DOI: 10.13718/j.cnki.xdzk.2020.11.005

基于互逆和对称关系补全的知识图谱数据扩展方法

应坚超，　蒲　飞，　徐晨鸥，　杨柏林

浙江工商大学 计算机与信息工程学院，杭州 310018

摘要：知识图谱表示学习方法旨在将知识图谱的实体与关系表示为低维、稠密的向量，并用于高效的语义计算，在知识图谱的构建、融合以及其他方面发挥重要作用. 传统的知识图谱表示学习模型通常考虑了知识图谱中已有的事实，而忽略了知识图谱中隐藏的语义信息，使得表示学习并不能充分地表达原知识图谱的信息. 目前的数据增强知识图谱表示学习模型需要借助第三方工具或者大量人工干预，增强数据的可靠性与稳定性有待加强. 该文基于集合论中所提出的互逆/对称关系概念，提出了关系统计扩展方法(Relationship Statistics Expansion，RSE)方法，即通过统计的方法从现有知识图谱中获取稳定且可靠的先验知识，将其用于数据集扩展. 同时，利用先验知识对互逆关系的表示模长施加约束，更加符合语义逻辑. 该研究分别在 WN18，FB15K，WN11，NELL-995 共 4 个常用数据集上进行链路预测任务来评价模型效果，采用了目前主流的 4 个具有代表性知识图谱表示学习模型 TransE，DistMult，RotatE，HAKE 作为基准. 结合该文提出的 RSE 方法后，RSE-TransE 的 MRR 值分别提高了 7.9%，12%，4.5%，7.2%；RSE-DistMult 的 MRR 值分别提高了 11.3%，5.8%，4.1%，1%；RSE-RotatE 的 MRR 值分别提高了 2.6%，6.7%，5.1%，1%；RSE-HAKE 的 MRR 值分别提高了 3.2%，4.5%，5.5%，11.3%. 实验结果表明，该文提出的基于互逆和对称关系补全的知识图谱数据扩展方法可以挖掘知识图谱中隐含的语义信息，并且能显著提升传统的知识图谱表示学习模型在链路预测任务上的准确率和性能.

关　键　词：知识图谱；表示学习；互逆关系；对称关系；统计方法；数据扩展

中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1673-9868(2020)11-0043-09

随着大数据时代的发展，互联网源源不断地产生不同结构形式的海量动态数据. 为了适应科学研究和生产生活的发展要求，需要找到一种合理有效的方式对产生的大数据进行表述、管理、组织、利用. 学术界和工业界将注意力放在了具有极强开放互通能力和文本处理能力的知识图谱上. 知识图谱最早起源于语义网络，语义网络旨在通过人与机器都可以理解的方式描述客观事物，达到人与机器沟通无障碍的目标. 构建的知识图谱具有高智能化水平网络，能模拟人的认知思维方式，并且支持知识推理. 世界各大互联网公司例如谷歌、百度、微软等均已投入构建相应的知识图谱，随着知识获取、知识融合、知识验证等知识图谱构建的主要技术手段的逐渐成熟，目前互联网已有一些大规模知识库可以获取使用，例如 LinkedData[1]，DBpedia[2]，WordNet[3]，Freebase[4]，YAGO[5]等，并在问答系统[6]、信息检索[7]、推荐系统[8]、自然语言处理等应用领域得到了广泛的使用. 从本质上来说，知识图谱就是一种大规模的语义网络和结构化的语义知识库，描述客观世界中存在的各种实体及关系，通常采用万维网联盟(W3C)发布的资源描述框架(Resource description framework，RDF)三元组(h，r，t)作为表示知识的存储格式，h 和 t 分别代表头实体和

收稿日期：2020-10-15
基金项目：浙江省重点研发项目(2019C01004).
作者简介：应坚超(1996－)，男，硕士研究生，主要从事知识图谱表示学习研究.
通信作者：杨柏林，教授.

