

知识图谱嵌入技术研究进展

舒世泰, 李松⁺, 郝晓红, 张丽平

哈尔滨理工大学 计算机科学与技术学院, 哈尔滨 150080

+ 通信作者 E-mail: lisongbeifen@163.com

摘要: 知识图谱嵌入(KGE)是知识图谱领域一个新的研究热点, 旨在利用词向量的平移不变性将知识图谱中实体和关系嵌入到低维向量空间, 进而完成知识表示。以解决实际问题的类型为划分依据, 首先, 阐述了四类主要的知识图谱嵌入方法, 包括基于深度学习的方法、基于图形特征的方法、基于翻译模型的方法以及基于其他模型的方法, 对每种模型的算法思想进行详细阐述, 总结了每种模型的优缺点; 其次, 从常用数据集、评价指标、算法、实验四方面对知识图谱嵌入算法实验进行分析与归纳, 对嵌入方法做了横纵向对比; 最后, 从解决实际问题的角度出发, 给出了知识图谱嵌入技术未来的发展方向。通过研究, 发现在基于深度学习的方法中, LCPE 模型的效果最好; 在基于图形特征的方法中, TCE 模型的效果最好; 在基于翻译模型的方法中, NTransGH 模型的效果最好。今后的研究可以在 LCPE、TCE、NTransGH 的基础上进行拓展, 不断提高链接预测和三元组分类的实验效果。

关键词: 知识图谱嵌入(KGE); 知识表示; 知识图谱补全(KGC); 链接预测; 三元组分类

文献标志码:A **中图分类号:**TP319

Knowledge Graph Embedding Technology: A Review

SHU Shitai, LI Song⁺, HAO Xiaohong, ZHANG Liping

School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China

Abstract: Knowledge graph embedding (KGE) is a new research hotspot in the field of knowledge graphs, which aims to apply the translation invariance of word vectors to embedding entities and relationships of the knowledge graph into a low-dimensional vector space to complete knowledge representation. In this paper, it is mainly concerned with the classification according to the types of practical problems to be solved. Firstly, it expounds four major types of embedding methods of knowledge graph, including deep learning-based methods, graphical features-based methods, translation model-based methods, and other model-based methods. The algorithm ideas of each model are elaborated, and the advantages and disadvantages of each model are concluded. Secondly, the algorithm experiment of knowledge graph embedding is analyzed and summarized from the four aspects of commonly used data sets, evaluation indicators, algorithms, and experiments, then a horizontal and vertical comparison of the embedding

基金项目: 国家自然科学基金(61872105, 62072136); 黑龙江省自然科学基金(LH2020F047); 黑龙江省留学归国人员科学基金项目(LC2018030); 黑龙江省教育厅科学技术研究项目(12531z004); 国家重点研发计划(2020YFB1710200)。

This work was supported by the National Natural Science Foundation of China (61872105, 62072136), the Natural Science Foundation of Heilongjiang Province (LH2020F047), the Scientific Research Foundation for Returned Scholars Abroad of Heilongjiang Province (LC2018030), the Technology Research Project of Heilongjiang Provincial Education Department (12531z004) and the National Key Research and Development Program of China (2020YFB1710200).

收稿日期: 2021-03-25 **修回日期:** 2021-06-30

method is made. Finally, from the perspective of solving practical problems, the future direction of knowledge graph embedding technology is given. Through research, it is discovered that in the deep learning-based method, LCPE achieves the best effect; in the graphical features-based method, TCE makes the best impression; whereas in the translation model-based method, NTransGH responds most optimistically. Future researches can be expanded on the basis of LCPE, TCE, and NTransGH to continuously improve the experimental effects of link prediction and triplets classification.

Key words: knowledge graph embedding (KGE); knowledge representation; knowledge graph completion (KGC); link prediction; triple classification

知识图谱(knowledge graph, KG)^[1]是一种用图模型来描述知识和建模世界万物之间关系的技术方法,属于语义层面的知识库(knowledge base, KB)^[2],通常以网络的形式出现。在网络中,节点表示实体,节点之间的边表示关系,并使用三元组(h, r, t)对知识进行表示。KG最早应用于提升搜索引擎的能力。随后,KG在辅助智能问答、自然语言处理、大数据分析、推荐计算、可解释性人工智能等方面展现出较大的应用价值^[3-5]。其中,知识表示是这些应用的基础。然而,由于知识的不断积累,KB的规模扩展非常快,知识形式也变得越来越多样化^[6-8]。以传统的形式对知识进行表示的弊端越来越明显,如实体间语义关系的推理变困难、数据稀疏性问题严重、计算的复杂性高、难以运用到大型KG上等。

近年来,以知识表示为代表的知识图谱嵌入(knowledge graph embedding, KGE)^[9]技术取得了新的研究进展,提出了将稀疏的三元组(h, r, t)语义信息映射到低维稠密的向量空间,将实体与关系的推理转变为对象间距离的计算,计算的距离越小,实体间的相似性越高^[10]。同时,KGE通过对实体和关系进行表示,能够解决传统表示方法面临的数据稀疏、知识推理困难等问题,并且促进了知识获取、实体消歧性能的不断提升^[11-13]。

根据KGE的典型特性,学者们提出了众多KGE模型,主要分为两类:平移距离模型^[14]、语义匹配模型^[15]。对于平移距离模型,比较经典的有TransE(translating embedding)^[16]、TransA(translating adaptive)^[17]、TransR(translating in relation space)^[18]、TransH(translating on hyperplanes)^[19]、TransM(translating with relational mapping)^[20]、KG2E(knowledge graphs with Gaussian embedding)^[21]以及SE(structured embedding)^[22]模型等,这些模型的共同点是评分函数采用距离进行衡

量。对于语义匹配模型,比较有代表性的是RESCAL模型^[23]、SME(semantic matching energy)模型^[24]、NTN(neural tensor network)模型^[25]等,这些模型的共同点是评分函数采用相似度进行衡量。然而,KGE模型仍然面临知识图谱补全(knowledge graph completion, KGC)完成不足、图形结构特征未被充分利用、无法处理复杂的语义关系等问题。本文将对现有的KGE方法进行分析与研究,通过整理与归纳,总结目前研究面临的挑战,并对未来的研究趋势进行展望。本文的主要贡献如下:

(1)对KGE方法进行了较为全面的分类,以解决问题类型作为分类依据,分为基于深度学习的方法、基于图形特征的方法、基于翻译模型的方法以及基于其他模型的方法。

(2)详细阐述每种模型的算法思想,归纳并分析KGE方法中每种算法的优点和存在的局限性;最后从方法分类、文献发表年份、模型优缺点、所用数据集、评价指标和算法思想等角度对知识图谱嵌入方法做了横纵向比较。

(3)列出所述方法常用的实体关系数据集;对方中常用评价指标进行说明;以WN18和FB15K数据集为例,对相关方法在该两种数据集上的链接预测结果进行对比和分析。

(4)讨论了当前研究的难点问题,并预测了未来的研究趋势。

1 知识图谱嵌入概述

1.1 知识图谱嵌入概念

KGE旨在将KG中包括实体和关系的内容映射到低维连续向量空间中,也称知识表示学习^[26]。在词向量的启发下,考虑如何将知识图谱中的实体和关系映射到连续向量空间中,并包含一些语义层面的

信息^[27-29],使得在下游任务中更充分应用知识图谱,如智能问答、关系抽取等任务。连续向量的表达蕴含着更多的语义信息,更容易被计算机理解和操作^[30-31]。

1.2 知识图谱嵌入的特点

传统的知识表示方法,如早期专家系统时代的知识表示方法都是以符号逻辑为基础进行知识表示,其特点是易于刻画离散、显性的知识,具有较好的可解释性。但仍有许多不能用符号来刻画连续、隐形的知识,在表示过程中失去鲁棒性,从而在下游任务中难以达到预期效果。通过嵌入(embedding)技术将KG中的实体和关系映射到低维连续向量空间中,提高了知识表示的能力。该方法主要有以下特点:

(1)使用向量的表示方式可以提高应用时的计算效率。将KG的实体和关系映射到向量空间中,实现语义相似度计算等复杂操作^[32-33],计算效率显著提高,有效解决数据稀疏等问题。

(2)增加了下游应用设计的多样性。用向量表示后,KG更适用于当前流行的机器学习算法,例如神经网络等方法^[34-35]。因为下游应用输入的并不是符号,所以可以考虑的方法也不仅限于图算法。

(3)将KGE作为下游应用的预训练向量输入,使得输入的信息不再是孤立的不包含语义信息的符号,而是已经经过一次训练,并且包含一定信息的向量^[36-37],还可用来监督神经网络的训练过程。

