面向大数据领域的事理认知图谱构建与推断分析

王军平^{1*}, 张文生¹, 王勇飞², 孙正雅¹

1. 复杂系统管理与控制国家重点实验室 (中国科学院自动化研究所), 北京 100190

2. 国电大渡河流域水电开发有限公司, 成都 610000

* 通信作者. E-mail: junping.wang@ia.ac.cn

收稿日期: 2019-12-04; 修回日期: 2020-02-04; 接受日期: 2020-04-28; 网络出版日期: 2020-07-13

科技创新 2030—“新一代人工智能”重大项目 (批准号: 2018AAA0102100) 和国家自然科学基金 (批准号: 61772525, 61876183, U1636220) 资助项目

摘要 随着大数据对全球生产、流通、分配、消费等国计民生领域产生重要影响, 事理作为认知智能重要概念, 不仅帮助人们发现大数据所蕴含的国计民生的线索与发展规律, 而且帮助人们更好认知人机物三元世界未来发展趋势. 考虑到人机物事理动态演化是工业大数据有别于其他行业大数据处理的本质区别, 本文从人机物事理动态演化特性智能认知着手, 结合传统知识图谱在互联网领域取得的成功经验, 提出了面向大数据领域的事理认知图谱构建与推断分析. 首先论述了事理认知图谱对推动认知智能研究发展的贡献, 剖析了它与知识图谱异同之处, 深度理解人类社会发展变化规律的重要价值. 其次阐述了事理抽取与泛化、多模态联合网络化合成表示、进化认知和推断分析等关键技术研究最新进展以及面临的挑战. 最后, 结合我们课题组在事理认知图谱的研究进展情况, 归纳总结了事理认知图谱在预防网络电信诈骗和安全生产管控等领域最新应用效果. 本文结尾总结和展望事理认知图谱的未来研究方向和发展前景.

关键词 工业大数据, 知识图谱, 事理认知, 人机行为演化分析

1 引言

随着信息技术与实体经济的深度融合引发的数据迅猛增长, 美国白宫率先发布了《美国白宫大数据白皮书》, 随后欧洲各国启动大数据研究计划, 我国也颁布了《促进大数据发展行动纲要》等战略举措, 推动大数据分析技术在生产、流通、分配、消费活动以及经济运行等重要环节的应用. 例如: 美国 GE 通用公司通过工业互联网技术实现对全球飞机发动机、发电装备的在线远程检测, 开启了全球制造业数字化升级转型. 截至 2018 年, 发达国家的 GDP 数字经济占比已经达到了 34.8%^[1]. 预示着大数据已成为众多产业的基础性战略资源, 挖掘与释放大数据价值成为新一代人工智能研究的重要方

引用格式: 王军平, 张文生, 王勇飞, 等. 面向大数据领域的事理认知图谱构建与推断分析. 中国科学: 信息科学, 2020, 50: 988–1002, doi: 10.1360/SSI-2019-0273
Wang J P, Zhang W S, Wang Y F, et al. Constructing and inferring event logic cognitive graph in the field of big data (in Chinese). Sci Sin Inform, 2020, 50: 988–1002, doi: 10.1360/SSI-2019-0273

向,归根结底大数据蕴含着推进产业供给侧改革、发展数字经济的重要决策线索,也就意味着谁率先掌握人类社会活动的基本规律数据(包括个体兴趣偏好、群体消费趋势、关系和行为分析等),谁就可以在市场营销、商业计划、社会规划、经济建设中做到科学施策,合理优化产业资源,重塑产业价值链,培育数字经济发展模式^[2]。

大数据技术与实体经济深度,催生了企业数字化升级转型,积累了规模巨大、来源分散、模态多样性的人机物运行机理数据资源。这类大数据特性使得现有机器学习、数据挖掘和数据仓库等技术难以胜任海量数据组织管理任务。Google 的知识图谱(knowledge graph)通过三元组对概念实体、属性和关系等核心要素进行形式化地描述,以最小的代价将大规模网络信息有效组织起来,开辟了互联网领域的数亿实体、属性和关联关系的可视化存储、组织和管理。知识图谱在网络搜索(Freebase, Dbpedia, Wikidata, YAGO)中广泛应用,满足了人们快速获取所需信息和全面掌握潜在价值的迫切需求。缺乏深度剖析人机物运行机理数据蕴含决策价值的能力,主要聚焦在:(1)知识图谱仅聚焦实体概念及关系组织管理,缺少洞察产业运行机理演化规律、资源合理配置和风险传播路径机制;(2)现有产业大数据以多源、多学科和多类型非结构化的形式存在、蕴藏人机物运行机理演化知识,使得知识图谱难以在有限时间内提取准确有效的决策知识,清晰地展示产业未来发展方向。

针对知识图谱面临困境,先后诞生了谷歌大脑、IBM Watson、微软 Azure、阿里大脑、百度大脑等,开启一种全新的海量数据智能计算模式,以其卓越处理结构化与非结构化数据和智能化人机交互能力,掀起了全球学术界和产业界研究高潮。Liu^[3]提出事理(event logic)概念,并采用有向图模型形式化表征人机物行为演化机理,推进事理概念与金融产业深度融合。伴随 AlphaGo 在围棋领域战胜人类,工业机器人、情感陪护机器人等问世,充分证明事理认知是人工智能从弱智能走向强智能的必有之路^[4]。本文总结了国内外已取得与事理认知密切相关的研究成果,从事理认知计算角度,提出了事理认知图谱,全面阐述了跟事理逻辑表示、认知学习、推断分析等密切相关技术的最新进展。结合我们团队在事理认知技术及应用的研究进展,梳理事理图谱在安全生成管控、预防诈骗等领域应用的推广途径。最后对事理认知图谱未来发展的挑战和趋势进行总结与展望。

