软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(10):4439−4462 [doi: 10.13328/j.cnki.jos.006884]

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

知识赋能的新一代信息系统研究现状、发展与挑战*

朱 迪[1], 张博闻[1], 程雅琪[1], 刘昕悦[1], 吴文隆[1], 王铁鑫[1], 文 浩[2], 李博涵[1]

[1](南京航空航天大学 计算机科学与技术学院, 江苏 南京 211106)

[2](南京航空航天大学 航空学院, 江苏 南京 210016)

通信作者: 李博涵, E-mail: bhli@nuaa.edu.cn

摘 要: 信息系统的发展目前正处于感知智能迈向认知智能的关键阶段, 传统信息系统难以满足发展要求, 数字化转型势在必行. 数字线索(digital thread)是面向全生命周期的数据处理框架, 通过连接生命周期的各阶段数据, 实现物理世界与数字空间的映射与分析. 知识图谱(knowledge graph)是结构化的语义知识库, 以符号形式描述物理世界中的概念及其相互关系, 通过知识驱动形成体系化的构建与推理流程. 两者对知识赋能的信息系统研究具有重要意义. 综述了知识赋能的新一代信息系统的研究现状、发展与挑战. 首先, 从数字线索系统出发, 介绍数字线索的概念和发展, 分析数字线索的六维数据构成和 6 个数据处理阶段; 然后介绍知识图谱系统, 给出普遍认同的知识图谱的定义和发展, 概括知识图谱的架构与方法; 最后, 分析和探索数字线索与知识图谱结合的方向, 列举 KG4DT (knowledge graph for digital thread)和 DT4KG (digital thread for knowledge graph)的受益方向, 对未来知识赋能的新一代信息系统提出开放问题.

关键词: 数字线索; 知识图谱; 知识赋能; 信息系统

中图法分类号: TP18

中文引用格式: 朱迪, 张博闻, 程雅琪, 刘昕悦, 吴文隆, 王铁鑫, 文浩, 李博涵. 知识赋能的新一代信息系统研究现状、发展与挑战. 软件学报, 2023, 34(10): 4439−4462. http://www.jos.org.cn/1000-9825/6884.htm

英文引用格式: Zhu D, Zhang BW, Cheng YQ, Liu XY, Wu WL, Wang TX, Wen H, Li BH. Survey on Knowledge Enabled New Generation Information Systems. Ruan Jian Xue Bao/Journal of Software, 2023, 34(10): 4439−4462 (in Chinese). http://www.jos.org.cn/1000-9825/6884.htm

Survey on Knowledge Enabled New Generation Information Systems

ZHU Di[1], ZHANG Bo-Wen[1], CHENG Ya-Qi[1], LIU Xin-Yue[1], WU Wen-Long[1], WANG Tie-Xin[1], WEN Hao[2], LI Bo-Han[1]

[1](College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

[2](College of Aerospace Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: The development of information system is in the critical stage from perceptual intelligence to cognitive intelligence. Traditional information systems are difficult to meet the development requirements, and digital transformation is imperative. Digital thread is a full-life-cycle data processing framework that maps and analyzes the physical world and digital space by connecting data from different life cycle stages. Knowledge graph is a structured semantic knowledge base that describes concepts and relationships in the physical world in the form of symbols, forming systematic construction and reasoning process driven by knowledge. Both of them are of great significance to the research of knowledge enabled information system. This study reviews the current research status, development and challenges of the new generation knowledge enabled information system. First, starting from the digital thread system, the concept and

* 基金项目: 国家重点研发计划(2020YFB1708100); “十四五”民用航天技术预先研究项目(D020101); 国家自然科学基金(62172351); 高安全系统的软件开发与验证技术工业和信息化部重点实验室课题(NJ2018014); 南京航空航天大学前瞻布局科研专项资金

本文由“知识赋能的信息系统”专题特约编辑高宏教授、陈华钧教授、赵翔教授、李瑞轩教授推荐.

收稿时间: 2022-07-05; 修改时间: 2022-08-18, 2022-12-14; 采用时间: 2022-12-28; jos 在线出版时间: 2023-01-13

development of digital thread, the six-dimensional data structure and six data processing stages of digital thread are introduced. Then, the generally accepted definitionand the development of knowledge graph system are given, and the structure and methods of knowledge graph are summarized. Finally, the direction of combining digital thread with knowledge graph is analyzed and explored, the benefits of KG4DT (knowledge graph for digital thread) and DT4KG (digital thread for knowledge graph) are listed, and the open questions are raised about the new generation of knowledge enabled information systems in the future.

Key words: digital thread; knowledge graph; knowledge enabling; information system

人工智能的深化, 使得信息系统的研究重点由“感知智能”转向“认知智能”, 要求在具备多信息感知能力的前提下, 有效实现人类思维理解、知识共享、因果推理等认知智能行为. 新一代信息系统以数据与知识为核心要素, 通过数据与知识双驱动形式实现知识赋能, 不仅使用常规数据驱动方法构建模型, 同时关联用户行为、常识知识以及知识推理等实现主动的认知与推理. 因此, 如何寻求结合数据使能与知识驱动的相关方法技术成为巨大的难题.

数字线索(digital thread, DT)拥有体系化的数据处理能力, 具备“全部元素建模定义、全部数据采集分析、全部决策仿真评估”的设计特性, 可以有效破解数据孤岛问题. 不同于常规的数据驱动方法, 数字线索通过先进的数据获取和虚拟建模技术, 建立物理世界与数字空间之间的映射关系, 最终实现全生命周期数据的整合、分析和决策. 知识图谱(knowledge graph, KG)提供强大的语义表达能力、存储能力和推理能力, 可以“自下而上”地形成知识数据挖掘、抽取、推理和应用的知识工程体系, 实现系统化的知识构建与推理流程.

新一代信息系统可以结合数字线索与知识图谱以实现数据与知识双驱动的知识赋能. 其中, 数字线索以多种传感器获取信息并孪生化实体数据, 通过虚实映射的数据处理和分析手段为信息系统提供底层支撑. 同时, 知识图谱以结构化的数据标签描述客观世界, 通过基于语义的上下文搜索形式为信息系统增强顶层认知能力并最终实现知识赋能. 如图 1 所示, 数字线索与知识图谱的有机融合为知识赋能的新一代信息系统提供了新的研究视野.

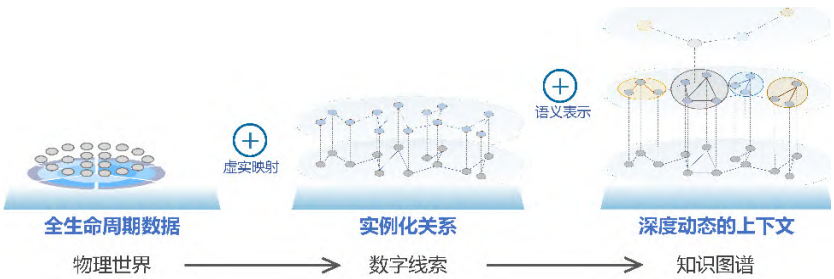

图 1 知识赋能的新一代信息系统发展形式

1 数字线索使能的信息系统

本节从数字线索的定义与发展起始, 然后依次介绍数字线索的体系架构、常见的数字线索数据构成、数据处理方法和应用领域.

1.1 数字线索的定义与发展

1.1.1 数字线索的定义

数字线索(也被译为数字线程、数字主线、数字纽带等, 本文统一采用数字线索翻译方式)是一种面向全生命周期的数据分析框架. 基于高精度的数字系统模型(digital system model, DSM), 数字线索提供面向系统全生命周期的数据管理、综合和分析能力, 具有“全部元素建模定义、全部数据采集分析、全部决策仿真评估”的技术特点. 数字线索也是数字孪生(digital twin)[1]的核心数据使能技术, 综合各类数据、信息和工程知识, 进行物理世界与数字空间的映射建模, 实现数据处理与信息交互. 本质上, 数字线索是建立物理实体的数字化镜像(digital mirror, DM)[2]. 数字线索已在多领域展开研究与应用, 包括航空航天、汽车驾驶、智能电网等[3−9],

是国家“十四五”规划纲要中构建“数字孪生城市”的重要数据赋能技术[10].

数字线索的主要表现形式为 DSM, 采用多种建模语言, 包括统一建模语言(unified modeling language, UML)[11]、系统建模语言(systems modeling language, SysML)[12]、本体语言(ontology language)[13]等, 对系统的结构模型、行为模型、需求模型和参数模型等[14]进行定义, 最终生成树状结构模型体系, 实现对工程系统的数字化表达.

1.1.2 数字线索发展

数字线索最早可追溯至 20 世纪 70 年代的美国军事航空航天领域, 早期承担数字孪生系统的数据使能部分, 经过技术演化和实践探索发展至今, 数字线索已在多领域展开应用, 数字线索的发展时间轴如图 2 所示.

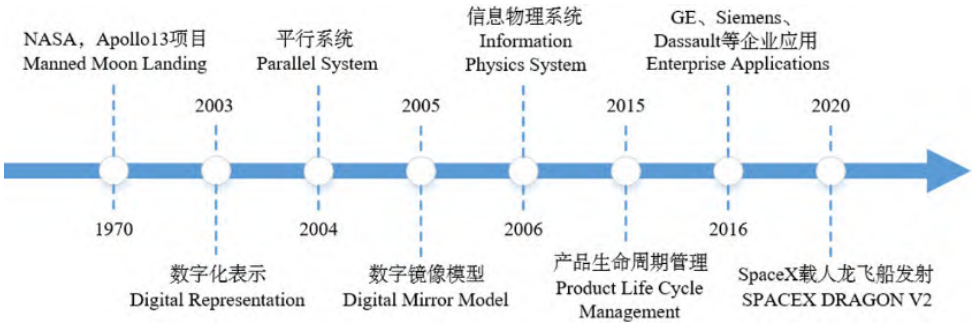

图 2 数字线索发展时间轴

1970 年, 美国航空航天局(National Aeronautics and Space Administration, NASA)在载人登月任务 Apollo13 项目中首次了使用物理实体作为孪生体, 模拟登月突发状况, 训练宇航员做出正确决策[15]. 2003 年, 美国 Michael Grieves 提出了“与物理产品等价的虚拟数字化表达”概念, 并于 2005 年提出了由物理空间、虚拟空间和物理-虚拟连接信息共同组成的信息镜像模型(information mirroring model). 数字线索雏形包含于数字孪生概念中.

2013 年, 美国空军(United States Air Force, USAF)发布了 Global Science and Technology Vision, 将数字线索拆分为独立领域[16]. 同年, NASA 与美国空军研究实验室(Air Force Research Laboratory, AFRL)联合提出了美国航天飞行器发展指导, 数字线索被定义为关键数据使能技术. 2015 年, NASA 任务中心创始人 Kraft 提出了应用数字线索的产品生命周期管理(product lifecycle management, PLM), 将数字线索技术抬到新的高度. 2020 年, SpaceX 载人龙飞船“自由号”成功发射, 数字线索技术实现落地应用.

2014 年左右, 美、德、英、法等国家将数字线索引入工业界的数字化转型和信息管理研究. 2016 年, GE、Siemens、Dassault 等公司尝试应用数字线索. 2017 年起, 国内学者开始研究数字线索. 陶飞提出了数字孪生模型[17]. 中航工业等研究机构在飞行器设计中实现了基于 MBSE 的数字线索应用[18]. 北京航空航天大学可靠性工程研究所搭建了 MBRSE 数字线索平台, 融入数字化研制环境, 具有国际领先水平.

此外, 美国国家科学基金会提出的信息物理系统(cyber-physical system, CPS)概念, 使用人机接口与物理世界进行交互[19]; 中国科学院王飞跃等人在《平行系统方法与复杂系统的管理和控制》中提出了平行系统概念, 利用现实和虚拟组建共同系统管理复杂系统[20]; 陶飞等人提出了数字孪生车间(digital twin shop-floor)等[21]. 这些系统本质上都是通过对复杂物理世界数据建立虚拟关联模型, 实现物理空间与数字空间的映射与交互.

1.2 数字线索体系架构

数字线索体系架构如图 3 所示, 根据流程逻辑可划分感知层、数据层、模型层、功能层和应用层.

感知层中, 数字线索利用物联网基础设施, 对物理实体进行精准测量感知. 数据层以多维异构数据为基础, 将数据和工程知识组成权威真相源(authoritative source of truth, AST), 进行数据采集、数据传输、数据管

理等操作. 模型层以统一建模为核心, 通过系统工程方法实现物理实体在数字空间的虚拟映射, 充分利用下层采集数据, 支撑上层功能要求. 功能层是数字线索的直接价值体现部分, 主要实现系统认知、状态分析、辅助决策等功能, 助力应用层中数字线索在各类场景的价值实现, 包括智慧城市、智慧医疗、智慧工业等.

