DOI 10.19656/j.cnki.1002-2406.20220301

信 息 工 程

一种面向中医医案知识图谱的链路预测模型

羊艳玲,李燕✉,钟昕妤
(甘肃中医药大学信息工程学院,甘肃 兰州 730000)

【摘要】知识图谱有助于实现智能化、个性化的中医药知识服务,其中链路预测可以解决知识图谱中缺失信息的发现与还原,也是目前知识图谱应用领域中的研究热点之一。但目前补全的医学知识图谱很少覆盖类型和层级结构,而且链路预测未考虑到关系三元组。本文创新性地提出一种新的归纳推理模型 HSTP(Hierarchical Structure Type),基于类型和层级结构获取信息,利用知识图谱中实体之间的语义关联,将类型、节点与关系进行融合,然后提出一个关系相关网络来学习不同模式对归纳链路预测的重要性,最终得到真实和完整的中医医案知识图谱。结果表明,该模型能够有效表达实体之间的语义关联,在链路预测任务的基准数据集上提高了 3.9% 左右,可以为解决知识图谱中缺失信息的发现与还原提供研究基础。

【关键词】知识图谱;链路预测;中医医案

【引用格式】
羊艳玲,李燕,钟昕妤. 一种面向中医医案知识图谱的链路预测模型[J]. 中医药信息, 2022, 39(3): 1-6,15.
YANG Y L, LI Y, ZHONG X Y. A link prediction model for knowledge graph of TCM medical records[J]. Information on TCM, 2022, 39(3): 1-6,15.

知识图谱(knowledge graph,KG)是大数据时代下针对海量知识产生的一种新型管理与服务模式,其属于语义网络范畴,是表示知识的一种新途径,用于描述真实世界中存在的各种实体、概念或属性,抽取并呈现出特定领域概念之间的语义关系[1]。近年来,因其有助于医学信息表达的分类和标准化,以及医学知识的共享、分布和应用,具有临床诊断、治疗、研究和教育应用价值,知识图谱在医学领域也逐渐得到关注与重视。它将医学知识映射纳入知识服务系统,以提高信息检索、智能问答、决策支持和知识可视化等多种服务的效果,从而提升知识服务能力[2]。

然而,随着深入研究以及将知识图谱应用到各种领域,研究人员发现在应用中仍存有一些问题,其中限制广泛应用的最主要因素是不完备性[3],即知识图谱中存在缺失的实体或信息,导致其应用存在一定约束,大大限制了用于检索和推理的准确性。因此,知识图谱链路预测是补全知识的一个重要基础,其首要目标是预测知识图谱中实体之间可能存在的关系,以及发现和恢复缺失信息[4]。

链路预测通过网络中已知节点的信息和网络结构,预测两个无限连接节点之间存在链接关系的可能性,为缺失信息恢复和错误信息检测提供技术支持[5]。链路预测是信息科学与复杂网络之间的重要联系,近年来,国内外学者们就知识图谱的链路预测应用方面开展了众多研究工作,已形成较为全面且系统的成果[6]。如 GETOOR 等[7]对链路预测实现数据挖掘的相关概念和研究进行了梳理归纳,重点叙述了其定义、存在的问题和经典方法。DRUMOND 等[8]针对 KG 更新,提出利用张量分解的方法实现对缺失 RDF 三元组数据的补充。SOCHER 等[9]在预测中引入神经网络方

基金项目:中国高校产学研创新基金-异构智能计算项目(2020HYA02008);甘肃中医药大学研究生创新基金项目(2021CX76)
第一作者简介:羊艳玲(1995-),女,2019级生物医学工程专业硕士研究生。
✉通讯作者简介:李燕(1976-),女,教授,主要研究方向:中医药数据挖掘。

法,但存在模型复杂和参数调优等不足。目前,在链路预测研究中主要面临以下两大难题,一是现有大规模KG存储数据量极大;二是KG构建形成单一的实体属性和关系而忽视了相关联的外部信息,而这些外部信息中包含了极为丰富的先验知识,因此融合关联外部信息的KG才是更为全面、真实的。如何将KG提供的数据与外部知识相结合也是面临的难点之一[10]。