2 事理认知图谱

2.1 基本概念

谷歌大脑、IBM Watson、微软 Azure、DeepMind AlphaGo、阿里大脑、百度大脑等问世,表明事理^[3]是人类识别、推理和理解社会发展的重要概念。他们以卓越处理结构化与非结构化的数据能力,帮助企业掌握人机物要素行为机理演化规律,洞察行业发展趋势,应对内外部风险挑战,深受社会各行各业的青睐。本文结合现有研究成果进行比较和提炼,从人工智能的认知计算角度,提出了事理认知图谱,采用五元组〈事件、状态、行为、时间、关联影响〉概率图模型,形式化描述现实世界中人机物几何结构、状态和行为等事理(如图1)^[5]。对事理逻辑进行推理、分析、对比、归纳、总结和论证等加工处理,就像人类大脑一样通过视、听、触、理解等多种感知通道把外界信息转换为内部判断机理模型一样,使得事理认知图谱具备高动态、高维度、多模式分布式大场景人机物事理演化的感知、智能描述和跨媒体推理分析等能力。

事理认知图谱与知识图谱相比,它是人工智能 2.0 时代崭新研究领域,从认知智能角度出发,以理解现实世界人机物的行为演化机理为研究对象,重点刻画事件之间的顺承关系、因果关系、条件关系和上下位等事理逻辑,真实刻画企业价值链的运行规律和模式。具体区别如下:

图 1 (网络版彩图) 事理认知图谱概念、核心技术及应用生态
Figure 1 (Color online) The event cognitive vault concepts, key techniques and applications

- (1) 事理认知图谱从现实世界人机物活动状态、行为、事件、场景等多模态数据中自动抽取事理逻辑知识, 以刻画事理演化规律为研究对象, 而知识图谱以从大规模结构化数据抽取实体概念及关系为研究对象, 以图形式化方式分类整理成结构化知识, 方便人们检索各种类型知识。
- (2) 事理认知图谱的边表示事理逻辑规则, 即顺承关系、因果关系、条件关系和上下位等, 边上标注概率信息、逻辑规则概率分布, 而知识图谱的边表示实体间确实存在链接关系, 也就是说某一条属性或关系要么成立, 要么不成立。

2.2 基本组成结构

事理认知图谱基本结构 (如图 2), 按照从底向上思路, 划分为异构数据层、事理表示层、推理计算层和智能服务层等 4 部分, 具体如下:

- (1) 异构数据层, 主要汇聚了现实世界人机物活动过程中产生的状态、行为、事件、场景等多模态海量异构数据。这类数据以结构化、半结构化和非结构化等形态存在, 完整提供了现实世界中数以亿计人机物的所言所行多模态数据。
- (2) 事理表示层, 采用网络表示方法实现对数以亿计人机物的所言所行多模态异构数据跨媒体统一表征, 提取数以亿计人机物事理、特征及逻辑关系, 在信息空间构建数以亿计人机物实体、状态和行为等演化语义网络, 能够支持对数以亿计人机物事理节点删减、新增等动态变化, 对节点间时空关联关系动态变化等进行动态存储、组织和管理。
- (3) 推理计算层, 采用利用统计分析和图神经网络的逻辑推理技术, 不断优化事理表示层时空演化网络的结构和参数, 按照用户意图从上下文情景中获取不同事理主题之间各种各样复杂逻辑网络结构, 并采用跨模态信息完整性分析技术, 解决事理图谱补全和事理图谱去噪等问题, 进而使事理图谱

图 2 (网络版彩图) 事理认知图谱分层结构

Figure 2 (Color online) The hierarchical structure of event cognitive vault

更加丰富和完善, 形成实时计算大脑系统.

(4) 智能服务层, 主要面向行业提供智能问答、智能搜索、舆情分析、用户画像、智能设计、智能诊断、智能生产、反欺诈等智能服务工具, 让机器具备人类认知事件发展规律能力, 从杂乱无章多模态数据中, 获取人类认知世界的常识, 掌握人类发展规律, 同时让人类借理事理演化图谱更好地从客观世界中去挖掘、获取和沉淀知识, 这些知识和人工智能系统形成正循环, 做到人类与机器自然交互途径, 构建共融共生的混合增强智能生态.

3 事理认知图谱自动构建技术

事理认知图谱自动构建技术, 以多模态数据 (包括文本、音视频、图片、文字、消息等) 为源头, 按照事理认知图谱的 4 层基本结构体系, 重点归纳总结事理抽取、多模态逻辑关系融合表示、事理进化认知推理等自动构建关键技术研究进展, 详细阐述如何对现实世界蕴含的事理、属性特征和逻辑关系的自动抽取与泛化, 生成各种事理及逻辑规则, 并消除不同情境的逻辑概念的歧义, 剔除冗余和误差等, 形成有向图形式的事理知识库.