2 知识图谱嵌入方法

目前,学者们的研究主要集中在基于深度学习的方法、基于图形特征的方法、基于翻译模型的方法三方面,少数学者的研究集中在基于其他模型的方法。这四方面虽然已经提出了一些成果,但是,很少有学者能够在链接预测、三元组分类以及实体对齐任务上取得实质性的突破,提升的性能也并不是很明显。因此,这四方面在今后很长一段时间内仍然是研究的热点。

2.1 基于深度学习的方法

KGC问题是指在知识图谱中某些链路往往是不完整的,即图中缺少链接。例如,在Freebase和DBpedia中,超过66%的人条目缺少出生地。先前有关KGC的工作主要集中在浅、快速的模型上,这些模型可以缩放到大型KG上。但是,与深度多层模型相比,这些模型学习的表达功能较少,可能会限制其性能^[38]。

针对此问题,研究者提出若干模型,比较有代表性的是ComplEx模型^[39]、ConvE模型^[40]、LCPE(local combination projection embedding)模型^[41]。

ComplEx模型的核心是利用潜在因子分解和复数的嵌入。算法思想是利用复数空间嵌入的组合来处理多种二元关系,包括对称和反对称关系,解决链接预测问题。该方法首先证明了在实体之间只有单一关系的方阵情况下使用复向量嵌入方法,然后推广到三阶张量中的一组堆叠方阵,以表示多个关系。该方法还认为只要使用正确的表示方法,标准的嵌入点积可以是一个非常有效的复合函数;当使用复向量时,由于它涉及两个向量之一的共轭转置,点积通常被称为Hermitian点积,此时点积不再是对称的,而反对称关系可以根据所涉及实体的顺序得到不同的分数。因此复向量可以有效地捕获反对称关系,同时保留点积的计算优势,保证在空间和时间复杂度上都是最优的。由于只使用Hermitian点积,与神经张量网络等模型相比,该方法更简单^[42-44]。

嵌入的点积可很好地缩放,并且可以处理关系的对称性和自反性,使用适当的损失函数甚至可以实现传递性。同时,该方法可扩展到大规模数据集上。在基于FB15K数据集的链接预测实验中,该模型性能比TransE^[16]提高19.9%,具有较好的实验效果。然而,此模型也存在一些缺陷,如预测性能偏低,生成的负采样较少,训练时间偏长,不能同时对所有的关系模式和复杂关系进行建模和推理^[45]。

ConvE模型的核心是使用多层次卷积神经网络。算法思想是使用2D卷积来预测KG中缺失的链接,通过卷积层和全连接层对输入实体和关系之间的相互联系进行建模。如图1所示,具体过程如下:

- (1)将实体和关系嵌入进行重塑和连接;
- (2)将所得矩阵作为卷积层的输入;
- (3)将所得的特征映射张量矢量化并投影到k维空间中;
- (4)与所有候选对象嵌入匹配。

该模型的主要特点是得分函数由二维嵌入上的卷积定义。得分函数为:

$$\psi_r(\mathbf{e}_s, \mathbf{e}_o) = f(vec(f([\bar{\mathbf{e}}_s; \bar{\mathbf{r}}_r]^*\omega))\mathbf{W})\mathbf{e}_o \quad (1)$$

式中, \mathbf{e}_s 和 \mathbf{e}_o 分别为头实体 s 和尾实体 o 的嵌入表示, $r_r \in \mathbb{R}^k$ 是依赖于 r 的一个关系参数, \mathbf{W} 为权值矩阵, $\bar{\mathbf{e}}$, 和 $\bar{\mathbf{r}}$, 分别表示 \mathbf{e}_s 和 \mathbf{r}_r 的 2D 重塑。为了训练模型参

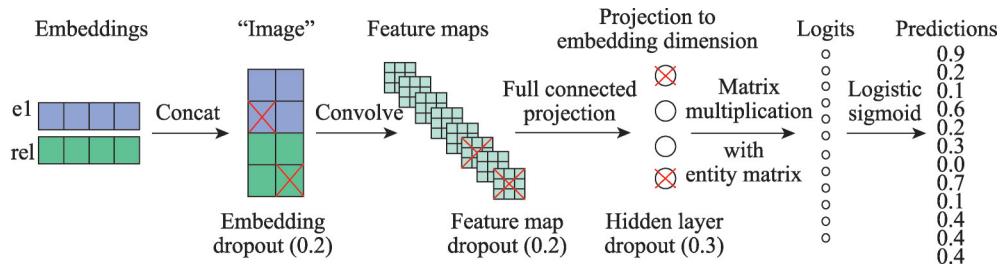


Fig.1 Process diagram of ConvE

图1 ConvE流程图

数,该模型使用逻辑回归函数来表示得分情况,即:

$$p = \sigma(\psi_r(e_s, e_o)) \quad (2)$$

并最小化交叉熵损失函数，即：

$$\mathcal{L}(p, t) = -\frac{1}{N} \sum_i (t_i \cdot \log_a p_i + (1-t_i) \cdot \log_a (1-p_i)) \quad (3)$$

式中, N 表示实体数量, t 表示维度为 $\mathbf{R}^{1 \times 1}$ 或维度为 $\mathbf{R}^{1 \times N}$ 的标签向量, 判别关系是否存在, 若不存在则为 0。同时, 该方法为更快训练文献[46]模型, 使用修正线性单元作为非线性函数 f , 并在每一层之后进行批量归一化处理, 以提高收敛速度。此外, 在嵌入运算和卷积运算后的特征映射和全连接层上使用了 Dropout 算法及 Adam 优化器, 对标签进行平滑处理, 防止过拟合现象。

在 WN18 和 FB15K 数据集上, ConvE 性能优于 ComplEx, 具有较高的参数效率, 在对具有高度关联性的节点建模时也非常有效。但是, 它仍然存在一些局限性, 例如嵌入之间的交互次数偏低, 模型卷积深度偏浅。

LCPE 模型的核心是使用参数共享的神经网络。该模型思想是将 ProjE 模型和实体相似度信息相结合,先判断两个实体是否有关系,后判断该关系的具体类型,由于相似的实体嵌入向量在向量空间中距离更近,可根据实体分布稠密的局部空间来判断实体之间的具体关系类型。如图 2 所示,该模型是由判断两个实体之间是否存在关系和实体之间相似度的网络构成。

在图 2 中, \mathbf{W}^c 是由候选实体向量组成的矩阵, \mathbf{W}^E 是由实体向量组成的矩阵, \mathbf{W}^R 是由关系向量组成的矩阵, E_i 和 E_j 分别是从 \mathbf{W}^c 和 \mathbf{W}^E 中提取出的一个实体向量, \mathbf{R} 是从 \mathbf{W}^R 中提取出的关系向量, D_e 和 D_r 分别代表组合矩阵。

该模型将 ProjE 模型和实体间的相似度结合，定义如下得分函数：

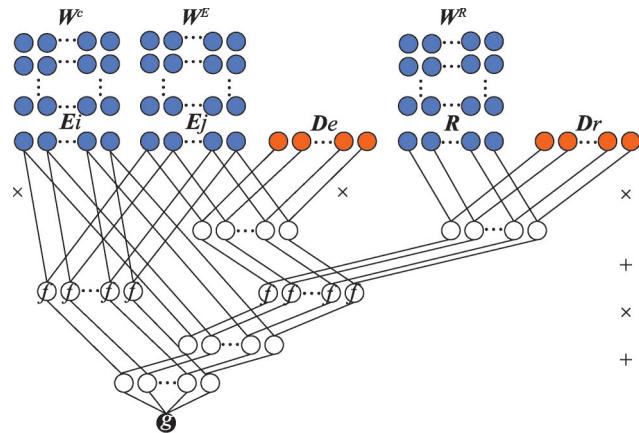


Fig.2 LCPE model structure

图2 LCPE模型结构

$$h(e, r)_i = g(\mathbf{W}_{[i,:]}^c f(e \oplus r) + f(\mathbf{W}_{[i,:]}^c \cdot e) + b_p) \quad (4)$$

式中, $h(e, r)_i$ 指代实体集中第 i 个实体的得分, f 和 g 表示激活函数, $\mathbf{W}^c \in \mathbb{R}^{s \times k}$ 表示候选实体矩阵, b_p 表示偏移量, $e \oplus r$ 表示实体和关系的组合运算, $\mathbf{W}_{[i,:]}^c \cdot e$ 表示实体向量 e 和候选实体集中第 i 个实体的嵌入向量的内积。用 softmax 和 tanh 作为激活函数替换后得到的得分函数为:

$$h(e, r)_i = \frac{\exp(\mathbf{W}_{[i,:]}^c \tanh(e \oplus r) + \tanh(\mathbf{W}_{[i,:]} e_j) + b_p)}{\sum_i \exp(\mathbf{W}_{[i,:]}^c \tanh(e \oplus r) + \tanh(\mathbf{W}_{[i,:]} e_j) + b_p)} \quad (5)$$