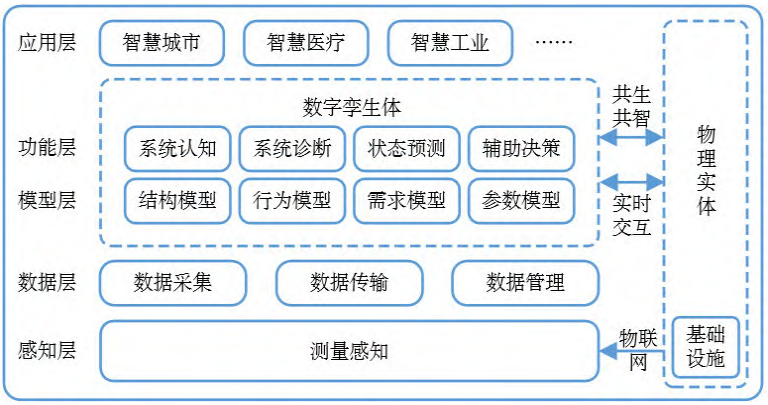

图 3 数字线索体系架构

1.3 数字线索数据构成

数字线索的主要数据构成可分为 6 个组成部分, 包括物理实体、虚拟模型、服务数据、融合数据、连接数据和领域知识, 数字线索的六维数据构成概念如公式(1)所示:

$$M_{TD}=(PE,VE,SD,DK,FD,CD) \tag{1}$$

其中, *PE* 表示物理实体, *VE* 表示虚拟实体, *SD* 表示服务数据, *DK* 表示领域知识, *FD* 表示融合数据, *CD* 表示各数据部分间的连接数据. 根据公式(1), 数字线索的数据构成如图 4 所示.

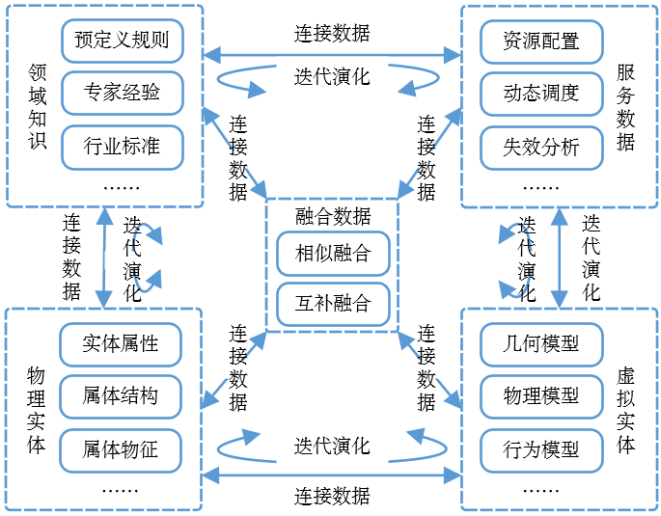

图 4 数字线索的数据构成

物理实体(physical entity, PE), 是指在物理世界中的具有特定属性、结构和特征的对象[22]. PE 是数字线索的主要数据采集对象, 分为静态信息(例如标识、特征、时间表等)和动态信息(例如状态、位置、关系等), 根据功能结构, 可划分为单元级(unit) PE, 系统级(system) PE, 复杂系统级(system of systems) PE.

虚拟实体(virtual entity, VE), 是指在数字空间中从不同时空尺度描述和复现物理实体, 包括几何模型、物理模型、行为模型和维护规则等. 本质上是构建与物理实体数据耦合的数字模型.

服务数据(service data, SD), 是指服务化封装各类数字线索所需的数据、算法、仿真结果等, 可分为业务服务数据(business services, BServices)和功能服务数据(functional services, FServices)[23]: 前者面向跨领域场景中的操作指导、资源调度、决策支持等业务活动; 后者通过模型管理、数据处理、综合连接等服务, 支撑数字线索内部功能运行.

领域知识(domain knowledge, DK), 是指在跨领域系统内综合各行业基础知识, 用于指导系统搭建、信息交互、数据处理等. 主要知识包括专家经验、预定义规则、行业标准以及系统运行过程中产生的新知识等.

融合数据(fusion data, FD), 是指融合处理上述 4 种数据的混合数据, 通过相似或互补融合的方式, 结合加权方法、神经网络等融合算法, 使底层多源数据从多角度相互补充, 形成统一描述, 实现信息的共享与增值.

连接数据(connection data, CD), 是指其他数据的导出和传输数据, 主要用于各数据部分的互联互通, 通过不断的迭代演化实现高一致性, 发现并消除数字线索各生命周期数据中的扰动因素, 保持数据一致和联动.

1.4 数字线索方法与技术

数字线索对大规模复杂数据具有强大的处理与分析能力, 需要运用多种数字线索数据处理方法与技术.

1.4.1 数据获取

数据获取(data collection), 数字线索主要获取物理实体、虚拟实体、服务数据、领域知识这 4 类数据, 在后续过程中衍生出融合数据和连接数据. 首先, 借助用于仿真的微机电系统(micro-electro-mechanical system, MEMS)和 MATLAB、3DMAX、ProE 等模型软件[24]对几何、物理、行为和规则等数据进行采集. 然后, 运用有限状态机(finite state machine, FSM)、马尔可夫链(Markov chain, MC)、神经网络(artificial neural networks, ANNs)等方法, 基于本体进行数字建模, 涉及问题模型、评估模型、决策模型等. 最后, 建立基础信息物理系统(CPS)以抽取重要信息要素, 例如产品生命周期管理(product lifecycle management, PLM)系统、产品数据管理(product data management, PDM)系统、制造执行系统(manufacturing execution system, MES)等覆盖物理实体生命周期的数据系统[25].

1.4.2 数据存储

数据存储(data storage), 数字线索需将不同领域场景和对象中采集的各类数据转化为统一共享模式, 以支撑数据库管理技术(database management technology, DMT)的各项功能, 例如增加、删除、更改和查询等. 首先, 以标准表示形式进行数据记录, 包括内容、格式、接口等. 然后进行数据格式转化(例如: Dang 等人[26]提出的利用中间数据格式设计弹性数据转换框架, 进一步解析数据的方法; Marjan Hossein 等人[27]提出的基于语义处理数据, 设计数据格式转换的自动映射(automatic mapping)等), 并使用多种建模语言进行信息建模, 包括 UML、SysML、Ontology Language 等.

1.4.3 数据连接

数据连接(data connection), 数字线索将不同数据部分进行连接以支持实时交互: 首先, 通过系统建模与仿真的校核、验证与确认活动(verification, validation and accreditation, VV&A), 确保模型仿真的效率和准确性; 然后, 利用过滤算法(filtering algorithm, FA)、降维算法(descending algorithm, DA)、回归算法(regression algorithm, RA)等方法, 去除数据噪声和冗余. 同时, 数字线索不断添加、处理、合并新的数据, 借助 ANSYS、ABAQUS、MARC 等大型通用有限元分析软件(finite element analysis, FEA), 可实现信息数据更新. 此外, 也可以利用机器学习算法对已对齐数据进行修正和补充, 计算信息熵(information entropy)以评估更新效率.

1.4.4 数据融合

数据融合(data fusion), 数字线索面向物理世界与数字空间进行的数据融合[28], 通过对前期数据的统一分类、关联、集成和融合等操作, 实现信息共享与增值. 数据融合可分为前端融合(early-fusion, EF)或数据水平融合(data-level fusion)、后端融合(late-fusion, LF)或决策水平融合(decision-level fusion)和中间融合(intermediate-fusion, IF)[29]等方法, 通过降低数据融合的信息熵, 以减少数据的不确定性、随机性和模糊性. 此外, 神经网络、贝叶斯方法、集成学习(ensemble learning, EL)等技术的引入, 同样可以提高数字线索数据融合的准确性和可靠性.

1.4.5 数据演化

数据演化(data evolution), 数字线索集成各类数据, 进行多领域多尺度融合建模, 以支持内在知识发现. 首先, 对初始数据进行数据拟合, 包括最小二乘法(least square, LS)、差分进化法(differential evolution, DE)、粒子群优化法(particle swarm optimization, PSO)等. 然后, 从不同领域视角融合建模, 保持采集数据与系统数据高度一致, 并依据各时间尺度和空间尺度模拟众多科学问题, 同时满足空间尺度、时间尺度和耦合范围要求. 最后, 运用 *K*-means 算法[30]、Apriori 算法[31]等方法挖掘数据关系, 基于逻辑规则、表示学习等知识推理方法, 推导内在知识, 抽象数据关联以支持顶层数字服务.

1.4.6 数字服务

数字服务(data servitization), 数字线索封装数据资源并根据用户需求提供访问服务. 首先, 通过语义网络(semantic Web, SW)、资源描述框架(resource description framework, RDF)等技术[32], 封装数字线索系统内的各类数据资源, 将其转化为相应服务. 然后, 根据约束条件(例如时间、成本、收益等)将各个独立的服务进行组合, 提供集成解决方案. 相关方法包括需求分解、相似度匹配、多目标优化等[33]. 此外, 数字线索也正逐步运用虚拟现实(virtual reality, VR)[34]、增强现实(augmented reality, AR)[35]和混合现实(mixed reality, MR)[36]等技术, 提供物理实体与虚拟模型的可视化映射.

综上, 数字线索数据的处理方法和技术分析总结见表 1.

表 1 数字线索数据处理方法与技术

处理方法	核心思路	相关技术
数据获取	采集物理世界的各类数据, 衍生融合数据和连接数据	仿真微机电系统 MEMS, 建模软件(MATLAB, 3DMAX, ProE 等), 建模方法(FSM, MC, ANNs), CPS 系统(PLM, PDM, MES, SCADA 等)
数据存储	多模态数据转化和统一存储	数据库管理技术 DMT, 资源描述框架 RDF, 建模语言(UML, SysML 等)
数据连接	数据传输, 数据降噪和一致性评估	验证、确认和认可活动(VV&A), 连接算法(FA, DA, RA 等), FEA 软件(ANSYS, ABAQUS, MARC 等)
数据融合	物理世界与数字空间的数据融合	数据融合方法(EF, IF, LF 等),
数据演化	数据集成, 多领域 多尺度融合建模, 知识发现	数据拟合方法(LS, DE, PSO 等), 数据挖掘算法(*K*-means, Apriori 等)
数字服务	集成数字功能, 提供个性化数字服务	资源封装(SW, RDF 等) 功能集成技术(需求分解、相似度匹配等) 3R 技术(AR, VR, MR)

1.5 数字线索应用

数字线索具有全生命周期的综合数据处理能力, 能够有效加速数据、信息与通信的融合. SIEMENS、DASSAULT、PTC、Ansys、AVEVA、Microsoft、Unity 等国际企业不断推进数字线索应用落地, 国内的主流厂商主要包括 51world、优诺科技、美云智数、华力创通、华龙迅达等. 国内外企业基于自身技术特性, 构建了包括 GEPredix、SIEMENS COMOS、PTCThingWorx 和 Jupiter Digital Twin Platform 等综合性数字线索信息系统[37−41], 典型应用例如 GEPredix[42], 通过整合 APM、OPM、iFIX、Proficy、Historian 等软件服务, 并结合设备模型与数据分析构建数字孪生体; SIEMENS COMOS Platform[38]覆盖工业生命周期, 通过多种软件通信接口, 将需求分析、设计加工及后期运维管理数字化转移至同一工程平台进行综合管理. 基于数字线索的信息系统已在多个应用场景展开技术应用, 可总结为城市治理、工业制造和医疗保健等.

(1) 城市治理

现代城市的人口与规模飞速增长, 引入数字线索的智能管理结构支撑数字孪生的模拟仿真, 带来新的城市治理格局[43,44]. 借助城市基础设施和人类行为学习, 数字线索检测城市系统内的状态变化并预测可能的未来行为[45]. 数字线索同样可以优化城市车辆交通, 实现实时数据传输与管理, 例如车辆行驶速度、位置和路

线[46]. 应用于自动驾驶和车辆互联, 实现高效的车群管理和安全的交通疏导[47]. 此外, 数字线索的灵活应用, 也可以助力城市环境治理, 通过构建基于数字线索的环境评估模型, 检测空气污染物和温室气体排放[48], 可以有效地将污染排放降低至合格水平, 保护城市生态系统.