3.1 事理及规则自动抽取

事理及逻辑关系自动抽取是构建事理认知图谱最重要的步骤, 从多模态、高维度、高噪声、多样性等特性的数据层, 自动抽取各种类型的事理概念、属性和逻辑关系 (顺承、转折、因果、递进、关联、解释等), 并对它们进行抽象与泛化. 这部分研究工作重点聚焦在主题事件及属性自动抽取和逻辑关系

自动抽取等主题上。

主题事件及属性自动抽取。最初, Liu 等^[6]采用启发式算法与人工编写规则相结合的方法从文本、语音和视频中自动抽取文本关键词实体。然而, 基于规则的方法具有明显的局限性, 不仅需要耗费大量人力, 而且可扩展性较差, 难以适应上下文数据的变化。随后, 人们开始尝试采用统计机器学习方法辅助自然语言处理方法, 解决命名实体抽取问题, Liu 等^[7]利用 K-最近邻算法和条件随机场模型, 实现了对 Twitter 文本、语音和视频数据中包含实体的识别。然而迄今为止, 单纯基于有监督学习的实体抽取方法, 在准确率和召回率上的表现都不够理想, 且算法的性能依赖于训练样本的规模, 对此类方法的发展形成了制约。随着深度学习技术的发展, Rawat 等^[8]提出了一种条件 Wasserstein 生成式对抗网络模型 (conditional Wasserstein generative adversarial networks, CWGAN), 该模型借鉴条件生成式对抗网络以文本描述为条件的图像概率分布的思想, 完成命名实体识别以句子序列为条件获得标注序列概率分布的任务。该模型的生成器和判别器都采用双向 LSTM 结构, 不同的是生成器生成命名实体标签的概率分布, 判别器则为生成器的生成质量打分并反馈给生成器, 生成器根据反馈更新梯度从而提升生成标签概率的质量。另外, CWGAN 采用梯度惩罚的方法来保证梯度在后向传播的过程中保持平稳。实验表明 CWGAN 方法在命名实体识别的事理抽取的准确率和召回率上最优。构建跨越自然语言理解与计算视角相融合的智能描述与生产方法体系, 实现从图像、文本、视频和语音等多模态数据从单一抽取向多模态并行自动抽取的转变。

逻辑关系自动抽取就是从数以亿计人机物事理节点全概率分布环境中可视化探索出事理之间的逻辑关系, 例如, 顺承、转折、因果、递进、关联、解释等。当前研究成果主要集中在事件因果关系抽取上, Zhou 等^[9]提出采用多实例多标签 (multi-instance multi-label) 方法对关系抽取进行建模, 刻画一个实体对可能存在多种关系的情况。Socher 等^[10]提出使用递归神经网络解决关系抽取问题。该方法首先对句子进行句法解析, 然后为句法树上的每个节点学习向量表示。通过递归神经网络, 可以从句法树最低端的词向量开始, 按照句子的句法结构迭代合并, 最终得到该句子的向量表示, 并用于关系分类。该方法能够有效地考虑句子的句法结构信息, 但同时该方法无法很好地考虑两个实体在句子中的位置和语义信息。Zeng 等^[11]提出用卷积神经网络 (convolutional neural network, CNN) 进行关系抽取, 采用词汇向量和词的位置向量作为卷积神经网络的输入, 通过卷积层、池化层和非线性层得到句子表示。实体的位置向量和其他相关的词汇特征使得句子中的实体信息能够被较好地考虑到关系抽取中。

但是由于 CNN 不适合学习长距离的语义信息, Zhang 等^[12]使用双向循环神经网络抽取文本中的关系, 虽然该模型能利用上下文信息, 但是上下文的范围受梯度消散问题的影响, 难以取得预期效果。为此, Miwa 等^[13]提出了一种基于端到端神经网络的关系抽取模型。该模型使用双向长短时记忆模型 (long-short term memory, LSTM) 和树形 LSTM 同时对实体和句子进行建模, 在数据集 SemEval-2010Task8 上取得了最好的效果, 与之前的模型相比, 该方法效果取得较大幅度提升。同时, 考虑到上下文信息以及训练速度。但多数现有的因果关系抽取方法都需要人工定义模式和约束, 且严重依赖知识库。Feng 等^[14]利用生成式对抗网络 (generative adversarial network, GAN) 的对抗学习特性, 将带注意力机制的双向门控循环单元神经网络 (bidirectional gated recurrent units network, BGRU) 与对抗学习相融合, 通过重定义生成模型和判别模型, 基本的因果关系抽取网络能够与判别网络形成对抗, 进而从因果关系解释信息中获得高区分度的特征。事实上, 事理逻辑关系蕴藏于不同模态和类型的信息源中。我们需要探索如何利用多语言文本、图像和音频信息进行关系抽取。事理顺承关系抽取受语料标注的限制, 研究进展相当缓慢。

图 3 (网络版彩图) 多模态联合网络化合成表示技术路线

Figure 3 (Color online) The multi-mode joint network synthesis representation technology map

3.2 多模态逻辑网络化合成表示

随着数字设备的普及,特别是以智能平板电脑、手机为代表的便携式设备的普及,人们更加方便地通过这些设备拍摄照片,这些记录信息常常不是由单一类型的数据组成的,而更多的是由文本、音频、图像、视频等多模态形式的数据构成。我们通过事理抽取与泛化方法从海量多源异构的文本、图像、视频等多模态数据中,提取以亿计人物事理节点及逻辑关联关系。由于这些事理节点及关联关系以碎片化、独立分布、噪声大、信息不完备等形式存在,我们采用多模态联合网络化合成表示(如图3)将这些海量不同模态事理节点及关联关系转换成计算机可识别和处理的内容,并进行统一表征描述、组织、管理、搜索、分析等,形成统一的特征描述,支撑事理智能描述与处理。主流方法是网络表示学习,划分为网络节点嵌入表示、事理属性嵌入表示和多模态异构网络合成。