并利用 listwise 方法进行训练。LCPE 模型的损失函数定义如下：

$$L(e, r, y) = -\sum_i \frac{1(y_i = 1)}{\sum_j 1(y_j = 1)} \log_a h(e, r)_i \quad (6)$$

在公开数据集WN18中，LCPE正例三元组的MeanRank比ProjE提高了11.0，而正例三元组的Hit@10比ProjE提升了0.20%；在FB15K中，MeanRank提前了7.5，Hits@10平均提升了3.05%。然而，此模型也存在一些局限性，例如共享变量神经网络层数偏浅。

无论是 ComplEx, 还是 ConvE 与 LCPE, 这些都是非常新颖的嵌入模型, 能够较好地用于 KGC 工作, 只是每个模型的侧重点不同, ComplEx 模型侧重于使用 Hermitian 点积, ConvE 模型侧重于使用多层卷积网络, LCPE 模型侧重于将 ProjE 模型和实体相似度信息相结合。除此之外, 也可以考虑使用卷积神经网络改进最先进的模型, 每个三元组都表示为一个三列矩阵, 其中每个列向量代表一个元素, 再将此三列矩阵输入到卷积层, 在卷积层上对矩阵操作多个过滤器以生成不同的特征图; 将这些特征图连接到一个表示输入三元组的单个特征向量中, 通过点积将特征向量与权重向量相乘以返回分数, 来预测三元组是否有效。同时, 也可以考虑融合稀疏的关注机制, 通过共享概念探索三元组中隐藏的关系。

2.2 基于图形特征的方法

知识图谱是由若干实体和关系组成的一个复杂的多关系图形网络结构。图形特征问题是指当前知识图谱嵌入方法不能很好地利用 KG 中图形结构特征。传统的 KGE 模型在学习时注重将 KG 视为一组独立的三元组, 这样会忽略图结构中的重要信息, 包括实体特征信息、三元组上下文等。为解决此问题, 研究者们提出若干模型, 比较有代表性的包括 TCE (triple-context-based knowledge embedding) 模型^[47]、DPTTransE 模型^[48]。

TCE 模型也称三重上下文的知识嵌入模型。算法思想是充分利用 KG 中的图形结构特征, 尤其围绕三元组的局部结构, 即由邻居上下文和路径上下文组成的三元组上下文, 在统一的框架中表示三元组及其上下文的结构信息(如图 3), 这样三元组上下文中的结构信息就可以结构化。

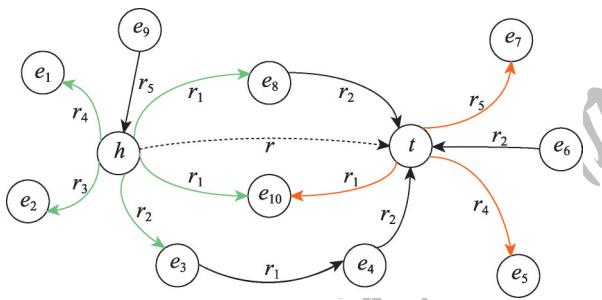


Fig.3 Description of triple context of triples in KG

图3 KG 中三元组的三重上下文的说明

对于邻居上下文, 给定一个实体 e , 实体 e 的邻

居上下文是一个集合:

$$C_N(e) = \{(r, t) | \forall r, t(e, r, t) \in \mathbb{K}\} \quad (7)$$

即以 e 为头实体的三元组中出现的所有关系-尾实体对。例如在图 3 中, 实体 h 的邻居上下文是:

$$C_N(h) = \{(r_1, e_{10}), (r_1, e_8), (r_2, e_3), (r_3, e_2), (r_4, e_1)\} \quad (8)$$

对于路径上下文, 是指 KG 中的一个实体到另一个实体的一组路径。这种结构特征有利于对实体对之间的关系进行建模并捕捉到它们的信息交互。例如在图 3 中, 实体 h 和实体 t 的路径上下文是:

$$C_p(h, t) = \{(r_1, r_2), (r_2, r_1, r_2)\} \quad (9)$$

因此该模型可形式化表示为:

$$C(h, r, t) = C_N(h) \cup C_p(h, t) \quad (10)$$

传统模型的得分函数只与实体和关系的嵌入有关, 如 TransE 的得分函数为:

$$\|h + r - t\|_{L1/L2} \quad (11)$$

而该模型的得分函数是在此基础上融合三重上下文得到的, 形式化为:

$$f(h, r, t) = P((h, r, t)|C(h, r, t); \Theta) \quad (12)$$

式中, $C(h, r, t)$ 表示 (h, r, t) 的三重上下文。一个三元组得分越高, 表明它在更大程度上具有不变性。但是, 此模型的局限性是在基线上不如经典的翻译模型。

DPTTransE 模型的核心是利用判别路径进行嵌入表示。算法思想是采用联合学习机制, 将基于图像特征模型和基于潜在特征模型相结合。

对于潜在特征模型, 在 KG 中, 多步关系和中间实体带来了更多的交互信息, 因此在多步关系和直接关系之间建立强相关性是非常必要的。为此该方法提出语义关联假设, 假设直接关系可以通过多步关系路径的线性组合来重构, 这些判别路径可以提高多个关系之间的嵌入能力。但是, 基于潜在特征模型的语义信息丢失是不可避免的, 须根据上述假设, 将内在损失与语义损失一起最小化, 公式如下:

$$f_e(h, r_k, t) = \|h + r_k - t\| + \frac{\lambda_1}{N} \|\alpha_k P(h, t) - r_k\| \quad (13)$$

式中, $r_k = \alpha_k P(h, t)$ 表示直接关系的语义可以通过给定相似实体对的多步关系线性组合来近似重构, $P(h, t) = \{p_1, p_2, \dots, p_a\}$ 表示多个关系路径的集合, $N = \sum_{p, j \in P(h, r_k, t)} \prod_{\theta \in \alpha_{k,j}} \theta$

为归一化因子, λ_1 用于平衡左右两部分, 函数得分越小, 说明三元组被翻译的效果越好。

对于图像特征模型, 为衡量每条路径的可靠性,

引入PRA-style方法,该方法不同于原始PRA,而是使用路径作为特征来预测实体对之间的关系。在关系聚类阶段,使用K-means算法将相似的关系分成一组;在特征计算阶段,基于实体相似性假设计算每条路径特征值,相似性定义如下:

$$S(h_i, t_i) = \sum_{z \in \Gamma(h_i) \cap \Gamma(t_i)} \frac{1}{d_z} \quad (14)$$

式中, z 是给定实体对 (h_i, t_i) 的每条路径的中间实体, d_z 表示实体 z 的度。

基于图特征在发现语义相关性和为实体-关系提供更精确、更有判别性的语义嵌入方面起着关键作用。该模型可同时从潜在特征和图形特征中学习,建立这两个特征之间的相互联系。不足之处在于基于图形特征必须与三元组交互。

TCE模型、DPTransE模型都能够较好地处理图形特征问题。TCE模型的核心是利用三元组上下文,虽然这种思想并不是第一次出现,但是,将其利用到评分函数中还是首次出现,改进的空间仍然很大,不足之处是在基线上不如经典的翻译模型。DPTransE模型的核心是利用判别路径的嵌入,将不同的模型进行融合,互相取长补短已经成为当下研究的新常态,此模型今后的发展空间较广阔,不足之处是基于图的特征必须与三元组交互。

2.3 基于翻译模型的方法

为了解决KG中复杂关系问题,许多基于翻译的模型将KG中的实体和关系嵌入到连续的向量空间中,并将关系编码为该空间中的翻译操作,从而获得更好的性能^[49]。这些模型在表达KG的自反性、1-N、N-1和N-N等复杂关系时都有局限性。针对这个问题,比较有代表性的有NTransGH模型^[50]、STransH模型^[51]、TransG模型^[52]。

NTransGH模型的核心是将神经网络与广义超平面的翻译机制相结合。算法思想是首先通过广义超平面转换机制,将关系建模为转换操作,使得实体在不同的关系中有不同的角色;然后设计一个神经网络以捕获复杂的关系模式,使用一组特殊矢量而不是TransH中的一个法向量来确定广义超平面^[53-54]。为了表达复杂的关系,引入了两层神经网络来定义得分函数:

$$f(h, r, t) = \text{sigmoid}(\mathbf{W}_2^\top [\text{ReLU}(\mathbf{W}_1 \mathbf{m})]) \quad (15)$$

式中, $\mathbf{W}_1 \in \mathbb{R}^{n \times n}$ 、 $\mathbf{W}_2 \in \mathbb{R}^n$ 为权重参数, $\mathbf{m} = \mathbf{M}_r \mathbf{h}_\perp + \mathbf{r} - \mathbf{M}_t \mathbf{t}_\perp$

