计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(12)-2999-11
doi: 10.3778/j.issn.1673-9418.2208105

噪声知识图谱表示学习:一种规则增强的方法

邵天阳,肖卫东,赵　翔+
国防科技大学 信息系统工程重点实验室,长沙 410073
+ 通信作者 E-mail: xiangzhao@nudt.edu.cn

摘　要:知识图谱用于存储结构化事实,这些事实以三元组的形式表示,即(头实体,关系,尾实体)。当前大规模知识图谱的构建通常采用(半)自动化的方法进行知识抽取,过程中不可避免地会引入噪声,这可能会影响知识表示的效果。然而,多数传统表示学习方法假设知识图谱中的三元组都是正确的,并据此对知识进行分布式表示。因此,对知识图谱进行噪声检测是一项至关重要的工作。此外,知识图谱的不完整问题也备受人们关注。对以上问题进行了研究,提出了一种逻辑规则和关系路径信息相结合的知识表示学习框架,它在检测可能存在的噪声的同时,还能生成无噪的知识表示,实现相互辅助增强的效果。具体而言,该框架分为三元组嵌入模块和三元组可信度估计模块。在三元组嵌入模块中,在三元组结构信息的基础上引入关系路径信息和逻辑规则信息以构造更为完善的知识表示,其中后者用于增强关系路径推理的能力和表示学习的可解释性;在三元组可信度估计模块中,进一步利用三种信息对三元组进行可信度判断以检测可能存在的噪声。在三个公开评测数据集上进行了实验验证,结果表明,与所有的基线方法相比,该模型在知识图谱噪声检测和知识补全等任务上均取得了显著的性能提升。

关键词:知识图谱;知识图谱补全;噪声检测;三元组可信度;三元组嵌入
文献标志码:A　**中图分类号:**TP391

Noisy Knowledge Graph Representation Learning: a Rule-Enhanced Method

SHAO Tianyang, XIAO Weidong, ZHAO Xiang+
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China

Abstract: Knowledge graphs are used to store structured facts, which are presented in the form of triples, i.e., (head entity, relation, tail entity). Current large-scale knowledge graphs are usually constructed with (semi-) automated methods for knowledge extraction and the process inevitably introduces noise, which may affect the effectiveness of the knowledge representation. However, most traditional representation learning methods assume that the triples in knowledge graphs are correct and represent knowledge in a distributed manner accordingly. Therefore, noise detection on knowledge graphs is a crucial task. In addition, the incompleteness of knowledge graphs has also attracted people's attention. The above problems are studied and a knowledge representation learning framework combining logical rules and relation path information is proposed, which accomplishes knowledge representation learning and achieves a mutual enhancement effect while detecting possible noise. Specifically, the framework is divided into a triple embedding part and a triple trustworthiness estimation part. In the triple embedding part, relation path

基金项目:国家自然科学基金面上项目(61872446);湖南省自然科学基金杰出青年项目(2019JJ20024)。
This work was supported by the National Natural Science Foundation of China (61872446), and the Natural Science Foundation of Hunan Province (2019JJ20024).
收稿日期:2022-08-29　**修回日期:**2023-05-05

information and logical rule information are introduced to construct a better knowledge representation based on the triple structure information, the latter of which is used to enhance the ability of relation path reasoning and the interpretability of the representation learning. In the triple trustworthiness estimation part, three types of information are further utilized to detect possible noise. Experiments are conducted on three public evaluated datasets and the results show that the model achieves significant performance improvement in tasks such as knowledge graph noise detection and knowledge complementation compared with all baseline methods.

Key words: knowledge graph; knowledge graph completion; noise detection; triple trustworthiness; triple embedding

近年来,人工智能在各个领域蓬勃发展,如问题回答[1]和推荐系统[2]等,它对人们的日常生活产生了广泛的影响。在这些领域中,人们希望人工智能智能体能够具有理解、推理和解决问题的能力。而知识图谱(knowledge graph,KG)可以为这种能力的实现提供坚实的基础。知识图谱旨在描述现实世界中存在的各种事物(实体)以及它们之间的关系,它通常以三元组(头实体,关系,尾实体)的形式存储知识,记作 (h,r,t)。

尽管知识图谱在现实世界中被广泛使用,但如Yago[3]、WordNet[4]和Freebase[5]等包含了数十亿三元组的大规模知识图谱仍然受到不完整问题的困扰。具体来说,在Freebase中,300万人中有75%缺失国籍[6]。不完整问题会对某些知识图谱应用场景产生负面影响。例如,对于问题回答系统而言,不完整的知识图谱会导致错误答案。因此,知识构建和知识补全对于下游的应用场景是必要的。

对于知识构建,目前自动机制和众包发挥着越来越大的作用,但缺点是会引入噪声,一些研究工作已经发现了知识图谱中存在的噪声[7-8]。例如,在Benchmark上开放的信息抽取模型在67%的召回率下只达到了24%的准确率[8]。对于知识补全,目前主流方法之一是知识表示学习[9-17],即将实体和关系投射到一个连续的低维空间,以获得其表示(特征)。然而这些方法大都假设知识图谱中没有噪声,这显然不符合事实。忽略知识图谱中的噪声得到的知识表示将包含不正确的信息,这会对下游的应用产生不利影响,因此考虑噪声的存在是必要的。

最近,Xie等人[12]提出了一个名为CKRL(confidence-aware knowledge representation learning)的模型,该模型利用三元组置信度来进行噪声检测,同时构建知识表示。为了判断一个三元组是否可信,其借鉴PTransE[13]模型并根据结构信息和关系路径信息获得一个置信度分数。然而,CKRL中的三元组置信度估计模块忽略了辅助信息,这些辅助信息会使得获得的知识表示更为全面。Xie等人提到,在噪声检测的实验中PTransE[13]的效果远不如TransE[14],实验结果也证明了这一点。经过文献[15]和研究分析发现,因为路径表示完全是基于嵌入空间的数值计算来实现的,这导致了误差传播进而使得路径嵌入的准确性受限,最后影响了整个表示的学习,而这个问题在噪声知识图谱上会变得更加严重。因此,尽管利用路径信息来扩展三元组的结构信息是可行的,但噪声三元组的存在使得通过关系路径进行推理的误差增大且缺乏可解释性。

为了解决上述问题,本研究提出了一个逻辑规则和关系路径信息相结合的知识表示学习框架RPKRL(logic rules and relation path information knowledge representation learning framework),以检测知识图谱中的噪声并构造无噪的知识表示。该模型考虑引入逻辑规则来提高关系路径推理的精度和可解释性,同时利用三元组可信度对三元组质量进行判断。图1显示了RPKRL模型框架的简要说明,在进行知识抽取和自动知识构建之后,知识图谱中包含噪声且存在不完整的问题。该模型可以在检测图谱中存在的噪声的同时生成无噪知识表示以进行知识补全。

