计算机科学
COMPUTER SCIENCE

DL$^+$:一种增强型双层知识图谱推理框架

武月佳, 周建涛

引用本文

武月佳, 周建涛. DL$^+$:一种增强型双层知识图谱推理框架[J]. 计算机科学, 2023, 50(12): 302-313.

WU Yuejia, ZHOU Jiantao. DL$^+$:An Enhanced Double-layer Framework for Knowledge Graph Reasoning [J]. Computer Science, 2023, 50(12): 302-313.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000170

DL$^+$:一种增强型双层知识图谱推理框架

武月佳　周建涛
内蒙古大学计算机学院(软件学院)　呼和浩特 010021
生态大数据教育部工程研究中心　呼和浩特 010021
蒙古文智能信息处理技术国家地方联合工程研究中心　呼和浩特 010021
内蒙古自治区云计算与服务软件工程实验室　呼和浩特 010021
内蒙古自治区社会计算与数据处理重点实验室　呼和浩特 010021
内蒙古自治区纪检监察大数据重点实验室　呼和浩特 010021
内蒙古自治区大数据分析技术工程实验室　呼和浩特 010021
(wuyuejia@imudges.com)

摘　要　知识图谱是图数据库的一个重要研究领域,它可以形式化地描述现实世界中的事物及其关系,但其不完整性和稀疏性阻碍了其在诸多领域中的应用。知识图谱推理技术旨在根据知识图谱中已有的知识来推断新的知识或识别错误的知识以完善知识图谱。尽管现有的各类推理方法可以获得部分有效知识,但仍然存在获取路径不全、忽略局部信息和引入噪声等问题。基于此,发现且明确提出路径连通性差问题并证明推理有效性与实体间路径连通比率呈正相关规律,进一步提出一种用于增强现有推理方法性能的双层框架 DL$^+$。模型第一层是知识增广器,主要利用社区发现算法在初始知识图谱上提取实体邻域信息,构建新知识以增广知识规模,然后设计社区剪枝优化去除构建时引入的噪声,最后将增广后的知识图谱抽取还原为与初始图谱表示相同的结构并输出到第二层以保证模型"即插即用"的特性。第二层是知识推理机,通过在知识增广后的图谱上进行学习推理以达到增强现有知识图谱推理模型的目的,使模型可以在图谱路径连通性比率较高的情况下获得更优的推理结果。最终在 4 个标准知识图谱数据集上进行的大量实验结果表明 DL$^+$ 算法可以有效缓解实体间路径连通性差的问题,与 9 类基准方法相比,所提算法的预测精度平均提高了 4.798%。

关键词:知识图谱;知识图谱推理;社区发现;路径连通性;链接预测
中图法分类号　TP391

DL$^+$:An Enhanced Double-layer Framework for Knowledge Graph Reasoning

WU Yuejia and ZHOU Jiantao
College of Computer Science(College of Software),Inner Mongolia University,Hohhot 010021,China
Engineering Research Center of Ecological Big Data,Ministry of Education,Hohhot 010021,China
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian,Hohhot 010021,China
Inner Mongolia Engineering Laboratory for Cloud Computing and Service Software,Hohhot 010021,China
Inner Mongolia Key Laboratory of Social Computing and Data Processing,Hohhot 010021,China
Inner Mongolia Key Laboratory of Discipline Inspection and Supervision Big Data,Hohhot 010021,China
Inner Mongolia Engineering Laboratory for Big Data Analysis Technology,Hohhot 010021,China

Abstract　As an important research field of the graph database,the knowledge graph(KG) can formally describe things and their relationships in the real world. However,its incompleteness and sparsity hinder its application in many fields. The knowledge graph reasoning(KGR) technology aims to complete the knowledge graph by inferring new knowledge or identifying wrong knowledge according to the existing knowledge in the knowledge graph. Although existing reasoning methods can obtain partially effective knowledge paths,there are still some problems such as incomplete acquisition paths,ignoring local information,introdu-

到稿日期:2022-10-23　返修日期:2023-03-06
基金项目:国家自然科学基金(62162046);内蒙古科技攻关项目(2021GG0155);内蒙古自然科学基金重大项目(2019ZD15);内蒙古自然科学基金(2019GG372);内蒙古大学研究生创新创业专项经费(11200-5223737)
This work was supported by the National Natural Science Foundation of China(62162046),Inner Mongolia Science and Technology Project(2021GG0155),Natural Science Foundation of Major Research Plan of Inner Mongolia(2019ZD15),Natural Science Foundation of Inner Mongolia,China(2019GG372) and Special Fund for Graduate Innovation and Entrepreneurship of Inner Mongolia University(11200-5223737).
通信作者:周建涛(cszjtao@imu.edu.cn)

cing noise. Based on this, this paper finds and explicitly proposes the problem of poor path connectivity, proves that the reasoning validity is positively correlated with the path connectivity ratio between entities, and further proposes a double-layer framework DL$^+$ which is used to enhance the performance of existing reasoning methods. The first layer is a knowledge augmenter, which mainly uses the community discovery algorithm to extract the entity neighborhood information on the initial KG and build new knowledge to expand the knowledge scale, and then designs a community pruning optimization method to remove the noise introduced in the construction. Finally, the augmented KG is extracted and restored to the same structure as the initial KG representation and output to the second layer to ensure the "plug-and-play" feature of the model. The second layer is a knowledge reasoner, which can enhance the existing KGR model by learning and reasoning on the KG after knowledge augmentation, so that the model can obtain better reasoning results when the graph path connectivity ratio is high. Finally, a large number of experimental results on four standard KG datasets show that the DL$^+$ can effectively alleviate the problem of poor path connectivity between entities, and improve the average prediction accuracy by 4.798% compare with nine types of benchmark methods.

Keywords Knowledge graph, Knowledge graph reasoning, Community discovery, Path connectivity, Link prediction

1 引言

知识图谱(KG)本质上是揭示实体之间关系的语义网络,它可以形式化地描述现实世界中的事物及其关系。基于知识图谱的研究方向主要有:知识抽取、知识表示、知识融合和知识推理。现有的知识图谱,如 Freebase[1],DBpedia[2] 和 WordNet[3]已经为许多自然语言处理(NLP)任务提供了有效支持,如知识问答[4]、知识抽取[5]和语义检索[6]等。然而,知识图谱通常是不完整且稀疏的,这一性质使其在上述任务中依然存在一定的局限性,例如知识图谱中知识的缺失可能导致知识问答不准确和无法检索对应语义等问题。

知识图谱推理(Knowledge Graph Reasoning,KGR)是解决上述不完整性问题的主要方法之一,即根据知识图谱中已有的知识来推断新知识,主要目标包括链接预测、属性预测、实体预测和关系预测等[7]。根据推理方法的不同,知识图谱推理可划分为基于规则的推理、基于分布式表示的推理、基于神经网络的推理以及混合推理[7]。

目前,知识图谱推理算法可以有效学习缺失的知识,但仍然存在一些问题。其中,基于规则的推理使用的传递性规则可能存在噪声,从而误导推理;基于分布式表示的推理将三元组向量化建立约束,可能导致级联误差;基于神经网络的推理缺乏可解释性;混合推理目前仅局限于混合两种方法且混合模式大多较浅层,如何混合多种互补方法还有待研究。

除上述问题外,本文发现并提出各类知识图谱推理方法中存在的重要问题为推理路径的不连通性,即直接在全局知识图谱上进行知识的挖掘与推理,而没有足够关注实体间的路径信息,导致路径不连通,进而忽略了实体间隐含的局部领域信息,使得三元组预测精度受限。着眼于此问题,本文进一步提出了一种"即插即用"的用于增强任意现有知识图谱推理方法性能的双层推理框架 DL$^+$(double-layer$^+$)。本文的主要工作如下:

1)发现并明确提出和定义现有各类知识图谱推理算法中存在的问题,即实体间路径连通性差的问题。

2)将知识图谱看作知识社区网络,通过划分知识图谱中的社区并利用剪枝优化后的实体路径以增广知识图谱。

3)提出"即插即用"型 DL$^+$ 框架,该框架有两层:第一层是知识增广器,通过在初始知识图谱上利用社区发现算法抽取实体集合实现路径可连通性并剪枝优化社区内实体关系去除噪声,使得 DL$^+$ 模型可以有效缓解路径连通性差的问题。最后将知识增广后的图谱抽取还原为与初始图谱表示相同的结构,使得 DL$^+$ 能够适应现有的不同推理算法的变化并能与其直接嵌套应用,进一步保证了框架"即插即用"的特性。第二层是知识推理机,通过将第一层的输出作为输入,在增广后的知识图谱上进行三元组学习可以增强任意现有知识推理方法的性能,即 DL$^+$ 模型可以在路径连通性比率较高的情况下更有效地获取得分以提高推理精度。

4)通过将各类经典知识图谱推理方法嵌入 DL$^+$ 框架后获取的大量实验结果表明,DL$^+$ 可以有效缓解知识图谱中路径连通性差的问题,并分别在 WN18,WN18RR,FB15K 以及 FB15K-237 数据集[8]上获得比基准平均高出 2.797%,18.274%,3.456%和 6.595%的预测精度,进一步证明了其具有较强的普适性,可以有效增强任意知识图谱推理框架的性能,从而证明了该模型的有效性。