为投影后经过平移操作得到的三元组。

NtransGH模型可视化如图4所示。在模型训练时,通过替换语义相似实体来提高生成负例的质量。该模型的损失函数为:

$$L = - \left\{ \sum_{(h, r, t) \in P} y \log_a f(h, r, t) + \sum_{(h, r, t) \in N} (1 - y) \log_a (1 - f(h, r, t)) \right\} + C(A_1 + A_2) \quad (16)$$

其中, A_1 、 A_2 表示如下:

$$A_1 = \sum_{e \in E} [\|\mathbf{e}\|_2^2 - 1]_+ + \sum_{r \in R} [\|\mathbf{r}\|_2^2 - 1]_+ \quad (17)$$

$$A_2 = \sum_{r \in R} \left[\frac{\sum \mathbf{W}_r^{i\top} \mathbf{W}_r^j}{\|\mathbf{W}_r^j\|_2} - \varepsilon^2 \right]_+ \quad (18)$$

式中, C 是一个用于衡量软约束重要性的超参数, P 是正例三元组的集合, N 是负例三元组的集合;在式(17)、式(18)中, \mathbf{W}_r^i 和 \mathbf{W}_r^j 表示向量, ε 是小标量, $[x]_+ \triangleq \max(0, x)$ 。

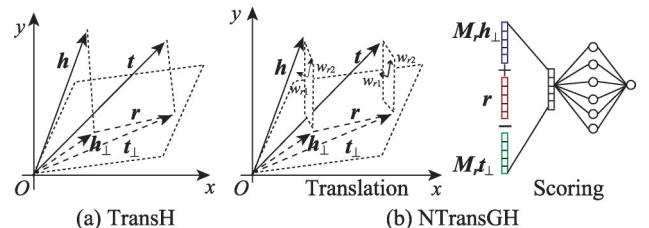


Fig.4 Comparison of TransH and NTransGH models
图4 TransH 和 NTransGH 模型比较

与TransH^[19]相比,该模型基于FB15K的链接预测实验中,Hits@10提高了17.1%。该模型缺陷是需要操作的参数偏多,训练难度大。

STransH模型的核心是分别在实体空间和关系空间中建模。算法思想是将SE模型与TransE模型进行结合,借鉴TransH模型的思想,引入投影到特定关系超平面的机制。具体方法:首先将三元组中的头实体 h 和尾实体 t 映射到给定关系的超平面 \mathbf{w}_r 上,分别用 \mathbf{h}_\perp 和 \mathbf{t}_\perp 表示,用关系向量 \mathbf{r} 将超平面上的 \mathbf{h}_\perp 和 \mathbf{t}_\perp 联系起来。即:

$$\begin{cases} \mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_\perp^\top \mathbf{h} \mathbf{w}_\perp \\ \mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_\perp^\top \mathbf{t} \mathbf{w}_\perp \end{cases} \quad (19)$$

由此得到该模型的得分函数,即:

$$f_r(h, t) = g(\|\mathbf{W}_{r,1} \mathbf{h}_\perp + \mathbf{r} - \mathbf{W}_{r,2} \mathbf{t}_\perp\|_{L1/L2}) \quad (20)$$

式中, $W_{r,1}$ 、 $W_{r,2}$ 用于三元组中头实体和尾实体的投影操作, $g(x)$ 为 \tanh 函数。实验表明使用 $L1$ 距离效果更好。

其次采用单层神经网络的非线性操作来刻画实体与关系之间的语义联系。该方法中, 对负例三元组的抽样策略进行改进, 以不同的概率替换头实体和尾实体, 定义如下参数:

$$p = \frac{tph}{tph + hpt} \quad (21)$$

采用参数 p 的伯努利分布抽样策略, 在式(21)中, tph 指每个头实体对应的尾实体的平均数量, hpt 指每个尾实体对应的头实体的平均数量。并选择最相近的实体进行替换, 实体和关系之间的语义相似度通常使用向量之间的相似度表示。定义实体的相似度为:

$$dis(h, h') = \sqrt{\sum_{i=1}^k (h_i - h'_i)^2} \quad (22)$$

式中, h 和 h' 分别表示正例三元组和负例三元组中的头实体。

实验表明, 该模型性能比 TransH^[19] 提高 3.4%。缺陷是没有将关系路径考虑在内, 在今后研究中可将关系路径考虑在内, 以达到更好的表示效果。

在知识图谱的一个三元组 (h, r, t) 中, 一个关系可能与该三元组中多个实体存在潜在联系, 即产生多重关系语义问题。为解决该问题, 研究者提出一种模型即 TransG 模型。该模型的核心算法是使用高斯混合模型来刻画三元组中头实体 h 和尾实体 t 的联系, 并结合贝叶斯非参数无限混合嵌入模型。关系 r 代表的每种语义用高斯分布来描述, 由此形成多个高斯分布, 从而区分出正确和错误三元组。

下面从几何角度解释该思想: 对于给定的三元组 (h, r, t) , TransE 模型希望 $h + r \approx t$, 即头向量加关系向量约等于尾向量, 而 TransG 模型将该几何关系推广为:

$$\begin{cases} m_{(h,r,t)}^* = \arg \max_{m=1,2,\dots,M_r} \left(\pi_{r,m} \exp \left(-\frac{\|u_h + u_{r,m} - u_t\|_2^2}{\sigma_h^2 + \sigma_t^2} \right) \right) \\ h + u_{r,m_{(h,r,t)}}^* \approx t \end{cases} \quad (23)$$

式中, $m_{(h,r,t)}^*$ 为主要成分的指数, 由于存在指数效应, 该项影响最大; σ_h 和 σ_t 分别表示头实体和尾实体分布的方差, $\pi_{r,m}$ 是混合因子, 表示第 i 个分量的权重, M_r 是关系 r 的语义分量数。对于多数三元组, 只有一个

具有显著非零值的成分, 即 $\pi_{r,m} \exp \left(-\frac{\|u_h + u_{r,m} - u_t\|_2^2}{\sigma_h^2 + \sigma_t^2} \right)$,

由于该式呈指数衰减, 最终结果会足够小, 从而减少其他语义成分的干扰, 更好地表示多个语义关系。

相比基于 TransE 的若干模型, TransG 可根据三元组的特定语义自动选择最佳的翻译向量, 更专注于特定的语义嵌入, 避免其他不相关的语义成分带来的干扰。缺陷是不适用于多特征关联挖掘。

在处理复杂关系问题上, NTransGH 模型、STransH 模型、TransG 模型都表现出了很大的发展潜力。NTransGH 模型的核心是将神经网络与广义超平面的翻译操作相结合, 当前, 在 KGE 方法中, 将神经网络运用于链接预测和三元组分类任务中仍然属于比较热门的研究点; 同样, 将翻译机制运用到 KGE 中也属于比较热门的研究点, 这是第一次将两者结合在一起, 具有很大的研究价值, 不足之处是需要操作的参数偏多。STransH 模型的核心是分别在实体空间和关系空间建模, 将两种简单、常见的模型融合在一起, 通过特定的机制进而提升预测的性能, 仍有很大的研究空间, 不足之处是没有将关系路径考虑在内。对于 TransG 模型, 它的核心是贝叶斯非参数无限混合嵌入, 不足之处是不适用于多特征关联挖掘。

2.4 基于其他模型的方法

除上述三种方法外, 少数学者正在拓展其他方面的研究, 如关于实体的层次类型中的丰富信息、关于实体简洁的描述、关于三元组的相互联系与相互依赖、关于逻辑规则的背景信息等。这些方面的研究虽然取得突破性进展的不多, 但极大丰富了 KGE 方法的研究, 拥有广阔的研究前景。

基于翻译模型的知识表示方法在 2013 年首次被提出, 研究者们针对各种问题提出了许多解决方案。现有的大多数方法集中于对三元组结构化信息的表示, 忽略了层次类型中丰富的语义信息。文献 [55] 利用层次实体类型的优势, 提出 TKRL (type-embodied knowledge representation learning) 模型。在遵循 TransE 模型的假设下, TKRL 模型首先将头、尾实体投影到对应的类型空间, 然后通过两种类型编码器对层次结构进行建模, 最后根据得分函数优化 TKRL。该模型能够很好地利用层次类型信息, 与 TransE^[16]、TransR^[18] 相比, 性能分别提高 11.3%、6.2%。但缺陷是只将类型信息考虑到 KG 的表示学习中。在未来的研究中可将图像和文本形式的语义信息融