(2) 工业制造

对于现有工业制造生产, 数字线索提供更高效率、更低成本的数据治理选择. 结合物联网基础设备, 现有模块化的制造业产品设计和生产单元在引入数字线索智能框架后, 可以融入数字线索概念并串联产品制造的各生命周期阶段[49], 提供制造设备的零部件到完整组件的全生命周期可视化显示[50], 智能高效地预测设备生产活动, 创建自主生产系统. 同时, 数字线索支持跨周期跨领域的设计人员信息交互, 协助专业人员进行生产、编程和控制, 包括智能订单调度、系统决策支持等操作, 对生产生命周期不同阶段的信息流进行整合、分析和处理, 更高效地实现数据交流. 此外, 数字线索也应用于设备设计, 以持续评估生产系统并实现自动化的独立数据采集.

(3) 医疗保健

医疗保健领域中的数字线索系统借助医疗物联网设备, 可以形成新型医学模拟方法, 利用多学科、多物理和多尺度模型相结合的数字线索技术, 配合医疗体系, 提供稳健、精确和有效的医疗服务. 大规模数字模型的处理分析可以深度挖掘病理关联, 额外补充医疗人员的医疗领域知识. 借助患者、云服务、健康中心等系统性组合, 数字线索可以对患者的身体数据进行实时感知[51], 跟踪饮食、睡眠、运动等医疗保健数据, 评估健康状态, 给出诊断结果和治疗方案. 此外, 结合 VR、AR 以及 MR 等技术, 数字线索使得采用虚拟仿真形式实现远程手术成为可能[52].

2 知识图谱驱动的信息系统

本节以知识图谱的定义与发展为出发点, 然后依次介绍知识图谱的体系架构、方法技术和应用方向.

2.1 知识图谱的定义与发展

2.1.1 知识图谱的定义

知识图谱(knowledge graph, KG)是一种以图形式表现客观世界的实体(例如人、概念、事物等)及实体关系的知识库, 其中, 节点(实体)和边(实体间关系)组成多边关系图, 本质上是一种具有有向图结构的语义网络(semantic network, SN), 是关系的最有效的表示方式之一. 知识图谱旨在对多结构类型的复杂数据进行概念、实体和关系抽取, 构建实体关系的可计算模型. 根据覆盖范围和应用领域, 知识图谱可分为通用知识图谱和行业知识图谱: 前者侧重于构建行业常识性知识, 应用于搜索引擎或推荐系统, 注重广度, 强调融合更多的实体; 后者面向特定领域, 依靠特定数据构建不同的行业知识图谱, 对企业提供内部的知识化服务, 实体的属性与数据模式丰富. 目前, 知识图谱已在多领域得到广泛应用, 包括语义搜索[53]、智能问答[54]、智能推荐[55]等方面, 是认知智能信息系统的重要发展技术.

知识图谱的主要表现形式为三元组形式, 即 $G=\{E,R,F\}$, 其中, E 表示 $\{e_1,e_2,\ldots,e_E\}$, 是知识库中的实体集合, 是具有可区别性且独立存在的某种事物, 共包含 $|E|$ 种不同的实体概念; R 表示关系集合 $\{r_1,r_2,\ldots,r_R\}$, 即知识图谱中的边集合, 代表知识图谱中节点之间的各种联系, 共包含 $|R|$ 种不同关系; F 表示事实集合, $\{f_1,f_2,\ldots,f_F\}$ 中每一个事实 f 可定义为一个三元组 $(h,r,t)\in f$, 其中, h 表示头实体, t 表示尾实体, r 表示二者之间的关系.

2.1.2 知识图谱发展

知识图谱可追溯至 20 世纪 60 年代提出的语义网络, 期间经历一系列演化, 形成如今的现代知识图谱. 知识图谱技术发展时间轴如图 5 所示.

- 1965 年, Feigenbaum 提出了专家系统(expert system, ES)[56], 基于专家知识做出决策. 1968 年, Quillian 提出了语义网络, 由相互连接的节点(概念或者对象)和边(节点的关系)组成. 此后, 知识库(knowledge base, KB)和知识表示(knowledge representation, KR)成为了研究热点;
- 1977 年, Feigenbaum 提出了知识工程(knowledge engineering, KE)[57], 采用人工智能, 以知识作为处理

对象, 使用计算机解决问题. 1980 年, McCarthy 提出了本体论(ontology), 通过构建本体描述世界. 1989 年, Berners-Lee 发明了万维网(World Wide Web, WWW)[58], 并在 1998 年提出了语义网(semantic Web, SW)[59], 使得网络数据变得机器可读. 万维网结合资源描述框架(resource description framework, RDF)后获得知识表示与推理能力. 2006 年, Berners-Lee 提出了链接数据(linked data)[60], 以构建数据公开和数据链接的语义网生态;

- 2012 年, Google 公司提出了知识图谱的概念[61], 提升了搜索引擎的返回质量和查询效率, 提供围绕主题的结构化信息. 此后, 知识图谱与 ML、NLP、DL 等技术不断结合, 2015 年左右, 知识表示学习(knowledge representation learning, KRL)的概念被提出, 将实体和关系在低维连续向量空间中加以表征. 2020 年, 多模态知识图谱(multi-modal knowledge graph, MMKG)的概念被提出, 关注视觉数据并关联传统 KG 中的符号知识(包括实体、概念、关系等)与图像, 成为知识图谱系统未来发展重要方向.

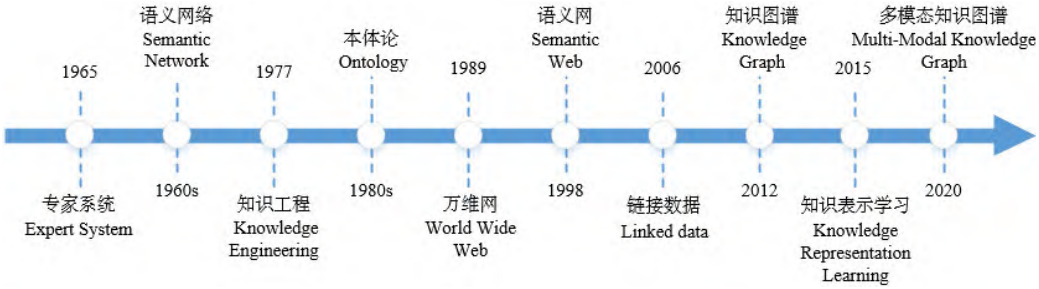

图 5 知识图谱发展时间轴

2.2 知识图谱体系架构

知识图谱按逻辑划分, 可分为数据层和模式层: 数据层以事实(fact)三元组为单位, 存储具体的数据信息; 模式层面向概念和关系, 存储知识数据, 包括实体(entity)、关系(relation)、属性(attribute)等知识定义.

知识图谱首先对原始数据(非结构化、半结构化和结构化)进行获取与处理, 提取信息要素. 然后, 通过知识抽取、知识融合、知识加工等技术方法, 从原始数据库和第三方数据库中提取知识事实, 构建知识图谱. 最后进行知识推理和应用, 知识推理是知识图谱能力输出的主要方式, 知识应用将知识图谱与其他特定业务或领域相结合, 利用知识图谱的技术特性提高业务效率. 知识图谱的体系架构如图 6 所示.

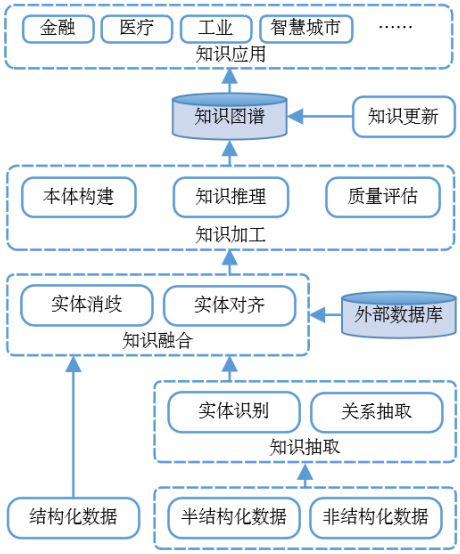

图 6 知识图谱的体系架构

2.3 知识图谱方法与技术

知识图谱运用多种方法与技术对原始数据进行挖掘处理, 主要可分为知识抽取、知识融合和知识推理.

2.3.1 知识抽取

知识抽取(knowledge extraction, KE)作为构建知识图谱的第 1 步, 旨在从异构数据中抽取重要信息要素. 依据发展顺序, 其主要技术分为基于规则字典的知识抽取、基于统计机器学习的知识抽取、基于深度学习的知识抽取, 具体过程分为命名实体识别和关系抽取两个阶段.

(1) 命名实体识别

命名实体识别(name entity recognition, NER)也称为实体抽取, 在文本中识别出特定含义或强指代性的实体, 主要方法可分为: 基于规则和字典的 NER 方法, 领域专家根据领域特点构造出规则模板, 凭借模式匹配和字符串匹配的规则设计词典模板[62]; 基于统计机器学习的 NER 方法, 将识别问题作为序列标注问题, 在标签序列中预测强相互依赖关系; 基于深度学习的 NER 方法, 避免大量的人工特征构建, 通过梯度传播训练优化网络结构模型. NER 的典型方法与技术见表 2.

表 2 命名实体识别的典型方法与技术

NER 分类	核心思路	使用场景	方法优、缺点	相关技术
基于规则字典	领域专家 构建规则集	小规模知识 图谱处理	高精度、低召回、低移植性, 无法覆盖所有语言特征规则	NERA[63], LaSIE-II[64], FASTUS[65]等
基于统计学习	对序列标注 问题预测 强相互依赖关系	新领域知识 图谱, 通用性 知识图谱构建	易于改动, 通用性强, 但依赖人工构造特征, 训练 时间长, 特征选取依赖高	HMM[66], ME[67], TMN[68], SVM[69]等
基于深度学习	梯度传播 训练优化 网络结构模型	目前主流方法, 复杂知识 图谱构建	数据驱动, 学习能力强, 覆盖广, 可移植性高, 但依赖硬件支撑	LSTM-CNN[70], LSTM-CRF[71] W2NER[72], RNN-T[73] CNN-LSTM-CRF[74], PO-TreeCRFs[75] CNN-BI-LSTM-CRF[76]等

此外, 部分研究者也正将注意力模型、迁移学习、半监督学习等方法引入 NER, 以提高实体识别效率, 减少人工标注成本.

(2) 关系抽取

关系抽取(relation extraction, RE), 在 NER 之后, 识别文本语料中离散化实体间的语义关系, 建立实体间的语义链接, 主要方法可分为: 基于规则和字典的 RE 方法, 基于文本词语、词性或语义的模式集合, 以人工构造形式构成语法和语义规则, 在后续对特定领域字典进行扩充; 基于统计机器学习的 RE 方法, 以数据是否标注作为分类标准, 分为有监督、半监督、无监督这 3 种关系抽取方法, 提升召回率, 增强跨领域通用性; 基于深度学习的 RE 方法, 通过神经网络训练数据构建模型. RE 的典型方法与技术见表 3.

表 3 关系抽取的典型方法与技术

RE 分类	核心思路	方法细分	方法优、缺点	相关技术
基于 规则字典	通过手写 规则匹配 文本数据	基于触发词(基于模式); 基于依存关系(语法树)	高准确度, 特定领域, 小数据集灵活, 召回率低, 鲁棒性差	RelExt[77], Pore[78]等
基于 统计学习	处理数据 标注问题	监督学习(特征向量/核函数); 半监督学习; 无监督学习(聚类方法)	数据依赖低, 领域 无关性, 但存在提取 误差传播问题	BI-LSTM[79], Att-RCNN[80], Self-Att-CNN[81]等
基于 深度学习	使用多种 神经网络 训练模型	流水线; 联合学习; 远程监督方法	数据驱动, 高预测精度, 可移植性高	PCNN[82], SDP-LSTM[83], APCNNs[84], SGCN[85], DSGAN[86], JRE_TRL[87]等

2.3.2 知识融合

知识融合(knowledge fusion, KF), 旨在消歧、加工、整合知识抽取阶段获得的扁平化形式知识, 确定知识图谱中等价的实例、类别和属性. 去除冲突和重叠的知识数据, 更新知识图谱, 主要方法包括实体消歧和实体对齐.

(1) 实体消歧

实体消歧(entity disambiguation), 解决一词多义问题, 确保知识图谱中的同名实体指称项具有明确定义和区分. 实体消歧可分为聚类消歧和链接消歧: 前者将所有实体指称项按其指向的目标实体进行聚类, 即每一个实体指称项对应到一个单独的类别; 后者将实体指称项与目标实体列表中的对应实体进行链接实现消歧. 例如: LoG[88]提出了关键字提取方法 Sent2Word, 从局部提取全局特征, 检测每个文档的关键字; DSRM[89]对实体语义相关性建立模型; EDKate[90]用于实体和文本的联合嵌入等.

