Journal of Computer Applications
计算机应用,2018,38(11):3089-3093
ISSN 1001-9081
CODEN JYIIDU
2018-11-10
http://www.joca.cn

文章编号:1001-9081(2018)11-3089-05
DOI:10.11772/j.issn.1001-9081.2018041238

基于实体相似度信息的知识图谱补全算法

王子涵[1],邵明光[1],刘国军[1],郭茂祖[2],毕建东[1],刘　扬[1*]
(1.哈尔滨工业大学 计算机科学与技术学院,哈尔滨 150001;
2.北京建筑大学 建筑大数据智能处理方法研究北京市重点实验室,北京 100044)
(*通信作者电子邮箱 yliu76@hit.edu.cn)

摘　要:为了解决知识图谱的链接预测问题,提出了一种共享变量的神经网络模型(LCPE),该模型通过将实体和关系嵌入到向量空间中实现对链接的预测。通过分析 Unstructured Model,推导出在向量空间中两个有关系的实体嵌入距离更近,即相似的实体之间更可能具有关系,LCPE 模型将 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。三元组预测实验中,LCPE 模型在与 ProjE 模型参数规模相同的情况下,在公开数据集 WN18 中,正例三元组的平均得分排名(Mean Rank)比 ProjE 提前了 11,而正例三元组在前 10 名中出现的概率 Hit@10 比 ProjE 提升了 0.2 个百分点;在 FB15k 中,Mean Rank 提前了 7.5,Hits@10 平均提升了 3.05 个百分点;证明了 LCPE 模型能够将实体相似度信息融入 ProjE 中并有效提升预测准确度。

关键词:知识图谱;链接预测;嵌入向量;神经网络;相似度
中图分类号:TP301.6　　**文献标志码**:A

Knowledge graph completion algorithm based on similarity between entities

WANG Zihan[1], SHAO Mingguang[1], LIU Guojun[1], GUO Maozu[2], BI Jiandong[1], LIU Yang[1*]
(1. *College of Computer Science and Technology, Harbin Institute of Technology, Harbin Heilongjiang* 150001, *China*;
2. *Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing* 100044, *China*)

Abstract: In order to solve the link prediction problem of knowledge graph, a shared variable network model named LCPE (Local Combination Projection Embedding) was proposed, which realized the prediction of links by embedding entities and relationships into vector space. By analyzing the Unstructured Model, it was derived that the distance between related entities' embedding was shorter in the vector space, in other words, similar entities were more likely to be related. In LCPE model, ProjE model was used based on similarity between two entities to judge whether the two entities were related and the relation type between them. The experiment shows that with the same number of parameters, the LCPE improves Mean Rank by 11 and lifts Hit@10 0.2 percentage points in dataset WN18 while improves Mean Rank 7.5 and lifts Hit@10 3.05 percentage points in dataset FB15k, which proves that the similarity between entities, as auxiliary information, can improve predictive capability of the ProjE model.

Key words: knowledge graph; link prediction; embedding vector; neural network; similarity

0　引言

在知识图谱中,知识以三元组〈头实体,关系,尾实体〉的形式存储,使得知识能够得到结构化整理[1],但是知识图谱中的知识称不上完善,尤其是在大规模知识图谱中,需要不断补充实体关系进行知识图谱补全[2],而由于信息量过大,实体之间的关系多且复杂,传统的利用链接预测的方法已经不能完成大规模知识图谱的补全[3],所以对知识图谱的补全往往采用学习知识表示,并定义得分函数、采用对三元组进行打分的方法实现关系预测。这样知识图谱的补全算法就成了对三元组的得分进行排序的算法[4]。

