面向大规模网络安全知识图谱的快速表示学习模型

韩忠明[1,2],熊峙冰[3],陈福宇[3],杨伟杰[4*],张珣[2,3]

1. 北京工商大学国际经管学院,北京 100048
2. 食品安全大数据技术北京市重点实验室,北京 100048
3. 北京工商大学计算机学院,北京 100048
4. 北京工商大学人工智能学院,北京 100048

摘要　针对大规模网络安全知识图谱表示学习训练速度慢、对头尾实体的关系表达缺乏的问题,提出一种基于随机游走的快速训练模型。该模型首先通过关系路径下的随机游走对整体知识图谱的实体进行初步训练表示;设计了主宾语嵌入,联合关系特定主语嵌入与关系特定宾语嵌入,学习知识图谱中关系的语法含义;再次通过关系路径下的随机游走辅助知识图谱的快速训练。在多个数据集上进行了大量实验,并与多个现有模型进行对比,结果表明,提出的模型能够缩短1/3的训练时间,提升约3%的表示效果,在加快知识图谱表示学习训练速度的同时,有效改善了表示学习的效果。

关键词　知识图谱;知识图谱嵌入;表示学习;随机游走

网络安全情报分析是网络安全的重要基础,知识图谱则为网络安全情报分析提供了强大的手段,学术界和工业界,如Palantir、MITRE等,都对网络安全领域知识图谱进行了深入研究。然而,网络安全领域知识图谱具有规模大、关系复杂多样等特点,这对知识图谱的快速训练提出了很大挑战。

现有的典型知识图谱表示方法中,不使用神经网络的方法,如TransE、ComplEx、DistMult、TransR、RESCAL、RotatE等,都需要大量时间的训练。而基于神经网络的编码模型,在构建方法上从实体和关系的分布式表示出发,利用复杂的神经网络结构,如张量网络、图卷积网络和变换器等,可以学习到更丰富的表示方法。这些深度模型虽然取得了有竞争力的效果,但对算力和空间的要求非常大,无法满足实际应用中大规模知识图谱的快速训练要求。现有的分布式训练方法主要基于数据并行实

收稿日期:2023-02-23;修回日期:2023-05-12
基金项目:国家重点研发计划项目(2019YFC0507800);北京市自然科学基金项目(4172016)
作者简介:韩忠明,教授,研究方向为互联网数据挖掘,电子信箱:hanzm@th.btbu.edu.cn;杨伟杰(通信作者),副教授,研究方向为大数据分析与信息检索,电子信箱:ywj2123@126.com
引用格式:韩忠明, 熊峙冰, 陈福宇, 等. 面向大规模网络安全知识图谱的快速表示学习模型[J]. 科技导报, 2023, 41(13): 23-31; doi: 10.3981/j.issn.1000-7857.2023.13.003

现,通过增加GPU数量或者提升GPU内存大小的方法加速训练,这些方法都要求使用者提供相应的算力完成训练。而针对单个计算设备,知识图谱嵌入训练加速问题的研究仍处于初步阶段,有部分研究者采用基于多线程的并行优化提升训练速度,但这些方法的加速仍然依赖于机器的基础性能,受算力的限制,并不能从根本上提升表示学习的速度。

另外,知识图谱中的实体表达不仅与其本身语义有关,也与它所在的三元组有关。三元组中的关系可以为三元组的头尾实体附加一层隐藏语义,如对网络安全知识图谱的关系"攻击"而言,其所有尾实体都存在一个隐藏的共同属性"主机"。对于某个关系的所有头实体或尾实体而言,也应当有部分相似的语义,而现有的方法均没有对这种关系隐含的共同特征进行表达,造成实体嵌入信息的缺失,导致下游任务效果降低。有部分研究者基于实体本身的分类特征为实体嵌入增加属性,但这些方法需要知识图谱本身提供实体的分类信息,无法对没有分类信息的知识图谱的自适应进行表达。基于此,提出了一种基于随机游走的知识图谱快速学习方法,通过关系路径下的随机游走加快表示学习收敛速度,同时引入主宾嵌入的关系编码,以解决大规模知识图谱的训练效率问题,同时提升表示学习的效果。

