; intelligent methods; data cleaning; data linking; data discovery

现实世界中产生的大规模原始数据普遍表现出多源、异构、高噪声等复杂特征。数据工程作为统一多源数据、构建数据湖、支撑数据驱动系统的核心技术,通过对多源、异构、高噪声原始数据进行清洗、连接、发现等处理,支撑数据从原始形态转化为有用形态,形成可用的高质量数据集。

但是,传统人工主导、规则驱动的数据工程方法,通过手工构建数据处理管道,基于静态规则设计清洗策略并进行实体对齐与记录融合,在处理大数据时,面临着高成本、低效率、难以应对频繁变化的数据处理业务需求等巨大挑战。

近年来,人工智能特别是大模型技术的持续发展,为破解传统数据工程瓶颈提供了新的技术路径和契机。数据工程智能化方法的核心是学习驱动的数据理解与处理机制,实现对数据清洗、连接、发现等环节的自动化,并能在需要时与大数据计算、可视化分析等外围平台协同,以增强数据处理能力与系统适配性。从传统范式向智能主导、大模型驱动的新范式演进,既减少了人工参与,又显著提升了数据处理准确性、鲁棒性,特别是增强了跨场景迁移、拓宽数据处理范围和数据应用边界的能力。

数据工程智能化方法领域已有部分综述工

收稿日期:2026-02-10
基金项目:国家部委基金资助项目(2024QD00400)
第一作者:张海粟(1982—),男,安徽巢湖人,教授,博士,博士生导师,E-mail:zhanghaisu@nudt.edu.cn
***通信作者**:颜登程(1987—),男,湖北襄阳人,副教授,博士,博士生导师,E-mail:yanzhou@ahu.edu.cn
引用格式:张海粟,王龙,颜登程,等. 数据工程智能化方法研究综述[J]. 国防科技大学学报, 2026, 48(3): 211-227.
Citation: ZHANG H S, WANG L, YAN D C, et al. Review of intelligent methods for data engineering[J]. Journal of National University of Defense Technology, 2026, 48(3): 211-227.

作。例如,Paton 等[1]主要聚焦于数据发现任务,对数据搜索、数据导航、数据注释及模式推断等方向的相关方法进行了系统总结;Zhu 等[2]和郭志懋等[3]围绕数据清洗问题,对数据错误类型、问题定义及典型解决方法进行了梳理与分析;Zhou 等[4]则关注大模型在数据处理流程中的应用潜力,探讨了其在数据处理、存储与分析等环节中的赋能方式。总体而言,现有综述工作多从单一任务或单一技术视角对数据工程相关研究进行总结,对贯穿数据工程多个关键环节的综合性分析仍相对有限。在上述研究基础上,本文从数据工程核心流程出发,聚焦数据清洗、数据连接与数据发现三个关键环节。这三个任务直接面向多源异构数据的质量提升与语义整合问题,是数据工程中最典型的数据理解与数据处理任务,也是近年来人工智能技术应用最为活跃的研究方向。因此,本文从这三个任务视角出发,提出一种贯穿多个任务阶段的综合综述视角。同时,本文系统梳理了从传统机器学习方法到大模型方法在上述任务中的研究进展,构建了数据工程智能化方法的技术分类体系,并总结了各类方法的主要特点及其在常用数据集上的实验表现。通过上述工作,本文旨在为研究者提供对数据工程智能化研究现状的系统性认识,并为相关方法的比较分析与应用选择提供参考依据。

1 数据工程总体研究框架

数据工程总体研究框架及其智能化方法如图 1 所示。其中,数据清洗,主要针对源数据中的缺失、错误及冗余问题,研究基于表示学习和大模型等技术实现错误检测、数据修复、数据补全等方法以提高数据质量。数据连接,主要针对多源、异构数据集之间语义关系模糊,难以实现数据集成的问题,研究基于人工智能的数据模式匹配和实体匹配方法,以实现数据集间语义的对齐与关联。数据发现,主要针对海量数据集难以组织、搜索困难的问题,研究实现数据搜索、数据导航和数据注释等方法,促进跨数据集之间的主题汇集并提高对大数据的组织和运用能力。

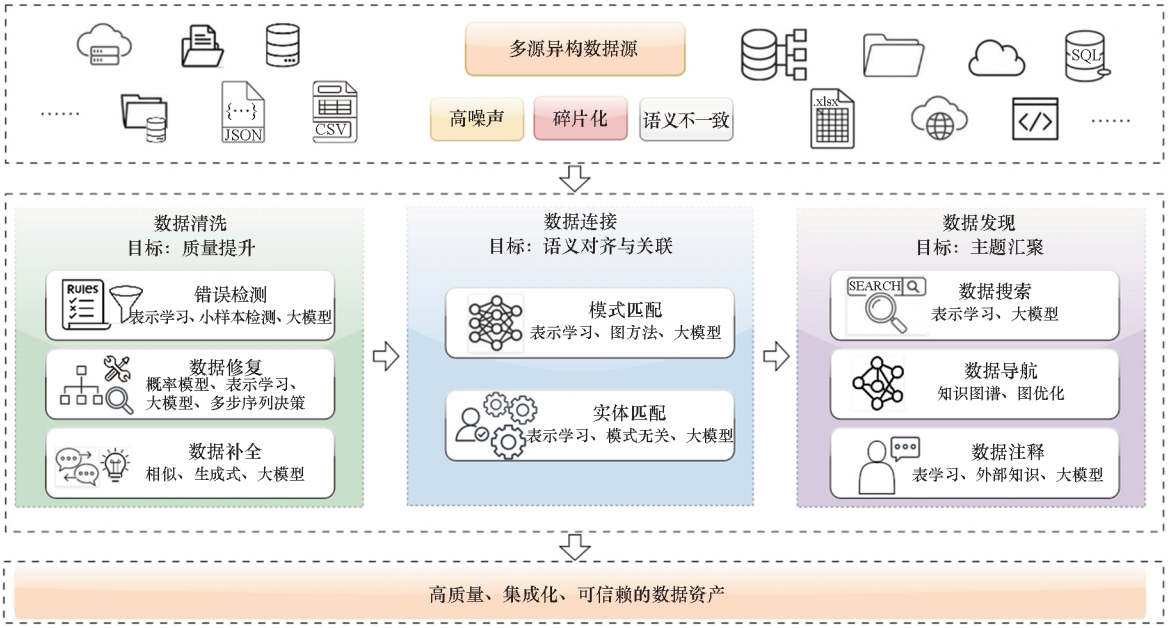

图 1 数据工程总体研究框架及其智能化方法

Fig. 1 General research framework for data engineering and its intelligent methods

2 智能化数据清洗方法

数据清洗的目标是识别出数据集中值缺失、值异常和约束冲突等错误问题,并进行修复与补全以提升数据质量。数据清洗如图 2 所示。智能化数据清洗方法,通过机器学习、深度学习和大模型等技术对数据错误模式和数据间潜在关系进行自动学习与推理,完成数据中的错误检测、数据修复与数据补全,实现数据清洗由传统人工规则范式向潜在规则自动学习与推理的范式转变,提高了数据清洗效率,增强了数据清洗方法在不同场景下的迁移性。数据清洗方法研究常在表 1 所示的各类数据集上进行验证。

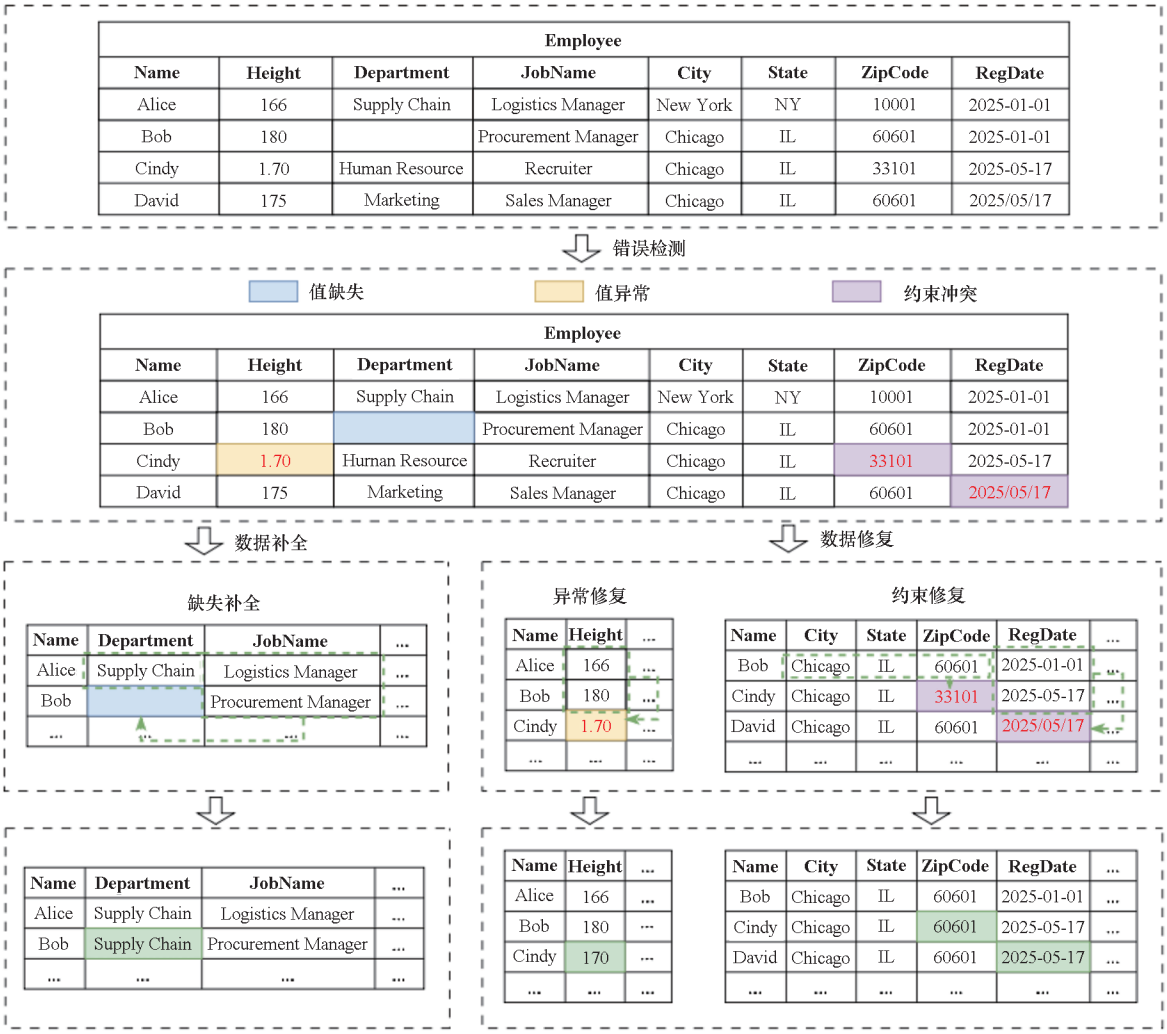

图 2　数据清洗示意图

Fig. 2　Schematic diagram of data cleaning

表 1　数据清洗研究常用数据集

Tab. 1　Data cleaning research common datasets

数据集	行数	列数	主题	测试场景
EEG[5]	14 980	15	神经科学	错误检测
Adult[6]	48 842	15	社会科学	错误检测 数据补全
Credit[7]	30 000	24	商业	数据补全
Animal[8]	26 729	10	动物	错误检测 数据修复 数据补全
Movie[9]	9 329	7	电影	
Nashville[10]	211 945	22	地理信息	错误检测 数据修复
Titanic[11]	891	12	生存预测	
Beers[12]	2 410	8	啤酒	
Cancer[13]	3 048	34	公共卫生	错误检测 数据修复 数据补全