尾实体，r 代表关系. 知识图谱构建和应用依赖于知识表示，但是基于三元组的存储方式面临着计算复杂度高、可重用率低、数据稀疏等问题，无法完全充分地刻画客观世界中实体之间的语义关系. 近年来，在自然语言处理、语音识别、图像和文本分析等领域中逐步发展成熟的深度学习技术为大数据时代下的表示学习发展提供了良好基础. 知识图谱表示学习技术的目的是通过机器学习的手段用低维稠密的向量来表示知识图谱中的实体和关系，不但数据稀疏性等问题得到有效解决，而且使其在知识图谱应用层面发挥重要作用. 由于现有的知识图谱通常是不完整的，预测缺失关系是知识图谱中存在的一个基本问题. 最近，以TransE模型[9]为代表的知识图谱表示学习模型对学习实体和关系的低维表示得到了广泛的研究，这些方法被证明用于缺失链路预测是可扩展且有效的. 这些方法的一般直觉是根据观察到的知识事实来建模和推断知识图谱中的关系类型. 例如，一些关系是对称的(如婚姻)，而另一些是反对称的(如亲子关系)；有些关系是其他关系的逆关系(例如，老师和学生)；有些关系可能由其他人组成(例如，我母亲的丈夫是我的父亲). 从观察到的事实中找到建模和推断这些模式的方法是至关重要的，即对称/反对称、互逆和组合，以便预测缺失的环节. 事实上，由于知识图谱基本都是自底向上构建而成，绝大部分表示学习模型在对实体和关系进行向量投影时仅仅考虑知识图谱中现有的直接信息，而忽略了隐藏在知识图谱中可发掘的可靠语义信息，从而导致在表达原知识图谱语义关系时不够准确. 本文对现有知识图谱进行统计，获取各关系下满足互逆/对称关系三元组的数目并计算比例，在设置可信阈值的基础上扩展现有实体和关系的新三元组，实现对现有原知识图谱数据的扩充，直接提高了嵌入学习的向量表示效果. 在此基础上，对已获取的互逆关系添加模长约束，使其语义表达更符合客观逻辑. 实验结果表明基于互逆和对称关系补全的知识图谱数据扩展方法可以显著提高已有知识图谱表示学习模型的准确率和性能.

1 相关工作

1.1 基本的知识图谱表示学习模型

TransE模型受word2vec[10]启发，首次提出将三元组中的关系看作是头实体到尾实体的某种翻译过程，即词向量的平移不变现象，例如：C(king)－C(queen)≈C(man)－C(woman)，其中C(w)就是word2vec学习到的词向量表示. TransE模型结构简单，兼具良好的表达准确率和扩展性能，但是在1-N，N-1和N-N的复杂关系上效果并不尽如人意.

TransH模型[11]在TransE模型的基础上，通过引入特定关系的超平面，将头尾实体的向量投影在该超平面上，并在该超平面上完成从头实体到尾实体的翻译过程. 该模型在复杂关系的链路预测问题上比TransE模型表现更好.

TransR模型认为每一个关系都有其特有的语义空间，即同一个关系中的实体距离相近，而非同一个关系下的实体距离应较远. 该模型将头实体和尾实体通过关系的矩阵映射到一个新的语义空间中再完成翻译过程，进一步提升了TransE在复杂关系中的建模效果. 此外，还提出了CTransR模型，对TransR在聚类分组问题上进行了扩展，但是参数过多的问题导致其不适用于大规模的知识图谱.

TransD模型[12]针对TransR模型参数过多的问题提出了将映射矩阵替换为自适应矩阵乘积，有效减少了参数，提高了计算效能. TransA模型在TransH模型的基础上引入了自适应度量方法，为表示向量中的每个维度增加了一个权重系数，由非负矩阵表示一个关系，提升了模型表示效果. DistMult模型[13]是一种双线性模型，每个关系放弃全矩阵表示而采用对角矩阵表示，具有与TransE相同的可扩展性与更好的表达性能. ComplEx模型将TransE模型实数范围的表示改为复数范围的表示，进一步提高了模型表示性能. RotatE模型[14]相比于TransE模型提供了一种全新的翻译思路，将头实体到尾实体的翻译过程由平面上向量的平移转为空间中向量的旋转，进一步提升了向量表示效果. 并且RotatE模型提出了一种新的基于概率的自对抗负样本采样方法，该方法对绝大部分表示学习模型的准确率都有显著提高.

上述模型均基于TransE模型的嵌入表示思想，从实体和关系的翻译过程以及语义空间角度不断提升模型的适用范围和表达能力，在知识图谱中更多的复杂关系上表现良好. 但是单纯从空间建模出发，仅仅考虑了知识图谱中现有的事实数据，忽略了隐含的逻辑关联和语义信息，使得表示学习模型不能更准确充分表达知识图谱的语义信息.

