DOI 10.15302/J-SSCAE-2023.02.018

大数据知识工程发展现状及展望

郑庆华[1]，刘欢[2,3]，龚铁梁[2]，张玲玲[2]，刘均[2,3*]

（1. 西安交通大学，西安 710049；2. 西安交通大学计算机科学与技术学院，西安 710049；
3. 陕西省大数据知识工程重点实验室，西安 710049）

摘要：大数据知识工程是人工智能的“基础设施”、诸多行业和领域面临的共性需求、信息化迈向智能化的必由之路。本文阐述了大数据知识工程产生的背景与概念内涵，提出了“数据知识化、知识体系化、知识可推理”的研究框架；梳理了知识获取与融合、知识表征、知识推理等大数据知识工程关键技术和智慧教育、税务风险管控、智慧医疗等典型场景中的工程应用；总结了大数据知识工程面临的挑战，研判了大数据知识工程的未来研究方向，包括复杂大数据知识获取、知识+数据混合学习、脑启发知识编码记忆等。研究建议，引导多学科交叉融合，设立重大和重点研发专项，推动大数据知识工程基础理论与技术攻关；加强企业和研究机构间交流合作，推广前沿研究成果并形成应用示范，建立大数据知识工程行业标准体系；以重大需求应用为导向，探索校企协同育人模式，加快大数据知识工程技术在重要行业的落地应用。

关键词：大数据知识工程；知识获取；知识融合；知识表征；知识推理

中图分类号：TP319　　**文献标识码：**A

Development and Prospect of Big Data Knowledge Engineering

Zheng Qinghua[1], Liu Huan[2,3], Gong Tieliang[2], Zhang Lingling[2], Liu Jun[2,3*]

(1. Xi’an Jiaotong University, Xi’an 710049, China; 2. School of Computer Science and Technology, Xi’an Jiaotong University, Xi’an 710049, China; 3. Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi’an 710049, China)

Abstract: Big Data Knowledge Engineering is the infrastructure of artificial intelligence, a common requirement faced by various industries and fields, and the inevitable path for the digitalization to intelligence. In this paper, we firstly elaborate on the background and connotation of big data knowledge engineering and propose a research framework of “data knowledgeization, knowledge systematization, and knowledge reasoning”. Secondly, we sort out the key technologies of knowledge acquisition and fusion, knowledge representation, and knowledge reasoning and introduce engineering applications in typical scenarios such as smart education, tax risk control, and smart healthcare. Thirdly, we summary the challenges faced by big data knowledge engineering and predict the future research directions including complex big data knowledge acquisition, knowledge+data hybrid learning, and brain-inspired knowledge coding and memorizing. Finally, several suggestions are given by the research: guiding interdisciplinary integration and establishing major and key R&D projects to promote the basic theory and technological breakthroughs of big data knowledge engineering; strengthening communication and cooperation between enterprises and research institutions as well as promoting cutting-edge research results to form application demonstrations, so as to establish an industry-standard system for big data knowledge engineering;

收稿日期：2023-01-16；**修回日期：**2023-03-02

通讯作者：*刘均，西安交通大学计算机科学与技术学院教授，研究方向为自然语言理解、计算机视觉、智慧教育；
E-mail: liukeen@xjtu.edu.cn

资助项目：国家自然科学基金项目(62250009)；中国工程科技知识中心项目(CKCEST-2022-1-40)

本刊网址：www.engineering.org.cn/ch/journal/sscae

exploring school-enterprise cooperation in line with market demands, orienting towards major application needs, and accelerating the landing application of big data knowledge engineering technology in the country's important industries.

Keywords: Big Data Knowledge Engineering; Knowledge Acquisition; Knowledge Fusion; Knowledge Representation; Knowledge Reasoning

一、前言

我国信息化建设经过40多年的发展，在教育、政务、金融、医疗等领域积累了海量数据，如何将这些数据进一步转化为相关知识、反哺领域发展、破解实际工程难题，逐渐成为各领域的共性需求。知识工程旨在研究人类知识的机器表征与计算问题，是人工智能领域重要的分支，其目的是将人类或专家的知识输入到计算机中并建立推理机制，让机器也能拥有知识并能进行计算和推理，据此解决实际问题。当前，我国知识工程的发展经历了以专家系统为代表的传统知识工程和以主流深度学习技术为代表的现代知识工程等阶段，显著推动了各领域的发展，但在解决各领域实际工程问题中仍存在一定的局限性。例如，传统知识工程在20世纪七八十年代得到迅猛发展后，在90年代进入“寒冬期”，主要原因在于“知识获取”主要来自领域专家，面临人工成本过高、专家经验局限、无法动态解决复杂工程问题等[1]；在现代知识工程阶段，深度学习模型（尤其是大规模预训练模型）在自然语言处理、计算机视觉等方面取得显著发展，但这类数据驱动的模型存在数据依赖性强、算力/能源消耗过度[2-4]等挑战，难以应对实际工程问题中的高阶、多跳推理任务，也难以满足医疗、信息安全等关键领域的可解释性需求。

大数据知识工程可以从多源大数据中挖掘碎片知识，融合成人类可理解、机器可表征与可推理的知识库/知识图谱[1]，可显著缓解上述技术的局限性，为求解实际工程问题提供支撑。与传统知识工程不同，大数据知识工程的知识获取过程以机器为主、人工为辅，有效地缓解了传统知识中“知识获取”的瓶颈问题；同时，大数据知识工程生成的符号化知识有助于弥补现有深度学习的局限性，两者融合有望实现“符号+神经”的推理方式，可以同时应对实际工程问题中普遍存在的直觉系统（System 1）与逻辑分析系统（System 2）的推理任务[5]。

为推动大数据知识工程的进一步发展，本文将梳理大数据知识工程的发展现状，总结该领域面临的挑战和未来的研究方向，提出我国大数据知识工程科技与产业高质量发展的对策建议，助力大数据知识工程的落地应用，服务我国经济社会发展。

二、大数据知识工程的发展现状

（一）大数据知识工程概述

数据–信息–知识–智慧体系（DIKW）模型[6]自底向上刻画了从数据、信息、知识到智慧的层次关系以及不断增值的过程，广泛应用于知识管理领域。据此，本文提出了大数据知识工程研究框架（见图1），包含数据知识化、知识体系化、知识可推理3个阶段。

数据知识化旨在实现数据增值。首先，从多源海量的大数据中挖掘出能够用于问题求解的碎片知识，形式包括文本片段、图像、逻辑规则等；其次，通过去冗消歧，实现碎片知识的量质转换；最后，采用表征学习方法，将不同模态的碎片知识表征到一个低维稠密的公共空间中，为后续推理计算的跨模态互操作提供支撑。碎片知识与输入数据相比，不仅从规模上得到约简，而且实现了由低质向

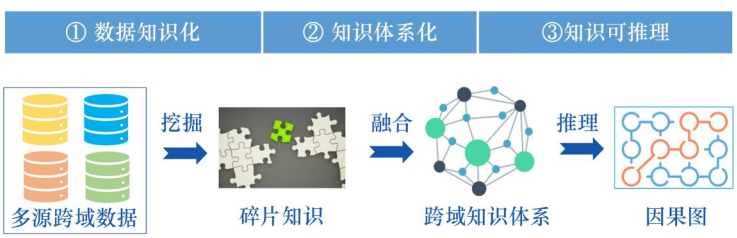

图1 大数据知识工程的研究框架

可信、非结构化向结构化的转化，由此提升了数据的价值密度。

知识体系化是根据实际工程问题将跨域的碎片知识融合成知识体系，实现知识增值的过程。首先，挖掘出碎片知识之间的因果、前序等语义关系。例如，在计算机领域中，与“线性表”相关的碎片知识与“堆栈”就存在前序关系，必须先学习前者，再学后者。其次，通过对碎片知识及语义关系进行非线性融合，产生不同于已有碎片知识的新知识，实现“整体大于部分之和”。