2.1　错误检测

错误检测旨在识别数据集中存在的值缺失、值异常和约束冲突等问题,是数据清洗流程中的关键环节。随着数据规模和复杂度不断提升,错误检测面临错误类型多样、标注数据稀缺以及应用场景差异显著等挑战。现有研究总体呈现从依赖数据分布学习的传统方法,到缓解标注不足的小样本学习方法,再到利用大模型语义理解与推理能力的智能化方法的发展趋势。基于此,当前智能化错误检测研究主要可分为基于学习的错误检测、基于小样本的错误检测以及融合大模型的错误检测三类。

1)基于学习的错误检测。真实世界中数据错误类型多种多样,不同字段的数据约束条件不同,数据违反约束条件的错误也不同。基于学习的错误检测方法,主要通过监督或无监督学习方式去发现数据分布规律,实现对违反约束条件或值异常等错误的检测。对于数据违反约束条件的

情况，通常基于学习样本集，构建机器学习模型归纳出约束条件，再检测违反约束条件的错误，典型的研究包括：SCODED[14] 通过概率依赖引入列间统计约束条件，Uni-Detect[15]、Auto-Test[16]、杜岳峰等[17]和 Guardrail[18] 则自动学习约束条件。对于值异常检测的情况，Dai 等[19]以无监督方式，采用深度神经网络方法学习正确数据的正常分布，进而通过比较检出分布异常的数据。基于学习的错误检测，对自动高效地探查数据中复杂的潜在错误起到重要作用。

2）基于小样本的错误检测。基于监督学习的错误检测常面临训练数据不足的问题，多种基于小样本的错误检测方法被提出。HoloDetect[20] 采用深度神经网络进行小样本训练实现错误检测；SAGED[21] 和 ZeroED[22] 利用大模型零样本推理能力，对少量错误样本进行学习并进行错误检测推理，降低了对大规模数据标注的要求，提升错误检测的效率。

3）融合大模型的错误检测。依托大模型语义理解能力和推理能力，近来研究开始以大模型为基座，设计融合大模型的错误检测方法。通过提示词工程方式使大模型对数据样本进行理解，并推理生成约束规则，思路如图 3 所示。Narayan 等[23]将数据清洗任务转换成提示词任务，并基于上下文学习引入示例样本，通过大模型进行错误检测推理。GIDCL[24] 利用大模型自动生成具有可解释性的清洗规则并指导特征工程，通过对错误检测模型进行迭代优化，实现可靠且高效的错误检测。为了进一步增强大模型对业务事实的理解，RetClean[25] 将错误检测建模为“检索 - 匹配”的验证过程，通过引入外部知识增强大模型对业务事实一致性的验证能力，大模型根据外部知识库中检索到的支持性证据数据进行错误检测。总体而言，融合大模型的错误检测从语义层面去理解和推理数据中存在的错误，可以方便地引入业务事实，并生成可解释的清洗规则，可以很好地提高数据错误检测的可靠性。

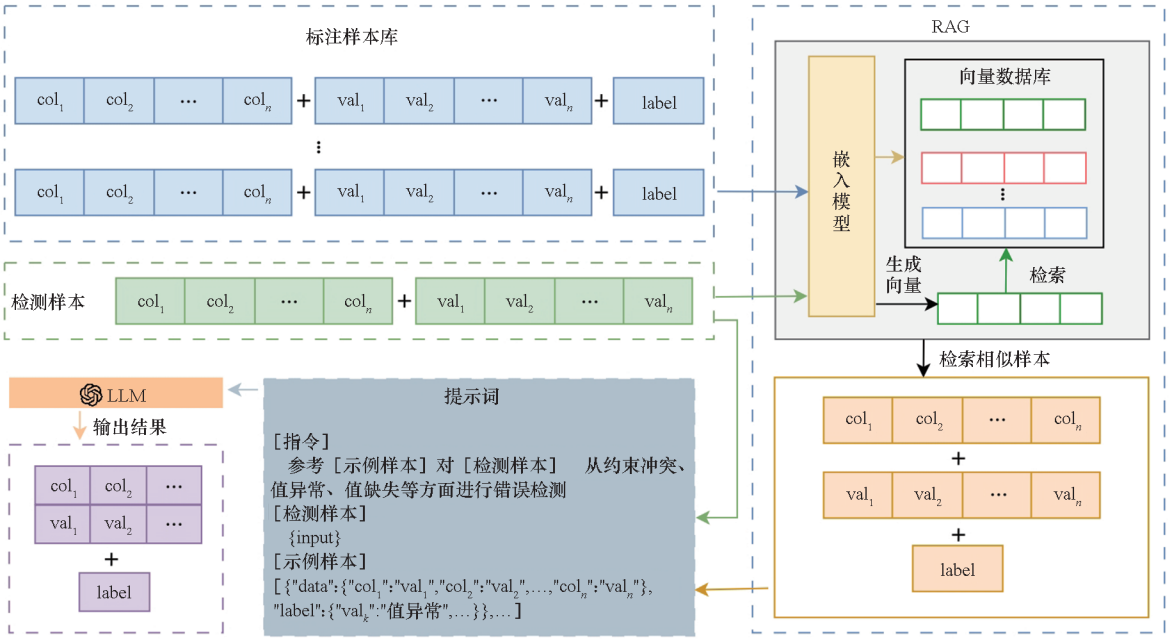

图 3　融合大模型的错误检测原理

Fig. 3　Principles of error detection with integrated large language model

现有研究尽管在自动化与智能化方面取得了显著进展，但仍存在一些不足。例如，不同数据场景下错误类型差异较大，现有方法在跨领域泛化能力方面仍有提升空间；同时，大模型方法在成本控制、稳定性以及与结构化数据特征的深度融合方面仍有待进一步研究。未来，错误检测研究有望在领域知识融合、多模态数据理解以及高效低成本的大模型应用等方向取得进一步突破。

2.2　数据修复

数据修复旨在针对检测出的值缺失、值异常和违反约束等错误进行修正，使修正后的数据符合分布或约束要求。其挑战在于数据约束之间往往相互依赖且关系复杂，修复规则数量多、搜索空间大且修复规则序列长。现有研究围绕如何更有效地刻画数据之间的约束关系以及如何自动生成修复策略展开，方法体系也呈

现从显式规则建模到隐式关系学习,再到生成式与决策式方法的演进趋势。目前相关研究路径包括基于概率模型的数据修复、基于表示学习的数据修复、基于大模型的数据修复以及多步序列决策数据修复等方向。

1)基于概率模型的数据修复。传统数据修复方法,通常针对预定义好的约束或约束组合进行修复,难以发现和使用潜在存在的未定义约束或约束组合。基于概率模型的数据修复,将众多完整性约束,甚至是数据统计特征及外部参考信息统一采用概率模型进行建模,通过统计推断的方式寻找更全面的潜在约束关系,并用于数据修复,典型工作包括 HoloClean[26]、MLNClean[27]、BClean[28]和 PClean[29]等。基于概率模型的数据修复,可以自动挖掘出潜在约束关系,具有较可靠的理论基础。

2)基于表示学习的数据修复。基于概率模型的方法,依然需要对数据间的约束关系进行手工抽取。基于表示学习的方法,将数据映射到潜向量空间,以数据在向量空间中的位置表征其关系,进而在潜向量空间对错误数据的潜向量表示进行修复。Sudowoodo[30]通过对比表示学习框架学习数据的向量表示,根据数据记录之间的向量相似性抽取错误数据的相似数据集,并用于实现数据修复。Lopster[31]则直接在潜向量空间进行数据修复,通过学习潜空间算子,将错误数据转换到正确的数据空间区域,最后再将错误数据被修复后的潜空间表示解码成数据。上述方法将数据映射到潜向量空间并利用向量关系进行修复,能够更灵活、高效地捕捉数据间的复杂依赖,实现对错误数据的精准修复。

3)基于大模型的数据修复。基于概率模型和基于表示学习的数据修复均需要针对具体约束组合预定义修复规则。借助大模型强大的生成和推理能力,IterClean[32]、RetClean[25]和 ZeroEC[33]以提示词或思维链驱动大模型生成修复规则和修复结果。GIDCL[24]在数据修复过程中,还使用了生成数据对大模型进行微调。基于大模型的数据修复方法,将数据修复由规则挖掘范式转向规则生成范式,进一步提高了数据修复方法的灵活性。

4)多步序列决策数据修复。真实环境中的数据集常存在多种错误,需要多步修复,该过程可被建模成序列决策过程。GARF[34]和 GARF +[35]采用序列生成对抗网络,在训练阶段学习依赖关系,在推理阶段生成修复规则序列。UniClean[36]则从优化角度出发,以最小化所有修复器检测分数为优化目标来编排修复步骤。多步序列决策数据修复,实现了数据工程场景下多步工作流自动化的尝试,同时也为数据清洗、连接、发现等整体工作流自动化提供了启发。

总体来看,现有数据修复研究已经从基于规则的约束建模逐步发展到基于表示学习与生成式模型的自动化修复。然而,在复杂真实场景下仍面临若干挑战,例如如何在缺乏高质量标注数据的情况下实现可靠修复、如何在多源异构数据环境中统一建模数据依赖关系,以及如何在保证可解释性的同时提升自动化程度。未来研究可进一步探索结合大模型与结构化约束知识的混合修复框架,以及面向数据工程工作流的端到端自动化修复方法,以提升数据修复在真实应用场景中的实用性与可扩展性。

2.3　数据补全

数据补全的目标是对缺失数据进行估计和填补,以恢复数据完整性,它是数据质量管理中的重要研究方向。由于真实数据往往存在值域范围广、数据稀疏以及语义信息利用不足等问题,数据补全面临较大挑战。现有研究方法主要经历从基于数据相似性的近邻推断方法,到通过学习整体数据分布的生成式方法,再到利用语义信息与外部知识的大模型方法的发展过程。这一演进过程体现了数据补全研究从依赖局部统计特征逐步转向融合全局分布与语义知识的趋势。当前主流研究路径主要包括基于近邻的数据补全、基于对抗生成的数据补全以及基于大模型的数据补全。

