DOI:10.7544/issn1000-1239.20211264
59(8):1694-1722,2022

基于强化学习的知识图谱综述

马 昂[1] 于艳华[1] 杨胜利[2] 石 川[1] 李 劼[1] 蔡修秀[1]
[1](北京邮电大学计算机学院(国家示范性软件学院) 北京 100876)
[2](中国人民解放军国防大学 北京 100091)
(ang@bupt.edu.cn)

Survey of Knowledge Graph Based on Reinforcement Learning

Ma Ang[1], Yu Yanhua[1], Yang Shengli[2], Shi Chuan[1], Li Jie[1], and Cai Xiuxiu[1]
[1](*School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing* 100876)
[2](*National Defence University of People's Liberation Army, Beijing* 100091)

Abstract Knowledge graph (KG) is a form of data representation that uses graph structure to model the connections between things. It is an important foundation for realizing cognitive intelligence and has received extensive attention from academia and industry. The research of knowledge graph includes four parts: knowledge representation, knowledge extraction, knowledge fusion, knowledge reasoning. Currently, there are still some challenges in the research of knowledge graphs. For example, knowledge extraction methods face difficulty in obtaining labeled data, while distantly supervised training samples have noise problems. The interpretability and reliability of the knowledge reasoning methods need to be further improved. Knowledge representation methods also have problems such as relying on manually defined rules or prior knowledge. Knowledge fusion methods fail to fully model the interdependence between entities. Environment-driven reinforcement learning (RL) algorithms are suitable for sequential decision-making problems. By modeling the research problem of the knowledge graph into a path (sequence) problem, and applying reinforcement learning methods, the above-mentioned problems in the knowledge graph can be solved, which has important application value. The basic knowledge of KG and RL are introduced firstly. Secondly, a research of KG based on RL are comprehensively reviewed. Then, it focuses on how the KG method based on RL can be applied to practical application areas such as intelligent recommendation, conversation system, game, biology, medicine prediction, finance and cybersecurity. Finally, the future directions of KG and RL are discussed in detail.

Key words knowledge graph; reinforcement learning; named entity recognition; relation extraction; knowledge reasoning; knowledge representation; knowledge fusion

摘 要 知识图谱是一种用图结构建模事物及事物间联系的数据表示形式,是实现认知智能的重要基础,得到了学术界和工业界的广泛关注.知识图谱的研究内容主要包括知识表示、知识抽取、知识融合、知识推理4部分.目前,知识图谱的研究还存在一些挑战.例如,知识抽取面临标注数据获取困难而远程

收稿日期:2021-12-24;**修回日期**:2022-03-10
基金项目:国家自然科学基金项目(U1936104);国家重点研发计划项目(2020YFB2104503)
This work was supported by the National Natural Science Foundation of China (U1936104), and the National Key Research and Development Program of China (2020YFB2104503).
通信作者:于艳华(yuyanhua@bupt.edu.cn)

监督训练样本存在噪声问题,知识推理的可解释性和可信赖性有待进一步提升,知识表示方法依赖人工定义的规则或先验知识,知识融合方法未能充分建模实体之间的相互依赖关系等问题.由环境驱动的强化学习算法适用于贯序决策问题.通过将知识图谱的研究问题建模成路径(序列)问题,应用强化学习方法,可解决知识图谱中的存在的上述相关问题,具有重要应用价值.首先梳理了知识图谱和强化学习的基础知识.其次,对基于强化学习的知识图谱相关研究进行全面综述.再次,介绍基于强化学习的知识图谱方法如何应用于智能推荐、对话系统、游戏攻略、生物医药、金融、安全等实际领域.最后,对知识图谱与强化学习相结合的未来发展方向进行展望.

关键词　知识图谱;强化学习;命名实体识别;关系抽取;知识推理;知识表示;知识融合

中图法分类号　TP391

自谷歌在2012年推出“知识图谱”(knowledge graph, KG)后,知识图谱技术已迅速成为数据挖掘、数据库和人工智能等领域的研究热点.知识图谱采用图结构来描述知识和建模事物及事物间关系[1].它将信息表达成更接近人类认知的形式,提供了一种组织、管理和认知理解海量信息的能力[2].知识图谱本质是一种大规模语义网络,既包含了丰富的语义信息,又天然具有图的各种特征,其中,事物或实体属性值表示为“节点”,事物之间的关系或属性表示为“边”.目前,知识图谱相关的知识自动获取[3-5]、知识推理[6-8]、知识表示[9-10]、知识融合[11]已成为搜索问答[12]、大数据分析[4]、智能推荐[6]和数据集成[11]的强大资产,被广泛应用于多个行业领域.

目前,大部分知识图谱的研究是基于监督学习的方法[3,6,13-14].然而,为模型获得足够的标注数据成本较高.为此部分学者提出使用远程监督的方法来减少数据标注[15],远程监督指的是借助外部知识库为数据提供标签[16].但远程监督获得的训练样本中存在噪声.此外,现有方法还存在依赖人工预定义的规则和先验知识或模型缺乏可解释性等问题.强化学习(reinforcement learning, RL)适用于贯序决策问题,通过学习如何与环境交互,进而辅助人类决策.它在进行策略选择时更关注环境状态,对行为的选择进行更好地理解和解释.将知识图谱研究的问题建模成路径或序列相关的问题,例如,将基于远程监督的命名实体识别中干净样本的选择建模成序列标注任务、将关系推理建模成路径查找问题等,应用强化学习算法可以避免依赖人工预定义的规则或先验知识,解决模型缺乏可解释性或仅提供事后可解释性(post-hoc explanation)的问题,具有重要的研究和应用价值.

近年来,学术界和工业界对知识图谱、强化学习2个领域进行了深入研究,有不少分别聚焦知识图谱和强化学习的综述性文章.文献[1,3-4,6-8,11,14,17]分别围绕知识图谱的表示学习、知识获取、知识推理、知识图谱构建与应用、多模态知识融合等进行综述.文献[18-22]分别对基于价值的和基于策略的强化学习、深度强化学习算法、多智能体算法进行综述.文献[23-24]对强化学习在综合能源管理和金融交易领域的研究进行阐述.然而,尽管已有诸多的知识图谱、强化学习综述文献,但仍缺乏对知识图谱和强化学习相结合的研究进行系统地梳理和总结的工作.与现有的工作相比,本文工作的不同主要体现在2个方面:1)通过系统调研已发表的基于强化学习的知识图谱相关研究的论文,全面总结了基于强化学习的知识图谱研究,包括知识抽取、知识推理、知识表示、知识融合等研究成果.2)介绍了基于强化学习的知识图谱如何应用于智能推荐、游戏攻略、生物医药、金融、网络安全等实际领域.本文是第1篇系统介绍该研究方向的综述论文.

1　知识图谱研究进展

知识图谱作为大数据时代重要的一种结构化的知识表示形式,引起了学术界和工业界的广泛关注与研究.“知识图谱”由谷歌于2012年正式提出,其目的是为了支撑语义搜索任务而建立的知识库.随着知识图谱技术的不断发展和进步,知识图谱的概念也不断被丰富和深化.知识图谱定义为 $G=\{E,R,F\}$,其中,E,R 和 F 分别表示实体、关系、事实的集合,事实被定义为一个三元组 $(h,r,t)\in F$,其中,h 和 t 分别代表头实体和尾实体,r 代表头尾实体间的关系.图1是名著《水浒传》的一个知识图谱片段,图中节点表示实体,边表示关系,三元组(宋江,结拜,武松)表达了宋江与武松是结拜兄弟的事实.

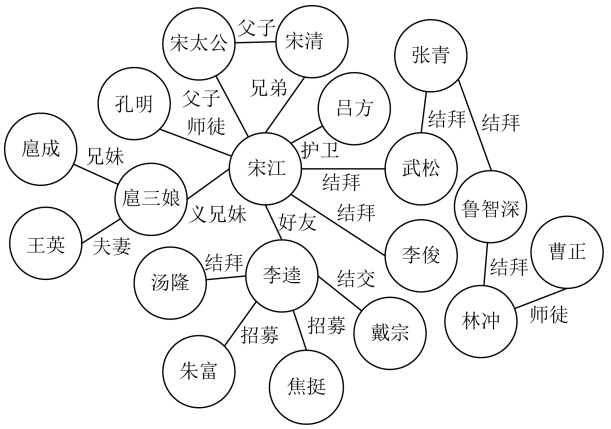


Fig. 1 Example of knowledge graph

图 1 知识图谱示例

目前,学术界和工业界已构建了一大批知识图谱.学术界最具代表性的有 DBpedia[25],YAGO[26],ConceptNet[27],Wikidata[28],以及国内学术团队构建的 AMiner[29],CN-DBpedia[30],Zhishi.me[31] 等,涉及通用常识、科技文献、语言翻译等领域.在工业界,谷歌、微软、阿里、美团等公司都投入了大量资源来构建各自的领域知识图谱.例如,阿里利用来自于淘宝、天猫等多个平台的商品数据构建了一个包含了百亿级别三元组数据的商品知识图谱,用于搜索、前端导购、智能问答等业务,帮助行业人员进行选品,提高消费者购物体验.知识图谱已经在搜索、社交、商业、医疗等领域有了一定的实践与应用,并取得了较好的成效.

通常,知识图谱的研究主要包括知识抽取、知识推理、知识表示、知识融合等方面,表 1 围绕这些方面分别从传统模型与基于深度学习的模型 2 个角度,对知识图谱研究中的常见算法进行分类总结.

Table 1 Classic Knowledge Graph Algorithms

表 1 知识图谱相关研究算法

类别		传统模型	基于深度学习的模型	
			基于神经网络的模型	基于图神经网络的模型
知识抽取	实体识别	LaSIE-II[32],FACILE[33],NetOwl[34],SRA[35],LTG[36]	CNN-CRF[37],RNN-CRF[38],BiLSTM-CRF[39], TagLM[40]	CGN[41],GraphRel[42],BiFlaG[43]
	关系抽取	Bootstrap[44],Snowball[45]	CR-CNN[46],Bi-LSTM+CNN[47], PCNN+ATT[48]	GP-GNNs[49],C-GCN[50],AGGCN[51], KATT[52],GCNN[53],GAIN[54]
知识推理		FOIL[55],AMIE[56],PRA[57]	NTN[58],ProjE[59],Path-RNN[60]	SACN[61],RGHAT[62], RGCN[63],CompGCN[64]
知识表示		TransE[65],TransR[66],TransD[67], DistMult[68],ComplEx[69],PTransE[70]	CapsE[71],ConvKB[72], ConvE[73],ConvR[74]	TransGCN[75],ReInceptionE[76]
知识融合		MTransE[77]	ED-LNA[78]	GNED[79],MuGNN[80], OAG[81],AliNet[82]

1) 知识抽取是从不同来源、结构的数据中提取知识,形成结构化数据存入知识图谱.对于结构化和半结构化的数据,可以直接利用映射、转换等操作.但对于非结构化数据而言,知识抽取较为困难.一般知识抽取任务包括命名实体识别(named entity recognition, NER)、关系抽取(relation extraction, RE)(实体属性抽取、实体关系抽取)等.

2) 知识推理是从已有的知识中推理实体间可能存在的关系或属性值.知识图谱通常是不完整的,例如,实体间路径缺失、实体属性值缺失等.因此,知识推理常用于知识图谱补全(knowledge graph completion),也可用于知识图谱去噪(knowledge graph cleaning)等任务.

3) 知识表示是对现实世界的一种抽象表达.知识表示方式主要分为符号表示和数值表示[2],符号表示,如网络本体语言(web ontology language, OWL),RDF(resource description framework)等,符号表示方便易于理解,但基本符号性质使知识图谱难以操作[1].因此,提出了知识图谱嵌入(knowledge graph embedding, KGE)或知识表示学习(knowledge representation learning, KRL)方法,将知识图谱的实体和关系嵌入到连续向量空间中[1],从而实现对其语义信息和固有结构的表示.

4) 知识融合是将从不同来源得到的同一实体或概念的描述信息融合起来[11].描述信息可以是同种类型,也可以是不同类型.例如图片、文字、音频、视频等.

近年来,针对知识图谱的研究已经取得了很大进展.文献[13-14]基于深度学习方法分别对实体识别、实体关系抽取进行了全面综述.文献[17]利用浅层语言分析中的基础语言信息和关系结构信息 2 个层面特征对自动术语抽取问题进行分类总结.文献

[8]将知识图谱推理分为单步推理、多步推理,分别从基于规则的、基于表示学习的、基于神经网络的以及混合推理4个方面对知识推理的最新研究进行了归纳总结.文献[4]围绕事理认知图谱的构建与推断进行总结归纳,梳理了事理认知图谱的最新应用效果.文献[83]从知识图谱构建过程出发,将知识图谱补全问题分为概念补全和实例补全2个层面,对知识图谱补全技术进行了系统的回顾与探讨.文献[1]对知识图谱嵌入技术进行梳理总结.文献[11]对多源知识融合相关研究技术和最新进展进行了归纳总结.

虽然基于深度学习的方法在知识图谱的研究已经取得了不错的效果,但还存在一些问题,主要表现在4个方面:

1) 标记数据缺乏.为监督学习获取领域标记数据成本较高.为此,部分学者引入远程监督学习,虽然减少了数据标注的成本,但所构造的训练样本中噪声较高.

2) 数据常识信息匮乏[4].生活中存在很多约定俗成的常识知识,这些知识几乎不会显式地出现在大部分语料中.

3) 知识图谱存在不完整性[32,83].不完整性主要表现在2个方面:显式不完整,即实体之间路径缺失;隐式不完整,即2个实体之间存在过长的路径,现有的推理模型很难推断.

4) 现有基于深度学习的方法缺乏可解释性[7,83].目前,应用深度学习进行知识图谱的推理或者推荐方法更关注于结果的准确性,结果的透明性和可解释性较差.

随着深度学习技术的快速发展,人们希望赋予知识图谱更高的能力,即赋予知识图谱更强的推理、理解、表达能力.强化学习是一种从试错过程中发现最优行为策略的技术,已经成为解决环境交互问题的通用方法[84].不同于通过数据学习规律的方法,强化学习是通过与环境的交互来学习,这种方式更接近于人类的学习认知过程.因此,强化学习方法具备强大的探索能力和自主学习能力.知识图谱与强化学习的结合主要有3种思路:

1) 将知识图谱的相关问题建模成路径(序列)问题,利用强化学习的方法来解决.例如,将命名实体识别建模为序列标注任务,使用强化学习方法来学习标注策略;将知识推理建模为路径推理问题,利用强化学习方法进行关系和节点选择.

2) 将强化学习方法用于有噪声训练样本的选择或过滤,减少远程监督方法所带来的噪声,利用高质量的样本提高知识图谱命名实体识别和关系抽取方法的性能.

3) 将知识图谱所包含的信息作为外部知识,编码进强化学习的状态或奖励中,提升强化学习智能体的探索效率,应用于关系抽取和知识推理等场景.知识图谱与强化学习的结合对于提高模型的可解释性和推理能力,提升训练数据质量,具有重要的研究与应用价值.

2　强化学习研究进展

强化学习研究智能体(agent)与环境(environment)的相互作用,通过不断学习最优策略(policy),做出序列决策并获得最大奖励(reward)[85-86].强化学习的过程可以由Markov决策过程(Markov decision process, MDP)来描述,使用四元组来表示(A,S,$\boldsymbol{P}$,R).其中,动作空间A表示智能体对环境施加的动作集合,状态空间S表示环境状态集合,$\boldsymbol{P}$为状态转移矩阵,奖励R表示环境对动作做出的反馈.策略π为状态空间到动作空间的映射.智能体与环境交互,如图2所示,其中,A_t,S_t,R_t分别表示在时刻t的动作、状态和奖励.通常,状态与奖励的设置与实际问题密切相关.强化学习的核心目标是使长期累积奖励最大化.累计奖励被定义为奖励序列的一些特定函数,由于未来奖励的总和往往是无穷大的,一种常见的做法是引入折扣因子$\gamma\in[0,1]$,用于平衡最近的奖励与未来的奖励,时刻t以后的累积奖励为

$$G_t=R_{t+1}+\gamma R_{t+2}+\cdots=\sum_{\tau=0}^{+\infty}\gamma^{\tau}R_{t+\tau+1}. \quad (1)$$

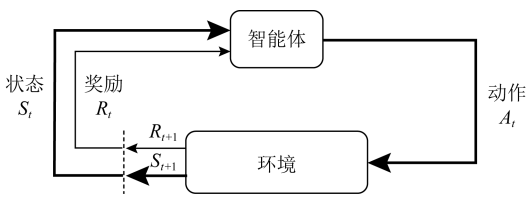


Fig. 2　Interaction between agent and environment

图2　智能体与环境交互图

依据是否对环境建模,强化学习方法可分为基于模型的强化学习方法和无模型的强化学习方法[23,85].

1) 基于模型的强化学习方法.假定任务对应的Markov决策过程四元组均为已知,即机器已对环境进行了建模,然后再利用该模型做出动作规划或者策略选择,方法对模型十分敏感.