知识可推理是根据知识融合生成的知识体系，找出求解工程问题所需的推理路径的过程。传统符号系统擅长确定性推理，易于刻画显性知识，具有可组合、可解释等特性，但也存在组合爆炸问题[7]，并在不确定性推理、隐性知识刻画等方面存在局限性。基于深度学习的机器推理模型具有较强的表征与学习能力，但泛化能力有限，且大多属于黑盒模型，存在可解释性问题[8]。因此，仅依靠传统符号系统或深度学习模型难以满足实际中的复杂推理需求，需要融合符号推理和深度学习[9]。此外，推理过程涉及诸多优化目标，包括精准度、时效性、可解释性，这些目标又可以分解为多个子目标，因此，在实际工程问题中的机器推理是一个多步骤、多目标组合优化难题。

（二）大数据知识工程的关键技术

基于大数据知识工程研究框架，本文给出了大数据知识工程的技术体系（见图2）。该技术体系包括知识获取与融合、知识表征、知识推理等核心技术。具体来看，知识获取与融合包含知识图谱构建、逻辑公式抽取、基于知识森林的知识融合等技术；知识表征包含符号化表征和分布式表征等技术；知识推理包含知识检索推理、自动问答推理、有记忆推理和可解释推理等技术。其中，知识获取与融合技术和知识表征技术能够解决数据知识化和知识体系化问题，知识推理技术能够解决知识可推理问题。

1. 知识获取与融合

知识获取是从单个或多个数据源中提取知识并形成知识库的过程，是后续知识表征与知识推理的前提和基础。知识图谱与逻辑公式是当前两种主流的知识库组织形式。① 知识图谱最初是由谷歌公司提出，用来优化搜索引擎的技术，用于描述现实世界中的概念及其相互关系。知识图谱用资源描述框架（RDF）三元组和属性图表示知识，其目的在于从数据中识别、发现并推断事物与概念间的复杂关系。知识图谱的构建涉及实体抽取、关系抽取、事件抽取等内容。其中，实体抽取是从文本中检测出命名实体并将其分类到预定义的类别中，如人物、组织、地点、时间等[10]；关系抽取指根据文本上下文识别两个或者多个实体之间的关系，如“出生于”“首都”“夫妻”关系等[11]；事件抽取指识别文本中关于事件的信息，并以结构化的形式呈现。此外，一个高质量的知识图谱还需要实体融合、关系推理等步骤，可用于知识问答、语言理解、决策分析等多个领域。② 逻辑公式是一种通过谓词、量化符、操作符、参数等来描述客观事物逻辑关系的形式化语言。逻辑公式包括命题逻辑公式、一阶逻辑公式以及高阶逻辑公式。逻辑公式抽取旨在将大量知识通过修改和扩充逻辑表达式的方式来完成对碎片知识的归纳。与知识图谱相比，逻辑公式是对知识更高层次的总结与归纳，且具有更好的可解释性。在知识图谱基础上抽取泛化性更强的一阶逻辑公式是当前的研究热点。例如，从“统计量”角度出发，首先生成一阶逻辑公式候选集，然后根据特定的评估函数筛选符合要求的一阶逻辑公式[12]。此外，可以通过构建可微分网络模型来同时学习一阶逻辑公式的置信度和结构信息，使其具备良好的通用性和泛化能力[13]。

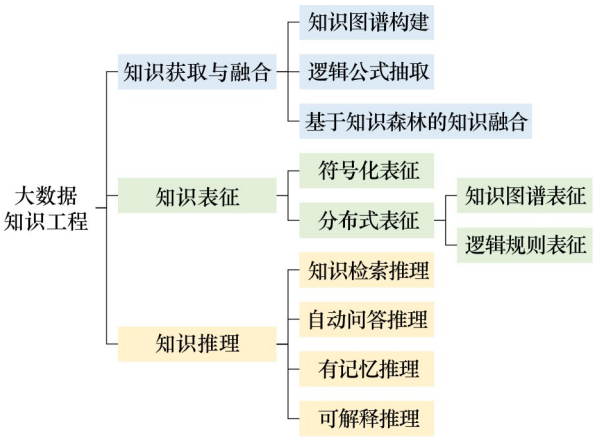

图2　大数据知识工程的技术体系

在知识获取的基础上，知识融合的根本性问题在于如何对不同来源的数据进行合并，去除冗余知识，实现对知识的最优整合。为解决这一问题，西

安交通大学研究团队提出了一种创新的知识融合模型[14]，即知识森林。如图 3 所示，知识森林采用“分面聚合”与“导航学习”相结合的策略，形成由主题分面树（右图树形结构）与学习依赖关系（左侧森林中路径）结合的知识层次结构。知识森林的构建包括主题分面树生成、碎片化知识装配、认知关系挖掘 3 个步骤。其中，分面树生成旨在挖掘领域内具有饱满内容信息的知识主题及其更细粒度的分面结构，可以先通过主题分面联合学习算法生成主题和分面集合，然后基于基序（Motif）结构挖掘每个主题的分面层级结构[15]。碎片化知识装配旨在学习文本、图像等碎片知识和主题分面树的映射关系，形成实例化的主题分面树，以图文并茂的知识表达形式为学习者提供更加全面的主题内容[16]。上述知识森林构建过程可以通过运用自然语言处理、计算机视觉、跨媒体挖掘等技术得以实现。学习依赖关系表现为在学习某项知识主题之前必须要掌握该知识主题的前提知识，这类关系的挖掘可通过分析知识主题的分布特性与语义特性、认知关系的局部性与非对称性来实现[17]。知识森林是一种创新的知识库形态，可为知识检索、智能问答、问题生成等推理任务提供知识支撑，在教育、税务、医疗等多个领域具有应用前景。

整体来看，当前的知识获取和融合方法已取得显著成效，但这些方法大都是基于封闭域的方法，预先设定了特定知识类型集合，难以满足实际应用中新知识不断衍生、更新的需要。因此，如何实现开放域知识的获取与融合仍是未来研究的一项挑战。

2. 知识表征

传统的基于符号逻辑的知识表征方法，包括产生式规则、霍恩逻辑、脚本理论等，能够刻画显式、离散的知识。这类方法的计算和推理能力弱，难以挖掘复杂知识实体间的语义关系。与之不同，分布式的知识表征将知识转化为便于计算机存储和计算的向量形式，更有利于后续的复杂推理，是实现高效人工智能系统的关键。

知识分布式表征经历了从浅层表征到深层表征的过程。20 世纪初，研究人员主要关注浅层知识表征方法，包括主成分分析、线性判别分析、流形学习、多层感知机等；21 世纪初，面向神经网络的贪婪分层预训练和参数微调方法掀起了深层知识表征的热潮[18]。与浅层表征相比，深层表征方法的网络隐层数明显增多、参数量增大，可以更准确地学习大数据内部隐藏的规律，进而准确刻画知识在语义、结构等方面的特性。近年来，计算机硬件资源的提升又进一步推动了基于深度网络的知识表征方法的发展。

知识分布式表征主要分为知识图谱表征和逻辑规则表征两类。① 知识图谱的表征学习旨在将知识图谱中的实体和关系嵌入到连续的低维向量空间中，主要分为直推式学习与归纳式学习两类[19]。直推式学习旨在挖掘知识图谱中实体和关系的特征信