1)基于近邻的数据补全。无论是地址、产品分类等离散型字段,还是价格、尺寸等连续型字段,其值域范围往往很广,仅凭数据记录本身难以决定补全值。基于近邻的数据补全方法,通过寻找相似的近邻数据记录,并根据近邻数据记录取值来缩小待补全数据记录的取值范围,直至决断出补全结果。Song 等[37]定义了一种基于属性值的相似性规则,通过该规则检索近邻数据记录,然后使用近邻数据记录属性值补全数据。MIDIA[38]学习缺失记录与非缺失记录之间的非线性关系预测缺失值。针对离散型字段和连续型字段值同时存在缺失的情况,GRIMP[39]提出一种基于异构图建模并计算近邻关系的方法,通过聚合近邻字段值补全数据。基于近邻的数据补全方法,将缺失值推断从数据记录本身扩展到更大范围的近邻数据记录集合上,提高其效果的关键是如何度量数据记录之间的相似性进而利用近邻记

录值补全数据。

2)基于对抗生成的数据补全。由于真实数据维度一般较多,基于近邻的数据补全方法得到的近邻数据记录集合,往往比较稀疏,给数据补全带来困难。基于对抗生成的数据补全方法,通过学习整体数据分布,并通过伪数据生成方式解决数据稀疏问题。GAIN[40]和IFGAN[41]引入生成对抗网络(generative adversarial network,GAN)框架,通过生成器生成补全数据,并利用判别器区分观测值与生成值的方式进行对抗优化以实现数据补全。VGAIN[42]进一步结合自编码器与GAN的优势,对潜在分布添加高斯噪声增强特征学习的鲁棒性,并利用重构误差指导模型训练,从而生成更精确的补全值。基于对抗生成的方法提供了一种学习分布规律并用来生成补全数据的范式,可以缓解数据维度高造成的数据稀疏问题。

3)基于大模型的数据补全。基于近邻的数据补全方法和基于对抗生成的数据补全方法一般仅考虑数据值本身,往往忽略了数据字段语义信息的指导作用。CRILM[43]利用预训练语言模型为缺失特征生成语义化描述,并作为辅助信息微调下游小型预测模型以实现补全。RetClean[25]和LakeFill[44]采用检索增强策略,从外部知识库中获取相似信息辅助大模型补全数据。LLM-Forest[45]则考虑单个大模型的可靠性问题,通过随机森林的思想融合多个大模型的补全结果。基于大模型的数据补全方法,从语义层面对数据上下文和外部知识进行理解并生成数据,可以提升补全数据的解释性和可靠性。

总体来看,现有数据补全研究虽然在算法模型和应用场景上取得了显著进展,但仍存在一些值得进一步探索的问题。例如:多数方法仍侧重于单表数据补全,对于跨表关联数据和复杂数据生态环境中的缺失问题关注不足;同时,不同类型数据(如结构化数据与文本数据)的协同补全机制仍有待深入研究。未来研究可进一步探索结合结构化数据关系、语义知识以及生成模型能力的统一数据补全框架,并加强在真实复杂数据环境中的鲁棒性与可解释性研究。

表2对前述智能化数据清洗方法进行了总结。这些方法通常在精确率(P)、召回率(R)、F1分数(F1)、准确率(ACC)、受试者工作特征曲线下的面积(AUC)、均方根误差(RMSE)等核心指标上进行验证和比较。

表2 数据清洗方法对比

Tab. 2 Comparison of data cleaning methods

领域	基本范式	方法	指标	核心方法
错误检测	学习	SCODED[14]	P,F1	统计约束
		Uni-Detect[15]	F1	自动学习约束
		Auto-Test[16]	P,R,F1	自动学习约束
		杜岳峰等[17]	R	函数约束
		Guardrail[18]	P,R,F1	自动学习约束
		Dai等[19]	AUC	深度神经网络
	小样本	HoloDetect[20]	P,R,F1	深度神经网络
		SAGED[21]	F1	语义感知与推理
		ZeroED[22]	F1	语义感知与推理
	融合大模型	Narayan等[23]	F1,ACC	提示词工程
		GIDCL[24]	P,R,F1	语义感知与推理
		RetClean[25]	ACC	检索增强生成
数据修复	概率模型	HoloClean[26]	P,R,F1	概率推断模型
		MLNClean[27]	P,R,F1	马尔可夫逻辑网络
		BClean[28]	P,R,F1	自动贝叶斯网络
		PClean[29]	P,R,F1,ACC	概率编程模型
	表示学习	Sudowoodo[30]	R,F1	对比表示学习
		Lopster[31]	F1	潜空间算子学习
	大模型	IterClean[32]	P,R,F1	提示词工程
		RetClean[25]	ACC	检索增强生成
		ZeroEC[33]	P,R,F1	思维链
		GIDCL[24]	P,R,F1	大模型微调
	多步序列决策	GARF[34]	P,R,F1	生成对抗网络
		GARF +[35]	P,R,F1	生成对抗网络
		UniClean[36]	F1	多路径修复优化
数据补全	近邻	Song等[37]	ACC	值相似
		MIDIA[38]	ACC,RMSE	非线性缺失预测
		GRIMP[39]	ACC	异构图
	对抗生成	GAIN[40]	AUC,RMSE	生成对抗网络
		IFGAN[41]	RMSE	生成对抗网络
		VGAIN[42]	RMSE	生成对抗网络
	大模型	CRILM[43]	ACC	预训练模型
		RetClean[25]	ACC	检索增强生成
		LakeFill[44]	R	检索增强生成
		LLM-Forest[45]	ACC	树状多模型预测

3　智能化数据连接方法

数据连接目标是识别来自不同数据集的数据模式之间、数据记录之间的语义关系并进行匹配,实现数据属性的横向扩展和数据记录的纵向扩展,思路如图 4 所示。智能化数据连接方法通过表示学习和大模型等技术,对数据模式、数据记录等进行嵌入空间语义编码并在语义层面进行推理,完成数据模式匹配和数据实体匹配,实现数据连接由传统规则式连接范式向隐空间式、语义推理式连接范式转变,提高了复杂语义场景下数据连接的准确性和全面性。数据连接方法的相关研究,常在表 3 和表 4 所列数据集上验证。

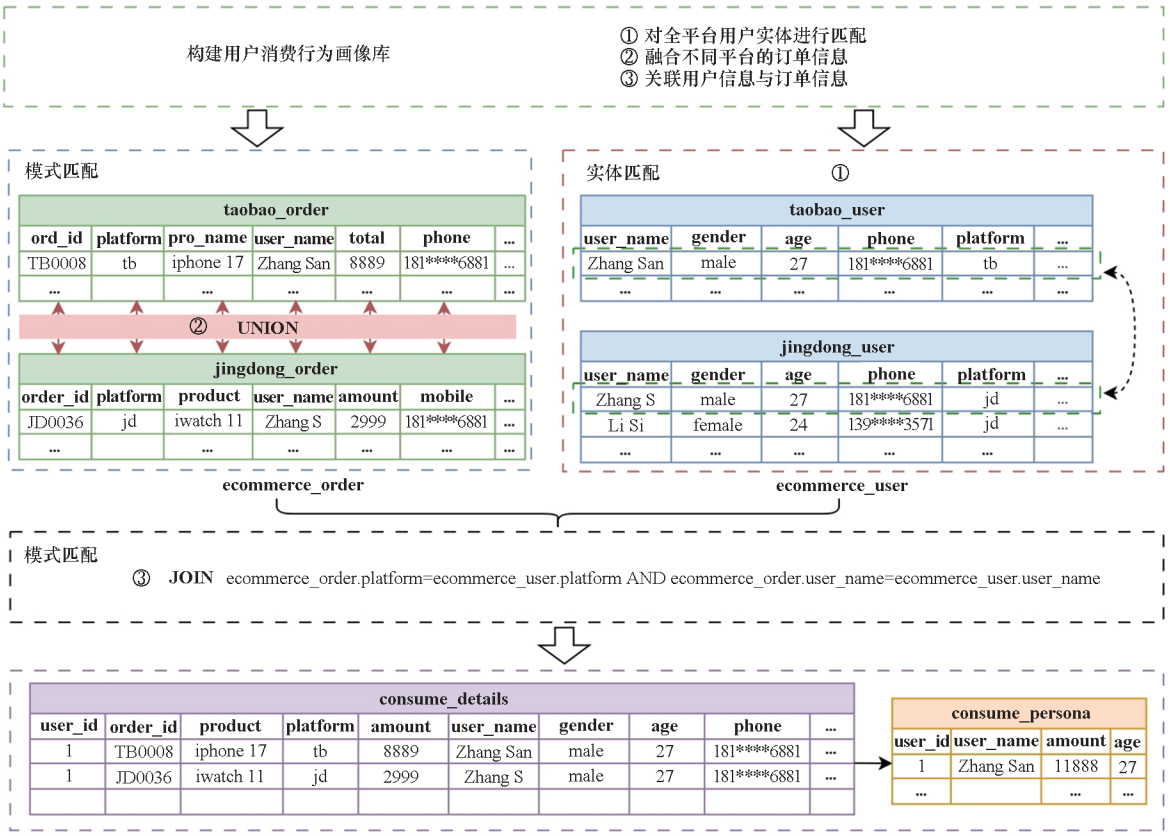

图 4　数据连接示意图

Fig. 4　Schematic diagram of data linking

表 3　数据连接 – 模式匹配研究常用数据集

Tab. 3　Data linking-schema matching research common datasets

数据集	表对	行数	列数
WikiData[46]	4	5 423 ~ 10 846	13 ~ 20
ChEMB[46]	180	7 500 ~ 15 000	12 ~ 23
Open Data[47]	180	11 628 ~ 23 255	26 ~ 51
TPC-DI[48]	180	7 492 ~ 14 983	11 ~ 22
Magellan[49]	7	864 ~ 131 099	3 ~ 7

表 4　数据连接 – 实体匹配研究常用数据集

Tab. 4　Data linking-entity matching research common datasets

数据集	列数	实体对	匹配实体对
Walmart-Amazon[50]	5	10 242	962
Amazon-Google[50]	3	11 460	1 167
DBLP-Scholar[50]	4	28 707	5 347
DBLP-ACM[50]	4	12 363	2 220
Fodors-Zagats[50]	6	946	110
iTunes-Amazon[50]	8	532	132
Abt-Buy[51]	3	9 575	1 028

3.1　模式匹配

模式匹配旨在识别不同数据表中的相关字段,从而确定连接字段以支撑数据融合。该问题的核心挑战在于:字段属性与字段值形式多样,字段语义表达复杂,传统基于人工特征的方法难以充分刻画字段之间的语义关系;同时,多表之间往往存在复杂的结构关联,而高质量标注数据又相对稀缺。随着数据表示学习和大模型技术的发展,模式匹配方法逐渐从基于人工特征的相似性

匹配范式，演进到基于深度表示学习的语义匹配范式，并进一步发展为融合结构关系建模和大模型语义推理的智能化匹配范式。现有研究大致可分为三类：基于表示学习的方法、基于图方法以及基于大模型的方法。