1.2　数据增强的知识图谱表示学习模型

AMIE[15]关联规则挖掘工具，通过统计的方法将知识图谱中隐藏的语义逻辑规则提取出来并赋予置信度. 例如知识图谱中有这样 2 个三元组："张三，出生于，上海"和"张三，工作于，上海"；通过 AMIE 工具可以获取一条规则：关系"出生于"和关系"工作于"存在一定的关联，且具有一定可信度. 那么可以从已有三元组(李四，出生于，杭州)推导出可能存在的三元组(李四，工作于，杭州). RUGE 模型[16]在 ComplEx 模型的复数表达基础上，借助了 AMIE 工具挖掘获取的大量关联规则，并通过软标签的方式同时迭代更新标注三元组和未标注三元组. 通过这个迭代过程，RUGE 模型将分布式知识表示学习和逻辑推理二者合理融合，发掘了更多隐藏语义信息并提高了模型整体的表达效果.

在选取样本的过程中，为了将一些不相关的错误样本进行过滤，限制了关系的定义域和值域，并将其作为先验知识加入到已有的表示学习模型中去，只有那些满足限制关系条件的三元组参与了建模[17]. 提出不同类型的实体应该具有不同的表示，通过编码器将新添加的实体层次信息加入到表示学习模型中，增大了拥有相同类型实体之间的差距，区分了相似实体类型的表示. 受远程监督思想启发，将丰富的文本信息融入知识图谱中，通过联合已有的三元组事实和文本的额外实体描述，进行复杂关系的表示训练. 并且每一个关系对于不同的头实体和尾实体具有不同的表示，以此解决 1-N,N-1,N-N 的复杂关系表示问题.

上述模型均采用了通过直接添加文本信息、借助挖掘工具、联合训练模型等方法提升了模型训练效果，但是存在 2 个缺点：一个是添加信息过程中很有可能加入了错误信息，无法最大限度提高训练效果；另一个是这些信息的来源都需要经过较复杂的提取和人工干预，可扩展性不够强.

2　基于互逆和对称关系补全的知识图谱数据扩展方法

2.1　符号说明

G 表示为知识图谱，E 表示知识图谱中实体的集合，N 表示实体的总数目，R 表示知识图谱中关系的集合，M 表示关系的总数目. h_i，r_j，t_k 表示知识图谱中的事实三元组，其中 h_i，$t_k \in E$，$r_j \in R$，$0 \leqslant i$，$k \leqslant N$，$0 \leqslant j \leqslant M$. 可信度阈值 $\lambda \in \{0.3, 0.5, 0.7\}$，互逆/对称关系集合 S：$\{s_j \mid (r_i, r_k, p)\} 0 \leqslant i$，$k \leqslant N$，$p \in (0, 1)$，$j > 0$，$p$ 为满足互逆/对称关系三元组的比例.

2.2　对称关系和互逆关系定义

关系的分类方式多种多样，集合论中常见的有 2 种分类依据. 根据对称性可以分为对称关系、反对称关系、非对称关系；根据传递性可以分为传递关系、反传递关系、非传递关系. 而互逆关系是集合论的另一个基本概念之一，本文主要对知识图谱中的互逆关系和对称关系进行统计分析，并给出如下定义：

互逆关系：在特定的知识图谱 G 中存在一定数量比例的事实三元组(h_i，r_j，t_k)与事实三元组(t_k，r_l，h_i)$0 \leqslant l \leqslant M$，那么可以称关系 r_j 与关系 r_l 为互逆关系.

对称关系：在特定的知识图谱 G 中存在一定数量比例的事实三元组 h_i，r_j，t_k 与事实三元组 t_k，r_j，h_i，那么可以称关系 r_j 为对称关系.

通过观察不难看出，对称关系和互逆关系在表现形式上几近相同，当互逆关系中 $r_j = r_l$ 时，即表现为对称关系，这为后续的统计工作简化了思路.

2.3　扩展逻辑原理

现有的知识图谱往往存在内容不够完整的问题，包括实体的补充和关系的添加 2 个部分. 而关系的添加又可以通过已有的知识图谱事实关系去分析计算获得，其中对称关系和互逆关系三元组可以借助统计方法判定并且通过统计结果添加额外新三元组信息.