1 相关工作

根据知识图谱表示学习模型的得分函数的类型,知识图谱嵌入方法被分为以下3类。

基于翻译的模型,认为对于一个三元组(h, $\boldsymbol{r}$, t),关系$\boldsymbol{r}$可以当作是从头实体h到尾实体t的一个翻译(translation)操作。TransE[1]是最具代表性的转化距离模型,在TransE中,它将实体和关系都表示为同一空间中的向量,给定一个fact(h, $\boldsymbol{r}$, t),关系被解释为一个翻译向量$\boldsymbol{r}$,实体h和t通过$\boldsymbol{r}$以较低的误差连接起来。TransH[2]在TransE的基础上引入了特定于关系的超平面;TransR[3]引入了特定于关系的空间;TransD[4]和TranSparce[5]为TransR的简化;TransM[6]、ManifoldE[7]、TransF[8]和TransA[9]则放宽了对于$h+\boldsymbol{r}\approx t$过分严格的要求。除了TransE及其变体,基于高斯嵌入的方法KG2E[10]将实体和关系表示为从多维高斯分布中抽取的随机向量,TransG[11]用混合高斯来表示关系。另一类基于翻译思想的变种模型将关系建模为头尾实体之间的旋转,代表性的模型为RotatE[12],HAKE[13]模型也采用了旋转建模的方式,使用同心圆表达实体之间的层级语义信息。

基于语义匹配的模型,利用基于相似度的评分标准,通过匹配实体关系内的潜在语义信息来度量三元组的可信性。其中RESCAL[14]将实体表示为向量,每种关系表示为一个矩阵,用来代表实体之间所有潜在成分之间的相互作用。TATEC[15]建模双向交互,即一个实体与一个关系之间的交互。DistMult[16]通过将关系矩阵限制为对角矩阵来简化RESCAL。Hole[17]结合了RESCAL的表现力和DistMult的效率和简单性,用循环相关操作将实体表示组成$h\star t\in R^d$。ComplEx[18]在DistMult的基础上引入了复值嵌入。

基于神经网络的模型,将神经网络方法引入知识图谱嵌入中,挖掘知识图谱中的深层隐藏信息,提升嵌入表达效果。SME[19]使用神经网络架构进行语义匹配。神经张量网络(NTN)[20]将实体投影成输入层的向量嵌入,提出了特定于关系的向量$\boldsymbol{M}_r$用于组合实体h和t,多层感知机MLP[21]中每个实体和关系都表示为一个单独的向量,参数为所有关系共享。神经关联模型NAM[22]使用"deep"架构进行语义匹配,使用L层线性隐藏层组成深层神经网络。近年来,ConvE[23]使用2D卷积将头实体和关系重塑为2维矩阵,通过多层非线性层来学习语义信息。R-GCN[24]则首次将GCN框架引入了知识图谱的表示学习中。SACN[25]在ConvE的基础上引入了GCN,由加权卷积网络WGCN编码器和ConvE-TransE解码器组成。CompGCN[26]将知识图嵌入技术中的实体—关系组合操作与图卷积模型结合,在聚合邻域信息时引入关系信息。考虑到神经网络模型无法解释网络如何捕捉到语义信息的潜在结构,Carl等[27]基于词嵌入的最新理论将知识图谱关系分为3种类型,并将其运用到知识图谱表示学习中。

近年来越来越多中国研究人员投入到知识图谱表示学习的研究中。方阳等[28]在2018年提出了基于TransE的改进模型TransAH,引入面向特定关系的超平面模型,彭敏等[29]在TransE的基础上提出了聚合邻域信息的联合知识表示模型TransE-NA,李鑫超等[30]提出了基于改进向量投影距离的知识表示学习模型SProjE,该模型引入自适应度量方法,降低了噪声信息的影响。为了加强嵌入向量的语义表达,文洋等[31]提出了基于实体相似性的表示学习方法SimE,利用实体的结构邻域度量实体的相似性。陈恒等[32]将关系定义为超复数空间中头实体到尾实体的旋转,用于推理和建模各种关系模式,包括对称/反对称、反转和组合。

总体而言,双线性模型主要基于乘法运算,使得其参数要比基于神经网络方法使用的参数少,但会有很多性能上的局限性。而基于神经网络的编码模型,从实体和关系的分布式表示出发,利用复杂的神经结构来学习更丰富的表示方法,但也存在无法体现网络安全知识图谱中复杂的多重边问题。