合到该模型中,同时融合多元知识库信息如维基百科,以引入更深层次的内容。

表示学习旨在将实体和关系投影到连续低维空间中,大多数方法专注于表示三元组中实体和关系之间的联系。实际上,在许多知识图谱中,通常都有对实体的简洁描述,而现有方法都无法很好地表示这些信息。因此,文献[56]根据实体描述的特点提出了DKRL(description-embodied knowledge representation learning)模型。算法思想是将词袋模型(continuous bag-of-words, CBOW)和卷积神经网络模型(convolutional neural networks, CNN)相结合来表示语义信息。DKRL模型对相应的三元组和实体描述的信息进行建模,并将每个三元组中的关系视为从头实体到尾实体的转换;同时,CNN考虑了文本中单词之间复杂的局部交互作用,即文本词序。特别的,有些实体并不存在,只有其描述信息,而DKRL模型可以根据实体的描述自动构建实体的表示。与TransE^[16]模型相比,性能提升了5.3%,具有良好的泛化能力和鲁棒性,可尝试应用于大规模知识图谱及Web领域。此模型的局限性是仅考虑用于表示学习的实体描述,未考虑各种关系或实体类型的文本信息。

现有的大多数方法将KB视为一组三元组,再分别学习它们的表示形式。但是,三元组是相互联系并相互依赖的。针对此问题,文献[57]提出一种图感知KGE方法,即GAK(E(graph aware knowledge embedding))模型。算法思想是首先将KB表示为有向图,利用图的结构信息来学习任何顶点或边的表示形式;其次引入邻居上下文、路径上下文和边缘上下文三种类型的图上下文进行嵌入,每种类型都从不同的角度反映知识的属性;最后提出一种注意力机制,以学习不同实体和关系的表示能力。在链路预测和三元组分类实验中,GAK(E)模型的性能优于TransE、TransH等模型。缺陷是未使用实体描述或来源于其他文本信息来构建文本上下文。

逻辑规则包含丰富的背景信息,但往往并未得到很好的研究。针对此问题,文献[58]将知识图谱和逻辑规则相结合,即KALE(knowledge and logic embedding)模型,核心是将三元组和逻辑规则在统一的框架中表示和建模。算法思想是将三元组看作原子,根据TransE模型假设对其进行建模;逻辑规则看作由原子与逻辑连接词组成的公式,采用 t 范数模糊逻辑

进行建模。通过这种方式以获得更具预测性的实体和关系嵌入,有利于知识获取和知识推理。与TransE^[16]相比,性能提升了1.4%。该方法可以获得更多的预测性实体嵌入,甚至可以在纯逻辑推理范围之外做出更好的预测。缺陷是未考虑合并其他类型的逻辑规则。

TKRL模型、DKRL模型、GAK(E)模型及KALE模型分别在处理实体的层级类型中的丰富信息、实体简洁的描述、三元组的相互联系与相互依赖、逻辑规则的问题时,都发挥了各自的优势,提高了预测的效果。但是,由于这些方面的研究属于KGE领域的难点,研究结果并不是很理想。然而,这些模型对于实体识别、关系抽取等任务仍然具有较大的作用,因此,这些方面的探索仍具有一定的研究价值。

3 知识图谱嵌入实验比较

本章介绍了上述模型所使用的数据集,介绍了实验中常用的评价指标,对上述模型的算法思想、模型优缺点、所用数据集及评价指标进行了总结。最后以WN18和FB15K数据集为例,展示了上述模型在链接预测上的实验结果。

3.1 常用数据集

随着人工智能和大数据技术的飞速发展,研究者已构建出多个大规模知识库,如语言知识库WordNet^[59]和世界知识库Freebase^[60]等。以上模型所用数据集都基于这两种知识库,使用其中的部分子集开展实验。为了加强对上述KGE方法性能的评估,本文收集了比较常用的实体关系数据集,总共达11个,每个数据集的实际数据见表1。

Table 1 Often-used data sets of entity relationships

表1 实体关系常用数据集

数据集	关系	实体	训练集	验证集	测试集
WN18 ^[17]	18	40 943	141 442	5 000	5 000
FB15K ^[16]	1 345	14 951	483 142	50 000	59 071
WN11 ^[51]	11	38 696	112 581	2 609	10 544
FB13 ^[52]	13	75 043	316 232	5 908	23 733
FB122 ^[58]	122	9 738	91 638	9 595	5 057
FB20K ^[56]	19 923	57 803	18 753	11 586	151
FB15K-237 ^[40]	237	14 541	272 115	17 535	20 466
YAGO3-10 ^[40]	123 182	37	—	—	—
FB15K+ ^[55]	1 855	14 951	486 446	50 000	62 374
WN18-IMG ^[59]	18	5 684	13 634	466	766
WN9-IMG ^[59]	9	6 555	11 741	1 337	1 319

3.2 评价指标

对于KGE算法实验,主要的评价指标有MeanRank、Hits@ n 以及ACC。对于链接预测中的关系预测和实体预测,常用的评价指标为MeanRank、Hits@ n ;对于三元组分类效果的预测,常用的指标为ACC。

(1) MeanRank

MeanRank表示正确实体的平均排序得分,MeanRank值越小表示排名越靠前,实体向量化结果越好,是衡量链接预测的重要指标。

(2) Hits@ n

Hits@ n 表示正确实体排在前 n 名的概率,该值越高表示效果越好。常见的有Hits@10、Hits@3和Hits@1。此外,在Hits@ n 中,将未经处理的实验设置称为“Raw”,将剔除了对实验有干扰的损坏三元组的实验设置称为“Filt”。

(3) ACC

ACC是评价三元组分类任务准确率的重要指标,值越高,表明模型在三元组分类这一任务上的效果越好。公式如下:

$$ACC = \frac{T_p + T_n}{N_{pos} + N_{neg}} \quad (24)$$

式中, T_p 表示预测正确的正例三元组的数量; T_n 表示预测正确的负例三元组数量; N_{pos} 和 N_{neg} 分别表示训练集中的正例三元组和负例三元组的数量。

3.3 方法比较

本文将KGE算法分为四类,分别对比了每一种算法的分类、名称、发表年份、数据集、评价指标、算法思想、局限性,具体的比较结果见表2。

3.4 实验结果

为了加深对基于深度学习的方法、基于图形特征的方法、基于翻译模型的方法及基于其他模型的

Table 2 KGE algorithm comparison

表2 KGE算法比较

分类	名称	发表年份	优缺点	数据集	评价指标	算法思想	整体思想
	ComplEx ^[39]	2016	优点:使用Hermitian点积,嵌入的点积可很好地缩放,可处理对称性和自反性问题 缺点:预测性能偏低,训练时间长	WN18 FB15K	MeanRank Hits@10	将潜在因子分解和复向量嵌入相结合,利用该组合处理多种二元关系	通过构建具有多隐层的机器学习模型和海量的训练数据,来提升链接预测或三元组分类的准确性
基于深度学习的方法	ConvE ^[40]	2018	优点:具有较高的参数效率,使用标签平滑防止过拟合 缺点:嵌入之间的交互次数偏低,模型卷积深度偏浅	WN18 FB15K YAGO3-10 WN18RR FB15K-237	MeanRank Hits@10 Hits@3 Hits@1	使用2D卷积来预测知识图中的缺失链接,通过卷积层和全连接层对输入实体和关系之间的相互联系建模	将ProjE模型和实体之间的相似度信息进行融合,在判断两个实体是否有关系的基础上判断具体关系类型
	LCPE ^[41]	2020	优点:参数量小并且能够更为出色地预测到实体之间的各种关系 缺点:共享变量神经网络层数偏浅	WN18 FB15K	MeanRank Hits@10	定义由邻居上下文和路径上下文组成的三重上下文,并定义得分函数来评估其相关性	充分利用图结构特征,如三元组上下文,为实体和关系提供更精准、更具解释性的语义嵌入
基于图形特征的方法	TCE ^[47]	2017	优点:充分利用KG的物理结构,尤其围绕三元组的局部结构,即三元组上下文 缺点:在基线上不如经典的翻译模型	FB15K	MeanRank Hits@10	采用联合学习策略,结合了两个模型:基于图的特征模型和基于潜在的特征模型	
	DPTransE ^[48]	2018	优点:能很好完善基于图的模型的学习过程 缺点:基于图的特征必须与三元组交互	WN11 WN18 FB13 FB15K	MeanRank Hits@10		

表2(续)