2 相关工作

知识图谱本质上是语义网络的知识库,可以对事物及其相互关系进行形式化描述,其知识是对图谱中实体、关系和事实的统称。事实用三元组(h,r,t)表示,其中 h,r 和 t 分别代表头实体、关系和尾实体。虽然现有的知识图谱规模庞大,但其不完整性和稀疏性阻碍了其在语义分析[9]、知识问答[4]和个性化推荐[10]等诸多领域的应用。其中,知识图谱的不完整性通常指缺失整个三元组(h,r,t)或缺失部分头实体 h、尾实体 t 和它们之间的关系 r;知识图谱的稀疏性包括实体稀疏性和关系稀疏性。实体稀疏性指实体数量占事实总数的比例较小,关系稀疏性指关系数量占事实总数的比例较小[11]。如 Lv 等[12]所提到的,如表 1 所列,度代表每个实体的出度,可以用来指示知识图谱的不完整性和稀疏度水平。首先,以表 1 中数据集为例,当存在以万为数量级的实体数时,所对应的度数至少应为万数级,而实际知识图谱中的实体度数为十位数级甚至个位数级,由此可知知识图谱是非常不完整的且缺失大量的三元组信息。其次,表 1 中所列数据集的实体数及关系数占总事实数的比例较小,由此可知知识图谱实际上是极其稀疏的,尤其表现在关系稀疏性方面。

表 1 部分知识图谱数据集统计[12]

Table1 Partial knowledge graph dataset statistics

数据集	#实体	#关系	#事实	度	
				平均数	中位数
FB15K-237	14541	237	272115	19.27	14
WN18RR	40943	11	86835	2.19	2
NELL23K	22925	200	35358	2.21	1
WD-singer	10282	135	20508	2.35	2

知识图谱推理技术是缓解上述问题的有效方法之一，其包括两类知识，即已知的知识和由已知的知识推理得出的新知识，并通常以三元组（头实体，关系，尾实体）的形式进行表达[7]。根据推理角度的不同，基于规则的推理通常利用传递性规则挖掘路径，例如最典型的 PRA（Path Ranking Algorithm）算法[13]通过将路径看作特征来进行预测。基于分布式表示的推理主要利用低维向量表示三元组，将推理预测转换为向量操作，例如 Dettmers 等[8]提出的 ConvE 模型利用嵌入的二维卷积和多层非线性特征对知识图谱进行建模，使其可伸缩到大规模知识图谱。基于神经网络的推理直接使用神经网络建模知识图谱并在得到向量表示后进一步推理，例如，ProjE[14]是一个共享可变神经网络模型，其通过学习知识图谱实体和边的联合嵌入以及改变标准损失函数来填补 KG 中缺失的信息。混合推理通过混合多种推理方法进行推理，例如，为了解决关系路径的非均匀性问题，Sen 等[15]将基于规则的知识图谱推理方法与图嵌入相结合，进一步改善了推理结果并实现了两者的最佳效果。

尽管现有知识图谱推理方法能够挖掘丰富的知识，但仍然存在获取知识不完整、推理时间受限以及忽略局部信息等问题。本文着眼于研究知识图谱中边的不完整性，即关系的不完整性，发现且明确提出路径连通性差的问题并证明推理有效性与实体间路径连通性呈正相关规律，即路径连通性差会直接导致三元组的预测精度下降。由于大多数知识推理方法首先会对知识图谱中的路径进行搜索，实体间路径不连通可能会导致路径特征学习出现偏差，从而导致预测精度下降。基于此，本文进一步提出了一种“即插即用”型双层推理框架 DL$^+$，通过抽取并优化实体邻域信息进行知识增广以达到增强现有知识图谱推理方法性能的目的，使模型能够在缓解路径连通性差问题的同时提高推理预测精度。

3 问题定义

3.1 问题提出与定义

实体局部领域信息中包含大量有效信息[16]，计算实体间相关性是获取局部领域信息的有效方法之一。但在计算获取过程中，本文发现其获取路径受实体间路径连通性的影响，即在同一知识图谱中不同的路径连通性比率在同一方法下会产生不同的知识推理结果，且两者呈正相关规律。

本文基于 100 个实体及不同路径连通比率（10%，20%，40%，60%，80%，100%）在同一方法下进行模拟，以平均精度均值（Mean Average Precision，MAP）、平均倒数排名（Mean Reciprocal Rank，MRR）和 Hits@10 这 3 个衡量标准进行观测，得分越高代表预测结果越精确。模拟结果如图 1 所示，随着路径连通比率的增长，MAP，MRR 以及 Hits@10 得分均有所增长，尤其在路径连通比率为 10%～40%之间时。出现该现象的原因是：在路径稀疏程度较高时增广知识可以最大化获取有效路径；而随着稀疏程度的降低，即原有路径比率增高，之后进行增广的知识不可避免地会引入一定的噪声，即无关路径，从而导致衡量标准得分的增长比率下降。综上，模拟结果表明知识推理的预测精度随着路径连通性比率的增长而增长，证明路径连通性问题是影响推理结果的因素之一。

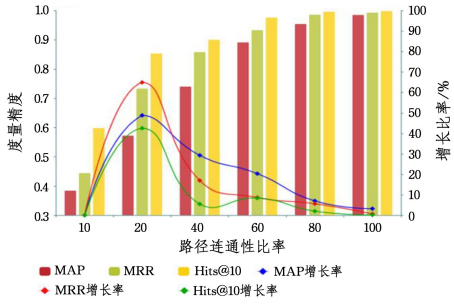


图 1 推理结果与路径连通性比率关系图

Fig. 1 Relational diagram of reasoning results vs. path connectivity ratio

本文进一步给出知识图谱通用性定义以及本文发现并提出的知识图谱中路径连通性的形式化定义。

定义 1（知识图谱） 给定一个知识图谱 $\mathcal{G}=(\mathcal{E},\mathcal{R},\mathcal{F})$，其中 $\mathcal{E}$ 表示实体集，$\mathcal{R}$ 表示关系集，$\mathcal{F}$ 表示事实三元组集。对于 $h,t\in\mathcal{E},r\in\mathcal{R}$ 存在 $h\xrightarrow{r}t$ 组成三元组 (h,r,t)，则事实三元组可表示为 $\mathcal{F}=\{(h,r,t)\mid h,t\in\mathcal{E},r\in\mathcal{R}\}$。

例如在图 2 所示知识图谱中，“发烧”和“隔离”等椭圆表示实体，“传播方式”和“症状”等有向箭头代表关系，（“新型冠状病毒肺炎”，“症状”，“发烧”）等三元组代表事实。

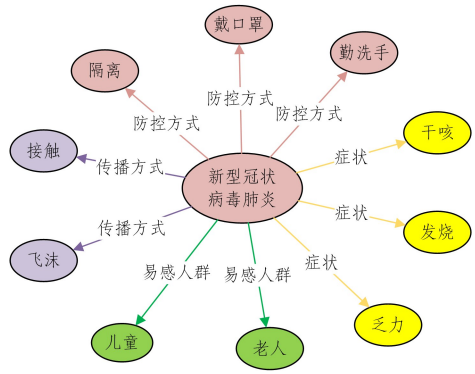


图 2 知识图谱示例图

Fig. 2 Sample graph of knowledge graph

定义 2（知识图谱路径连通性） 给定一个知识图谱 $\mathcal{G}=(\mathcal{E},\mathcal{R},\mathcal{F})$，对于任意实体 $\mathcal{E}_{\text{head}}\in\mathcal{E}$ 存在指向实体 $\mathcal{E}_{\text{tail}}\in\mathcal{E}$ 的可达路径 $p=\mathcal{E}_{\text{head}}\xrightarrow{r_1}\mathcal{E}_1\xrightarrow{r_2}\mathcal{E}_2\xrightarrow{r_3}\cdots\xrightarrow{r_n}\mathcal{E}_{\text{tail}}$，则 $\mathcal{E}_{\text{head}}$ 和 $\mathcal{E}_{\text{tail}}$ 是连通的，且称 $r_1r_2r_3\cdots r_n\in\mathcal{R}$ 为其连通路径；若不存在可达路径，则称其为不连通的。

例如图 2 中“新型冠状病毒肺炎”与“发烧”和“干咳”之间存在可达路径“症状”，因此称它们是连通的；而“发烧”与“老人”之间不存在可达路径，则称它们是不连通的。

3.2 问题实例

以 WN18 数据集中的数据片段为例，图 3（a）代表初始 KG 中的部分实体及关系分布情况，图 3（b）代表实体

A（“15286249”）的邻接子图信息，图 3（c）代表实体 B（“05250659”）的邻接子图信息。在知识图谱未进行知识增广前，A 实体与 B 实体之间不存在可达路径，即 A 与 B 不连通，这同时意味着 A 的所有邻接实体与 B 的所有邻接实体之间的路径也互不连通。这种现象在各类知识图谱中普遍存在且可能会导致丢失大量有效的局部邻域信息，尤其体现在基于规则的知识推理模型中，路径之间的不连通性直接影响推理规则的游走与生成，从而影响三元组的预测精度。