具体来说,RPKRL可分为两部分:三元组嵌入模块和三元组可信度估计模块。在三元组嵌入模块中,引入逻辑规则来指导路径的构成,从而提高其精确性和可解释性,该模块相比PTransE[13]而言构造了更为完善的知识表示。在三元组可信度估计模块中,进一步利用关系路径信息和逻辑规则信息得到三元组可信度从而对三元组可信度进行判断。通过结合这两部分,该模型能够检测到知识图谱中可能存在的噪声,并构建无噪的知识表示。在三个数据集上评估了模型,结果显示与基线相比,该模型具有较好的有效性和稳健性。

这项工作的主要贡献可总结如下:

(1)针对路径推理在噪声知识图谱中存在的问题,提出了一个新颖的RPKRL框架,用于同时进行知

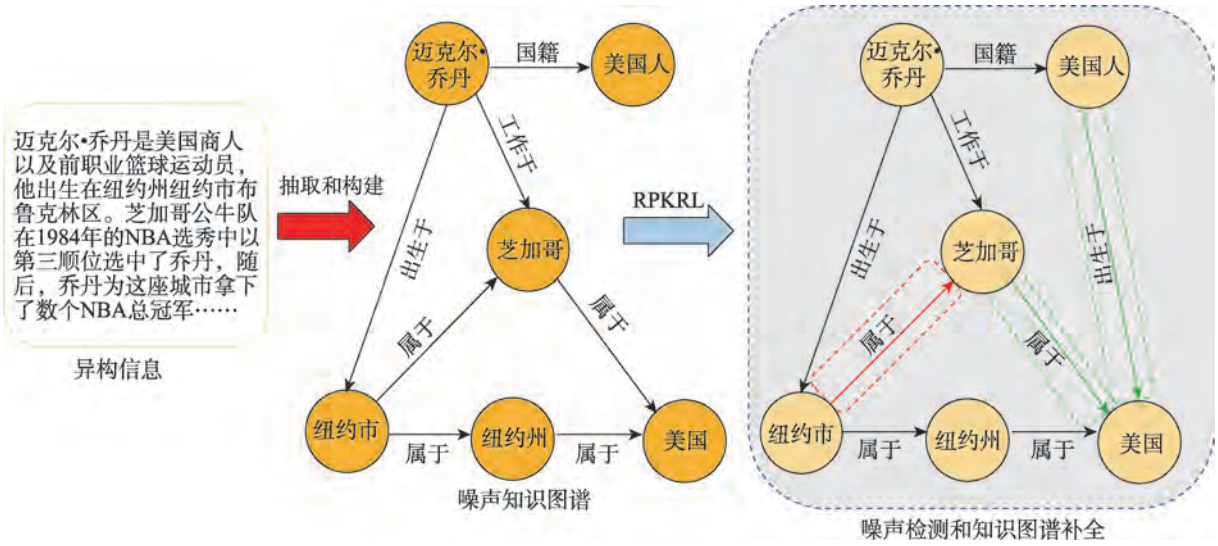

图1 RPKRL模型框架的简要工作流程描述

Fig.1 Brief description of RPKRL model framework workflow

识图谱噪声检测和知识表示学习，该框架大幅度提高了使用路径信息进行噪声检测和知识图谱补全的效果。

(2)引入了逻辑规则，以便能够在噪声检测中区分噪声。由于路径推理会导致误差的传播，而这个问题在有噪声的知识图谱上会更加严重。因此，试图通过逻辑规则的准确性来解决这个问题。

(3)逻辑规则可以增强关系路径的可解释性。关系路径推理得到的关系通常通过关系的表示之间的运算，例如相加和相乘等，缺乏可解释性，逻辑规则具有的可解释性很好地补足了这一缺陷。

1 相关工作

1.1 知识图谱噪声检测

尽管近年来知识图谱在许多领域得到了广泛的应用，但噪音问题的存在对知识的获取产生了负面的影响[16]。最近，一项名为“针对知识库中的破坏性检测”的任务引起了广泛的关注，它的目的在于解决故意破坏知识图谱的问题[17]。人们逐渐意识到噪声检测对于知识获取和知识应用的重要性越来越高。大多数知识图谱的噪声检测工作是在知识图谱构建时完成的[7,18]。例如，YAGO2[19]是人们在人工监督下从维基百科中提取知识所形成的数据集，因此可以评估这些知识的正确性。Wikidata也是通过众包的人力管理软件提取的数据集，软件使用者可以审核数据以删除错误的信息[20]。小型知识图谱上或许可以进行人工噪音检测，但在大规模的知识图谱上，这将是耗时耗力的。

近年来，研究人员开始关注知识图谱噪声的自动检测[21-22]。Dong等人[23]利用知识图谱的先验知识构建了一个概率知识库，并将其与网络内容相结合，以共同判断三元组的质量。然而，这种方法是为某个知识图谱构建量身定做的，并不具备泛化能力。Li等人[24]使用神经网络方法为不可见的三元组提供置信度分数以进行知识库补全，但这种方法忽略了知识库中的其他信息。Xie等人[12]介绍了进行噪声检测和构建知识表示的三元组置信度框架，它结合了三元组结构信息和关系路径信息来判断三元组质量。然而，这种方法忽略了其他有用的信息，而且利用路径进行推理也存在可解释性的问题。

相比之下，RPKRL模型在三元组结构信息的基础上引入逻辑规则信息来增强关系路径的推理表达能力和模型的可解释性，进而提高模型的噪声检测能力。

1.2 知识表示学习模型

近年来，知识表示学习受到越来越多的关注，许多研究人员在知识表示学习方面做了大量的工作[25-26]，主要可以分为三种类型：(1)基于平移的模型，这类模型源自词嵌入的平移不变原理[27]，TransE[14]是最具代表性的基于平移的模型，它将实体和关系投影到同一空间，并将关系视为头实体和尾实体之间的平移，后续基于TransE模型，又衍生出了许多扩展模型。(2)张量分解模型，RESCAL[28]利用张量分解，将关系表示为矩阵，将实体表示为向量。在此基础上，