目前,学习知识表示的代表模型有距离模型、单层神经网络模型、能量模型、双线性模型、张量神经网络模型[5]、矩阵分解模型和翻译模型等[6]。其中效果较好的张量神经网络模型[5]和基于文档特征向量空间模型[7]都是在现有模型中加入了多层神经网络来进行知识图谱的补全[4]。这些方法虽然提高了链接预测能力,却增大了参数规模,并且一次仅能考虑一种关系,导致模型的扩展性较差、效率较低,无法应用于大型知识库[8],所以,翻译模型因其简单和准确性引起了广泛的关注。Mikolov 等[9-10]通过类比实验发现,词向量空间的平移不变现象普遍存在于词汇的语义关系和句法关系中[5],该现象的发现使得翻译模型 TransE[11]被提出。该模型的主要思想是将关系作为对头实体的翻译(Translation),并根据尾实体和关系对头实体的翻译的相似度来定义得分函数。TransE 模型的参数少、计算复杂度低却能够直接建立起实体和关系之间的复杂语义联系,在大规模稀疏知识图谱上,

收稿日期:2018-04-30;**修回日期**:2018-06-07;**录用日期**:2018-06-29。
基金项目:国家自然科学基金资助项目(61671188,61571164,61502122);国家重点研发计划项目(2016YFC0901902)。
作者简介:王子涵(1995—),女,江苏徐州人,硕士研究生,主要研究方向:知识图谱;　邵明光(1995—),男,山东济宁人,硕士研究生,主要研究方向:知识图谱;　刘国军(1979—),男,天津宝坻人,副教授,博士,CCF 会员,主要研究方向:机器学习、计算机视觉、图像处理、模式识别;郭茂祖(1966—),男,山东夏津人,教授,博士,CCF 会员,主要研究方向:机器学习、生物信息学、城市计算;　毕建东(1964—),男,黑龙江齐齐哈尔人,副教授,博士,主要研究方向:机器学习;　刘扬(1976—),男,黑龙江哈尔滨人,副教授,博士,主要研究方向:机器学习。

《计算机应用》www.joca.cn

TransE 的性能更是惊人[5]。为了能够更加精确地描述实体和关系之间的语义联系，Trans 系列模型[3,12-16]都在 TransE 的基础上被提出。虽然这些模型的预测能力相较 TransE 有所提升，但是增大了参数规模，部分模型[3,14-16]还需要用 TransE 对参数进行预训练，这些都增加了计算的复杂度，使得模型扩展性变差。文献[4]提出 ProjE（Embedding Projection for Knowledge Graph）模型成功避免了这些问题，该模型不需要预训练，参数规模小且预测能力较强。然而，由于 ProjE 的关注点在于实体与关系之间的联系导致它并没有能够充分利用神经网络感知和分析实体向量所代表的语义信息。通过研究 Unstructured Model（UM）[17]，得到具有关系的实体可能集中在一些小区域内的结论。由于向量空间可以看作是实体的语义空间，也就是说比较相似的实体之间更可能具有关系，所以实体是否相似可作为是否存在关系的一个衡量标准。利用这个标准和 ProjE 模型，神经网络可以同时分析两个实体之间是否存在关系和具体存在什么关系，相当于侧重分析在实体分布集中的那些局部空间中的实体之间的具体关系，所以本文提出了一种新的算法——LCPE（Local Combination Projection Embedding），并通过在标准数据集上的实验证明 LCPE 的预测能力优于 ProjE。

综上所述，本文的主要贡献是：

1）发现了一种可以用于辅助判断实体之间是否存在关系的辅助信息——实体之间的语义相似度，而语义相似度可以利用实体向量在实体嵌入空间中的距离来判断；

2）提出了 LCPE 算法，将 ProjE 模型和实体相似度信息融合，该模型可以同时判断两个实体是否存在关系和具体存在什么关系，并通过实验验证了 LCPE 在与 ProjE 参数规模相同的情况下预测能力提升。

1 LCPE 模型

1.1 相关模型

翻译模型 TransE[11]的中心思想是将关系看作是头实体到尾实体的翻译，头实体 $\boldsymbol{h}$、关系 $\boldsymbol{r}$ 和尾实体 $\boldsymbol{t}$ 之间的关系表示如下：

$$\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t} \tag{1}$$

因此，TransE 的得分函数定义为：

$$E(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_{L1/L2} \tag{2}$$

即向量 $\boldsymbol{h}+\boldsymbol{r}$ 和 $\boldsymbol{t}$ 之间的 L1 或 L2 距离，其本质上是衡量 $\boldsymbol{h}+\boldsymbol{r}$ 和 $\boldsymbol{t}$ 之间的相似度。