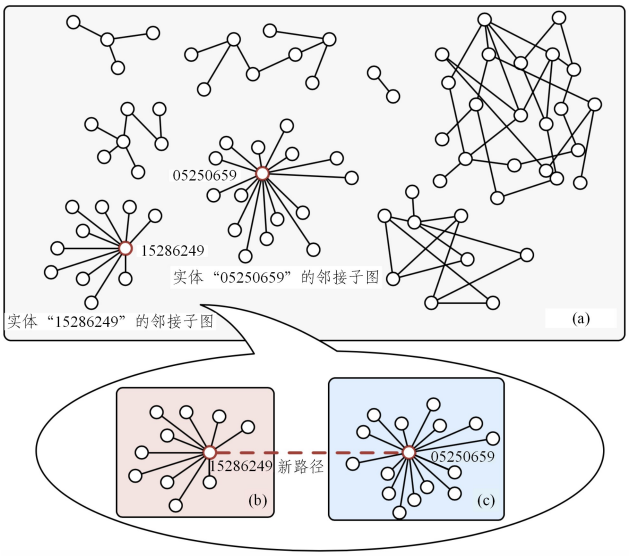


图 3 WN18 数据集中路径连通性差问题示例图（电子版为彩图）

Fig. 3 Example diagram of poor path connectivity problems in WN18 dataset

为解决上述路径连通性差的问题，本文提出一种增强型双层知识推理框架 DL$^+$，通过构建社区发现操作可以识别如图 3 中的实体 A 与实体 B 属于同一社区的现象，即实体 A 与实体 B 之间存在隐含关系，并通过增广实体 A 与实体 B 之间的新知识以强化初始 KG 的整体结构，如图 3 中红色虚线所示，以此提高 KG 中的路径连通性比率，并使其能够增强现有知识图谱推理方法的性能，达到提高推理精度的目的。

4 DL$^+$ 双层知识图谱推理方法

4.1 框架总览

为了解决路径连通性差的问题，本文进一步提出了一种用于增强任意现存知识推理方法性能的双层推理框架 DL$^+$，整体框架如图 4 所示，包括知识增广器以及知识推理机。对于一个特定的知识图谱，首先将三元组导入 DL$^+$ 框架第一层知识增广器，通过将知识图谱看作知识社区网络，经过三元组处理器、社区初始化和社区发现将关联性较强的实体划分到相同社区以增广初始知识图谱规模；然后利用社区剪枝优化去除冗余知识以缓解引入噪声的问题；接着通过社区抽取还原，使知识增广后的知识图谱保持与初始知识图谱相同的表示结构以保证框架“即插即用”的特性，即在无需额外操作的情况下，将第一层知识增广器的输出直接当作第二层知识推理机的输入，并通过在知识增广后的图谱上学习并推理获取得分，来达到增强现有知识推理方法性能的目的。

具体地，DL$^+$ 具有双层结构：第一层知识增广器首先利用社区发现算法提取并优化实体社区特征对缺失知识进行增广，然后利用社区剪枝优化方法去除噪声后将新知识添加到初始知识图谱结构中以提高路径连通性比率，接着利用社区抽取还原保持知识增广前后知识图谱表示结构的一致性，最后通过将知识增广后的图谱输出到第二层以增强后续推理性能；第二层知识推理机将第一层的输出当作输入，即在知识增广后的知识图谱上对知识进行学习并推理以增强现有推理方法的效果，使其在路径连通性比率较高的情况下获得精度更高的推理效果，具体如算法 1 所示。首先将预处理后的每个节点 E_i 划分到其邻接节点所在的社区中，当模块度不变时停止；然后将划分后的社区聚合成为一个点，重新构建网络；再根据社区内节点相关程度优化和构建新知识来增广知识图谱并设置其与增广前知识图谱表示结构相同，使得框架不受算法和参数变化的影响；最后在知识增广后的知识图谱上学习并获取推理结果得分，以增强现有知识图谱推理方法的性能。

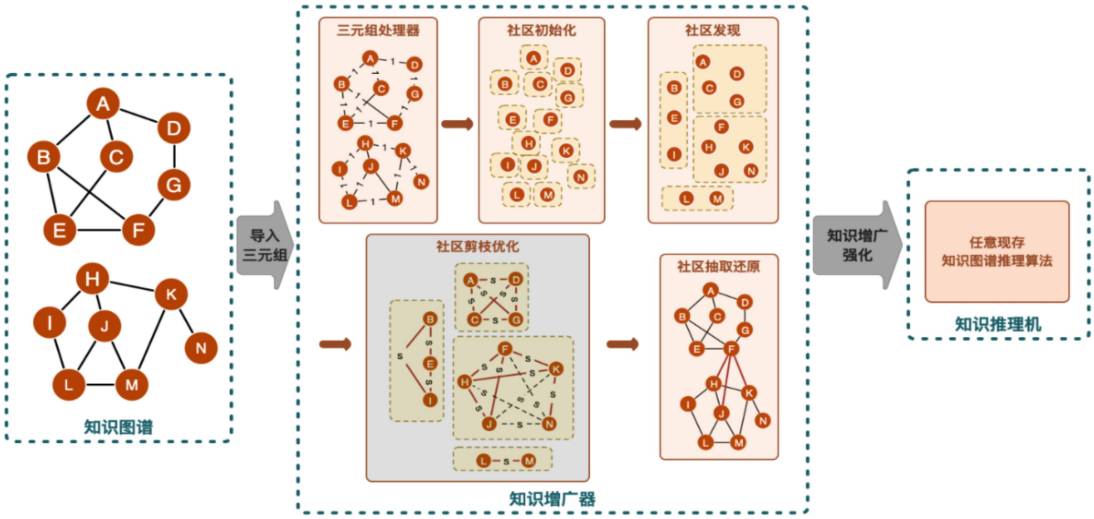


图 4 DL$^+$ 框架总览图

Fig. 4 Overview of DL$^+$ framework

算法 1 DL$^+$ 双层框架算法描述

输入：知识图谱 $\mathcal{G}=(\mathcal{E},\mathcal{R},\mathcal{F})$，其中 $\mathcal{E}$ 为实体集，$\mathcal{R}$ 为关系集，$\mathcal{F}=(\mathcal{E}_{\text{head}},\mathcal{R},\mathcal{E}_{\text{tail}})$ 为事实集，K 为剪枝数

输出：更新后的知识图谱 $\mathcal{G}'$

```
1. / * 预处理 * /
2. foriin Lenth(ℰ) do:
3.    if ℰi→ℰi+1 : / * 由定义 2:ℰi与ℰi+1之间存在连通路径 * /
4.       scorei←1
5.    else scorei←0
6.    ℱ. AddWeight(scorei)
7. end for
8. / * 社区发现 * /
9. loop:
10.   for i in Lenth(ℰ) do:
11.      ℰj←Divided(ℰi) / * 将节点划分为相邻节点ℰj的社区 * /
12.      Calculate_Modularity() / * 计算节点移动后模块度的变化 * /
13.         MoveNode() / * 移动到变化最大的社区 * /
14.      if 没有变化:break
15.   end for
16.   Calculate_Modularity_Partition() / * 计算分区模块度 * /
17.   Cluster() / * 使用分区对初始节点进行聚类 * /
18.   for i in Ci do: / * Ci为划分的社区子集 * /
19.      Recode_Number() / * 重新编码社区号 * /
20.      Edges←Building_Edges()/ * 重新构建边 * /
21.      for e in Edges do:
22.         Recalculate(ki) / * 重新计算边权值ki * /
23.         Save(e)
24.      end for
25.      Reset()
26.   end for
27.   if 模块度迭代没有变化:break
28. end loop
29. loop:
30.   for i in Cdo:
31.   j←i+1
32.      for 〈ℰ. (i),ℰ. (j)〉 in ℰ do:
33.            score〈i,j〉← Calculate_Score(ℰ. (i),ℰ. (j)) / * 计算实体
              间相似度 * /
34.            score_total. append(score〈i,j〉)
35.      end for
36.      Sort(score_total) / * 按节点相关度排序 * /
37.      for k in range(0,K) do:
38.         for s in score_total:
39.            NewKnowledge←〈ℰ. (s. i),ℰ. (s. j)〉 / * 在社区内实体
              之间构建新知识 * /
40.            Graph. Add(NewKnowledge) / * 将新知识增广到知识
              图谱中 * /
41.         end for
42.      end for
43.   endfor
44. end loop
45. / * 知识推理 * /
46. Reasoning()/ * 任意知识推理方法 * /
```

4.2 知识增广器

知识增广器位于本文所提 DL$^+$ 双层知识图谱推理框架的第一层，通过将知识图谱看作知识社区网络并使用社区发现算法将具有较强相关性的知识划分到相同社区后，为其添加新关联以提高知识图谱路径连通性比率；然后利用社区剪枝优化方法去除冗余关系，缓解噪声引入问题；最后使用社区抽取还原将新知识添加到初始知识图谱中完成知识增广并保持增广前后知识图谱表示结构不变，使模型具有“即插即用”特性，即能够直接将知识增广后的知识图谱输出到 DL$^+$ 框架第二层知识推理机中进行后续推理，从而增强现有知识图谱推理方法的性能。