2　基于随机游走的大规模知识图谱快速训练模型

为了对大规模的知识图谱进行快速训练,实现更加高效的知识图谱表示学习,将图表示学习中的随机游走思想引入本研究,提出了一种基于随机游走的大规模知识图谱表示学习快速模型RWRel,整体模型框架如图1所示。模型通过关系路径下的随机游走,融合知识图谱全局结构信息,对知识图谱中实体嵌入进行初步的训练,加速后续表示学习收敛速度。同时,为了使得知识图谱表示学习模型得到的向量能够较好地刻画关系中的主谓语义,模型联合关系特定主语嵌入与关系特定宾语嵌入,学习了知识图谱中关系的语法含义,最后,再次通过关系路径下的随机游走辅助知识图谱的快速训练。另外,模型还能与现有的多种知识图嵌入方法进行结合,如TransE、TransR和RotatE等,具有极强的可扩展性。

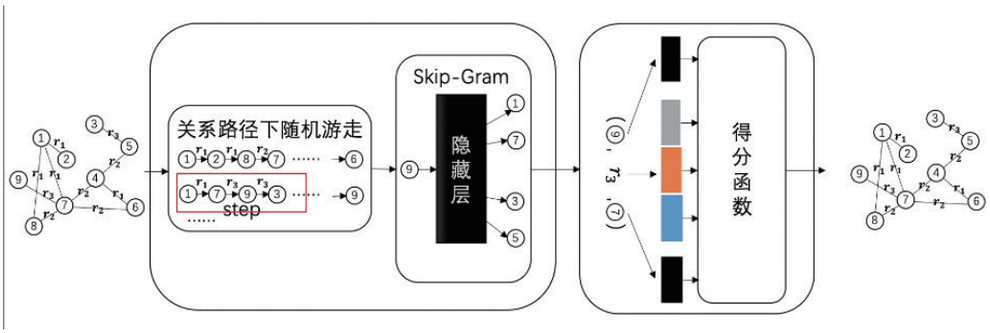


图1　RWRel模型示意

2.1　关系路径下的随机游走

为了在随机游走过程中融入知识图谱中的关系信息,针对知识图谱的多关系特点设计了关系路径下随机游走的方法,目标是使游走时搜索的目标节点在语义上与当前节点更相似。为了实现采样时的语义接近,设计了一种基于关系类型的有偏采样策略,采样时根据之前游走的关系路径在当前节点的邻居节点中选择采样节点。具体来说,对于当前节点v,若该节点为初始节点,下一步游走沿关系r的概率为

$$P\left(rp_0=r\middle|c_0=v\right)=\begin{cases}\dfrac{1}{\left|R_v\right|}, if\ r\in R_v\\ 0,\quad \text{otherwise}\end{cases}\tag{1}$$

R_v代表有节点v参与的关系类型。

通过这种处理,可以确保初始游走方向不会因为大量重复关系类型而忽视节点v周围只出现1次的稀有关系。选择完游走方向后,在该关系下继续确定下一步游走的节点,对于当前初始节点v,若选

择的游走路线是r,下一步采样节点是x的概率为

$$P\left(c_1=x\middle|c_0=v,rp_0=r\right)=\begin{cases}\dfrac{1}{\left|N_{vr}\right|},if\ (v,x)\in r\\0,\qquad otherwise\end{cases}\tag{2}$$

N_{vr}代表节点v在关系$\boldsymbol{r}$下的邻居节点。

根据初始节点选择完第一步的游走方向和具体游走节点后,接下来游走的方向与节点上一步游走的关系路径以及上一步的节点相关。即,对当前节点v,若上一步游走是经过$(t,\boldsymbol{rl},v)$路径,则下一步游走的采样节点是x的概率为

$$P\left(c_i=x\middle|c_{i-1}=v,\boldsymbol{r}p_{i-1}=\boldsymbol{rl},c_{i-2}=t\right)=\begin{cases}\dfrac{\pi_{(v,rl,x)}}{Z},if\ (v,\boldsymbol{rl},x)\in R\\0,\qquad otherwise\end{cases}\tag{3}$$

$\pi_{(v,\,rl,\,x)}$是节点v和节点x之间的未归一化转移概率,Z是归一化常数。

希望游走方向整体趋向于更相似的节点的同时,保证游走中的节点多样性,因此引入了超参数α,β,γ。通过超参数α控制关系多样性,通过β,γ控制整体游走的深度和广度。节点v和节点x之间的转移概率$\pi_{(v,\,rl,\,x)}$计算公式如下

$$\pi_{(v,rl,x)}=\begin{cases}\dfrac{\omega_{rl}(t,x)}{\alpha},if\ (v,x)\in\boldsymbol{rl}\\\omega(t,x),\qquad otherwise\end{cases}\tag{4}$$