在中医临床领域,构建知识图谱的一个核心知识源是中医医案。由于实际医案数据普遍存在歧义性和多样性的问题[11],使临床领域知识网络中可能存在一些缺失的医疗实体和实体之间的链接,或者实体之间可能存在不正确的关联。这些关系可以利用临床领域知识图谱链路预测进行补充或校正,得到更加全面、真实的知识图谱[12]。医疗领域中医案数据通常具有语义关联,并且医案之间的语义具有很明显的强关联性。与此同时,关于高血压病领域知识图谱的链路预测少之又少,且未考虑到中医知识图谱三元组体系和类型以及信息缺失等问题。为了应对这一挑战,本课题组提出了一种新的归纳推理模型,即HSTP(Hierarchical Structure Type),旨在将中医知识图谱从两方面进行补全优化,一是利用中医三元组类型,如<疾病,处方,药物>和层级结构进行补全;二是利用新提出的模型判断图谱中两个节点是否一致,加强实体类型形成拓扑层级结构。

1 相关工作

知识图谱中的链路预测是利用已有的关系推断出新的关系,从而建立一个更完整的知识图谱任务。为了补充KG中实体之间缺失的信息,知识图谱的解决方案是利用现有知识推断潜在知识。换句话说,KG是用现有事实来预测知识图谱中实体之间的潜在关系。在某种程度上,KG实质就是复杂网络,其类似于复杂网络中的链路预测,但更复杂的是不仅要预测节点之间可能的链接关系,而且能够推断这些链接关系中包含的各种信息[13]。尽管归纳链路预测在实际应用中的重要性不言而喻,但现有的研究大多集中在演绎链路预测,无法应对从未观察到的实体。链路预测问题是复杂网络的一个经典问题,当前已有了丰富的成果,总体是通过分析节点之间的相似关系来进行预测,比如基于相似性的预测方法、基于似然估计的预测方法、基于概率模型的预测方法等。对于深度模型来说,更多工作将链路预测作为深度模型的评价方法来使用,本质也是来挖掘两节点的相似性。现较为成熟的链路预测方法有以下几种。

一是基于规则学习的方法。这一方法是基于观察到的关系共现模型,学习规则一般是通过归纳得到的,而且能够自然过渡到其他实体,因为它们和实体之间是独立的。Neural LP提出了一种端到端可微框架来学习逻辑规则的结构和参数[14]。DRUM通过挖掘更正确的逻辑规则,进一步改进了神经网络[15]。然而,基于规则学习的方法主要集中在挖掘horn规则,限制了它们对知识图谱中更复杂的语义关系建模的能力。

二是基于嵌入的方法。该方法已被证明是知识图谱推理的一个有前途的方向[16],一些基于嵌入的方法可以为未见的实体生成嵌入。GraIL等提出基于GNN的预测框架,通过推理局部子图实现实体独立方式的归纳预测,但该方法无法实现常见实体的关系获取[17]。

三是基于GNNs的链路预测。由于KG自身的图模式表达特点,基于GNNs的链路预测方法在近几年展现出巨大潜力。ZHANG等[18]利用GNN结合层次注意力实现对实体领域信息的有效利用,但该网络的训练依赖于实体嵌入而难以对不可见实体间缺失链接进行补充。

四是基于关系矩阵的方法。近期的KG嵌入方法研究开始考虑引入关系间相关性。DO等[19]实现对关系投影空间的跨越基分解并共享给所有关系。ZHU等[20]尝试将关系矩阵分解成两个低维矩阵相乘来学习。

2 基于HSTP模型的构建方法

2.1 构建模型

本文提出一种新的归纳推理模型HSTP,它能有效利用相邻的关系三元组。具体来说是从关联模型(correlation patterns)和相关系数(correlation coefficients)两个方面对语义关联进行了建模。根据不同的结构特征将所有关系划分为多个关联模式组,然后将原始知识图转换为关系相关图(RCG),其中节点表示关系,边表示任意两个关系之间的关联模式。