分类	名称	发表年份	优缺点	数据集	评价指标	算法思想	整体思想
	NTransGH ^[50]	2019	优点:在复杂关系的映射属性中具有很强的表达能力 缺点:需要操作的参数偏多	WN11 WN18 FB13 FB15K	MeanRank Hits@10 ACC	结合了用于将关系建模 为广大超平面的翻译操作机制,以及用于捕获实体与关系之间更复杂的神经网络	以 $h+r \approx t$ 翻译原则为基础,通过引入广义超平面
基于翻译模型的方法	StransH ^[51]	2019	优点:可以应用到大规模知识图谱补全和推理等任务上 缺点:没有将关系路径考虑在内	WN11 WN18 FB13 FB15K	MeanRank Hits@10 ACC	引入投影到特定关系超平面的机制,使得实体在不同的关系中有不同的角色	引入高斯混合模型来刻画三元组中头实体 h 和尾实体 t 的联系,并结合贝叶斯非参数无限混合嵌入模型
	TransG ^[52]	2016	优点:可自动发现关系的潜在语义,并利用关系组件的混合进行嵌入 缺点:不适用于多特征关联挖掘	WN11 WN18 FB13 FB15K	MeanRank Hits@10 ACC	使用高斯混合模型来刻画三元组中头实体 h 和尾实体 t 的联系,并结合贝叶斯非参数无限混合嵌入模型	引入低维向量空间中,有效地完成知识的表示
	TKRL ^[53]	2016	优点:能够捕获层次结构类型的信息 缺点:只将类型信息考虑到 KG 的描述学习中,而未探索图像和文本形式的丰富信息	FB15K FB15K+	MeanRank Hits@10 Hits@1 ACC	将层次结构类型视为实体的投影矩阵,并使用两种编码器来对层次结构进行建模	在该方法中,充分利
基于其他模型的方法	DKRL ^[54]	2016	优点:能自动构建 KG 中新颖实体的表示形式 缺点:仅考虑用于表示学习的实体描述,未考虑各种关系或实体类类型的文本信息	FB20K FB15K	MeanRank Hits@10 Hits@1	将词袋模型和卷积神经网络模型相结合来表示实体描述的语义信息	用 KG 的附加属性,如层次结构信息、实体信息
	GAKE ^[57]	2016	优点:可以灵活地处理除图形上下文以外的其他信息 缺点:未使用实体描述或来自其他来源的文本信息来构建文本上下文	FB15K	MeanRank Hits@10 ACC	将 KB 表示为有向图,并利用图的结构信息来学习任何顶点或边的表示形式	描述信息、逻辑背景等进行建模,丰富了研究内容
	KALE ^[58]	2016	优点:不仅与三元组兼容,而且与规则兼容,对于知识的获取和推理,更具预测性 缺点:未考虑合并其他类型的逻辑规则	WN18 FB122	MeanRank Hits@10 Hits@5 Hits@3 ACC	在统一框架中表示和建模三元组和逻辑规则	

方法算法的理解,对比了每种算法在 FB15K 数据集上的链接预测实验,具体结果见表 3。

从表 3 可以看出,在基于深度学习的方法中,LCPE 模型 MeanRank(Filt) 比 ComplEx 和 ConvE 分别低 14.5、11.0, 区分度明显; ComplEx 的 Hits@10 最高, 为 84.0%, 但与后两种算法结果差距不大, 总体上, LCPE 模型的实验效果相对较好。在基于图形特征的方法中, TCE 模型 MeanRank 比 DPTransE 低 99.0, 而两者 Hits@10 相差不大, 综合来讲, TCE 模型略胜

一筹。在基于翻译模型的方法中, NTransGH 模型的 MeanRank 最低且 Hits@10 最高, 故 NTransGH 模型的效果最好。

从表 4 可得, 以 WN18 子集做链接预测时, 在基于深度学习的方法中, LCPE 模型的 MeanRank 最低, ComplEx 的 Hits@10 最高, 综合评定, LCPE 模型相对较好。在基于图形特征的方法中, TCE 的 MeanRank 最低, DPTransE 的 Hits@10 最高, 两者的实验效果平分秋色。在基于翻译模型的方法中, NTransGH 模型

Table 3 Experimental results of link prediction on FB15K
表3 在FB15K上链接预测实验结果

Method	MeanRank(名次)		Hits@10/%	
	Raw	Filt	Raw	Filt
ComplEx ^[39]	24.2	69.2	—	84.0
ConvE ^[40]	—	65.7	—	83.1
LCPE ^[41]	176.4	54.7	52.1	82.3
TCE ^[47]	110.0	25.0	55.3	83.1
DPTTransE ^[48]	209.0	77.0	57.8	84.1
NTransGH ^[50]	139.0	51.0	61.7	85.6
StransH ^[51]	196.0	68.0	46.6	69.5
TransG ^[52]	203.0	98.0	52.8	79.8
TKRL ^[55]	184.0	68.0	49.2	69.4
DKRL ^[56]	181.0	91.0	49.6	67.4
GAKE ^[57]	228.0	119.0	44.5	64.8
KALE ^[58]	—	—	—	—

Table 4 Experimental results of link prediction on WN18
表4 在WN18上链接预测实验结果

Method	MeanRank(名次)		Hits@10/%	
	Raw	Filt	Raw	Filt
ComplEx ^[39]	—	—	—	94.7
ConvE ^[40]	374.0	—	—	95.6
LCPE ^[41]	234.2	216.6	78.9	95.2
TCE ^[47]	110.0	25.0	55.3	83.1
DPTTransE ^[48]	233.0	193.0	80.1	90.3
NTransGH ^[50]	165.0	150.0	87.1	97.8
StransH ^[51]	347.0	330.0	77.1	90.6
TransG ^[52]	483.0	470.0	81.4	93.3
TKRL ^[55]	—	—	—	—
DKRL ^[56]	—	—	—	—
GAKE ^[57]	—	—	—	—
KALE ^[58]	—	—	—	82.1

MeanRank 最低并且 Hits@10 最高,因此,NTransGH 的实验效果最好。在今后的研究中,可尝试将这几种模型进行融合,以得到更好的实际应用。

4 研究与展望

本文对现阶段 KGE 的研究现状进行了综述,对目前 KGE 领域已有的研究成果总结归纳。本章讨论了目前研究的难点问题,并对知识图谱嵌入技术的未来发展进行了展望。

从知识表示概念首次被提出到现在,基于距离模型、能量模型、单层神经网络模型、双线性模型、张量神经网络模型、矩阵分解模型及翻译模型等几类模型的研究一直在持续,尤其 2013 年 Bordes 等人提出 TransE 模型后,针对 TransE 模型存在的局限性,学者们从不同角度改进该模型,但仍未得到广泛的应

用。目前该领域仍存在以下研究难点:

(1) 表示空间的选择。表示空间在编码实体的语义信息并获取相关属性等方面起着关键作用。目前研究人员在 Point-wise 空间、复向量空间、高斯分布和流形空间都取得了研究进展。在建模时,如何设计一个合适的表示空间,以提高实体和关系的表示能力并降低计算的复杂性,仍是今后研究难点。

(2) 少样本的不确定性。KG 遵循长尾分布,尤其在领域知识图谱中,许多实体只有少量的三元组描述。在链接预测实验中往往需要足够的样本进行训练,显然少量的三元组无法完成有效的模型训练,在下游任务中难以达到预期效果。此外这些小样本存在不确定性,实体和关系之间的语义非常模糊,导致无法准确学习到实体和关系之间的语义表示。如何解决少样本的不确定性问题亟待研究。

随着 KGE 技术的不断发展,越来越多的研究者开始关注此领域,未来的发展趋势如下:

(1) 基于图卷积神经网络的嵌入

随着对知识图谱研究的不断深入,KG 得到了广泛的应用,包括智能问答和信息检索。研究者对知识库和知识图谱的创建及维护上投入了大量的精力,但即使最大的知识库也存在知识缺失问题。针对该问题,文献[61]首次提出将图卷积神经网络(graph convolutional network, GCN)应用于建模知识图谱中多关系数据。过程如图 5 所示。

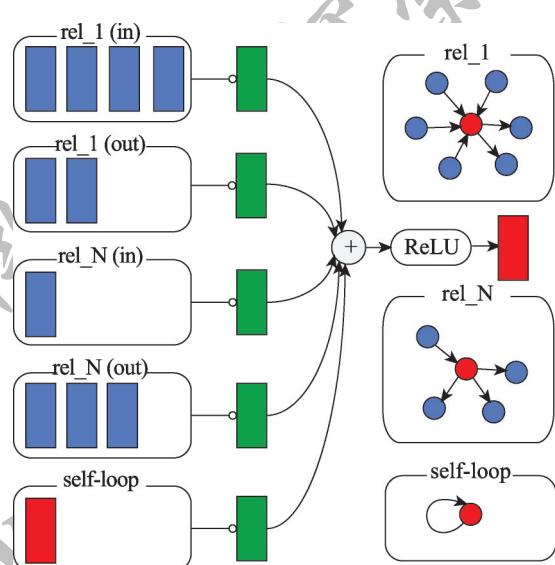


Fig.5 Entity update graph based on GCN

图5 基于GCN的实体更新图

具体的,在对每个实体进行表示时,和当前实体有直接关联的所有关系分别用GCN进行聚合,聚合后的每个结果通过激活函数生成新的实体。该方法引入了参数共享和加强稀疏约束机制,对于KG中的其他实体,可通过共享参数实现新实体的形成。