1）基于表示学习的模式匹配。传统方法主要根据字段值人工构建特征并基于预定义相似性度量指标进行匹配。基于表示学习的方法，通过深度神经网络强大特征提取能力，将字段属性和值等信息生成向量表示并用于后续基于相似性的模式匹配中。PEXESO[52] 和 PolyJoin[53] 分别采用 fastText 和双向编码表示（BERT）对字段进行向量表示。而 OmniMatch[54] 和 Unicorn[55] 则在提取字段特征向量基础上，分别采用图神经网络和混合专家网络模块进行特征增强。LSM[56]、DeepJoin[57] 和 DiscoverGPT[58] 利用 BERT 模型已有的文本表征能力，采用标注模式匹配样本对预训练模型进行微调。基于表示学习的方法，可捕捉复杂的字段模式关系，降低了对人工特征工程的依赖，提升了匹配的准确性和模型的泛化能力。

2）基于图方法的模式匹配。基于表示学习提取特征，通常仅关注匹配字段对之间的信息，而真实数据表之间通常存在复杂关联关系，此类关联关系可为模式匹配提供增强信息。基于图方法的模式匹配，将数据表和字段作为图的节点，构建数据表及字段间的关联关系图，并通过图优化和图搜索等方法，查找字段之间的匹配关系。Auto-BI[59] 通过模式匹配预测方法在局部视角对字段对的匹配概率进行预测，基于字段对概率构建全局连接图，对连接图进行优化，通过优化后的图中连边指示模式是否匹配。Auto-Prep[60] 除使用数据表的原始字段构建连接图之外，还将数据表转换过程中生成的新的字段构建在连接图中，通过图搜索来确定匹配模式。基于图方法的模式匹配，采用显式图建模多种数据关系，并利用多跳关系数据增强模式匹配能力，是结构化知识在模式匹配上的有用尝试。

3）基于大模型的模式匹配。基于表示学习和图方法的模式匹配都需要大量标注数据。大模型的少样本学习能力和语义理解能力可以在较少标注样本下进行模式匹配。其思路如图 5 所示，是通过提示词方式，将模式匹配任务指令、示例样本、待匹配样本等信息输入大模型，通过提示词工程或大模型微调方式，由大模型进行推理给出匹配结果。GRAM[61] 和 JD-SCOPE[62] 均通过构造模式匹配任务提示词并采用多种基座大模型进行模式匹配。不同的是，GRAM 引入 RAG 技术进行属性相同实体的字段检索和示例样本检索，而 JD-SCOPE 采用提示词模板自动生成技术构建更适合匹配任务的提示词。Magneto[63] 使用大小模型结合方式，训练阶段用大模型生成训练样本对小模型进行微调，推理阶段小模型负责实现候选匹配对检索和粗排，大模型负责实现候选匹配对评估和精排。基于大模型的方法，将模式匹配从模式特征挖掘范式带入模式特征增强范式，能有效丰富模式语义，提高模式匹配效果。

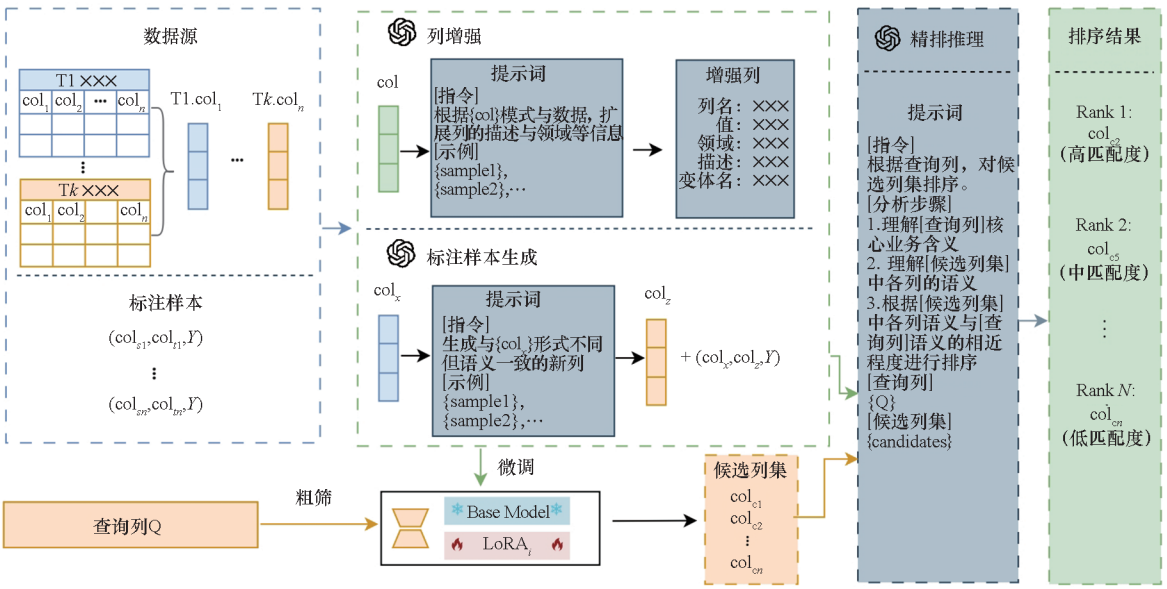

图 5 基于大模型的模式匹配原理图

Fig. 5 Schematic diagram of integrated large language model pattern matching

总体来看,现有模式匹配方法在语义建模能力方面已取得明显进展,但仍存在若干研究挑战:例如,如何在缺乏高质量标注数据的情况下提升模型泛化能力,如何有效融合字段语义信息与数据表结构关系,以及如何在大模型推理能力与计算成本之间取得平衡。未来研究可进一步探索大模型与结构化表示学习的深度融合、弱监督或自监督模式匹配方法,以及面向真实数据湖环境的大规模模式匹配框架,以提升模式匹配方法在开放复杂数据环境中的适用性。

3.2　实体匹配

实体匹配旨在识别两个数据记录是否指向同一个真实世界实体,是数据集成与数据连接中的核心问题。由于现实数据通常存在属性表达差异、语义歧义以及结构不一致等问题,实体匹配面临实体语义和结构特征复杂、实体模式信息不完善以及标注成本高等挑战。随着研究的不断发展,实体匹配方法经历了从依赖特征工程和规则设计,逐步向表示学习和深度语义建模演进的过程。近年来,随着预训练语言模型和大模型技术的发展,研究者开始探索利用深度语义表示与上下文理解能力提升匹配效果。当前研究路径主要包括基于表示学习的实体匹配、模式无关的实体匹配以及基于大模型的实体匹配等方向。

1)基于表示学习的实体匹配。由于实体语义和结构复杂,匹配规则构建成本高且灵活性差。基于表示学习的方法,通过学习捕获实体语义和结构特征,并嵌入空间向量,再通过计算向量相似度来判定实体是否匹配。DAEM[64]、Seq2SeqMatcher[65]、CorDEL[66]分别采用不同表示学习方法,捕获实体在句法和语义层面的关系实现匹配。其中,Seq2SeqMatcher 着重捕获词元间的语义相关性,而 DAEM 和 CorDEL 着重捕获语法和记录之间的相似性与差异性。

2)模式无关的实体匹配。现实情况下,待实体匹配时可能没有对齐的模式,此时模式或属性语义是未知的。模式无关的实体匹配,不依赖数据表模式对齐关系,直接基于记录中包含的内容来判断两个记录是否指向同一实体。樊峰峰等[67]计算记录序偶的属性相似度并将其映射到特征空间,利用离群点检测识别潜在匹配对,从而实现自动实体匹配。Teong 等[68]、Ditto[69]和 StruBERT[70]通过预训练语言模型将未对齐模式下的实体匹配问题视为自然语言处理问题,通过将元组拼接为句子来消除对模式的依赖。其中 Teong 等和 Ditto 主要通过微调预训练语言模型的方式,而 StruBERT 采用行列双向自注意力的结构感知 BERT,统一融合表格文本与结构信息,用于表格检索和相似度计算,扩大了实体匹配方法的适用范围。

3)基于大模型的实体匹配。预训练语言模型方法需要依赖大量标注的匹配实体对。大模型可利用上下文学习能力捕获复杂实体间的语义关系。Fan 等[71]与 Peeters 等[72]采用不同提示策略提升实体匹配效率。为突破传统大模型只专注于二元实体匹配的局限,COMEM[73]将多个候选记录整体输入,引入比较、选择机制建模记录间的关系,在匹配决策中显式利用记录交互信息实现全局一致的实体匹配。基于大模型的上下文学习能力,为大规模实体匹配奠定了基础。

总体来看,现有研究在提升实体语义理解能力和降低模式依赖方面取得了显著进展。然而,当前方法存在若干挑战。例如,大模型方法在实际应用中仍面临推理成本高、可解释性不足以及幻觉等问题;同时,复杂多源数据环境下的跨模态实体匹配、弱监督或无监督实体匹配等问题仍有待进一步研究。未来研究可在高效大模型推理、跨领域知识迁移以及多源异构数据融合等方向进一步探索,以提升实体匹配在真实数据环境中的可扩展性与鲁棒性。

表 5 总结了前述智能化数据连接方法。这些方法通常在 P、R、F1、ACC、平均精度均值(MAP)、平均倒数排名(MRR)等核心指标上进行验证。

表 5　数据连接方法对比

Tab. 5　Comparison of data linking methods

领域	范式	方法	指标	核心方法
模式匹配	表示学习	PEXESO[52]	P,R,F1	向量相似性检索
		PolyJoin[53]	R,MAP	向量相似性检索
		OmniMatch[54]	F1	向量 + 图神经网络
		Unicorn[55]	R,F1,ACC	向量 + 混合专家网络
		LSM[56]	ACC	预训练模型
		DeepJoin[57]	P,R,F1	预训练模型
		DiscoverGPT[58]	P,R,F1	预训练模型
	图方法	Auto-BI[59]	P,R,F1	图优化
		Auto-Prep[60]	P,R,F1	图搜索
	大模型	GRAM[61]	ACC	基于输入特征的动态提示选择
		JD-SCOPE[62]	P,R,F1,ACC	提示词工程
		Magneto[63]	R,MRR	大小模型协作

续表

领域	范式	方法	指标	核心方法
实体匹配	表示学习	DAEM[64]	F1	神经网络 + 对抗学习
		Seq2Seq Matcher[65]	P,R,F1	“对齐 - 比较 - 聚合”网络
		CorDEL[66]	F1	对比深度学习框架
	模式无关	樊峰峰等[67]	F1,ACC	离群点检测
		Teong 等[68]	P,R,F1	预训练模型微调
		Ditto[69]	F1	预训练模型微调
		StruBERT[70]	P,R,F1	双向自注意力的结构感知 BERT
	大模型	Fan 等[71]	F1	上下文学习
		Peeters 等[72]	F1	上下文学习
		COMEM[73]	F1	多源交互信息融合

4 智能化数据发现方法

数据发现的目标是,根据用户不同形式的表达需求搜索大规模数据集(如数据湖),通过数据导航和数据注释等方式,得到用户所需数据集,如图 6 所示。通过数据发现,可实现对清洗、连接后的大规模数据集的高效组织运用。传统基于关键词和浅层相似度的数据发现方法难以深入理解“数据集 - 数据集”之间、“用户查询意图 - 数据集”之间的深层次语义关系。智能化数据发现方法通过知识图谱、表示学习、大模型等技术对用户查询意图、数据集元信息进行语义增强并生成数据集的高效组织结构,实现由浅层关键词相似性匹配范式向深层业务意图匹配范式的转变。数据发现方法相关研究常在表 6 所示数据集上进行验证。