DisMult[29]将关系矩阵简化为对角矩阵，ComplEx[30]引入了复数以扩展DisMult，以便更好地对非对称关系进行建模。此时，实体和关系都在复数空间。(3)神经网络模型，NTN(neural tensor network)[31]首先将实体的向量作为神经网络的输入，然后将这两个实体由关系特有的关系张量(以及其他参数)组合，并映射到一个非线性隐藏层，最后一个特定于关系的线性输出层给出了三元组的评分。此外，还有ConvE[32]和ConvKB[33]等神经网络模型。在这三类模型中，基于平移的模型既简单又有效，同时还能够达到最好的性能。这类模型将实体和关系都投影到一个连续的低维向量空间中，并根据基于距离的评分函数进行建模，从而获得知识表示。与其他方法相比，TransE能够实现简单性和有效性的平衡。然而，由于其结构简单，在处理1-N、N-1和N-N这样的复杂关系时，它的效果并不理想。对于此，人们提出了许多改进的知识表示方法[34-35]。例如，DualE[36]在对偶四元数空间建模，Nayyeri等人[15]引入了复平面上的莫比乌斯变换。

平移假设只集中在三元组上，这可能会忽略其他有效信息。PTransE[13]提出实体对之间的路径嵌入可以通过多步骤的关系推理得到。AutoETER[37]提出将关系看作实体类型之间的转换操作，进而学习实体的表示。此外，还有许多其他类型的信息可以利用，如视觉信息、属性信息、逻辑规则等。

大多数传统方法都假设知识图谱中的所有三元组都是完全正确的，因此，它们无法检测到知识图谱中可能存在的噪声。与它们不同，RPKRL引入了三元组可信度的概念来区分含有噪声的三元组和正例三元组。

2 方法

本章将详细介绍模型RPKRL，由三元组嵌入模块和三元组可信度估计模块组成。首先给出文中使用的符号：给定一个正例三元组(h,r,t)，考虑头部和尾部实体h，$t\in E$和$r\in R$，其中E和R是实体和关系的集合。T表示包含噪声三元组的所有训练三元组。下面详细介绍整体模型结构及其组成部分结构。

2.1 背景知识

基于平移的模型有很多，其中，TransE[14]是最基础的也是最具代表性的基于平移的模型之一。它将知识图谱中的实体和关系投影到同一个低维连续向量空间中。具体而言，对于一个正例三元组(h,r,t)，TransE[14]认为其实体向量和关系向量应满足$\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$，因此，TransE[14]的模型框架如下：

$$E(h,r,t)=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_2 \tag{1}$$

其中，$\boldsymbol{h}$、$\boldsymbol{r}$和$\boldsymbol{t}$分别代表头实体、关系和尾实体的向量。若三元组(h,r,t)为正例三元组时，则分数$E(h,r,t)$较低，若三元组(h,r,t)为负例三元组时，则分数$E(h,r,t)$较高。

2.2 模型框架

RPKRL模型可以在检测知识图谱中噪声的同时构建无噪的知识表示。首先给出模型公式如下：

$$E(T)=\sum_{(h,r,t)\in T} RP(h,r,t)\cdot LTT(h,r,t) \tag{2}$$

其中，$RP(h,r,t)$是三元组嵌入函数，而$LTT(h,r,t)$是三元组可信度函数。它们利用结构信息作为主体。此外，添加了关系路径信息和逻辑规则信息。较低的$RP(h,r,t)$分数表示实体和关系在三元组更适合嵌入框架。与传统的嵌入式模型不同，该模型考虑了知识图谱中的噪声，针对于此引入了三元组可信度衡量。一个更高的三元组可信度得分意味着三元组更可靠，即越有可能是正例。将在下面的两部分介绍三元组嵌入模块和三元组可信度估计模块。

2.3 三元组嵌入模块

传统的路径推理方法利用的路径表示是由基于嵌入空间的数值计算得到，这会导致误差的传播，从而影响整个表示学习。此外，这些方法在路径表示的获取过程中缺乏可解释性。受RPJE(rule and path-based joint embedding)[38]模型的启发，引入逻辑规则及其置信度$\mu\in[0,1]$(Horn规则)，并将其与路径相结合，以提高路径推理的精度和可解释性(任何知识图谱规则提取算法或工具都可以自动挖掘Horn规则)。

这些规则可以分为长度为1和长度为2的两种类型，分别命名为R1和R2。图2显示了规则指导路径中关系的合成进行推理的过程。规则R1通过规则主体和规则头部将两个关系联系起来，规则R2则可以用来指导路径中关系的合成。对于规则R1来说，当$\forall x,y:r_2(x,y)\Leftarrow r_1(x,y)$成立时，关系R1和关系R2在训练过程中具有较高相似性。对于规则R2，必须使规则主体的组成部分形成顺序路径，从而可以组成关系路径。因此，如表1所示，共总结了8种不同类型的规则转换模式，然后对它们进行编码以与路径组合。在进行路径中关系的合成时，尝试用规则指导合成，直到不能合成为任何关系为止。特别的，将由规则指导关系的合成称为$R(p)$，这也是路径p的