2) 无模型的强化学习方法.不需要对环境建模,通过和环境交互来学习到一个价值函数或者策略函数.依据智能体策略计算方式,分为基于价值(value-based)、基于策略(policy-based)以及基于Actor-Critic的方法3类,表2对无模型经典强化学习算法进行了简要对比.

Table 2 Comparison of Classic Model-Free Reinforcement Learning Methods

表2 经典无模型强化学习方法对比

类别	基于价值的方法	基于策略的方法	Actor-Critic方法
原理	通过学习价值来指导策略	直接学习策略	既学习价值函数又学习策略
经典算法	Q-learning[87],SARSA[88],Deep Q-learning[89],Deep Q-Network (DQN)[90],Double DQN[91],Prioritized Replay DQN[92],Dueling DQN[93]	REINFORCE[94]	Actor-Critic[95],A3C[96],DDPG[97],TRPO[98],PPO[99]
优势	样本利用率高、价值函数估值方差小,不易陷入局部最优	适用于现阶段的行动对未来决策影响较深的任务,例如围棋、象棋等	价值函数估值方差小、样本利用率高,算法整体的训练速度快
劣势	适用离散的动作集合,且最优策略通常是确定性策略	需要大量的采样训练,收敛性差,容易收敛到局部最优	

2.1 基于价值的强化学习

基于价值的强化学习方法通过学习价值来指导策略,通过选取最大价值函数对应的动作,隐式地构建最优策略.Watkins等人[87]和Rummery等人[88]将状态与动作构建成一张Q-table来存储Q值,根据Q值选择获得最大收益的动作,分别提出Q-learning和SARSA.这类方法虽然简单,但面对复杂状态集合问题时,需要维护一张巨大的Q-table.一种有效的解决方法是对价值函数近似表示.Deep Q-learning[89]利用深度神经网络对动作价值函数进行拟合,神经网络的输入是状态,输出是近似Q函数.Mnih等人[90]使用结构一样的主网络(main)和目标网络(target)替换原有神经网络,提出了DQN(deep Q-network).主网络用来选择动作,更新模型参数;目标网络用于计算Q'值,目标网络的参数采用延时更新.DDQN(deep double Q-network)[91]通过解耦动作选择和Q值的计算,先在主网络中找出最大Q值对应的动作a',然后利用该动作在目标网络中计算Q值来消除过估计(over estimation).DQN和DDQN都使用了经验回放(experience replay),Schaul等人[92]提出优先级经验回放(prioritized experience replay),使用时序差分误差(temporal difference error, TD-error)来衡量优先级(权重),按照权重采样有利于加快学习速度.Dueling DQN(dueling deep Q-network)[93]进一步将Q网络分为价值函数、优势函数.价值函数仅与状态有关,优势函数同时与状态和动作有关,Q价值函数可以通过价值函数、优势函数计算得到.但需要注意的是这类方法通常适用于处理离散的动作集合,且最优策略通常是确定性策略.

2.2 基于策略的强化学习

对于动作空间连续或策略是随机的问题,可以利用基于策略的强化学习方法.基于策略的方法是对策略函数近似,使用含参函数$\pi(a|s,\theta)$来计算最优策略,模型由参数θ控制得到最优策略.对于离散动作空间,可以使用softmax计算动作概率;对于连续空间,通常使用高斯分布计算动作概率.基于策略的目标函数$J(\theta)$有3种常用计算方式:基于初始状态期望、基于平均价值、基于平均奖励,如表3所示.初始状态期望是计算从某一初始状态开始,智能体依据策略一直到回合结束,所获得的奖励之和.平均价值是指对于没有初始状态的任务.例如,连续性任务,从某时刻起,计算其所有可能的状态价值函数的均值.平均奖励指每一时间步的平均奖励,即所有可能状态在该策略下,所能获得的奖励的加权平均.在

Table 3 Common Objective Functions Used in Policy-Based Reinforcement Learning

表3 基于策略的强化学习常用目标函数

类别	目标函数	说明
初始状态期望	$J_1(\theta)=V_{\pi_\theta}(s_1)$	$V_{\pi\theta}(s_1)$记为状态值函数,表示从某一初始状态s_1开始,依据策略π_θ直到回合结束的累计奖励
平均价值	$J_2(\theta)=\sum_s d_{\pi_\theta}(s)V_{\pi_\theta}(s)$	所有可能的状态价值函数的均值.其中,$d_{\pi_\theta}(s)$是基于策略生成的Markov链关于状态的静态分布
平均奖励	$J_3(\theta)=\sum_s d_{\pi_\theta}(s)\sum_a \pi_\theta(s,a)R(s,a)$	所有可能状态在该策略下,所能获得的及时奖励的加权平均值,其中,$\pi_\theta(s,a)$表示在状态s下选择动作a所获得的奖励

确定目标函数后,对目标函数进行优化,例如,采用梯度上升更新参数,即可确定最优策略.蒙特卡洛策略梯度 REINFORCE[94]是一种经典的基于策略的算法,但该算法需要使用从当前时刻开始到结束的所有奖励,策略梯度计算:

$$\nabla_\theta J(\theta)=E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)G_t], \quad (2)$$

其中,$\nabla_\theta J(\theta)$表示目标函数对 θ 求导,E 表示期望.

REINFORCE 导致方差较高,从而降低智能体的学习速度.为了解决这一问题,研究者提出了一些方法,例如,在计算累计奖励时减去基线,策略梯度计算:

$$\nabla_\theta J(\theta)=E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)(G_t-b(s))], \quad (3)$$

其中,$b(s)$表示只与状态有关,不随动作变化的任意函数,常使用状态价值函数作为基线.基于策略的强化学习方法通过策略梯度方法直接优化用深度神经网络参数化表示的策略.这类方法中,所有更新都只有在回合结束后才能进行,梯度方差较大,学习速率较为缓慢.

2.3　基于 Actor-Critic 的强化学习

Actor-Critic 算法[95]也是降低基于策略的强化学习方差的一种方式.它将基于价值和基于策略的强化学习相结合,由 Actor 和 Critic 网络组成.Actor 根据价值函数训练策略,选择动作得到反馈;Critic 根据状态训练价值函数,用于评价策略的好坏.REINFORCE 等基于策略的强化学习方法通过采样,利用实际累积奖励计算策略梯度.而 Actor-Critic 使用价值函数的估计值,计算策略梯度.Actor-Critic 经典算法与策略梯度计算,如表 4 所示.A3C(asynchronous advantage Actor-Critic)[96]通过同时生成多个 AC 算法线程,同步进行训练,共享参数,提高了算法效率.DDPG(deep deterministic policy gradient)[97]构建了 4 个神经网络:主 Actor 网络、目标 Actor 网络、主 Critic 网络和目标 Critic 网络.DDPG 借鉴了 DQN 中双网络的思想,通过双网络和经验回放,解决了 Actor-Critic 收敛困难的问题.为了保证策略的优化总是朝着不变坏的方向进行,研究者们提出了 TRPO(trust region policy optimization)[98],PPO(proximal policy optimization)[99]等算法,来提高策略梯度的收敛速度.基于 Actor-Critic 的强化学习方法在基于策略的和基于价值的强化学习方法之间找到一种平衡,降低策略梯度求解时的梯度(估计)方差.

Table 4　Classical Algorithms based on Actor-Critic and Policy Gradient Calculation

表 4　基于 Actor-Critic 的经典算法与策略梯度计算

代表性算法	策略梯度计算$\nabla_\theta J(\theta)$	说明
Actor-Critic	$\nabla_\theta J(\theta)=E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)Q_\pi(s,a)]$	用值函数 $Q_\pi(s,a)$策略梯度(见式(2))中的累计奖励值 G_t
Advantage Actor-Critic	$\nabla_\theta J(\theta)=E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)A(s,a)]$	用优势函数 $A(s,a)=Q_\pi(s,a)-V_\pi(s)$代替策略梯度(见式(2))中的累计奖励值 G_t,$V_\pi(s)$为状态值函数(如表 3 所示)
TD Actor-Critic	$\nabla_\theta J(\theta)=E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\sigma]$	用时序差分 $\sigma=R_{t+1}+\gamma V_\pi(s_{t+1})-V_\pi(s)$ 代替策略梯度(见式(2))中的累计奖励值 G_t

2.4　其　他

除 2.1～2.3 节所述的经典强化学习方法外,还有多智能体强化学习方法、分层强化学习方法、对抗强化学习方法等.多智能体强化学习(mulit agent reinforcement learning, MARL)指至少拥有 2 个智能体的强化学习方法[100-101].Lowe 等人[100]提出了多智能体深度确定性策略梯度方法(mulit agent deep deterministic policy gradient, MADDPG)方法,每个智能体的学习需考虑其他智能体的动作策略,进行中心化训练和非中心化执行.Sunehag 等人[101]提出考虑协作任务的多智能体强化学习算法价值分解网络(value-decomposition networks, VDN).所有的智能体共享同一个奖励值,智能体之间共享网络参数,算法收敛速度快.多智能体强化学习状态空间和联合动作空间随智能体数量指数增长,计算复杂度较大.面对维度灾难和稀疏奖励延迟问题,研究者提出分层深度强化学习方法(hierarchical deep reinforcement learning, HDRL)[102].分层的思想有利于减小问题规模,降低奖励稀疏和延迟问题.Eysenbach 等人[102]提出无监督框架下的策略学习算法 DIAYN(diversity is all you need),该方法在无奖励的环境中,自适应地产生奖励函数.基于互信息的目标函数学习到一些有用的技能,用来控制智能体访问状态.面对模拟环境和现实环境存在差异,策略难以迁移的问题,研究者提出对抗强化学习方法(generative adversarial reinforcement learning, GARL)[103-104].Pinto 等人[103]提出鲁棒的对抗强化学习方法(robust adversarial reinforcement learning, RARL),通过同时训练 2 个智能体使强化学习更好地泛化到真实环境.Protagonist 智能体做出决策,

Adversary 智能体产生扰动干扰 Protagonist 智能体决策，使用交替过程优化 2 个智能体.Chen 等人[104]提出了 Cascading-DQN，利用生成对抗网络同时学习用户行为模型以及奖励函数.将用户行为模型作为强化学习的环境，得到候选物品组合推荐策略，解决推荐系统利用用户的在线反馈来训练推荐策略，消耗大量交互成本，影响用户体验的问题.

3 基于强化学习的知识图谱研究

目前，大多数知识图谱的相关方法基于监督学习，但对数据进行标注费时费力.为了解决标注困难的问题，有学者提出了远程监督的方法.远程监督减少了数据标注成本，但又在训练数据中引入了噪声[15].虽然，目前知识图谱的研究方法在准确率、精度、召回率等性能上取得了很好的效果，但这些方法结果的透明性、可解释性、可信赖性等还有待进一步研究[7,84,105].强化学习方法不同于一般的监督学习，它把相关问题建模为序列决策问题，近年来在知识图谱领域得到应用，可以帮助解决远程监督的噪音问题、知识推理结果可解释性差[105]等问题.本节将分别从命名实体识别、关系抽取、知识推理、知识表示、知识融合等 5 个方面，详细介绍强化学习方法在各类研究中的进展，如图 3 所示：

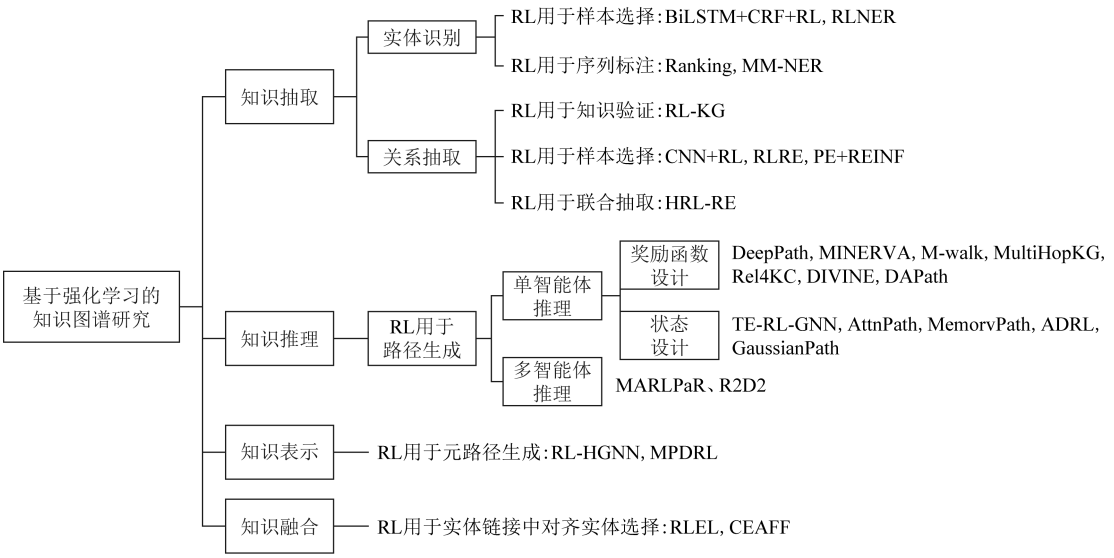


Fig. 3 Classification of knowledge graph research methods based on reinforcement learning

图 3 基于强化学习的知识图谱研究方法分类

3.1 命名实体识别

命名实体识别旨在对序列进行命名实体标注，判断输入句子中的词是否属于人名、地名、组织机构名等.现有命名实体识别方法依赖人工标注数据，但标注成本较高.远程监督方法可以降低标注成本[15]，但远程监督获得的训练样本中又存在噪声.强化学习方法可以通过自主学习选择高质量的训练样本数据，解决上述问题.目前，基于强化学习的命名实体识别方法思路主要有 2 类：1)使用深度强化学习模型自动学习样本选择策略，过滤掉训练数据中的噪声.2)将命名实体识别任务利用强化学习来建模，即将序列标注任务转换为序列决策问题.通过利用 Markov 决策过程模型来进行序列标注，即为序列中的每个元素分配一个标签.

基于将强化学习用于命名实体识别中的训练样本选择这一思路，Yang 等人[106]采用基于策略的强化学习来解决远程监督方法中训练数据存在噪声的问题，设计了一个从远程监督方法得到的部分标注(partial annotation)数据中获得干净实例的算法.算法框架如图 4 所示，包括命名实体(NE)标记器(图 4 左部分)、实例选择器(图 4 右部分).NE 标记器基于双向 LSTM(bi-directional LSTM, BiLSTM)和条件随机场(conditional random fields, CRF)模型进行命名实体识别，其训练数据包括部分手工标注的监督数据和实例选择器从部分标注数据中选择的干净实例.实例选择器采用基于策略的强化学习方法，首先选取训练数据的一个子集作为包.智能体需要从包中的部分标注数据中选择正确标记的句子，作为干净实例输入 NE 标记器.具体来看，对于每个句子，智能体根据策略网络决策对其执行的

动作(选择或不选择).状态如图 4 的虚线框所示,包括由 BiLSTM 编码的当前实例的向量表示以及由 MLP 根据当前实例向量表示计算的该实例词序列中的标签分数.包中所有句子处理完毕智能体得到一个奖励,奖励是 NE 标记器对包中所有句子标签序列的条件概率取对数平均值.实例选择器根据 NE 标记器提供的奖励进行学习,优化实例选择器的策略网络.该模型利用强化学习的思想进行远程监督样本数据选择或去噪,提升了命名实体识别性能.更进一步,Wan 等人[107]为了处理噪音数据对命名实体识别模型带来的影响,利用基于策略的 REINFORCE 算法对样本数据进行纠正.给定一个句子,状态被定义为当前的输入(词)和以前的上下文,动作被定义为是否对该词的标签进行修改.该模型由 2 个模块组成:标签修改器和标签预测器.标签修改器模块作为强化学习中的智能体,能够纠正标签错误的训练数据.标签预测器采用 BiLSTM+CRF 模型,用来完成序列标注任务.2 个模块在训练过程中相互影响.标签修改器的状态表示从标签预测器生成,同时也会从标签预测器获得奖励来指导策略的学习.而标签预测器模块又依赖于从标签修改器获得的最终标签进行训练.与 Yang 等人[106]不同,Wan 等人[107]的模型是学习一个独立的标签修改器来纠正错误的标签,标签预测器的性能变化作为标签修改器的奖励来直观地反映标签修改器的效果.通过利用高质量的训练数据,以提高命名实体识别模型的性能.