2.2 相关定义

定义1(关系相关模块):基于任何两个关系之间的语义相关性与其拓扑结构高度相关的模块。

定义2(关系相关网络,RCN):模拟不同相关模式在链路预测中的重要性。它由相关模式和相关系数两个部分组成。

定义3(关联模式):任意两个关系之间的关联都与其拓扑结构相关。

定义4(相关系数):阐述两种关系之间的语义关联程度。

定义5(关系相关图,RCG):为达到对任意两个关系之间的相关模式进行建模的目的,将所有关系对分为七类拓扑模式。其中节点表示关系,边表示原始知

识图中任意两个关系之间的关联模式。

定义6(图形结构模块):对于三元组(u,r_t,v),周围的局部图包含了关于三元组如何与其邻域连接的信息。

2.3 理论基础

HSTP旨在以独立于实体的方式对给定的三元组(u,r_t,v)进行评分,其中r_t是u和v之间的目标关系。其中HSTP包括两个模块:关系关联模块和图形结构模块。关系关联模块输出嵌入向量r_t^N和e_u,将这两个模块组织在一个统一的框架中,框架见图1。利用一个评分网络将这两个模块的输出结合起来,得到给定三元组(u,r_t,v)的分数。

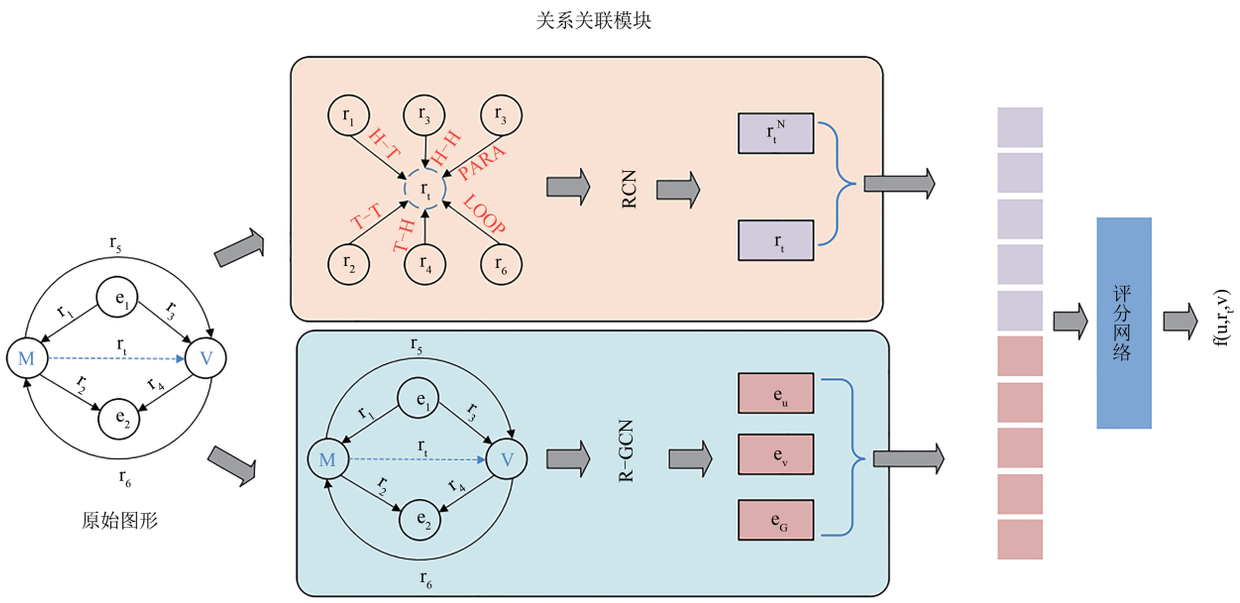

图1 评分框架图

得分函数:$f(u,r_t,v)$定义为:

$$f(u,r_t,v)=\left[r_t^f \oplus e_S\right]W_S \quad (1)$$

其中,$W_S \in R^{4d\times 1}$代表权重参数。

损失函数:进行负采样并使用对比较链接损失对模型进行训练,使其得分正样本高于负样本。损失函数定义为:

$$L=\sum_{i\in[n],(u,r_t,v)\in\Gamma} max(0,f(u_i',r_{t,i}',v_i')-f(u,r_t,v)+\gamma) \quad (2)$$

其中,γ代表超参数;$(u_i',r_{t,i}',v_i')$代表表示事实三元组(u,r_t,v)的第i个负样本;$[n]$代表$\{1,2,\cdots,n\}$,n是负样本个数。

3 实验数据集及结果分析

3.1 数据集

为验证本文所提出的HSTP模型的效果及解决高血压病中医医案KG的补全,需要在数据集上通过实验验证。笔者使用了文献[17]中提出的归纳链路预测基准数据集作为公开数据集用作训练,这些数据来自WN18RR[21]、FB15k-237[22]和NELL-995[23]。将前期收集的高血压病中医医案数据集作为私有数据集用作预测,其中有1 345个关系三元组用来训练,共包含632个实体和495关系。