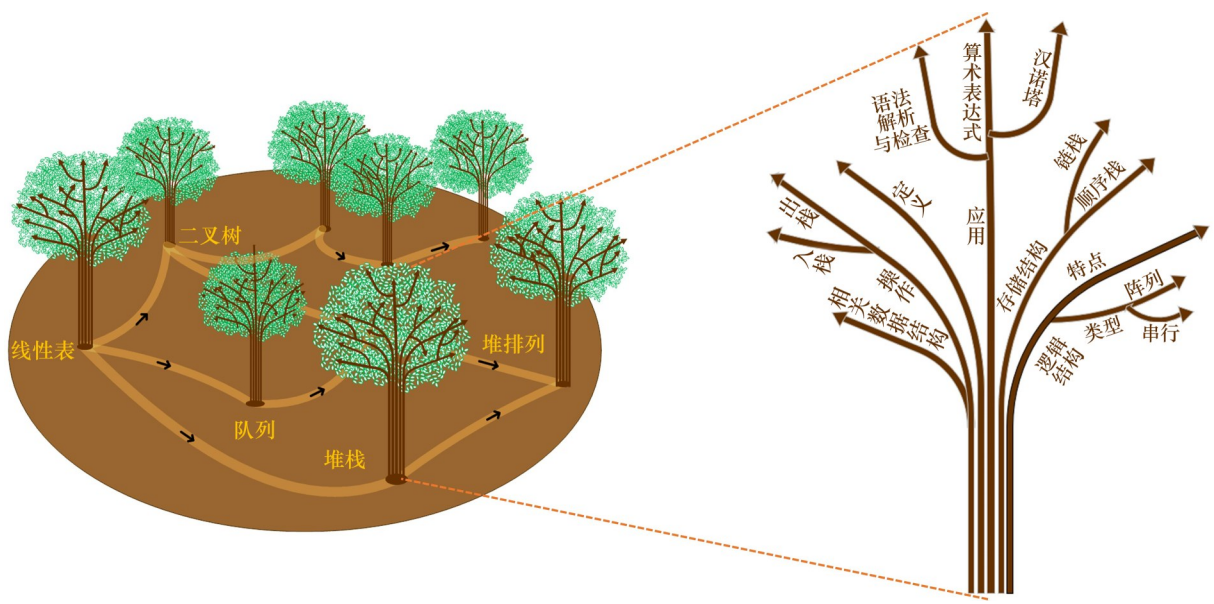

图 3　知识森林

息，同时将特征信息用于知识图谱中隐藏链接的补全。以 TransE[20]、RESCAL[21]等为代表，直推式方法对知识图谱中已知的实体、关系或整条三元组进行特征表示，并设计合理的得分函数以衡量三元组特征嵌入的合理性。归纳式学习主要用于提取当前知识图谱外部的实体和关系的潜在表征，需要模型拥有更高的泛化能力。以 GraIL[22]方法为例，利用知识图谱的实体无关性挖掘三元组的局部子图拓扑结构信息进行潜在特征提取，是知识归纳式学习的主要方法。② 逻辑规则表征旨在将离散的符号化逻辑公式映射到低维连续空间，是连接符号主义和联结主义的纽带之一。逻辑公式表征的过程虽然可能存在部分信息损耗，但由于输入样本一般含有噪声，将其嵌入到低维空间可以过滤掉部分噪声，提升模型的泛化能力，并有效减少逻辑公式的存储与计算成本。逻辑公式表征学习首先将逻辑公式转换为对应的句法结构，之后使用神经网络模型进行嵌入。根据使用句法结构和表征网络的不同，逻辑规则表征学习研究可以分为基于序列、基于树结构和基于图结构的逻辑公式表征方法[23]。其中，基于序列的方法将逻辑规则视为简单的符号序列形式，之后通过神经网络进行嵌入[24]；基于树结构的方法通过句法解析工具将逻辑规则转换为树结构并进行嵌入[25]；基于图结构的方法常采用图卷积神经网络强化逻辑规则中节点之间的信息交互，以便捕获更深层的结构信息[26]。

近年来，深度学习技术已在知识深层表征学习方面取得重大突破，但仍然存在训练成本偏高、可解释性弱、动态演化难等挑战，未来还需更加深入研究。

3. 知识推理

知识推理是根据已有的知识推断出新知识或识别错误知识的过程。在大数据知识工程中，知识推理以知识表征学习的结果为输入，以计算机视觉、自然语言处理、跨模态学习等技术为手段，输出推理结果。典型的知识推理技术包括知识检索推理、自动问答推理、有记忆推理和可解释推理等 4 种。

知识检索推理是在知识组织的基础上从知识库中检索出知识的过程。给定一组查询，知识检索技术需要通过对问题进行解析、理解，进而在知识库中完成查询、推理、比较等逻辑运算。最初的知识检索方法由信息检索发展而来，经历了信息检索 – 特定知识库检索 – 知识图谱检索的发展过程。随着知识库规模的不断增大，未来的知识检索将面临知识图谱模式复杂性高、检索算法复杂性高和泛化性弱等问题。

自动问答推理是根据用户的自然语言问题在已有资源上进行查询与推理，最终将精准答案返回给用户。根据推理空间中不同的资源组织形态，可将自动问答分为自然语言问答[27]、跨模态问答[28]和视觉问答[29]等。例如，教科书式问答[28]是面向智慧教育的智能答疑，是教育领域的一项跨模态问答推理任务。作为自动问答的对偶问题，问题生成可以为自动问答系统提供必要或额外的数据，能够与问答系统有机地结合在一起而互相促进。

模型具备推理能力的关键前提是模型具有记忆能力。相比其他推理模型，有记忆推理模型能够保存更多的信息，可以在后续推理任务中加以使用。有记忆推理模型的发展经历了长短期记忆（LSTM）网络[30]、神经图灵机[31]、记忆网络[32,33]和可微神经计算机（DNC）[34]等阶段。其中，DNC 采用外部存储矩阵作为神经网络的“记忆”，采用一个变体的 LSTM 作为“控制器”，具有强大的记忆管理能力，可以选择性地写入和读取记忆，允许对记忆内容进行反复地修改[34]。因此，DNC 在某种程度上更接近于人类大脑的能力。

深度学习模型的高复杂性和黑盒属性使模型无法为推理的结果作出解释。根据解释产生的方法，推理模型一般分为事前解释和事后解释模型[35]。最近，为了实现推理过程的可控制和可干预，研究人员提出了符号化分层的可解释推理模型（SHiL）[1]。该模型属于事前和事后解释推理模型的融合。SHiL 的核心思想是“分层递阶可控+符号化知识驱动”，即基于介科学理论[36]，将多层次多尺度动态时空关联的复杂数据系统划分成若干介区域，形成分层递阶结构。同时，针对每个介区域的功能和状态特点，构建内嵌着物理学或社会学知识的符号化控制机制（如常识、规则等）。SHiL 模型具有可理解、可编程、可干预的特点，实现了知识驱动的数据计算及推理。

最近，知识推理的发展已经进入融合符号主义与联结主义的阶段，即利用前者规则的逻辑推理能力和后者深度学习的自主学习能力，构建更加强大的知识推理模型。

（三）大数据知识工程的工程应用现状

1. 智慧教育

智慧教育旨在运用现代信息技术改变传统教育模式，促进教育改革与发展。教育大数据指整个教育活动过程中所产生的、根据教育需要所采集到的、一切用于教育发展并可创造潜在价值的数据集合。大数据驱动的教学范式具有高效率、有智慧、产业化等优点[37]。在教育资源方面，可采用知识图谱等技术聚合多个地区、多种形态的优质资源，并对这些资源进行表征和深度分析，为教师教学和学生自主学习提供资源支撑；在“教”方面，利用教育大数据，可以生成教学方案、模拟教师作出决策等，大幅减轻教师工作负荷，实现优质师资的快速、规模化“复制”[38]；在“学”方面，通过分析学生的兴趣、能力、学习状态和知识掌握能力，精准规划学生学习路径、学习资源等，实现因材施教[39]。