通过α控制下一步游走时仍选择与上一步相同的关系路线的概率,当α较大时,游走倾向于选择与上一步不同的关系,当α小时,游走倾向于选择与上一步相同的关系路线。$\omega_{rl}(t,\,x)$和$\omega(t,\,x)$通过β,γ控制整体走向,具体计算公式如下

$$\omega_{rl}(t,x)=\begin{cases}\dfrac{1}{\beta},if\ t=x\\1,if\ x\in N(t)\\\dfrac{1}{\gamma},if\ x\notin N(t)\end{cases}\tag{5}$$

参数β控制反走回访问过节点的概率,当β值较高,则游走进行反走的概率就会降低,反之,反走的概率变高。然后使用参数γ控制游走进行广度搜索或是深度搜索。当$\gamma>1$,游走倾向于与上一步游走节点t接近的节点,当$\gamma<1$,游走倾向于当前节点v接近的节点。最后,为了减少游走过程中存在度为1的节点而出现的反走,我们考虑采用一种跳跃式的游走策略,使得当游走到度为1的节点时,向一定范围内其他与当前节点类似的节点上跳跃,这时需要依次反向搜索路径上节点的邻居。另外,我们认为其特定关系下的度占总度数比例越大则该节点与当前关系游走到的节点越像,在关系$\boldsymbol{r}$下游走到的v节点和搜索节点x之间的转移概率$\pi_{(v,\,rl,\,x)}$计算公式如下

$$\pi_{(v,rl,x)}=\frac{1}{d_{vx}}\cdot\frac{D^x_{rl}}{D^x}\tag{6}$$

式中,d_{vx}代表节点x和节点v之间的距离。

通过这个参数控制反向搜索的距离,回退的节点数越多,转移概率越小。式(6)中D^x_{rl},D^x则分别代表节点x关系$\boldsymbol{rl}$下的邻居数和节点x在所有关系下的邻居数,这样当节点x在关系$\boldsymbol{rl}$下的邻居数占总邻居数的比重越大,下一步游走到x节点的概率也就越大。

通过基于关系类型的有偏采样策略对知识图谱进行采样后,本研究使用Word2Vec学习实体的向量,得到知识图谱的初步实体嵌入。

2.2 引入主宾嵌入的关系编码

关系作为三元组“事件”中的“谓词”,在语法中对主语和宾语起到一个连接作用,同时作为一个“谓词”,它也能为主语和宾语提供特定的语法信息,同一个谓语词所修饰的主语应当具有一些相似特征,如谓词“出生于”,其主语通常为人物,宾语通常为地点。为了表示个“谓词”对于其头尾实体所附带的“主语”和“宾语”特征,我们对于关系p,除了对其本身设计了向量表示$\boldsymbol{p}$,同时为其“主语”和“宾语”设计了主语嵌入$\boldsymbol{p}_s$和宾语嵌入$\boldsymbol{p}_o$。如图2所示,对于三元组$f=(s,\boldsymbol{p},o)$,其头实体s在这个三元组中的嵌入$\boldsymbol{emb}_s$为实体s的初始嵌入$\boldsymbol{s}$和关系$\boldsymbol{p}$的主语嵌入$\boldsymbol{p}_s$之和,即$\boldsymbol{emb}_s=\boldsymbol{s}+\boldsymbol{p}_s$。同理,尾实体$o$在这个三元组中的嵌入$\boldsymbol{emb}_o$为实体$o$的初始嵌入$\boldsymbol{o}$和关系$\boldsymbol{p}$的宾语嵌入$\boldsymbol{p}_o$之和,即$\boldsymbol{emb}_o=\boldsymbol{o}+\boldsymbol{p}_o$。这种设置让编码可以在原有表示的基础上,体现关系中的主谓语义。而且此编码方式可以直接融入现有的表示学习方法中,如典型的TransE和RotatE等,提升表示学习的效果。