(2) 实体对齐

实体对齐(entity alignment), 解决同义异名问题, 判断多个实体是否指向真实世界中同一客观对象, 利用实体的属性信息判定不同实体是否可进行对齐. 基于机器学习的实体对齐方法主要采用监督和无监督学习方式, 依据知识的属性相似度匹配方式进行实体对齐, 例如决策树(decision tree, DT)、支持向量机(support vector machine, SVM)等. 依赖实体的属性信息, 通过属性相似度进行跨平台实体对齐关系的推断. 基于知识表示学习的方法通过将知识图谱中的实体和关系都映射低维空间向量, 使用数学方法对各实体间的相似度进行计算, 例如 Trans 模型[91−96]方法等.

KF 的典型方法与技术见表 4.

表 4 知识融合的典型方法与技术

KF 方法	核心思路	方法目标	方法细分	相关技术
实体消歧	利用特征值或相似度实现实体消歧	解决一词多义问题	聚类消歧 链接消歧	Sent2Word, DSRM, EDKate 等
实体对齐	通过嵌入表示实现实体对齐	解决同义异名问题	有监督实体对齐 无监督实体对齐	DT, SVM, Trans 模型等

2.3.3 知识推理

知识推理(knowledge reasoning, KR), 根据知识库中现有的实体关系数据推测和构建实体之间新关系, 进而丰富和扩大知识库网络. 主要方法可分为: 基于逻辑规则的推理方法, 利用知识的符号性和简单规则及特征推理得到新知识; 基于嵌入表示的推理方法, 将图结构中的隐含关联信息映射向量化表示, 发现内在关联关系; 基于神经网络的推理方法, 利用各种神经网络建模非线性复杂关系, 挖掘隐含语义和结构特征. KR 的典型方法与技术见表 5.

表 5 知识推理的典型方法与技术

KR 分类	核心思路	方法细分	方法优、缺点	相关技术
基于逻辑规则	使用一阶谓词逻辑/机器学习方法/图结构特征, 挖掘隐含关系	逻辑方法, 统计方法, 图结构方法	准确度高, 可解释性强, 但计算复杂度高, 难以处理关系稀疏的数据	AMIE[97], SFE[98], HIRI[99], PRA[100], CPRA[101]等
基于嵌入表示	向量化表示复杂数据结构	张量分解方法, 距离模型方法, 语义匹配方法	数据处理量大, 拟合能力强, 但可解释性弱, 模型训练难度大	RESCAL[102], TransE[103] ManifoldE[104], DistMul[105], Complex[106], ANALOGY[107]等
基于神经网络	使用多种神经网络训练模型, 学习语义特征和结构特征	卷积神经网络, 循环神经网络, 图神经网络等	强推理能力和泛化能力, 但解释性不足, 缺乏全局语义的考虑	ConvE[108], InteractE[109], KGDL[110], SACN[111], DeepPath[112]等

2.4 知识图谱应用

结构化的知识图谱具有高效的数据处理和知识推理能力, 伴随信息化与数字化建设的发展和自然语言处理技术的进步, 基于知识图谱的信息系统从通用知识图谱衍生出语言、常识、领域等多种知识图谱应用形式. 国内外互联网公司如 Microsoft、Google、Facebook、Amazon 以及腾讯、阿里巴巴、美团、百度等积极布局知识图谱系统, 构建了包括 Satori、Probase、Google Knowledge Graph、腾讯云知识图谱、阿里云、美团大脑和百度智能云等综合知识图谱信息系统[113−115], 典型应用如: TKG[116], 由腾讯开发的集成了图数据库、图计算引擎及图可视化分析的一体化平台; AliMe MKG[117], 由阿里巴巴开发构建的以内容为中心的多模态商品知识

图谱, 为消费者提供商品认知画像以辅助消费决策. 目前, 基于知识图谱的信息系统在互联网领域得到广泛应用, 其中的技术应用类型可归纳为自然语言理解、知识问答、智能推荐等.

(1) 自然语言理解

自然语言理解(natural language understanding, NLU), 通过对语法、语义、语用的分析, 获取自然语言的语义表示. 基本自然语言处理(natural language processing, NLP)任务的语言建模, 预测序列中给定单词的下一个单词. 传统的语言建模不能利用文本语料库中经常观察到的实体的事实知识, 在引入知识图谱后, 基于知识感知的 NLU 通过注入统一语义空间的结构化知识, 增强语言表示. 知识图谱将知识融入语言表征, 利用明确的事实知识和内隐语言表示实现知识赋能, 在 NLU 任务中取得了不错的效果. 例如: Chen 等人[118]提出了基于两个知识图谱的双图随机游走算法, 即基于时隙的语义知识图和基于单词的词汇知识图, 以考虑口语理解中的顺序关系; Wang 等人[119]通过加权词概念嵌入, 增强基于知识的概念化的短文本表示学习; Peng 等人[120]整合了外部知识库, 建立面向事件分类的异构信息图.

(2) 知识问答

基于知识图谱的问答(knowledge graph-based question answering, KGQA), 使用知识图谱的事实知识回答自然语言问题. 传统知识问答系统以大规模语法规则为基础, 泛化性差. KGQA 利用知识图谱中的事实回答自然语言问题, 具有包含实体详细关系的知识库, 面向隐式推理时, 可有效增强问答质量、扩大问答范围、提高问答效率. KGQA 可分为单一事实问答(single-fact QA)和多跳推理问答(multi-hop reasoning QA). 其中,

- 单一事实问答以知识图谱作为外部知识来源, 回答小规模知识图谱事实问题. 例如: BAMnet 等人[121]提出的双向注意机制, 模拟问题和知识图谱之间的双向交互; Mohammed 等人[122]利用复杂深度模型, 包括门控循环单元(gated recurrent unit, GRU)、LSTM 等进行问答操作, 效果良好;
- 多跳推理问答将基于深度学习的知识图谱与神经编码器-解码器模型相结合, 具备多跳常识推理能力, 结构化的知识促进了问答系统内多跳推理的符号空间和语义空间之间常识性知识融合的研究. 例如: KagNet 等人[123]通过 GCN、LSTM 和基于层次路径的关系表示构建模式图; Zhang 等人[124]构造图网络结构并提取粗细粒度特征, 结合句子信息和实体信息回答多跳推理问题; Fu 等人[125]提出的 RERC 模型, 基于复杂问题分解的三阶段框架, 以获得完整的推理证据路径.

(3) 智能推荐

推荐系统(recommender system, RS), 旨在联系没有物品和没有明确目标的用户, 解决信息过载的问题, 帮助用户找到其感兴趣的内容. 传统推荐系统包括基于协同过滤的推荐系统、基于内容的推荐系统以及混合推荐系统. 其主要考虑用户的偏好序列, 忽视用户的偏好细节, 通常无法解决稀疏性问题和冷启动问题. 引入知识图谱的智能推荐系统, 提供知识信息, 例如实体、关系和属性等. 例如: KPRN[126]将用户和项目之间的交互视为知识图谱中的实体关系路径, 并对 LSTM 路径进行偏好推理以捕获顺序依赖关系; PGPR[127]对基于知识图的用户项交互执行强化策略的路径推理; KGAT[128]等将图形注意网络应用于实体关系和用户项目图的协作知识图, 通过嵌入传播和基于注意力的聚合来编码高阶连通性; SEGAR[129]利用图卷积网络和知识图注意网络增强会话推荐.

3 数据与知识双驱动的新一代信息系统

本节从数字线索与知识图谱融合的信息系统出发展开介绍, 然后归纳了 KG4DT 和 DT4KG 的发展思路与新兴技术, 最后总结了知识赋能的新一代信息系统所面临的机遇与挑战.

3.1 数字线索与知识图谱融合的信息系统

新一代信息系统要求以数据与知识双驱动, 数字线索和知识图谱的融合, 可以有效实现知识赋能. 近年来, 伴随数字线索与知识图谱技术的发展和各行业领域要求的提升, 基于两者融合的信息系统越来越受到重视, 在保留两者原有系统特性的基础上, 通过多种技术融合形式, 利用知识驱动的数字孪生来管理复杂的互联信息系统. 数字线索与知识图谱融合的信息系统前期主要应用于航空航天领域, 例如: NASA 针对信息资源

管理, 将数字线索与基于知识图谱的知识架构(knowledge architecture)相融合, 结合数字孪生体以实现知识管理、信息架构和数据科学的领域融合[130,131]. 信息化和数字化进程, 使得数据与知识双驱动的信息系统被推广至智能制造、智慧城市等多个领域. SIEMENS、DASSAULT、PTC 等企业在原有数字线索信息系统的基础上, 融合知识图谱技术, 扩展关联数据范围, 映射多源生产模型, 提供复杂网络的可视化[132−134]. 例如: GE 在数字化风场中, 基于 PTC 的数字孪生解决方案, 对风场资源、风机设备等构建关联知识图谱, 实现风机设施可视化、远程实时监控和健康诊断等操作[135]; Turku City Data 通过智慧城市知识图谱支撑数字孪生, 从而解决节能减排、流量分析等城市管理的关键事项[136].

数字线索与知识图谱融合的信息系统双驱动应用形式如图 7 所示, 通过数字线索与知识图谱双向交叉, 一方面, 知识图谱具有高度语义抽象的图结构、推理能力以及可解释性等特点, 可以赋能数字线索(knowledge graph for digital thread, KG4DT), 实现结构化设计、可靠性增强、知识化增值等; 另一方面, 数字线索基于完整的生命周期信息进行全数据采集、全元素建模、全决策仿真, 提供物理空间与数字空间的精准映射与实时交互, 可以驱动知识图谱(digital thread for knowledge graph, DT4KG), 提供数据链治理、形式化建模和数字孪生化等服务.

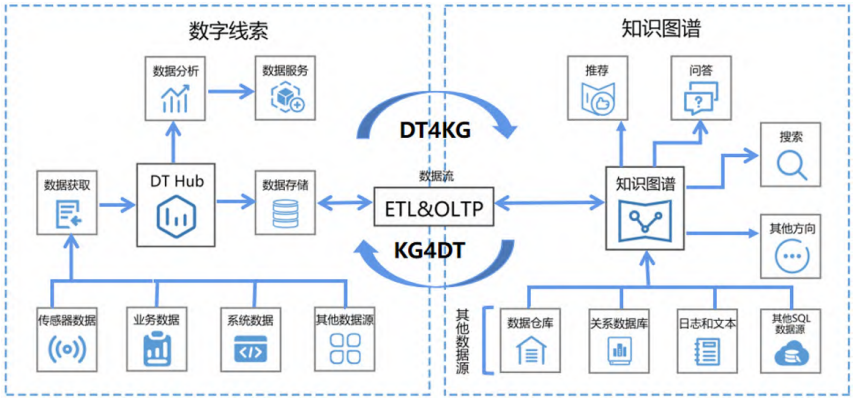

图 7 新一代信息系统双驱动引擎

3.2 知识图谱赋能数字线索(KG4DT)

现有的数字线索设计主要基于经验方法与领域规则, 需要人工调整与维护. 引入知识图谱可以优化基于数字线索的信息系统设计, 提供更合理的解决方案.

3.2.1 结构化设计

面对跨领域跨生命周期的数据存储, 单一的数字线索系统依赖传统数据存储方式, 在存储和查询时具有一定的优化空间, 结合知识图谱可以实现数字线索结构化设计, 包括数据整合、模型识别和图谱构建等.

(1) 数据整合

数字线索的数据来源多领域、多尺度、多周期, 数据的结构差异影响实际的数据处理与交互. 知识图谱具有强大的数据同构能力, 可提供异构数据同一化处理能力, 通过 GOPPRRE、OWL 等方法, 知识图谱可以对数据进行规则化预处理, 后期构建面向全生命周期的结构化知识图谱, 实现对各模块数据的一致性和完整性处理. 例如: Zuheros 等人[137]提出了不借助知识库等外部资源, 利用 LSTM 对目标实体进行上下文信息编码; Ganea 等人[138]提出了(local context attention mechanism, LCATT)深度学习框架, 结合实体嵌入和局部注意力机制, 可有效处理文档级实体消歧问题.