近年来，西安交通大学研究团队将知识森林理论成功应用于在线教育，研发了知识森林导航学习系统，解决了散、杂、乱碎片知识的结构化和体系化描述问题，优化了海量在线教学资源的组织方式，提高了在线学习效率和备课质量[14]。在此，以“万有引力”知识的获取、学习和备课为例进行简要介绍。① 将以往利用搜索引擎在互联网上漫无边际查找学习资料的方式转变为知识森林导航学习系统指导下的学习资源查找。在查找“万有引力”知识点时，系统将给出与“万有引力”相关的知识体系，实现“既见树木，又见森林”，既能方便地获取某个特定知识点的知识，也能从宏观上得到与之相关的知识点。② 知识森林提供了个性化的导学路径推荐。在开展在线教育时，利用知识森林为学生提供一系列导学功能。例如，能够为学生生成一条符合“万有引力”学习目标以及认知能力的学习路径，避免无目标、无头绪的乱学，即解决所谓的“学习迷航”问题；能够解答学生学习中与课程知识相关的问题，帮助学生答疑。

知识森林导航学习系统已在高等继续教育及国际教育培训领域进行了应用，验证了大数据知识工程在教育领域的应用价值。在高等继续教育领域，基于知识森林构建技术及导航学习技术研发建成的“慕课（MOOC）中国”学习平台，促进了我国MOOC平台的做大做强，抢占全球MOOC智能导学技术制高点。在国际教育培训领域，基于知识森林构建技术和导航学习技术创建了国际工程科技知识中心（IKCEST）丝路工程科技发展专项培训系统，服务于俄罗斯、泰国、吉尔吉斯斯坦、乌兹别克斯坦等“一带一路”国家，培养了来自100余个国家的4万多留学生以及在华涉外企业人员。

2. 税务风险管控

智慧税务旨在推动现代信息技术新成果与税收工作深度融合，促进纳税服务进一步便捷普惠、税收征管进一步提质增效、税收执法进一步规范透明，最终目标是全方位提高税务服务能力、监管能力和执政能力。事实上，税务场景包含政策法规、报表、发票、预算、结算等相关数据，如何有效利用此类海量、低质、无序的碎片信息，并实现自动化辅助决策是智慧税务治理面临的重要挑战。运用大数据知识工程方法，一方面可自动化地从海量税务数据中获取蕴含的法规、经济、行业等知识，另一方面能够对提炼的知识进行推理和应用，解决税务领域面临的智能化决策支撑、可解释的税收监管等关键难题[40]。

从税收服务的角度来看，运用大数据知识工程，可有效实现税收政策与纳税人的双向精准匹配，以应对税收政策文本、纳税人经营情况实时变化带来的挑战。首先，从税收政策文本中获取多类规则和条件（包括行业属性、纳税人属性、税种信息、涉税约束等），并采用知识融合技术对知识库中的规则进行重复合并、失效剪裁，构建规则知识库；随后，对相关知识进行规则编码，构建决策表；最后，可结合实际业务需求，将获取纳税人数据交由规则计算引擎自动计算税额、自动填写申报等，最大限度地减少纳税人的时间及心理成本，并保证各类税收政策的应享尽享[41]。

从税收监管的角度来看，大数据知识工程方法可以从企业资金流、发票流、合同流、物流中抽取碎片化知识，结合财税行业特征知识，构建面向税务部门的财税知识库。随后，通过运用知识表征和符号化知识推理技术，将风险线索依据时序、依赖、因果等关系进行动态融合，生成推理路径和证据链，提高涉税违法行为稽查结果的可解释性，从而主动发现潜在涉税违规企业，帮助税务部门有效控制企业的犯罪风险，减少偷税漏税带来的财政损失，促进精准监管和精确执法，同时避免对诚信纳

税人的打扰。此外，对于涉税违规企业，不仅可以得到识别结果，还可以给出相关的证据链以保证可信性、公信力和执行力。

3. 智慧医疗

智慧医疗是一种以居民健康医疗数据为核心，融合物联网、云计算、人工智能等新兴技术的综合服务模式。“十三五”以来，随着医疗信息化的高速发展，包括以电子病历为核心的临床系统建设、以控费为目的医保控费系统建设、“互联网+”医疗信息系统改进以及以医联体为载体的区域卫生信息化建设，积累了海量的医学数据。如何从这些数据中提取信息，进行有效管理、分析和应用，是实现医学知识检索、临床诊断、医疗质量管理以及电子健康档案智能化分析处理的基础。构建医学知识图谱则是实现上述目标的关键手段。

中文医学知识图谱CMeKG[42]是基于大规模医学文本数据，以人机结合的方式研发而来的。该知识图谱的构建参考了国际疾病分类体系（ICD）[43]、解剖学治疗学及化学分类系统（ATC）[44]、医学系统术语表（SNOMED）[45]、医学主题词表（MeSH）[46]等权威国际医学标准以及规模庞大的临床指南、行业诊疗规范以及医学百科知识等信息。CMeKG 1.0（2019年1月）包括6000多种疾病、10 000多种药物（西药、中成药、中草药）、1200余种诊疗技术及设备的结构化知识描述，涵盖疾病的临床症状、发病部位、药物治疗、手术治疗、鉴别诊断、影像学检查以及药物成分、适应症、用法用量、有效期、禁忌症等30多种常见关系类型；CMeKG 1.0中有描述医学知识的概念关系实例及属性三元组超过100万。CMeKG 2.0（2019年9月）则针对多源异构的医学资源进行了知识融合，新增了症状类知识，并对儿科疾病进行了详细描述。拓展后的CMeKG 2.0目前包含超过10 000种疾病、20 000种药物、10 000种症状以及3000种诊疗技术的结构化知识描述，相应的医学知识三元组达156万。

基于医学知识图谱进行医疗信息检索能够提高检索精度，克服传统医疗搜索响应速度慢、存储消耗大等缺点。例如，中医药学语言系统结合“知识卡片”嵌入以及“知识地图”展示，可将中医领域概念知识进行可视化，方便用户针对具体概念进行查询和搜索。国外著名的专用医疗信息搜索引擎有WebMed[47]、Healthline[48]以及Google Health[49]等，其中Google Health在面对具体疾病及症状的搜索请求时，能够提供超过400种健康状况的数据，同时能给出相应的症状描述。

基于医学知识图谱，结合患者症状表现及化验信息，临床决策支持系统（CDSS）可自动生成诊断报告以及治疗方案，并能对医生给出的诊疗方案进行查漏补缺，减少甚至避免误诊情况的发生。我国代表性的CDSS开发者有神州医疗、迈瑞医疗等，国际上有DiagnosisOne、DXplain[50]、Micromedex[51]等。当前，将知识图谱应用于CDSS已成为研究热点，但仍面临全科医学知识图谱不完备、医疗决策置信度不高、基于人工智能方法得到的预测结果缺乏可解释性等挑战。