具体地，在知识图谱中，实体与实体之间的联系构成了整个图结构，有的实体之间的连接较为紧密，有的实体之间则较为稀疏。因此，本文将这样的知识图谱看作知识社区网络，其中连接较为紧密的部分可以看作一个社区，其内部的节点之间有较为紧密的连接，而两个社区间的连接则相对稀疏。基于此，本文提出利用社区发现算法计算实体间相关度来补充缺失路径并利用知识增广器增强现有知识图谱推理方法的性能。具体地，如图 4 中知识增广器组件所示，三元组处理器主要导入三元组进行预处理并为其添加实体间权重数据，即根据定义 2(知识图谱中路径连通性)来判断节点对之间是否具备连通性，若连通则为其赋权重值为 1，若不连通则赋值为 0；社区初始化主要将各个节点代表的实体划分到不同社区中，如图 4 所示，每个黄色虚线框均代表不同的社区，可以发现初始状态下每个节点均独立地分布在不同的社区内等待下一步划分；社区发现通过计算模块度并迭代更新网络将不同实体划分到不同社区中，例如图 4 中节点按照模块度计算方法被划分到了 4 个不同的社区以增强节点间的关联程度；社区剪枝优化对划分到同一社区内的实体进行相关度计算，然后根据相关度数值仅保留其排名靠前的 K 条边完成剪枝，以提高实验速度并减少噪声的引入，图 4 中以 K 等于 6 为例，其中红色加粗实线代表最终保留的前 K 条边，黑色虚线代表被剪枝的边；最后社区抽取还原将新路径增广到知识图谱中以提高路径连通性，如图 4 中将社区发现并剪枝后的新知识(“F”，“H”)，(“F”，“J”)和(“F”，“K”)添加到知识图谱中以增广初始图谱数据，使得初始图谱中的两个子图之间增加了新的关系，从而缓解了其路径连通性差的问题。

特别地，为了评价社区划分的优劣，Newman[17] 提出模块度概念，即用模块度衡量社区划分的好坏，模块度越大代表社区划分效果越好。模块度公式如下所示：

$$Q=\frac{1}{2m}\sum_{i,j}\left[A_{i,j}-\frac{k_i k_j}{2m}\right]\delta(c_i,c_j) \tag{1}$$

其中，$m=\frac{1}{2}\sum_{i,j}A_{i,j}$ 代表图中的权重之和；$A_{i,j}$ 表示节点 i 和 j 之间的权重；$k_i=\sum_j A_{i,j}$ 表示与顶点 i 连接的边权重；c_i 代表顶点被分配到的社区；$\delta(c_i,c_j)$ 判断顶点 i 与 j 是否被划分到同一社区，如果是则返回 1，否则返回 0。模块度计算的简化形式为：

$$Q=\sum_c\left[\frac{\Sigma_{\text{in}}}{2m}-\left(\frac{\Sigma_{\text{tot}}}{2m}\right)^2\right] \tag{2}$$

其中，Σ_{in} 代表社区 c 内部的权重；Σ_{tot} 代表与社区 c 内部的点连接的边的权重，包括社区内部的边以及社区外部的边。

在划分社区后，为了缓解引入噪声和冗余边的问题，本文进一步提出了一种社区剪枝优化方法。首先计算实体间相关性分数。

$$score(\boldsymbol{i},\boldsymbol{j})=\text{sigmoid}(P\cdot\text{RELU}(\boldsymbol{i}\oplus\boldsymbol{j})+b) \tag{3}$$

其中，$\boldsymbol{i}$ 和 $\boldsymbol{j}$ 为实体向量，P 为训练参数，$\oplus$ 为组合运算，sigmoid 和 RELU 为激活函数，b 表示偏移量。然后对初始知识

图谱社区进行剪枝，即仅保留每个社区中节点对相关度最高的 K 条边。这样的设置可以保证相同节点划分的特征相同且显著降低了图的密度，由此可以降低计算内存成本并有效缓解引入噪声的问题。

最后，为保证 DL$^+$ 框架“即插即用”的特性，本文将经过社区发现和剪枝优化后的新知识以与初始知识图谱表示相同的结构增广到初始图谱中，并将经过三元组处理器的知识重新遍历并抽取还原为初始形态，这样的设置可以使所提模型不受算法和参数变化的影响，即可以在知识增广后的图谱中以“即插即用”式的嵌套方式与现有知识图谱推理模型结合应用，从而增强其推理性能。

值得注意的是，本文所提 DL$^+$ 框架可以使用任意社区发现算法，其中经典的社区划分 FastUnfolding 算法[18] 为无监督方法，具有易于实现和复杂度为线性等优点，并且其固有的多层次特性可以规避模块化的分辨率限制问题。由于本文实验数据集中三元组数量达到十万级，因此 DL$^+$ 利用其来划分社区。

综上，针对知识图谱中路径连通性差的问题，本文提出了一种“即插即用”型双层知识图谱推理框架 DL$^+$。其第一层知识增广器通过将知识图谱看作社区网络并利用社区发现算法建立实体间新关联以强化初始知识图谱网络，然后利用社区剪枝优化操作去除冗余知识，缓解引入噪声问题，最后利用社区抽取还原组件将增广知识后的知识图谱以与初始知识图谱表示结构相同的形态输出到第二层知识推理机中，从而增强现有知识图谱推理方法的性能。实验结果表明，本文所提 DL$^+$ 框架中第一层知识增广器组件可以构建同属相同社区的新知识以增强现有推理方法的性能，能够有效提高路径连通比率以及预测精度。

4.3 知识推理机

知识推理机位于 DL$^+$ 双层知识图谱推理框架的第二层，其输入为第一层知识增广器的输出，这样的设置可以保证其在知识图谱路径连通性比率较高的情况下获取三元组得分进行学习推理，以增强任意现有知识图谱推理方法的预测性能。

具体地，得益于 DL$^+$ 框架第一层知识增广器中社区抽取还原组件的设置，本文将增广后的知识图谱还原为与初始知识图谱表示相同的结构，使其具有“即插即用”特性，即 DL$^+$ 框架第一层知识增广器的输出可以直接作为第二层知识推理机的输入，由此可以在框架第二层中基准实验设置不变的情况下直接插入第一层进行嵌套推理，使得知识增广器能够适应不同算法和参数等变量的变化且能够嵌套应用于任意现有知识图谱推理模型，进一步保证了 DL$^+$ 框架的灵活性及普适性。特别地，与各类知识图谱推理模型进行嵌套实验后的增强效果将在第 5 节实验及结果中详细展示。

综上，本文提出的 DL$^+$ 框架由两层结构组成：知识增广器和知识推理机。其中，知识增广器位于框架的第一层，其通过社区发现、社区剪枝优化和社区抽取还原等 5 部分组件有效缓解了知识图谱中路径连通性差的问题以及在提高路径连通比率中引入噪声的问题，接着将增广后的图谱还原为与初始图谱相同的表示结构以保证其“即插即用”特性并输出到第二层以增强推理性能；知识推理机位于框架的第二层，其通过将第一层的输出直接应用为其输入，在现有任意基准算法参数无需变换的情况下保证了 DL$^+$ 框架的灵活性和普适性，使其能够在路径连通比率较高，即在有效缓解知识图谱不完整性和稀疏性的情况下获取三元组得分以提高各类推理方法预测精度。最终实验结果表明，本文提出的“即插即用”型双层知识图谱推理框架 DL$^+$ 可以有效缓解知识图谱中路径连通性差的问题并进一步提高任意现有知识图谱推理方法的预测精度。

5 实验及结果

5.1 数据集与数据预处理

为了充分验证 DL$^+$ 框架的有效性，实验采用 4 个不同的知识图谱推理任务标准数据集：WN18，WN18RR，FB15K 以及 FB15K-237[13]，具体如表 2 所列。其中，WN18RR 和 FB15K-237 为低维数据集，它们分别是在 WN18 和 FB15K 高维数据集的基础上去掉了有逆三元组而得到的数据集，通过添加这两种数据集可以在不同数据维度下更好地评估模型性能。

表 2 实验数据集

Table 2 Experimental datasets

数据集	# 实体	# 关系	训练集	验证集	测试集
WN18	40943	18	141442	5000	5000
WN18RR	40943	11	86835	3034	3134
FB15K	14951	345	483142	50000	59071
FB15K-237	14541	237	272115	17535	20466

为了更方便且快速地对知识图谱进行社区发现，实验首先对知识图谱中的三元组数据进行预处理，处理流程如图 5 所示。具体地，实验首先将数据集中的三元组文件进行拆分，然后根据实体节点列表中各节点的编号，将每对三元组的头尾节点进行编号处理并根据路径连通性初始化其关系权重为 1 或 0。以 WN18 数据集为例，具体预处理结果如表 3 所列。

表 3 WN18 数据集中数据预处理示例

Table 3 Example of data preprocessing in WN18 dataset

处理前三元组 〈头节点，尾节点，关系〉	处理后三元组 〈头编号，尾编号，权重〉
〈03964744，04371774，_hyponym〉	〈 27536，33729，1〉
〈00260881，00260622，_hypenym〉	〈 25546，10838，1〉
〈02199712，02188065，_meber_holonym〉	〈1213，38780，1〉
…	…