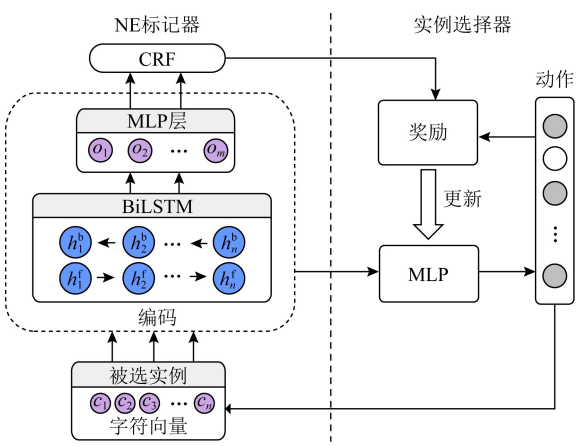


Fig. 4　Framework of NER based on reinforcement learning[106]

图 4　基于强化学习的 NER 算法框架图[106]

命名实体识别也可以直接建模为序列决策问题.Maes 等人[108]利用强化学习来解决序列标注问题.利用 Markov 决策过程对标签序列构建过程建模,引入了基于蒙特卡洛思想的 Rollout 算法,替换 SARSA 算法中的 Q 价值函数,以便有效地学习如何标记新序列.状态包括当前句子以及已经标注的标签序列,动作空间是由当前词所有可能的标签所构成的集合.提出了一种可以按任何顺序来预测序列标签的算法.首先预测置信度更高的标签,以丰富上下文信息,帮助减少在预测置信度较低的标签时的歧义.该模型在保证不增加复杂度的前提下,提升了命名实体识别的性能.Lao 等人[109]提出一种利用强化学习进行命名实体识别的算法 MM-NER (MCTS enhanced MDP for NER).在 AlphaGoZero 方法的启发下,MM-NER 是一种基于策略梯度的强化学习算法,并将蒙特卡洛树搜索 MCTS 用于对策略提升,进行命名实体识别的模型.状态包括当前词的上下文和已经标注的标签序列.动作空间由当前词的所有可能标签构成的集合.具体来说,使用了 2 个 LSTM 网络分别编码上下文,通过将 LSTM 网络的输出拼接起来输入到全连接层 MLP 得到状态表示.在训练过程中,MM-NER 利用策略网络的输出和价值函数来指导蒙特卡洛树的搜索,最后输出一个更准确的搜索策略.不同于直接利用策略函数进行序列标注,MM-NER 利用 MCTS 在生成的策略函数和值函数的指导下进行探索,降低命名实体识别任务的时间复杂度和模型陷入局部最优解的可能性,并取得了较好性能.

3.2　关系抽取

关系可以定义为实体之间或实体与属性之间的某种联系,关系抽取就是自动识别实体(或实体与属性)之间具有的某种语义关系.现有关系抽取方法大多基于神经网络模型[46-54],通过监督学习或远程监督学习来完成抽取任务.为了降低标注成本,学者们提出使用远程监督的方法.远程监督方法虽然有效,但在训练样本中引入了噪声[15].强化学习方法可以通过知识引导来避免噪声数据带来的影响.基于强化学习的关系抽取方法主要可以分为 3 类:1)使用强化学习模型对抽取结果进行知识验证;2)利用强化学习模型进行训练样本选择;3)将实体识别与关系抽取 2 个任务联合建模,互为增强.

知识图谱中的实体属性往往是嘈杂、不完整的,甚至是缺失的.例如,知识库 DBpedia 中有近一半的实体包含的关系(包括属性)少于 5 条[110].针对关系抽取任务中的属性抽取,Liu 等人[111]提出利用强化学习方法为开放域新实体补充可靠属性关系的算法

RL-KG.为了有效地过滤文章不正确或信息提取系统错误而产生的嘈杂答案,Liu 等人[111]提出了一个知识引导的强化学习框架,来进行开放域属性提取.在该框架中,属性抽取任务首先利用模板转化为搜索引擎的搜索问题.除信息搜索系统为强化学习智能体提供的候选答案外,知识库中的相关知识也被强化学习智能体用作背景知识,辅助决策.属性提取任务被建模为 Markov 决策过程.首先,给定 2 个候选答案,人为指定第 1 个作为当前最佳答案,状态包括由搜索引擎给出的 2 个答案的置信度、2 个答案与知识库给出的相关知识的相似度、2 个答案间的相似度.动作包括停止搜索、保留当前的最佳答案并继续搜索、替换当前的最佳答案并继续搜索.采用基于价值的 DQN 算法,根据当前状态对动作价值进行估计.该算法框架可以适用于不同的信息提取系统,并且通过知识的引导可以显著提高属性抽取的性能.

现有基于远程监督的关系抽取模型假设只要同时包含 2 个实体的句子,都在描述同一种关系.这一假设会产生很多错误标签,需要通过一些样本标签过滤方法,提升训练样本质量.例如,利用深度强化学习策略来选择正确的句子作为训练数据,尽量避免错误标签对模型的影响[112-114].Feng 等人[112]和 Qin 等人[113]利用强化学习智能体作为样本实例选择器,选择正确的训练样本.其中,Feng 等人[112]提出了一种新的关系抽取模型 CNN+RL,如图 5 所示,该算法由实例选择器、关系分类器 2 部分构成.CNN+RL 采用经典的基于策略的 REINFORCE 算法来训练实例选择器,目的是尽量选择正确的句子进行学习.给定一个由若干条句子组成的句子包,状态包括当前句子、已经选择的句子和实体对,动作被定义为是否选择当前句子.实例选择器对句子进行选择,然后使用所选择的句子训练关系分类器.关系分类器应用 CNN 模型获得句子的抽象表示,并基于句子级的关系分类概率计算包中所选句子的联合概率的几何平均数,以此作为奖励传递给实例选择器,通过计算梯度更新策略网络参数.与 Feng 等人的工作类似,Qin 等人[113]提出了一种噪音训练数据指示器,通过指示器能够自动识别出标记错误的实例并对其进行过滤.奖励设计与 Feng 等人不同,Feng 等人的奖励是从关系分类概率中计算得到,而该模型的奖励是直接通过关系分类器的分类效果,即 $F1$ 值的变化计算得到.分类效果度量值 $F1$ 的变化直观地反映了奖励,这为奖励函数的设计提供了一种新的思路.不同于利用经典分类方法完成关系分类的工作,Zeng 等人[114]利用强化学习智能体直接进行关系抽取,提出了一种将远程监督与强化学习相结合来训练关系抽取器的模型 PE+REINF(position enhanced REINFORCE).模型首先利用 PCNN 对句子进行特征提取作为智能体的状态;其次,定义所有包含同一个实体对的句子集合为一个句子包.关系抽取器采用基于策略的 REINFORCE 算法,读取包中的句子,预测其关系作为动作.根据包中大多数句子的关系预测包的关系标签,并与真实包的关系标签进行比较,以最大化长期奖励.PE+REINF 使用长期奖励来训练关系抽取器,利用 PCNN 对状态进行编码时融入位置因素可以提高关系抽取准确性.

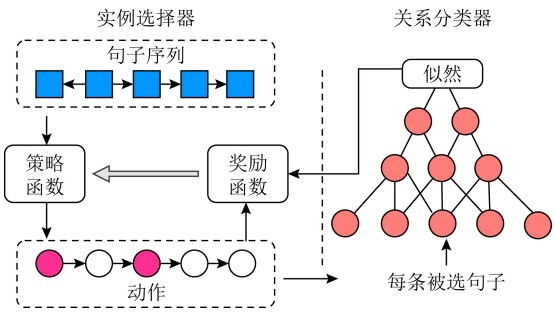


Fig. 5 Framework of RE based on reinforcement learning[112]

图 5 基于强化学习的 RE 算法框架图[112]

大多数关系抽取方法是将所有实体识别后再确定关系类型.与这类方法不同,Takanobu 等人[115]提出了一种关系实体联合抽取算法 HRL-RE,使用分层强化学习框架来增强实体抽取与关系检测之间的交互.整个抽取过程分为 2 层级强化学习模型,每层均采用基于策略的强化学习算法,分别用于关系检测和实体抽取.给定一条关系和实体均待标注的句子,首先,高层级强化学习智能体依据高层级状态依次对句子中的每个单词进行关系标注,其中,高层级状态由 Bi-LSTM 编码的当前隐层向量、关系类型向量、上一时刻状态构成.当某个单词被检测为某一关系的触发词后,就将关系信息传递到低层级强化学习智能体.低层级智能体依据低层级状态和已标注的关系类型对句子进行 NER 序列标注,其中,低层状态由 Bi-LSTM 编码的当前隐层向量、实体类型向量、上一时刻状态、由 MLP 编码的高层级状态向量构成.句子结束时,智能体的控制权将移返回给高层级强化学习智能体,以对下个词进行关系检测.在进行重叠关系(一个实体同时存在于 2 个或以上

关系中或同一实体对存在 2 个或以上关系)抽取时,可以基于不同关系类型为同一单词分配不同实体标签,增强关系检测和实体抽取的交互,提高关系抽取的性能.

3.3　知识推理

知识图谱通常是不完整的.知识推理是指根据知识图谱中已有的知识,采用某些方法,推理出新的知识,包括实体预测和关系预测.传统的推理方法,例如基于规则的推理[55-56]会引入一些人类先验知识,专家依赖度过高.目前,大部分知识推理是基于神经网络模型[58-64].神经网络模型通常更关注于推理结果,模型可解释性、可信赖性有待进一步提升.

除基于规则、基于神经网络的推理方法外,知识推理问题也可以建模成知识图谱中路径查找问题,其中节点代表实体,关系代表边.强化学习智能体根据当前环境(所在节点)通过策略学习或价值函数学习,来决定下一步的行动(通常为关系或(关系,实体)),从而完成推理任务.因此,基于强化学习的知识图谱推理在学术界和工业界得到广泛研究.基于强化学习的知识图谱推理方法依据智能体的个数可以分为单智能体推理方法、多智能体推理方法.多智能体推理方法指至少拥有 2 个智能体的基于强化学习的知识推理方法.多智能体之间存在着一定的关系,如合作、竞争或同时存在竞争与合作的关系.我们将分别从单智能体推理、多智能体推理 2 个方面进行详细介绍.

3.3.1　单智能体推理

单智能体推理指利用一个强化学习智能体进行推理的一类任务.问题的关键是构造合适的奖励函数以及用合适的方式对状态进行表示.

1) 奖励函数设计

通常,强化学习方法对奖励函数都非常敏感,奖励值的微小变化可能会导致推理性能的波动,因此奖励函数的设计非常重要.

DeepPath[116]是将强化学习应用于知识图谱推理的研究工作,利用强化学习解决大规模知识图谱中多跳关系路径推理问题(实体间除存在直接关系外还存在间接关系),为知识图谱推理提供了一种新思路.DeepPath 利用经典的基于翻译的知识图谱表示学习模型 TransE 或 TransH 得到实体和关系的向量表示.状态向量由当前节点表示向量、目标节点与当前节点表示向量的差构成.奖励设计考虑了可能影响智能体探索路径质量的 3 种因素:准确性、路径长度、路径多样性,人工设定奖励函数为三者的加权平均.DeepPath 采用基于策略的 REINFORCE 算法,使用一个全连接神经网络来参数化策略函数 $\pi(a|s,\theta)$,将状态映射到动作(关系)对应的概率分布中.训练智能体在知识库中寻找推理路径.与基于随机游走的路径查找模型相比,DeepPath 可以控制路径质量.Das 等人[117]提出 MINERVA (meandering in networks of entities to reach verisimilar answers).状态由查询节点、要查询的关系、当前所在节点、查询答案构成.动作空间由当前节点所有出边构成.利用基于策略的 REINFORCE 算法在知识图谱中进行搜索,以找到答案路径.在实现过程中,还为每个节点加入了自环以及反向关系,确保在找到正确答案后智能体可以采取“停止”操作以及撤消一个潜在的错误决策.与 DeepPath 不同的是,MINERVA 设计了一个随机历史依赖策略,由状态和动作构成的历史序列通过 LSTM 进行编码.策略网络采用 2 层全连接网络,输入为 LSTM 当前时间步隐层编码、当前节点表示、查询节点表示和关系向量,输出为可能的动作(关系)的概率分布.与 DeepPath 算法相比,推理性能得到了提升.但 MINERVA 假设推理路径一定存在,因此 MINERVA 无法应对当推理路径不存在的情况.

针对奖励稀疏的问题,Shen 等人[118]设计了一个名为 M-Walk 的智能体,它由一个深度循环神经网络(recurrent neural network, RNN)和蒙特卡洛树搜索(Monte Carlo tree search, MCTS)组成.首先,利用 RNN 编码历史路径信息,将历史信息、当前节点的邻居信息、当前节相连接边的信息作为状态,动作空间由当前节点的所有邻居节点构成.为了解决奖励稀疏的问题,M-Walk 采用改进的蒙特卡洛树搜索来生成路径,以产生更多正奖励的路径.由于生成的路径所采用的策略不同于原始策略,因此采用 Q-learning 算法进行 Q 值计算,进而更新策略参数,在 MCTS 路径生成和策略改进之间交替训练,以迭代地改进策略.基于 MCTS 的 M-Walk 能够产生具有更多积极奖励的路径,以缓解在图中游走的奖励稀疏的问题,提高推理的准确性.

很多强化学习推理算法采用硬奖励设计,即如果预测与真实数据一致则奖励记为 1,否则,记为 0 或 −1.然而,实际中针对不同的数据集设计合理奖励函数,可以获得更好的性能.上述工作[116-118]都在奖励方面进行了相应设计.除上述工作,Godin 等人[119]在知识图谱上使用强化学习进行知识问答,并指出现有基于强化学习的推理在问答任务上的奖励设计

的局限性,即简单地返回正确或不正确的答案是不全面的,还应该允许智能体对于不切实际的问题不予回答.针对这一问题,Godin 等人[119]对智能体奖励进行改进,将奖励分为 3 种类型:对正确答案给予肯定奖励;对不正确答案给予否定奖励;对不回答问题给予中立奖励 0.允许模型不回答问题.使用基于策略的 REINFORCE 算法,通过在 3 种回答可能性之间进行权衡,以获得最大累计奖励.此外,因为允许模型不回答,因此引入了新的性能指标回答率、精度以及二者的结合指标 QA 评分(可视为 $F1$ 评价指标的变体)来评价智能体的性能.并显著提高了回答的准确度.Lin 等人[120-121]对智能体奖励进行改进,提出了不使用基于智能体是否达到正确目标节点的二进制奖励,而是采用了一种软奖励机制,即利用基于嵌入的预训练模型 DistMult[68]或 ComplEx[69]的节点表示来计算正确目标节点与最终节点的相似性来作为无法确定正确性的目标实体的软奖励.此外,文献[120]为避免虚假路径误导模型,采用了一种执行动作退出机制,即在根据状态计算动作概率后,对每个动作应用伯努利分布采样,确定其是否被"屏蔽",以便对不同的路径进行有效的探索.通过对奖励函数的设计和对搜索空间进行更彻底的探索(引入动作退出机制)提高推理性能.

不同于上述人工设计奖励的方法,Li 等人[122]提出了一种自适应的强化学习算法 DIVINE(deep inference via imitating non-human experts),对于不同的数据集,可以自动调整奖励函数以逼近最佳性能,从而消除了额外的人工干预.DIVINE 模型包括一个生成对抗推理器和示例采样器.具体地,生成对抗推理器中的生成器是一个基于策略梯度的智能体,判别器是一个自适应的奖励函数.示例采样器用于自动地从知识图谱中抽取模仿示例.生成器用于生成推理路径,判别器被用来衡量生成的推理路径和模仿示例之间语义相似度,使用语义相似度来对生成器进行更新.判别器(自适应的奖励函数)从整体上对推理示例进行模仿,而不仅局限于其中包含的状态-动作对,从而引导智能体可以找到更多样化的路径,提升了推理准确性.大多数强化学习推理工作并没有考虑到为图中相同位置分配不同的奖励,例如,当智能体从不同的路径到达特定的位置时,都会以相同概率选择下一个动作,推理路径单一.针对这一问题,Tiwari 等人[123]提出一种距离感知奖励的强化学习算法 DAPath(distance-aware path),可以根据图中某一特定位置分配不同的奖励.人们通常认为在靠近目标实体的位置上采取的行动比之前采取的行动影响更大.基于这一假设,奖励由距离感知因子和全局奖励(1 或 −1)的乘积计算.距离感知因子考虑了路径长度和当前节点的位置 2 部分因素.当前节点越靠近目标实体,距离感知因子越大;当节点位置固定,路径长度越短,距离感知因子越大.模型中的策略网络使用图自注意力机制(graph self-attention, GSA)和门控循环单元(gate recurrent unit, GRU)的记忆机制,能够捕捉到路径邻域内更全面的实体和关系信息.通过应用距离感知因子,模型能够挖掘更可靠的路径以及发现一些常识性的推理路径.

2) 状态设计

针对强化学习的状态设计,以往的研究表明[124-125],融入节点类型信息有助于提高推理的准确度.针对现有强化学习没有对节点类型信息和节点在图中的拓扑结构建模的问题,Saebi 等人[126]提出了一种基于图神经网络类型增强的强化学习算法(type enhanced RL-GNN, TE RL-GNN).模型首先采用 mean/max pooling 算法计算融合了节点类型信息的节点表示.其次,利用 GCN 考虑知识图谱中关系和拓扑结构信息得到更丰富的节点表示.状态包括给定的查询实体及关系、智能体当前所在实体以及智能体所遍历的实体和关系的历史轨迹信息,使用 LSTM 编码.动作被定义为当前节点及其所有邻居节点.最后利用基于策略的强化学习算法 REINFORCE 进行学习.在强化学习状态表示中引入节点类型信息,有助于丰富节点表示、提升推理能力.