当表示机器学习推动了新一代人工智能在网络信息自组织领域广泛应用,网络节点嵌入表示算法将网络信息转化为低维稠密的实数向量,并将其用作已有的机器学习算法的输入。最典型例子是DeepWalk^[15],它将事理视为概念并生成短随机游走,从而弥补网络嵌入和单词嵌入之间的差距。然后,可以将诸如Skip-gram之类的神经语言模型应用于这些随机游走以获得网络嵌入。DeepWalk的优点可以概括为:首先按需生成随机游走,对每个事理及逻辑样本进行了优化;然后随机游走和Skip-gram的组合使DeepWalk成为在线算法;最后DeepWalk是可扩展的,生成随机游走和优化Skip-gram模型的过程都是高效且平凡的并行化。最重要的是,DeepWalk引入了深度学习图形的范例。Node2vec是DeepWalk的扩展,它引入了一个偏向的随机步程序,结合了广度优先搜索(breadth first search, BFS)风格和深度优先搜索(deep first search, DFS)风格的邻域探索,然而计算训练复杂度越来越复杂。Petar等^[16]提出了一种图注意力网络(graph attention networks, GAT)模型,使用多尺度网络表示方法,能够从原始样本中,不预先确定超参数来控制上下文节点分布,而是自动学习对图转换矩阵的幂集的事理逻辑网络,然而很难探索节点样本探索。SDNE^[17]方法通过深度自动编码器保持两跳邻居之间的接近度。它通过最小化它们的表示之间的欧几里德距离,进一步保持相邻节点之间的接近度。他们采用随机冲浪策略捕获图形结构信息,将这些结构信息转换为PPMI矩阵,并训练堆叠去噪自动编码器(stacked denoising auto encoder, SDAE)以嵌入节点。

为了模拟不同类型的节点和边缘,大多数异构网络嵌入方法通过联合最小化每种模态的损失学习

图 4 (网络版彩图) 事理情境进化认知技术路线

Figure 4 (Color online) The event logical situational evolutionary cognition technology map

节点嵌入。这些方法要么直接在相同的潜在空间中学习所有节点嵌入, 要么事先为每个模态构建嵌入, 然后将它们映射到相同的潜在空间。Chang 等^[18]提出了异构网络的深度嵌入框架, 首先为每种模态(如图像、文本)构建一个特征表示, 然后将不同模态的嵌入映射到同一个嵌入空间。优化目标是最大化链接节点的嵌入之间的相似性, 以及最小化未链接节点的嵌入, 可实现边在相同模态内的两个节点之间以及来自不同模态的节点之间。HEBE^[19]是一种嵌入大规模异构事件网络的算法, 其中事件被定义为网络中一组节点(可能是不同类型)之间的交互。虽然先前的工作将事件分解为事件中涉及的每对节点之间的成对交互, 但 HEBE 将整个事件视为超边界并同时保留所有参与节点之间的接近度。EOE^[20]是用于耦合异构网络的网络嵌入方法, 其中两个同构网络与网络间边缘连接。EOE 学习两个网络的潜在节点表示, 并利用和谐嵌入矩阵将不同网络的表示转换为相同的空间。为了构建随机漫游, Metapath2vec 使用基于元路径的漫游来捕获不同类型节点之间的关系, 在模型优化期间考虑节点类型信息。

3.3 事理规则认知学习

事理情景认知推理(如图 4)是指面向多模态统一表征网络, 通过不确定性推理与决策、深度强化迁移学习、主动学习等认知推理方法, 充分挖掘事理异构网络模型之间的多层次语义关联、时空关联等, 实现事理属性及逻辑校正、消除概念的歧义、剔除冗余和错误概念, 从而实现基于明确规则与特定领域的“计算智能”, 到语音、图像、视频识别解释的“感知智能”, 再到具备理解、推理和解释的“认知智能”, 使得计算机自然表达与人类认知理解基本统一, 进而使得机器与人类的自然交互成为现实。为事理认知图谱广泛应用到客户画像、精准分析、智慧搜索、智能推荐、智能解释、自然人机交互等领

域提供智能模型,以解放人类脑力,显著提高机器生产力.这部分研究工作主要分为3类.

(1) 基于规则模板的主题事件归纳方法.由于相同事理逻辑来源于多个不同文本、音频、视频和图像等渠道,造成了包含事理、属性和逻辑关系在不同场景的冗余问题,当前研究主要采用语义模板生成一个完整事理逻辑关系的标准模板,建立不同来源事理关系的标准共享空间,然后将所有数据投影到该空间.比如早期的CCA方法,通过线性函数尽可能地将成对出现的图像和文本数据投影到共享空间的同一位置.该方法的缺点是,只分析了不同组变量间每个维度 X 与每个维度 Y 之间的关系,而忽略了 X 与 Y 这两组变量内部的关系.RSKCCA算法^[21]是一种事理逻辑关系统计分析工具,能特征化两组多维变量的非线性关系.可以提高原始向量空间的维度,并增加其灵活性.常用的核函数包括:线性核函数、 Q 阶多项式核函数、高斯径向基础函数(radial basis function, RBF)核函数和多层感知器(multi-layer perceptron, MLP)核函数.

(2) 基于关系迁移的不同情景逻辑规则认知方法.随着人机物事件之间在时间、空间上相继发生的态势变化,上下文情景事理逻辑关系出现不确定性演化,由于先验数据不足,难以有效认知未来事理逻辑演化趋势.Mihalkova等^[22]提出了一种关系知识迁移马尔科夫(Markov)网络算法,该算法将关系知识通过马尔科夫逻辑网络跨关系域迁移,将实体及其关系从源域连接到目标域,比方说生物病毒传播规律到计算机病毒传播规律的迁移.然而难以处理非欧几里德数据空间情景演化,Parisotto等^[23]提出了一种基于行为模拟(actor-mimic)的多关系深度迁移方法,它可以基于一组相关的源任务预训练一个深度策略网络,通过监督信号的指导,使得单一的策略网络学会各自任务相对应的策略,并将学习到的知识迁移到相似的新任务中.Rusu等^[24]提出了一种基于渐进式神经网络的迁移深度强化学习方法,解决从仿真环境中迁移知识到真实环境的难题.Silver等^[25]提出了一种同时泛化状态和目标空间的通用值函数逼近器,实现了已学习到的关系知识迁移到那些未知不确定性的动态演化环境.但随着目标不同的新任务规模增长,算法性能不够理想.Wang等^[26]提出基于共享子空间的多任务在线学习与快速决策方法,有效解决了在大规模连续状态-行为空间的多任务学习泛化和快速准确决策问题.