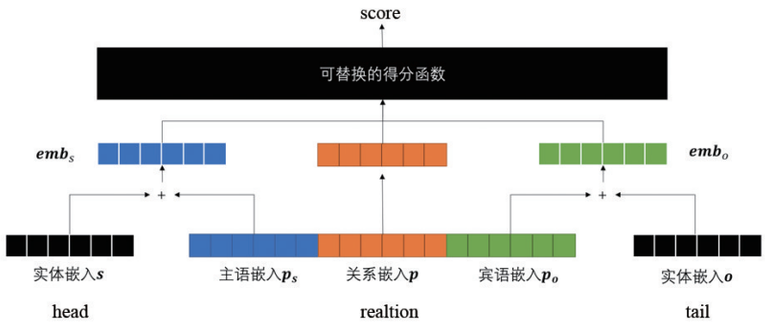


图2　引入主宾嵌入的关系编码示意

3　实验与结果分析

3.1　实验设置

为验证提出的模型在训练时间和准确度上的效果,在链接预测任务上将其与多个得分函数组合进行了对比性实验分析,使用Pytorch实现模型,并在1660TiGPU上运行。

首先使用FB15k-237和WN18RR两个数据集进行评估,为验证本模型在大规模数据集上的有效性,选择部分典型的基线模型在FB15k数据集上进行对比实验。数据集的具体统计信息如表1所示。

表1　数据集FB15k-237和WN18RR说明

数据集	FB15k-237	WN18RR	FB15k
实体数量	14541	40943	14951
关系类型数量	237	11	1345
训练集边数量	272115	86835	483142
验证集边数量	17535	3034	5000
测试集边数量	20466	3134	59071

3.2　实验结果

在链路预测实验中使用2个评价指标:平均倒数排名(MRR)和命中率(Hits@k)。

$$MRR = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{7}$$

其中,$|Q|$是三元组集合的个数,$rank_i$是指第i个三元组在链接预测中的排名,该指标越大越好。

Hits@k是指在链接预测中排名小于k的三元组的平均占比。

$$Hits@k = \frac{1}{Q}\sum_{i=1}^{|Q|}\Pi\left(rank_i \leqslant k\right) \tag{8}$$

其中,$|Q|$是三元组集合的个数,$rank_i$是指第i个三元组在链接预测中的排名,$\Pi(\cdot)$是indicator函数(若条件真则函数值为1,否则为0)。一般的取k等于1、3或者10,该指标越大越好。

分别对比了RWRel框架和各基线模型组合在两个数据集的结果,具体如表2、表3所示。

表2、表3中,第1部分为非神经网络方法,这类方法均可直接与RWRel框架结合;第2部分为神经网络方法,这类方法的耗时普遍比非神经网络方法高,但结果比非神经网络好一些。在FB15k-237数据集上,引入RWRel框架后,各非神经网络模型在各项指标上均优于原始结果,在MRR指标上使用RWRel框架的TransE、DistMult、ComplEx和RotatE平均提升了2%。其中表现最优秀的得分函数为RotatE,和这些方法原本的表现一致。在神经网络的方法上,RWRel方法略弱于最新的神经网络方

表2　FB15k-237数据集上链接预测性能对比

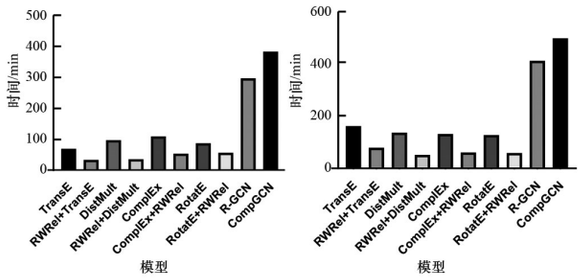
模型	MRR	Hits@1	Hits@3	Hits@10
TransE	0.292	0.192	0.325	0.478
DistMult	0.241	0.155	0.263	0.419
RotatE	0.301	0.211	0.331	0.483
ComplEx	0.248	0.149	0.283	0.423
KBGAN	0.278	—	—	0.458
R-GCN	0.244	0.149	0.382	0.413
ConvE	0.325	0.237	0.356	0.501
ConvKB	0.243	0.155	0.371	0.421
CompGCN	0.334	0.239	0.379	0.525
RWRel+TransE	0.322	0.232	0.353	0.504
RWRel+DistMult	0.284	0.194	0.316	0.463
RWRel+ComplEx	0.290	0.201	0.320	0.469
RWRel+RotatE	0.334	0.245	0.367	0.523