(2) 模型识别

跨领域的数字线索系统在数据标识、模型查询等阶段仍基于领域经验依赖人工构建, 存在同义异名的模型定义问题, 并且, 传统数据查询方式难以满足大规模的动态规划探索要求. 基于深度学习(例如 CNN[139]、RNN[140]、混合模型[141]等)的知识图谱技术可在领域知识稀缺的情况下, 从 DT 模型中自动学习复杂的隐藏特

征, 识别同义模型, 对数字线索的模型识别与确定具有参考价值, 例如多模态大规模概念本体协作编辑任务(large-scale concept ontology for multimedia, LSCOM)[142]、多模态本体(core ontology for multimedia, COMM)[143]等方法.

(3) 图谱构建

数字线索会产生海量数据, 在数据处理、模型输出等阶段, 存在数据维持、冗余文档和理解歧义等隐含问题. 知识图谱具有高可解释性, 可通过 RDF[144]描述 DT 知识资源, 并挖掘、分析、构建和显示内在知识关系, 提供兼顾关系表物理结构与图模型逻辑结构的存储方案, 如 DB2RDF[145]、SQLGraph[146]等. 知识图谱可以实现对数字线索的图谱化存储, 例如 IBM DB2[147]数据库的“关系-图谱”混合存储方案, 基于关系模式灵活配置图谱视图, 实现关系和图谱的统一分析查询.

3.2.2 可靠性增强

可靠性设计是数字线索的重要研究方向, 以产品、故障、环境等模型为核心, 综合集成通用与质量特性是研究的难点, 知识图谱的引入, 将有效增强数字线索的可靠性设计.

(1) 敏感数据发现

数字线索以模型数据为核心驱动, 并且数字线索的海量数据可能存在多种设备端口. 敏感数据泄露会导致严重的信息丢失和安全问题. 基于注意力机制(ATT)的知识图谱可根据不同 Attention-CNN 机制对数字线索系统内重要安全要素加以挖掘, 确定其与各领域的安全关联. 例如: APCNNs[148]通过 PCNN 和句子级注意力机制引入实体描述; HATT 等线索系统[149]通过连接每个层次的注意表示, 捕获关系层次结构等.

(2) 失效影响分析

数字线索在进行潜在失效模式及后果分析(failure mode and effects analysis, FMEA)[150]操作时, 存在故障关系割裂问题, 传统的故障分析模式难以直接确定失效原因机理. 可以借助 NLP 训练模型, 通过构建失效分析知识图谱, 实现多源异构知识数据管理, 形成专家经验, 挖掘失效产品、失效模式、失效原因、检测方法、改善措施等知识. 并且, 在后续过程中, 实现智能故障诊断和失效归因分析, 以降低故障带来的损失.

(3) 攻击事件调查

数字线索面临多种网络攻击威胁, 传统的单步攻击告警无法满足网络空间预警要求, 安全推理知识图谱可有效确定攻击源、攻击介质(中间点)、攻击路径, 是数字线索安全体系从被动防御到主动防御的重要技术, 例如 CVE[151]、CCE[152]、CVSS[153]、CAPEC[154]等安全相关的异构知识库. 在相关知识补全后, 安全知识图谱可以关联日志语义, 通过图分析方法实现攻击路径分析. 例如, Zeng 等人[155]提出的通过对攻击行为语义进行提取, 可推理安全知识图谱节点语义, 枚举所有行为子图.

3.2.3 知识化增值

数字线索系统的模型数据量级不断增大, 知识图谱通过对海量数据的知识挖掘, 实现数字线索系统数据的知识化增值将成为发展趋势.

(1) 模型知识表示

数字线索面对各领域多维异构数据, 使用 SysML、UML、DSL 等语言统一建模, 并通过多种开发工具及数据库软件(如 PLM、CAD/CAE、ALM 等)实现整合数据. 以此为基础引入知识图谱, 对已有模型数据使用基于深度学习的知识表示方法, 可以有效实现维度压缩, 同时对实体增加语义描述并增加模型的知识表示能力, 增强模型及关系的可解释性, 如 TransE[156]、TransH[157]、TransR[158]、TransA[159]等基于向量的知识表示方式.

(2) 模型知识存储

数字线索目前仍依赖于传统数据存储形式, 存在相似模型高度重复等问题, 模型复用与存储成为难题. 知识图谱可以对现有模型数据实现知识融合, 有效解决知识重复和关联不明等问题, 在对模型数据整合、消歧、加工、推理验证、更新等步骤后, 系统性地建立高质量的数字线索知识库[160].

(3) 模型知识推理

数字线索具有多领域、多尺度、多生命周期的复杂数据来源, 数据规模通常非常庞大. 该数据背景下, 知

识图谱具有强大的知识挖掘与推理能力, 可以实现数字线索内智能搜索、智能问答、推荐系统、对话系统等多种知识推理应用[161], 挖掘隐含知识, 丰富并拓展内在知识库, 充分利用数字线索的大规模数据量, 可有效提高知识推理正确率[162], 实现大规模知识图谱的推理活动[163].

3.3 数字线索驱动知识图谱(DT4KG)

知识图谱系统伴随人工智能发展已在多领域展开深度研究, 数字线索的有机结合为实现进一步的知识赋能提供了充分的数据准备, 包括基于数字线索全生命周期数据的数据链治理、形式化建模和数字孪生化.

3.3.1 数据链治理

大规模的通用或特定行业领域知识图谱, 通常依赖于高质量数据, 以实现知识的准确抽取和快速聚合. 数字线索在知识图谱数据治理领域存在结合的可能, 可用于发现、清理、集成和标记数据以提高数据质量.

(1) 全周期数据获取

面向全周期数据的数字线索可以提供完整的数据描述, 根据应用程序和用户需求自动地从数据仓库(data warehouse, DWH)中查找相关数据集, 为知识图谱提供完整的数据来源与访问接口, 并且数字线索中的 SQL 可以进一步扩展至支持 AI 模型[164], 以设计迎合用户需求的服务. 此外, 基于机器学习的数据发现可以增强查找相关数据的能力[165−167], 从而有效地查找大量数据源中的相关数据. 例如, Fernandez 等人提出的 Aurum[165], 这是一种基于超图形式的数据发现系统, 根据用户的需求为搜索数据集提供灵活的查询, 利用企业知识图(enterprise knowledge graph, EKG)捕获各种关系以支持各种查询等.

(2) 全系统数据清理

知识图谱在处理缺失数据、重复数据、错误数据和不可用数据等脏数据时, 没有较好的处理方式, 会影响到后续的模型训练. 数字线索可以对全系统的各周期阶段数据集提前进行检测和修复等预处理, 引入数据清理和集成技术[168], 检测和修复脏数据并集成多源数据, 为知识图谱系统提供高质量数据. 例如, Wang 等人[169]提出的用于机器学习任务的清洁活动框架, 清洁给定的具有凸损失函数的数据集和机器学习模型, 选择能够最大程度提高模型性能的记录并迭代地处理这些记录.

(3) 全领域数据标注

知识图谱模型训练中的训练效果严重依赖于数据集质量, 尽管近年来提出了无监督、小样本的深度学习, 但仍有部分知识图谱研究依赖基于统计意义的大数据模型, 数据标注效果与性能存在密切的关联性. 数字线索系统可以提供优质的标注服务, 利用数字线索涵盖的各领域专家、行业知识库和企业众包, 为机器学习标记大量训练数据, 通过数千名人员标记数据的数据标注众包模式[170]为知识图谱系统实现数据标注, 例如 Mechanical Turk、Figure-eight、CrowdFlower、Mighty AI 等平台[171], 或国内的百度众测、阿里众包、京东微工等数据标注平台.

3.3.2 形式化建模

面对大规模行业知识图谱的构建、推理及应用, 目前尚无较好的方式处理灵活、动态的企业数据和业务变化. 数字线索有效管控行业的全周期数据, 可在数据预处理和模型存储阶段, 通过形式化建模方式规范行业知识图谱等.

(1) 统一的数据预处理

知识图谱的原始数据形式分为结构化、半结构化和非结构化, 对于外部知识库构建和多维异构数据处理方法尚未统一. 数字线索具有循环反馈的全生命周期数据追踪[172]能力, 可以为知识图谱系统提供丰富的数据来源和输入数据的结构化预处理, 降低后期数据处理难度. 首先, 数字线索可利用各领域专家知识与经验, 以补充领域知识库的不足; 其次, 数字线索具有先天建模优势, 可满足跨领域跨周期数据的结构化要求, 自动优化并建模数据, 提供架构模型、逻辑模型、物理模型、功能模型等[173−175], 实现知识图谱系统的结构化数据输入; 最后, 对于非结构化的重要信息要素, 可自动生成关联文本, 降低后续关联抽取难度.

(2) 通用的模型存储与训练

目前, 知识图谱的深度学习训练以数据库形式支持模型存储、模型更新和并行训练, 需要提供模型存储

方案和多用户训练使用. 数字线索可以通过利用 UML、SysML 等统一建模语言, IBM-Rationa Harmony-SE、INCOSE 等系统工程方法和 OOSEM、Vitech-MBSE 等建模工具, 在动态更新数据时实现复杂工程建模, 提供各类系统模型的存储与访问接口, 为解决知识图谱系统的模型存储问题提供思路. 此外, 知识图谱的深度学习模型训练研究多集中在算法有效性上, 数字线索可以通过领域专家标注、重要事件加权等操作, 生成合适的训练数据集, 并通过模拟工具对其进行自动或人工标注, 不需要大量使用真实数据进行扩展和交叉验证, 可以有效提高知识图谱的训练效率.

3.3.3 数字孪生化——以虚促实

知识图谱面对大规模实际工程体系, 缺乏强可读性的跨领域跨生命周期显示模式, 而数字线索存在多种显示形式和结构模型, 可以以数字孪生形式为知识图谱提供安全感知、系统分析等可视化研究新角度, 并最终以虚拟形式促进实际系统实现.

(1) 数字孪生安全感知

知识图谱信息系统主要依赖传统的安全检测方式, 对隐含威胁判定、动态威胁评估、威胁事件挖掘等事件处理能力不足. 数字线索通过 MBSE 方法安全感知全生命周期数据, 以数字孪生的可视化形式融合直观特征有效分析数据和监视系统, 利用数据驱动的威胁事件建模等方式对知识图谱系统内的网络安全领域信息(heterogeneous cyber security information, HCSI)[172]进行加工、处理、整合, 并进一步转化为结构化的智慧安全领域知识库, 实现对复杂知识图谱系统的可视化安全感知和检测.

(2) 数字孪生系统分析

结合感知手段, 数字线索可实现知识图谱系统的数字孪生仿真, 直观地提高信息可解释性, 对各周期阶段数据状态、属性、关联形象化展示. 以此为基础, 在知识图谱应用的数字仿真系统中进行聚类方法、因子分析、共词分析等处理, 对知识图谱特征向量运用中心抽样方法, 对节点重要等级分级, 使得模型具有可解释性, 最后以可视化形式对相关信息系统进行深度理解.

3.4 知识赋能信息系统的机遇与挑战

人工智能 3.0 正迈向新的阶段, 数据与知识双驱动的新一代信息系统以数字线索和知识图谱为核心方法实现知识赋能, 成为“感知智能”转向“认知智能”的关键, 存在着诸多机遇与挑战.

知识赋能的信息系统如第 3.1 节、第 3.2 节所述, 对于数据与知识双驱动的结合, 不仅是系统升级与技术变革, 更重要的是将产生巨大的社会效能: 一方面, 知识赋能的信息系统将有力推动传统技术的革新, 数据与知识双驱动的信息系统通过复杂方案关联、事件回溯分析、知识事实可视化等知识化处理能力, 将为不同领域提供标准化的信息交互体系、基础设备安全保障、跨层次跨学科的知识资源共享平台, 进一步扩展系统控制范围与控制深度, 对产业结构调整产生深远影响; 另一方面, 知识赋能的信息系统将推动数字化转型升级, 数字化转型已成为国家推动创新的有效途径. 围绕《“十四五”国家信息规划》提出的数字化转型战略, 知识赋能的新一代信息系统有机关联物理世界与数字空间的全生命周期数据, 通过建立重要资料的数字孪生体, 以数字化、网络化、智能化形式, 将逐步完善新兴产业设计概念, 塑造未来生产样式.

同时, 知识赋能的信息系统围绕数据与知识双驱动的发展形式也面临着诸多挑战.

(1) 激活数据要素, 将数字线索作为数据支撑技术并充分激活数据要素潜能所面临的新挑战.