(3) 基于CNN-RNN的规则智能描述方法.基本思想是利用CNN提取文本、语音和图像在特定场景下的事理特征编码,然后利用RNN作为解码器,将文本、语音和图像逐句生成对应的顺承、转折、因果、递进、关联、解释等智能描述.Wang等^[26]提出了一种多模RNN模型,首先使用CNN提取图像的顺承、转折、因果、递进、关联、解释等特征,并形成事理逻辑嵌入向量,然后将文本、视频和图像等多模态关联逻辑嵌入向量,形成多模态智能要素表达.Xu等^[27]提出了自然语言处理的注意力机制,将注意力机制应用到多模态事理网络认知建模,使得在生产相关单词时将人的注意力更多集中在对应的物体上.另外最近深度强化学习被提出后,成为自动描述文本、视频和图像等多模态分层关联网络的重要方法.在视频、图像、音频和文本等内容的跨模态描述方面,Karpathy等^[28]提出了使用基于区域CNN的概念实体及属性抽取和双向RNN技术,考虑到人类对外界信息的认知依靠不同感官信息在线自动合成,已达到整体理解外部环境目标,本文借鉴人类认知机理,对机器认知模型信息合成和全局泛化理解能力进行改进,实现了机器认知从一般化到针对特定人特定情景的个性化描述转变,学习机制从单纯的数据驱动向数据与知识交互驱动.但是当前方法主要集中在图像语义智能解析与理解,Zhou等^[29]提出了一个基于协同学习框架的孪生神经网络模型,来优化和补充地理兴趣点(point of interest)的标签生成方法,解决不同数据类型的数据源的协同学习、关联认知,从而实现了跨模态的知识认知、表达和知识联合交互建模.

4 规则演化推断分析技术

事理演化推断分析是面向特征场景,生成各种客户画像、精准分析、智慧搜索、智能推荐、智能解释、自然人机交互、深层关系推理等模型的重要工具,发掘和释放特定行业大数据的潜在价值的根本途径,研究面向特定场景的事理模型推断分析技术是当前事理认知图谱领域的重要任务。

4.1 基于统计学习的演化推断分析

大多数事理应用所解决的问题都可以形式化地表述为概率模型中的推断问题。在推断问题中,针对实际应用事理模型随机变量集合 S 中变量的取值分布,采用生成式模型或判别式模型为问题建模,并运用针对具体模型的推断算法计算事理模型概率分布情况。常用的基于团树的精确推理,能够计算查询变量的边缘分布或条件分布的精确值,但其计算复杂度随着图模型的树宽呈指数增长,因此适用范围有限。David 等^[30]提出了相关主题模型 (correlated topic model)、层次贝叶斯逻辑斯蒂回归模型 (Bayesian logistic regression model) 等非共轭模型,以及基于坐标上升优化迭代过程的拉普拉斯 (Laplace) 变分推理和 delta 变分推理方法。Ranganath 等^[31]提出黑盒变分推理 (black box variational inference) 算法,把目标函数的导数作为变分近似的期望,将仅适用于特定模型上的一般变分推理算法进行了推广,从变分近似抽样得到梯度值,用以更新模型的参数值,使变分推理算法适用于更广泛的领域。Sun 等^[32]针对噪声聚合数据,诸如列联表数据或计数数据上的集合图模型 (collective graphical models),提出期望最大近似边际推理算法。Bernhard 等^[33]对图模型联合分布的析取范式 (disjunctive normal form) 作为查询条件的近似推理问题,提出了近似边际推理 (approximate marginal inference) 算法,从而推广了仅局限于简单边际查询的经典推理技术。在生成式模型中,我们已知包含感兴趣变量集合在内的一些相互联系的变量的联合分布,以及其中可观测变量的观测值 (或真实值),目标是以可观测变量为条件计算目标变量的条件概率。在判别式模型中,我们已知包含感兴趣变量集合 S 在内的一些相互联系的变量与另一些可观测变量之间的联系,即以可观测变量为条件的条件分布,以及可观测变量的观测值,目标同样是计算感兴趣变量集合 S 中的变量的条件概率。

4.2 基于图神经网络的演化推断分析

图神经网络 (graph neural networks, GNN)^[34]开启了深度学习在非欧几里德数据上的处理能力,在社会科学、社交网络、知识图谱、推荐系统、生命科学等领域得到了广泛应用。GNN 不仅对动态知识演化特性建模能力表现出优越特性,而且以其令人信服的性能和较高的可解释性,已成为事理图谱逻辑关系推断分析的重要研究方向。Levie 等^[35]在深入分析了 GFT 模型后,提出了序列图神经网络模型,实现了多模态异构事理模型多粒度融合与演化学习推理机制,有效降低了由于虚假关联带来的推理预测虚警问题,对于演化学习问题,将形式化推理转化成时序学习或者跨域学习问题。DeepMind 研究团队^[25]将图神经网络与强化学习相结合应用于围棋机器人,升级 AlphaGo 的事理认知系统,推断出尚未被人类棋手发现的知识,在各种人机对抗当中取得了胜利。为了控制泛化风险及推理错误,Joshua 等^[36]提出了图马尔科夫神经网络 (graph Markov neural networks, GMNN),结合了关系马尔科夫网络的预测和图神经网络强大特征表示二者的优势。在网络模型 E-step 推理过程中,关系马尔科夫网络有效求解了对象标签的最大似然后验分布;在 M-step 特征学习中,使用图神经网络对局部标签依赖关系进行建模,二者协作 N 次迭代优化事理逻辑推理过程中参数约简和先验逻辑共享问题。GMNN 模型被广泛应用到大规模图模型邻接关系预测,然而现实世界中多模态数据具有多样性和动态演化特性,仅依靠文本推理很难充分挖掘多模态事理蕴含的语义信息。Michelle 等^[37]提出了多模态