对于不同的查询关系,为了使智能体更多地关注与查询关系密切相关的关系和邻居信息,Wang 等人[127]和 Li 等人[128]利用了图注意力机制来编码节点的邻居信息,分别提出了 2 种基于强化学习的推理算法 AttnPath 和 MemoryPath.智能体当前的状态由当前所在实体信息、历史信息、当前实体邻居节点信息 3 部分组成.2 种算法都综合利用预训练模型 TransD,LSTM 和图注意力.其中,使用 TransD 来计算实体向量表示,利用 LSTM 编码历史信息,利用图注意力网络编码节点的邻居信息.动作与奖励的定义与 Lin 等人[120]类似,另外添加了对于无效动作的惩罚机制,即无效动作奖励为 −1.通过基于策略的强化学习算法 REINFORCE 控制节点(动作)选择.在状态设计中引入注意力机制,使得智能体更关注与查询关系密切相关的实体与关系,信息惩罚机制的引入能够避免智能体在毫无意义的状态

下停滞不前,提高推理效率.Wang 等人[129]提出一种基于自注意力的深度强化学习框架 ADRL(attention-based deep reinforcement learning).首先,利用 CNN 和 LSTM 来编码实体信息和智能体历史路径信息作为状态.其次,由于知识图谱中实体之间存在丰富的语义信息,ADRL 设计了一个基于自注意力机制的实体间的关系模块,依据这些实体和实体间关系的重要性来指导策略网络学习.LSTM 和注意力机制相结合,提高了推理的可解释性.

针对现有方法假设实体-关系表示遵循单点分布,但事实上,不同实体与关系可能包含不同的不确定性的问题,Wan 等人[130]提出了一个贝叶斯多跳推理范式 GaussianPath,旨在捕捉推理路径的不确定性.GaussianPath 使用高斯分布来表示一个实体或关系.通过训练智能体,高斯分布的后验概率将会收敛,从而减少实体或关系的不确定性.由于知识图谱中的状态-动作组合空间过大,难以直接得到 Q 函数.GaussianPath 使用 Bayesian LSTM 编码当前状态,使用了贝叶斯线性回归层来近似 Q 函数.通过知识图谱补全和实体链接等任务来验证推理性能,实验结果表明 GaussianPath 可以利用预先训练的高斯分布形式的先验知识,加速训练的收敛速度,提升推理结果的准确性.

3.3.2　多智能体推理

多智能体推理是利用 2 个或以上的智能体进行推理的一类工作.智能体之间可以存在合作、竞争或同时存在合作与竞争的关系.相比于单智能体强化学习,多智能体强化学习具有易于实现和易于任务分配的优点[131].

Li 等人[132]提出了一种基于多智能体强化学习的路径推理模型(multi-agent reinforcement learning based method for path reasoning, MARLPaR).模型共同训练2 个智能体,一个用于关系选择,另一个用于实体选择,如图 6 所示.关系选择智能体由三元组(S^{RS},A^{RS},π^{RS}) 构成,用于寻找特定查询关系,$S_t^{RS}=e_t$ 表示当前实体,$A_t^{RS}=\{r\in R:(e_t,r,v)\in E,e_t,v\in V\}$表示连接当前实体的所有关系,$E$ 代表三元组集合,π_t^{RS} 被定义为一个随机历史依赖策略.实体选择智能体与关系选择智能体类似,用于从关系的尾部实体集中选择最合适的实体,以准确地找到候选实体.2 个智能体进行交替训练,以最大化期望奖励.不同于以往通过单一策略网络进行关系或(关系,实体)选择的工作,MARLPaR 分别采用 2 个策略网络进行实体和关系选择,当路径选择过程中出现一对多或多对多关系时,充分考虑了实体选择的重要性,利用智能体的合作提升推理性能.

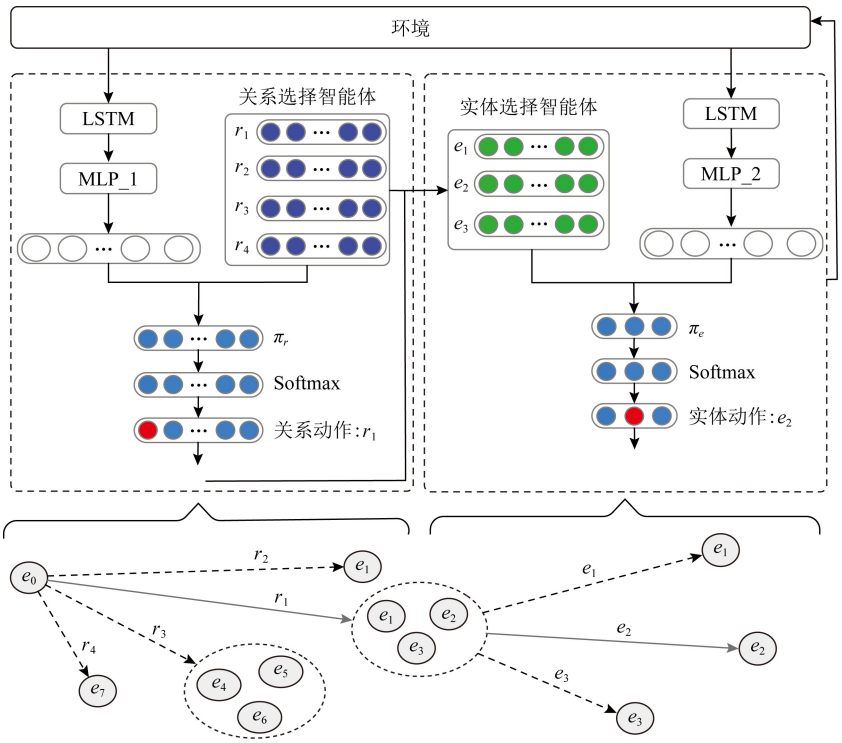


Fig. 6　MARLPaR reasoning method based on multi-agent reinforcement learning[132]

图 6　多智能体强化学习推理方法 MARLPaR[132]

不同于 Li 等人[132]的工作,受博弈论的启发,Hildebrandt 等人[133]提出了一种基于辩论动力学的知识图谱自动推理算法 R2D2(reveal relations using debate dynamics).其主要思想是将知识图谱补全问题转化为 2 个强化学习智能体之间的博弈过程.模型由 2 个智能体和 1 个判别器构成.具体来看,对于待查询三元组,2 个智能体分别寻找可以证明三元组为真和为假的证据,智能体间相互对立.判别器为一个二进制分类器,用于整合所有证据,并计算最终的置信概率以及预测三元组最终得分.模型采用交替训练的方式,每一次仅训练智能体或判别器,以最大化累积奖励的期望.与其他黑箱方法相比,这 2 个智能体之间的辩论博弈不仅提升了推理性能还为知识图谱推理提供了可解释的依据.

大多数传统方法假设目标实体和源实体之间的距离很短.然而,在真实知识图谱上并非如此.Zhang 等人[134]在推理中引入了外部语料库,提出了一个基于知识图谱和外部语料库的协同推理框架.模型由推理智能体、信息提取智能体构成.推理智能体根据当前节点状态选择下一步的动作.信息提取智能体对语料库的知识进行排序,将最高排名的三元组加入知识图谱中.设计了 4 种奖励机制:1)完全合作,成功完成任务后 2 个智能体可获得相同的奖励;2)合作,但信息提取智能体的奖励取决于推理智能体是否采纳其建议;3)软奖励机制,即使代理没有得到最终答案,也会分配软奖励,奖励通过输出实体和答案实体之间向量表示的余弦相似度计算得到;4)博弈,每个智能体都希望将其成本降到最低.推理智能体的目标是最小化其推理跳数.信息提取智能体的目标则是最小化其提案被推理智能体拒绝的次数.针对前 3 种奖励策略与第 4 种奖励策略,分别采用遗忘算法(forget algorithm)和响应目标对手算法(respond to target opponents algorithm)来学习策略网络.该方法为多智能体强化学习推理提供了一个通用结构框架,并且能够捕获实体之间的长距离关系,适用于大规模知识图谱推理任务.

3.4 知识表示

知识图谱在表示结构化数据方面非常有效,但这种三元组的基本符号性质使知识图谱难以操作[135].为了解决这一问题,提出了知识表示学习[1].知识表示学习旨在将知识图谱丰富的结构和语义信息嵌入到低维节点表示中.目前,常用的知识表示学习方法[1]有基于翻译模型 Trans 系列的方法[69-71]、基于语义匹配的方法[68-69]、基于神经网络的方法[71-76].基于翻译模型的方法简单易于理解,但是基于翻译模型的方法不能处理复杂关系,模型复杂度较高.基于语义匹配的方法需要大量参数且复杂度较高,限制了其在大规模稀疏知识图谱上的应用.基于神经网络的方法虽然建模能力较强,但是结果缺乏一定的可解释性.基于图的随机游走模型[136-138]也是用于知识表示学习的一类方法.这类方法依赖于人工设置元路径来捕获图的语义信息.然而,人工设置元路径需要丰富的专家领域知识,对于大规模、复杂且语义丰富的知识图谱来说,是一件充满挑战的任务.

Zhong 等人[139]针对需要人工设定元路径的问题,提出了一种基于异质图神经网络的强化学习的算法(reinforcement learning based on heterogeneous graph neural networks, RL-HGNN).模型由 2 部分构成,即强化学习智能体模块和图神经网络模块.强化学习智能体模块利用基于价值的 DQN 算法,学习根据当前状态选择动作的策略,生成路径实例;图神经网络模块对生成的路径实例进行信息聚合以学习节点表示.并将更新后的节点表示应用于下游任务中,利用下游任务的性能改进计算奖励,对动作值函数的估计 Q 进行优化,提高生成元路径的质量和效率.但是文献[139]没有全面考虑一个实体可能属于多个类型,例如 Obama:{Writer,President,Activist,Person},而这些信息可以提供对图中节点之间关系的丰富语义解释.因此,Wan 等人[140]提出基于强化学习的元路径发现算法(meta-path discovery with reinforcement learning, MPDRL).模型包含 2 部分:路径实例生成和元路径生成.具体来看,MPDRL 利用经典的策略梯度 REINFORCE 算法,通过 GRU 编码历史信息.策略网络采用 2 层全连接网络,输入为 GRU 当前时间步隐层编码、当前状态信息,输出为可能的动作(关系)的概率分布,从而推断从源节点到目标节点的路径,生成大量路径实例.其次,在路径实例基础上,采用 LCA(lowest common ancestor)算法,为节点分配实体类型,自动生成元路径.MPDRL 可以在大规模的异质信息网络中,自动挖掘出被人类专家忽略的有用元路径.将这些元路径应用于下游任务(链路预测)中,实验表明将这些元路径应用于下游任务中,显著提高了下游任务的性能.

3.5 知识融合

知识图谱中的知识来源广泛,具有多源、异构等特点,需要构建统一的大规模知识库来支撑推理和理解任务.知识融合研究如何将来自多个来源的关于同一个实体或概念的描述信息融合起来[11],形成高

质量统一的知识图谱的一类任务.通常,知识融合包括本体匹配(ontology matching)、本体对齐(ontology alignment)、实体链接(entity linking)、实体消歧(entity disambiguation)、实体对齐(entity alignment)等.现有的知识融合方法还存在受噪声数据以及对齐种子对数量的限制[141],或者未能充分建模实体之间的相互依赖关系等问题.

为了充分建模实体间的相互依赖关系,Fang 等人[142]将实体链接任务建模成一个序列决策问题,利用一篇文章或一个句子前面提到的实体提供的信息来消除后面提到的实体歧义,提出了一个端到端的强化学习模型 RLEL.如图 7 所示,模型由 3 部分组成,局部编码器、全局编码器、实体选择器.局部编码器编码指称和候选实体的局部特征,以获得潜在向量表示.具体地,对于每个指称及其候选实体,局部编码器先利用 LSTM 对指称的上下文进行编码.再利用 LSTM 编码候选实体的描述信息与预训练得到的实体嵌入拼接.为了丰富词汇和统计特征,对包括实体的流行度、实体描述与提及上下文之间的编辑距离等特征进行编码.全局编码器由一个 LSTM 构成,对指称和时刻 0～t 所选择的实体进行编码,输出历史决策信息.状态由局部编码器得到的当前信息和全局编码器得到的历史决策信息组成.动作空间由当前指称所指向的所有可能的目标实体构成.实体选择器采用经典的策略梯度 REINFORCE 算法,从候选实体集中选择目标实体.实体选择器的策略网络不仅考虑当前指称及候选实体,还充分利用先前相关实体的信息来消除歧义,并探索了当前选择对后续决策的长期影响,从全局的角度做出决策,避免错误传播,提升了实体链接的效果.同样 Zeng 等人[143]也利用强化学习进行序列决策的思想来建模实体链接问题,提出了具有特征框架的集体实体对齐(collective entity alignment with features framework, CEAFF).CEAFF 采用 Actor-Critic 算法.考虑到实体链接中的决策连贯性(依据相似度最大值匹配)和排他性(1 对 1 约束),状态由局部相似性、排他性和连贯性 3 部分构成,局部相似性考虑了当前源实体与候选目标实体之间的相似性,排他性与目标实体相关,利用 one-hot 向量表示目标实体是否已被选择,连贯性考虑到当前候选目标实体和前面选择的目标实体之间的相关性.动作空间由当前源实体所对应的所有目标实体构成.Actor 与 Critic 均采用 2 层 MLP,模型通过充分考虑实体之间的相互依赖关系提升实体对齐任务的性能.

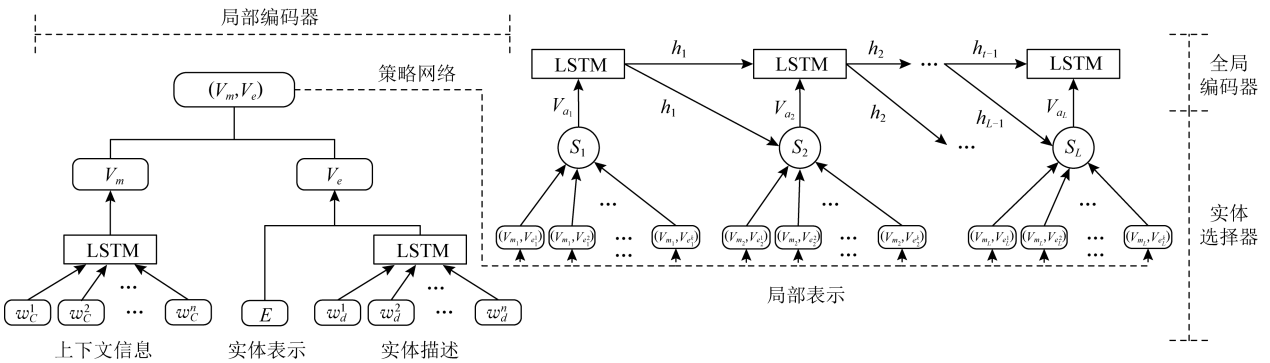


Fig. 7　Entity linking model RLEL based on reinforcement learning[142]

图 7　基于强化学习实体链接模型 RLEL[142]

4　基于强化学习的知识图谱的应用

知识图谱可以为各领域提供丰富的信息和先验知识,强化学习方法拥有强大的探索能力和自主学习能力.基于强化学习的知识图谱相关技术能够降低噪声数据的干扰、自动选择高质量的样本数据、更好地理解环境和提供可信解释.因此,基于强化学习的知识图谱在很多领域得到应用.强化学习与知识图谱的结合,从结合方式上来看,可以分为 2 类.1)将实际问题建模为包含多种节点类型和关系类型的知识图谱,强化学习在知识图谱上进行探索学习策略.2)将知识图谱作为外部信息引入强化学习框架中,用来指导强化学习的探索过程.我们将介绍如何将知识图谱与强化学习结合解决实际应用中的问题,包括智能推荐、对话系统、游戏攻略、生物医药、金融、网络安全等.

4.1　智能推荐

1) 单步推荐

推荐系统常面临数据稀疏、可解释性等问题以

及个性化定制、新型推荐任务等新的需求.知识图谱可以向推荐系统中引入辅助信息,如商品内容、跨领域信息等.与常用的推荐方法不同,基于强化学习的知识图谱推荐是在知识图谱中探索图(路径查找)来找到从用户到商品的有意义的路径.强化学习智能体在探索过程中进行决策,解决数据稀疏,提高推荐可解释性,使得推荐结果更符合用户需求.