假设某知识图谱中含有三元组(李明，老师，肖亮)，根据常理推断可以获得原知识图谱中不包含的新三元组(肖亮，学生，李明)，那么"老师"和"学生"是一对互逆关系，这是先验知识并且能扩充到其他符合关系"老师"和关系"学生"的三元组中. 但是机器在对知识图谱进行表示学习的时候并不具备人所具有的先验知识，无法自行获取合理可靠的新三元组，缺少这一部分的信息将影响整个知识图谱的嵌入表达.

同理，假设某知识图谱中含有三元组(李明，同学，高琦)，根据常理推断可以获得原知识图谱中不包含的新三元组(高琦，同学，李明)，那么"同学"是对称关系，这是先验知识并且能扩充到其他符合关系"同

学”的三元组中. 若是能为机器添加这些先验知识并将其用于三元组的训练过程，那么整个知识图谱的表达效果将获得更多提升.

首先是先验知识的判断过程，机器无法单独理解某一个关系的含义，但是可以理解关系与关系(自身)之间的联系，这是利用好对称关系和互逆关系特殊性的前提. 先计算满足互逆/对称关系三元组数量占该关系所有三元组数量的比例，再与预先设置的阈值进行比较来判定某2个关系是否满足互逆关系(某一个关系是否满足对称关系).

其次是将获得的先验知识用于新信息的扩展，机器可以将已经获得的某2个互逆关系(某一个对称关系)对原有知识图谱中未成对的三元组通过关系替换、头尾实体互换的方式形成新三元组.

2.4 互逆和对称关系补全与扩展方法

如图1所示，以关系 r_1 与关系 r_2 为例，现有知识图谱三元组中存在一定数量满足互逆关系形式的三元组对，且满足条件的成对三元组数量占关系 r_1 与关系 r_2 所有三元组数量的比值超过阈值 λ 时，就可以认定关系 r_1 与关系 r_2 为互逆关系.

如图2所示，以关系 r_1 为例，现有知识图谱三元组中存在一定数量满足对称关系形式的三元组对，且满足条件的成对三元组数量占关系 r_1 所有三元组数量的比值超过阈值 λ 时，就可以认定关系 r_1 为对称关系.

互逆/对称关系先验知识的获取依赖于互逆关系三元组的比例计算.

互逆关系比例计算公式：

$$p=\frac{numr_i,\ r_j}{num(r_i)+num(r_j)}$$

其中 $numr_i,\ r_j$ 表示关系 r_i 与关系 r_j 为互逆关系形式的三元组数量，$num(r_i)+num(r_j)$ 表示关系 r_i 与关系 r_j 所有的三元组数量和.

对称关系比例计算公式：

$$p=\frac{numr_i,\ r_i}{num(r_i)}$$

其中 $numr_i,\ r_i$ 表示关系 r_i 为对称关系形式的三元组数量，$num(r_i)$ 表示关系 r_i 所有的三元组数量.

将得到的 p 与 λ 进行比较，若 $p\geqslant\lambda$，则关系 r_i 与关系 r_j 可以认定为互逆关系，并添加到互逆/对称关系集合中，当 $r_i=r_j$ 时表现为对称关系，$s_k=(r_i,\ r_j,\ p)\in S$，反之则不可认定.