表3 WN18RR数据集上链接预测性能对比

模型	MRR	Hits@1	Hits@3	Hits@10
TransE	0.227	0.162	0.233	0.501
DistMult	0.43	0.39	0.44	0.49
RotatE	0.465	0.428	0.492	0.571
ComplEx	0.440	0.410	0.460	0.510
KBGAN	0.214	—	—	0.472
R-GCN	0.435	0.409	0.478	0.547
ConvE	0.430	0.400	0.440	0.520
ConvKB	0.249	0.057	0.417	0.524
CompGCN	0.479	0.443	0.494	0.546
RWRel+TransE	0.252	0.192	0.286	0.510
RWRel+DistMult	0.467	0.414	0.461	0.509
RWRel+ComplEx	0.460	0.409	0.491	0.579
RWRel+RotatE	0.485	0.438	0.499	0.583

法CompGCN,但在Hits@1上表现略好于CompGCN。在WN18RR数据集上,引入RWRel框架后各非神经网络模型均有了明显提升,其中表现最好的RotatE得分函数在RWRel框架下能获得比最新的CompGCN在论文上报告的结果高出0.6%。这表明提出的引入主宾嵌入的关系编码方法使原始的非神经网络方法能获得对标神经网络方法的效果。

为验证所提出的方法在训练时间的改善效果,将其与基线模型收敛的时间进行对比,具体结果如图3所示。

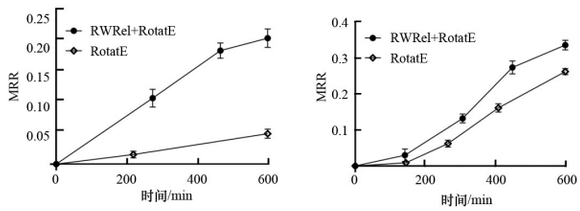


(a) FB15k-237训练时长 (b) WN18RR训练时长

图3 训练时间对比

可以看到在FB15k-237数据集上,在相同的学习率时,使用RWRel的方法均比不使用时训练时长缩短1/2。其中效果最突出的是在DistMult模型上,时间缩短了近64%,同时在准确率上提升了4%。在WN18RR数据集上表现更为优秀,使用RWRel框架后训练时长缩短接近2/3。基于神经网络的方法在2个数据集上训练普遍耗时在4 h以上,其中效果最好的CompGCN在FB15k-237数据集上训练时长达到6.5 h,在WN18RR数据集上训练时长接近8 h。而效果表现最接近CompGCN模型的RWRel+RotatE方法在2个数据集上仅需训练38 min和58 min,可见本文方法在快速学习方面的有效性。

为了分析RWRel中随机游走方法对于实体嵌入初始化表示的有效性,使用表现最好的RotatE模型,将其本身与结合RWRel框架后的2种模型在开始训练的前600 s的MRR指标变化情况进行对比,具体情况如图4所示。



(a) FB15k-237训练情况 (b) WN18RR训练情况

图4 前600 s训练情况对比

从图4可以看出,在FB15k-237数据集上,使用RotatE作为解码器的RWRel模型在训练开始的10 min内就可以获得最终收敛结果约70%的效果。而这时原始的RotatE方法仅达到最终效果的20%左右。在WN18RR数据集上,使用RotatE作为解码器的RWRel模型在10 min时表现出的性能要比原始的RotatE模型相对提升25%以上,这证明了RWRel模型中的随机游走能有效提升实体嵌入初始化表达的效果,可以加快训练速度,使模型在较短的时间内获得良好的表示向量。

最后,为验证RWRel模型中引入主宾嵌入的关系编码对表示学习效果的提升,设计了一个消融实验,分别将模型中的随机游走模块和关系编码模块移除,观察剩余模块对实验结果的影响。具体结果如表4所示。其中,RWRel+RotatE(1)代表模型中初始化嵌入为随机生成,保留了主宾嵌入,RWRel+RotatE(2)代表模型保留了关系路径下的随机游走,移除了主宾嵌入的关系编码。实验中各模型运行时间对比如图5所示。

表4　FB15k-237数据集消融实验对照

模型	MRR	Hits@1	Hits@3	Hits@10
RotatE	0.301	0.211	0.331	0.483
RWRel+RotatE(1)	0.339	0.241	0.366	0.531
RWRel+RotatE(2)	0.295	0.218	0.329	0.503
RWRel+RotatE	0.334	0.245	0.367	0.523