图 5 (网络版彩图) 基于事理认知图谱的预防网络诈骗应用

Figure 5 (Color online) The event cognitive vault-based prevent network fraud system

空间演化推理学习算法,采用图神经网络表示学习方法,按照时间序列对齐的多感官异构多模态数据,将不同视角感知输入编码为一个紧凑联合表示函数,生成了多视角图神经网络模型,基于该模型进行联合推理,丰富接触的操作任务的价值,增强了多模态事理自我监督推断训练能力,消除了手工注释的限制.所取得成果表明图神经网络是跨媒体事理演化推断的重要途径.Zhu 等^[38]提出了基于强化学习的因果关系发现算法,将强化学习应用到多模态变量因果评价打分法中,通过基于自注意力机制的自编码图神经网络探索多模态数据之间的关系,结合因果结构的条件,并使用策略梯度的强化学习算法对多层次自编码神经网络参数进行训练,最终得到多变量因果图结构.所取得研究成果为认知事理之间的顺承、因果、条件和上下位等关系开辟新的研究思路.简而言之,上述成果为研究多模态事理逻辑推断分析方法体系开辟新的研究思路,然而探索揭示现实世界的事理演化规律推断分析方法是人工智能新的研究方向.

5 事理认知图谱的前沿应用

事理认知图谱作为人工智能时代崭新的研究领域,其核心技术及应用模式尚处于探索阶段.本节考虑如何向学术界和产业界展现事理认知图谱对实体经济产生重要影响,同时结合国内外对事理认知应用的思考,结合我们团队在事理认知图谱理论及应用研究的最新进展,选取预防网络诈骗和安全生产管控系统作为事理认知图谱的前沿应用范例,帮助大家深刻理解事理认知图谱行业应用价值.

5.1 基于事理认知图谱的预防网络诈骗系统

近些年来互联网金融蓬勃发展,特别是移动互联网和智能移动终端的兴起,使得传统商业线上交易模式得以实现.这种模式出现给大众带来快速便捷的金融服务,同时伴随大量个人信息泄露问题日趋严重,引发网络电话诈骗事件层出不穷.传统的反欺诈信息手段主要依赖于人工审核信息来源真实性,而现有身份证、手机号码、银行流水等材料的伪造成本非常低,反诈骗机构难以有效获取网络电话诈骗信息传播规律,难以准确识别诈骗线索和手段演化路径.本案例结合我们团队在事理认知图谱研究的最新进展,选取现有已发生电话诈骗 100 万条左右文本、音频等信息为训练数据集,采用事理特征提取、网络化合成表示、归纳推理分析、误差评估等事理图谱构建技术,识别出覆盖校园贷、网络贷款、冒名办卡透支欠款、冒充公检法机关处理违法事件、帐户钱款不翼而飞、个人名誉无端受损等常见网络电话诈骗信息传播行为特征,形成网络电话诈骗事理认知图谱(如图 5).选取中国某运营

图 6 (网络版彩图) 基于事理认知图谱的全生产隐患风险主动预警系统

Figure 6 (Color online) The event cognitive vault-based active warning system for all production risks

商 10 万级用户群在线交互信息作为测试数据, 采用事理相关性推断分析、博弈演化评估等技术, 代替人工审核, 进而总结出各种网络电话诈骗手段运作规律, 从历史序列值和当前序列值中合理推断出未来发生可能的概率, 找出欺诈者的蛛丝马迹, 挖掘其数据的矛盾点和可疑点, 取得了实时准确识别出 95000 条电话诈骗信息, 及时阻止受害财产损失的效果. 将网络电话诈骗事件消灭在萌芽状态, 帮助运营商提升事前主动预防网络电话诈骗发生.

5.2 基于事理认知图谱的安全生产管控系统

现有常规性预警技术受安全生产非常规态的前兆特征真假难辨、征兆隐蔽、波动性强等干扰, 造成了预警滞后困局. 本案例以我们团队和国家能源集团某水电开发公司联合承担项目为背景, 选取近五年已发生的人员、设备、大坝、地灾、防汛、环境污染等事故历史数据作为训练样本, 采用卷积神经网络 CNN、双向 LSTM 和双向循环等深度学习, 从文本、视频和图像等多模态的事故训练数据集中, 萃取了人员事故、设备事故、大坝事故、地质危害、汛期损害、环境侵蚀等时空特征机理. 复盘已发生的人员事故、设备事故、大坝事故、地质危害、汛期损害、环境侵蚀等历史情景, 对历史数据进行因素分析、模式识别、误差评估, 总结客观规律和模式, 发现重要的线索与模式, 建立梯级水电站群安全预警事理认知图谱 (如图 6). 采用因果关联推断分析、归纳推理分析、参数估计、假设检验等事理推断分析技术, 构建特定场景发生地质灾害、大坝变形、大型设备运行状态、人员技术事故、隐患点突变、水质污染等事故预警规则库, 结合产生人员、设备、大坝、地质、防汛、环境等运行状态实时数据, 预测未来安全态势发展趋势. 形成基于事理认知图谱的全生产隐患风险主动预警系统, 主动发现潜在安全事件线索, 挖掘其安全线索的矛盾点和可疑点, 实施后取得了人员误操作毫秒级实时预警, 设备故障提前 10 天预警, 地质灾害、环境损害等识别准确率提升到 97% 左右的效果.