三、大数据知识工程技术面临的挑战与未来研究方向

随着人工智能、物联网、云计算及区块链等技术的飞速发展，各领域产生了记录人类生产、生活行为的海量数据。基于这些海量数据，如何挖掘其中的模式和规律知识，实现从数据到知识、从知识到决策的转化，是第四范式科学研究要解决的核心问题[52]。最近，受AlphaFold[53]这一任务的启发，研究人员提出了“科学研究的第五范式雏形”的思想[54]，指出需要将领域知识（包括人类先验/专家知识等）融入到算法以及模型的设计中，以更好地解决领域问题。据此，本文分析了大数据知识工程在知识获取、知识表征和知识推理等方面面临的挑战，并探讨了解决这些挑战潜在的未来研究方向。

（一）知识获取

传统的知识获取技术更专注于从海量文本数据中挖掘潜在知识，在模态多样性和知识类型上存在较大的局限性。未来，如何获取蕴含信息更加丰富的视觉知识和隐匿性强的常识知识将是知识获取技术的发展方向。下面对这两类知识进行介绍，并分析其潜在的研究方向。

1. 视觉知识获取

视觉知识是一种有望提高跨媒体知识表达能力，进一步推动人工智能发展的新框架[55]。认知心理学理论表明，视觉记忆是区别于语言记忆的特殊存在，人类可以对脑内的视觉记忆根据需要进行折

叠、旋转、扫描、类比等操作[56]。这类记忆被认知心理学家称为“心象”，在人工智能领域被称为视觉知识。视觉知识具有以下特性：① 能表达对象的空间形状、大小、空间关系以及色彩和纹理；② 能表达对象的动作、速度及时间关系；③ 能进行对象的时空变换、操作与推理，包括形状变换、动作变换、速度变换、场景变换、各种时空类比、联想和基于时空推理结果预测。如何有效处理并合理运用视觉知识成为人与信息及信息机器交流最重要的途径。

视觉知识具有多种表达形式，根据知识的连续与离散表达可以将其划分为静态视觉知识和动态视觉知识。静态视觉知识又称为视觉常识，指从真实世界场景中可收集到的静态视觉事实以及社会主体根据该事实可预知的信息或做出的推论。计算机对于视觉常识知识的研究是极其困难的。一方面视觉常识知识的广度巨大，且计算机缺乏类似人类对于常识知识积累的先验知识。另一方面，除了视觉元素上低级的识别类任务外，计算机需要对图像中隐含的上下文信息进行更深入的理解。动态视觉叙事是指由一组连续的静态视觉知识组成的、以时间关系或空间关系为序列的知识表达。空间关系表达为场景结构，描述各对象之间的上下、左右、前后等方位关系以及距离关系、里外关系、大小关系；时间关系表达为动态结构，表达对象的生长、位移、动作、变化、竞赛、协同等。

另外，近年来，学术界开始关注示意图这一种高级的静态视觉知识。示意图是一种采用图形化元素来呈现的视觉表示形式，通常用于表达某些专业领域中特定知识主题或知识概念的内在规则 / 逻辑信息。示意图广泛分布在 MOOC 网站、开放知识库、技术论坛等知识源中。对这类特殊图像的分析与理解是知识库构建、智能答疑等知识密集型任务的基础，也是跨媒体智能的重要组成部分。在底层视觉特征方面，示意图的颜色、纹理、背景等信息远不如自然图像丰富，该视觉特征的稀疏特点导致在模型训练阶段易出现过拟合、难收敛等问题。在高层语义表达方面，示意图具有不同于自然图像的“同形不同义、同义不同形”现象。以图 4 为例，“太阳系”与“原子”示意图的形状相似但意义完全不同。示意图的这一现象使理解面临更为严峻的语义鸿沟问题。

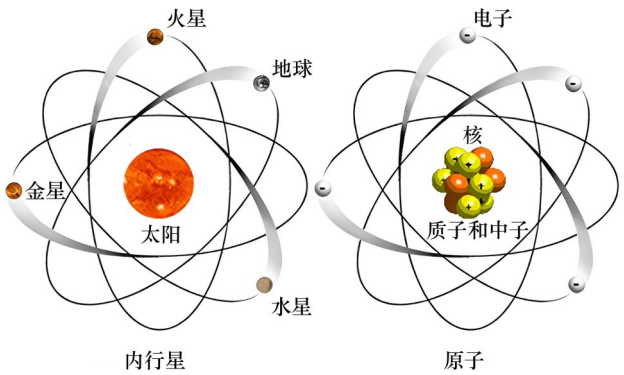

图 4 “同形不同义”的示意图示例

视觉知识理论不仅可以促进跨媒体表达的研究，也可以支撑和提升诸如智能创作、逻辑推理等更为广泛的人工智能领域的研究和应用。目前，很多研究尚未正式引入视觉知识的概念，同时视觉知识在结构化表示、操作与推理、重建与生成等方面都存在一定的局限性。

2. 常识知识获取

常识知识指人们对现实世界中不同事物间的联系达成的有效共识，涵盖大量人类经验，被广泛接受、无需解释和论证[57]。常识知识可以使计算机尽可能像人一样熟悉所有的事实和信息并进行推理决策，在机器问答、会话情感识别、故事结尾生成等方面发挥巨大作用。

常识知识具有以下 3 种特性。① 概念性，绝大多数常识知识是概念知识，表示某一类事物的共有特征，而非某一实体的独有特征。② 一般性，常识知识蕴含的概念被广泛接受并具有一般性。例如，“人呼吸需要氧气”是常识知识，而“细胞膜的组成需要胆固醇”只是被特定领域的专家所知，更具有专业性，因此不是常识知识。③ 隐含性，常识知识是普遍共享的，在人们的口头或书面交流中通常被省略。常识知识的类型表现非常多样。例如，ConceptNet[58]、ATOMIC[59]是典型的常识知识图谱，它们将常识知识表示为关系三元组，并把这些关系三元组组织成网络结构。以 WordNet[60]、Roget[61]为代表的词汇数据库是按照一定规则需求由知识专家人工编撰构建的知识源，也是常识知识。BERT[62]等预训练语言模型也被认为是常识知识的一种表达形式。这些模型通常基于大型语料库训练得到，可以有效地捕获句法特征、语义信息和事实知识。在自然语言处理的相关研究中，上述常识知识可作为

背景语义，显著增强上下文语义信息；在计算机视觉相关研究中，常识知识可以改善导航、操纵、识别等各项下游任务的性能，从而实现真正意义的人工智能。

对常识知识的认知水平不足仍是人工智能发展的重要瓶颈。常识知识是多元化的，包括但不限于直觉、心理、视觉、情感等多种形式以及文本、图像、语音等多种模态。因此，如何将跨语言、跨模态的多源数据中对某一事件、概念及关系要素进行链接与融合，以获得丰富的常识知识和表示将是一个重要研究方向。另外，目前的大规模常识库虽然包含了人类的一些情绪状态、隐含语义和可能行为等，但很少强调人类在日常生活中广泛采用的社交互动模式，如怎样以同理心的方式回应别人等。因此，如何利用网络上丰富的动态对话资源来构建社交常识知识库以更有利于各类机器对话、问答、聊天等下游任务的建设，是一个重要的研究方向。

（二）知识表征

在海量标注数据和超强计算能力的推动下，现有知识工程技术在众多领域与任务上的性能已经全面接近甚至超越了人类。然而，知识表征技术仍然存在模型复杂度高和可解释性差等现实挑战。具体表现为：首先，深层表征和推理模型结构复杂、参数量庞大、训练难度极大。例如，文本表征模型GPT-3包含超过1700亿个内部参数，使用了45 TB数据来训练[63]。其次，大多数深层表征模型都属于黑盒模型，难以对模型内部机制和结果进行理解，其对应的优化方案也无法明确。