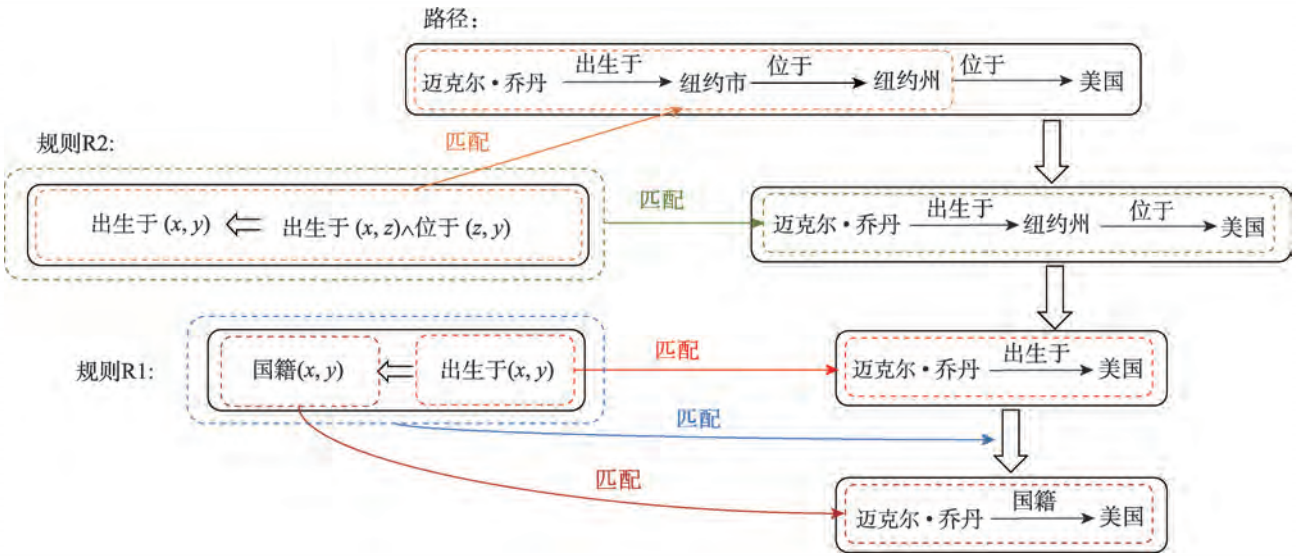

图2 规则指导路径中的关系的合成示例

Fig.2 Example of relations composition in rule-guided path

表1 规则R2的转换模式列表

Table 1 List of rules R2 conversion mode

原规则	编码规则
$r_3(a,b)\Leftarrow r_1(a,e)\wedge r_2(e,b)$	$r_3\Leftarrow(r_1,r_2)$
$r_3(a,b)\Leftarrow r_1(e,b)\wedge r_2(a,e)$	$r_3\Leftarrow(r_2,r_1)$
$r_3(a,b)\Leftarrow r_1(e,b)\wedge r_2(e,a)$	$r_3\Leftarrow(r_2^{-1},r_1)$
$r_3(a,b)\Leftarrow r_1(e,a)\wedge r_2(e,b)$	$r_3\Leftarrow(r_1^{-1},r_2)$
$r_3(a,b)\Leftarrow r_1(a,e)\wedge r_2(b,e)$	$r_3\Leftarrow(r_1,r_2^{-1})$
$r_3(a,b)\Leftarrow r_1(b,e)\wedge r_2(a,e)$	$r_3\Leftarrow(r_2,r_1^{-1})$
$r_3(a,b)\Leftarrow r_1(e,a)\wedge r_2(b,e)$	$r_3\Leftarrow(r_1^{-1},r_2^{-1})$
$r_3(a,b)\Leftarrow r_1(b,e)\wedge r_2(e,a)$	$r_3\Leftarrow(r_2^{-1},r_1^{-1})$

嵌入表示。利用规则R2对路径进行建模，其计算公式如下：

$$E_2(p,r)=R(p|h,t)\|\boldsymbol{R}(\boldsymbol{p})-\boldsymbol{r}\| \tag{3}$$

其中，$R(p|h,t)$是给定实体对(h,t)间关系路径p的可靠度，该可靠度可以由路径约束资源分配机制（path-constraint resource allocation，PCRA）[13]计算得到，$\mu(p)=\{\mu_1,\mu_2,\cdots,\mu_n\}$是规则R2的置信度的集合。

对于逻辑规则的可解释性，表2展示了一些例子。表中前面部分为规则，后面部分为规则置信度。原本的关系路径推理中，关系的合成通过关系向量间的计算，如加、减、乘和除得到，关系的推理则通过关系向量间的相似度计算等方法得到，由于是数值间的计算，可解释性较差，而规则的引入则补足了这一点。由规则来指导路径中关系的合成及关系推理，不仅增加了其正确性，也提高了其可解释性。

最后，设计了一种新的结合关系路径信息和逻辑规则信息的三元组嵌入模型。模型公式如下：

$$RP(h,r,t)=E_1(h,r,t)+E_2(p,r) \tag{4}$$

其中，$E_1(h,r,t)=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|$是TransE模型的评分函数。这里使用TransE模型的评分函数作为主嵌入函数，使得可以将其替换为其他优化后的翻译模型或者引入辅助信息的翻译模型。

2.4 三元组可信度模块

受CKRL[12]和DSKRL（dissimilarity-support-aware

表2 规则R1和R2的例子

Table 2 Examples of rules R1 and R2

规则	例子
R1	/location/hud_county_place/county(a,b) <= /location/us_county/hud_county_place^{-1}(a,b) 1
	/sports/sports_team/roster./sports/sports_team_roster/player(a,b) <=/soccer/football_team/current_roster./soccer/football_roster_position/player(a,b) 0.87
R2	/film/film/estimated_budget./measurement_unit/dated_money_value/currency(a,b) <=/film/film/release_date_s./film/film_regional_release_date/film_release_region(a,b)/\/location/statistical_region/gni_in_ppp_dollars./measurement_unit/dated_money_value/currency(a,b) 0.95

knowledge representation learning)[39]模型的启发,在三元组可信度模块中,对三元组的质量进行判断,计算三元组质量的公式如下:

$$Q_1(h,r,t)=-(\gamma+E_1(h,r,t)-E_1(h'+r'-t')) \tag{5}$$

在训练开始时,将所有三元组的局部三元组可信度 $LTT(h,r,t)$ 初始化为1。在训练过程中,数值会发生变化。形式上,局部三元组可信度 $LTT(h,r,t)$ 随其三重质量 $Q(h,r,t)$ 变化如下:

$$LTT(h,r,t)=\eta LTT(h,r,t),\ Q_1(h,r,t)\leqslant 0 \tag{6}$$

其中,η 是确保 $LTT(h,r,t)>0$ 和 $LTT(h,r,t)<1$ 的超参数。$LTT(h,r,t)$ 的值将以线性速率减小,因为当 $Q(h,r,t)\leqslant 0$ 时,这个三元组更可能包含噪声,所以应该具有较低的三元组可信度。

此外,引入逻辑规则以加强对三元组质量的判断效果。具体的,利用规则R1找到关系 r 的相似关系 r_R,然后将三元组 (h,r,t) 替换为 (h,r_R,t),进行质量计算:

$$Q_2(h,r,t)=-(\gamma+\mu E_1(h,r_R,t)-E_1(h'+r'-t')) \tag{7}$$

其中,μ 是规则R1的置信度。

通过进一步计算三元组 (h,r,t) 的质量后,三元组可信度 $LTT(h,r,t)$ 也将随之变化:

$$LTT(h,r,t)=LTT(h,r,t)-\alpha,\ Q_2(h,r,t)\leqslant 0 \tag{8}$$

其中,α 是确保 $LTT(h,r,t)>0$ 和 $LTT(h,r,t)<1$ 的超参数。

2.5 损失函数及优化

根据TransE[14]可以将RPKRL的损失函数形式化为一组成对得分函数的和,该损失函数会使得正例三元组的得分低于负例三元组,损失函数公式如下:

$$L=\sum_{(h,r,t)\in T}\sum_{(h',r',t')\in T'}(L_1(h,r,t)+\beta\sum_{p\in P(h,t)}L_2(p,r))+\lambda LTT(h,r,t) \tag{9}$$