6 总结与展望

事理认知图谱是人工智能领域崭新的研究领域, 它是网络、工业、交通和医疗等领域构建从多模态大数据到有价值知识转化的重要载体, 通过事理认知图谱源源不断地给行业提供实体、场景、事件

和主题等关联有价值知识,完成从知识到决策的计算过程.事理认知图谱完美解决了人工智能领域从数据到决策的技术瓶颈.因此,开展面向大数据领域的事理认知图谱构建与推断分析研究,是新一代人工智能的重要前沿研究课题,具有重要的学术意义和广泛的应用价值.

事理认知图谱构建与推断分析研究虽然已经取得了丰富的成果,但是在以下方面仍然面临着巨大的挑战.

(1) 事理模式感知与理解.以文本、音视频、图片、文字、消息等多模态为载体的大数据已经成为电信网络、工业、交通等行业应用数据主流形态,那么这些行业大数据还包含着高维分类标注、多样性内容标签、格式编码复杂、内容集成、流化处理、隐私安全保护等诸多高噪声信息.这些数据除了数据量巨大以外,还具有多源异构、语义时变、时效性强的特性.这些特性使得对现有处理高噪声多模态数据方法难以胜任,造成非结构化数据识别很难有效应用推广,需要探讨新的事件语义特征提取与编码方法.

(2) 多模态事理演化统一表示.现有电信网络、工业、交通等行业大数据随着时间变化数据节点模态及关联结构不断动态变化,使得现有网络表示学习方法,如: Network Embedding, Graph Embedding 和 DeepWalk 等,造成了“信息多但用不了,有信息但找不到”的局面.特别地,产业大数据具有模式多样、内容难于理解、关联复杂多变、价值难以有效识别、质量良莠不齐、真伪难以判定等复杂特性,使得现有度量并刻画数据复杂性和厘清内在因果逻辑等技术难以应对.

(3) 跨模态语义理解与挖掘.事理认知图谱通过跨媒体数字孪生、演化表示等方法,快速构建了网络化、数字化、虚拟化的电信网络、工业、交通环境,给人们带来了前所未有的信息自主权,使得人类社会的信息化水平进入了一个全新阶段.通过分析事理演化规律给人们极大认知现实世界事理发展规律、社会发展不断创新、产业升级改造等带来前所未有的进步,我们沉浸在这种喜悦环境的同时,也带来了信息产业社会化、信息内容碎片化和信息传播网络化的问题,给现有基于策略推理、近似推理、跨模态关联分析等语义理解方法带来前所未有的挑战.例如大数据中的群体行为模式尚未得到深刻理解和充分掌握,导致产业发展信息的可信性、传播的可预测性、群体行为的可控性等方面仍处于一种无序状态,造成人们创造大量产业发展数据却对其知之甚少的现状.在工业大数据领域,如果我们不深入分析机器参数演化及群体行为的原动力和本质特征,那我们就很难准确预测设备运行寿命.

参考文献

- 1 Siebel T M. Digital Transformation Survive and Thrive in an Era of Mass Extinction. New York: RosettaBooks, 2019. 23-50
- 2 Tang J, Chen W G. Deep analytics and mining for big social data. Chin Sci Bull, 2015, 60: 509-519
- 3 Ding X, Li Z Y, Liu T, et al. ELG: an event logic graph. 2019. ArXiv: 1907.08015
- 4 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529-533
- 5 Wang J P, Shi Y K, Zhang W S, et al. Multitask policy adversarial learning for human-level control with large state spaces. IEEE Trans Ind Inf, 2019, 15: 2395-2404
- 6 Liu Q, Li Y, Yang D H, et al. Knowledge graph construction techniques. J Comput Res Develop, 2016, 53: 582-600
- 7 Liu Z Y, Sun M S, Lin Y K, et al. Knowledge representation learning: a review. J Comput Res Develop, 2016, 53: 1-16
- 8 Rawat S, Shen M H H. A novel topology optimization approach using conditional deep learning. 2019. ArXiv: 1901.04859
- 9 Zhou Z H, Zhang M L, Huang S J, et al. Multi-instance multi-label learning. Artif Intell, 2012, 176: 2291-2320
- 10 Socher R, Milind G, Christopher D M, et al. Zero-shot learning through cross-modal transfer. In: Proceedings of