通过以上介绍可知，TransE 模型的参数规模为 $n_e k+n_r k$，其中 n_e 和 n_r 分别是知识库中实体和关系的数量，k 是特征维数。

ProjE 模型[4]的基础模型是共享变量的神经网络模型，它利用组合矩阵 $\boldsymbol{D}$ 对实体和关系进行组合作为关系对头实体的翻译，组合运算如式(3)所示：

$$\boldsymbol{e}\oplus\boldsymbol{r}=\boldsymbol{D}_e\boldsymbol{e}+\boldsymbol{D}_r\boldsymbol{r}+\boldsymbol{b}_c \tag{3}$$

假设实体和关系被嵌入到 k 维向量空间中，那么 $\boldsymbol{e}\in\mathbf{R}^k$ 表示实体向量，$\boldsymbol{r}\in\mathbf{R}^k$ 表示关系向量，$\boldsymbol{D}_e$、$\boldsymbol{D}_r\in\mathbf{R}^{k\times k}$ 是组合矩阵，由于在这类问题中考虑同一实体和关系的不同的特征维度之间的相互作用关系意义并不大[5]，所以 $\boldsymbol{D}_e$、$\boldsymbol{D}_r$ 被设为对角阵，$\boldsymbol{b}_c$ 表示偏置量。

利用(3)定义的组合规则，结合关系作为头尾实体的翻译的核心思想，ProjE 提出得分函数(4)用于衡量尾实体与 $\boldsymbol{e}\oplus\boldsymbol{r}$ 之间的相似度：

$$h(\boldsymbol{e},\boldsymbol{r})=g(\boldsymbol{W}^c f(\boldsymbol{e}\oplus\boldsymbol{r})+b_p) \tag{4}$$

其中：f 和 g 是激活函数，$\boldsymbol{W}^c\in\mathbf{R}^{s\times k}$ 是候选实体集组成的矩阵，其中 s 表示候选实体的数量，b_p 表示偏置。

ProjE 模型与 TransE 模型相比，参数量只多了 $5k+1$，其中，得分函数中的偏移量、组合矩阵和组合运算中的偏移向量分别占 1、$4k$ 和 k 个参数。不仅如此，ProjE 模型不需要预训练，相较需要通过预训练提高预测精确度的模型[3,15-17]，它训练模型所需要的时间更少，模型的扩展性也得以增强。

1.2 LCPE 算法概述

由于 ProjE 模型只是简单地考虑向量 $\boldsymbol{t}$ 和向量 $\boldsymbol{h}\oplus\boldsymbol{r}$ 之间的相似度，并没有充分利用实体嵌入向量的语义信息，导致神经网络可以分析的信息不足，优势无法完全发挥。

Unstructured Model[17]是 $r=0$ 时的 TransE 模型，在 UM 中，所有实体之间的关系都视为单关系，即不考虑实体之间的具体关系类型，只考虑这实体之间是否存在关系。由于 $r=0$，结合 TransE 的得分函数式(2)可知，在 UM 中，当两个实体之间具有某种关系时，这两个实体嵌入向量之间的距离会比较小，并且两个实体之间具有的关系越多，得到的嵌入向量之间的距离就会越小。直观来说，实体是由多个属性描述的，实体在语义空间中的嵌入向量就是用于描述实体的属性值集合，实体嵌入向量之间的距离越小，就代表这两个实体越相似。综上所述，实体越相似，实体之间存在关系的可能性越大。这个结论适用于大多数事实，例如父子关系，这个关系产生于同类实体之间，并且父子之间具有很多相同或相似的属性，比如所在地、长相、家庭等，这些属性值相同或接近都缩小了父子这两个实体的嵌入向量之间的距离。