其中,λ 是超参数,T' 表示负例三元组的集合,$L_1(h,r,t)$、$L_2(p,r)$ 是关于三元组 (h,r,t) 和路径对 (p,r) 的损失函数:

$$L_1(h,r,t)=\max(0,\gamma_1+E_1(h,r,t)-E_1(h',r',t')) \tag{10}$$

$$L_2(p,r)=\max(0,\gamma_2+E_2(p,r)-E_2(p,r')) \tag{11}$$

其中,γ_1 和 γ_2 是超参数。

在训练过程中,由于知识图谱中没有显式的负例三元组,将训练三元组中的实体或关系进行随机替换,且替换后得到的负例三元组不在训练三元组集合中,负三元组采样规则如下:

$$T'=(h',r,t)\cup(h,r',t)\cup(h,r,t') \tag{12}$$

对于优化,使用小批量随机梯度下降(stochastic gradient descent, SGD)来最小化损失函数。

2.6 复杂度分析

首先给出所使用的符号。N_T 是训练三元组的数量,N_P 是关系路径的数量,N_L 是关系路径的长度,N_r 是规则的数量,K 是实体和关系向量的维度。参考PTransE[13]给出的复杂度分析,在每个迭代循环中,TransE的复杂度为 $O(N_TK)$,PTransE的复杂度为 $O(N_TKN_PN_L)$。RPKRL模型使用了规则信息和关系路径信息,复杂度为 $O(N_T^2KN_rN_L)$。

3 实验

为验证模型及其各部分的有效性,在公开数据集上进行了充分评测。

3.1 数据集

实验验证在FB15K数据集上进行,FB15K数据集是一个典型的基准知识图谱,它是从现实世界中广泛使用的大规模知识图谱Freebase中提取出来的。在FB15K数据集中,有14 951个实体和1 345个关系,以及对应的592 213个三元组。其中训练集含有483 142个三元组,验证集含有50 000个三元组,测试集含有59 071个三元组。大多数现实世界的知识图都包含噪声,但FB15K中没有明显标记的噪声,为此,使用了CKRL[14]的3个公开可用的数据集。3个数据集分别命名为FB15K-N1、FB15K-N2和FB15K-N3。它们之间的不同之处在于含有不同的噪声率,分别为10%、20%和30%。

事实上,现实世界知识图谱中的许多噪音都源于同类实体之间的误解[14]。它表明,在现实世界的知识图谱中,噪声(姚明,出生地,加拿大)比(姚明,出生地,足球)更有可能发生。具体来说,给定知识图谱中的一个正例三元组 (h,r,t),随机地将相同类型的头或尾实体与后者替换以形成负例三元组 (h',r,t) 或 (h,r,t')。例如,正例三元组(姚明,出生地,中国)将被负例三元组(姚明,出生地,澳大利亚)或(姚明,出生地,英国)所替换。3个含有噪声的数据集与FB15K共享相同的实体、关系、验证集和测试集。具体的数据如表3所示。

3.2 实验设置

选择TransE[14]、PTransE[13]、TransH[33]、TransR[34]、CKRL[12]和RPJE[38]作为不同实验比较的基线。使用小批量SGD训练RPKRL模型。边际 γ_1 和 γ_2 均被设置为1。将学习率 δ 设置为动态,并在开始时从{0.001,

表3 噪声数据集统计

Table 3 Statistics of noise datasets

数据集	负例三元组	训练三元组	验证三元组	测试三元组
FB15K-N1	46 408	529 550	50 000	59 071
FB15K-N2	93 782	576 924	50 000	59 071
FB15K-N3	187 925	671 067	50 000	59 071

0.002,0.003,0.004} 中选择，最后在 {0.000 1, 0.000 2} 中选择。对于三元组可信度，下降控制速率 η 和 α 分别设置在 {0.80,0.85,0.90} 和 {0.10,0.01} 之间。该模型的最优配置是：δ 以0.001开始，以0.000 1结尾，$\eta=0.9$，$\alpha=0.01$，在验证集上进行了优化。为了进行公平比较，所有模型中实体和关系嵌入的维度均设置为50。

3.3 知识图谱噪声检测

为了验证RPKRL模型在检测知识图谱中存在的噪声的性能，进行了知识图谱噪声检测任务。该任务旨在基于三元组得分来检测知识图谱中可能存在的噪声。

3.3.1 评测准则

使用TransE的能量函数作为RPKRL模型和基线模型的评分函数，然后根据评分对训练集中所有的三元组进行排序。如果一个三元组得分较高，那么它更有可能是一个噪声三元组。根据排名计算并绘制准确率和召回率曲线，以显示RPKRL模型和基线模型的噪声检测能力。

3.3.2 实验结果

图3~图5分别展示了模型在3个数据集上的噪声检测性能结果，从中可以观察到：(1)本研究模型RPKRL在不同噪声率(10%、20%、40%)的所有3个数据集上都获得了最好的性能。这有力地证明了其