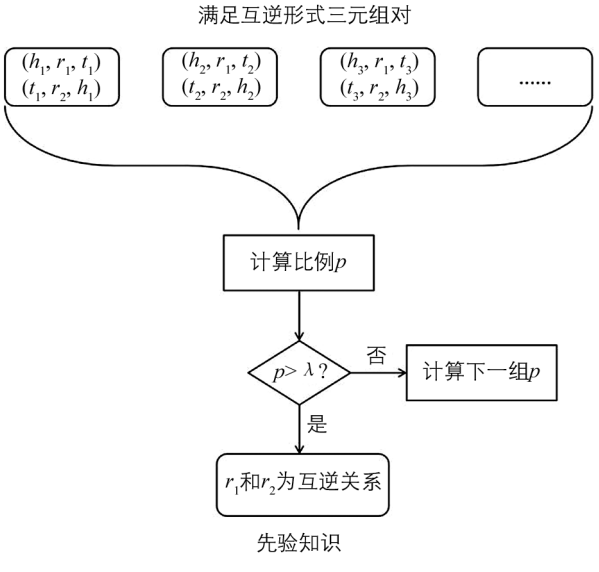

图1 互逆关系统计

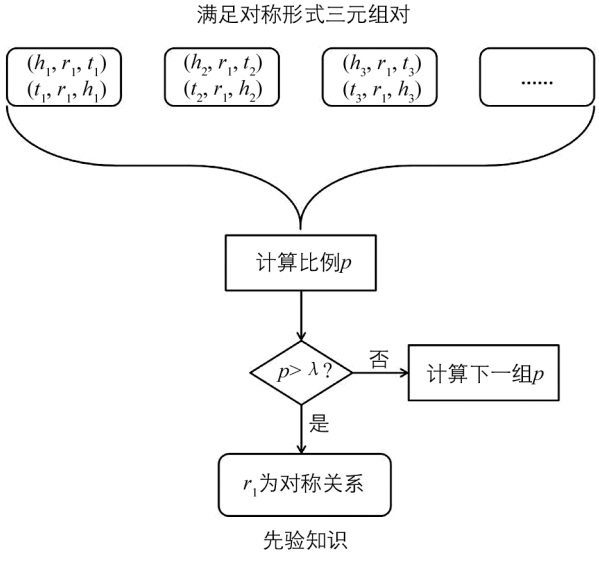

图2 对称关系统计

如图3所示，在已获得认定的互逆关系基础上，将关系 r_1 与关系 r_2 中未成对的三元组通过关系替换

和头尾实体对调的方法形成新的三元组. 如图 4 所示，在已获得认定的对称关系基础上，将关系 r_1 中未成对的三元组通过头尾实体对调的方法形成新的三元组.

对整个知识图谱完成统计后获得互逆/对称关系集合 S，找出原知识图谱中不满足互逆/对称关系的独立三元组，然后根据先验知识集合 S 对这些独立三元组进行扩展.

原独立三元组为 h_i，r_j，t_k，若集合 S 中有互逆关系：$(r_j, r_k, p)p>\lambda$，则新三元组为 t_k，r_k，h_i；若集合 S 中有对称关系：$(r_j, r_j, p)p>\lambda$，则新三元组为 t_k，r_j，h_i.

至此，对原知识图谱互逆和对称关系的统计与扩展方法已完成，新获得的三元组来自知识图谱本身的隐含语义信息，稳定性好，可靠性高，计算复杂度低，能为知识图谱的训练带来更大提升.

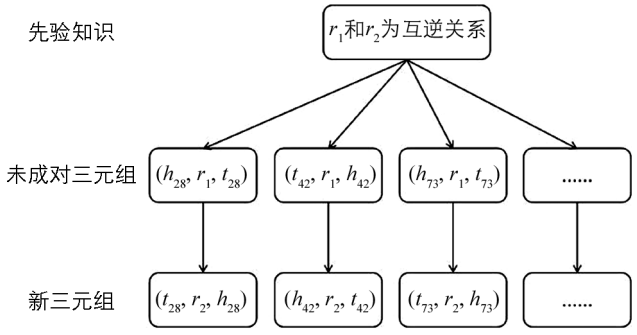

图 3　互逆关系扩展

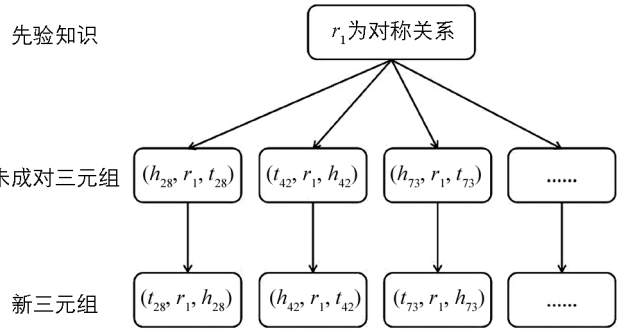

图 4　对称关系扩展

2.5　语义逻辑约束

在对原知识图谱进行统计后，获得的互逆/对称集合 S 不仅可以用来扩展数据集，还可以作为添加语义逻辑约束的条件.

事实上，任何一个互逆关系在语义空间里的表达应该是相互关联的，例如“李明，老师，肖亮”和“肖亮，学生，李明”这一对互逆关系三元组，在语义空间里可以理解为，从“李明”这个实体点出发走到“肖亮”这个实体点的距离和从“肖亮”这个实体点出发走到“李明”这个实体点的距离，二者的距离应该是近乎相等的，这是符合语义逻辑的判断. 所以在已有互逆关系集合信息的基础上，限制如下约束条件：

$$L = || r_i | - | r_j ||, r_i, r_j \in (r_i, r_j, p)$$

3　实验验证

实验通过链路预测任务结果的好坏来评估整个知识图谱嵌入表示准确率. 在表示学习的主流模型下，来验证借助本文的基于互逆/对称关系补全的知识图谱数据扩展方法(简称 RSE 方法)是否达到更好的效果.