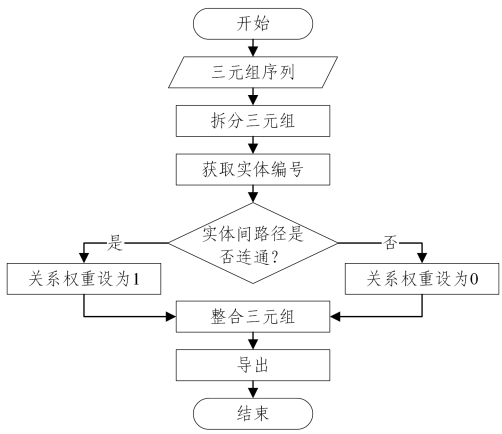


图 5 数据预处理流程图

Fig. 5 Flow chart of data preprocessing

5.2 实验环境配置与参数设置

本文实验基于 Windows 操作系统，使用 Python 编程语言、PyCharm 编程工具以及 Pytorch 架构进行实验，显卡为 NVIDIA 双 3090 Ti，内存为 Kingston 32GB，具体环境与配置如表 4 所列。

表 4 实验环境与配置总览

Table 4 Overview of experimental environment and configuration

配置名称	配置值
操作系统	Windows
编程语言	Python
编程工具	PyCharm
架构	PyTorch
显卡	NVIDIA 3090 Ti * 2
内存	Kingston 32GB * 2
固态硬盘	WesternDigital 2TB
机械硬盘	WesternDigital 14TB
处理器	Intel 13900K

为了更有效地验证所提 DL$^+$ 框架的性能，得益于其“即插即用”的特性，实验在推理机中保持与不同基准方法相同的参数设置以获得更加准确的对比结果，具体实验参数设置如表 5 所列。特别地，实验将实体训练维度设为 500 维，神经网络层数设为 6，隐藏层维度设为 512，内层维度设为 512，注意力头设为 32，优化器超参数设为 1×10^{-8} 等。另外，为了加快训练速度并避免过拟合现象，实验使用 L2 正则化约束剪枝参数 K 并将惩罚系数设为{0，0.0005，0.001}。

表 5 实验参数设置

Table 5 Experimental parameter settings

参数名称	参数值
训练维度	500 维
神经网络层数	6
隐藏层维度	512
内层维度	512
注意力头	32
批量大小	1024
dropout	{0.1，0.2，…，0.6}
注意力 *dropout*	{0.1，0.2，…，0.6}
最高学习率	3×10^{-4}
学习率衰减	线性
ϵ	1×10^{-8}
惩罚系数	{0，0.0005，0.001}

5.3 基准与评估标准

实验将不同种类知识图谱推理方法中具有代表性的经典或最先进的算法作为基准，并在执行链接预测任务时将其与 DL$^+$ 框架嵌套后进行对比以充分验证模型的有效性。其中，链接预测任务的目的是预测缺失的头实体 h、尾实体 t 或一个三元组(h，r，t)的关系 r。

为了更加全面地评估所提 DL$^+$ 推理框架的性能，实验基准使用各类知识图谱推理实现方法下的经典或最优方法，包括基于规则的传统局部推理算法 SFE[19]、基于分布式表示的最优推理算法 ConvE[8]、基于神经网络中强化学习的经典推理算法 DeepPath[20] 以及基于混合图神经网络与分布式表示的经典推理算法 R-GCN[21]、KG-BERT[22] 和最优算法 AWR-GCN[23]。特别地，为了进一步证明 DL$^+$ 框架的有效性，本文额外增加了目前知识图谱推理领域的最新算法 SimKGC[24]，IBL[25] 以及 PUDA[26] 作为实验基准。具体描述如下：

1)SFE：一种基于规则的经典局部推理算法。SFE 对节点周围的子图进行特征刻画，并从子图中提取特征。其表达能力很强，允许提取比图中两个节点之间的路径丰富得多的特征以提高推理精度。

2)ConvE：一个神经链接预测模型。其中输入实体和关系之间的相互作用由卷积和全连接层进行建模，其主要特点是分数由二维形状嵌入的卷积所定义。

3)DeepPath：首次将强化学习引入知识图谱路径建模。其使用一个完全连接的神经网络参数化其策略，并使用人工奖励标准，包括全局精度、效率和多样性来评估路径质量。

4)R-GCN：首次将图神经网络用于关系图建模并混合分布式表示方法以增强推理性能。其通过引入参数共享技术和加强稀疏性约束使其可以应用到具有大量关系的知识图谱中。

5)KG-BERT：将知识图谱中的三元组视为文本序列并提出双向编码器对三元组进行建模。其以三元组的实体和关系描述作为输入，然后利用语言模型计算三元组的评分函数进行推理。

6)AWR-GCN：一种注意力加权关系图卷积网络。其遵循“编码器-解码器”架构，且被用作编码器为每个邻域中的实体分配不同的权重以获得更加丰富的关系信息进行推理。

7)SimKGC：通过引入批内负采样(In-batch Negatives)、批前负采样(Pre-batch Negatives)和困难负样本的自我负采样(Self-negatives)这 3 种类型的负采样以提高对比学习效率实现推理。

8)IBL：基于分布式表示模型中的翻译模型来建立实例在某一属性上的等价关系，即实例等价性，并与翻译模型相结合学习实体-关系表示以实现知识图谱推理任务。

9)PUDA：通过正标签-无标签学习(Positive-unlabeledLearning)和对抗数据增强来解决负样本伪阴性问题和正样本稀疏性问题，从而提高知识图谱推理精度。

实验使用知识推理任务的标准度量：平均倒数排名 MRR 以及 Hits@K 来评价实验的性能，对于每一项指标，分数越高表明效果越好。*MRR* 描述如下：

$$MRR=\frac{1}{|T|}\sum_{i=0}^{|T|}\frac{1}{rank_i}=\frac{1}{|T|}\left(\frac{1}{rank_1}+\frac{1}{rank_2}+\cdots+\frac{1}{rank_{|T|}}\right) \tag{4}$$

其中，T 是三元组集合，$|T|$ 表示三元组集合个数，$rank_i$ 是第 i 个三元组的链接预测排名。

Hits@K 指标描述如下：

$$Hist@K=\frac{1}{|T|}\sum_{i=1}^{|T|}\mathbb{I}(rank_i\leqslant K) \tag{5}$$

其中，$\mathbb{I}$ 为指示函数，当 $rank_i\leqslant K$ 时，$\mathbb{I}$ 值为 1；当 $rank_i>K$ 时，$\mathbb{I}$ 值为 0。特别地，为了更全面地与各类基准进行对比实验，本文实验中 K 值的设置与基准方法保持一致并设置为 1，3 和 10。

5.4 实验结果与分析

5.4.1 知识增广器实验结果与分析

知识增广器位于 DL$^+$ 双层框架的第一层，其在 WN18

数据集上的部分运行示例结果如表 6 所列，实验共发现 65 个社区，经过剪枝处理后在社区内部实体之间构建新知识来强化初始知识图谱的结构，以缓解路径连通性差的问题，从而进一步增强任意现有知识推理方法的性能。

表 6 WN18 数据集上获取社区信息部分数据

Table 6 Partial data of community information obtained on WN18 dataset

社区编号	社区成员数量		社区成员
	剪枝前	剪枝后	剪枝后
1	946	597	05250659,15286249,00866702…
2	533	269	10480253,04581829,11722466…
3	187	114	01947887,10218043,06453849…
4	246	152	11672400,02186360,12861892…
5	103	85	01904845,02578510,09073938…
…	…	…	…

为证明知识增广器对于增强知识图谱中路径连通性的有效性，如图 6 所示，实验在不同数据集上使用知识增广器构建缺失路径。计算并对比构建前后路径连通性比率的增长结果可以发现，增广后的知识图谱中路径连通性比率较初始状态均有所上升并呈正相关趋势。另外，由于 WN18RR 数据集相较于 WN18 数据集更为稀疏，因此 WN18RR 数据集上的连通性增长比率明显高于 WN18 数据集上的增长比率，同理 FB15K-237 数据集上的连通性增长比率要高于 FB15K 数据集上的增长比率。由此证明了所提 DL$^+$ 框架中知识增广器组件通过将实体划分到不同社区并根据实体相关性系数构建新知识以增广初始知识图谱结构可以显著提升实体间路径连通性比率，从而有效缓解知识图谱中路径连通性差的问题。

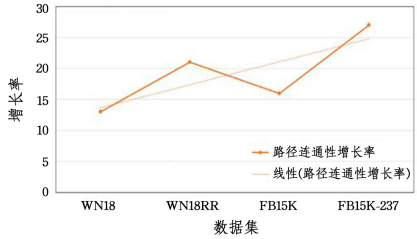


图 6 知识图谱增广前后路径连通性增长率对比图

Fig. 6 Comparison diagram of path connectivity growth rate before and after knowledge graph augmentation

5.4.2 知识推理机实验结果与分析

知识推理机位于 DL$^+$ 双层推理框架的第二层，实验将第一层知识增广器的输出作为知识推理机的输入并对 4 个数据集进行链接预测的整体结果如表 7 所列，其中“O”代表使用基准方法进行实验，“DL$^+$”代表将基准方法嵌入所提 DL$^+$ 框架进行实验。

表 7 各数据集下嵌入 DL$^+$ 框架与基准方法的性能比较

Table 7 Performance comparison between embedded DL$^+$ framework and baseline methods on each dataset