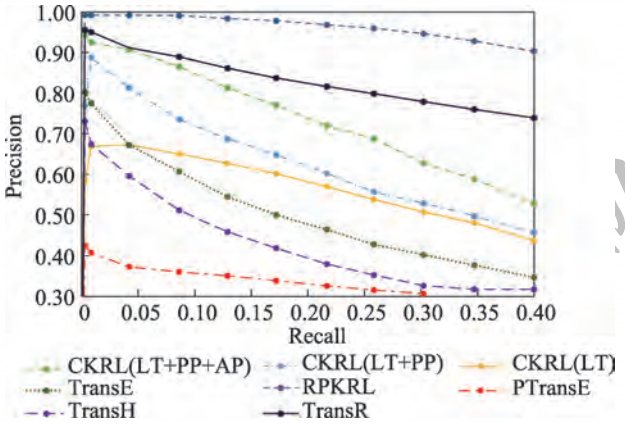

图3 FB15K-N1数据集上噪声检测结果

Fig.3 Noise detection results on FB15K-N1 dataset

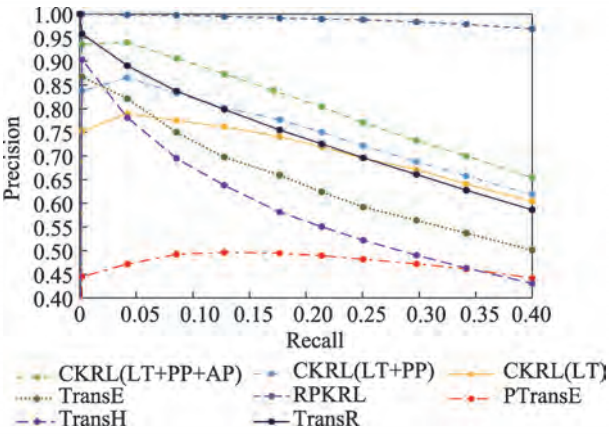

图4 FB15K-N2数据集上噪声检测结果

Fig.4 Noise detection results on FB15K-N2 dataset

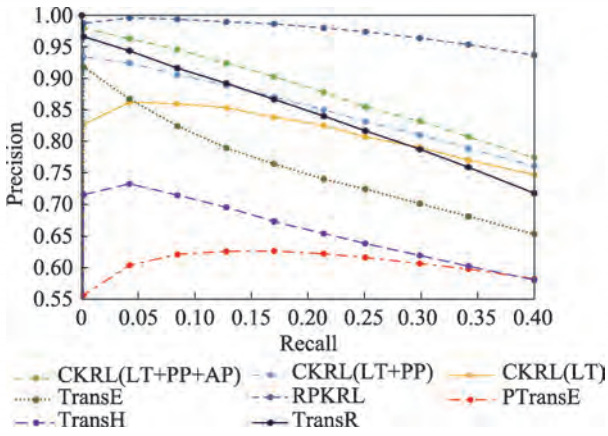

图5 FB15K-N3数据集上噪声检测结果

Fig.5 Noise detection results on FB15K-N3 dataset

检测知识图谱中的噪声的能力。(2)单纯的路径推理PTransE在噪声检测任务上表现非常差，RPKRL模型针对于此做出了改进，通过引入逻辑规则信息来指导关系路径中关系的合成，实验证明改进是有效的且实验效果提升较大。

3.4 知识图谱补全

知识图谱补全注重于知识表示学习的质量，其目标是在 h、r 和 t 中缺失任意一个的情况下补全三元组。

3.4.1 评测准则

本文主要关注实体预测。遵循TransE[14]中相同的设置，进行了两个典型的度量：(1)正确答案的平均排名；(2)Hits@10表示正确答案排在前10位的实体。此外，遵循TransE[14]中使用的不同的评估设置“Raw”和“Filter”。

3.4.2 实验结果

表4和表5展示了模型在3个数据集上的实体预

表4 实体Mean Rank预测结果

Table 4 Results of entity prediction on Mean Rank

模型	FB15K-N1		FB15K-N2		FB15K-N3	
	Raw	Filter	Raw	Filter	Raw	Filter
TransE	240	144	250	155	265	171
PTransE	225	95	231	111	247	123
TransH	257	163	263	170	277	185
TransR	261	167	280	187	311	218
CKRL(LT)	237	140	243	146	244	148
CKRL(LT+PP)	236	139	241	144	245	149
CKRL(LT+PP+AP)	236	138	240	144	245	150
RPKRL	**192**	**77**	**192**	**80**	**195**	**86**

表5 实体Hits@10预测结果

Table 5 Results of entity prediction on Hits@10 **单位** %

模型	FB15K-N1		FB15K-N2		FB15K-N3	
	Raw	Filter	Raw	Filter	Raw	Filter
TransE	44.9	59.8	42.8	56.3	40.2	51.8
PTransE	39.6	54.4	37.8	51.1	35.3	46.4
TransH	44.7	58.7	42.4	54.9	39.5	50.5
TransR	43.8	57.5	41.5	53.5	38.9	49.4
CKRL(LT)	45.5	61.8	44.3	59.5	42.7	56.9
CKRL(LT+PP)	45.3	61.6	44.2	59.4	42.8	56.8
CKRL(LT+PP+AP)	45.3	61.6	44.2	59.3	42.8	56.6
RPKRL	**45.6**	**63.4**	**45.4**	**62.9**	**45.0**	**61.7**

测结果,可以发现:在所有3个噪声数据集上,RPKRL模型在所有评估指标上都优于所有的基线模型,尤其是平均排名(Mean Rank)的提升幅度很大。与CKRL(LT+PP+AP)相比,RPKRL平均提高55。这证实了RPKRL模型所获得的知识表示的质量,因为它不仅可以检测知识图中的噪声,在知识图谱补全方面也具有更好的性能。

3.5 消融实验

为了衡量模型各个组件的影响,比较了当模型处于不同子模块设置时两个任务的性能。RPKRL(RP)表示只考虑三元组嵌入而不考虑三元组可信度的策略。RPKRL(E_l)表示在三元组嵌入模块中只利用三元组本身结构信息的策略。评测准则的执行方式与以前相同。

3.5.1 知识图谱噪声检测结果

图6~图8分别展示了模型在3个数据集上的噪声检测性能结果,从中可以观察到:(1)RPKRL在3个数据集上都取得了不错的结果,这证实了模型中各个子模块的有效性。(2)RPKRL与RPKRL(E_l)的效果差异随着数据集噪声率的增加,先增加后减少,

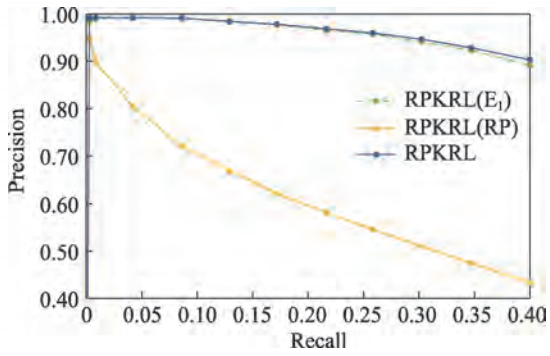

图6 消融实验:FB15K-N1数据集上噪声检测结果

Fig.6 Ablation study: noise detection results on FB15K-N1 dataset

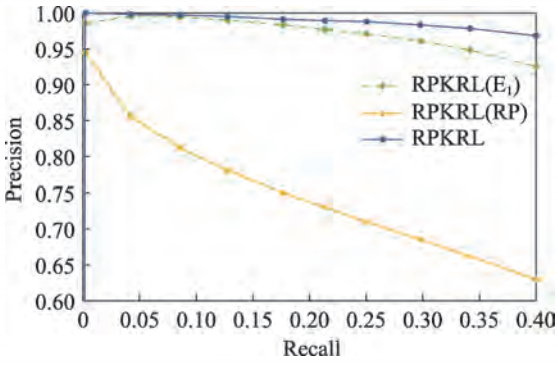

图7 消融实验:FB15K-N2数据集上噪声检测结果

Fig.7 Ablation study: noise detection results on FB15K-N2 dataset

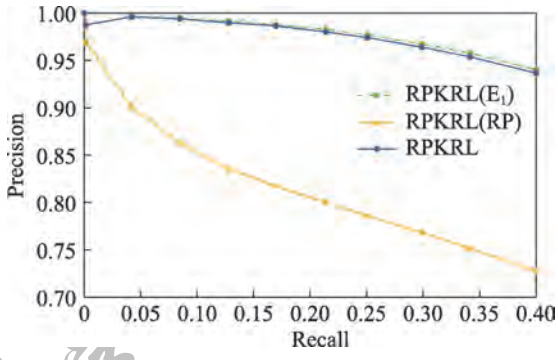

图8 消融实验:FB15K-N3数据集上噪声检测结果

Fig.8 Ablation study: noise detection results on FB15K-N3 dataset

这意味着模型需要随着噪声率的变化而进行调整。(3)RPKRL和RPKRL(E_l)比RPKRL(RP)具有更好的性能,这在实际的噪声检测系统中更为重要,这意味着虽然仅仅靠三元组嵌入模块已经可以进行噪声检测,但三元组可信度模型的引入将大大提升这一效果。