综上，两个实体是否相似可以作为实体之间是否存在关系的一个判断条件，并通过判断两个实体之间是否存在关系来加强模型的链接预测能力。由于相似的实体嵌入向量在向量空间中距离更近，所以相似的实体嵌入向量会集中在向量空间中的一些小区域中，利用两个实体是否相似作为加强预测能力的辅助信息就意味着侧重于在那些实体分布稠密的局部空间判断实体之间的具体关系类型，因此本文提出了 LCPE 模型，它将 ProjE 模型和实体相似度信息相结合，充分发挥了神经网络模型的优势，提高了模型的链接预测能力。

本文将实体之间的相似度作为辅助信息加入 ProjE 模型，定义得分函数：

$$h(\boldsymbol{e},\boldsymbol{r})_i=g(\boldsymbol{W}^c_{[i,:]}f(\boldsymbol{e}\oplus\boldsymbol{r})+f(\boldsymbol{W}^c_{[i,:]}\cdot\boldsymbol{e})+b_p) \tag{5}$$

其中：$h(\boldsymbol{e},\boldsymbol{r})_i$ 表示候选实体集中第 i 个实体的得分，f 和 g 是激活函数，$\boldsymbol{W}^c\in\mathbf{R}^{s\times k}$ 是候选实体矩阵，其中 s 表示候选实体的数量，b_p 表示偏移量。$h(\boldsymbol{e},\boldsymbol{r})$ 计算的同样也是 $\boldsymbol{W}^c$ 中的实体和组合运算 $\boldsymbol{e}\oplus\boldsymbol{r}$ 之间的相似度。$\boldsymbol{W}^c_{[i,:]}\cdot\boldsymbol{e}$ 是实体向量 $\boldsymbol{e}$ 和候选实体集中第 i 个实体的嵌入向量的内积，内积越大表示余弦距离越小，这两个实体就越相似。由于是共享变量的神经网

《计算机应用》www.joca.cn

络模型,本文算法增加的辅助项并没有增加原 *ProjE* 模型的参数数量。如表 1 所示,本文算法相较 Trans 系列模型在参数规模上仍具备优势。

表 1　各模型的参数规模对比

Tab. 1　Comparison of parameter size of each model

知识图谱补全模型	参数规模	知识图谱补全模型	参数规模
UM	$n_e k$	TransH	$n_e k + 2n_r k$
TransE	$n_e k + n_r k$	TransR	$n_e k + n_r(k + k^2)$
ProjE	$n_e k + n_r k + 5k$	TransD	$2n_e k_e + 2n_r k_r$
LCPE	$n_e k + n_r k + 5k$	TransF	$n_e k_e + n_r k_r + s k_e k_r + 2s$

注:n_e 表示实体数,k 表示对应实体向量和关系向量的参数个数,s 表示关系基空间的个数。

如图 1 所示,LCPE 模型是一个由判断两个实体之间相似度的网络和判断两个实体之间是否具有某种关系的网络共同构成的神经网络。$\boldsymbol{W}^c$ 是由候选实体向量组成的矩阵,$\boldsymbol{W}^E$ 表示实体向量构成的矩阵,$\boldsymbol{W}^R$ 表示关系向量构成的矩阵,$\boldsymbol{E}_i$ 和 $\boldsymbol{E}_j$ 分别是从 $\boldsymbol{W}^c$ 和 $\boldsymbol{W}^E$ 中提取出的一个实体向量,$\boldsymbol{R}$ 是从 $\boldsymbol{W}^R$ 中提取出的关系向量,$\boldsymbol{D}_e$ 和 $\boldsymbol{D}_r$ 分别代表组合矩阵。