推荐中存在一些商品或项目几乎很少有用户或根本没有用户交互.针对数据稀疏和冷启动问题,Song 等人[144]提出了 Ekar 算法(explainable knowledge aware recommendation),将推荐问题建模成图上的路径推理问题.具体来看,该方法将用户-商品交互图与商品-实体知识图谱通过商品相关联形成用户-商品-实体图.目标用户被定义为初始状态,所连接的图上的节点构成定动作空间.奖励设计与 Lin 等人[120]的想法一致.采用经典的策略梯度 REINFORCE 算法来训练一个策略函数,用于决策下一步动作的选择,最终形成一条用户到商品的路径完成推荐.Ekar 通过引入知识图谱作为补充信息缓解了数据稀疏的问题,提高了推荐的准确度和效率.数据稀疏性还表现在负样本的缺乏.目前的工作大多对未观测到的数据进行负采样.然而,无论是静态负采样策略还是自适应负采样策略,都不足以产生高质量的负样本.因此,Wang 等人[145]提出了一种新的负采样算法 KGPolicy(knowledge graph policy network).模型采用经典的策略梯度 REINFORCE 算法,状态定义为用户和智能体当前所在节点.动作被定义为一个 2 跳路径.设计了一个邻居注意力模块,该模块指定了 1 阶和 2 阶邻居的不同重要性,以便自适应地捕获对节点的偏好,并产生潜在的负例.KGPolicy 可以从与正样本的交互中自适应接收带有知识的负信号,产生潜在的负样本来训练推荐模型,提高模型处理缺失数据的能力,提升了模型的准确度.

现有推荐系统更关注于推荐结果的准确度,结果常常缺乏可解释性或仅能提供事后可解释性(post-hoc explanation).Xian 等人[146]提出 PGPR 算法(policy-guided path reasoning algorithm).采用经典的策略梯度 REINFORCE 算法,并引入状态值函数以减小方差.状态包括用户、智能体当前所在节点、智能体历史路径信息.动作被定义为与当前节点所连接的(关系,实体)对.最终形成一条用户到商品的路径完成推荐.但由于奖励信号的稀疏性以及知识图谱中动作空间巨大,这种试错性方法具有较差的收敛性.因此,Zhao 等人[147]提出了一个基于知识引导的推理框架 ADAC(adversarial Actor-Critic).模型由 Actor、Critic、路径判别器和元路径判别器 4 部分构成,通过 3 种启发式的策略:最短路径、元路径、用户感兴趣的实体,提取满足要求(较少标注、可解释、准确)的演示路径.具体来看,Actor 用于生成路径,路径判别器、元路径判别器用于判断 Actor 所生成的路径是否符合启发式的策略并给出奖励.Critic 用于估计动作值函数 Q,采用时间差分(TD)方法来学习 Critic 网络.最后,4 部分进行联合优化,通过知识引导限制强化学习智能体的探索过程加速训练过程,提升推理的性能.

在个性化推荐领域,如个性化学习路径推荐需要为用户按顺序推荐个性化的学习项目,例如课程、讲座等,以满足每个学习者的独特需求.针对现有工作没有同时建模认知结构(例如学习者的知识水平)和所学项目的知识结构(例如项目之间的先修关系)的问题,Liu 等人[148]提出了一种认知结构增强的自适应学习算法 CSEAL(cognitive structure enhanced framework for adaptive learning).模型采用经典的 Actor-Critic 算法,状态由学习目标和当前知识水平 2 部分构成,其中,学习目标采用 one-hot 编码,若是最终学习项目记为 1,否则记为 0,知识水平利用 LSTM 来编码.动作由当前学习项目所连接的所有学习项目构成.奖励被定义为学习周期结束后成绩的变化.Actor 为一个策略网络,用于根据当前状态计算所有动作对应的概率分布,Critic 为一个价值网络,用于对状态值函数进行估计.CSEAL 综合了个性和共性 2 方面信息,不仅考虑到用户的学习能力还考虑了知识体系结构,设计了一种基于知识结构的认知导航算法,以确保学习路径的逻辑性,减少了决策过程中的搜索空间,从而为用户制定个性化的学习方案.

顺序推荐(sequential recommendation)旨在根据用户的顺序交互行为,依次推荐下一个或接下来的几个商品.在这类推荐中,奖励函数不仅应该考虑单个预测的性能,还需要根据推荐序列来衡量整体性能.现有的深度学习的方法仅关注当前所推荐商品的准确性,并未考虑该商品对于推荐序列长期的影响.因此,Wang 等人[149]将知识图谱引入基于强化学习的顺序推荐场景中,提出了 KERL 算法(knowledge-guided reinforcement learning model).KERL 采用经典的策略梯度 REINFORCE 算法,状态包含了历史交互商品序列信息、用户当前偏好信息、预测出的用户未来偏好信息.具体来看,首先使

用 TransE 得到商品的向量表示.然后,使用 GRU 编码历史交互商品序列,使用 mean pooling 算法来聚合用户已经交互过的商品向量作为用户当前偏好,未来偏好基于当前的偏好,使用一个 MLP 来直接预测得到.动作空间由当前智能体所在节点的邻居节点构成.奖励函数由时序级奖励和知识级奖励 2 部分构成.其中,时序级奖励使用翻译中经典的 BLEU 算法度量实际交互的商品子序列和强化学习预测的商品子序列的相似性,知识级奖励由余弦相似度来度量预测的用户兴趣偏好序列和真实偏好序列的相似性.KERL 赋予了时序预测模型考虑推荐商品长期收益的能力,实现知识对强化学习探索过程的指导,提高了顺序推荐的准确度.

2) 多步推荐

随着抖音、快手和各类自媒体移动应用程序的广泛使用,新的推荐场景不断涌现出来.交互式推荐(interactive recommendation, IR)和对话式推荐(conversation recommendation, CR)受到了广泛的关注.与单步推荐不同,交互式推荐与对话式推荐系统是一个多步决策的过程.在每一步中,系统向用户推荐一个商品或者询问用户对于某种属性的偏好,并从用户那里接收反馈,这些反馈会影响下一步的推荐决策.推荐-反馈交互重复进行,直到用户访问会话结束.因此,这类任务可以很自然地利用强化学习来进行建模.在这类场景下推荐系统需要与用户进行多次信息交流,以此获得更多有利于明确用户兴趣偏好、真实意图和实际需求的信息.一种常见的方法是利用大量的辅助数据(如社交网络、知识图谱)来更好地解释用户意图[150-151].Zhou 等人[152]提出利用知识图谱的先验知识进行基于强化学习的交互式推荐算法 KGQR(knowledge graph enhanced Q-learning framework for interactive recommendation),如图 8 所示.图 8 中模型包含 4 部分:图卷积网络模块、状态表示模块、候选集选择模块和 Q-network 模块.模型采用经典的基于价值的 Dueling DQN 算法,状态被定义为用户交互的商品序列.首先图卷积网络模块通过图卷积神经网络 GCN 编码知识图谱中商品的语义相关性和拓扑结构信息,作为商品的表示向量;然后状态表示模块再利用 GRU 编码用户交互的商品序列作为最终状态,输入 Q-network 中.动作空间被限定为用户交互过商品的 k 阶邻居.KGQR 不仅考虑到了用户偏好的时序性,而且通过知识图谱结构信息精剪动作空间,显著提高强化学习采样效率.

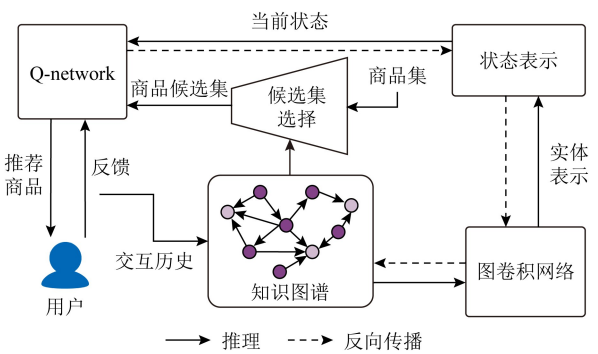


Fig. 8　Framework of KGQR[152]

图 8　KGQR 模型框架图[152]

在对话式推荐系统中,通常对用户商品数据构建用户-商品-属性知识图谱,一段对话可以被表示为图上的一条路径.强化学习方法用于决定下个动作(商品推荐或属性询问),帮助系统学习一个多回合的对话策略.Lei 等人[153]提出了一种对话路径推理算法 CPR(conversational path reasoning).模型采用经典的基于价值的 Deep Q-learning 算法,状态包括对话历史信息和候选商品数量 2 部分.动作只有 2 种选择,即“商品推荐”或“属性询问”. CPR 定义用户-属性偏好和用户-商品偏好用于决定推荐哪个商品或选择哪个属性进行询问.其中用户-属性偏好由信息熵建模,用户-商品的偏好由 EAR[154]中的 FM 变体计算.如果智能体选择属性询问,则直接依据用户-属性偏好从候选属性集中选择得分最高的属性.如果选择商品推荐,则依据用户-商品偏好从候选商品集中选择 top-k 的商品进行推荐.CPR 将对话推荐问题建模为基于图的路径推理问题,提高了对话推荐的可解释性,策略网络只需要决定何时询问和何时推荐,将动作空间减少到 2 个,减轻了策略网络的建模负担.与 Lei 等人[153]类似,Deng 等人[155]提出了 UNICORN 算法(unified conversational recommender).模型采用经典的基于价值的 Dueling DQN 算法.模型先通过图卷积神经网络 GCN 编码动态加权图中商品的语义和结构信息,作为商品的表示向量.但与 Lei 等人[153]不同,UNICORN 采用 Transformer 来编码用户历史对话序列信息作为状态.动作空间由商品集合和属性集合构成.实验表明,UNICORN 能够在更短的对话轮数下,了解用户偏好并为用户推荐合适商品.除了提高推荐的效率和准确性以外,基于知识图谱与强化学习相结合来进行对话推荐,还通过给出知识图谱推理路径,增加了对话推荐的可解释性.

4.2 对话系统

自然语言处理领域的一个重要应用就是人机对话系统,它是人机交互的核心功能之一.计算机想要顺利无障碍地和人类交流,必须具备丰富的背景知识以及强大的决策能力,结合背景知识去理解对话内容,从外部知识库中找出相关的知识并进行推理,从而生成合适的回答.知识图谱为对话系统提供了丰富的背景知识,而强化学习可以从相关知识集合中选出恰当的知识,并且可以利用用户的隐式负反馈信息,确保对话效果持续稳步提升.

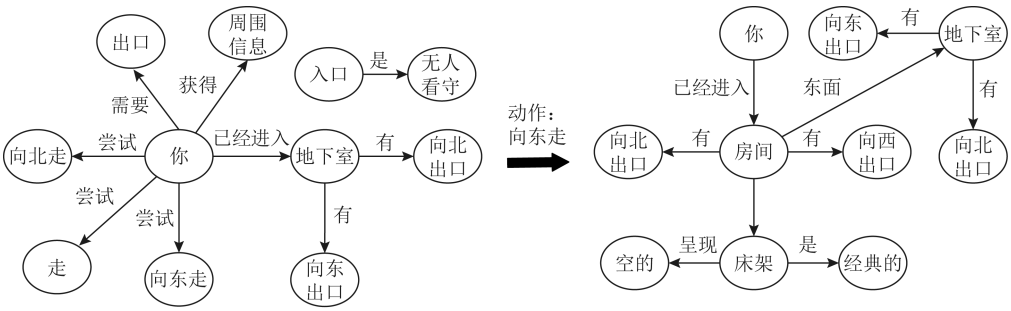
针对开放域对话和知识问答,大多数的研究只是利用用户输入和知识之间的相似度,将知识整合到了回复的生成过程中.然而,只依靠相似度是不能保证回复的合适准确的,例如人类在进行对话和知识问答时还需要借助大脑的决策能力,从相关知识中筛选出合适的知识进行回答.徐聪[156]提出了一个结合知识库的对话模型,将强化学习算法用于对相关知识的有效选择决策中.模型包括 2 部分:知识决策部分与回复生成部分.知识决策部分为对知识的粗选和精选.粗选指根据会话主题从完整的知识图谱中检索出所有相关子图作为候选集合;精选指利用强化学习中的策略梯度算法 REINFORCE,在子图中选择最合适的知识作为最优知识,其中,状态包含当前智能体所选择的知识、初始话题实体、目标话题实体、输入文本序列以及对话历史等 5 方面信息.动作被定义为图中与当前话题实体相连的所有边或切换话题.奖励被定义为真实回复和所选择知识的相关性.回复生成部分利用 Transformer 网络对输入文本和知识进行编码和解码.借助于强化学习的决策能力,该模型能够选择合适的知识并生成上下文连贯、回复内容准确、易于用户理解以及形式多样的回复.模型采用 $F1$ 值来评估模型的知识决策的精度-召回率,即输出回复相对于标准回复在字级别上的精度-召回率.在百度 Knowledge Driven Dialogue 数据集上 $F1$ 值提升了 5.28%.

现有 QA 方法只能从基准数据的显示问答对中学习.然而,用户很少会明确地将答案标记为正确或错误.针对这一问题,Kaiser 等人[157]提出了 Conqer 算法(conversational question answering with reformulations),该方法可以利用用户的重新提问这种隐式负反馈进行学习.模型采用经典的基于策略梯度的 REINFORCE 算法.状态包含用户当前所提的问题、用户最初的问题以及当前问题所关联的实体.动作被定义为当前问题所关联的实体所对应的边.奖励取决于用户的新问题与原问题是否同属一个主题.具体地说,首先将用户提出的问题所包含的实体通过实体消歧方法链接到外部知识库,将回答过程建模为多个智能体在知识图谱上并行游走过程,节点选择由策略网络输出的动作决定.该策略网络将当前所提的问题、用户最初的问题以及当前问题所关联的实体作为输入,通过分析用户新的提问与原来的问题相比是否表达出新的意图获得奖励进行训练.Conqer 是在用户提问未得到理想答案后重新提问时,利用隐式负反馈信息来学习对话策略的工作.实验表明:Conqer 优于当时最优方法 Convex[158],并且对各种噪声具有鲁棒性.

4.3 游戏攻略

文字类冒险游戏是一种玩家必须通过文本描述来了解世界,通过相应的文本描述来声明下一步动作的游戏.这类游戏中强化学习智能体根据接收到的文本信息进行自动响应,以实现规定的游戏目标或任务(例如拿装备、离开房间等).强化学习善于序列决策,知识图谱善于建模文本的语义和结构信息.因此,强化学习和知识图谱相结合在文字类冒险游戏中得到了成功的应用.基于强化学习的知识图谱方法在进行游戏策略学习时主要思路可分为 2 类:1)将游戏状态构建成一张知识图,利用强化学习技术进行游戏策略学习;2)将知识图谱作为外部知识辅助强化学习智能体进行决策.

文献[159-160]将每个时刻游戏中的状态表示为一张知识图谱,利用图结构特性以及图中的信息传递进行状态的表示学习.Ammanabrolu 等人[159]提出了一个基于深度强化学习的游戏策略学习算法 KG-DQN,它将每一时刻的游戏状态(文本描述)表示为一张状态图.采用图的形式有利于修剪动作空间,以实现更有效的探索.玩游戏时,智能体接收对当前游戏状态的观察(文本描述),根据给定的观察对状态图进行更新,如图 9 所示.采用 SBLSTM (sliding bidirectional LSTM)编码观察,同时利用图注意力机制对状态图进行编码.智能体每一时刻的状态可以通过对观察的编码和状态图的编码进行线性变换得到.采用 Q-learning 算法学习在当前状态下采取行动的策略.但 KG-DQN 的动作空间仍然很大,训练成本仍然较高.针对游戏动作空间巨大的问题,Ammanabrolu 等人[160]提出了利用知识图谱表示状态,通过预定义模板生成动作空间的算法 KG-A2C.模型由状态编码模块、动作解码模块 2 部分构成.状态的定义考虑到了观察的文本描述(包括当前



当前状态描述：你进入了地下室.你通过观察来获取周围环境的信息.你需要一个无人看守的出口吗?你应该尝试向东走.你不喜欢门?为什么不尝试向北走,那个入口无人看守.

下一时刻状态描述：你进入了一个房间.你可以看到一个床架,床架是经典的,床架似乎是空的.北边有出口.别担心,它已解锁.向西有一个畅通的出口.

Fig. 9　States updating in text adventure game[159]

图 9　文字冒险游戏中的状态图更新[159]

状态的环境描述、游戏需求、游戏反馈以及前一步采取的动作)、原始分数和状态图(与文献[159]类似).状态编码模块根据所有的观察数据利用多个 GRU 和多头图注意力机制编码游戏当前状态.动作解码模块利用状态编码信息通过多个 GRU 解码动作.模型训练阶段采用 Valid Action 检测算法[161]对动作空间进行了精简,利用优势演员评论家算法 A2C(advantage Actor-Critic)学习策略.KG-A2C 把组合巨大的行动空间限制在较小合理的动作空间中,提高了智能体的学习效率.