- 一方面, 中国数字线索研究仍处于发展阶段, 并且受限于仿真、建模及数据融合等技术瓶颈, 国内数字线索研究方案目前主要针对特定的应用场景或行业领域提供服务, 潜在价值尚未得到充分重视;
- 另一方面, 数字线索的多维数据处理存在难点. 例如: 在物理实体维度, 如何实时获取多维属性数据, 进而实现跨领域物理实体的互联互通与智能交互; 在虚拟实体维度, 如何构建动态的多尺度高精度模型, 保证虚拟实体的一致性/可靠性/统一性; 在服务数据维度, 如何合理调用数字线索模型与资源, 满足多领域/多层次/多形式的业务需求, 并实现数据资源增值增效; 在领域知识维度, 如何在特定行业获取大规模的领域知识, 处理差异数据和不精准数据并构建

深度知识结构; 在融合数据维度, 如何实现各维数据的深度融合与综合处理, 构建精准映射与智能交互; 在连接数据维度, 如何实现实时的跨协议/跨接口/跨平台的数据交互与迭代演化等. 此外, 在知识赋能的信息系统中, 数字线索部分的数据安全、系统算法优化等方面也存在一定的问题, 有待解决;

(2) 扩展知识内涵, 通过知识图谱实现认知推理并扩展知识内涵需要作进一步思考.

- 一方面, 在实际应用场景中, 复杂动态的知识图谱应用构建困难, 存在知识资产地位尚未确立、知识确权难题尚待破解、知识共享和交流困难、数据安全与隐私保护体系尚不健全等应用难点, 尚未确定高效且通用的解决方案;
- 另一方面, 对于知识图谱技术本身, 仍存在部分优化难点. 例如: 在知识融合方面, 如何实现动态知识评估以保证融合质量, 解决多领域多语言中的实体对齐和实体消歧问题; 在知识推理方面, 如何保护多元关系结构, 并完整利用其内在关联实现推理, 在领域小样本或零样本学习下实现推理; 在知识表达、存储和查询方面, 如何从半自动构建转向自动构建高质量知识图谱, 合理利用知识的动态约束信息并充分扩展知识的指导能力实现动态推理, 摆脱传统数据依赖困境等. 此外, 充分利用各类数据资源, 已有业务知识和人力要素进行验证标注, 有效实现知识图谱技术算法落地, 构建通用与行业知识图谱也是未来的研究方向之一;

(3) 领域深度结合, 推动数字线索与知识图谱技术交叉实现数据与知识双驱动, 并继续结合多元技术最终实现知识赋能的信息系统值得深思.

- 一方面, 缺乏交叉理论的深入认识, 在国家重大战略需求的驱动下, 多学科交叉与多技术融合成为常态. 数字线索与知识图谱的相关理论技术迅速发展的同时, 缺少对各自领域深入的双向了解, 无法及时更新和应用相关的理论研究, 难以有效形成交叉研究文化, 建立深度交叉合作. 同时, 数字与知识双驱动信息系统的实际应用场景模糊, 如何充分发挥源于军事航空领域的数字线索优势, 在各信息领域与知识图谱展开合作, 实现军转民技术的降维使用存在难点;
- 另一方面, 如何结合多元信息技术, 发展、运用和治理互联网体系, 构建万物互联的泛在网络已成为实现知识赋能必须面对和解决的重要问题. 例如: 结合物联网技术, 实现对物理世界的全面感知, 实现实时可靠的数据传输; 结合 3R (AR, VR, MR)技术, 实现虚拟模型的可视化呈现, 增强虚实实体的检测、验证及反馈能力; 结合区块链技术, 赋予信息系统数据的去中心化、不可篡改、追根溯源等特性. 此外, 如何进一步在金融、交通、医疗等领域展开实际的落地应用同样值得考虑.

4 结束语

数字化转型背景下, 数字线索提供面向全生命周期的海量数据处理架构, 成为破解数据孤岛问题和实现物理世界与数字空间映射交互的重要研究领域, 而知识图谱打破传统数据存储和使用方式, 以图结构抽取和运用知识, 是认知智能的关键赋能手段. 本文阐述、分析和总结了数字线索和知识图谱的研究与应用, 分别介绍了两者的基本概念和发展历程, 概括了数字线索的数据构成和数据处理方法以及知识图谱的体系架构和典型方法技术, 总结了各自应用方向与领域. 最后, 结合相关技术研究, 本文分析了二者有机结合下, 知识赋能的新一代信息系统可能的形式(KG4DT & DT4KG). 希望本综述能够为知识赋能的新一代信息系统研究与发展提供一定的理论参考和创新思路.

References:

[1] Tao J, Dai YC, Wei R, *et al*. Study on production lifecycle based on digital thread and digital twin. Aeronautical Manufacturing Technology, 2017, 21: 26–31 (in Chinese with English abstract). [doi: 10.16080/j.issn1671-833x.2017.21.026]

[2] Grieves M. Digital twin: Manufacturing excellence through virtual factory replication. White Paper, 2014, 1: 1–7.

[3] Liu W, Tao F, Cheng J, *et al*. Digital twin satellite: Concept, key technologies and applications. Computer Integrated Manufacturing Systems, 2019, 25(6): 1569−1575.

[4] Mandolla C, Petruzzelli AM, Percoco G, *et al*. Building a digital twin for additive manufacturing through the exploitation of blockchain: A case analysis of the aircraft industry. Computers in Industry, 2019, 109: 134−152.

[5] Pkr A, Nm B, Csr C, *et al*. Digital twin of an automotive brake pad for predictive maintenance—ScienceDirect. Procedia Computer Science, 2019, 165: 18−24.

[6] Zheng Y, Chen L, Lu X, *et al*. Digital twin for geometric feature online inspection system of car body-in-white. Int'l Journal of Computer Integrated Manufacturing, 2021, 34(7−8): 752−763.

[7] Coraddu A, Oneto L, Baldi F, *et al*. Data-driven ship digital twin for estimating the speed loss caused by the marine fouling. Ocean Engineering, 2019, 186(Aug.15): 106063.1−106063.14.

[8] Zhou M, Yan J, Feng D. Digital twin framework and its application to power grid online analysis. CSEE Journal of Power and Energy Systems, 2019, 5(3): 391−398.

[9] Peng Y, Zhao S, Wang H. A digital twin based estimation method for health indicators of DC-DC converters. IEEE Trans. on Power Electronics, 2020, PP(99): 1.

[10] The outline of the 14th five-year plan (2021−2025) for national economic and social development and the long-range objectives through the year 2035 of P. R. China. People's Daily, 2021-03-13(001) (in Chinese). [doi: 10.28655/n.cnki.nrmrb.2021.002455]

[11] Wang S, Guo X, Tie Y, *et al*. Weighted hybrid fusion with rank consistency. Pattern Recognition Letters, 2020, 138: 329−335.

[12] Guan D, Cao Y, Yang J, *et al*. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion, 2019, 50: 148−157.

[13] Manaa M, Akaichi J. Ontology-based modeling and querying of trajectory data. Data & Knowledge Engineering, 2017, 111: 58−72.

[14] Taylor CN, Bishop AN. Homogeneous functionals and Bayesian data fusion with unknown correlation. Information Fusion, 2019, 45: 179−189.

[15] Boschert S, Rosen R. Digital Twin—The Simulation Aspect. Mechatronic Futures: Springer, 2016. 59−74.

[16] Piascik R, Vickers J, Lowry D, *et al*. Technology area 12: Materials, structures, mechanical systems, and manufacturing road map. NASA Office of Chief Technologist, 2010. 15−88.

[17] Tao F, Cheng Y, Cheng JF, *et al*. Theories and technologies for cyber-physical fusion in digital twin shop floor. Computer Integrated Manufacturing Systems, 2017, 23(8): 1603−1611 (in Chinese with English abstract). [doi: 10.13196/j.cims.2017.08.001]

[18] Tao F, Zhang M, Cheng JF, *et al*. Digital twin workshop: A new paradigm for future workshop. Computer Integrated Manufacturing Systems, 2017, 23(1): 1−9 (in Chinese with English abstract). [doi: 10.13196/j.cims.2017.01.001]

[19] Hause M. Using MBSE to evaluate and protect the electrical grid as a system of systems. In: Proc. of the INCOSE Int'l Symp. 2017. 597−612.

[20] Wang FY. Parallel system methods for management and control of complex systems. Control and Decision, 2004, 19(5): 485−489, 514 (in Chinese with English abstract). [doi: 10.13195/j.cd.2004.05.6.wangfy.002]

[21] Yang LY, Chen SY, Wang X, *et al*. Digital twins and parallel systems: State of the art, comparisons and prospect. Acta Automatica Sinica, 2019, 45(11): 2001−2031 (in Chinese with English abstract). [doi: 10.16383/j.aas.2019.y000002]

[22] Kibira D, Shao GD, Weiss BA. Buiding a digital twin for robot workcell prognostics and health management. In: Proc. of the WSC. 2021. 1−12.

[23] Qin YD, Guo DK, Luo LL, Xu M. A joint orchestration of security and functionality services at network edge. Computer Networks, 2022, 212: 108951.

[24] Wen JQ, Gabrys B, Musial K. Toward digital twin oriented modeling of complex networked systems and their dynamics: A comprehensive survey. IEEE Access, 2022, 10: 66886−66923.

[25] Feng D. Data Deduplication for High Performance Storage System. Springer, 2022. 1−162.

[26] Dang TK, Ta MH, Hoang NL. Intermediate data format for the elastic data conversion framework. In: Proc. of the IMCOM. 2021. 1−5.

[27] Hosseini M, Kalwar S, Rossi MG, Sadeghi M. Automated mapping for semantic-based conversion of transportation data formats. In: Proc. of the SEM4TRA-AMAR@SEMANTiCS. 2019.

[28] Gao J, Li P, Chen Z, *et al*. A survey on deep learning for multimodal data fusion. Neural Computation, 2020, 32(5): 829−864.

[29] Wang X, Zheng S, Wang W, *et al*. Multi-rate data fusion for wireless sensor networks with time-delay based on improved cubature kalman filter. In: Proc. of the 2021 IEEE Asia-Pacific Conf. on Image Processing, Electronics and Computers (IPEC). IEEE, 2021. 716−720.

[30] Gantassi R, Masood Z, Choi YH. Enhancing QoS and residual energy by using of grid-size clustering, *K*-means, and TSP algorithms with MDC in LEACH protocol. IEEE Access, 2022, 10: 58199−58211.

[31] Tian M, Zhang L, Guo P, *et al*. Data dependence analysis for defects data of relay protection devices based on apriori algorithm. IEEE Access, 2020, 8: 120647−120653.

[32] Wu J, Orlandi F, O'Sullivan D, *et al*. An ontology model for climatic data analysis. In: Proc. of the 2021 IEEE Int'l Geoscience and Remote Sensing Symp. (IGARSS). IEEE, 2021. 5739−5742.

[33] Paschou T, Rapaccini M, Adrodegari F, *et al*. Digital servitization in manufacturing: A systematic literature review and research agenda. Industrial Marketing Management, 2020, 89: 278−292.

[34] Nee AYC, Ong SK. Virtual and augmented reality applications in manufacturing. IFAC (Proc. Volumes), 2013, 46(9): 15−26.

[35] Thanigaivel NK, Ong SK, Nee A. Augmented reality-assisted robot programming system for industrial applications. Robotics and Computer-integrated Manufacturing, 2019, 61.

[36] Rokhsaritalemi S, Sadeghi-Niaraki A, Choi SM. A review on mixed reality: Current trends, challenges and prospects. Applied Sciences, 2020, 10(2): 636.

[37] Chen Y. Integrated and intelligent manufacturing: Perspectives and enablers. Engineering, 2017, 3(5): 588−595.

[38] Scheeren I, Pereira CE. Combining model-based systems engineering, simulation and domain engineering in the development of industrial automation systems: Industrial case study. In: Proc. of the 17th IEEE Int'l Symp. on Object/Component/Service-oriented Real-time Distributed Computing. IEEE, 2014. 40−47.

[39] Singh S, Shehab E, Higgins N, *et al*. Towards information management framework for digital twin in aircraft manufacturing. Procedia CIRP, 2021, 96: 163−168.

[40] Zhu Q, Zhang LG, Ding YL, *et al*. Fromreal 3D modeling to digital twin modeling. Acta Geodaeticaet Cartographica Sinica, 2022, 51(6): 1040−1049. [doi: 10.11947/j.AGCS.2022.20210640]

[41] Lim KYH, Zheng P, Chen CH, *et al*. A digital twin-enhanced system for engineering product family design and optimization. Journal of Manufacturing Systems, 2020, 57: 82−93.

[42] Chen Y. Integrated and intelligent manufacturing: perspectives and enablers. Engineering, 2017, 3(5): 588−595.

[43] Fuller A, Fan Z, Day C, *et al*. Digital twin: Enabling technologies, challenges and open research. IEEE Access, 2020, 8: 108952−108971.