3.5.2 知识图谱补全结果

表6和表7展示了模型在3个数据集上的实体预

表6 消融实验-Mean Rank

Table 6 Ablation study-Mean Rank

模型	FB15K-N1		FB15K-N2		FB15K-N3	
	Raw	Filter	Raw	Filter	Raw	Filter
RPKRL(RP)	196	87	206	103	226	126
RPKRL(E_1)	**191**	**74**	**192**	**77**	**193**	**83**
RPKRL	192	77	**192**	80	195	86

表7 消融实验-Hits@10

Table 7 Ablation study-Hits@10 **单位:%**

模型	FB15K-N1		FB15K-N2		FB15K-N3	
	Raw	Filter	Raw	Filter	Raw	Filter
RPKRL(RP)	45.1	61.0	43.5	58.1	41.2	54.3
RPKRL(E_1)	44.9	62.6	44.7	62.1	43.6	61.0
RPKRL	**45.6**	**63.4**	**45.4**	**62.9**	**45.0**	**61.7**

测结果,从中可以观察到:(1)在所有3个数据集上,RPKRL都获得了最佳的Hits@10,这意味着模型的每个子模块都是有效的。(2)从表中看出,三元组可信度模块的加入对于模型效果的提升是巨大的,这说明在进行补全预测时,多重判断的设置极大地提升了路径推理的准确性。

3.6 案例分析

本节给出一个具体的案例以显示RPKRL模型在噪声检测方面的优越性。遵循3.3.1小节评测准则,在10%噪声率的数据集(噪声三元组共46 408个,正例三元组共483 142个,共529 550个三元组)中选取一个噪声三元组(The Motorcycle Diaries(film),/film/film/release_date_s./film/film_regional_release_date/film_release_region,Italy)。其中,The Motorcycle Diaries(film)是一部电影的名字,Italy为一个国家的名字,该电影是在美国上映的,而不是意大利,因此这是一个噪声三元组。

采用TransE的能量函数 $E(h,r,t)=|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}|$ 对该三元组进行判断,RPKRL模型得分为5.738 02,在噪声检测排名中为38 607名;PTransE模型得分为4.993 4,在噪声检测中排名为249 547;CKRL模型得分为4.514 21,在噪声检测中排名为327 618。可以看出3个模型中只有RPKRL将其判断为噪声三元组,而后两个模型将其判断为正例三元组,且排名较为靠后,即后两个模型认为该三元组是正例三元组的可能性很大。

4 结束语

本文提出了一种新的RPKRL模型,旨在检测知识图谱中的噪声,同时学习无噪声的知识表示。该模型利用三元组的结构信息和辅助信息(关系路径信息和逻辑规则信息)来估计三元组的可信度得分。针对知识图谱中的知识补全任务和噪声检测任务,对模型进行了评估实验。在三个公开数据集上的实验结果表明,RPKRL能够很好地利用结构信息和辅助信息来度量三元组可信度,这对噪声检测和表示学习具有重要意义。三元组可信度的利用对于真实世界中知识的构建和噪声检测也是有用的。

未来将探索以下研究方向:(1)增加更多的外部支持信息,以获得更好的实体和关系的嵌入,这对知识驱动的任务有积极的影响;(2)将可信度应用于知识构建中的噪声检测,以从根源降低噪声。

参考文献:

[1] MARION P, NOWAK P, PICCINNO F. Structured context and high-coverage grammar for conversational question answering over knowledge graphs[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Nov 7-11, 2021. Stroudsburg: ACL, 2021: 8813-8829.

[2] 歹杰, 李青山, 褚华, 等. 突破智慧教育: 基于图学习的课程推荐系统[J]. 软件学报, 2022, 33(10): 3656-3672.
DAI J, LI Q S, CHU H, et al. Breakthrough in smart education: course recommendation system based on graph learning[J]. Journal of Software, 2022, 33(10): 3656-3672.

[3] SUCHANEK F M, KASNECI G, WEIKUM G. YAGO: a core of semantic knowledge[C]//Proceedings of the 16th International Conference on World Wide Web, Banff, May 8-12, 2007. New York: ACM, 2007: 697-706.

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

[5] BOLLACKER K D, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, Jun 10-12, 2008. New York: ACM, 2008: 1247-1250.

[6] WEST R, GABRILOVICH E, MURPHY K, et al. Knowledge base completion via search-based question answering[C]//Proceedings of the 23rd International World Wide Web Conference, Seoul, Apr 7-11, 2014. New York: ACM, 2014: 515-526.

[7] HEINDORF S, POTTHAST M, STEIN B, et al. Vandalism detection in wikidata[C]//Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Indianapolis, Oct 24-28, 2016. New York:

ACM, 2016: 327-336.

[8] STANOVSKY G, MICHAEL J, ZETTLEMOYER L, et al. Supervised open information extraction[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Jun 1-6, 2018. Menlo Park: AAAI, 2018: 885-895.

[9] 彭敏, 黄婷, 田纲, 等. 聚合邻域信息的联合知识表示模型[J]. 中文信息学报, 2021, 35(5): 46-54.
PENG M, HUANG T, TIAN G, et al. Neighborhood aggregation for knowledge graph representation[J]. Journal of Chinese Information Processing, 2021, 35(5): 46-54.

[10] YANG H, LIU J F. Knowledge graph representation learning as groupoid: unifying TransE, RotatE, QuatE, ComplEx[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Queensland, Nov 1-5, 2021. New York: ACM, 2021: 2311-2320.

[11] ALLEN C, BALAZEVIC I, HOSPEDALES T. Interpreting knowledge graph relation representation from word embeddings[C]//Proceedings of the 9th International Conference on Learning Representations, Austria, May 3-7, 2021:1-16.