此外,针对应用于文本类游戏的强化学习智能体缺乏人类所具有的游戏推理能力.Xu 等人[162]提出基于堆叠分层注意力机制算法 SHA-KG(stacked hierarchical attention with knowledge graphs).利用知识图谱的结构信息进行显示推理,帮助智能体做出决策.状态包括观察的文本描述、原始分数和状态图 3 部分.不同于以往将游戏状态图表示成为一张图的工作,SHA-KG 对状态图依据不同的关系以及时间顺序进行子图划分,采用分层注意力机制对不同层级的特征进行提取,观察的文本描述利用 GRU 进行编码,与原始分数共同形成状态的表示.SHA-KG 采用优势演员评论家算法 A2C.通过将整个知识图谱划分为多个子图,并采用分层注意机制给出的不同层级的评分,帮助人类更好地解释智能体的推理依据和决策过程.Adhikari 等人[163]规范化了上述工作的状态图的概念,将其定义为信念图(belief graph),即在探索过程中学习到的图结构,提出了 GATA(graph-aided transformer agent)方法.模型采用经典的基于价值的 DQN 方法.信念图中节点可以表示玩家、物品、位置以及一些条件(例如,对于烹饪类游戏的开关、切片等操作),关系可以表示在特定时间下实体之间的关系(例如,在…正北方).GATA 是一个基于 Transformer 的智能体,首先通过对原始文本描述构造信念图作为状态,并基于该状态应用策略网络进行动作选择.然后,根据该动作和新的观察动态更新信念图.利用图结构形式的结构化表征可以提高强化学习智能体的可解释性,使决策过程更加透明.

现实生活中有些游戏还需要一些额外的常识知识作为补充信息.在这类游戏环境中,智能体需要借助于外部常识知识,例如,苹果放在冰箱中,盘子放在碗橱里,进而更好的完成游戏任务.在现有强化学习算法中引入外部知识,有利于减少强化学习智能体的动作空间,提高智能体的训练速度.Ammanabrolu 等人[164]进一步扩展了 KG-DQN 算法[159],探索了文字类游戏中游戏策略迁移的方法.模型利用了从游戏文本中提取出的知识图谱为强化学习智能体在同类游戏间的迁移提供先验知识.采用 DQN 网络参数权值有效地迁移知识.知识图谱通过为智能体提供不同游戏的状态和动作空间之间更明确且可解释的映射,能够在智能体间进行有效的迁移,以达到减少训练时间并提高所学习策略质量的目的.Murugesan 等人[165]将外部知识图谱 ConceptNet 作为补充信息应用在基于强化学习的文本游戏类任务中,提出了 Belief+KG_Evolve 算法.模型共包含 3 个部分,输入数据编码模块、基于图的知识融合模块、动作预测模块.输入数据编码模块将历史动作和游戏观察利用预训练模型 GloVe 和 GRU 网络进行编码.基于图的知识融合模块将状态图和外部知识图谱进行整合,利用 ConceptNet Numberbatch 词向量表示

和图卷积网络 GCN 进行图的表示学习.动作预测模块将数据编码模块得到的编码、基于图的知识融合模块得到的图的表示以及动作候选集作为输入,输出动作概率.Murugesan 等人[165]指出常识知识可以帮助智能体高效和准确地行动,但太多的常识知识也会对智能体起到干扰.如何确定并过滤掉那些无用常识是一个值得研究的方向.

4.4 药物/疾病预测

在生物医药领域,药物合成、新材料发现、疾病预测等在科技迅速发展的今天显得日益重要,给社会发展和人们生活带来巨大变化.引入强化学习方法,可以利用智能体在知识图谱中的自动探索做出最优决策,同时找到的路径可以为反应物生成或者疾病预测提供可解释性依据.目前,基于强化学习的知识图谱技术已经被应用于发现新的药物或材料、化学反应物预测以及药物组合预测、疾病预测等领域.

同时结合高度复杂和不可微的规则,设计模型以找到所需特性的分子是一项具有挑战性的任务.You 等人[166]提出了图卷积策略网络(graph convolutional policy network, GCPN).GCPN 将分子图生成的问题建模为一个序列决策问题,即在一个具有化学感知的环境中迭代地向分子图添加子结构和边.将图表示学习、对抗训练(用于奖励设计)等技术融入强化学习框架中,采用经典的基于 Actor-Critic 框架的 PPO[94],以优化由分子属性目标和对抗性损失组成的奖励.研究结果表明,GCPN 生成的分子在 Penalized logP 指标上比原有方法高出 61%.针对化学反应产物预测(chemical reaction prediction)问题,Do 等人[167]提出了图变换策略网络(graph transformation policy network, GTPN).GTPN 采用 Actor-Critic 算法,状态被定义为包含输入反应物和试剂分子系统的标记图,动作由序列结束信号、反应物分子节点、新的关系类型组成.GTPN 采用图神经网络来表示输入的反应物和试剂分子,并使用强化学习来寻找最佳的化学键变化序列,将反应物转化为产物.GTPN 不需对图变换的长度或顺序做任何假设.实验结果表明,在大型数据集 USPTO 中,GTPN 比原有方法准确度提高了约 3%.基于电子健康记录(electronic health record, EHR)的药物组合预测(medicine combination prediction, MCP)可以帮助医生为复杂病患者开药.针对 MCP 研究要么忽略了药物之间的相关性,即 MCP 被定义为二元分类任务,要么假设药物之间存在顺序相关性,即 MCP 被定义为序列预测任务的问题.Wang 等人[168]考虑到还应考虑药物之间的相互作用,即患者用药安全,提出了一种基于图卷积的强化学习模型 CompNet(combined order-free medicine prediction network),模型将药物组合预测问题建模为一个 Markov 决策过程,即时刻 t 的药物选择取决于之前时刻 $t-1$ 中选择的药物.利用 Deep Q-learning 来学习药物之间的相关性和相互作用.状态由患者表征和医学知识图谱表征经过非线性变换得到.动作空间是由所有药物构成的集合.具体地,首先使用 Dual-CNN 来获取基于 EHR 的患者表征;然后,引入与预测药物相关的医学知识来创建动态医学知识图谱,使用关系图卷积网络 R-GCN 对其进行编码;最后,CompNet 通过融合患者信息和药物知识图谱来进行动作(药物)选择.实验结果表明,在数据集 MIMIC-III 中,CompNet 显著优于现有方法,Jaccard 和 $F1$ 值分别提高了 3.74% 和 6.64%.Sun 等人[169]将医学知识和医学数据相结合构建疾病知识图谱,将疾病预测任务建模在知识图谱上的游走问题.疾病知识图谱中节点表示疾病(例如冠心病),边表示疾病之间的关系(例如引发).状态定义同时考虑到了病人的信息和在知识图谱中历史游走的信息(当前所在节点以及走过的节点),动作空间是由智能体当前节点的所有邻居节点(疾病)构成的集合.采用 Actor-Critic 算法,学习智能体的游走策略.最终智能体所在的节点代表病人所患的疾病,游走的路径表示可解释的疾病进展路线,可作为预测病人所患疾病的解释性依据.实验表明在 MIMIC 数据集中疾病的预测准确率可达 63.9%.

4.5 其 他

除了推荐、对话系统、游戏、生物医药等领域,基于强化学习的知识图谱方法还可以应用于金融、网络安全等其他领域.

在金融领域,Miao 等人[170]将文献[112]所提方法应用于动态金融知识图谱构建的关系抽取任务中.实验表明,该方法可以降低噪声数据的干扰,提高关系提取模型的准确度.动态金融知识图谱可对大量金融数据进行标准化和可视化,辅助金融从业人员进行分析与决策.在网络安全领域,Piplai 等人[171]结合网络安全知识图谱,将强化学习算法应用于恶意软件检测.算法模拟安全专业人员使用自身背景知识来识别攻击,将从描述相同或相似的恶意软件攻击的文本中挖掘出的知识应用于强化学习算法的动作选择概率和奖励函数的设计中.实验表明,使用先验信息源的加权均值的奖励函数在恶意攻击检测中效果最好.

5　未来发展方向

近几年来,针对知识图谱和强化学习的相关研究已经成为人工智能领域的热点方向.知识图谱可以同时建模数据的拓扑结构和语义信息,强化学习是一种从试错过程中发现最优行为策略的技术[84],适用于解决贯序决策问题.知识图谱与强化学习的结合有利于提升训练样本质量,还有利于提高可解释性和可信赖性.但是,强化学习方法在知识图谱领域应用也存在一些不足,主要表现在2个方面:1)对强化学习状态的表示,文献[134]提到目前强化学习状态表示大多使用预训练得到的节点嵌入.然而,当知识图谱中增加新三元组时,节点的嵌入也需要重新训练,计算成本较大.文献[126]提到除了结构信息以外,节点的文本描述信息、层次结构的类型信息也十分重要.在知识图谱表示学习领域,文献[172]和文献[173]分别将文本描述信息、关系路径等信息,用于构建更加精准的知识表示.然而,这些方法还未广泛应用于强化学习状态的表示中.2)强化学习的奖励函数设计,与人工定义奖励函数相比,文献[122]和文献[147]已经开始尝试利用知识图谱中的信息结合抗性学习来生成自适应的奖励函数.如何自动生成更合理的奖励函数还有待进一步研究.

目前围绕强化学习与知识图谱结合的研究还处于起步阶段,有广阔的发展空间.未来值得关注5个方向:

1) 基于强化学习的动态时序知识图谱研究

随着应用的深入,人们不仅关注实体关系三元组这种简单的知识表示,还需要掌握包括逻辑规则、决策过程在内的复杂知识.目前基于强化学习的知识图谱研究主要围绕静态知识图谱.然而,知识随着时间的推移往往是动态变化的.如何利用强化学习在解决序列决策问题方面的优势,来建模知识图谱的动态性,学习知识图谱的变化趋势,解决实际应用中的复杂问题是一个值得研究的课题.Li等人[174]研究了动态时序知识图谱的时序推理问题.受人类推理方式的启发,CluSTeR(clue searching and temporal reasoning)包含线索搜索和时序推理2部分.线索搜索模块采用随机集束搜索算法,作为强化学习的动作采样方法,从历史事件中推断多条线索.时序推理模块使用基于R-GCN进行编码,并应用GRU进行时序预测,实现从线索中推理答案.

2) 基于强化学习的多模态知识图谱研究

面对越来越复杂多样的用户诉求,单一知识图谱已不能满足行业需求.多模态数据[11]可以提供更丰富的信息表示,辅助用户决策,提升现有算法的性能.目前,基于强化学习的知识图谱研究主要针对文本数据.如何利用强化学习技术进行多模态知识图谱的构建与分析仍是一个值得研究的方向.He等人[175]将强化学习方法应用于视频定位(video grounding),即给定一段文本描述将其与视频片段相匹配的任务中.He等人将这个任务建模为一个顺序决策的问题,利用Actor-Critic算法学习一个逐步调节时间定位边界的代理,完成视频与文本的匹配.

3) 基于新的强化学习方法的知识图谱研究

强化学习作为人工智能领域研究热点之一,其研究进展与成果也引发了学者们的关注.强化学习领域最近提出了一系列新的方法和理论成果,例如,循环元强化学习[176]、基于Transformer的强化学习[177]、逆强化学习[178]等相关的理论.如何将这些新的理论方法应用在知识图谱的构建或研究应用中,值得深入思考.Hou等人[179]在强化学习动作选择中引入了知识图谱中隐含的规则来约束动作选择,进一步精简了动作空间,提高了强化学习效率.Hua等人[180]提出了一种元强化学习方法来进行少样本复杂知识库问答,以减少对数据注释的依赖,并提高模型对不同问题的准确性.

4) 基于强化迁移学习的知识图谱研究

基于强化学习的知识图谱方法具有一定的可解释性和准确性.但强化学习不同于监督学习,样本数据来源于智能体与环境的交互,会导致收集大量无用且重复的数据,成本较高.一种解决思路是将迁移学习应用到强化学习中,通过将源任务学习到的经验应用到目标任务中,帮助强化学习更好地解决实际问题.文献[164]、文献[170]将迁移学习和强化学习结合起来,分别应用于同类游戏策略学习以及动态金融知识图谱构建领域,并取得了不错的效果,缓解了特定领域因训练数据不足所带来的挑战,提高了模型举一反三和融会贯通的能力.因此,基于强化迁移学习的知识图谱研究也是未来一个重要的研究方向.

5) 算法可解释性度量研究

由于知识图谱能够提供实体间的语义和结构信息,强化学习智能体的学习过程和人类认知世界的过程比较相似,产生的解释更易于人类理解.因此,一些研究者利用强化学习和知识图谱开展可解释性

的研究.然而,这些研究工作可解释性的效果只能通过实例分析来进行评测.目前,针对解释性还没有统一或者公认的衡量标准[84],如何衡量模型的可解释性是未来需要研究的问题之一.

6 总 结

知识图谱既包含图的拓扑结构信息又包含丰富的语义信息,得到越来越多研究者的关注.然而,目前知识图谱研究面临标注数据获取困难、模型依赖人工定义的规则和先验知识、方法缺乏可解释性等问题.环境驱动的强化学习方法学习过程更接近于人类认知,产生的解释更易于人类理解,具有十分重要的研究意义.本文首先简要介绍了知识图谱和强化学习的基础知识.其次,对基于强化学习的知识图谱相关研究,包括强化学习在知识抽取、知识推理、知识表示、知识融合等方面的研究进行了全面综述.最后,介绍了基于强化学习的知识图谱研究在智能推荐、对话系统、游戏、生物医药、金融、网络安全等领域的实际应用.在此基础上,对未来的发展方向,包括基于强化学习的动态知识图谱、基于强化学习的多模态知识图谱、基于新强化学习方法的知识图谱研究、基于强化迁移学习的知识图谱以及算法可解释性度量等进行了展望.

作者贡献声明:马昂负责调研并完成论文撰写;于艳华负责论文审阅,并给出详细修改指导意见;杨胜利、石川、李劼对论文提出指导意见;蔡修秀负责论文格式修订.

参 考 文 献

[1] Zhang Tiancheng, Tian Xue, Sun Xianghui, et al. Overview of research on knowledge graph embedding technology [J/OL]. Journal of Software, 2021 [2021-11-15]. http://www.jos.org.cn/1000-9825/6429.htm (in Chinese)
(张天成, 田 雪, 孙相会, 等. 知识图谱嵌入技术研究综述[J/OL]. 软件学报, 2021 [2021-11-15] http://www.jos.org.cn/1000-9825/6429.htm)

[2] Xiao Yanghua. Knowledge Graph Concept and Technology [M]. Beijing: Electronic Industry Press, 2020 (in Chinese)
(肖仰华. 知识图谱概念与技术[M]. 北京: 电子工业出版社, 2020)

[3] Ji Shaoxiong, Pan Shirui, Cambria E, et al. A survey on knowledge graphs: Representation, acquisition and applications [J]. arXiv preprint, arXiv: 2002.00388, 2020

[4] Wang Junping, Zhang Wensheng, Wang Yongfei, et al. Constructing and inferring event logic cognitive graph in the field of big data [J]. SCIENTIA SINICA Informationis, 2020, 50(07): 988-1002 (in Chinese)
(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[J]. 中国科学: 信息科学, 2020, 50(07): 988-1002)

[5] Wu Zongyou, Bai Kunlong, Yang Linrui, et al. Review on text mining of electronic medical record [J]. Journal of Computer Research and Development, 2021, 58(3): 513-527 (in Chinese)
(吴宗友, 白昆龙, 杨林蕊, 等. 电子病历文本挖掘研究综述[J]. 计算机研究与发展, 2021, 58(3): 513-527)

[6] Qin Chuan, Zhu Hengshu, Zhuang Fuzhen, et al. A survey on knowledge graph-based recommender systems [J]. SCIENTIA SINICA Informationis, 2020, 50(07): 937-956 (in Chinese)
(秦川, 祝恒书, 庄福振, 等. 基于知识图谱的推荐系统研究综述[J]. 中国科学: 信息科学, 2020, 50(07): 937-956)

[7] Chen Xiaojun, Jia Shengbin, Xiang Yang. A review: Knowledge reasoning over knowledge graph [J]. Expert Systems with Applications, 2020, 141: 112948

[8] Guan Saiping, Jin Xiaolong, Jia Yantao, et al. Knowledge reasoning over knowledge graph: A survey [J]. Journal of Software, 2018, 29(10): 2966-2994 (in Chinese)
(官赛萍, 靳小龙, 贾岩涛, 等. 面向知识图谱的知识推理研究进展[J]. 软件学报, 2018, 29(10): 2966-2994)

[9] Yao Siyu, Zhao Tianzhe, Wang Ruijie, et al. Rule-guided joint embedding learning of knowledge graphs [J]. Journal of Computer Research and Development, 2020, 57(12): 2514-2522 (in Chinese)
(姚思雨, 赵天哲, 王瑞杰, 等. 规则引导的知识图谱联合嵌入方法[J]. 计算机研究与发展, 2020, 57(12): 2514-2522)