对于感应链路预测,训练组和测试组应重叠实体。WN18RR、FB15k-237和NELL-995归纳出4种类型的归纳数据集,且其大小不断增加。数据集详细信息见表1。

3.2 实验参数设置

将HSTP与几种经典的方法进行比较,包括Neural LP[14]、DRUM[15]和GraIL[17]。使用Adam优化器[24]进行训练,初始学习率为0.01,批量大小为16。在训练和测试时,随机抽取每个三元组的两跳封闭子图,并使用一个两层的GCN来计算子图的嵌入。对于WN18RR、FB15k-237和NELL-995,损失函数中的margins分别设置为8、16、10,最大训练时epochs设置为10。

3.3 三元组分类及结果分析

三元组分类是一个简单的二分类问题,即对一个三元组(u,r_t,v)判断它是正样本还是负样本。链路预测是用实体集中的实体替换掉头实体或尾部实体,计算所有三元组的得分,然后得到原三元组在所有三元组中的排名。三元组分类任务在很多补全模型中被当作评测任务,其方法是通过三元组(u,r_t,v)的两个阶段模型传播和输出模型计算(u,r_t,v)的得

分函数,如果评分函数小于指定阈值划分为正样本,否则为负样本。由于这是一项二元分类任务,使用准确率作为评估指标。三元组分类的实验结果如表2、图2所示。

表1 归纳基准的统计数据表

No.		WN18RR			FB15k-237			NELL-995		
		#R	#E	#TR	#R	#E	#TR	#R	#E	#TR
v1	Train	9	2 746	6 678	183	2 000	5 226	14	10 915	5 540
	Test	9	924	1 991	146	1 500	2 404	14	225	1 034
v2	Train	10	6 954	18 968	203	3 000	12 085	88	2 564	10 109
	Test	10	2 923	4 863	176	2 000	5 092	79	4 937	5 521
v3	Train	11	12 078	32 150	218	4 000	22 394	142	4 947	20 117
	Test	11	5 084	7 470	187	3 000	9 137	122	4 921	9 668
v4	Train	9	3 861	9 842	222	5 000	33 916	77	2 092	9 289
	Test	9	7 208	15 157	204	3 500	14 554	61	3 294	8 520

注:#E、#R、#TR分别表示实体、关系和三元组的数量。

表2 三元组分类任务的准确率(%)

Method	WN18RR	FB15k-237	NELL-995
Neural LP	78.4	75.5	76.9
DRUM	79.0	76.4	77.6
GraIL	79.2	77.2	79.1
HSTP	81.2	82.8	83.7

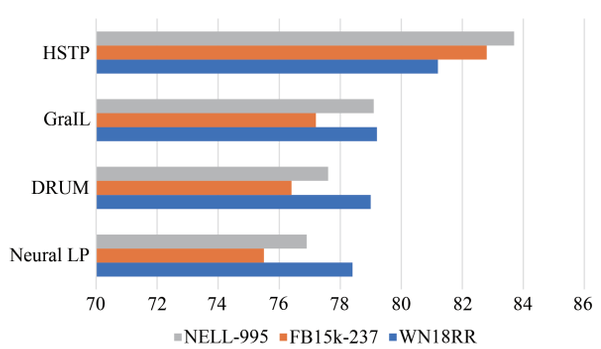

图2 三元组分类任务准确率

由以上结果可知:

①在三元组分类任务上,HSTP性能优于Neural LP、DRUM和GraIL模型。说明HSTP模型获取实体向量的方式较其他复杂,能有效地捕捉序列特征。

②比较数据集WN18RR、FB15k-237和NELL-995发现,随着实体增多,训练数据的减少,模型的分类效果都会降低。但是在同一个数据集下,各种模型性能的相对关系基本保持不变。