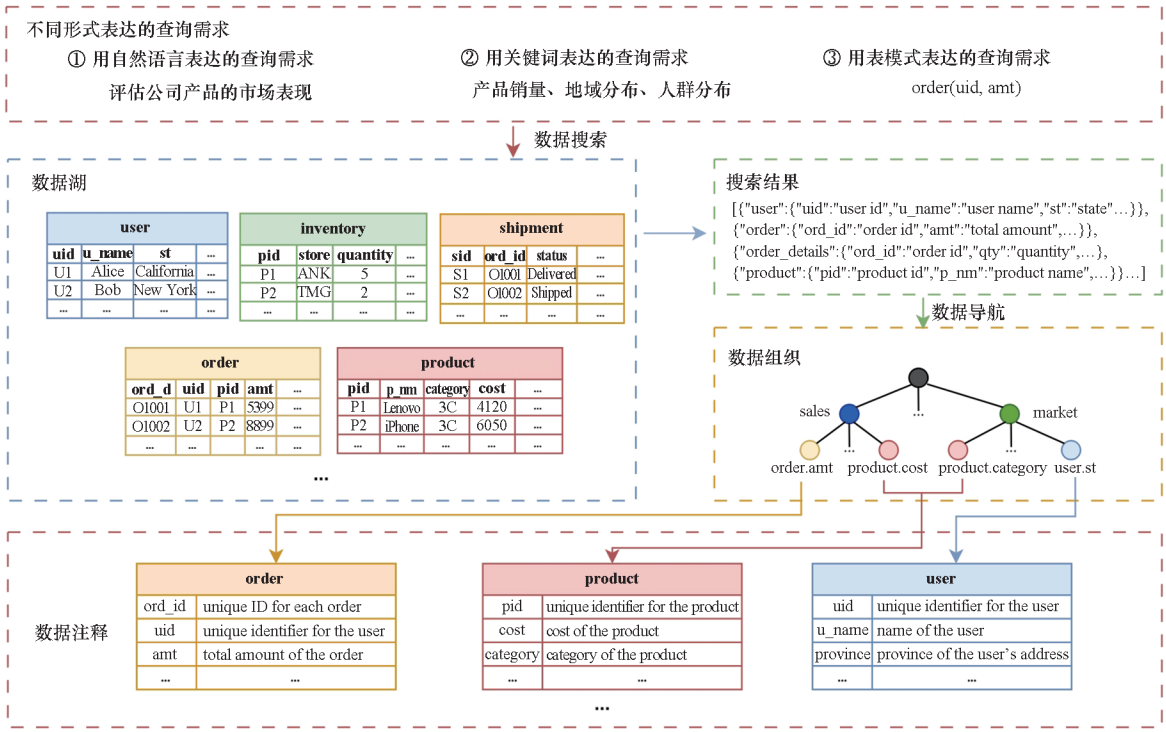

图 6 数据发现示意图

Fig. 6 Schematic diagram of data discovery

表 6 数据发现研究常用数据集

Tab. 6 Data discovery research common datasets

数据集	表数量	平均行	平均列	测试场景
TUS-SANTOS[74]	1 127	6 033.90	12.65	数据搜索
ECB Union[74]	4 226	311.76	35.95	
CKAN Subset[74]	40 594	1 823.24	24.92	数据导航
Wiki Jaccard[74]	8 489	47.30	2.70	数据搜索
Wiki Containment[74]	10 318	47.15	2.69	数据注释
Wiki Union[74]	40 752	51.05	2.62	

4.1 数据搜索

数据搜索旨在根据用户查询请求从数据湖等环境中获取匹配的数据表,是数据发现任务中的核心环节。查询请求通常包括关键词、自然语言语句,或包含表结构、表内容等信息的查询描述。其中,后者通常用于发现可连接(JOIN)或可合并(UNION)的数据表,以支持后续数据整合与分析。然而,用户查询请求和数据表信息通常呈现为半结构化或非结构化形式,同时用户查询意图

表达简略且可能存在歧义，而数据表元信息往往缺乏完整注释或语义描述，这使得查询意图与数据表语义之间存在显著的语义鸿沟。针对这一问题，现有研究大体经历了从基于表示学习的数据搜索方法到利用大模型进行语义增强的数据搜索方法的发展过程，逐步提升了对查询意图与数据表语义关系的建模能力。

1）基于表示学习的数据搜索。数据搜索本质是将用户查询请求与数据表进行匹配，但是用户查询请求和数据表信息通常都是非结构化的。基于表示学习的方法，通过深度神经网络将用户查询请求和数据表相关信息进行向量化表示，并通过相似性计算等方式进行匹配，得到符合查询请求的数据表。Chen 等[75]和 Lin 等[76]分别使用 BERT 和 text-embedding-3-small 等预训练模型对数据信息进行向量表示，而 Starmie[77]和 TabSketchFM[78]则提出预训练模型并进行通用表格任务预训练，再用数据发现任务进行模型微调以增强对数据的表征能力。基于表示学习的方法，提出将数据集隐式表示的方法，为更多引入深度学习方法奠定了基础。

2）基于大模型的数据搜索。用户查询意图往往以若干个关键词或短句等形式出现，并且还可能存在歧义等意图模糊问题，而数据表在表级和字段级上的注释信息通常缺失，且一般没有对数据表中数据的概要性、清晰性描述。基于大模型的方法，通过提示词等方式扩展用户查询，生成数据表元信息和数据概要信息，用于增强数据搜索性能，思路如图 7 所示。Pneuma[79]通过检索增强技术生成数据元和数据概要，TableRAG[80]则通过 RAG 技术同时扩展用户查询意图、生成数据元信息和概要信息。基于大模型的数据搜索，可以对用户查询意图和数据集信息进行双向增强，从业务层面提升了数据搜索准确度。

总体来看，现有研究已从单纯依赖语义表示匹配逐渐发展到结合生成式模型进行语义增强的数据搜索范式。然而，当前方法仍主要依赖生成的元信息质量，并在大规模数据湖环境下的效率、可解释性以及跨领域泛化能力方面仍存在不足。未来研究可进一步探索结构化表信息与大模型语义推理能力的深度融合，以及面向复杂数据分析任务的意图感知型数据搜索方法。

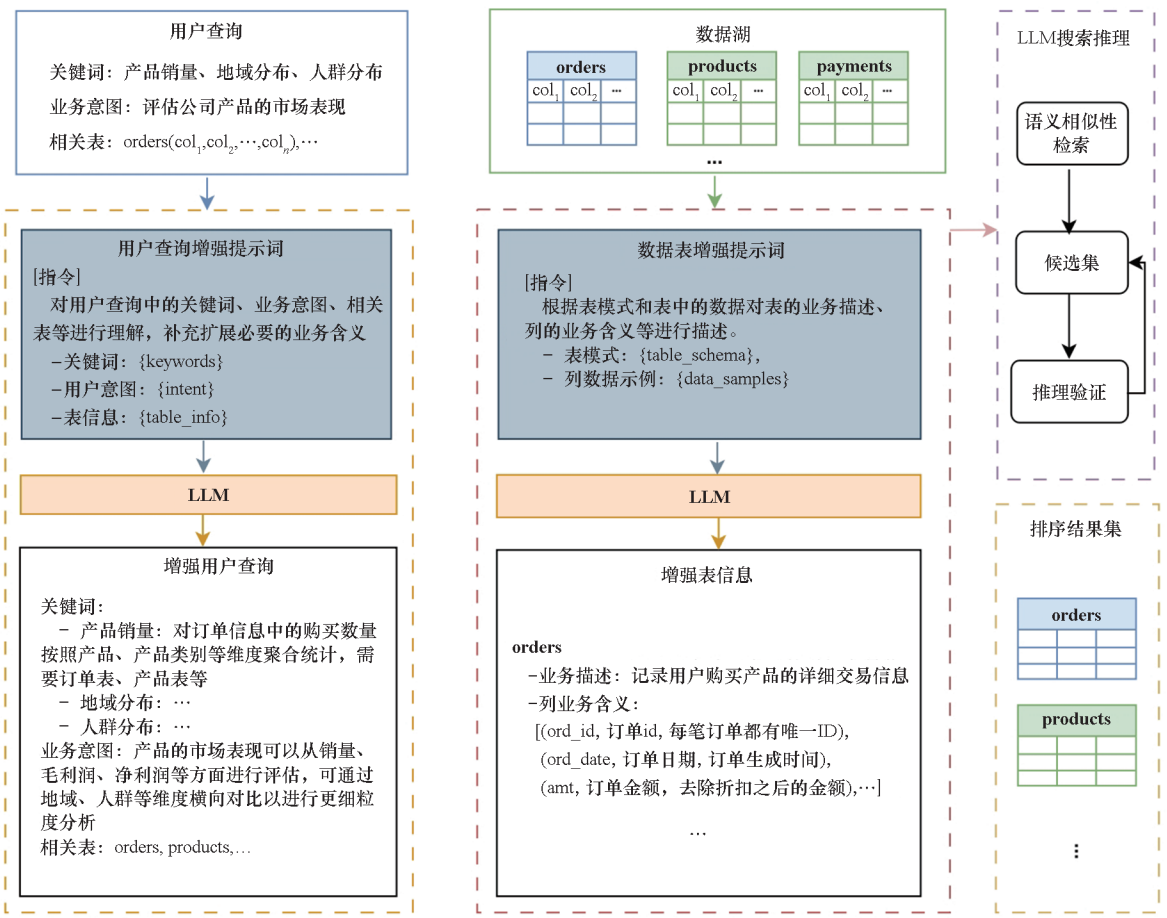

图 7　融合大模型的数据搜索原理图

Fig. 7　Schematic diagram of LLM-integrated data search

4.2 数据导航

数据导航旨在根据数据集之间的关系构建一种支持用户在多个数据集之间高效跳转和探索的数据组织形式。在数据湖等大规模数据环境中，用户往往需要通过数据之间的结构或语义联系逐步发现潜在相关数据，因此如何构建清晰、可探索的数据关系结构成为数据导航的核心问题。然而，该任务面临多方面挑战：一方面，数据集之间的关联既可能来自模式结构、字段包含关系，也可能来自语义相似性或业务语义。另一方面，不同研究在数据组织形式上也呈现从语义网络到结构化导航图的多样化设计。现有研究沿着两条技术路径展开：一类方法利用知识图谱等语义网络显式表达数据之间的语义关联，以支持灵活的数据发现与导航；另一类方法则将数据导航问题转化为图结构优化问题，通过构建具有明确层次或组织结构的导航图，提高用户探索效率。