算法	设置	数据集															
		WN18				WN18RR				FB15K				FB15K-237			
		MRR	*Hits@K*			*MRR*	*Hits@K*			*MRR*	*Hits@K*			*MRR*	*Hits@K*		
			@1	@3	@10		@1	@3	@10		@1	@3	@10		@1	@3	@10
SFE	O	0.870	0.530	0.790	0.910	0.690	0.480	0.730	0.850	0.810	0.660	0.720	0.880	0.570	0.490	0.600	0.730
	DL$^+$	**0.910**	**0.660**	**0.850**	**0.940**	**0.780**	**0.650**	**0.840**	**0.910**	**0.900**	**0.730**	**0.860**	**0.920**	**0.700**	**0.630**	**0.740**	**0.810**
ConvE	O	**0.943**	0.935	0.946	0.956	0.430	0.400	0.440	0.520	0.657	0.558	0.723	**0.831**	0.325	**0.237**	0.356	0.501
	DL$^+$	0.937	**0.944**	**0.958**	**0.969**	**0.489**	**0.472**	**0.497**	**0.556**	**0.664**	**0.595**	**0.731**	0.829	**0.331**	0.231	**0.394**	**0.555**
Deep Path	O	0.648	0.703	0.832	0.864	0.224	0.131	0.157	0.207	0.625	0.534	0.749	0.856	0.619	0.52	0.74	0.835
	DL$^+$	**0.674**	**0.715**	**0.844**	**0.87**	**0.231**	**0.158**	**0.201**	**0.294**	**0.633**	**0.548**	**0.764**	**0.862**	**0.627**	**0.534**	**0.763**	**0.886**
R-GCN	O	0.819	0.697	0.929	0.964	0.123	0.08	0.137	0.207	0.696	**0.601**	0.760	0.842	0.248	0.153	0.258	0.414
	DL$^+$	**0.825**	**0.702**	**0.933**	**0.977**	**0.218**	**0.126**	**0.144**	**0.210**	**0.697**	0.594	**0.762**	**0.871**	**0.253**	**0.161**	**0.263**	**0.495**
KG-BERT	O	0.958	0.944	0.951	0.952	0.438	0.412	0.465	0.524	0.811	0.761	0.84	0.902	0.268	0.197	0.289	0.42
	DL$^+$	**0.961**	**0.947**	**0.955**	**0.960**	**0.441**	**0.425**	**0.472**	**0.539**	**0.823**	**0.771**	**0.852**	**0.914**	**0.269**	**0.208**	**0.299**	**0.427**
AWR-GCN	O	0.961	0.951	0.963	0.961	0.446	0.395	0.484	0.561	0.803	0.752	0.821	0.897	0.452	0.334	0.461	**0.572**
	DL$^+$	**0.962**	**0.960**	0.963	**0.967**	**0.449**	**0.401**	**0.490**	**0.568**	**0.818**	**0.760**	**0.834**	**0.906**	**0.456**	**0.337**	**0.465**	0.570
Sim KGC	O	0.963	0.945	0.969	0.972	0.665	0.573	0.719	0.808	0.678	0.493	0.694	0.805	0.333	0.238	**0.359**	0.507
	DL$^+$	**0.965**	**0.948**	**0.970**	0.972	**0.666**	**0.575**	**0.721**	**0.813**	**0.679**	**0.497**	**0.696**	**0.810**	**0.335**	**0.242**	**0.365**	**0.511**
IBL	O	0.956	0.948	0.959	0.962	0.432	0.405	0.459	0.493	0.630	0.502	0.659	0.743	0.319	0.238	**0.359**	0.498
	DL$^+$	**0.959**	**0.950**	**0.962**	**0.963**	**0.433**	**0.409**	**0.462**	**0.500**	**0.634**	**0.504**	**0.662**	**0.745**	**0.322**	**0.241**	0.357	**0.505**
PUDA	O	0.957	0.944	0.961	0.965	0.478	**0.435**	0.496	0.583	0.752	0.585	0.793	0.886	0.368	0.266	0.403	0.577
	DL$^+$	**0.959**	**0.945**	**0.963**	**0.968**	**0.480**	0.434	**0.497**	**0.586**	**0.754**	**0.589**	**0.802**	**0.890**	**0.369**	**0.272**	**0.409**	**0.581**

实验结果表明，使用 DL$^+$ 框架可以分别在 WN18，WN18RR，FB15K 以及 FB15K-237 数据集上获得比各类基准平均高出 1.667%，10.438%，2.244% 和 4.841% 的预测精度，并且在所有数据集上分别获得了比 SFE，ConvE，DeepPath，R-GCN，KG-BERT，AWR-GCN，SimKGC，IBL 和 PUDA 这 9 类基准平均高出 15.124%，5.243%，7.591%，10.997%，1.671%，0.886%，0.540%，0.565% 以及 0.562% 的推理精度，证明了 DL$^+$ 框架第一层知识增广器可以有效缓解知识图谱中路径连通性差的问题，且将其以“即插即用”型方式应用于第二层知识推理机可以在不同数据维度和不同基准方法下提升链接预测任务的推理精度，进一步证明了所提框架的有效性。特别地，从各类基准获得的提升精度百分比可以发现，DL$^+$ 框架与基于路径规则生成的 SFE 算法、DeepPath 算法以及 R-GCN 算法相结合可以获得更加显著的改进效果，这是由于初始知识图谱中路径之间的不连通性直接影响了该类方法下推理规则的游走与生成，证明通过 DL$^+$ 知识增广器组件对知识图谱中缺失的知识进行增广后再学习推理可以强化知识图谱的连通性，从而在路径连通性较高的情况下达到增强任意现有知识图谱推理方法性能的目的，以此获得更优的预测效果。

特别地，对于 WN18 数据集中的每一种关系，表 8 列出了嵌入 DL$^+$ 框架与基准的 *MRR* 分数对比结果。总体上，9 类基准的预测精度平均提升了 1.154%，且在每种关系下分别提升了 1.075%，3.200%，0.943%，0.958%，1.302%，0.64%，2.509%，−0.001%，1.062%，0.880%，0.925%，0.421%，2.106%，1.369%，1.263%，0.251%，0.742%，1.132%。从对比结果可知，在绝大多数关系中，通过嵌入 DL$^+$ 框架可以取得比基准算法更优的预测结果，证明了其通过知识增广能够增强现有知识图谱推理方法的性能，即在有效缓解路径连通性差问题的同时提高链接预测精度。另外，嵌套 DL$^+$ 框架后的实验模型在每种关系下获得的推理精度比 SFE，ConvE，DeepPath，R-GCN，KG-BERT，AWR-GCN，SimKGC，IBL 和 PUDA 这 9 类基准分别平均提升了 5.035%，−0.669%，4.089%，0.693%，0.331%，0.105%，0.218%，0.344%以及 0.242%，证明了 DL$^+$ 框架与基于路径生成的 SFE 算法和 DeepPath 算法相结合时可以获得更加显著的提升效果，进一步证明了 DL$^+$ 框架通过引入知识增广器可以显著提升知识图谱中的路径连通性比率并缓解其不完整性和稀疏性问题，使得该类基准算法可以在生成游走路径时学习到更丰富的特征，益于推理。

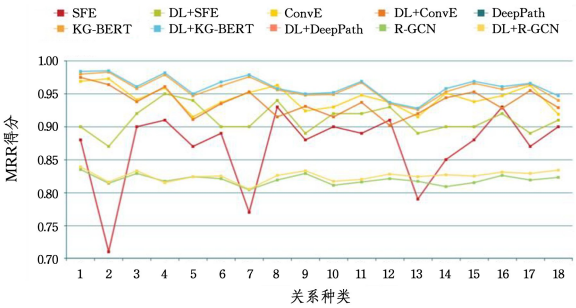
表 8 WN18 数据集每种关系下嵌入 DL$^+$ 框架与基准的 *MRR* 分数对比结果

Table 8 Comparison results of *MRR* scores between DL$^+$ and baselines embedded in each relation on WN18 dataset