[44] Deng T, Zhang K, Shen ZJM. A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. Journal of Management Science and Engineering, 2021, 6(2): 125−134.

[45] Anthopoulos L. Smart utopia VS smart reality: Learning by experience from 10 smart city cases. Cities, 2017, 63: 128−148.

[46] Khan A, Aslam S, Aurangzeb K, *et al*. Multiscale modeling in smart cities: A survey on applications, current trends, and challenges. Sustainable Cities and Society, 2022, 78: 103517.

[47] Chang BJ, Chiou JM. Cloud computing-based analyses to predict vehicle driving shockwave for active safe driving in intelligent transportation system. IEEE Trans. on Intelligent Transportation Systems, 2019, 21(2): 852−866.

[48] Oxley T, Dore AJ, Apsimon H, *et al*. Modelling future impacts of air pollution using the multi-scale UK integrated assessment model (UKIAM). Environment Int'l, 2013, 61: 17−35.

[49] Rosen R, Von Wichert G, Lo G, *et al*. About the importance of autonomy and digital twins for the future of manufacturing. IFAC-PapersOnLine, 2015, 48(3): 567−572.

[50] Kritzinger W, Karner M, Traar G, *et al*. Digital twin in manufacturing: A categorical literature review and classification. IFAC-PapersOnLine, 2018, 51(11): 1016−1022.

[51] Viola J, Chen Y. Parallel self optimizing control framework for digital twin enabled smart control engineering. In: Proc. of the 1st IEEE Int'l Conf. on Digital Twins and Parallel Intelligence (DTPI). 2021. 358−361.

[52] Wang X, Zou L, Wang CK, *et al*. 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]

[53] Dong X, Gabrilovich E, Heitz G, *et al*. 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. 2014. 601−610.

[54] Wang X, Chen WX, Yang YJ, *et al*. Research on knowledge graph partitioning algorithms: A survey. Chinese Journal of Computers, 2021, 44(1): 235−260 (in Chinese with English abstract).

[55] Gong F, Wang M, Wang H, *et al*. SMR: Medical knowledge graph embedding for safe medicine recommendation. Big Data Research, 2021, 23: 100174.

[56] Liao SH. Expert system methodologies and applications—A decade review from 1995 to 2004. Expert Systems with Applications, 2005, 28(1): 93−103.

[57] Studer R, Benjamins VR, Fensel D. Knowledge engineering: Principles and methods. Data & Knowledge Engineering, 1998, 25(1): 161−197.

[58] Berners-Lee T, Groff JF. WWW. ACM SIGBIO Newsletter, 1992, 12(3): 37−40.

[59] Shadbolt N, Berners-Lee T, Hall W. The semantic Web revisited. IEEE Intelligent Systems, 2006, 21(3): 96−101.

[60] Bizer C, Heath T, Berners-Lee T. Linked data—The story so far. Int'l Journal on Semantic Web and Information Systems (IJSWIS), 2009, 5(3): 1−22.

[61] Singhal A. Introducing the knowledge graph: things, not strings. 2020. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

[62] Rau LF. Extracting company names from text. In: Proc. of the 7th IEEE Conf. on Artificial Intelligence Application. 1991. 29−32.

[63] Shaalan K, Raza H. NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 2009, 60(8): 1652−1663.

[64] Humphreys K, Gaizauskas R, Azzam S, *et al*. University of sheffield: Description of the LaSIE-II system as used for MUC-7. In: Proc. of the Conf. on Message Understanding. Association for Computational Linguistics, 1995.

[65] Appelt DE, Bear J, Hobbs JR, *et al*. SRI Int'l FASTUS system. In: Proc. of the 4th Conf. 1992.

[66] Bikel DM, Miller S, Schwartz R, *et al*. Nymble: A highperformance learning name-finder. In: Proc. of the 5th Conf. on Applied Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 1997. 194−201.

[67] Borthwick AE. A maximum entropy approach to named entity recognition [Ph.D. Thesis]. New York University, 1999.

[68] Lin BYC, Lee DH, Shen M, *et al*. TriggerNER: Learning with entity triggers as explanations for named entity recognition. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020. 8503−8511.

[69] Sun L, Han X. A feature-enriched tree kernel for relation extraction. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol.2: Short Papers). 2014. 61−67.

[70] Chiu JP, Nichols E. Named entity recognition with bidirectional LSTM-CNNs. Trans. of the Association for Computational Linguistics, 2016, 4: 357−370.

[71] Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition. arXiv:1603.01360, 2016.

[72] Li J, Fei H, Liu J, *et al*. Unified named entity recognition as word-word relation classification. arXiv e-prints, 2021.

[73] Soltau H, Shafran I, Wang M, *et al*. RNN transducers for nested named entity recognition with constraints on alignment for long sequences. arXiv.2203.03543, 2022.

[74] Wu FZ, Liu JX, Wu CH, *et al*. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation. In: Proc. of the World Wide Web Conf. New York: ACM, 2019. 3342−3348.

[75] Fu Y, Tan C, Chen M, *et al*. Nested named entity recognition with partially-observed TreeCRFs. 2020. [doi: 10.3115/1699510.1699529]

[76] Ronran C, Lee S. Effect of character and word features in bidirectional LSTM-CRF for NER. In: Proc. of the 2020 IEEE Int'l Conf. on Big Data and Smart Computing (BigComp). IEEE, 2020.

[77] Schutz A, Buitelaar P. Relext: A tool for relation extraction from text in ontology extension. In: Proc. of the Int'l Semantic Web Conf. Berlin, Heidelberg: Springer, 2005. 593−606.

[78] Wang G, Yu Y, Zhu H. PORE: Positive-only relation extraction from wikipedia text. In: Proc. of the Semantic Web. LNCS 4825, Shanghai: Apex Data & Knowledge Management Lab., Department of Computer Science and Engineering, Shanghai Jiaotong University, 2007.

[79] Shu Z, Zheng D, Hu X, *et al*. Bidirectional long short-term memory networks for relation classification. In: Proc. of the 29th Pacific Asia Conf. on Language, Information and Computation. 2015. 73−78.

[80] Guo X, Zhang H, Yang H, Xu L, Ye Z. A single attention-based combination of CNN and RNN for relation classification. IEEE Access, 2019, 7: 12467−12475. [doi: 10.1109/ACCESS.2019.2891770]

[81] Yan X, Duan YX, Zhang ZH. Entity relationship extraction fusing self-attention mechanism and CNN. Computer Engineering & Science, 2020, 42(11): 2059−2066.

[82] Kuang J, Cao YX, Zheng JB, *et al*. Improving neural relation extraction with implicit mutual relations. In: Proc. of the 36th IEEE Int'l Conf. on Data Engineering. Piscataway: IEEE, 2020. 1021−1032.

[83] Xu Y, Mou LL, Li G, *et al*. Classifying relations via long short term memory networks along shortest dependency paths. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015. 1785−1794.

[84] Ji GL, Liu K, He SZ, *et al*. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2017. 3060−3066.

[85] Sahu SK, Thomas D, Chiu B, *et al*. Relation extraction with self-determined graph convolutional network. In: Proc. of the 29th ACM Int'l Conf. on Information and Knowledge Management. New York: ACM, 2020. 2205−2208.

[86] Qin PD, Xu WR, Wang WY. DSGAN: Generative adversarial training for distant supervision relation extraction. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018. 496−505.

[87] Xiao Y, Tan CX, Fan ZJ, *et al*. Joint entity and relation extraction with a hybrid transformer and reinforcement learning based model. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2020. 9314− 9321.

[88] Xin KX, *et al*. LoG: A locally-global model for entity disambiguation. World Wide Web, 2020, 1−23.

[89] Huang H, Heck L, Ji H. Leveraging deep neural networks and knowledge graphs for entity disambiguation. arXiv:1504.07678, 2015.

[90] Fang W, Zhang J, Wang D, *et al*. Entity disambiguation by knowledge and text jointly embedding. In: Proc. of the 20th SIGNLL Conf. on Computational Natural Language Learning. 2016. 260−269.

[91] Bordes A, Weston J, Collobert R, *et al*. Learning structured embeddings of knowledge bases. In: Proc. of the 25th AAAI Conf. on Artificial Intelligence. 2011.

[92] Bordes A, Usunier N, Garcia-Duran A, *et al*. Translating embeddings for modeling multi-relational data. Advances in Neural Information Processing Systems, 2013, 26.

[93] Xiao H, Huang M, Zhu X. From one point to a manifold: Knowledge graph embedding for precise link prediction. arXiv:1512.04792, 2015.

[94] Feng J, Huang M, Wang M, *et al*. Knowledge graph embedding by flexible translation. In: Proc. of the 15th Int'l Conf. on Principles of Knowledge Representation and Reasoning. 2016. 557−560.

[95] Qian W, Fu C, Zhu Y, *et al*. Translating embeddings for knowledge graph completion with relation attention mechanism. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence {IJCAI-18}. 2018.

[96] Che C, Liu D. Crosslanguage entity alignment combined with attribute information through bidirectional alignment. Computer Engineering, 2021. https://doi.org/10.19678/j.issn.1000-3428.0060540

[97] Galárraga LA, Teflioudi C, Hose K, *et al*. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. In: Proc. of the 22nd Int'l Conf. on World Wide Web. New York: ACM, 2013. 413−422

[98] Gardner M, Mitchell T. Efficient and expressive knowledge base completion using subgraph feature extraction. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. 2015. 1488−1498.

[99] Liu Q, Jiang L, Han M, *et al*. Hierarchical random walk inference in knowledge graphs. In: Proc. of the 39th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2016. 445−454.

[100] Lao N, Mitchell T, Cohen W. Random walk inference and learning in a large scale knowledge base. In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing. 2011. 529−539.

[101] Wang Q, Liu J, Luo Y, *et al*. Knowledge base completion via coupled path ranking. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Vol.1: Long Papers). 2016. 1308−1318.

[102] 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. Madison: Omnipress, 2011. 809−816.

[103] Karetnikov A, Ehrlinger L, Geist V. Enhancing TransE to predict process behavior in temporal knowledge graphs. In: Proc. of the Int'l Conf. on Database and Expert Systems Applications. Cham: Springer, 2022. 369−374.

[104] Xiao H, Huang ML, Yu H, *et al*. From one point to a manifold: orbit models for knowledge graph embedding. arXiv preprint, 2017.

[105] Yang BS, Yih WT, He XD, *et al*. Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint, 2015.

[106] Trouillon T, Welbl J, Riedel S, *et al*. Complex embeddings for simple link prediction. In: Proc. of the 33rd Int'l Conf. on Machine Learning (ICML). New York, 2016. 2071.

[107] Liu HX, Wu YX, Yang YM. Analogical inference for multirelational embeddings. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney, 2017. 2168.

[108] Dettmers T, Minervini P, Stenetorp P, *et al*. Convolutional 2D knowledge graph embeddings. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2018. 1811−1818.

[109] Vashishth S, Sanyal S, Nitin V, *et al*. InteractE: Improving convolution-based knowledge graph embeddings by increasing feature interactions. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2020. 3009−3016.

[110] Zhao MJ, Zhao YW, Xu B. Knowledge graph completion via complete attention between knowledge graph and entity descriptions. In: Proc. of the 3rd Int'l Conf. on Computer Science and Application Engineering. New York: ACM, 2019. 47.

[111] Shang C, Tang Y, Huang J, *et al*. End-to-end structureaware convolutional networks for knowledge base completion. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2019. 3060−3067.

[112] Xiong WH, Hoang T, Wang WY. DeepPath: A reinforcement learning method for knowledge graph reasoning. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017. 564−573.

[113] Xue Y, Zhang H, Ma H. Performance evaluation of image and video cloud services. In: Proc. of the 20th IEEE Int'l Conf. on High Performance Computing and Communications; the 16th IEEE Int'l Conf. on Smart City; the 4th IEEE Int'l Conf. on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2018. 733−741.

[114] Wu W, Li H, Wang H, *et al*. Probase: A probabilistic taxonomy for text understanding. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. 2012. 481−492.

[115] Zou X. A survey on application of knowledge graph. Journal of Physics: Conf. Series. IOP Publishing, 2020, 1487(1): 012016.

[116] Melnik J. China's "National Champions" Alibaba, Tencent, and Huawei. Education About Asia, 2019, 24(2): 28−33.

[117] Xu G, Chen H, Li FL, *et al*. AliMe MKG: A multi-modal knowledge graph for live-streaming E-commerce. In: Proc. of the 30th ACM Int'l Conf. on Information & Knowledge Management. 2021. 4808−4812.