③在三个数据集上,本文提出的HSTP模型相对于其他模型分别提高了2.4%、2.8%和3.2%,整体性能优于上述模型。

3.4 链路预测模型

3.4.1 基准模型

为评估提出的关系相关模块的有效性,课题组提出了一个称为HSTP-base的基线,该基线得分三元组(u, r_t, v),仅依赖于关系相关模块的输出,因此,HSTP-base的得分函数为:

$$f_{base}(u, r_t, v) = r_t^F W_{base} \tag{3}$$

其中,$W_{base} \in R^{d \times 1}$代表权重参数。

3.4.2 评价及分析

使用精度召回曲线(AUC)下的面积作为分类度量,AUC(Area Under Curve)被定义为ROC曲线下的面积。用随机实体替换每个测试三元组的头部或尾部,以对相应的负三元组进行采样。然后用相等数量的负三元组对正三元组进行评分,用不同的随机种子进行实验,并报告平均结果。从WN18RR、FB15k-237和NELL-995中提取的归纳基准数据集的AUC-PR,结果见表3、图3。

表3 链路预测任务实验结果

Method	WN18RR				FB15k-237				NELL-995			
	v1	v2	v3	v4	v1	v2	v3	v4	v1	v2	v3	v4
Neural LP	86.00	82.56	61.72	81.08	66.34	74.55	72.15	72.24	62.12	81.23	85.12	83.69
DRUM	85.02	83.05	62.10	81.08	68.71	74.44	71.03	71.10	57.86	82.78	86.12	81.45
GraIL	91.57	91.26	81.60	90.12	80.72	82.12	90.43	90.16	81.15	90.86	90.12	84.46
HSTP-base	97.11	95.23	87.25	95.15	84.35	92.23	95.24	97.13	91.00	91.22	93.12	92.64
HSTP	96.35	95.95	90.11	94.34	86.53	91.20	95.10	98.11	91.22	94.34	94.60	95.53

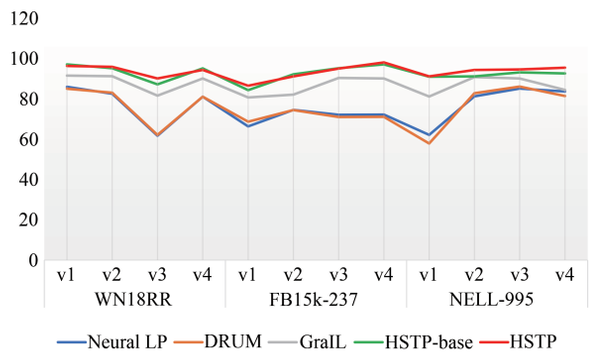

图 3　链路预测的 AUC – PR 结果图

由图 3 链路预测的 AUC – PR 结果可知，从三元组任务和链路预测两个任务总体来说，课题组的基线模型 HSTP – base 在所有数据集上都优于归纳基线。由于 HSTP – base 完全依赖于关系相关模块来执行链路预测，此结果证明课题组提出的归纳链路预测模型大大提高了 HSTP 模型的性能，在大多数数据集上比 GraIL 提高了 3.9% 左右，验证了在归纳链路预测任务中 HSTP 模型的有效性。

3.5　补全高血压病中医医案知识图谱

古代中医学中并无“高血压病”概念，现代高血压病在中医辨证理论体系中所对应的疾病有“眩晕”“头痛”，对应的病机为“肝阳上亢”。如图 4 和图 5 所示，其描述了关于高血压病中医医案知识图谱的补全前和补全后，因为未知链路预测的任务是判断图中实体之间的连线是否真实存在，所以补全后的知识图谱可以考虑到相邻的关系三元组，展示更多缺失的信息，使高血压病在中医的辨证论治中更加系统化、全面化。首先体现在高血压病的证治分型上，在肝火上炎、肝肾亏虚、气虚血瘀、阴虚阳亢和痰湿壅盛证的研究基础上增加了其他脏腑、气血津液和情志证等方面的辨证，同时也对应增加了每个证型的脉证、舌象等具体临床症状表现。补全优化后的优势主要体现以下三点：第一，中医学的基本特点是整体观念和辨证论治，一个脏腑或者一个部位的病变往往会累及其他的脏腑和部位，通过补全此图，可以比较直观地看到相同疾病不同证型之间会出现有共同的证候，体现中医学整体观念和辨证论治的特点；第二，使本病的辨证更加精确完善且具有连贯性和统一性；第三，强化表达了疾病 – 症状 – 证候之间的联系，体现了中医辨证以五脏为中心的整体观，辨证方式包含了脏腑辨证、八纲辨证和气血津液辨证。