与之相比，人类生来就具备对知识进行编码和记忆的能力，这依托于人类大脑复杂的结构与机制。人脑可以自主地表征知识、归纳学习、推理知识，并可以并行执行多项不相关的任务；此外，相比于知识工程技术所需的庞大计算成本，人脑能够做到在保持相对较高效率的同时维持低能耗。因此，人类大脑仍然是目前唯一的真正智能系统，学习大脑的各项复杂机制，建立更强大和更通用的知识表征模型是非常有前景的。接下来，介绍大脑在知识表征与序列记忆处理方面的最新进展，为大数据知识表征技术的下一步发展方向提供借鉴。

知识在大脑中如何表征一直是科学研究的前沿问题。认知神经科学家已经证明，空间位置信息和抽象知识信息在大脑中都是以认知地图的形式存储在海马体中的。为了探究复杂活动中大脑的知识编码机制，如同时涉及空间位置变化和抽象认知变量的任务，有研究[64]构建了小鼠在执行认知决策任务时背侧海马1区的神经活动空间。实验结果表明，神经元对空间位置信息和抽象认知变量的编码是同时进行的，且互相依存。此外，通过神经流形空间对虚拟场景下小鼠在运动状态时的群体神经元活动进行降维，发现海马体神经元群体活动对空间位置信息和抽象认知变量的表征都呈现出很强的几何结构特点；同时，这些表征知识的几何结构特定于具体的任务而存在。最后，研究还发现，富含抽象认知信息的神经元能够让生物做出预测和判断行为。这项研究揭示了大脑中复杂知识的表征具有明显的几何结构特点。因此，在设计新的知识表征模型时，可以借鉴流形学习方法对低维空间中表征的知识进行结构判断与评价，以提高模型的知识表征能力。

人类大脑无时无刻不在处理序列信息，不论是语言沟通、动作实施还是情景记忆，本质上都涉及对时序信息的表征，因此序列记忆是大脑的一项基本认知功能。为了探究时序记忆编码问题，在最新研究中[65]，研究人员利用在体双光子钙成像技术，记录了猕猴外侧前额叶皮层（负责工作记忆的区域）上数千个神经元的活动。实验结果表明，每个次序的信息都可以为其在高维的钙成像数据中找到一个对应的二维子空间。在每个子空间中，每个点所处的位置与猕猴看到的真实六边形结构相对应；而且3个不同次序的信息对应的子空间彼此接近正交，即序列中的每个信息在大脑中都有独立的存储空间。此外，研究人员还发现，靠后次序信息的子空间中六边形环状结构的半径小于靠前的次序空间，这一结构也对应了序列记忆的行为表现，即生活中要记忆的内容越多，越往后的信息越容易出错。这项序列工作记忆研究揭示了大脑神经元存储序列记忆的编码机制，其对应了一种将不同次序子空间内的结构信息嵌入高维向量空间的表征方式，将对脑启发的知识编码与记忆提供重要的借鉴。

（三）知识推理

伴随深度学习的发展，知识推理模型的设计越来越复杂，并被广泛应用于诸多领域。实践表明，

这些复杂模型在推理速度、精度以及稳定性上都已超越人类水平，但仍然面临一定的挑战。具体表现为：用户难以直观理解模型中的参数、结构以及产生的特征，无法精确掌握模型在推理和决策时的依据。这促使学术界和工业界对新型知识推理框架进行探索。近年来，反事实推理和可解释推理模型逐渐引起研究人员的重视，成为大数据知识推理技术的下一步发展方向。下面对这两种推理模型进行介绍，并分析其今后研究方向。

1. 反事实推理

反事实推理又称反事实思维，指对过去已经发生的事实进行否定和重新表征，构建一种可能性假设的思维活动。反事实逻辑推理能力是人类智能的重要表现之一，在当前人工智能的研究热潮中，研究者们意识到，具有像人类一样的因果推断和反事实推理的能力，是从弱人工智能走向强人工智能的象征。因果关系具有自下而上的3个层次，分别是关联、干预和反事实[66]。反事实处于“因果关系之梯”的最顶层，如图5所示。

反事实推理需要基于观测数据执行，为此，研究者们设计了多种反事实推理框架，其中最为著名的是潜在结果框架（POF）[67]和结构因果模型（SCM）[66]。POF借鉴了统计学中的随机对照试验和潜在结果的概念，构建了基于因果推断的分析框架，其核心思想是“没有假设就没有因果”，即如果现实情况不能满足基本假设，潜在结果的结论就不成立。POF

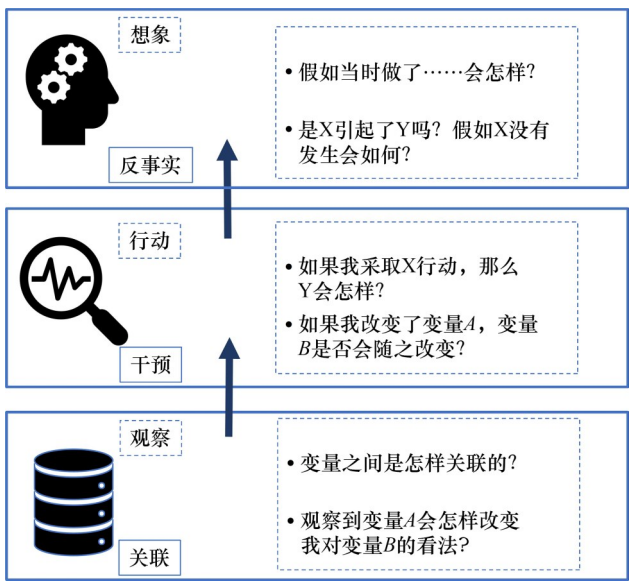

图5 因果关系之梯

中常用的3个基本假设为：研究对象干预值稳定假设、可忽略性假设以及正值假设。在这3个假设基础上，研究者们设计出对应的因果推理方法如匹配法、逆概率加权法以及分层方法等。SCM则是通过构建因果图与结构方程，探究反事实因果关系。在该体系下，因果关系的推断依托于有向无环图的3种基本路径结构：链状结构、叉状结构和对撞结构。3种结构各有不同的信息传递方式，所有的因果图均能拆解为这3种结构的组合。SCM通过对变量间的因果关系参数化，并利用结构方程模型做推理。

在因果推断与大数据知识工程两个领域呈现交织融合的背景下，反事实推理也在大数据知识工程领域发展迅速，并在视觉问答（利用反事实推理消除语言偏差）、重复问题识别（利用反事实推理取代传统统计量分析）等任务领域等取得了成功。尽管如此，基于反事实推理的通用理论体系仍未建立，如何有效整合实际数据、明确评价指标和目的，如何基于多模态数据设计可拓展的推理模型等问题亟待解决。

2. 可解释推理

近年来，可解释推理已成为学术界和工业界的研究热点。然而，对于可解释性的定义，目前尚未形成统一认识，一种业界比较认可的定义是：可解释性是一种以人类知识、理解的方式给人类提供解释的能力。在一些低风险情形下（如电影推荐），人们可以不关注模型为何作出这种判断，但在高风险情形（如自动驾驶、药物推荐等），除了获得高准确率的预测结果，模型还必须解释如何作出当前的预测。这种对模型高可靠性要求进一步提高了对可解释性研究的需求。