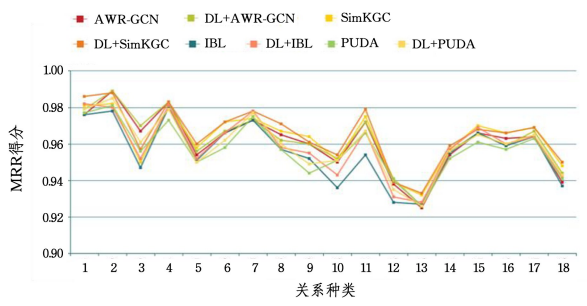
#	关系名称	算法															
		SFE		ConvE		DeepPath		R-GCN		KG-BERT		AWR-GCN		SimKGC		IBL		PUDA	
		O	DL$^+$	O	DL$^+$	O	DL$^+$	O	DL$^+$	O	DL$^+$	O	DL$^+$	O	DL$^+$	O	DL$^+$	O	DL$^+$
1	_member_of_domain_topic	0.88	**0.90**	0.969	**0.975**	0.658	**0.686**	0.835	**0.839**	0.980	**0.984**	0.976	**0.979**	0.981	**0.986**	0.976	**0.982**	0.977	**0.979**
2	_member_meronym	0.71	**0.87**	**0.973**	0.964	0.634	**0.669**	0.814	**0.816**	0.983	**0.985**	0.989	0.989	0.982	**0.988**	0.978	**0.980**	0.981	**0.985**
3	_derivationally_related_form	0.90	**0.92**	**0.941**	0.938	0.651	**0.675**	0.829	**0.833**	0.958	**0.961**	0.967	**0.970**	0.950	**0.957**	0.947	**0.952**	0.956	**0.961**
4	_member_of_domain_region	0.91	**0.95**	0.959	**0.961**	0.649	**0.668**	**0.817**	0.815	0.979	**0.982**	0.983	0.983	0.980	**0.983**	0.980	**0.981**	0.973	**0.979**
5	_similar_to	0.87	**0.94**	**0.915**	0.911	0.643	**0.664**	0.824	0.824	0.947	**0.950**	0.954	**0.956**	0.958	**0.960**	0.951	**0.952**	0.950	0.950
6	_hypernym	0.89	**0.90**	**0.937**	0.935	0.652	**0.673**	0.821	**0.825**	0.962	**0.968**	0.966	**0.967**	0.972	0.972	0.966	0.966	0.958	**0.962**
7	_member_holonym	0.77	**0.90**	0.952	**0.953**	0.645	**0.671**	0.804	**0.805**	0.976	**0.979**	0.973	0.973	0.974	**0.978**	0.973	**0.978**	0.975	**0.977**
8	_instance_hypernym	0.93	**0.94**	**0.963**	0.915	0.648	**0.663**	0.819	**0.826**	0.956	**0.958**	**0.965**	0.962	0.967	**0.971**	0.957	**0.958**	0.957	**0.960**
9	_member_of_domain_usage	0.88	**0.89**	0.924	**0.931**	0.637	**0.678**	0.829	**0.833**	0.948	**0.950**	0.960	0.960	**0.964**	0.961	0.952	**0.955**	0.944	**0.949**
10	_synset_domain_topic_of	0.90	**0.92**	**0.930**	0.915	0.643	**0.675**	0.811	**0.817**	0.949	**0.952**	0.950	**0.952**	0.951	**0.954**	0.936	**0.943**	0.951	0.951
11	_hyponym	0.89	**0.92**	**0.948**	0.937	0.649	**0.672**	0.816	**0.820**	0.967	**0.969**	0.972	0.972	0.975	**0.979**	0.954	**0.967**	0.966	**0.967**
12	_instance_hyponym	0.91	**0.93**	**0.937**	0.902	0.651	**0.678**	0.821	**0.828**	0.935	**0.937**	0.938	**0.941**	0.939	0.939	0.928	**0.931**	**0.940**	0.935
13	_synset_domain_usage_of	0.79	**0.89**	0.915	**0.920**	0.643	**0.671**	0.817	**0.824**	0.926	**0.928**	0.925	**0.926**	0.932	**0.933**	0.927	**0.928**	0.926	0.926
14	_has_part	0.85	**0.90**	**0.951**	0.944	0.652	**0.674**	0.809	**0.827**	0.953	**0.958**	0.954	**0.956**	0.957	**0.959**	0.955	**0.957**	0.952	**0.956**
15	_verb_group	0.88	**0.90**	0.938	**0.953**	0.644	**0.679**	0.815	**0.825**	0.966	**0.969**	0.966	0.966	**0.970**	0.968	0.966	**0.969**	0.961	**0.965**
16	_part_of	**0.93**	0.92	**0.947**	0.928	0.649	**0.677**	0.826	**0.831**	0.957	**0.961**	**0.963**	0.959	0.966	0.966	0.959	**0.960**	0.957	**0.960**
17	_synset_domain_region_of	0.87	**0.89**	**0.963**	0.955	0.653	**0.675**	0.819	**0.829**	0.965	**0.966**	0.964	**0.967**	0.969	0.969	0.964	0.964	0.963	**0.965**
18	_also_see	0.90	**0.91**	0.919	**0.929**	0.656	**0.685**	0.823	**0.834**	0.94	**0.947**	0.939	**0.944**	0.948	**0.950**	0.937	**0.942**	0.941	**0.943**
	总体	0.87	**0.91**	**0.943**	0.937	0.648	**0.674**	0.819	**0.825**	0.958	**0.961**	0.961	**0.962**	0.963	**0.965**	0.956	**0.959**	0.957	**0.959**

5.5 路径连通性影响分析

为了进一步证明 DL$^+$ 框架解决路径连通性差问题的有效性，本文将表 8 中的数据进一步转换为折线图，更直观地对比嵌入 DL$^+$ 框架后与基准之间的 MRR 分数，结果如图 7 所示。



(a)不同种类基准方法对比结果



(b)最新基准方法对比结果

图 7 WN18 数据集上每种关系下嵌入 DL$^+$ 框架与基准的 MRR 分数对比结果

Fig. 7 Comparison results of MRR scores between DL$^+$ and baselines embedded in each relation on WN18 dataset

可以发现，在连通性增强后的知识图谱上进行预测分析可以挖掘更多有效信息来辅助推理，以提高预测精度，证明了 DL$^+$ 框架利用知识增广器可以有效缓解知识图谱推理算法中存在的路径不连通性问题并由此提高链接预测精度，达到增强现有推理方法性能的目的。

由图 8 可以进一步发现，在具有不同连通性比率的知识图谱中进行推理的整体效果各不相同，其中 WN18RR 数据集相较于 WN18 数据集去除了其中的逆三元组，且事实数也有所减少，因此其连通性较差，在相同基准方法下整体数据预测结果也相对较差；同理，由于路径连通性比率降低，在相同基准方法下，FB15K-237 数据集上的预测结果与 FB15K 数据集相比性能也有所降低，由此可以证明在同一个知识图谱下不同的路径连通性比率会直接影响三元组的预测精度，即证明路径连通性是影响知识图谱推理精度的因素之一。

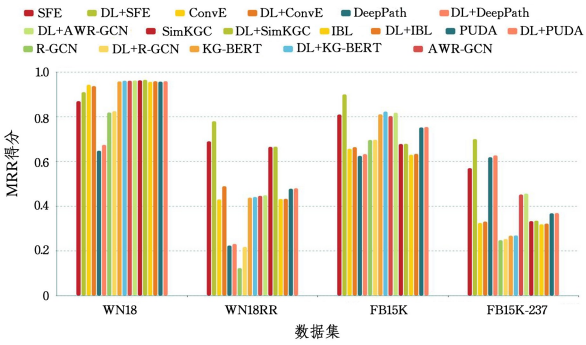


图 8 不同数据集下嵌入 DL$^+$ 框架与基准的 MRR 分数对比结果

Fig. 8 Comparison results of MRR scores between DL$^+$ and baselines embedded on different datasets

但利用 DL$^+$ 框架进行知识增广后，9 类基准方法在 WN18RR 数据集上的预测结果（平均提升 10.438%）相较于在 WN18 数据集（平均提升 1.667%）上的预测结果提升效果更加显著；同理，由于 DL$^+$ 框架使用知识增广器可以有效提升知识图谱中路径连通性比率，因此各基准模型在 FB15K-237 数据集上的预测结果（平均提升 4.841%）与在 FB15K 数据集上的预测结果（平均提升 2.244%）相比提升比率更加显著，证明本文提出的 DL$^+$ 框架可以通过知识增广器有效提升知识图谱中路径连通性比率，使得链接预测精度在路径更稀疏的知识图谱下能够获得更显著的提升效果，进一步证明 DL$^+$ 双层推理框架可以通过第一层知识增广器来缓解初始图谱中路径连通性差的问题，以增强第二层知识推理机中现有任意知识图谱推理方法的预测性能。

5.6 DL$^+$ 框架普适性证明

通过保持知识增广前后知识图谱表示结构的一致性可以保证框架具有“即插即用”的特性，使其可以在不改变参数等变量的前提下直接与任意现有知识图谱推理模型进行嵌套使用。在这一背景下，为了进一步证明本文所提框架的普适性，即证明现有任意知识图谱推理模型以“即插即用”式嵌套方式与所提框架结合推理后其性能是否均能有所提升，本文在 FB15K 数据集上，将 DL$^+$ 与现有具有代表性的且覆盖全部推理类型的 27 个知识图谱推理方法进行嵌套推理，其中包括：

1）基于规则的推理方法：PRA[13]，CPRA[27]，SFE[19] 以及 HiRi[28]。其中 PRA 以及 CPRA 基于全局知识图谱结构进行推理，而 SFE 和 HiRi 则引入更有助于推理的局部结构来降低推理代价。

2）基于分布式表示的推理方法：TransE[22]，TransH[22]，TransR[22]，TransD[22]，TransA[29]，RESCAL[30]，TransG[31] 以及 ConvE[8]。这类方法均通过向量化知识图谱进行推理，使得其学习的向量表示更有助于进行实体和关系的预测。

3）基于神经网络的推理方法：ProjE[14]，DistMult[22]，ComplEx[22]，DeepPath[20]，MINERVA[32]，DIVINE[32] 以及 PUDA[26]。该类方法利用神经网络模型建模并学习实体间的推理过程，使其能够充分学习路径的向量表示以强化推理。

4）混合推理方法：R-GCN[21]，RA-GCN[33]，CompGCN[34]，AWR-GCN[23]，Att _ GCN[35]，KG-BERT[22]，SimKGC[24] 和 IBL[25]。该类方法通过混合两类不同的推理方法以实现优势互补，从而达到有效提升知识推理效果的目的。