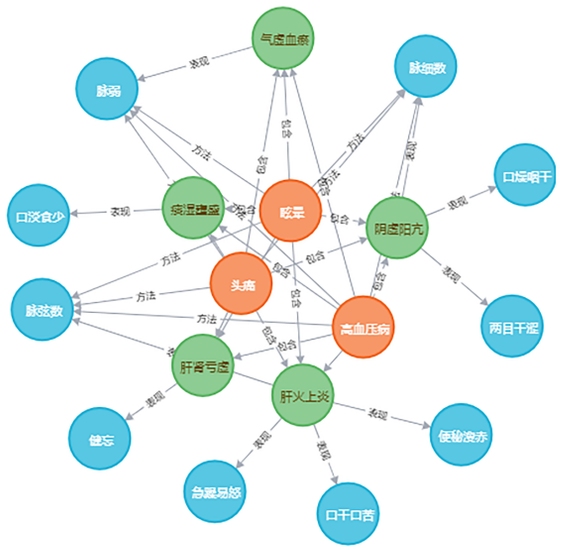

图 4　“疾病 – 症状 – 证候”可视化图补全前

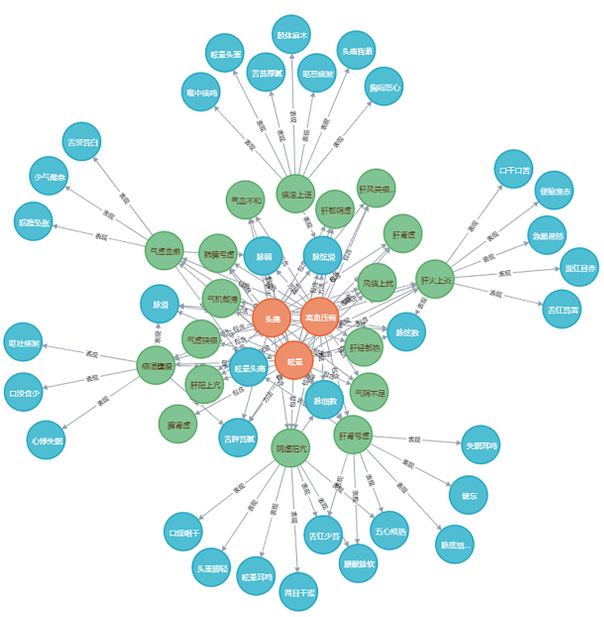

图 5　“疾病 – 症状 – 证候”可视化图补全后

4　总结

本文以医学领域为例，针对知识图谱描述中医高血压病病例的特点，提出将 KG 与相邻三元组相结合，充分描述实体节点的属性，并构建了一个关联关系描述属性的模型。然后基于 HSTP 预测模型实现 KG 的信息补全，从而找到其中缺失的信息。基于真实数据集的实验，验证了该方法的有效性，实验结果在一定程度上具有可行性。笔者就知识图谱补全研究中面临的信息覆盖不全面及相邻三元组信息丢失两大问题，提出以下解决方法。一是结合实体类型和层级结构信息（如中医知识图谱中的疾病 – 子病 – 类型结构）补全知识图谱；二是融合实体信息与拓扑结构形成增量，实现模型结构优化。经实验验证，新提出的 HSTP 归纳

推理模型能够有效建模语义关联并对比其他方法获得了更优的链路预测性能。

本文提供了一种针对医学领域知识图谱未知链路预测思路,但只限于初步探寻。以知识图谱链路预测问题的特点和应用为出发点,今后要开展的工作主要为如何处理大规模的知识图谱和海量标签数据集,并将预测未知链接扩展到医学的其他方面。

【参考文献】

[1] SEQUEDA J F. Integrating relational databases with the semantic web: a reflection[C]//Birkbeck, University of London. Reasoning Web 2017: Reasoning Web. Semantic Interoperability on the Web. London, UK: Springer International Publishing AG, 2017: 68 – 120.

[2] 于彤,李敬华,朱玲,等. 中医临床知识图谱的构建与应用[J]. 科技新时代,2017(4):51 – 54.

[3] 实体识别与链接[C]//中国中文信息学会语言与知识计算专业委员会. 2018中国中文信息学会语言与知识计算专业委员会学术研讨会会议论文集:知识图谱发展报告(2018). 广州:中国中文信息学会语言与知识计算专业委员会,2018:31 – 38.