1）基于知识图谱的数据导航。数据导航本质上是根据数据集之间的关系在数据集之间构建一个可导航的链路。基于知识图谱的方法，通过语义相似性在数据表、数据字段等之间构建连接关系，形成可导航的知识图谱。Aurum[81]根据数据表字段同属同一个数据表或数据源的关系，以超图形式构建连接数据表字段的知识图谱。SEMPROP[82]引入更多类型节点，包括数据表、字段、本体等，并通过词向量方法识别节点之间的语义连接，提升了数据查找与访问的效率，同时为复杂数据湖中的智能导航提供了有效支持。

2）基于图优化的数据导航。面对数据湖中众多的数据表，甚至是通过数据搜索方法过滤之后仍然较多的数据表，即使有已经构建好的知识图谱，但是由于知识图谱的非层次化结构特性，用户一般难以选择一个好的出发点以高效地进行数据探索。Nargesian 等[83-84]将数据导航建模成一个求解数据组织的有向无环图结构的图优化问题，图中节点为数据表的字段集合，连边为字段集合之间的子集关系，子节点为父节点的子集，叶子节点为仅含一个字段的集合。其提出一种用户在图结构上进行导航的概率模型，并提出了一种基于局部搜索的构建数据组织有向无环图结构的近似求解算法。Ouellette 等[85]进一步提出一种在数据搜索结果之上动态构建有向无环图的方法RONIN，为用户在海量数据表中提供高效、结构化的探索路径，同时提升了数据发现和分析的效率。

总体而言，现有数据导航研究已经从早期依赖语义关联的知识图谱方法，逐渐发展到结合结构优化的数据组织方法，以提升用户在大规模数据环境中的探索效率。然而，当前研究仍存在一定局限。例如，多数方法主要依赖数据模式或文本语义信息，难以充分利用数据内容、用户行为以及任务上下文等多维信息。同时，现有导航结构往往在系统构建阶段静态生成，缺乏对用户探索过程的动态适应能力。未来研究可进一步探索结合用户交互行为的自适应数据导航机制，以及融合语义关系、数据内容与使用上下文的多层次数据组织模型，以提升复杂数据环境中的数据发现与探索效率。

4.3 数据注释

数据注释旨在为数据表中的字段附加语义描述，从而提升数据的可理解性与可复用性。然而，在开放数据环境中，数据表往往存在元数据缺失、数据记录属性值稀疏以及人工标注成本高等问题，使得列语义识别和标注具有较大挑战。现有研究主要围绕如何充分利用表格内部信息与外部知识资源展开，并逐渐形成了从基于表示学习的方法、基于知识图谱的方法到基于大模型的方法的技术演进路径。前者侧重于从表格内部上下文中学习语义表示，随后通过引入外部知识图谱弥补上下文信息不足的问题，近年来则进一步借助大模型的知识推理能力提升列注释的自动化与泛化能力。

1）基于表示学习的列注释。主要思路是利用表自身的上下文信息对列进行标注，提升标注效率。Doduo[86]获取表格中所有列的序列化列值来表征表格上下文，从而兼顾列内和列间关系实现列标注。ColNet[87]通过混合神经网络学习表的局部特征。进一步，针对整合复杂表上下文的困难问题，RECA[88]提出一种表命名实体模式，对模式相似的表进行对齐，并在对齐后的表中提取有用的上下文信息对列注释。上述方法减少了对表格元数据的依赖。

2）基于知识图谱的列注释。当表的内部信息不足以探索正确的列信息时，表示学习会面临上下文缺失的问题。KGLink[89]与 C^2[90]通过构建知识图谱的方式，实现列注释。EXACTA[91]基于逆强化学习进行多跳知识图谱推理，以实现可解释的列注释。基于知识图谱的方法，可以有效引入外部结构化知识来辅助补充表的上下文信息。

3）基于大模型的列注释。CHORUS[92]和LLMCTA[93]利用提示词工程，在不需或仅需少量标注的情况下，通过精心设计上下文提示引导模

型生成列注释。RACOON[94]通过 RAG 方式将外部知识整合到大模型生成过程中,以提升列注释准确性。基于大模型的方法,将大模型固有的常识知识和对外部数据检索理解的能力融入列注释生成过程,可以减少人工标注成本,提升大规模数据集上列数据标注的可行性。

总体来看,现有研究在提升列注释自动化程度方面取得了一定进展,但仍存在一些不足。例如,不同数据域之间的语义差异可能影响模型的泛化能力,同时外部知识资源与表格上下文的有效融合仍有待进一步探索。未来研究可进一步关注跨领域数据语义对齐、可解释列注释以及大模型与结构化知识深度融合等方向,以提升数据注释方法在开放数据环境中的鲁棒性与适用性。

表 7 对前述智能化数据发现方法进行了总结,这些方法通常在 P、R、F1、ACC、MAP、效率(EFF)、主观评分(SR)等核心指标上进行验证和比较。

表 7　数据发现方法对比

Tab. 7　Comparison of data discovery methods

领域	范式	方法	指标	核心方法
数据搜索	表示学习	Chen 等[75]	MAP	预训练模型
		Lin 等[76]	SR	预训练模型
		Starmie[77]	MAP,R	对比学习 + 预训练模型
		TabSketchFM[78]	F1	预训练模型微调
	大模型	Pneuma[79]	EFF	检索增强生成
		TableRAG[80]	P,R,F1	检索增强生成
数据导航	知识图谱	Aurum[81]	F1,P	知识图谱
		SEMPROP[82]	P,R,F1	知识图谱
	图优化	Nargesian 等[83-84]	EFF	DAG 优化
		Ouellette 等[85]	EFF	DAG 优化
数据注释	表示学习	Doduo[86]	F1	序列化列值
		ColNet[87]	P,R,F1	混合神经网络
		RECA[88]	F1	表命名实体模式
	知识图谱	KGLink[89]	F1	知识图谱
		C^2[90]	ACC	知识图谱
		EXACTA[91]	F1	逆强化学习 + 知识图谱
	大模型	CHORUS[92]	P,R,F1	提示词工程
		LLMCTA[93]	F1	提示词工程
		RACOON[94]	F1	大模型 + 知识图谱

5　讨论

本综述致力于围绕数据清洗、数据连接和数据发现三个核心环节展开,按照从传统规则驱动方法转向以机器学习、表示学习和大模型为核心的发展趋势进行论述,着重体现数据工程智能化的演进路径。其中,大模型助力数据工程智能化贯穿数据清洗、数据连接和数据发现三个环节,也是目前的主流趋势。

然而,大模型在数据工程中的应用仍面临若干关键挑战,本文着重论述其幻觉、隐私和计算资源问题及其对应的解决方案。首先是模型可靠性问题。大模型在生成过程中可能产生“幻觉”,在数据工程场景中可能导致数据处理结果出现偏差。为缓解这一问题,现有研究通常通过引入 few-shot 示例,并结合向量检索或图检索等检索增强生成技术[95-96],为模型提供可靠的上下文信息;同时通过规则约束或后处理校验机制对生成结果进行验证,以降低幻觉带来的影响。其次是隐私与安全问题。数据工程任务往往涉及企业文档、医疗记录或客户服务日志等私有数据,其中可能包含个人身份信息或商业机密。为降低数据泄露风险,研究通常在数据预处理阶段采用隐私识别与脱敏技术识别并替换敏感信息[97-98]。在 RAG 构建过程中,生成合成数据替代原始敏感数据,在保留语义信息的同时去除具体敏感细节[99]。此外,还可通过规则匹配或大模型辅助判断等内容级过滤机制,剔除有毒、偏见或敏感内容,以保障数据的安全与合规。大模型的计算与存储开销也是实际应用中的重要挑战。现有研究通过异构存储机制减少 GPU 显存占用[100],通过键值(KV)缓存管理提升推理效率,并通过任务卸载与流水线化执行优化数据处理流程[101],从而提高系统整体效率。

综上所述,大模型为数据工程智能化提供了新的技术路径,使数据清洗、数据连接与数据发现等任务具备更强的自动化与语义理解能力。但在实际应用中,仍需在模型可靠性、数据隐私安全以及系统资源效率等方面进行综合权衡。

6　总结与展望

传统的数据工程方法主要依赖统计规则或统计学习手段,对数据中潜在的复杂语义信息以及数据间复杂的约束和关联关系,往往难以实现低成本、高效率和全面的挖掘。后来,数据整理[102-103](data wrangling)的出现,将数据清洗、数

据连接和数据发现逐步实现了端到端的自动化处理,为数据工程的效率和质量提升提供了新途径。近年来,人工智能特别是大模型的发展,为解决上述问题提供了新方法,数据工程与人工智能交叉,形成了数据工程智能化这一重要方向。展望未来的数据工程智能化方法研究,还需着眼于数据处理任务面临的复杂多样场景,重点加强自主决策、全流程智能协同、数据基础设施的智能接入等方面研究,不断提升数据工程对大规模复杂异构数据的处理能力。

1)自主决策数据智能体研究。不同领域数据、不同下游任务等都决定着数据清洗、数据连接和数据发现等任务的方法选择与策略选择。现有的智能化数据工程研究中,大模型等智能化方法通常固定在特定模块,在面临不同任务场景时缺少灵活性。未来研究可引入具有任务分解和自主决策能力的智能体技术[104],形成数据智能体。数据智能体根据任务场景和数据情况,对任务进行分解,选定适合方法与策略,并可根据执行结果进一步规划决策后续流程。例如在数据清洗中,智能体可以根据数据质量动态选择清洗策略,并在发现异常模式时进行迭代决策以选择后续清洗策略。自主决策数据智能体将数据工程由固定流程框架提升到任务自主分解与决策的动态流程新水平。未来需要重点研究基于智能体的数据工程任务分解、方法选择与策略决策等问题。

2)数据智能体通信协议研究。数据工程是包含数据清洗、数据连接、数据发现,甚至数据分析与应用等多个环节的复杂任务。端到端的数据工程智能化方法,需要建立不同环节数据智能体之间的通信协议,以此衔接各环节。现有代理间(agent-to-agent, A2A)、代理通信协议(agent communication protocol, ACP)等定义了通用智能体之间的协作协议[105],如元信息定义、通信传输方式、任务生命周期管理等。数据智能体之间要实现数据工程任务协作,还需研究数据任务定义、数据交换格式、数据权限等问题。未来研究需在现有智能体通信协议基础上,形成覆盖多种数据任务、面向灵活数据格式和精细控制数据权限的数据智能体通信协议。

3)数据基础设施接口协议研究。数据通常存放在数据湖、关系数据库等基础设施中,数据工程任务需要与基础设施交互才能对数据进行操作。现有数据工程方法研究通常关注算法改进,将数据简单抽象为 CSV 等格式的数据集,忽略了算法到数据基础设施的接口转换与封装。虽然模型上下文协议[106](model context protocol, MCP)可为智能体与外部工具、数据源提供通用交互标准,但实际应用中往往仍需针对具体任务开发代码。未来研究中,需要强化数据智能体衔接数据基础设施的接口协议及其执行能力。如在数据清洗任务中,生成针对某个特定错误的修复策略,并将其动态转换成针对数据执行引擎用户自定义函数(user-defined function, UDF),然后由数据智能体调用执行。

参考文献(References)

[1] PATON N W, CHEN J Y, WU Z Y. Dataset discovery and exploration: a survey[J]. ACM Computing Surveys, 2024, 56(4): 1 - 37.