依据解释产生的方法，可将推理模型大体划分为两类：事先解释和事后解释。前者主要指利用模型架构自带的解释而不借助额外的解释方法；后者主要指利用不依赖于模型自身的解释方法对推理结果进行解释。若一种方法可以解释黑盒模型，那么该方法可以：① 利用透明模型（如决策树、规则列表及线性模型等）近似模型推理的过程；② 能够对模型基于特定样例进行预测并作出解释；③ 能够了解模型内部的特定属性（如深度神经网络中神经元在某一决策中的作用）。值得注意的是，事后解释方法也可用于事前解释方法中。

尽管当前可解释推理模型在医疗（如临床决策

支持系统）、金融（如偷税 / 逃税 / 骗税检测）、交通（如自动感知 / 控制 / 决策）等民生领域展示出良好潜力，但总体研究还处于起步阶段，仍面临诸多挑战。例如，推理模型性能不足；一些表现较好的推理模型与领域强相关，可拓展性差；如何在同一任务 / 场景下，评判不同可解释性方法的优劣等。对这些问题的突破将推动可解释性推理的快速发展。

四、我国大数据知识工程发展建议

（一）多学科交叉融合，推动大数据知识工程的理论与技术攻关

多学科交叉融合是科技创新和理论创造的重要源泉，能够推动我国大数据知识工程技术的高质量发展。首先，建设大数据知识工程前沿交叉研究特区，设立大数据知识工程的重大 / 重点研发专项。以大数据知识工程相关联合实验室建设为抓手，促进计算机科学、人工智能与其他学科的深度交叉融合。其次，为学科交叉融合提供强有力的体制机制保障。做好学科交叉的顶层规划，理顺交叉学科学位授予机制体制，成立学科交叉服务平台，探索新兴交叉学科的评价方法。

（二）建立大数据知识工程的行业标准体系

大数据知识工程相关术语和适用准则等标准的建立是衡量行业技术发展水平的重要标志，是创新发展的引领和推动力量。首先，通过加强沟通、深化合作，整合并充分利用国内外大数据知识工程相关企业、研究机构的优势资源，重点突破知识获取、融合、表征、推理技术等。其次，推广相关的前沿研究成果，形成应用示范效应，打造行业应用标杆，优选出市场认可的通用标准和规范，从而促进行业技术标准体系的不断发展与完善。

（三）以需求为牵引，推动大数据知识工程在各个行业的工程应用

以大数据知识工程理论和技术攻关以及行业标准制定为契机，面向市场需求，打造基于“基础研究 – 技术创新 – 产业化”路径的“产学研”协同发展机制。首先，在高校及科研机构层面，发挥办学特色，集合院校优势学科，探索符合时代以及市场需求的校企协同育人模式。同时，在大数据知识工程及其应用技术方向投入相关资源，制定并完善相应的人才培养方案，增强技术推广过程中应用型人才的培育，注重培养学生的创新潜能。其次，在企业层面，紧扣市场需求，深化市场调研并积极布局，瞄准国际领先的发展目标，坚持以应用为主导开展研发，前瞻论证大数据知识工程交叉领域创新性研究的重点方向，通过示范效应带动整个产业链的深化拓展。

利益冲突声明

本文作者在此声明彼此之间不存在任何利益冲突或财务冲突。

Received date: January 16, 2023; **Revised date:** March 2, 2023

Corresponding author: Liu Jun is a professor from the School of Computer Science and Technology of Xi’an Jiaotong University. His major research fields include natural language understanding, computer vision, smart education. E-mail: liukeen@xjtu.edu.cn

Funding project: National Natural Science Foundation of China (62250009); China Engineering Science and Technology Knowledge Center Project (CKCEST-2022-1-40)

参考文献

[1] 郑庆华, 张玲玲, 龚铁梁, 等. 大数据知识工程 [M]. 北京: 科学出版社, 2022.
Zheng Q H, Zhang L L, Gong T L, et al. Big data knowledge engineering [M]. Beijing: Science Press, 2003.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]. Long Beach: The 31st International Conference on Neural Information Processing Systems, 2017.

[3] Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP [C]. Florence: The 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[4] Xie Z K, He F X, Fu S P, et al. Artificial neural variability for deep learning: On overfitting, noise memorization, and catastrophic forgetting [J]. Neural Computation. 2021, 33(8): 2163–2192.

[5] Bengio Y. The consciousness prior [EB/OL]. (2019-12-02)[2022-12-20]. https://arxiv.org/abs/1709.08568.

[6] Ackoff R L. From data to wisdom [J]. Journal of Applied Systems Analysis, 1989, 16(1): 3–9.

[7] Marcus G. The next decade in AI: Four steps towards robust artificial intelligence [EB/OL]. (2020-02-19)[2023-02-23]. https://arxiv.org/abs/2002.06177.

[8] 张钹, 朱军, 苏航. 迈向第三代人工智能 [J]. 中国科学: 信息科学, 2020, 50(9): 1281–1302.
Zhang B, Zhu J, Su H. Toward the third generation of artificial intelligence [J]. Scientia Sinica Informationis, 2020, 50(9): 1281–1302.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436–444.

[10] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition [C]. San Diego: The Confer-

ence of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[11] Ji G, Liu K, He S, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions [C]. San Francisco: The AAAI Conference on Artificial Intelligence, 2017.

[12] Galárraga L A, Teflioudi C, Hose K, et al. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases [C]. Rio de Janeiro: The 22nd International Conference on World Wide Web, 2013.

[13] Cohen W W. TensorLog: A differentiable deductive database [EB/OL]. (2016-07-19)[2022-12-20]. https://arxiv.org/abs/1605.06523.

[14] 郑庆华, 刘均, 魏笔凡, 等. 知识森林: 理论、方法与实践 [M]. 北京: 科学出版社, 2021.
Zheng Q H, Liu J, Wei B F, et al. Knowledge forest: Theory, method, and application [M]. Beijing: Science Press, 2003.

[15] Wei B, Liu J, Ma J, et al. Motif-based hyponym relation extraction from wikipedia hyperlinks [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(10): 2507–2519.

[16] Wu B, Wei B, Liu J, et al. Faceted text segmentation via multitask learning [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(9): 3846–3857.

[17] Liang C, Wu Z, Huang W, et al. Measuring prerequisite relations among concepts [C]. Lisbon: The Conference on Empirical Methods in Natural Language Processing, 2015.

[18] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation. 2006, 18(7): 1527–1554.

[19] Zha H, Chen Z, Yan X. Inductive relation prediction by BERT [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC: Association for the Advancement of Artificial Intelligence(AAAI), 2022: 5923–5931.

[20] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C]// Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: Massachusetts Institute of Technology Press, 2013: 2787–2795.

[21] Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data [C]//Proceedings of the International Conference on International Conference on Machine Learning. Washington DC: Association for Computing Machinery, 2011: 809–816.

[22] Teru K, Denis E, Hamilton W. Inductive relation prediction by subgraph reasoning [C]//Proceedings of the International Conference on Machine Learning. Washington DC: Association for Computing Machinery, 2020: 9448–9457.

[23] Lin Q, Liu J, Zhang L, et al. Contrastive graph representations for logical formulas embedding [J/OL]. IEEE Transactions on Knowledge and Data Engineering. 2021 [2022-12-20]. https://ieeexplore.ieee.org/abstract/document/9667296.

[24] Irving G, Szegedy C, Alemi A A, et al. Deepmath-deep sequence models for premise selection [C]//Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: Massachusetts Institute of Technology Press, 2016: 2235–2243.

[25] Evans R, Saxton D, Amos D, et al. Can neural networks understand logical entailment? [C]. Vancouver: The International Conference on Learning Representations, 2018.

[26] Xie Y, Xu Z, Kankanhalli M S, et al. Embedding symbolic knowledge into deep networks [C]//Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: Massachusetts Institute of Technology Press, 2019: 4233–4243.