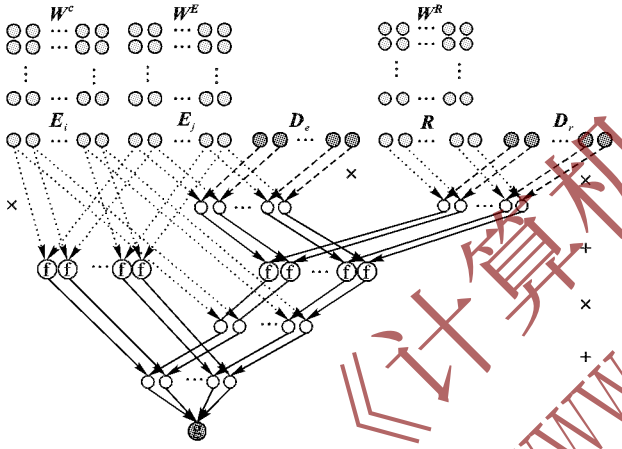

图 1　LCPE 模型结构

Fig. 1　Structure of LCPE model

1.3　模型训练

在本文算法中,候选实体集虽然因为共享实体变量而没有增加参数的数量,但是如果每次都用全部实体集进行训练,也会导致巨大的运算量,所以要使用候选抽样的方法来减小候选实体集 $\boldsymbol{W}^c$ 的规模[18-20],并且利用 Word2Vec[19] 的规则对候选集进行负例抽样效果最好[4]。具体方法是对于一个给定的实体 $\boldsymbol{e}$,其训练所使用的候选实体集由全部的正例中的实体集和一部分负例中的实体集构成,为了简单起见,利用二项分布 $B(1,Py)$ 来表示某个负例中的实体是否被选中,即 Py 表示该负例被选中的概率,而 $1-Py$ 表示未被选中的概率,实验表明最优的负例抽样概率为 25%[4]。

Trans 系列模型通常都采用了 pairwise 方法对模型进行训练,损失函数 L 定义为如式(6)形式:

$$L = \sum_{(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})\in S}\ \sum_{(\boldsymbol{h}',\boldsymbol{r}',\boldsymbol{t}')\in S'} \max\left(0,[\gamma + E(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}) - E(\boldsymbol{h}',\boldsymbol{r}',\boldsymbol{t}')]\right) \tag{6}$$

其中:E 表示三元组的得分函数;S 表示正例三元组的集合,正例三元组是在原知识库中存在的三元组;S' 表示负例三元组的集合,负例三元组是通过用知识库中其他的实体或关系替换正例三元组中的 $\boldsymbol{h}$、$\boldsymbol{r}$ 或 $\boldsymbol{t}$ 产生的在原知识库中不存在的三元组。

近年来,softmax 回归在多标签图像注释任务中取得了良好的效果,这是因为多标签图像注释以及许多其他分类任务应该综合考虑其预测出的候选集中的所有分数[5],因此,模型采用 softmax 函数以便考虑所有候选实体的分数,用 softmax 和 tanh 作为激活函数,将式(5)中的 g 和 f 分别用 softmax 函数和 tanh 函数替换可以将得分函数写成式(7)所示形式:

$$h(\boldsymbol{e},\boldsymbol{r})_i = \frac{\exp\left(\boldsymbol{W}^c_{[i,:]}\tanh(\boldsymbol{e}\oplus\boldsymbol{r}) + \tanh(\boldsymbol{W}^c_{[i,:]}e_j) + b_p\right)}{\sum_j \exp\left(\boldsymbol{W}^c_{[i,:]}\tanh(\boldsymbol{e}\oplus\boldsymbol{r}) + \tanh(\boldsymbol{W}^c_{[i,:]}e_j) + b_p\right)} \tag{7}$$

并利用 listwise 方法进行训练。LCPE 模型的损失函数定义为式(8):

$$L(\boldsymbol{e},\boldsymbol{r},\boldsymbol{y}) = -\sum_i^{|y|} \frac{1(y_i = 1)}{\sum_i 1(y_i = 1)} \mathrm{lb}\,(h(\boldsymbol{e},\boldsymbol{r})_i) \tag{8}$$

其中:$\boldsymbol{y}\in\mathbf{R}^s$ 是一个二元标签向量;s 是候选实体集的个数;$y_i = 1$ 代表第 i 个候选实体和实体 $\boldsymbol{e}$,关系 $\boldsymbol{r}$ 组成的三元组是一个正例三元组,$1(\cdot)$ 代表当括号内的等式成立时,$1(\cdot) = 1$,否则 $1(\cdot) = 0$;$h(\boldsymbol{e},\boldsymbol{r})$ 表示三元组的得分函数(式(5)),而 $h(\boldsymbol{e},\boldsymbol{r})_i$ 表示第 i 个候选实体和实体 $\boldsymbol{e}$,关系 $\boldsymbol{r}$ 组成的三元组的得分。