3.1　数据集描述

本文在 WN18，FB15K，WN11 和 NELL-995 共 4 个知识图谱广泛使用的标准数据集上评估了本实验的模型，分别将其分为训练集、验证集和测试集，训练集用于训练模型，验证集用于调整参数，测试集则用于评价模型性能，关于这 4 个数据集的信息见表 1.

表 1　WN18，FB15K，WN11 和 NELL-995 数据集

数据集	＃关系	＃实体	＃训练集	＃验证集	＃测试集
WN18	18	40 943	141 442	5 000	5 000
FB15K	1 345	14 951	483 142	50 000	59 071
WN11	11	38 588	112 581	2 609	10 544
NELL-995	200	75 492	149 679	543	3 992

3.2 实验对比方法

本文选取了4个具有代表性的主流表示学习模型TransE，DistMult，RotatE和HAKE作为对照组；分别与RSE方法结合后简称为RSE-TransE，RSE-DistMult和RSE-RotatE，作为实验组. 通过带过滤(filter)的方式进行链路预测实验来确定模型的性能.

其中RSE方法中参数λ具有两方面的作用：在λ可以取得较高的数值时，它可以确保数据扩展中先验知识具有较高的可靠度；在λ取得较低的数值时，它可以让数据集获得更多额外的三元组数目. 经过多次实验发现，在不同数据集中达到最好效果的阈值具有差别性，需要多次实验找到不同数据集对应的最佳λ值. 例如，在WN18中λ可以取0.9；而在NELL-995中，λ需要取0.1才能让扩展方法表现最好.

关于用于模型性能评价的指标，本实验共选用3种. ① MRR：平均倒数排名；② MR：平均排名；③ HITS@N(N=1,3,10)：前N个占比.

3.3 实验设置

为了使本次实验结果在最优的情况下进行对比，所有的对照组和实验组都采用了RotatE的自对抗负样本采样方法. 模型设置的超参数包括7项. 嵌入维度k为：500，1 000，2 000；负样本比例：256，512，1 024；batch size：256，512，1 024；margin：3，6，9，12；自对抗采样参数：0.5，1.0；学习率：0.001，0.000 5，0.000 1，0.000 05，0.000 01；训练迭代次数：150 000. 实体和关系的初始化均采取随机初始化方式.

3.4 实验结果

表2显示了模型在数据集WN18上链路预测任务的实验结果. 在结合本文提出的RSE方法后，MRR指标显示，TransE模型提高了7.9%，DistMult模型提高了11.3%，RotatE模型提高了2.6%，HAKE模型提高了3.2%，MR和HITS@N指标也均有明显提高.

表3显示了模型在数据集FB15K上链路预测任务的实验结果. 在结合本文提出的RSE方法后，MRR指标显示，TransE模型提高了12%，DistMult模型提高了5.8%，RotatE模型提高了6.7%，HAKE模型提高了3.2%，MR和HITS@N指标也均有明显提高.

表4显示了模型在数据集WN11上链路预测任务的实验结果. 在结合本文提出的RSE方法后，MRR指标显示，TransE模型提高了4.5%，DistMult模型提高了4.1%，RotatE模型提高了5.1%，HAKE模型提高了4.5%，MR和HITS@N指标也均有明显提高.

表5显示了模型在数据集NELL-995上链路预测任务的实验结果. 在结合本文提出的RSE方法后，MRR指标显示，TransE模型提高了7.2%，DistMult模型提高了1%，RotatE模型提高了1%，HAKE模型提高了11.3%. 在DistMult和RotatE结合RSE方法测试NELL-995数据集上时提高效果并不明显，可能原因如下：① 硬件限制，模型的参数并未能调到最佳，仅以同参数下的结果用于体现RSE方法的效果. ② 参数相互影响，阈值的设定和嵌入维度、最小步长、负样本数量等其余参数共同影响实验结果.