[27] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [C]//Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: Massachusetts Institute of Technology Press, 2020: 1877–1901.

[28] Kim D, Yoo Y J, Kim J S, et al. Dynamic graph generation network: Generating relational knowledge from diagrams [C]. Salt Lake City: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[29] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C]. Salt Lake City: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[30] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation. 1997, 9(8): 1735–1780.

[31] Graves A, Wayne G, Danihelka I. Neural turing machines [EB/OL]. (2014-12-10)[2022-12-20]. https://arxiv.org/abs/1410.5401.

[32] Weston J, Chopra S, Bordes A. Memory networks [EB/OL]. (2015-11-29)[2022-12-20]. https://arxiv.org/abs/1410.3916.

[33] Sukhbaatar S, Szlam A, Weston J, et al. End-to-end memory networks [C]//Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: Massachusetts Institute of Technology Press, 2015: 2440–2448.

[34] Graves A, Wayne G, Reynolds M, et al. Hybrid computing using a neural network with dynamic external memory [J]. Nature, 2016, 538(7626): 471–476.

[35] 侯中妮, 靳小龙, 陈剑赟, 等. 知识图谱可解释推理研究综述 [J]. 软件学报, 2022, 33(12): 4644–4667.
Hou Z N, Jin X L, Chen J Y, et al. Survey of interpretable reasoning on knowledge graphs [J]. Journal of Software, 2022, 33(12): 4644–4667.

[36] Huang W, Li J, Edwards P P. Mesoscience: Exploring the common principle at mesoscales [J]. National Science Review, 2018, 5(3): 321–326.

[37] 刘淇. 大数据驱动的教育变革: 从线下到线上、从人工到智能 [EB/OL]. (2022-05-28)[2022-12-20]. https://www.huaweicloud.com/cloudplus/fourthphase/detail_03.html.
Liu Q. Education reform driven by big data: From offline to online, from artificial to intelligent [EB/OL]. (2022-05-28)[2022-12-20]. https://www.huaweicloud.com/cloudplus/fourthphase/detail_03.html.

[38] Buenaño-Fernandez D, Villegas-CH W, Luján-Mora S. The use of tools of data mining to decision making in engineering education—A systematic mapping study [J]. Computer Applications in Engineering Education, 2019, 27(3): 744–758.

[39] Wang H, Fu W. Personalized learning resource recommendation method based on dynamic collaborative filtering [J]. Mobile Networks and Applications, 2021, 26(2): 473–487.

[40] 郑庆华, 师斌, 董博. 面向智慧税务的大数据知识工程技术及其应用 [J/OL]. 中国工程科学, [2022-12-09]. https://kns.cnki.net/

kcms/detail/11.4421.G3.20221208.1118.002.html.
Zheng Q H, Shi B, Dong B. Technologies and applications of big data knowledge engineering for smart taxation systems [J/OL]. Stragetic Study of CAE, [2022-12-09]. https://kns.cnki.net/kcms/detail/11.4421.G3.20221208.1118.002.html..

[41] 余红艳, 孙丽, 刘亚利. 减税政策: 动因追溯、制度约束与路向选择 [J]. 税务研究, 2022 (7): 32–37.
Yu H Y, Sun L, Liu Y L. Tax reduction policy: Motive tracing, institutional constraint and direction choice [J]. Tax Research, 2022 (7): 32–37.

[42] 奥德玛, 杨云飞, 穗志方, 等. 中文医学知识图谱CMeKG构建初探 [J]. 中文信息学报. 2019, 33(10): 1–7.
Byambasuren O, Yang Y F, Sui Z F, et al. Preliminary study on the construction of Chinese medical knowledge graph [J]. Journal of Chinese Information Processing, 2019, 33(10): 1–7.

[43] Sundararajan V, Henderson T, Perry C, et al. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality [J]. Journal of Clinical Epidemiology, 2004, 57(12): 1288–1294.

[44] Chen L, Zeng W M, Cai Y D, et al. Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities [J]. PloS One, 2012, 7(4): e35254.

[45] Stearns M Q, Price C, Spackman K A, et al. SNOMED clinical terms: Overview of the development process and project status [C]// Proceedings of American Medical Informatics Association Annual Symposium. Washington DC: American Medical Informatics Association, 2001: 662–666.

[46] Lipscomb C E. Medical subject headings (MeSH) [J]. Bulletin of the Medical Library Association, 2000, 88(3): 265–270.

[47] Suriya S, Nivetha S. Design of UML diagrams for WEBMED-healthcare service system services [J]. EAI Endorsed Transactions on e-Learning, 2023, 8(1): e5.

[48] Sherwani J, Ali N, Mirza S, et al. Healthline: Speech-based access to health information by low-literate users [C]. Bangalore: International Conference on Information and Communication Technologies and Development, 2007.

[49] Ayers J W, Althouse B M, Allem J P, et al. Seasonality in seeking mental health information on google [J]. American Journal of Preventive Medicine, 2013, 44(5): 520–525.

[50] Barnett G O, Cimino J J, Hupp J A, et al. DXplain: An evolving diagnostic decision-support system [J]. The Journal of the American Medical Association, 1987, 258(1): 67–74.

[51] Chatfield A J. Lexicomp online and micromedex 2.0 [J]. Journal of the Medical Library Association, 2015, 103(2): 112–113.

[52] Hey T, Tansley S, Tolle K, et al. The fourth paradigm: Data-intensive scientific discovery [M]. Mountain View: Microsoft Research, 2009.

[53] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583–589.

[54] 李国杰. 关于人工智能的若干认识问题 [J]. 中国计算机学会通讯, 2021, 17(7): 44–50.
Li G J. Some issues on understanding artificial intelligence [J]. Communications of the CCF, 2021, 17(7): 44–50.

[55] Pan Y H. Miniaturized five fundamental issues about visual knowledge [J]. Frontiers of Information Technology & Electronic Engineering, 2021, 22(5): 615–618.

[56] Anderson J R, Crawford J. Cognitive psychology and its implications [M]. San Francisco: WH Freeman, 1980.

[57] Ilievski F, Oltramari A, Ma K, et al. Dimensions of commonsense knowledge [J]. Knowledge-Based Systems, 2021, 229(11): 107347.

[58] Speer R, Chin J, Havasi C. Conceptnet 5.5: An open multilingual graph of general knowledge [C]. San Francisco: The AAAI Conference on Artificial Intelligence, 2017.

[59] Sap M, Le Bras R, Allaway E, et al. Atomic: An atlas of machine commonsense for if-then reasoning [C]. Hawaii: The AAAI Conference on Artificial Intelligence, 2019.

[60] Miller G A. WordNet: An electronic lexical database [M]. Cambridge: Massachusetts Institute of Technology Press, 1998.

[61] Roget P M. Roget's Thesaurus of English words and phrases [M]. New York: Thomas Y. Crowell Company, 1911.

[62] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [C]. Minneapoli: The North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[63] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [C]//Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: Massachusetts Institute of Technology Press, 2020: 1877–1901.

[64] Nieh E H, Schottdorf M, Freeman N W, et al. Geometry of abstract learned knowledge in the hippocampus [J]. Nature, 2021, 595(7865): 80–84.

[65] Xie Y, Hu P Y, Li J R, et al. Geometry of sequence working memory in macaque prefrontal cortex [J]. Science, 2022, 375(6581): 632–639.

[66] Pearl J, Mackenzie D. The book of why: The new science of cause and effect [M]. New York: Basic Books, 2018.

[67] Pearl J. Causality [M]. Cambridge: Cambridge University Press, 2009.