- Advances in Neural Information Processing Systems, San Francisco, 2013. 1–10
- 11 Zeng D J, Liu K, Lai S W, et al. Relation classification via convolutional deep neural network. In: Proceedings of the 25th International Conference on Computational Linguistics, 2014. 2335–2344
- 12 Zhang D X, Wang D. Relation classification via recurrent neural network. 2015. ArXiv: 1508.01006
- 13 Miwa M, Banso M. End-to-end relation extraction using LSTMs on sequences and tree structures. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, 2016. 1105–1116
- 14 Feng C, Kang L Q, Shi G, et al. Causality extraction with GAN. Acta Autom Sin, 2018, 44: 811–818 [冯冲, 康丽琪, 石戈, 等. 融合对抗学习的因果关系抽取. 自动化学报, 2018, 44: 811–818]
- 15 Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2014. 701–710
- 16 Petar V, Guillem C, Arantxa C, et al. Graph attention networks. 2017. ArXiv: 1710.10903
- 17 Wang D, Cui P, Zhu W. Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 1225–1234
- 18 Chang S, Han W, Tang J, et al. Heterogeneous network embedding via deep architectures. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015. 119–128
- 19 Gui H, Liu J L, Tao F B, et al. Large-scale embedding learning in heterogeneous event data. In: Proceedings of IEEE 16th International Conference on Data Mining (ICDM), Beijing, 2016
- 20 Xu L C, Wei X K, Cao J N, et al. Embedding of embedding: joint embedding for coupled heterogeneous networks. In: Proceedings of the 10th ACM International Conference on Web Search and Data Mining, 2017. 741–749
- 21 Mehrkanoon S, Suykens J A K. Regularized semipaired kernel CCA for domain adaptation. IEEE Trans Neural Netw Learn Syst, 2018, 29: 3199–3213
- 22 Mihalkova L, Huynh T N, Raymond J M. Mapping and revising Markov logic networks for transfer learning. In: Proceedings of the 22nd Conference on Artificial Intelligence, 2007. 608–614
- 23 Parisotto E, Ba J L, Salakhutdinov R. Actor-mimic: deep multitask and transfer reinforcement learning. In: Proceedings of the International Conference on Learning Representations, Puerto Rico, 2016. 156–171
- 24 Rusu A A, Colmenarejo S G, Gulcehre C, et al. Policy distillation. 2016. ArXiv: 1511.06295
- 25 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489
- 26 Wang J P, Sun Y C, Zhang W S, et al. Large-scale online multitask learning and decision making for flexible manufacturing. IEEE Trans Ind Inf, 2016, 12: 2139–2147
- 27 Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning, 2015. 2048–2057
- 28 Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 664–676
- 29 Zhou J B, Gou S H, Hu R J, et al. A collaborative learning framework to tag refinement for points of interest. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019. 1752–1761
- 30 David M B, John D L. Correlated topic models. In: Proceedings of the Advances in Neural Information Processing Systems, 2006, 18: 147–154
- 31 Ranganath R, Gerrish S, Blei D M. Black box variational inference. In: Proceedings of the 17th International Conference on Artificial Intelligence and Statistics, Reykjavik, 2014. 814–822
- 32 Sun T, Sheldon D, Kumar A. Message passing for collective graphical models. In: Proceedings of the 32nd International Conference on Machine Learning, France, 2015. 853–861
- 33 Bernhard P. Disjunctive normal form. In: Encyclopedia of Machine Learning. Boston: Springer, 2017. 371–372
- 34 Wu Z H, Pan S R, Chen F W, et al. A comprehensive survey on graph neural networks. 2019. ArXiv: 1901.00596
- 35 Levie R, Monti F, Bresson X, et al. CayleyNets: graph convolutional neural networks with complex rational spectral filters. IEEE Trans Signal Process, 2019, 67: 97–109

- 36 Meng Q, Yoshua B, Jian T, et al. GMNN: graph markov neural networks. In: Proceedings of the 36th International Conference on Machine Learning, 2019. 5241–5250
- 37 Michelle A L, Yuke Z, Peter Z, et al. Making sense of vision and touch: learning multimodal representations for contact-rich tasks. 2019. ArXiv: 1907.13098
- 38 Zhu S Y, Ng I, Chen Z T. Causal discovery with reinforcement learning. 2019. ArXiv: 1906.04477

Constructing and inferring event logic cognitive graph in the field of big data

Junping WANG^{1*}, Wensheng ZHANG¹, Yongfei WANG² & Zhengya SUN¹

1. *State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;*

2. *Guodian Dadu River Basin Hydropower Development Co. LTD, Chengdu 610000, China*

* Corresponding author. E-mail: junping.wang@ia.ac.cn

Abstract Along with the increasing impact of big data on the economic operation mechanism, including global production, circulation, distribution and consumption, event logic (EL) is becoming an important concept in mining and releasing the potential value of big data. Besides its critical role in exploring the clues and rules of digital economy, EL will also help people better understand the future development of the ternary human-cyber-physical universe. Considering the dynamic evolution of the EL is the essential difference between industrial big data and other big data. Therefore, we propose a constructing and inferring EL cognitive graph in the field of big data. Firstly, we introduce the great opportunities offered by the knowledge graph in promoting the high-quality development of the manufacturing industry, leading the economic development of the new digital manufacturing industry, and promoting the cross-development of sociology and information science. Secondly, we present the construction technology of EL cognitive graph and expound the progress and critical technical challenges in the research of the EL cognitive graph at home and abroad, in the aspects of multi-granularity acquisition and representation of EL, multimodal association and fusion, evolution analysis and mining of EL behavior. Finally, combined with the research progress of our works in EL cognitive graph, we summarize the effect of EL cognitive graph in the prevention of network fraud and the control of safe production as well as summarize and anticipate the future research direction and development prospect of the industry.

Keywords industrial big data, knowledge graph, event logic cognitive, analysis of human-machine behavior evolution

Junping WANG is a professor at Laboratory of Precision Sensing and Control Center, Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree in computer science and technology from Beijing University of Posts and Telecommunications in 2013, China. His research interests include cognitive computing, machine learning, and industrial big data analysis.

Wensheng ZHANG received his Ph.D. degree in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences, in 2000. He joined Institute of Software, Chinese Academy of Sciences, in 2001. He is a professor of Machine Learning and Data Mining and the director of Research and Development Department, Institute of Automation, Chinese Academy of Sciences.

Yongfei WANG was born in Kaifeng, Henan Province in 1983. He is a Ph.D. student, senior engineer, and the most responsible safety manager in Sichuan Province in 2017. His research interests include structural health intelligent monitoring, artificial intelligence product development, and hidden danger prediction.

Zhengya SUN was born in 1982. She received her Ph.D. from Institute of Automation, Chinese Academy of Sciences. She is an associate professor and master supervisor in Machine Learning and Big Data Lab, Institute of Automation, Chinese Academy of Sciences. Her research interests include statistical relational learning, probabilistic graphical model, data stream classification, and approximate reasoning.