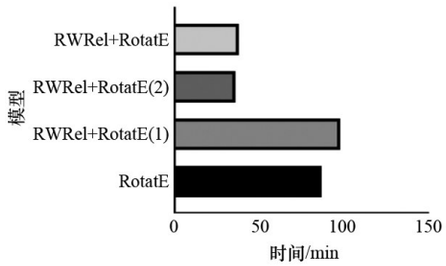


图5　FB15k-237数据集消融实验时间对比

从表4可以看出,去除引入主宾嵌入的关系编码后,RWRel框架得到的表示结果与原来得分函数RotatE的结果基本一致,证明RWRel框架在提升训练速度的同时能基本保留原始得分函数的性能。同时,仅保留关系编码模块的RWRel框架则比RotatE模型提升了超过3%的性能,证明引入主宾嵌入的关系编码能有效表达知识图谱中关系的语法信息,为RWRel框架提供良好的性能支撑。从图5中各模型的运行时间可以看出,不使用随机游走的RotatE和RWRel+RotatE(1)训练时长均超过了80 min,而使用了随机游走的RWRel+RotatE(2)和RWRel+RotatE的训练时长均在40 min以下,证明RWRel框架的训练速度提升主要来源于随机游走模块。通过消融实验可以看出,完整的RWRel框架中,关系路径下的随机游走加快表示学习方法的训练速度,引入主宾嵌入的关系编码提升了表示学习的表示性能。

4　结论

针对现有大规模知识图谱表示学习需要大量时间进行训练的现状,提出了一种基于随机游走的知识图谱快速训练模型RWRel,该模型包含了关系路径下的随机游走策略和针对关系语法含义的主宾嵌入编码,能够缩短大规模知识图谱表示学习时训练所需的时间,并有效改善知识图谱关系中蕴含的语法信息的表达。同时该方法还拥有良好的可扩展性,可直接运用于未来提出的其他得分函数上。但是,计算效率和模型表现力之间权衡,始终是大规模知识图谱的构建需要考虑的问题。随着预训练模型的发展,将预训练模型移植到图谱表示学习上,提前捕捉知识图谱中的实体关系和相关知识,可能会在下游应用中缩短时间、提升用户体验。另外,对于动态知识图谱中的快速表示学习也是未来的重要研究方向。

参考文献(References)

[1] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for embeddings for modeling multi-relational data[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems 2013. Red Hook, NY, USA: Curran Associates Inc., 2013: 2782-2795.

[2] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence 2014. Québec City, Québec, Canada: AAAI Press, 2014: 1112-1119.

[3] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence 2015. Austin, Texas: AAAI Press, 2015: 2181-2187.

[4] Ji G, He S, Xu L, 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 2015. Beijing, China: Association for Computational Linguistics, 2015: 687-696.

[5] Ji G, Liu K, He S, et al. Knowledge graph completion with adaptive sparse transfer matrix[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence 2016. Phoenix, Arizona: AAAI Press, 2016: 985-991.

[6] 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 Computing 2014. Chulalongkorn University, Phuket, Thailand: Department of Linguis-

tics, 2014: 328-337.

[7] Xiao H, Huang M, Zhu X. From one point to a manifold: Knowledge graph embedding for precise link prediction [C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence 2016. New York, USA: AAAI Press, 2016: 1315-1321.

[8] Feng J, Huang M, Wang M, et al. Knowledge graph embedding by flexible translation[C]//Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning 2016. Cape Town, South Africa: AAAI Press, 2016: 557-560.

[9] Xiao H, Huang M, Hao Y, et al. TransA: An adaptive approach for knowledge graph embedding[J]. arXiv preprint, arXiv:1509.05490, 2015.

[10] He S, Liu K, Ji G, et al. Learning to represent knowledge graphs with gaussian embedding[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management 2015. NY, USA: Association for Computing Machinery, New York, 2015: 623-632.

[11] Xiao H, Huang M, Zhu X. TransG: A generative model for knowledge graph embedding[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics 2016. Berlin, Germany: Association for Computational Linguistics, 2016: 2316-2325.

[12] Sun Z Q, Deng Z H, Nie J Y, et al. Rotate: Knowledge graph embedding by relational rotation in complex space [C]//In International Conference on Learning Representations, Ernest N 2019. New Orleans: Morial Convention Center, 2019: 1-18.

[13] Zhang Z Q, Cai J Y, Zhang Y D, et al. Learning hierarchy-aware knowledge graph embeddings for link prediction[C]//The Thirty-Fourth AAAI Conference on Artificial Intelligence 2020. New York, USA: AAAI Press, 2020: 3065-3072.