预测问题归根结底是一个对得分进行排序的问题,而 pairwise 的训练方法并没有很好地解决实体预测排名的问题,所以利用考虑了整体的得分排名的 listwise 方法对模型进行训练效果更好[4]。

2　实验结果与分析

本章展示了实体相似度作为辅助项的可行性以及 LCPE 与之前提到的模型的预测结果对比。关系预测实验在两个公开数据集 FB15k[21] 和 WN18 上进行,FB15k 是一个知识图谱的重要子集,WNID 是一个 WordNet[22] 的子集,ID 表示在 WordNet 中的 ID,用于在 WordNet 中唯一标记一个子集。FB15k 和 WN18 的实体、关系和三元组数如表 2 所示。

表 2　实验使用的数据集

Tab. 2　Datasets in the experiments

数据集	关系数	实体数	三元组数		
			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

2.1　参数设置

为了对这两个模型进行更加准确的对比,在实验中,本文根据 ProjE 中的参数设置,将两个模型的学习率设为 0.01,训练块大小为 200,正则化权重为 1E-5,dropout 概率为 0.5,向量空间的维度为 200,负例抽样概率为 0.25,按照均匀分布 $U[-6/\sqrt{k},6/\sqrt{k}]$ 初始化实体向量和关系向量的参数,并利用 Adam 方法[23] 作为损失函数的梯度下降算法。

2.2　实体相似度对比

为了防止嵌入向量的大小的影响,本文将计算实体之间的余弦距离用以衡量实体之间的相似度。随机提取 LCPE 中发生关系的实体对,计算它们的嵌入向量的余弦距离,同时对

《计算机应用》www.joca.cn

于上述实体,计算它们在 ProjE 中的余弦距离,最后计算二者的比值。实验在 WN18 数据集上进行。

实验结果如图 2 所示,纵坐标代表同一实体对在两个模型中的余弦距离比值(ProjE 中的实体之间的余弦距离/LCPE 中的实体之间的余弦距离)。实验结果显示距离比值大多大于 1,这说明 LCPE 中具有关系的实体的余弦距离更小,和本文预期相符,也说明了实体相似度的辅助信息可以用作增强预测能力的辅助项。

2.3 链接预测结果

链接预测是为了预测三元组中缺失的头实体和尾实体的任务,而实体的预测问题归根结底是实体的得分排序问题,所以使用 Raw Mean Rank、Filtered Mean Rank 和 Raw Hits@10、Filtered Hits@k 作为对链接预测的预测效果评估指标。Raw Mean Rank 是正确的实体在得到的得分序列中的平均排名;Raw Hits@k 是指正确的实体在得分为前 k 的元素中出现的概率,但是,有其他的正确实体排在待测的正确实体之前的情况,所以还有一种衡量标准就是将其他的已知的正确实体删去之后再排序得到 Filtered Mean Rank 和 Filtered Hits@k,它们分别代表删去了其他的正确实体后目标实体的平均排名和在前 k 个元素中出现的概率。由上述介绍可知,Raw Mean Rank 和 Filtered Mean Rnak 越低越好,Raw Hits@k 和 Filtered Hits@k 越高越好。本文将 CLPE 与 Trans 系列模型[3,11-17]及 ProjE 模型[4]进行对比。实验结果如表 3 所示。