表2 WN18-链路预测结果

模型	MRR	MR	HITS@1	HITS@3	HITS@10
TransE	0.779	251	0.709	0.820	0.930
RSE-TransE	0.858	122	0.744	0.969	0.977
DistMult	0.685	270	0.530	0.829	0.931
RSE-DistMult	0.798	130	0.673	0.921	0.965
RotatE	0.949	273	0.944	0.952	0.960
RSE-RotatE	0.975	134	0.973	0.976	0.981
HAKE	0.942	186	0.931	0.948	0.960
RSE-HAKE	0.974	76	0.969	0.977	0.982

表 3　FB15K-链路预测结果

模型	MRR	MR	HITS@1	HITS@3	HITS@10
TransE	0.673	48	0.582	0.735	0.827
RSE-TransE	0.793	29	0.715	0.857	0.906
DistMult	0.764	37	0.696	0.809	0.883
RSE-DistMult	0.822	26	0.767	0.860	0.915
RotatE	0.781	42	0.725	0.817	0.877
RSE-RotatE	0.848	29	0.807	0.876	0.919
HAKE	0.664	43	0.555	0.746	0.838
RSE-HAKE	0.709	34	0.607	0.787	0.869

表 4　WN11-链路预测结果

模型	MRR	MR	HITS@1	HITS@3	HITS@10
TransE	0.098	6 495	0.010	0.142	0.264
RSE-TransE	0.143	5 307	0.051	0.191	0.318
DistMult	0.121	7 770	0.063	0.135	0.239
RSE-DistMult	0.162	6 448	0.095	0.183	0.300
RotatE	0.170	7 104	0.113	0.191	0.277
RSE-RotatE	0.221	5 790	0.162	0.242	0.335
HAKE	0.194	4 243	0.128	0.219	0.322
RSE-HAKE	0.249	3 163	0.181	0.272	0.384

表 5　NELL-995-链路预测结果

模型	MRR	MR	HITS@1	HITS@3	HITS@10
TransE	0.279	7 162	0.133	0.401	0.494
RSE-TransE	0.351	10 154	0.259	0.417	0.501
DistMult	0.280	17 296	0.237	0.303	0.358
RSE-DistMult	0.290	13 943	0.245	0.313	0.368
RotatE	0.350	12 519	0.308	0.372	0.414
RSE-RotatE	0.360	12 531	0.320	0.380	0.424
HAKE	0.372	14 726	0.326	0.401	0.447
RSE-HAKE	0.485	9 289	0.420	0.518	0.610

结果表明，本实验的方法能有效且稳定提高三大模型在两大数据集链路预测任务上的结果，整体的表示学习准确率得到了提升. 由此可见，本实验的方法具备以下特点：① RSE 方法能大幅提升知识表示学习模型的嵌入效果；② RSE 方法可以有效结合几乎所有表示学习模型，结果均有提高，具有良好的扩展性；③ RSE 方法可以在不同数据集上均表现良好，具备良好的适应性.

4　结　论

知识图谱除了有丰富的现有事实信息，还有一部分隐藏信息可以通过推理、统计、计算等方式获取并加以利用，提高整个知识图谱的表示效果. 本文针对传统知识图谱表示学习方法未使用隐藏信息的缺点，通过统计的方法对原知识图谱的信息进行分析，利用互逆/对称 2 种较为特殊的关系整理出比例规则，作为

原知识图谱的先验知识添加到训练集的扩展中，同时利用获取的比例规则对互逆关系的模长进行约束，分别结合3个主流表示学习模型在链路预测任务上进行对比实验并获得了明显的效果提升. 本实验结果表明，本文提出的基于互逆/对称关系的知识图谱数据扩展方法能有效提升常见知识表示学习模型的效果，并且在不同的数据集上均适用，说明本方法具有良好的可扩展性和适应性.

参考文献：

[1] BIZER C, HEATH T, IDEHEN K, et al. Linked Data on the Web (LDOW2008) [C] //Proceeding of the 17th international conference on World Wide Web-WWW '08. April 21-25, 2008. Beijing, China. New York: ACM Press, 2008: 1265-1266.

[2] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: A Nucleus for a Web of Open Data [M]//The Semantic Web. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007: 722-735.

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

[4] BOLLACKER K, 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-SIGMOD '08. June 9-12, 2008. Vancouver, Canada. New York: ACM Press, 2008: 1247-1250.

[5] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a Core of Semantic Knowledge [C] //Proceedings of the 16th international conference on World Wide Web-WWW '07. May 8-12, 2007. Banff, Alberta, Canada. New York: ACM Press, 2007: 697-706.