[10] Niu Guanglin, Li Bo, Zhang Yongfei, et al. Joint semantics and data-driven path representation for knowledge graph inference [J]. arXiv preprint, arXiv: 2010.02602, 2020

[11] Zhao Xiaojuan, Jia Yan, Li Aiping, et al. A survey of multi-source knowledge fusion technology [J]. Journal of Yunnan University: Natural Sciences Edition, 2020, 42(207): 65-79 (in Chinese)
(赵晓娟, 贾焰, 李爱平, 等. 多源知识融合技术研究综述[J]. 云南大学学报: 自然科学版, 2020, 42(207): 65-79)

[12] Vakulenko S, Garcia J D F, Polleres A, et al. Message passing for complex question answering over knowledge graphs [C] //Proc of the 28th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2019: 1431-1440

[13] Li Jing, Sun Aixin, Han Jianglei, et al. A survey on deep learning for named entity recognition [J]. arXiv preprint, arXiv: 1812.09449, 2018

[14] E Haihong, Zhang Wenjing, Xiao Siqi, et al. Survey of entity relationship extraction based on deep learning [J]. Journal of Software, 2019, 30(6): 1793-1818 (in Chinese)
(鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019, 30(6): 1793-1818)

[15] Zeng Daojian, Liu Kang, Chen Yubo, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 1753-1762

[16] Li Dongmei, Zhang Yang, Li Dongyuan, et al. Review of entity relation extraction methods [J]. Journal of Computer Research and Development, 2020, 57(7): 1424-1448 (in Chinese)
(李冬梅, 张扬, 李东远, 等. 实体关系抽取方法研究综述[J]. 计算机研究与发展, 2020, 57(7): 1424-1448)

[17] Zhang Xue, Sun Hongyu, Xin Dongxing, et al. Survey on automatic term extraction research [J]. Journal of Software, 2020, 31(7): 2062-2094 (in Chinese)
(张雪, 孙宏宇, 辛东兴, 等. 自动术语抽取研究综述[J]. 软件学报, 2020, 31(7): 2062-2094)

[18] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey [J/OL]. Journal of Artificial Intelligence Research, 1996, 4: 237-285 [2021-11-15]. https://arxiv.org/pdf/cs/9605103.pdf

[19] Moerland T M, Broekens J, Jonker C M. Model-based reinforcement learning: A survey [J]. arXiv preprint, arXiv: 2006.16712, 2020

[20] Liu Jianwei, Gao Feng, Luo Xionglin. Survey of deep reinforcement learning based on value function and policy gradient [J]. Chinese Journal of Computers, 2019, 42(6): 1406-1438 (in Chinese)
(刘建伟, 高峰, 罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J].计算机学报, 2019, 42(6): 1406-1438)

[21] Liu Quan, Zhai Jianwei, Zhang Zongchang, et al. A survey on deep reinforcement learning [J]. Chinese Journal of Computers, 2018, 41(1): 1-27 (in Chinese)
(刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27)

[22] Sun Changyin, Mu Zhaoxu. Important scientific problems of multi-agent deep reinforcement learning [J]. Acta Automatica Sinica, 2020, 46(7): 1301-1312 (in Chinese)
(孙长银, 穆朝絜. 多智能体深度强化学习的若干关键科学问题[J]. 自动化学报, 2020, 46(7): 1301-1312)

[23] Xiong Luolin, Mao Shuai, Tang Yang, et al. Reinforcement learning based integrated energy system management: A survey [J]. Acta Automatica Sinica, 2021, 47(10): 2321-2340 (in Chinese)
(熊珞琳, 毛帅, 唐漾, 等. 基于强化学习的综合能源系统管理综述[J]. 自动化学报, 2021, 47(10): 2321-2340)

[24] Liang Tianxin, Yang Xiaoping, Wang Liang, et al. Review on financial trading system based on reinforcement learning [J]. Journal of Software, 2019, 30(3): 845-864 (in Chinese)
(梁天新, 杨小平, 王良, 等. 基于强化学习的金融交易系统研究与发展[J]. 软件学报, 2019, 30(3): 845-864)

[25] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a Web of open data [C] //Proc of the 6th Int Semantic Web Conf. Berlin: Springer, 2007: 722-735

[26] Mahdisoltani F, Biega J, Suchanek F M. YAGO3: A knowledge base from multilingual Wikipedias [C/OL] //Proc of the 7th Biennial Conf on Innovative Data Systems Research. New York: ACM, 2015 [2022-03-08]. https://pure.mpg.de/rest/items/item_2077946/component/file_2077968/content

[27] Liu H, Singh P. ConceptNet: A practical commonsense reasoning tool-kit [J]. BT Technology Journal, 2004, 22(4): 211-226

[28] Vrandecic D, Krötzsch M. Wikidata: A free collaborative knowledgebase [J]. Communications of the ACM, 2014, 57(10): 78-85

[29] Tang Jie, Zhang Jing, Yao Limin, et al. ArnetMiner: Extraction and mining of academic social networks [C] //Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 990-998

[30] Xu Bo, Xu Yong, Liang Jiaqing, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system [C] //Proc of the 30th Int Conf on Industrial Engineering and Other Applications of Applied Intelligent Systems. Berlin: Springer, 2017: 428-438

[31] Niu Xin, Sun Xinrong, Wang Haofen, et al. Zhishi.me-weaving Chinese linking open data [C] //Proc of the 10th Int Semantic Web Conf. Berlin: Springer, 2011: 205-220

[32] Humphreys K, Gaizauskas R J, Azzam S, et al. University of sheffield: Description of the LaSIE-II system as used for MUC-7 [C/OL] //Proc of the 7th Conf on Message Understanding. New York: ACM, 1998 [2021-11-15]. https://aclanthology.org/M98-1007/

[33] Black W J, Rinaldi F, Mowatt D. FACILE: Description of the NE system used for MUC-7 [C/OL] //Proc of the 7th Conf on Message Understanding. New York: ACM, 1998 [2021-11-15]. https://aclanthology.org/M98-1014/

[34] Krupka G R, Hausman K. IsoQuest Inc: Description of the NetOwl™ extractor system as used for MUC-7 [C/OL] //Proc of the 7th Conf on Message Understanding. New York: ACM, 1998 [2021-11-15]. https://aclanthology.org/M98-1015/

[35] Aone C, Halverson L, Hampton T, et al. SRA: Description of the IE2 system used for MUC-7 [C/OL] //Proc of the 7th Conf on Message Understanding. New York: ACM, 1998 [2021-11-15]. https://aclanthology.org/M98-1012/

[36] Mikheev A, Moens M, Grover C. Named entity recognition without Gazetteers [C] //Proc of the 9th Conf of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 1999: 1-8

[37] Yao Lin, Liu Hong, Liu Yi, et al. Biomedical named entity recognition based on deep neural network [J]. International Journal of Hybrid Information Technology, 2015, 8(8): 279-288

[38] Jiang Yufang, Hu Chi, Xiao Tong, et al. Improved differentiable architecture search for language modeling and named entity recognition [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2019: 3583-3588

[39] Huang Zhiheng, Xu Wei, Yu Kai. Bidirectional LSTM-CRF models for sequence tagging [J]. arXiv preprint, arXiv: 1508.01991, 2015

[40] Peters M E, Ammar W, Bhagavatula C, et al. Semi-supervised sequence tagging with bidirectional language models [C] //Proc of the 55th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2017: 1756-1765

[41] Sui Dianbo, Chen Yubo, Liu Kang, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2019: 3828-3838

[42] Fu T J, Li P H, Ma Wenyun. GraphRel: Modeling text as relational graphs for joint entity and relation extraction [C] //Proc of the 57th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 1409-1418

[43] Luo Ying, Zhao Hai. Bipartite flat-graph network for nested named entity recognition [C] //Proc of the 58th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 6408-6418

[44] Brin S. Extracting patterns and relations from the World Wide Web [C] //Proc of the 1998 Int Workshop on the Web and Databases. Berlin: Springer, 1998: 172-183

[45] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections [C] //Proc of the 5th ACM Conf on Digital Libraries. New York: ACM, 2000: 85-94

[46] Santos C N D, Xiang Bin, Zhou Bowen. Classifying relations by ranking with convolutional neural networks [C] //Proc of the 53rd Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2015: 626-634

[47] Katiyar A, Cardie C. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees [C] //Proc of the 55th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2017: 917-928

[48] Lin Yankai, Shen Shiqi, Liu Zhiyuan, et al. Neural relation extraction with selective attention over instances [C] //Proc of the 54th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 2124-2133

[49] Zhu Hao, Lin Yankai, Liu Zhiyuan, et al. Graph neural networks with generated parameters for relation extraction [C] //Proc of the 57th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 1331-1339

[50] Zhang Yuhao, Qi Peng, Manning C D. Graph convolution over pruned dependency trees improves relation extraction [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 2205-2215

[51] Guo Zhijiang, Zhang Yan, Lu Wei. Attention guided graph convolutional networks for relation extraction [C] //Proc of the 57th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 241-251

[52] Zhang Ningyu, Deng Shumin, Sun Zhanlin, et al. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks [C] //Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 3016-3025

[53] Sahu S K, Christopoulou F, Miwa M, et al. Inter-sentence relation extraction with document-level graph convolutional neural network [C] //Proc of the 57th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4309-4316

[54] Zeng Shuang, Xu Runxin, Chang Baobao, et al. Double graph based reasoning for document-level relation extraction [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 1630-1640

[55] Quinlan J R. Learning logical definitions from relations [J/OL]. Machine Learning, 1990, 5: 239-266 [2021-11-15]. https://link.springer.com/content/pdf/10.1007/BF00117105.pdf

[56] Galárraga L A, Teflioudi C, Hose K, et al. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases [C] //Proc of 22nd Int World Wide Web Conf. New York: ACM, 2013: 413-422

[57] Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks [J]. Machine Learning, 2010, 81(1): 53-67

[58] Socher R, Chen Danqi, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [C] //Proc of the 27th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 926-934

[59] Shi Baoxu, Weninger T. ProjE: Embedding projection for knowledge graph completion [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 1236-1242

[60] Neelakantan A, Roth B, McCallum A. Compositional vector space models for knowledge base inference [C/OL] //Proc of the 2015 AAAI Spring Symposia. Palo Alto, CA: AAAI, 2015 [2021-11-15]. https://arxiv.org/abs/1504.06662

[61] Shang Chao, Tang Yun, Huang Jing, et al. End-to-end structure-aware convolutional networks for knowledge base completion [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 3060-3067

[62] Zhang Zhao, Zhuang Fuzhen, Zhu Hengshu, et al. Relational graph neural network with hierarchical attention for knowledge graph completion [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 34(5): 9612-9619

[63] Schlichtkrull M S, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks [C] //Proc of the 15th Int Conf of Semantic Web. Berlin: Springer, 2018: 593-607

[64] Vashishth S, Sanyal S, Nitin V, et al. Composition-based multi-relational graph convolutional networks [C/OL] //Proc of the 8th Int Conf on Learning Representations. 2020 [2021-11-15]. https://arxiv.org/abs/1911.03082

[65] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of the 27th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2787-2795

[66] Lin Yankai, Liu Zhiyuan, Sun Maosong, et al. Learning entity and relation embeddings for knowledge graph completion [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 2181-2187

[67] Ji Guoliang, He Shizhu, Xu Liheng, et al. Knowledge graph embedding via dynamic mapping matrix [C] //Proc of the 53rd Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2015: 687-696

[68] Yang Bishan, Yih W, He Xiaodong, et al. Embedding entities and relations for learning and inference in knowledge bases [C/OL] //Proc of the 3rd Int Conf on Learning Representations. 2015 [2021-11-15]. https://arxiv.org/abs/1412.6575

[69] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 2071-2080

[70] Lin Yankai, Liu Zhiyuan, Luan Huanbo, et al. Modeling relation paths for representation learning of knowledge bases [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 705-714

[71] Nguyen D Q, Vu T, Nguyen T D, et al. A capsule network-based embedding model for knowledge graph completion and search personalization [C] //Proc of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 2180-2189

[72] Nguyen D Q, Nguyen T D, Nguyen D Q, et al. A novel embedding model for knowledge base completion based on convolutional neural network [C] //Proc of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 327-333

[73] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D knowledge graph embeddings [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 1811-1818

[74] Jiang Xiaotian, Wang Quan, Wang Bin. Adaptive convolution for multi-relational learning [C] //Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 978-987

[75] Cai Ling, Yan Bo, Mai Gengchen, et al. TransGCN: Coupling transformation assumptions with graph convolutional networks for link prediction [C] //Proc of the 10th Int Conf on Knowledge Capture. New York: ACM, 2019: 131-138

[76] Xie Zhiwen, Zhou Guangyou, Liu Jin, et al. ReInceptionE: Relation-aware inception network with joint local-global structural information for knowledge graph embedding [C] //Proc of the 58th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 5929-5939

[77] Chen Muhao, Tian Yingtao, Yang Mohan, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment [C] //Proc of the 26th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 1511-1517

[78] Ganea O E, Hofmann T. Deep joint entity disambiguation with local neural attention [C] //Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 2619-2629

[79] Hu Linmei, Ding Jiayu, Shi Chuan, et al. Graph neural entity disambiguation [J]. Knowledge-Based Systems, 2020, 195: 105620

[80] Cao Yixin, Liu Zhiyuan, Li Chengjiang, et al. Multi-channel graph neural network for entity alignment [C] //Proc of the 57th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 1452-1461

[81] Zhang Fanjin, Liu Xiao, Tang Jie, et al. OAG: Toward linking large-scale heterogeneous entity graphs [C] //Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2019: 2585-2595

[82] Sun Zequn, Wang Chengming, Hu Wei, et al. Knowledge graph alignment network with gated multi-hop neighborhood aggregation [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 34: 222-229

[83] Wang Shuo, Du Zhijuan, Meng Xiaofeng. Research progress of large-scale knowledge graph completion technology [J]. SCIENTIA SINICA Informationis, 2020, 50(04): 551-575 (in Chinese)
(王硕, 杜志娟, 孟小峰. 大规模知识图谱补全技术的研究进展[J]. 中国科学: 信息科学, 2020, 50(04): 551-575)

[84] Liu Xiao, Liu Shuyang, Zhuang Yunkai, et al. Explainable reinforcement learning: Basic problems exploration and a survey [J/OL]. Journal of Software, 2021 [2021-11-15]. http://www.jos.org.cn/1000-9825/6485.htm (in Chinese)

(刘潇, 刘书洋, 庄韫恺, 等. 强化学习可解释性基础问题探索和方法综述[J/OL]. 软件学报, 2021 [2021-11-15]. http://www.jos.org.cn/jos/article/abstract/6485)

[85] Sutton R S, Barto A G. Reinforcement Learning: An Introduction [M]. Cambridge, MA: MIT Press, 2018

[86] Ivanov S, D'Yakonov A. Modern deep reinforcement learning algorithms [J]. arXiv preprint, arXiv: 1906.10025, 2019

[87] Watkins C, Dayan P. Technical note Q-learning [J]. Machine Learning, 1992, 8: 279-292

[88] Rummery G A, Niranjan M. On-line Q-learning using connectionist systems [R/OL]. Cambridge: Cambridge University Engineering Department, 1994 [2021-11-15]. https://www.researchgate.net/publication/2500611_On-Line_Q-Learning_Using_Connectionist_Systems

[89] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning [J]. arXiv preprint, arXiv: 1312.5602, 2013

[90] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2019, 518(7540): 529-533

[91] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double Q-learning [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 2094-2100

[92] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay [C/OL] //Proc of the 4th Int Conf on Learning Representations. 2016 [2021-11-15]. https://arxiv.org/abs/1511.05952

[93] Wang Ziyu, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 1995-2003

[94] Williams R J. Simple statistical gradient following algorithms for connectionist reinforcement learning [J]. Machine Learning, 1992, 8(3-4): 229-256

[95] Konda V R, Tsitsiklis J N. On Actor-Critic algorithms [J]. SIAM Journal on Control and Optimization, 2003, 42(4): 1143-1166

[96] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 1928-1937

[97] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms [C] //Proc of the 31st Int Conf on Machine Learning. New York: ACM, 2014: 387-395

[98] Schulman J, Levine S, Moritz P, et al. Trust region policy optimization [C] //Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 1889-1897

[99] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms [J]. arXiv preprint, arXiv: 1707.06347, 2017

[100] Lowe R, Wu Yi, Tamar A, et al. Multi-agent Actor-Critic for mixed cooperative-competitive environments [C] //Proc of the 30th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 6379-6390

[101] Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward [C] //Proc of the 17th Int Conf on Autonomous Agents and MultiAgent Systems. New York: ACM, 2018: 2085-2087

[102] Eysenbach B, Gupta A, Ibarz J, et al. Diversity is all you need: Learning skills without a reward function [C/OL] //Proc of the 7th Int Conf on Learning Representations. 2019 [2021-11-15]. https://arxiv.org/abs/1802.06070

[103] Pinto L, Davidson J, Sukthankar R, et al. Robust adversarial reinforcement learning [C] //Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 2817-2826