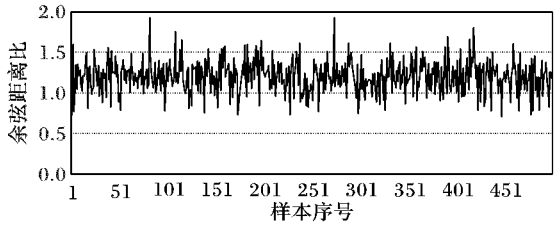

图 2 WN18 中实体之间的余弦距离

Fig. 2 Cosin distance between entities in WN18

表 3 各方法在 2 种数据集上的链接预测结果

Tab. 3 Link prediction results on two data sets by different methods

数据集	方法	Raw Mean Rank (头实体/尾实体)	Filtered Mean Rank (头实体/尾实体)	Raw Hits@10 (头实体/尾实体)	Filtered Hits@10 (头实体/尾实体)
WN18	UM	315	304	35.3	38.2
	TransE	263	251	75.4	89.2
	TransH	318/401	303/388	75.4/73.0	86.7/82.3
	TransR	232/238	219/225	78.3/79.8	91.7/92.0
	TranSparse(US)	233/223	221/211	79.6/80.1	93.4/93.2
	TranSparse(S)	235/224	223/221	79.0/79.8	92.3/92.8
	TransD	242/224	229/212	79.2/79.6	92.5/92.2
	TransF	—	198	—	95.3
	ProjE	248.9/254.3	231.2/238.4	78.7/80.2	95.3/95.0
	LCPE	234.2/238.3	216.6/222.4	78.9/80.2	95.2/95.0
FB15k	UM	1074	979	4.5	6.3
	TransE	243	125	34.9	47.1
	TransH	211/212	84/87	42.5/45.7	58.5/64.4
	TransR	226/198	78/77	43.8/48.2	65.5/68.7
	TranSparse(US)	216/190	66/82	50.3/53.7	78.4/79.9
	TranSparse(S)	211/187	63/82	50.1/53.3	77.9/79.5
	TransD	211/194	67/91	49.4/53.4	74.2/77.3
	TransF	—	62	—	82.3
	ProjE	278.7/181.8	83.8/58.7	41.1/48.7	74.5/79.6
	LCPE	269.2/176.4	75.5/54.7	44.4/52.1	77.2/82.3

神经网络模型由于其出色的信息感知和分析能力可以利用更少的参数实现更强的关系预测能力。LCPE 模型和 ProjE 模型充分发挥了这一优势,同时针对知识图谱补全问题采用 listwise 方法训练神经网络模型,考虑所有三元组的得分,尽可能保证所有正例的得分比负例的得分高,这也是它们的预测能力强的重要原因之一。

LCPE 比 ProjE 在 WN18 数据集上 Mean Rank 平均提前了 11,Hits@10 提升了 0.2 个百分点;在 FB15k 上 Mean Rank 提前了 7.5,Hits@10 平均提升了 3.05 个百分点。这也说明了实体相似度信息可以用于辅助判断实体间的具体关系类型,从而提高预测能力。

3 结语

知识图谱补全算法是对实体之间关系的预测算法,由于当今时代的信息量过大导致了预测模型可能具有巨大的参数量,所以如何利用尽可能少的参数量达到更高的预测精度就成了一个很重要的问题。ProjE 算法不仅实现了小参数量并且能够更为出色的预测到实体之间的各种关系。通过对 ProjE 算法和 Unstructured Model 的研究,本文提出了 CLPE 模型,它将利用了实体之间的相似度作为辅助信息优化了 ProjE 模型,并通过实验证明了实体之间的相似度作为辅助信息的合理性以及 CLPE 在没有增大参数规模的基础上将模型的预测能力进一步提升。

以简单的共享变量神经网络为基础的算法在知识图谱补全方面取得了很好的效果,而如何将更多看似简单但其实很有效的方法利用到知识图谱补全算法中,或如何优化现有模型以取得更好的效果仍然有待研究。

《计算机应用》www.joca.cn

参考文献(References)

[1] SINGHAL A. Introducing the knowledge graph: things, not strings [EB/OL]. [2012-01-08]. http://googleblog.blogspot.ie/2012/05/introducing-knowledge graph-things-not.html.

[2] 徐增林,盛泳潘,贺丽荣,等.知识图谱技术综述[J].电子科技大学学报,2016,45(4):589-606.(XU Z L, SHENG Y P, HE L R, et al. Review on knowledge graph techniques[J]. Journal of University of Electronic Science and Technology of China, 2016, 45(4):589-606.)

[3] LIN Y, LIU Z, ZHU X, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2015: 2181-2187.

[4] SHI B, WENINGER T. ProjE: embedding projection for knowledge graph completion[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2017: 1236-1242.