[6] HAO Y C, ZHANG Y Z, LIU K, et al. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge [C] //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 221-231.

[7] XIONG C Y, POWER R, CALLAN J. Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding [C] //Proceedings of the 26th International Conference on World Wide Web. Perth Australia. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017: 1271-1279.

[8] ZHANG F Z, YUAN N J, LIAN D F, et al. Collaborative Knowledge Base Embedding for Recommender Systems [C] //Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA. New York, NY, USA: ACM, 2016: 353-362.

[9] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating Embeddings for Modeling Multi-Relational Data [C/OL] //(2013-09-26) [2019-01-10]. https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.

[10] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space [EB/OL]. 2013: arXiv: 1301.3781 [cs.CL]. https://arxiv.org/abs/1301.3781.

[11] WANG Z, ZHANG J W, FENG J L, et al. Knowledge Graph Embedding by Translating on Hyperplanes [C/OL] //(2014-07-08) [2019-01-20]. https://openreview.net/forum?id=Sy-Cb1-ubS.

[12] JI G L, HE S Z, XU L H, et al. Knowledge Graph Embedding via Dynamic Mapping Matrix [C] //Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015: 687-696.

[13] YANG B S, YIH W T, HE X D, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases [EB/OL]. 2014: arXiv: 1412.6575 [cs.CL]. https://arxiv.org/abs/1412.6575.

[14] SUN Z Q, DENG Z H, NIE J Y, et al. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space [EB/OL]. 2019: arXiv: 1902.10197 [cs.LG]. https://arxiv.org/abs/1902.10197.

[15] GALÁRRAGA L A, TEFLIOUDI C, HOSE K, et al. AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases [C] //Proceedings of the 22nd international conference on World Wide Web-WWW '13. May 13-17, 2013. Rio de Janeiro, Brazil. New York: ACM Press, 2013: 413-422.

[16] GUO S, WANG Q, WANG L H, et al. Knowledge Graph Embedding with Iterative Guidance from Soft Rules [EB/OL]. 2017: arXiv: 1711. 11231 [cs. AI]. https: //arxiv. org/abs/1711. 11231.

[17] KROMPAß D, BAIER S, TRESP V. Type-Constrained Representation Learning in Knowledge Graphs [M] //The Semantic Web-ISWC 2015. Cham: Springer International Publishing, 2015: 640-655.

A Knowledge Graph Data Expansion Method Based on Reciprocal and Symmetric Relationship Completion

YING Jian-chao, PU Fei, XU Chen-ou, YANG Bai-lin

School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China

Abstract: The knowledge map representation learning method aims to represent the entities and relationships of a knowledge map as low dimensional and dense vectors, and is used for efficient semantic computing. It plays an important role in the construction, fusion and other applications of knowledge maps. The traditional knowledge map representation learning model usually considers the existing facts in the knowledge map, but ignores the semantic information hidden in it, which makes representation learning unable to fully express the information of the original knowledge map. The current data enhancement knowledge map representation learning model needs the help of third-party tools or a large number of manual interventions, so the reliability and stability of data need to be strengthened. Based on the concept of reciprocal/symmetric relation in the set theory, this paper proposes an RSE (relationship statistics expansion) method to obtain stable and reliable prior knowledge from the existing knowledge map by statistical methods, and then use it to expand the data set. At the same time, it is more consistent with the semantic logic to constrain the length of the representation module of the reciprocal relationship by using the prior knowledge. The effect of the model is evaluated through the link prediction task on three common datasets: wn18, fb15k and WN11. Four representative knowledge map representation learning models (TransE, DistMult, RotatE and HAKE) are used as the benchmark. Combined with the RSE method proposed in this paper, the MRR values of RSE-transE are increased by 7.9%, 12%, 4.5% and 7.2%; the MRR values of RSE-DistMult are increased by 11.3%, 5.8%, 4.1% and 1%; the MRR values of RSE-RotatE are increased by 2.6%, 6.7%, 5.1% and 1%; and the MRR values of RSE-HAKE are increased by 3.2%, 4.5%, 5.5% and 11.3%, respectively. The experimental results show that the proposed data expansion method based on reciprocal and symmetric relation completion can mine the semantic information hidden in the knowledge map, and can significantly improve the accuracy and performance of the traditional knowledge map representation learning models in the link prediction task.

Key words: knowledge graph; representation learning; reciprocal relationship; symmetric relationship; statistical method; data expansion

责任编辑　王新娟