[14] 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 International Conference on Machine Learning 2011. Madison, WI, USA: Omnipress, 2011: 809-816.

[15] García-Durán A, Bordes A, Usunier N. Effective blending of two and three-way interactions for modeling multi-relational data[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer, 2014: 434-449.

[16] Yang B, Yih W, He X, et al. Embedding entities and relations for learning and inference in knowledge bases [C]//International Conference on Learning Representations 2015. San Diego, CA, USA: Conference Track Proceedings, 2015: 141-153.

[17] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence 2016. Phoenix, Arizona: AAAI Press, 2016: 1955-1961.

[18] Trouillon T, Welbl J, Riedel S, et al. Complex Embeddings for Simple Link Prediction[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48 2016. New York, USA: JMLR.org, 2016: 2071-2080.

[19] Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data[J]. Machine Learning, 2014, 94(5): 233-259.

[20] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [C]//Twenty-seventh Conference on Neural Information Processing Systems 2013. Lake Tahoe, Nevada, USA: Curran Associates, 2013: 926-934.

[21] 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 2014. New York, USA: Association for Computing Machinery, 2014: 601-610.

[22] Liu Q, Jiang H, Evdokimov A, et al. Probabilistic reasoning via deep learning: Neural association models[C]// 25th International Joint Conference on Artificial Intelligence 2016. New York, USA: Deep Learning for Artificial Intelligence, 2016: 271-278.

[23] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D knowledge graph embeddings[C]//32nd AAAI Conference on Artificial Intelligence, AAAI 2018. New Orleans, Louisiana USA: AAAI Publications, 2018: 1811-1818.

[24] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]//European Semantic Web Conference. Cham: Springer, 2018: 593-607.

[25] Shang C, Tang Y, Huang J, et al. End-to-end structure-aware convolutional networks for knowledge base completion[C]. The Thirty-Third AAAI Conference on Artificial Intelligence. Honolulu, Hawaii, USA: AAAI Press, 2019, 33: 3060-3067.

[26] Vashishth S, Sanyal S, Nitin V, et al. Composition-based multi-relational graph convolutional networks[J]. arXiv preprint, arXiv:1911.03082, 2019.
[27] Carl A, Ivana Balažević, Timothy H. Interpreting knowledge graph relation representation from word embeddings [C]//The Ninth International Conference on Learning Representations 2021. USA: Virtual Conference, 2021: 1-16.
[28] 方阳, 赵翔, 谭真, 等.一种改进的基于翻译的知识图谱表示方法[J]. 计算机研究与发展, 2018, 55(1): 139-150.
[29] 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.
[30] 李鑫超, 李培峰, 朱巧明. 一种基于改进向量投影距离的知识图谱表示方法[J]. 计算机科学, 2020, 47(4): 189-193.
[31] 文洋, 张茂元, 周礼全, 等. 基于实体相似性的知识表示学习方法[J]. 计算机应用研究, 2021, 38(4): 1008-1012.
[32] 陈恒, 王维美, 李冠宇, 等. 四元数关系旋转的知识图谱补全模型[J]. 计算机科学, 2021, 48(5): 225-231.

A fast representation learning model for large-scale cybersecurity knowledge graphs

HAN Zhongming[1,2], XIONG Zhibing[3], CHEN Fuyu[3], YANG Weijie[4*], ZHANG Xun[2,3]

1. School of Economics and Management, Beijing Technology and Business University, Beijing 100048, China
2. Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing 100048, China
3. School of Computer Science, Beijing Technology and Business University, Beijing 100048, China
4. School of Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China

Abstract This paper comes up with a fast-training model based on random walk to address the problems of slow training speed for representation learning of large-scale cybersecurity knowledge graph and lack of relational representation of head and tail entities,. The model first performs an initial training representation of the entities of the overall knowledge graph by random walk under relational paths, then, a subject-object embedding is designed to learn the syntactical meaning of the relations in the knowledge graph by combining the relation-specific subject embedding with the relation-specific object embedding. Finally, fast training of the knowledge graph is again assisted by random wandering under relational paths. In this paper, extensive experiments are conducted on several datasets and the results are compared with those using several existing models. The results show that the model proposed in this paper can shorten the training time by 1/3 and improve representation by about 3%, effectively improving the representation learning effect while speeding up the training speed of knowledge graph representation learning.

Keywords knowledge graph; knowledge graph embedding; representation learning; random walk

（责任编辑　傅雪）