实验表明,该方法在链接预测方面比同类型方法提高29.8%,证明了该方法的有效性,可用于大规模多关系知识图谱的建模。该方法可推广到其他关系因子分解模型,有效地应用于关系提取和知识推理等工作。在未来的研究中,可将图形自动编码器模型和其他因子分解模型相结合,用一种依赖数据的注意力机制来取代该模型在邻居节点和关系类型上的聚合操作。基于GCN的表示学习研究方兴未艾,是一个重要的研究方向。

(2) 基于复杂上下文的嵌入

上下文是指在图谱中包含目标实体的相邻路径。路径(美国—_{公民}→Taylor-Swift—_说→英语)和(美国—_{公民}→Biden—_说→英语)表明这两个实体在国籍和语言方面是相似的,从向量表示角度来看,它们的嵌入向量应该也是相似度。

通过图上下文进行知识嵌入时有以下挑战:①在KG中,由于三元组都有独特的属性结构,会有许多不同类型的图上下文。如何设计一种框架来处理不同类型的图上下文是未来工作的一个挑战。②在同种类型的图上下文中,不同实体具有不同的表示能力。可考虑将实体描述信息和本地邻居上下文相结合形成“复合邻居”,利用一种新的图存储网络从复合邻居中提取潜在语义信息,设计一种更有效的邻居选择机制,避免选择过程中的遗漏;同时寻求其他类型的编码器,减少实验过程中的参数和内存开销,更好地解决KG的稀疏性问题,具有重要的意义。

(3) 基于时间模型的嵌入

KGE是一种预测KG缺失环节的有效方法,旨在学习实体和关系的分布式表示形式。现有的知识图嵌入模型主要考虑头、尾实体所在的空间具有相同的属性。但头、尾实体可以是不同类型的对象,不应该位于具有相同属性的向量空间中。文献[62]提供了一种新的研究思路。算法思想是引入时域和频域两个概念,将头实体表示为时域空间的一个点,尾实体表示为频域空间的一个点,将实体之间的转换和实体投影到时域空间的对角投影矩阵定义为一个组

合关系。

该模型性能比ComplEx提高约1.4个百分点,在链接预测任务上取得一定成效。此外,还可考虑不同关系模式下实体嵌入的多样性分布问题,对所有关系模式如对称/反对称、反演和合成的关系进行建模,也是一个具有研究价值的方向。在后续的研究中,一方面可利用时间捕捉实体之间的差异,另一方面可以充分利用KG的多模态信息(如文本、图像等),将其与KG的网络结构和时间这三部分融合,以获得更好的实验效果。

(4) 基于空间投影和复杂关系路径的嵌入

传统的KGE模型在进行知识表示时往往仅关注头、尾实体对之间直接的关系,而忽略掉了KG中可能包含的关系路径。PTransE模型提供了一种简单关系路径推理模式,如三元组(谷歌,是,人工智能公司)和三元组(人工智能公司,属于,高科技公司)具有潜在的关系路径,即(谷歌,是,高科技公司)。通过这种推理模式,提高了知识表示的有效性。

而在知识图谱中还有许多复杂的关系路径,如三元组(姚明,出生于,上海)和三元组(姚明,是,NBA球星)之间存在着复杂的推理关系,解决这种复杂的关系路径,一种比较新颖的思路是对关系进行建模,并且在空间进行投影,考虑结合关系的语义信息,探索一阶逻辑的分布式表示进行知识的嵌入,也是一个重要的研究方向^[63]。

参考文献:

- [1] AMIT S. Introducing the knowledge graph[R]. America: Official Blog of Google, 2012.
- [2] WIKIPEDIA. Knowledge graph[EB/OL]. [2021-03-21]. https://en.wikipedia.org/wiki/Knowledge_Graph.
- [3] LI M D, SUN Z Y, ZHANG S H, et al. Enhancing knowledge graph embedding with relational constraints[J]. Neurocomputing, 2020, 429: 33-40.
- [4] LI Z F, LIU H, ZHANG Z L, et al. Recalibration convolutional networks for learning interaction knowledge graph embedding[J]. Neurocomputing, 2021, 427: 118-130.
- [5] GONG F, WANG M, WANG H F, et al. SMR: medical knowledge graph embedding for safe medicine recommendation [J]. Big Data Research, 2021, 23: 1-8.
- [6] LU G M, ZHANG L Z, JIN M J, et al. Entity alignment via knowledge embedding and type matching constraints for knowledge graph inference[J]. Journal of Ambient Intelligence and Humanized Computing, 2021. DOI:10.1007/s12652-

- 020-02821-2.
- [7] WU Y L, ZHAO S L. Community answer generation based on knowledge graph[J]. *Information Sciences*, 2021, 545: 132-152.
- [8] WANG J B, KUAN N, CHEN X Y, et al. SUKE: embedding model for prediction in uncertain knowledge graph[J]. *IEEE Access*, 2021, 9: 3871-3879.
- [9] ZWIERZYNA M, FINAN C, DAVIES M, et al. A machine learning side effect prediction method using a comprehensive knowledge graph and network embeddings[C]//Proceedings of the British Pharmacology Society Meeting 2018, London, Dec 18-20, 2018. Hoboken: John Wiley & Sons, 2019: 3079.
- [10] KIM K, HUR Y, KIM G, et al. GREG: a global level relation extraction with knowledge graph embedding[J]. *Applied Sciences*, 2020, 10(3): 1-12.
- [11] CHAZARA P, NEGNY S, MONTASTRUC L. Flexible knowledge representation and new similarity measure: application on case based reasoning for waste treatment[J]. *Expert Systems with Applications*, 2016, 58: 143-154.
- [12] XIE Q Z, MA X Z, DAI Z H, et al. An interpretable knowledge transfer model for knowledge base completion[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Jul 30-Aug 4, 2017. Stroudsburg: ACL, 2017: 950-962.
- [13] SHI B X, WENINGER T. ProjE: embedding projection for knowledge graph completion[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, Feb 4-9, 2017. Menlo Park: AAAI, 2017: 1236-1242.
- [14] JIANG T S, LIU T Y, GE T, et al. Encoding temporal information for time-aware link prediction[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Nov 1-5, 2016. Stroudsburg: ACL, 2016: 2350-2354.
- [15] NGUYEN D Q, NGUYEN T D, NGUYEN D Q, et al. A novel embedding model for knowledge base completion based on convolutional neural network[C]//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. Stroudsburg: ACL, 2018: 327-333.
- [16] BORDES A, USUNIER N, GARCÍA-DURÁN A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Dec 5-8, 2013. Red Hook: Curran Associates, 2013: 2787-2795.
- [17] XIAO H, HUANG M L, YU H, et al. TransA: an adaptive approach for knowledge graph embedding[J]. arXiv:1509.05490, 2015.
- [18] 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.
- [19] 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, Québec City, Jul 27-31, 2014. Menlo Park: AAAI, 2014: 1112-1119.
- [20] FAN M, ZHOU Q, CHANG E, et al. Transition-based knowledge graph embedding with relational mapping properties [C]//Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, Phuket, Dec 12-14, 2014. Stroudsburg: ACL, 2014: 328-337.
- [21] HE S Z, LIU K, JI G L, et al. Learning to represent knowledge graphs with Gaussian embedding[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management, Melbourne, Oct 19-23, 2015. New York: ACM, 2015: 623-632.
- [22] BORDES A, WESTON J, COLLOBERT R, et al. Learning structured embeddings of knowledge bases[C]//Proceedings of the 25th AAAI Conference on Artificial Intelligence, San Francisco, Aug 7-11, 2011. Menlo Park: AAAI, 2011: 301-306.
- [23] NICKEL M, TRESP V, KRIEGEL H. A three-way model for collective learning on multi-relational data[C]//Proceedings of the 28th International Conference on Machine Learning, Washington, Jun 28-Jul 2, 2011. New York: ACM, 2011: 809-816.
- [24] BORDES A, GLOROT X, WESTON J, et al. A semantic matching energy function for learning with multi-relational data: application to word-sense disambiguation[J]. *Machine Learning*, 2014, 94(2): 233-259.
- [25] SOCHER R, CHEN D Q, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Dec 5-8, 2013. Red Hook: Curran Associates, 2013: 926-934.
- [26] XIAO H, CHEN Y D, SHI X D. Knowledge graph embedding based on multi-view clustering framework[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2021, 33(2): 585-596.
- [27] HUANG X, ZHANG J Y, LI D C, et al. Knowledge graph embedding based question answering[C]//Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Melbourne, Feb 11-15, 2019. New York: ACM, 2019: 105-113.
- [28] KANOJIA V, MAEDA H, TOGASHI R, et al. Enhancing knowledge graph embedding with probabilistic negative sampling