[5] SOCHER R, CHEN D, 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, Nevada: Curran Associates Inc., 2013: 926-934.

[6] 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.(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.)

[7] NEELAKANTAN A, ROTH B, MCCALLUM A. Compositional vector space models for knowledge base completion[J/OL]. arXiv Preprint, 2015, arXiv: 1504.06662(2015-04-24)[2015-05-27]. https://arxiv.org/abs/1504.06662.

[8] GARCIA-DURAN A, BORDES A, USUNIER N. Composing relationships with translations[EB/OL]. [2018-03-20]. http://www.emnlp2015.org/proceedings/EMNLP/pdf/EMNLP034.pdf.

[9] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J/OL]. arXiv Preprint, 2013, arXiv: 1301.3781(2013-01-16)[2013-09-07]. https://arxiv.org/abs/1301.3781.

[10] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J/OL]. arXiv Preprint, 2013, arXiv: 1310.4546[2013-10-16]. https://arxiv.org/abs/1310.4546.

[11] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[EB/OL]. [2018-03-20]. http://www.thespermwhale.com/jaseweston/papers/CR_paper_nips13.pdf.

[12] XIAO H, HUANG M, HAO Y, et al. TransA: an adaptive approach for knowledge graph embedding[J/OL]. arXiv Preprint, 2015, arXiv: 1509.05490(2015-09-18)[2015-09-28]. https://arxiv.org/abs/1509.05490.

[13] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]// Proceedings of the 28th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2014: 1112-1119.

[14] JI G, LIU K, HE S, et al. Knowledge graph completion with adaptive sparse transfer matrix[C]// Proceedings of the 13th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2016: 985-991.

[15] JI G, HE S, XU L, et al. Knowledge graph embedding via dynamic mapping matrix[EB/OL]. [2018-03-20]. http://www.anthology.aclweb.org/P/P15/P15-1067.pdf.

[16] DO K, TRAN T, VENKATESH S. Knowledge graph embedding with multiple relation projections[J/OL]. arXiv Preprint, 2018, arXiv: 1801.08641[2018-01-26]. http://arxiv.org/abs/1801.08641.

[17] Bordes A, Glorot X, Weston J, et al. Joint learning of words and meaning representations for open-text semantic parsing[EB/OL]. [2018-03-20]. http://www.thespermwhale.com/jaseweston/papers/otsm_aistats11.pdf.

[18] JEAN S, CHO K, MEMISEVIC R, et al. On using very large target vocabulary for neural machine translation[EB/OL]. [2018-03-20]. http://www.anthology.aclweb.org/P/P15/P15-1001.pdf.

[19] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[EB/OL]. [2018-03-20]. https://pub-tools-public-publication-data.storage.googleapis.com/pdf/44876.pdf.

[20] GUTMANN M, HYVÄRINEN A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models[EB/OL]. [2018-03-20]. http://proceedings.mlr.press/v9/gutmann10a/gutmann10a.pdf.

[21] KILICOGLU H, SHIN D, FISZMAN M, et al. SemMedDB: a PubMed-scale repository of biomedical semantic predications[J]. Bioinformatics, 2012, 28(23): 3158-3160.

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

[23] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2018-03-20]. http://yeolab.weebly.com/uploads/2/5/5/0/25509700/a_method_for_stochastic_optimization_.pdf.

This work is partially supported by the National Natural Science Foundation of China (61671188, 61571164, 61502122), the National Key R&D Program of China (2016YFC0901902).

WANG Zihan, born in 1995, M. S. candidate. Her research interests include knowledge graph.

SHAO Mingguang, born in 1995, M. S. candidate. His research interests include knowledge graph.

LIU Guojun, born in 1979, Ph. D., associate professor. His research interests include machine learning, computer vision, image processing, pattern recognition.

GUO Maozu, born in 1966, Ph. D., professor. His research interests include machine learning, bioinformatics, urban computing.

BI Jiandong, born in 1964, Ph. D., associate professor. His research interests include machine learning.

LIU Yang, born in 1976, Ph. D., associate professor. His research interests include machine learning.

《计算机应用》www.joca.cn