基于上述基准的总体实验结果如图 9 所示，可以发现在嵌套 DL$^+$ 框架后获取的 MRR 得分均比原始算法得分更高，由此可以充分且全面地论证所提 DL$^+$ 框架通过在知识推理前使用知识增广器为同一类实体构建新知识关联可以降低初始推理图谱的稀疏性，使得其无论在何种方法下（基于规则的、分布式表示的、神经网络的推理以及混合推理）均具有一定的普适性，即证明其可以通过添加知识增广器组件来增强任意现有知识图谱推理框架的性能。

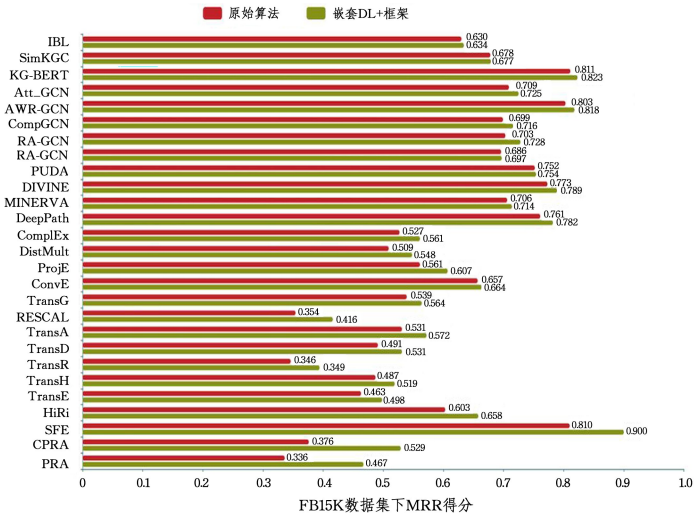


图 9 FB15K 数据集下嵌入 DL$^+$ 框架与各类算法 MRR 分数对比结果

Fig. 9 MRR scores comparison between embedded DL$^+$ framework and various algorithms on FB15K dataset

结束语 本文发现并提出知识图谱中路径连通性差问题是影响知识推理预测精度的因素之一，且连通性比率与预测结果成正比。为了缓解现有知识图谱路径连通性差的问题，本文进一步提出了一种用于增强任意现有知识推理方法性能的“即插即用”型双层推理框架 DL$^+$。该框架包括两层结构：知识增广器以及知识推理机。其中，第一层知识增广器通过社区发现算法增广社区内实体间路径来强化初始知识图谱，并以此增强第二层知识推理机中任意推理方法的性能。大量

实验结果表明，该框架能够有效缓解知识图谱中路径连通性差的问题并提高三元组预测精度。本文的贡献如下：

1)发现并明确定义知识图谱推理中存在路径连通性差的问题，并证明其与推理精度具有正相关性。

2)将知识图谱看作知识社区网络，利用社区发现算法划分知识图谱中的社区并利用剪枝优化后的实体路径来增广知识图谱。

3)提出了一种增强型双层知识图谱推理框架 DL^{+}。使用第一层知识增广器来增强第二层中任意现有知识图谱推理方法的性能，达到了有效缓解路径连通性差的问题以及提高三元组预测精度的目的。

今后的工作将着眼于发现更为细致的基于局部邻域信息以及知识社区的知识图谱推理算法来挖掘更多有效的本地知识，以缓解路径连通性差的问题并提高三元组推理的速度与精度。

参 考 文 献

[1] BOLLACKER K D, EVANS C, PARITOSG P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008.

[2] LEHMANN J. DBpedia: A Nucleus for a Web of Open Data [C]// Proceedings of the Semantic Web, International Semantic Web Conference, Asian Semantic Web Conference. Busan, Korea, 2007.

[3] MILLER G A. WordNet: a lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39-41.

[4] SUN Y W, CHENG G, LI X, et al. Complex Question Answering Method for Explainable Knowledge Graph Based on Graph Matching Network [J]. Journal of Computer Research and Development, 201, 58(12): 2673-2683.

[5] LU L, KONG F. Dialogue-oriented Entity Relation Extraction with Integrated Knowledge [J]. Computer Science, 2022, 49(5): 200-205.

[6] GE F B, SHEN X. Application research of knowledg-e graph technology in patent semantic retrieval [J]. China Inventions and Patents, 2022, 19(1): 10-18.

[7] GUAN S P, JIN X L, JIA Y T, et al. Research progress in Knowledge Reasoning based on Knowledge Graph[J]. Journal of Software, 2018, 29(10): 2966-2994.

[8] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D Knowledge Graph Embeddings[C]// Proceddings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18). New Orleans, LA, USA. 2017.

[9] ZHU G, IGLESIAS C A. Exploiting semantic similarity for named entity disambiguation in knowledge graphs[J]. Expert Systems with Application, 2018(101): 8-24.

[10] WANG X, HE X, CAO Y, et al. KGAT: Knowledge Graph Attention Network for Recommendation[C]// Procedding of the Knowledge Discovery and Data Mining. Anchorage: ACM, 2019.

[11] PUJARA J, AUGUSTINE E, GETOOR L. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017.

[12] LV X, HAN X, HOU L, et al. Dynamic Anticipation and Completion for Multi-Hop Reasoning over Sparse Knowledge Graph [C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Punta Cana, 2020.

[13] LAO N, WILLIAM W C. Relational retrieval using a combination of path-constrained random walks[J]. Machine Learning, 2010, 81: 53-67.

[14] SHI B, WENINGER T. ProjE: Embedding Projection for Knowledge Graph Completion[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, 2016.

[15] SEN P, CARVALHO B, ABDELAZIZ I, et al. Combining Rules and Embeddings via Neuro-Symbolic AI for Knowledge Base Completion[J]. arXiv: 2109.09566, 2021.

[16] ZHANG Z, ZHUANG F, ZHU H, et al. Relational Graph Neural Network with Hierarchical Attention for Knowledge Graph Completion[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020.

[17] NEWMAN M. Modularity and community structure in networks [C]// Proceedings of the National Academy of Sciences of the United States of America. 2006.

[18] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. arXiv: 0803.0476, 2008.

[19] GARDNER M, MITCHELL T. Efficient and Expressive Knowledge Base Completion Using Subgraph Feature Extraction [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

[20] XIONG W, HOANG T, WANG W Y. DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017.

[21] SCHLICHTKRULL M, KIPG T N, BLEOM P, et al. Modeling Relational Data with Graph Convolutional Networks[C]// Proceedings of the Extended Semantic Web Conference. Heraklion, 2018.

[22] YAO L, MAO C, LUO Y. KG-BERT: BERT for Knowledge GraphCompletion[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020.

[23] WANG S, ZHONG Y, WANG C. Attention Relational Graph Convolution Networks for Relation Prediction in Knowledge Graphs[C]// 2021 4th International Conference on Advanced Algorithms and Control Engineering(ICAACE 2021). 2021.

[24] WANG L, ZHAO W, WEI Z, et al. SimKGC: Simple Contrastive Knowledge Graph Completion with Pre-trained Language Models[C]// Proceedings of the Association for Computational Linguistics. 2022.

[25] CUI W Y, CHEN X R. Instance-based Learning for Knowledge Base Completion[C]// Proceedings of the Neural Information Processing Systems. 2022.

[26] TANG Z,PEI S,ZHANG Z,et al. Positive-Unlabeled Learning with Adversarial Data Augmentation for Knowledge Graph Completion[C]// Proceedings ofthe International Joint Conference on Artificial Intelligence. 2022.

[27] QUAN W,JING L,LUO Y,et al. Knowledge Base Completion via Coupled Path Ranking[C]// Proceedings of the Association for Computational Linguistics. 2016:1308-1318.

[28] QIAO L,JIANG L,HAN M,et al. Hierarchical Random Walk Inference in Knowledge Graphs[C]// Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. Pisa:ACM,2016:445-454.

[29] JIA Y,WANG Y,LIN H,et al. Locally adaptive translation for knowledge graph embedding[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2016:992-998.

[30] NICKEL M,TRESP V,KRIEGEL HP. A Three-Way Model for Collective Learning on Multi-Relational Data[C]// Proceedings of the International Conference on Machine Learning. 2011:809-816.

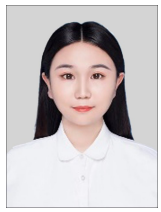
[31] HAN X,HUANG M,YU H,et al. TransG:A Generative Mixture Model for Knowledge Graph Embedding[J]. arXiv:1509.05488,2015.

[32] LI R,CHENG X. DIVINE:A Generative Adversarial Imitation Learning Framework for Knowledge Graph Reasoning[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019.

[33] TIAN A,ZHANG C,MIAO R,et al. RA-GCN:Relational Aggregation Graph Convolutional Network for Knowledge Graph Completion[C]// Proceedings of the 12th International Conference on Machine Learning and Computing. 2020:580-586.

[34] VASHISHTH S,SANYAL S,NITIN V,et al. Composition-based Multi-Relational Graph Convolutional Networks[C]// Proceedings of the International Conference on Learning Representations. 2020.

[35] WANG H,LIN H Z,LU L Y. Knowledge graph reasoning algorithm based on Att_GCN model[J]. Computer Engineering and Applications,2020,56(9):183-189.



WU Yuejia, born in 1996, Ph.D candidate, is a student member of China Computer Federation. Her main research interests include knowledge graph, knowledge graph representing and knowledge graph reasoning.

ZHOU Jiantao, born in 1974, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include cloud computing technology, social network recommendation technology, software engineering and so on.

(责任编辑:何杨)