[2] ZHU J Y, ZHAO X T, SUN Y, et al. Relational data cleaning meets artificial intelligence: a survey[J]. Data Science and Engineering, 2025, 10(2): 147 - 174.

[3] 郭志懋, 周傲英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002,13(11): 2076 - 2082.
GUO Z M, ZHOU A Y. Research on data quality and data cleaning: a survey[J]. Journal of Software, 2002, 13(11): 2076 - 2082. (in Chinese)

[4] ZHOU X H, HE J X, ZHOU W, et al. A survey of LLM × DATA[EB/OL]. (2025 - 05 - 24)[2026 - 01 - 27]. https://arxiv.org/abs/2505.18458.

[5] ROESLER O. EEG eye state dataset[EB/OL]. (2013 - 06 - 09)[2026 - 01 - 27]. https://archive.ics.uci.edu/ml/datasets/EEG + Eye + State.

[6] BECKER B, KOHAVI R. Adult dataset[EB/OL]. (1996 - 04 - 30)[2026 - 01 - 27]. https://doi.org/10.24432/C5XW20.

[7] YEH I C. Default of credit card clients dataset[EB/OL]. (2016 - 01 - 25)[2026 - 01 - 27]. https://doi.org/10.24432/C55S3H.

[8] Austin Animal Center. Shelter animal outcomes dataset[EB/OL]. (2016 - 03 - 22)[2025 - 12 - 24]. https://www.kaggle.com/competitions/shelter-animal-outcomes/data.

[9] The Movie Database (TMDb). TMDB 5000 movie dataset[EB/OL]. [2026 - 01 - 27]. https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata/data.

[10] Data. Nashville. gov. Traffic accidents dataset[EB/OL]. (2023 - 02 - 16)[2026 - 01 - 27]. https://datanashvillegov-nashville.hub.arcgis.com/datasets/1df26ca05f3f407e95ab5a4fc013eab7_0/explore.

[11] Kaggle. Titanic-machine learning from disaster[EB/OL]. [2026 - 01 - 27]. https://www.kaggle.com/competitions/titanic/data.

[12] HOULD J N. Craft beers dataset[EB/OL]. [2026 - 01 - 27]. https://www.kaggle.com/nickhould/craft-cans.

[13] Data. world. OLS regression challenge-cancer[EB/OL]. [2026 - 01 - 27]. https://data.world/nrippner/ols-regression-challenge.

[14] YAN J N, SCHULTE O, ZHANG M H, et al. SCODED: statistical constraint oriented data error detection[C]//Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020: 845 - 860.

[15] WANG P, HE Y Y. Uni-Detect: a unified approach to automated error detection in tables[C]//Proceedings of 2019 International Conference on Management of Data, 2019:

811 – 828.
[16] CHEN Q X, HE Y Y, WONG R C, et al. Auto-Test: learning semantic-domain constraints for unsupervised error detection in tables [J]. Proceedings of the ACM on Management of Data, 2025, 3(3): 1 – 27.
[17] 杜岳峰, 申德荣, 聂铁铮, 等. 基于关联数据的一致性和时效性清洗方法[J]. 计算机学报, 2017, 40(1): 92 – 106.
DU Y F, SHEN D R, NIE T Z, et al. A cleaning method for consistency and currency in related data[J]. Chinese Journal of Computers, 2017, 40(1): 92 – 106. (in Chinese)
[18] MA P C, WANG Z Y, JI Z L, et al. Guardrail: automated integrity constraint synthesis from noisy data[J]. Proceedings of the ACM on Management of Data, 2025, 3(4): 1 – 26.
[19] DAI W, HWANG K, FAN J C. Unsupervised anomaly detection for tabular data using deep noise evaluation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(11): 11553 – 11562.
[20] HEIDARI A, MCGRATH J, ILYAS I F, et al. HoloDetect: few-shot learning for error detection[C]//Proceedings of the 2019 International Conference on Management of Data, 2019: 829 – 846.
[21] ABDELAAL M, KTITAREV T, STäDTLER D, et al. SAGED: few-shot meta learning for tabular data error detection [C]//Proceedings of the 27th International Conference on Extending Database Technology (EDBT), 2024: 386 – 398.
[22] NI W, ZHANG K H, MIAO X Y, et al. ZeroED: hybrid zero-shot error detection through large language model reasoning[C]//Proceedings of 2025 IEEE 41st International Conference on Data Engineering (ICDE), 2025: 3126 – 3139.
[23] NARAYAN A, CHAMI I, ORR L, et al. Can foundation models wrangle your data? [J]. Proceedings of the VLDB Endowment, 2022, 16(4): 738 – 746.
[24] YAN M Y, WANG Y S, WANG Y, et al. GIDCL: a graph-enhanced interpretable data cleaning framework with large language models[J]. Proceedings of the ACM on Management of Data, 2024, 2(6): 1 – 29.
[25] NAEEM Z A, AHMAD M S, ELTABAKH M, et al. RetClean: retrieval-based data cleaning using LLMs and data lakes[J]. Proceedings of the VLDB Endowment, 2024, 17(12): 4421 – 4424.
[26] REKATSINAS T, CHU X, ILYAS I F, et al. HoloClean: holistic data repairs with probabilistic inference [J]. Proceedings of the VLDB Endowment, 2017, 10(11): 1190 – 1201.
[27] GE C C, GAO Y J, MIAO X Y, et al. A hybrid data cleaning framework using Markov logic networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(5): 2048 – 2062.
[28] QIN J B, HUANG S F, WANG Y S, et al. BClean: a Bayesian data cleaning system [C]//Proceedings of 2024 IEEE 40th International Conference on Data Engineering (ICDE), 2024: 3407 – 3420.
[29] LEW A, AGRAWAL M, SONTAG D, et al. PClean: Bayesian data cleaning at scale with domain-specific probabilistic programming [C]//Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, 2021: 1927 – 1935.
[30] WANG R H, LI Y L, WANG J. Sudowoodo: contrastive self-supervised learning for multi-purpose data integration and preparation [C]//Proceedings of 2023 IEEE 39th International Conference on Data Engineering (ICDE), 2023: 1502 – 1515.
[31] REIS E, ABDELAAL M, BINNIG C. Generalizable data cleaning of tabular data in latent space[J]. Proceedings of the VLDB Endowment, 2024, 17(13): 4786 – 4798.
[32] NI W, ZHANG K H, MIAO X Y, et al. IterClean: an iterative data cleaning framework with large language models[C]// Proceedings of the ACM Turing Award Celebration Conference-China 2024, 2024: 100 – 105.
[33] WU Y Y, YANG C, ZHU M Y, et al. A zero-training error correction system with large language models [C]// Proceedings of 2025 IEEE 41st International Conference on Data Engineering (ICDE), 2025: 2949 – 2962.
[34] PENG J F, SHEN D R, TANG N, et al. Self-supervised and interpretable data cleaning with sequence generative adversarial networks [J]. Proceedings of the VLDB Endowment, 2022, 16(3): 433 – 446.
[35] PENG J F, CUI H H, SHEN D R, et al. GARF +: self-supervised and interpretable data cleaning with sequence generative adversarial networks [J]. The VLDB Journal, 2025, 34: 64.
[36] DING X O, QIAN Z K, WANG H Z, et al. UniClean: a scalable data cleaning solution for mixed errors based on unified cleaners and optimized cleaning workflow [J]. Proceedings of the VLDB Endowment, 2025, 18(11): 4117 – 4130.
[37] SONG S X, SUN Y, ZHANG A Q, et al. Enriching data imputation under similarity rule constraints [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(2): 275 – 287.
[38] MA Q, LEE W C, FU T Y, et al. MIDIA: exploring denoising autoencoders for missing data imputation[J]. Data Mining and Knowledge Discovery, 2020, 34(6): 1859 – 1897.
[39] CAPPUZZO R, THIRUMURUGANATHAN S, PAPOTTI P. Relational data imputation with graph neural networks[C]// Proceedings of the 27th International Conference on Extending Database Technology (EDBT), 2024: 221 – 233.
[40] YOON J, JORDON J, VAN DER SCHAAR M. GAIN: missing data imputation using generative adversarial nets[C]//Proceedings of the 35th International Conference on Machine Learning, 2018: 5689 – 5698.
[41] QIU W, HUANG Y S B, LI Q Z. IFGAN: missing value imputation using feature-specific generative adversarial networks [C]//Proceedings of 2020 IEEE International Conference on Big Data (Big Data), 2020: 4715 – 4723.
[42] MIAO X Y, WU Y Y, CHEN L, et al. An experimental survey of missing data imputation algorithms [J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 6630 – 6650.
[43] HAYAT A, HASAN M R. A context-aware approach for enhancing data imputation with pre-trained language models[C]//Proceedings of the 31st International Conference on Computational Linguistics, 2025: 5668 – 5685.
[44] YANG C Y, LUO Y Y, CUI C X, et al. Data imputation with limited data redundancy using data lakes[J]. Proceedings of the VLDB Endowment, 2025, 18(10): 3354 – 3367.
[45] HE X R, BAN Y K, ZOU J R, et al. LLM-Forest: ensemble learning of LLMs with graph-augmented prompts for data imputation [C]//Findings of the Association for Computational Linguistics: ACL 2025, 2025: 6921 – 6936.
[46] Valentine. (Schema-) matching dataframes made easy[EB/OL]. [2026 – 01 – 27]. https://delftdata. github. io/