- [C]//Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Apr 3-7, 2017. New York: ACM, 2017: 801-802.
- [29] ZHANG W. Knowledge graph embedding with diversity of structures[C]//Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Apr 3-7, 2017. New York: ACM, 2017: 747-753.
- [30] XIA X Q, ZHANG D H, LIU Q, et al. Synergistic union of word embedding and knowledge graph for words semantic similarity measure[C]//Proceedings of the 4th International Conference on Computer and Communications, Chengdu, Dec 7-10, 2018. Piscataway: IEEE, 2018: 2349-2353.
- [31] TANG X L, YUAN R, LI Q Y, et al. Timespan-aware dynamic knowledge graph embedding by incorporating temporal evolution[J]. IEEE Access, 2020, 8: 6849-6860.
- [32] HAN X, ZHANG C H, SUN T T, et al. A triple-branch neural network for knowledge graph embedding[J]. IEEE Access, 2018, 6: 76606-76615.
- [33] PEI S C, YU L, HOEHNDORF R, et al. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference[C]//Proceedings of the 2019 World Wide Web Conference, San Francisco, May 13-17, 2019. New York: ACM, 2019: 3130-3136.
- [34] ZHANG Y, CAO W H, LIU J T. A novel negative sample generating method for knowledge graph embedding[C]//Proceedings of the 2019 International Conference on Embedded Wireless Systems and Networks, Beijing, Feb 25-27, 2019. New York: ACM, 2019: 401-406.
- [35] LEE G, KANG S, WHANG J J. Hyperlink classification via structured graph embedding[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, Jul 21-25, 2019. New York: ACM, 2019: 1017-1020.
- [36] ZHANG D H, HE Y C, WANG Y P, et al. Knowledge graph embedding based on multi-information fusion[C]//Proceedings of the 2019 IEEE International Conference on Computer Science and Educational Informatization, Kunming, Aug 16-19, 2019. Piscataway: IEEE, 2019: 310-314.
- [37] SHAN Y C, BU C Y, LIU X J, et al. Confidence-aware negative sampling method for noisy knowledge graph embedding[C]//Proceedings of the 2018 IEEE International Conference on Big Knowledge, Singapore, Nov 17-18, 2018. Washington: IEEE Computer Society, 2018: 33-40.
- [38] CHEN Q K, CHEN K, WU S, et al. Research about knowledge graph completion based on active learning[J]. Journal of Frontiers of Computer Science and Technology, 2020, 14(5): 769-782.
- 陈钦况, 陈珂, 伍赛, 等. 关于主动学习下的知识图谱补全研究[J]. 计算机科学与探索, 2020, 14(5): 769-782.
- [39] TROUILLO T, WELEB L, 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.
- [40] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D knowledge graph embeddings[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: 1811-1818.
- [41] WANG Z H, SHAO M G, LIU G J, et al. Knowledge map completion algorithm based on similar entity relationship[J]. Computer Application, 2018. DOI:10.11772/j.issn.1001-9081.2018041238.
- 王子涵, 邵明光, 刘国军, 等. 基于同类实体关系的知识图谱补全算法[J]. 计算机应用, 2018. DOI:10.11772/j.issn.1001-9081.2018041238.
- [42] REN L J, LU J, GUO W. Multi-source knowledge embedding research of knowledge graph[C]//Proceedings of the 3rd International Conference on Circuits, Systems and Devices, Chengdu, Aug 23-25, 2019. Piscataway: IEEE, 2019: 163-166.
- [43] WANG R J, WANG M, LIU J, et al. Graph embedding based query construction over knowledge graphs[C]//Proceedings of the 2018 IEEE International Conference on Big Knowledge, Singapore, Nov 17-18, 2018. Washington: IEEE Computer Society, 2018: 1-8.
- [44] YANG H, XIE G G, QIN Y, et al. Domain specific NMT based on knowledge graph embedding and attention[C]//Proceedings of the 21st International Conference on Advanced Communication Technology, PyeongChang, Feb 17-20, 2019. Piscataway: IEEE, 2019: 516-521.
- [45] NIU G L, LI B, ZHANG Y F, et al. AutoETER: automated entity type representation with relation-aware attention for knowledge graph embedding[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020: 1172-1181.
- [46] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
- [47] SHI J, GAO H, QI G L, et al. Knowledge graph embedding with triple context[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, Nov 6-10, 2017. New York: ACM, 2017: 2299-2302.
- [48] ZHANG M Y, WANG Q, XU W K, et al. Discriminative path-based knowledge graph embedding for precise link

- prediction[C]//LNCS 10772: Proceedings of the 40th European Conference on Information Retrieval, Grenoble, Mar 26-29, 2018. Cham: Springer, 2018: 276-288.
- [49] DU W Q, LI B C, WANG R. Representation learning of knowledge graph integrating entity description and entity type[J]. Journal of Chinese Information Processing, 2020, 34(7): 50-59.
杜文倩, 李弼程, 王瑞. 融合实体描述及类型的知识图谱表示学习方法[J]. 中文信息学报, 2020, 34(7): 50-59.
- [50] ZHU Q N, ZHOU X F, ZHANG P, et al. A neural translating general hyperplane for knowledge graph embedding[J]. Journal of Computational Science, 2019, 30: 108-117.
- [51] CHEN X J, XIANG Y. STransH: a revised translation-based model for knowledge representation[J]. Computer Science, 2019, 46(9): 184-189.
陈晓军, 向阳. STransH: 一种改进的基于翻译模型的知识表示模型[J]. 计算机科学, 2019, 46(9): 184-189.
- [52] XIAO H, HUANG M L, ZHU X Y. TransG: a generative model for knowledge graph embedding[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Aug 7-12, 2016. Stroudsburg: ACL, 2016: 2316-2325.
- [53] WANG R, LI B C, HU S W, et al. Knowledge graph embedding via graph attenuated attention networks[J]. IEEE Access, 2020, 8: 5212-5224.
- [54] GUO S, WNAG Q, WANG B, et al. SSE: semantically smooth embedding for knowledge graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(4): 884-897.
- [55] XIE R B, LIU Z Y, SUN M S. Representation learning of knowledge graphs with hierarchical types[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, Jul 9-15, 2016. Menlo Park: AAAI, 2016: 2965-2971.
- [56] XIE R B, LIU Z Y, JIA J, et al. Representation learning of knowledge graphs with entity descriptions[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, Feb 12-17, 2016. Menlo Park: AAAI, 2016: 2659-2665.
- [57] FENG J, HUANG M L, YANG Y, et al. GAKE: graph aware knowledge embedding[C]//Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Dec 11-16, 2016. Stroudsburg: ACL, 2016: 641-651.
- [58] GUO S, WANG Q, WANG L H, et al. Jointly embedding knowledge graphs and logical rules[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Nov 1-4, 2016. Stroudsburg: ACL, 2016: 192-202.
- [59] MILLER G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.
- [60] 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, Vancouver, Jun 9-12, 2008. New York: ACM, 2008: 1247-1250.
- [61] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//LNCS 10843: Proceedings of the 15th International Conference on Extended Semantic Web Conference, Heraklion, Jun 3-7, 2018. Cham: Springer, 2018: 593-607.
- [62] ZHANG Q J, WANG R G, JUAN Y, et al. Knowledge graph embedding by translating in time domain space for link prediction[J]. Knowledge-Based Systems, 2021, 212: 106564.
- [63] LIU Z Y, SUN M S, LIN Y K, et al. Knowledge representation learning: a review[J]. Journal of Computer Research and Development, 2016, 53(2): 247-261.
刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247-261.



舒世泰(1998—),男,硕士研究生,CCF学生会员,主要研究方向为知识图谱、知识表示。

SHU Shitai, born in 1998, M.S. candidate, student member of CCF. His research interests include knowledge graph and knowledge representation.



李松(1977—),男,博士,教授,CCF会员,主要研究方向为时空数据库、数据挖掘、数据信息安全。

LI Song, born in 1977, Ph.D., professor, member of CCF. His research interests include spatio-temporal database, data mining and data information security.



郝晓红(1969—),女,硕士,高级实验师,主要研究方向为时空数据库。

HAO Xiaohong, born in 1969, M.S., senior experimentalist. Her research interest is spatio-temporal database.



张丽平(1976—),女,硕士,副教授,CCF会员,主要研究方向为时空数据库、数据信息安全。

ZHANG Liping, born in 1976, M.S., associate professor, member of CCF. Her research interests include spatio-temporal database and data information security.