[12] XIE R B, LIU Z Y, LIN F, et al. Does William Shakespeare really write Hamlet? Knowledge representation learning with confidence[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, Feb 2-7, 2018. Menlo Park: AAAI, 2018: 4954-4961.

[13] LIN Y K, LIU Z Y, LUAN H B, et al. Modeling relation paths for representation learning of knowledge bases[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Sep 17-21, 2015. Stroudsburg: ACL, 2015: 705-714.

[14] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, Dec 5-8, 2013. Red Hook: Curran Associates, 2013: 2787-2795.

[15] NAYYERI M, VAHDATI S, AYKUL C, et al. 5* knowledge graph embeddings with projective transformations[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, Feb 2-9, 2021. Menlo Park: AAAI, 2021: 9064-9072.

[16] MANAGO M, KODRATOFF Y. Noise and knowledge acquisition[C]//Proceedings of the 10th International Joint Conference on Artificial Intelligence, Milan, Aug 23-28, 1987: 348-354.

[17] HEINDORF S, POTTHAST M, STEIN B, et al. Towards vandalism detection in knowledge bases: corpus construction and analysis[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Aug 9-13, 2015. New York: ACM, 2015: 831-834.

[18] MELO A, PAULHEIM H. Detection of relation assertion errors in knowledge graphs[C]//Proceedings of the 2017 Knowledge Capture Conference, Austin, Dec 4-6, 2017. New York: ACM, 2017: 22.

[19] HOFFART J, SUCHANEK F M, BERBERICH K, et al. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence, 2013, 194: 28-61.

[20] TANON T P, VRANDECIC D, SCHAFFERT S, et al. From freebase to Wikidata: the great migration[C]//Proceedings of the 25th International Conference on World Wide Web, Montreal, Apr 11-15, 2016. New York: ACM, 2016: 1419-1428.

[21] JIA S B, XIANG Y, CHEN X J, et al. Triple trustworthiness measurement for knowledge graph[C]//Proceedings of the 2019 World Wide Web Conference, San Francisco, May 13-17, 2019. New York: ACM, 2019: 2865-2871.

[22] HONG Y, BU C Y, WU X D. High-quality noise detection for knowledge graph embedding with rule-based triple confidence[C]//LNCS 13031: Proceedings of the 18th Pacific Rim International Conference on Artificial Intelligence, Hanoi, Nov 8-12, 2021. Cham: Springer, 2021: 572-585.

[23] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, Aug 24-27, 2014. New York: ACM, 2014: 601-610.

[24] LI X, TAHERI A, TU L F, et al. Commonsense knowledge base completion[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Aug 7-12, 2016. Stroudsburg: ACL, 2016: 1445-1455.

[25] 宁原隆, 周刚, 卢记仓, 等. 一种融合关系路径与实体描述信息的知识图谱表示学习方法[J]. 计算机研究与发展, 2022, 59(9): 1966-1979.
NING Y L, ZHOU G, LU J C, et al. A representation learning method of knowledge graph integrating relation path and entity description information[J]. Journal of Computer Research and Development, 2022, 59(9): 1966-1979.

[26] ZHANG Y Q, YAO Q M, DAI W Y, et al. AutoSF: searching scoring functions for knowledge graph embedding[C]//Proceedings of the 36th IEEE International Conference on Data Engineering, Dallas, Apr 20-24, 2020. Piscataway:

IEEE, 2020: 433-444.
[27] MIKOLOV T, YIH S W, ZWEIG G. Linguistic regularities in continuous space word representations[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Jun 9-14, 2013. Menlo Park: AAAI, 2013: 746-751.
[28] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data[C]//Proceedings of the 28th International Conference on Machine Learning, Bellevue, Jun 28-Jul 2, 2011. Madison: Omnipress, 2011: 809-816.
[29] YANG B S, YIH S W, HE X D, et al. Embedding entities and relations for learning and inference in knowledge bases [C]//Proceedings of the 3rd International Conference on Learning Representations, San Diego, May 7-9, 2015: 1-12.
[30] TROUILLON T, WELBL J, RIEDEL S, et al. Complex embeddings for simple link prediction[C]//Proceedings of the 33rd International Conference on Machine Learning, New York, Jun 19-24, 2016: 2071-2080.
[31] SOCHER R, CHEN D Q, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, Dec 5-8, 2013. Red Hook: Curran Associates, 2013: 926-934.
[32] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D knowledge graph embeddings[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, Feb 2-7, 2018. Menlo Park: AAAI, 2018: 1811-1818.
[33] NGUYEN T D, NGUYEN D Q, PHUNG D. A novel embedding model for knowledge base completion based on convolutional neural network[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Jun 1-6, 2018. Menlo Park: AAAI, 2018: 327-333.
[34] WANG Z, ZHANG J W, FENG J L, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence, Quebec City, Jul 27 -31, 2014. Menlo Park: AAAI, 2014: 1112-1119.
[35] LIN Y K, LIU Z Y, SUN M S, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, Jan 25-30, 2015. Menlo Park: AAAI, 2015: 2181-2187.
[36] CAO Z S, XU Q Q, YANG Z Y, et al. Dual quaternion knowledge graph embeddings[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, Feb 2-9, 2021. Menlo Park: AAAI, 2021: 6894-6902.
[37] NIU G L, LI B, ZHANG Y F, et al. AutoETER: automated entity type representation for knowledge graph embedding [C]//Findings of the Association for Computational Linguistics, Nov 16-20, 2020. Stroudsburg: ACL, 2020: 1172-1181.
[38] NIU G L, ZHANG Y F, LI B, et al. Rule-guided compositional representation learning on knowledge graphs[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, Feb 7-12, 2020. Menlo Park: AAAI, 2020: 2950-2958.
[39] SHAO T Y, LI X Y, ZHAO X, et al. DSKRL: a dissimilarity-support-aware knowledge representation learning framework on noisy knowledge graph[J]. Neurocomputing, 2021, 461: 608-617.

邵天阳(1998—),男,河南尉氏人,博士研究生,主要研究方向为知识图谱。
SHAO Tianyang, born in 1998, Ph.D. candidate. His research interest is knowledge graph.

肖卫东(1965—),男,湖南长沙人,博士,教授,博士生导师,主要研究方向为大数据分析、社会计算等。
XIAO Weidong, born in 1965, Ph.D., professor, Ph.D. supervisor. His research interests include big data analytics, social computing, etc.

赵翔(1986—),男,浙江金华人,博士,教授,博士生导师,主要研究方向为图数据管理、知识图谱构建等。
ZHAO Xiang, born in 1986, Ph.D., professor, Ph.D. supervisor. His research interests include graph data management, knowledge graph construction, etc.