[4] BAADER F, SERTKAYA B. Usability issues in description logic knowledge base completion[C]//Karl Erich Wolff, University of Applied Sciences, Darmstadt, Germany. ICFCA 2009: Lecture Notes in Computer Science, vol 5548. Darmstadt, Germany: Springer, 2009: 1 – 21.

[5] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5):651 – 661.

[6] 韩路,尹子都,王钰杰,等. 基于贝叶斯网的知识图谱链接预测[J]. 计算机科学与探索,2017,11(5):742 – 751.

[7] GETOOR L, DIEHL C P. Link mining: a survey[J]. Acm Sigkdd Explorations Newsletter, 2005, 7(2): 3 – 12.

[8] DRUMOND L, RENDLE S, SCHMIDT-THIEME L. Predicting RDF triples in incomplete knowledge bases with tensor factorization[C]// Microsoft Research – University of Trento Centre for Computational and Systems Biology. Proceedings of the 27th Annual ACM Symposium on Applied Computing. New York: Association for Computing Machinery, 2012: 326 – 331.

[9] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//NIPS. Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Lake Tahoe, Nevada, United States: NIPS, 2013.

[10] 韩路. 基于贝叶斯网的知识图谱链接预测[D]. 昆明:云南大学, 2017:22 – 24.

[11] 唐琳,郭崇慧,陈静锋. 中文分词技术研究综述[J]. 数据分析与知识发现,2020,4(Z1):1 – 17.

[12] 陈德华,殷苏娜,乐嘉锦,等. 一种面向临床领域时序知识图谱的链接预测模型[J]. 计算机研究与发展, 2017, 54(12):2687–2697.

[13] MA J, QIAO Y, HU G, et al. ELPKG: a high-accuracy link prediction approach for knowledge graph completion[J]. Symmetry, 2019, 11(9): 1096.

[14] YANG F, YANG Z, COHEN W W. Differentiable learning of logical rules for knowledge base reasoning[C]//NIPS. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, CA, USA: NIPS, 2017.

[15] SADEGHIAN A, ARMANDPOUR M, DING P, et al. Drum: End-to-end differentiable rule mining on knowledge graphs[J]. Advances in Neural Information Processing Systems, 2019.

[16] SUN Z, DENG Z H, NIE J Y, et al. Rotate: Knowledge graph embedding by relational rotation in complex space[J]. arXiv preprint arXiv:1902. 10197, 2019.

[17] TERU K, DENIS E, HAMILTON W. Inductive relation prediction by subgraph reasoning[C]//ICML. ICML 2020: 37th International Conference on Machine Learning. https://icml. cc/virtual/2020: PMLR, 2020: 9448 – 9457.

[18] ZHANG Z, ZHUANG F, ZHU H, et al. Relational graph neural network with hierarchical attention for knowledge graph completion[C]// The Association for the Advancement of Artificial Intelligence. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California USA: AAAI Press, 2020: 9612 – 9619.

[19] DO K, TRAN T, VENKATESH S. Knowledge graph embedding with multiple relation projections[C]//The Chinese Association of Automation. 2018 24th International Conference on Pattern Recognition. Beijing, China: IEEE, 2018: 332 – 337.

[20] ZHU J Z, JIA Y T, XU J, et al. Modeling the correlations of relations for knowledge graph embedding[J]. Journal of Computer Science and Technology, 2018, 33(2):323 – 334.

[21] TOUTANOVA K, CHEN D. Observed versus latent features for knowledge base and text inference[C]//The Association for Computational Linguistics and The Asian Federation of Natural Language Processing. Proceedings of the 3rd workshop on continuous vector space models and their compositionality. Beijing, China: Taberg Media Group AB, 2015: 57 – 66.

[22] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2d knowledge graph embeddings[C]//The Association for the Advancement of Artificial Intelligence. Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, Lousiana, USA: AAAI Press, 2018, 32(1).

[23] XIONG W, HOANG T, WANG W Y. Deeppath: A reinforcement learning method for knowledge graph reasoning[C]//ACL. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: ACL, 2017: 564 – 573.

[24] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]// ICLR. the 3rd International Conference for Learning Representations. San Diego, CA, USA, 2014, 12:6980.

(收稿日期:2021 – 12 – 04)

(下转第15页)