[104] Chen Xinshi, Li Shuang, Li Hui, et al. Generative adversarial user model for reinforcement learning based recommendation system [C] //Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 1052-1061

[105] Wang Xiang, Wang Dingxian, Xu Canran, et al. Explainable reasoning over knowledge graphs for recommendation [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 5329-5336

[106] Yang Yaosheng, Chen Wenliang, Li Zhenghua, et al. Distantly supervised NER with partial annotation learning and reinforcement learning [C] //Proc of the 27th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2159-2169

[107] Wan Jing, Li Haoming, Hou Lei, et al. Reinforcement learning for named entity recognition from noisy data [C] //Proc of Natural Language Processing and Chinese Computing. Berlin: Springer, 2020, 357-369

[108] Maes F, Denoyer L, Gallinari P. Sequence labeling with reinforcement learning and ranking algorithms [C] //Proc of the 18th European Conf on Machine Learning. Berlin: Springer, 2007: 648-657

[109] Lao Yadi, Xu Jun, Gao Sheng, et al. Name entity recognition with policy-value networks [C] //Proc of the 42nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2019: 1245-1248

[110] Shi B, Weninger T. Open-world knowledge graph completion [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 1957-1964

[111] Liu Ye, Zhang Sheng, Song Rui, et al. Knowledge-guided open attribute value extraction with reinforcement learning [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 8595-8604

[112] Feng Jun, Huang Minlie, Zhao Li, et al. Reinforcement learning for relation classification from noisy data [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 5779-5786

[113] Qin Pengda, Xu Weiran, Wang W Y. Robust distant supervision relation extraction via deep reinforcement learning [C] //Proc of the 56th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2137-2147

[114] Zeng Xiangrong, He Shizhu, Liu Kang, et al. Large scaled relation extraction with reinforcement learning [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 5658-5665

[115] Takanobu R, Zhang Tianyang, Liu Jiexi, et al. A hierarchical framework for relation extraction with reinforcement learning [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 7072-7079

[116] Xiong Wenhan, Hoang T, Wang W Y. DeepPath: A reinforcement learning method for knowledge graph reasoning [C] //Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 564-573

[117] Das R, Dhuliawala S, Zaheer M, et al. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning [C/OL] //Proc of the 6th Int Conf on Learning Representations. 2018 [2021-11-15]. https://arxiv.org/abs/1711.05851

[118] Shen Yelong, Chen Jianshu, Huang P S, et al. M-Walk: Learning to walk over graphs using Monte Carlo tree search [C] //Proc of the 31st Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018: 6787-6798

[119] Godin F, Kumar A, Mittal A. Using ternary rewards to reason over knowledge graphs with deep reinforcement learning [C/OL] //Proc of the 31st Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018 [2021-11-15]. https://www.researchgate.net/publication/331396810_Using_Ternary_Rewards_to_Reason_over_Knowledge_Graphs_with_Deep_Reinforcement_Learning

[120] Lin X V, Socher R, Xiong Caiming. Multi-hop knowledge graph reasoning with reward shaping [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 3243-3253

[121] Lin Xiao, Subasic P, Yin Hongfeng. Rel4KC: A reinforcement learning agent for knowledge graph completion and validation [C/OL] //Proc of the Workshop on Deep Reinforcement Learning for Knowledge Discovery. New York: ACM, 2019 [2021-11-15]. http://www.cse.msu.edu/~zhaoxi35/DRL4KDD/1.pdf

[122] Li Ruiping, Cheng Xiang. DIVINE: A generative adversarial imitation learning framework for knowledge graph reasoning [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2019: 2642-2651

[123] Tiwari P, Zhu Hongyin, Pandey H M. DAPath: Distance-aware knowledge graph reasoning based on deep reinforcement learning [J/OL]. Neural Networks, 2021, 135: 1-12 [2021-11-15]. https://pubmed.ncbi.nlm.nih.gov/33310193/

[124] Shen Ying, Ding Ning, Zheng Haitao, et al. Modeling relation paths for knowledge graph completion [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(11): 3607-3617

[125] Lei Kai, Zhang Jin, Xie Yuexiang, et al. Path-based reasoning with constrained type attention for knowledge graph completion [J]. Neural Computing and Applications, 2020, 32(11): 6957-6966

[126] Saebi H, Krieg S J, Zhang Chuxu, et al. Heterogeneous relational reasoning in knowledge graphs with reinforcement learning [J]. arXiv preprint, arXiv: 2003.06050, 2020

[127] Wang Heng, Li Shuangyin, Pan Rong, et al. Incorporating graph attention mechanism into knowledge graph reasoning based on deep reinforcement learning [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language. Stroudsburg, PA: ACL, 2019: 2623-2631

[128] Li Shuanyin, Wang Heng, Pan Rong, et al. MemoryPath: A deep reinforcement learning framework for incorporating memory component into knowledge graph reasoning [J/OL]. Neurocomputing, 2021, 419: 273-286 [2021-11-15]. https://www.sciencedirect.com/science/article/abs/pii/S0925231220312959

[129] Wang Qi, Hao Yongsheng, Cao Jie. ADRL: An attention-based deep reinforcement learning framework for knowledge graph reasoning [J/OL]. Knowledge Based System, 2020 [2021-11-15]. https://www.sciencedirect.com/science/article/abs/pii/S0950705120302525

[130] Wan Guojia, Du Bo. GaussianPath: A Bayesian multi-hop reasoning framework for knowledge graph reasoning [C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 4393-4401

[131] Zhang Zheng, Wang Dongqing, Gao Junwei. Learning automata-based multiagent reinforcement learning for optimization of cooperative tasks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(10): 4639-4652

[132] Li Zixuan, Jin Xiaolong, Guan Saiping, et al. Path reasoning over knowledge graph: A multi-agent and reinforcement learning based method [C] //Proc of the 2018 IEEE Int Conf on Data Mining Workshops. Piscataway, NJ: IEEE, 2018: 929-936

[133] Hildebrandt M, Serna J A Q, Ma Yunpu, et al. Reasoning on knowledge graphs with debate dynamics [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 4123-4131

[134] Zhang Yunan, Cheng Xiang, Gao Heting, et al. Cooperative reasoning on knowledge graph and corpus: A multi-agent reinforcement learning approach [J]. arXiv preprint, arXiv: 1912.02206, 2019

[135] Wang Quan, Mao Zhendong, Wang Bin, et al. Knowledge graph embedding: A survey of approaches and applications [J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743

[136] Dong Yuxiao, Chawla N V, Swami A. Metapath2vec: Scalable representation learning for heterogeneous networks [C] //Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 135-144

[137] Wang Xiao, Lu Yuanfu, Shi Chuan, et al. Dynamic heterogeneous information network embedding with meta-path based proximity [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(3): 1117-1132

[138] Fu T Y, Lee W C, Lei Zhen. HIN2Vec: Explore meta-paths in heterogeneous information networks for representation learning [C] //Proc of the 26th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2017: 1797-1806

[139] Zhong Zhiqiang, Li C T, Pang Jun. Reinforcement learning-based personalised meta-path generation for heterogeneous graph neural networks [J]. arXiv preprint, arXiv: 2010.13735v1, 2020

[140] Wan Guojia, Du Bo, Pan Shirui, et al. Reinforcement learning based meta-path discovery in large-scale heterogeneous information networks [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 34(4): 6094-6101

[141] Zhu Hao, Xie Ruobing, Liu Zhiyuan, et al. Iterative entity alignment via joint knowledge embeddings [C] //Proc of the Int Joint Conf on Artificial Intelligence. Piscataway, NJ: IEEE, 2017: 4258-4264

[142] Fang Zheng, Cao Yunan, Li Qian, et al. Joint entity linking with deep reinforcement learning [C] //Proc of the 2019 World Wide Web Conf. New York: ACM, 2019: 438-447

[143] Zeng Weixin, Zhao Xiang, Tang Jiuyang, et al. Reinforcement learning-based collective entity alignment with adaptive features [J]. ACM Transactions on Information Systems, 2021, 39(3): 26:1-26:31

[144] Song Weiping, Duan Zhijian, Yang Ziqing, et al. Explainable knowledge graph-based recommendation via deep reinforcement learning [J]. arXiv preprint, arXiv: 1906.09506, 2019

[145] Wang Xiang, Xu Yaokun, He Xiangnan, et al. Reinforced negative sampling over knowledge graph for recommendation [C] //Proc of the 2020 World Wide Web Conf. New York: ACM, 2020: 99-109

[146] Xian Yikun, Fu Zuohui, Muthukrishnan S, et al. Reinforcement knowledge graph reasoning for explainable recommendation [C] //Proc of the 42nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2019: 285-294

[147] Zhao Kangzhi, Wang Xiting, Zhang Yuren, et al. Leveraging demonstrations for reinforcement recommendation reasoning over knowledge graphs [C] //Proc of the 43rd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2020: 239-248

[148] Liu Qi, Tong Shiwei, Liu Chuanren, et al. Exploiting cognitive structure for adaptive learning [C] //Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2019: 627-635

[149] Wang Pengfei, Fan Yu, Xia Long, et al. KERL: A knowledge-guided reinforcement learning model for sequential recommendation [C] //Proc of the 43rd Int ACM SIGIR Conf on research and development in Information Retrieval. New York: ACM, 2020: 209-218

[150] Shi Yue, Larson M A, Hanjalic A. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges [J]. Computer Surveys, 2014, 47(1): 3:1-3:45

[151] Guo Qingyu, Zhuang Fuzhen, Qin Chuan, et al. A survey on knowledge graph based recommender systems [J/OL]. IEEE Transactions on Knowledge and Data Engineering, 2020 [2021-11-15]. https://arxiv.org/abs/2003.00911

[152] Zhou Sijin, Dai Xinyi, Chen Haokun, et al. Interactive recommender system via knowledge graph-enhanced reinforcement learning [C] //Proc of the 43rd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2020: 179-188

[153] Lei Wenqiang, Zhang Gangyi, He Xiangnan, et al. Interactive path reasoning on graph for conversational recommendation [C] //Proc of the 26th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2020: 2073-2083

[154] Lei Wenqiang, He Xiangnan, Miao Yisong, et al. Estimation-Action-Reflection: Towards deep interaction between conversational and recommender systems [C] //Proc of the 13th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2020: 304-312

[155] Deng Yang, Li Yaliang, Sun Fei, et al. Unified conversational recommendation policy learning via graph-based reinforcement learning [C] //Proc of the 44th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2021: 1431-1441

[156] Xu Cong. Research of dialogue model based on deep learning and reinforcement learning [D]. Beijing: University of Science and Technology Beijing, 2020 (in Chinese)
(徐聪. 基于深度学习和强化学习的对话模型研究[D]. 北京:北京科技大学, 2020)

[157] Kaiser M, Roy R S, Weikum G. Reinforcement learning from reformulations in conversational question answering over knowledge graphs [C] //Proc of the 44th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2021: 459-469

[158] Christmann P, Roy R S, Abujabal A, et al. Look before you Hop: Conversational question answering over knowledge graphs using judicious context expansion [C] //Proc of the 28th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2019: 729-738

[159] Ammanabrolu P, Riedl M. Playing text-adventure games with graph-based deep reinforcement learning [C] //Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 3557-3565

[160] Ammanabrolu P, Hausknecht M. Graph constrained reinforcement learning for natural language action spaces [C/OL] //Proc of the 8th Int Conf on Learning Representations. 2020 [2021-11-15]. https://arxiv.org/abs/2001.08837

[161] Hausknecht M, Ammanabrolu P, Côté M, et al. Interactive fiction games: A colossal adventure [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 7903-7910

[162] Xu Yunqiu, Fang Meng, Chen Ling, et al. Deep reinforcement learning with stacked hierarchical attention for text-based games [C/OL] //Proc of the 33rd Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2020 [2021-11-15]. https://arxiv.org/abs/2010.11655

[163] Adhikari A, Yuan Xingdi, Côté M-A, et al. Learning dynamic knowledge graphs to generalize on text-based games [J]. arXiv preprint, arXiv: 2002.09127, 2020

[164] Ammanabrolu P, Riedl M. Transfer in deep reinforcement learning using knowledge graphs [C] //Proc of the 13th Workshop on Graph-Based Methods for Natural Language Processing. Stroudsburg, PA: ACL, 2019: 1-10

[165] Murugesan K, Atzeni M, Shukla P, et al. Enhancing text-based reinforcement learning agents with commonsense knowledge [J]. arXiv preprint, arXiv: 2005.00811, 2020

[166] You Jiaxuan, Liu Bowen, Ying Zhitao, et al. Graph convolutional policy network for goal-directed molecular graph generation [C] //Proc of the 31st Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018: 6412-6422

[167] Do K, Tran T, Venkatesh S. Graph transformation policy network for chemical reaction prediction [C] //Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2019: 750-760

[168] Wang Shanshan, Ren Pengjie, Chen Zhumin, et al. Order-free medicine combination prediction with graph convolutional reinforcement learning [C] //Proc of the 28th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2019: 1623-1632

[169] Sun Zhoujian, Dong Wei, Shi Jinlong, et al. Interpretable disease prediction based on reinforcement path reasoning over knowledge graphs [J]. arXiv preprint, arXiv: 2010.08300, 2020

[170] Miao Rui, Zhang Xia, Yan Hongfei, et al. A dynamic financial knowledge graph based on reinforcement learning and transfer learning [C] //Proc of the 2019 IEEE Int Conf on Big Data. Piscataway, NJ: IEEE, 2019: 5370-5378

[171] Piplai A, Ranade P, Kotal A, et al. Using knowledge graphs and reinforcement learning for malware analysis [C] //Proc of the 2020 IEEE Int Conf on Big Data. Piscataway, NJ: IEEE, 2020: 2626-2633

[172] An Bo, Chen Bo, Han Xianpei, et al. Accurate text-enhanced knowledge graph representation learning [C] //Proc of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018, 745-755

[173] Niu Guanglin, Zhang Yongfei, Li Bo, et al. Rule-Guided compositional representation learning on knowledge graphs [C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 2950-2958

[174] Li Zixuan, Jin Xiaolong, Guan Saiping, et al. Search from history and reason for future: Two-stage reasoning on temporal knowledge graphs [J]. arXiv preprint, arXiv: 2106.00327, 2021

[175] He Dongliang, Zhao Xiang, Huang Jizhou, et al. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 8393-8400

[176] Alver S, Precup D. What is going on inside recurrent meta reinforcement learning agents? [C/OL] //Proc of the 9th Int Conf on Learning Representations. 2021 [2021-11-15]. https://arxiv.org/abs/2104.14644

[177] Chen Lili, Lu K, Rajeswaran A, et al. Decision transformer: Reinforcement learning via sequence modeling [J]. arXiv preprint, arXiv: 2106.01345, 2021

[178] Zhang Kaifeng, Yu Yang. Methodologies for imitation learning via inverse reinforcement learning: A review [J]. Journal of Computer Research and Development, 2019, 56(2): 254-261 (in Chinese)

(张凯峰, 俞扬. 基于逆强化学习的示教学习方法综述[J]. 计算机研究与发展, 2019, 56(2): 254-261)

[179] Hou Zhongni, Jin Xiaolong, Li Zixuan, et al. Rule-aware reinforcement learning for knowledge graph reasoning [C] // Proc of the 59th Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2021: 4687-4692

[180] Hua Yuncheng, Li Yuanfang, Haffari G, et al. Few-shot complex knowledge base question answering via meta reinforcement learning [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 5827-5837

Ma Ang, born in 1992. PhD candidate. Student member of CCF. Her main research interests include knowledge graph, reinforcement learning and recommendation system.

马 昂,1992 年生.博士研究生,CCF 学生会员.主要研究方向为知识图谱、强化学习和推荐系统.

Yu Yanhua, born in 1974. PhD, associate professor. Member of CCF. Her main research interests include data mining, machine learning, natural language processing and big data analysis.

于艳华,1974 年生.博士,副教授,CCF 会员.主要研究方向为数据挖掘、机器学习、自然语言处理和大数据分析.

Yang Shengli, born in 1968. PhD, professor. His main research interests include computer system simulation and evaluation, big data technology application.

杨胜利,1968 年生.博士,教授.主要研究方向为计算机系统仿真与评估和大数据技术应用.

Shi Chuan, born in 1978. PhD, professor. Member of CCF. His main research interests include data mining, machine learning, artificial intelligence and big data analysis.

石 川,1978 年生.博士,教授,CCF 会员.主要研究方向为数据挖掘、机器学习、人工智能和大数据分析.

Li Jie, born in 1977. PhD, lecturer. Member of CCF. His main research interests include cognitive science and knowledge graph.

李 劼,1977 年生.博士,讲师,CCF 会员.主要研究方向为认知科学和知识图谱.

Cai Xiuxiu, born in 1999. Master candidate. Student member of CCF. Her main research interests include knowledge graph, reinforcement learning.

蔡修秀,1999 年生.硕士研究生,CCF 学生会员.主要研究方向为知识图谱和强化学习.