valentine/.
[47] NARGESIAN F, ZHU E K, PU K Q, et al. Table union search on open data [J]. Proceedings of the VLDB Endowment, 2018, 11(7): 813 – 825.
[48] POESS M, RABL T, JACOBSEN H A, et al. TPC-DI: the first industry benchmark for data integration[J]. Proceedings of the VLDB Endowment, 2014, 7(13): 1367 – 1378.
[49] KONDA P, DAS S, SUGANTHAN G C P, et al. Magellan: toward building entity matching management systems [J]. Proceedings of the VLDB Endowment, 2016, 9(12): 1197 – 1208.
[50] MUDGAL S, LI H, REKATSINAS T, et al. Deep learning for entity matching: a design space exploration[C]//Proceedings of the 2018 International Conference on Management of Data, 2018: 19 – 34.
[51] BERTINO E, ATZENI P, TAN K L, et al. Evaluation of entity resolution approaches on real-world match problems[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 484 – 493.
[52] DONG Y Y, TAKEOKA K, XIAO C, et al. Efficient joinable table discovery in data lakes: a high-dimensional similarity-based approach [C]//Proceedings of 2021 IEEE 37th International Conference on Data Engineering (ICDE), 2021: 456 – 467.
[53] HU X M, LEI C, QIN X, et al. PolyJoin: semantic multi-key joinable table search in data lakes [C]//Findings of the Association for Computational Linguistics: NAACL 2025, 2025: 384 – 395.
[54] KOUTRAS C, ZHANG J N, QIN X, et al. OmniMatch: joinability discovery in data products[J]. Proceedings of the VLDB Endowment, 2025, 18(11): 4588 – 4601.
[55] TU J H, FAN J, TANG N, et al. Unicorn: a unified multi-tasking model for supporting matching tasks in data integration[J]. Proceedings of the ACM on Management of Data, 2023, 1(1): 1 – 26.
[56] ZHANG Y J, FLORATOU A, CAHOON J, et al. Schema matching using pre-trained language models[C]//Proceedings of 2023 IEEE 39th International Conference on Data Engineering (ICDE), 2023: 1558 – 1571.
[57] DONG Y Y, XIAO C, NOZAWA T, et al. DeepJoin: joinable table discovery with pre-trained language models[J]. Proceedings of the VLDB Endowment, 2023, 16(10): 2458 – 2470.
[58] HU X M, QIN X, LEI C, et al. DiscoverGPT: multi-task fine-tuning large language model for related table discovery[C]//Findings of the Association for Computational Linguistics: NAACL 2025, 2025: 358 – 373.
[59] LIN Y M, HE Y Y, CHAUDHURI S. Auto-BI: automatically build BI-models leveraging local join prediction and global schema graph [J]. Proceedings of the VLDB Endowment, 2023, 16(10): 2578 – 2590.
[60] LAI E Y, HE Y Y, CHAUDHURI S. Auto-Prep: holistic prediction of data preparation steps for self-service business intelligence [J]. Proceedings of the VLDB Endowment, 2025, 18(7): 2212 – 2225.
[61] LIU X Q, WANG R H, SONG Y, et al. GRAM: generative retrieval augmented matching of data schemas in the context of data security [C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024: 5476 – 5486.
[62] WANG Y H, DING B L, ZHU R, et al. Language models are explorers for join discovery on data lakes[C]//Proceedings of the 2025 SIAM International Conference on Data Mining (SDM), 2025: 398 – 408.
[63] LIU Y R, PENA E H M, SANTOS A, et al. Magneto: combining small and large language models for schema matching[J]. Proceedings of the VLDB Endowment, 2025, 18(8): 2681 – 2694.
[64] HUANG J C, HU W, BAO Z F, et al. Deep entity matching with adversarial active learning [J]. The VLDB Journal, 2023, 32: 229 – 255.
[65] NIE H, HAN X P, HE B, et al. Deep sequence-to-sequence entity matching for heterogeneous entity resolution [C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019: 629 – 638.
[66] WANG Z Y, SISMAN B, WEI H, et al. CorDEL: a contrastive deep learning approach for entity linkage [C]// Proceedings of 2020 IEEE International Conference on Data Mining (ICDM), 2020: 1322 – 1327.
[67] 樊峰峰, 李战怀, 陈群, 等. 一种基于离群点检测的自动实体匹配方法[J]. 计算机学报, 2017, 40(10): 2197 – 2211.
FAN F F, LI Z H, CHEN Q, et al. An outlier-detection based approach for automatic entity matching [J]. Chinese Journal of Computers, 2017, 40(10): 2197 – 2211. (in Chinese)
[68] TEONG K S, SOON L K, SU T T. Schema-agnostic entity matching using pre-trained language models[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020: 2241 – 2244.
[69] LI Y L, LI J F, SUHARA Y, et al. Deep entity matching with pre-trained language models [J]. Proceedings of the VLDB Endowment, 2020, 14(1): 50 – 60.
[70] TRABELSI M, CHEN Z Y, ZHANG S, et al. StruBERT: structure-aware BERT for table search and matching [C]// Proceedings of the ACM Web Conference 2022, 2022: 442 – 451.
[71] FAN M H, HAN X Y, FAN J, et al. Cost-effective in-context learning for entity resolution: a design space exploration[C]// Proceedings of 2024 IEEE 40th International Conference on Data Engineering (ICDE), 2024: 3696 – 3709.
[72] PEETERS R, STEINER A, BIZER C. Entity matching using large language models [C]//Proceedings of the 28th International Conference on Extending Database Technology (EDBT), 2025.
[73] WANG T S, CHEN X Y, LIN H Y, et al. Match, compare, or select? An investigation of large language models for entity matching [C]//Proceedings of the 31st International Conference on Computational Linguistics, 2025: 96 – 109.
[74] SRINIVAS K, DOLBY J, ABDELAZIZ I, et al. Lakebench: benchmarks for data discovery over datalakes [EB/OL]. (2023 – 07 – 09)[2026 – 01 – 27]. https://arxiv.org/abs/2307.04217.
[75] CHEN Z Y, TRABELSI M, HEFLIN J, et al. Table search using a deep contextualized language model[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020: 589 – 598.
[76] LIN R, CHOPRA B, LIN W J, et al. Rethinking dataset discovery with datascout[C]//Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology, 2025: 1 – 16.
[77] FAN G, WANG J, LI Y L, et al. Semantics-aware dataset discovery from data lakes with contextualized column-based representation learning [J]. Proceedings of the VLDB Endowment, 2023, 16(7): 1726 – 1739.
[78] KHATIWADA A, KOKEL H, ABDELAZIZ I, et al.

TabSketchFM: sketch-based tabular representation learning for data discovery over data lakes[C]//Proceedings of 2025 IEEE 41st International Conference on Data Engineering (ICDE), 2025: 1523 – 1536.

[79] BALAKA M I L, ALEXANDER D, WANG Q M, et al. Pneuma: leveraging LLMs for tabular data representation and retrieval in an end-to-end system[J]. Proceedings of the ACM on Management of Data, 2025, 3(3): 1 – 28.

[80] CHEN S A, CHEN Y F, EISENSCHLOS J, et al. TableRAG: million-token table understanding with language models[C]//Proceedings of Advances in Neural Information Processing Systems, 2024: 74899 – 74921.

[81] CASTRO FERNANDEZ R, ABEDJAN Z, KOKO F, et al. Aurum: a data discovery system[C]//Proceedings of 2018 IEEE 34th International Conference on Data Engineering (ICDE), 2018: 1001 – 1012.

[82] CASTRO FERNANDEZ R, MANSOUR E, QAHTAN A A, et al. Seeping semantics: linking datasets using word embeddings for data discovery[C]//Proceedings of 2018 IEEE 34th International Conference on Data Engineering (ICDE), 2018: 989 – 1000.

[83] NARGESIAN F, PU K Q, ZHU E K, et al. Organizing data lakes for navigation[C]//Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020: 1939 – 1950.

[84] NARGESIAN F, PU K, GHADIRI-BASHARDOOST B, et al. Data lake organization[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(1): 237 – 250.

[85] OUELLETTE P, SCIORTINO A, NARGESIAN F, et al. RONIN: data lake exploration[J]. Proceedings of the VLDB Endowment, 2021, 14(12): 2863 – 2866.

[86] SUHARA Y, LI J F, LI Y L, et al. Annotating columns with pre-trained language models[C]//Proceedings of the 2022 International Conference on Management of Data, 2022: 1493 – 1503.

[87] CHEN J Y, JIMÉNEZ-RUIZ E, HORROCKS I, et al. ColNet: embedding the semantics of web tables for column type prediction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 29 – 36.

[88] SUN Y S, XIN H, CHEN L. RECA: related tables enhanced column semantic type annotation framework[J]. Proceedings of the VLDB Endowment, 2023, 16(6): 1319 – 1331.

[89] WANG Y B, XIN H, CHEN L. KGLink: a column type annotation method that combines knowledge graph and pre-trained language model[C]//Proceedings of 2024 IEEE 40th International Conference on Data Engineering (ICDE), 2024: 1023 – 1035.

[90] KHURANA U, GALHOTRA S. Semantic concept annotation for tabular data[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021: 844 – 853.

[91] XIAN Y K, ZHAO H D, LEE T Y, et al. EXACTA: explainable column annotation[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021: 3775 – 3785.

[92] KAYALI M, LYKOV A, FOUNTALIS I, et al. CHORUS: foundation models for unified data discovery and exploration[J]. Proceedings of the VLDB Endowment, 2024, 17(8): 2104 – 2114.

[93] KORINI K, BIZER C. Evaluating knowledge generation and self-refinement strategies for LLM-based column type annotation[C]//Proceedings of Advances in Databases and Information Systems, 2025: 111 – 127.

[94] WEI L L, XIAO G R, BALAZINSKA M. RACOON: an LLM-based framework for retrieval-augmented column type annotation with a knowledge graph[EB/OL]. (2024 – 11 – 01)[2026 – 01 – 27]. https://arxiv.org/abs/2409.14556.

[95] EDGE D, TRINH H, CHENG N, et al. From local to global: a graph RAG approach to query-focused summarization[EB/OL]. (2025 – 02 – 19)[2026 – 01 – 27]. https://arxiv.org/abs/2404.16130.

[96] GUO Z R, XIA L H, YU Y H, et al. LightRAG: simple and fast retrieval-augmented generation[C]//Findings of the Association for Computational Linguistics: EMNLP 2025, 2025: 10746 – 10761.

[97] LIU Z L, HUANG Y, YU X W, et al. DeID-GPT: zero-shot medical text de-identification by GPT-4[EB/OL]. (2025 – 11 – 28)[2026 – 01 – 27]. https://arxiv.org/abs/2303.11032.

[98] LUKAS N, SALEM A, SIM R, et al. Analyzing leakage of personally identifiable information in language models[C]//Proceedings of 2023 IEEE Symposium on Security and Privacy (SP), 2023: 346 – 363.

[99] ZENG S L, ZHANG J K, HE P F, et al. Mitigating the privacy issues in retrieval-augmented generation (RAG) via pure synthetic data[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025: 24527 – 24558.

[100] RAJBHANDARI S, RUWASE O, RASLEY J, et al. ZeRO-infinity: breaking the GPU memory wall for extreme scale deep learning[C]//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2021: 1 – 14.

[101] GRAUR D, MRAZ O, LI M Y, et al. Pecan: cost-efficient ML data preprocessing with automatic transformation ordering and hybrid placement[C]//Proceedings of the 2024 USENIX Annual Technical Conference, 2024: 649 – 665.

[102] AKELLA A, MANATKAR A, NARAYANAM K, et al. CodeGenWrangler: data wrangling task automation using code-generating models[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, 2025: 949 – 960.

[103] LIU L, HASEGAWA S, SAMPAT S K, et al. AutoDW: automatic data wrangling leveraging large language models[C]// Proceedings of 2024 39th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2024: 2041 – 2052.

[104] WANG L, MA C, FENG X Y, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18: 186345.

[105] EHTESHAM A, SINGH A, GUPTA G K, et al. A survey of agent interoperability protocols: model context protocol (MCP), agent communication protocol (ACP), agent-to-agent protocol (A2A), and agent network protocol (ANP)[EB/OL]. (2025 – 05 – 04)[2026 – 01 – 27]. https://arxiv.org/abs/2505.02279.

[106] HOU X Y, ZHAO Y J, WANG S N, et al. Model context protocol (MCP): landscape, security threats, and future research directions[EB/OL]. (2025 – 03 – 30)[2026 – 01 – 27]. https://arxiv.org/abs/2503.23278.