[118] Chen YN, Wang WY, Rudnicky A. Jointly modeling inter-slot relations by random walk on knowledge graphs for unsupervised spoken language understanding. In: Proc. of the 2015 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015. 619−629.

[119] Wang J, Wang Z, Zhang D, *et al*. Combining knowledge with deep convolutional neural networks for short text classification. In: Proc. of the IJCAI. 2017.

[120] Peng H, Li J, Gong Q, *et al*. Fine-grained event categorization with heterogeneous graph convolutional networks. arXiv:1906.04580, 2019.

[121] Chen Y, Wu L, Zaki MJ. Bidirectional attentive memory networks for question answering over knowledge bases. In: Proc. of the NAACL-HLT. 2019. 2913−2923.

[122] Mohammed S, Shi P, Lin J. Strong baselines for simple question answering over knowledge graphs with and without neural networks. arXiv:1712.01969, 2017.

[123] Lin BY, Chen X, Chen J, *et al*. Kagnet: Knowledge-aware graph networks for commonsense reasoning. arXiv:1909.02151, 2019.

[124] Zhang M, Li F, Wang Y, *et al*. Coarse and fine granularity graph reasoning for interpretable multi-hop question answering. IEEE Access, 2020, 8: 56755−56765.

[125] Fu R, Wang H, Zhang X, *et al*. Decomposing complex questions makes multi-hop QA easier and more interpretable. arXiv:2110.13472, 2021.

[126] Wang X, Wang D, Xu C, *et al*. Explainable reasoning over knowledge graphs for recommendation. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2019. 5329−5336.

[127] Xian Y, Fu Z, Muthukrishnan S, *et al*. Reinforcement knowledge graph reasoning for explainable recommendation. In: Proc. of the 42nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2019. 285−294.

[128] Wang X, He X, Cao Y, *et al*. Kgat: Knowledge graph attention network for recommendation. In: Proc. of the 25th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. 2019. 950−958.

[129] Xu X, Tang Y, Xu Z. SEGAR: Knowledge graph augmented session-based recommendation. In: Proc. of the Int'l Conf. on Knowledge Science, Engineering and Management. Cham: Springer, 2021. 229−241.

[130] Kraft EM. The air force digital thread/digital twin-life cycle integration and use of computational and experimental knowledge. In: Proc. of the 54th AIAA Aerospace Sciences Meeting. 2016.

[131] Singh M, Fuenmayor E, Hinchy EP, *et al*. Digital twin: Origin to future. Applied System Innovation, 2021, 4(2): 36.

[132] Banerjee A, Dalal R, Mittal S, *et al*. Generating digital twin models using knowledge graphs for industrial production lines. In: Proc. of the Web Science. 2017.

[133] Hubauer T, Lamparter S, Haase P, *et al*. Use cases of the industrial knowledge graph at siemens. In: Proc. of the ISWC (P&D/Industry/BlueSky). 2018.

[134] Lim KYH, Zheng P, Chen CH, *et al*. A digital twin-enhanced system for engineering product family design and optimization. Journal of Manufacturing Systems, 2020, 57: 82−93.

[135] Tao F, Zhang H, Liu A, *et al*. Digital twin in industry: State-of-the-art. IEEE Trans. on Industrial Informatics, 2018, 15(4): 2405−2415.

[136] Träskman T. Smartness and thinking infrastructure: An exploration of a city becoming smart. Journal of Public Budgeting, Accounting & Financial Management, 2022.

[137] Zuheros C, Tabik S, Valdivia A, *et al*. Deep recurrent neural network for geographical entities disambiguation on social media data. Knowledge-based Systems, 2019, 173: 117−127.

[138] Ganea OE, Hofmann T. Deep joint entity disambiguation with local neural attention. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017. 2619−2629.

[139] Nguyen DQ, Nguyen TD, Nguyen DQ, *et al*. A novel embedding model for knowledge base completion based on convolutional neural network. arXiv:1712.02121, 2017.

[140] Socher R, Huval B, Manning CD, *et al*. Semantic compositionality through recursive matrix-vector spaces. In: Proc. of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012. 1201−1211.

[141] Takanobu R, Zhang T, Liu J, *et al*. A hierarchical framework for relation extraction with reinforcement learning. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2019. 7072−7079.

[142] Naphade M, Smith JR, Tesic J, *et al*. Large-scale concept ontology for multimedia. IEEE Multimedia, 2006, 13(3): 86−91.

[143] Arndt R, Troncy R, Staab S, *et al*. COMM: Designing a well-founded multimedia ontology for the Web. In: Proc. of the Semantic Web. Springer, 2007. 30−43.

[144] Klyne G. RDF concepts and abstract syntax W3C recommendation. 2004. http://www.w3.org/TR/rdf-concepts/

[145] Bornea MA, Dolby J, Kementsietsidis A, *et al*. Building an efficient RDF store over a relational database. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. 2013. 121−132.

[146] Sun W, Fokoue A, Srinivas K, *et al*. Sqlgraph: An efficient relational-based property graph store. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. 2015. 1887−1901.

[147] Tian Y, Xu EL, Zhao W, *et al*. IBM DB2 graph: Supporting synergistic and retrofittable graph queries inside IBM DB2. In: Proc. of the 2020 ACM SIGMOD Int'l Conf. on Management of Data. 2020. 345−359.

[148] Ji G, Liu K, He S, *et al*. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. 2017. 3060−3066.

[149] Fuller A, Fan Z, Day C, *et al*. Digital twin: Enabling technologies, challenges and open research. IEEE Access, 2020, 8: 108952−108971.

[150] Mangalathu S, Hwang SH, Jeon JS. Failure mode and effects analysis of RC members based on machine-learning-based shapley additive explanations (SHAP) approach. Engineering Structures, 2020, 219: 110927.

[151] Mell P, Grance T. Use of the common vulnerabilities and exposures (CVE) vulnerability naming scheme. Special Publication (NIST SP)-800-51, 2002.

[152] Mitre. Common configuration enumeration. https://cce.mitre.org/about/index.html

[153] First. Common vulnerability scoring system. https://www.first.org/cvss/specification-document

[154] Mitre. Common attack pattern enumeration and classification. https://http://capec.mitre.org/

[155] Zeng J, Zheng LC, Chen Y, *et al*. WATSON: Abstracting behaviors from audit logs via aggregation of contextual semantics. In: Proc. of the Network and Distributed System Security Symp. 2021.

[156] Bordes A, Usunier N, Garcia-Duran A, *et al*. Translating embeddings for modeling multi-relational data. Advances in Neural Information Processing Systems, 2013, 26.

[157] Wang Z, Zhang J, Feng J, *et al*. Knowledge graph embedding by translating on hyperplanes. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2014, 28(1).

[158] Lin Y, Liu Z, Sun M, *et al*. Learning entity and relation embeddings for knowledge graph completion. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. 2015.

[159] Jia Y, Wang Y, Jin X, *et al*. Knowledge graph embedding: A locally and temporally adaptive translation-based approach. ACM Trans. on the Web (TWEB), 2017, 12(2): 1−33.

[160] Chen Z, Wang Y, Zhao B, *et al*. Knowledge graph completion: A review. IEEE Access, 2020, 8: 192435−192456.

[161] Wu YB, Yang F, Lai GH, *et al*. Research progress of knowledge graph learning and reasoning. Journal of Chinese Computer Systems, 2016, 37(9): 2007−2013.

[162] Guan SP, Jin XL, Jia YT, *et al*. Knowledge reasoning over knowledge graph: A survey. Ruan Jian Xue Bao/Journal of Software, 2018, 29(10): 2966−2994 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5551.htm [doi: 10.13328/j.cnki.jos.005551]

[163] Shah H, Villmow J, Ulges A, *et al*. An open-world extension to knowledge graph completion models. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2019, 33(1): 3044−3051.

[164] Fernandes S, Bernardino J. What is bigquery? In: Proc. of the 19th Int'l Database Engineering & Applications Symp. 2015. 202−203.

[165] Fernandez RC, Abedjan Z, Koko F, *et al*. Aurum: A data discovery system. In: Proc. of the 34th IEEE Int'l Conf. on Data Engineering (ICDE). 2018. 1001−1012.

[166] Zheng Y, Li G, Li Y, *et al*. Truth inference in crowdsourcing: Is the problem solved? Proc. of the VLDB Endowment, 2017, 10(5): 541−552.

[167] Ilyas IF, Chu X. Data Cleaning. Morgan & Claypool, 2019.

[168] Fan J, Li G. Human-in-the-loop rule learning for data integration. IEEE Database Engineering Bulletin, 2018, 41(2): 104−115.

[169] Krishnan S, Wang J, Wu E, *et al*. Activeclean: Interactive data cleaning for statistical modeling. Proc. of the VLDB Endowment, 2016, 9(12): 948−959.

[170] Willis CG, Law E, Williams AC, *et al*. CrowdCurio: An online crowdsourcing platform to facilitate climate change studies using herbarium specimens. New Phytologist, 2017, 215(1): 479−488.

[171] Lu J, Chen Y, Herodotou H, *et al*. Speedup your analytics: Automatic parameter tuning for databases and big data systems. Proc. of the VLDB Endowment, 2019.

[172] Dahlan AG, Muhammad Naim S, Luqman Zulhilmi AA. The research of 3D modeling between visual & creativity. Int'l Journal of Innovative Technology and Exploring Engineering (IJITEE), 2019, 8: 180−186.

[173] Bandara M, Rabhi FA. Semantic modeling for engineering data analytics solutions. Semantic Web, 2020, 11(3): 525−547.

[174] Schindler T, Skornia C. Secure parallel processing of big data using order-preserving encryption on google bigquery. arXiv:1608.07981, 2016.

[175] Gao Y, Li XY, Hao P, *et al*. HinCTI: A cyber threat intelligence modeling and identification system based on heterogeneous information network. IEEE Trans. on Knowledge and Data Engineering, 2020.

附中文参考文献:

[1] 陶剑, 戴永长, 魏冉, 等. 基于数字线索和数字孪生的生产生命周期研究. 航空制造技术, 2017, 21: 26–31. [doi: 10.16080/j.issn1671-833x.2017.21.026]

[10] 中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要. 人民日报, 2021-03-13(001). [doi: 10.28655/n.cnki.nrmrb.2021.002455]

[17] 陶飞, 程颖, 程江峰, 张萌, 徐文君, 戚庆林. 数字孪生车间信息物理融合理论与技术. 计算机集成制造系统, 2017, 23(8): 1603–1611. [doi: 10.13196/j.cims.2017.08.001]

[18] 陶飞, 张萌, 程江峰, 戚庆林. 数字孪生车间——一种未来车间运行新模式. 计算机集成制造系统, 2017, 23(1): 1–9. [doi: 10.13196/j.cims.2017.01.001]

[20] 王飞跃. 平行系统方法与复杂系统的管理和控制. 控制与决策, 2004, 19(5): 485–489+514. [doi: 10.13195/j.cd.2004.05.6.wangfy.002]

[21] 杨林瑶, 陈思远, 王晓, 张俊, 王成红. 数字孪生与平行系统:发展现状、对比及展望. 自动化学报, 2019, 45(11): 2001–2031. [doi: 10.16383/j.aas.2019.y000002]

[52] 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]

[54] 王鑫, 陈蔚雪, 杨雅君, 张小旺, 冯志勇. 知识图谱划分算法研究综述. 计算机学报, 2021, 44(1): 235–260.

[162] 官赛萍, 靳小龙, 贾岩涛, 王元卓, 程学旗. 面向知识图谱的知识推理研究进展. 软件学报, 2018, 29(10): 2966–2994. http://www.jos.org.cn/1000-9825/5551.htm [doi: 10.13328/j.cnki. jos.005551]

朱迪(1997－), 男, 硕士, 主要研究领域为机器学习, 数据挖掘, 数字线索.

张博闻(1999－), 男, 硕士, CCF 学生会员, 主要研究领域为机器学习, 时空预测, 数据挖掘.

程雅琪(1997－), 女, 硕士, 主要研究领域为机器学习, 数据挖掘.

刘昕悦(2000－), 女, 硕士, 主要研究领域为机器学习, 知识图谱, 推荐系统.

吴文隆(2001－), 男, 学士, CCF 学生会员, 主要研究领域为机器学习, 数据挖掘.

王铁鑫(1987－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为基于模型的系统工程, 模型驱动方法, 领域本体与语义检测.

文浩(1979－), 男, 博士, 教授, 主要研究领域为航天器动力学与控制, 空间机器人, 数字线索.

李博涵(1979－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为时空数据库, 知识图谱, 自然语言处理, 推荐系统.