计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800195

结合三元组重要性的知识图谱补全模型

李忠文[1]　丁　烨[2]　花忠云[1]　李君一[1]　廖　清[1]
1 哈尔滨工业大学(深圳)计算机科学与技术学院　广东 深圳 518055
2 东莞理工学院网络空间安全学院　广东 东莞 523808
(18S151566@stu.hit.edu.cn)

摘　要　知识图谱是人工智能方向的一个热门研究领域。知识图谱补全是在给定头实体或者尾实体以及相应关系的条件下，补全缺失实体。基于翻译的模型如 TransE,TransH 和 TransR 是最常用的一类知识图谱补全方法。然而,大多数现有的补全模型在补全过程中都忽略了知识图谱中三元组重要性的特征。文中提出了一种新型的知识图谱补全模型 ImpTransE,该模型考虑了三元组中的重要性特征,设计了实体重要性排序方法 KGNodeRank 和多粒度关系重要性估计方法 MG-RIE,分别对实体重要性和关系重要性进行估计。具体来说,KGNodeRank 通过同时考虑关联结点的重要性及其重要性传递方向的概率来估计实体结点的重要性排名。MG-RIE 则同时考虑了关系的一阶重要性和高阶重要性,从而对关系的总体重要性进行合理估计。ImpTransE 同时考虑了三元组的实体重要性和关系重要性特征,使其在学习过程中对于不同的三元组信息可赋予不同的关注程度,提高了模型的表示学习性能,从而达到了更好的补全效果。实验结果表明,在两类知识图谱数据集中与 5 种对比模型相比,ImpTransE 模型在大部分指标上均具有最佳的补全性能,对不同数据集的补全效果获得了一致的提升。

关键词:知识图谱;关系重要性;实体重要性;链接预测
中图法分类号　TP391

Knowledge Graph Completion Model Based on Triplet Importance Integration

LI Zhong-wen[1],DING Ye[2],HUA Zhong-yun[1],LI Jun-yi[1] and LIAO Qing[1]
1 Department of Computer Science and Technology,Harbin Institute of Technology,Shenzhen,Shenzhen,Guangdong 518055,China
2 Department of Cyberspace Security,Dongguan University of Technology,Dongguan,Guangdong 523808,China

Abstract　Knowledge graph is a popular research area related to artificial intelligence. Knowledge graph completion is the completion of missing entities given head or tail entities and corresponding relations. Translation models (such as TransE,TransH and TransR) are one of the most commonly used completion methods. However,most of the existing completion models ignore the feature of the importance of the triplets in the knowledge graph during the completion process. This paper proposes a novel knowledge graph completion model,ImpTransE,which takes into account the importance feature in triplets,and designs the entity importance ranking method KGNodeRank and the multi-grained relation importance estimation method MG-RIE,to estimate the entity importance and relation importance,respectively. Specifically,the KGNodeRank method estimates the entity node importance ranking by considering both the importance of the associated nodes and the probability that their importance is transmitted,while the MG-RIE method considers multi-order relation importance to provide a reasonable estimate of the overall importance of the relation. ImpTransE takes into account the entity importance and relation importance features of triplets,so that different levels of attention are given to different triplets during the learning process,which improves the learning performance of the ImpTransE model and thus achieves better completion performance. Experimental results show that ImpTransE model has the best completion performance in most of the metrics on the two knowledge graph datasets compared with the five comparison models,and completion performance of different datasets is consistently improved.

Keywords　Knowledge graph,Relation importance,Entity importance,Link prediction

1　引言

知识图谱已经成为众多人工智能应用的重要信息来源，如百科类知识图谱 Freebase 被应用于 Google 搜索引擎,Facebook Social Graph 知识图谱被应用于 Facebook 的社交搜索产品中。常用的知识图谱有 WordNet[1],Freebase[2] 和 Ya-

到稿日期:2020-05-31　返修日期:2020-09-16　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(U1711261)
This work was supported by the National Natural Science Foundation of China(U1711261).
通信作者:廖清(liaoqing@hit.edu.cn)

go[3]等。知识图谱通常包含以三元组形式(h,r,t)存储的大量结构化数据，其中h表示头实体，t表示尾实体，r表示头实体h和尾实体t之间的某种关系。在真实世界中，即使一个最常用的知识图谱应用(如Freebase)中包含大量表示各类事实信息的三元组，但仍存在信息不完整、数据稀疏等问题。知识图谱补全的目的是基于知识图谱中现有的三元组信息预测缺失实体。根据预测实体位置的不同，知识图谱补全分为头实体预测$(?,r,t)$和尾实体预测$(h,r,?)$。已有学者在知识图谱补全领域提出了一些研究模型，如距离模型[4]、单层神经网络模型[5]等。距离模型是较早的知识表示模型，在距离模型中，所有实体被投影到相同的d维向量空间之中，使实体用d维向量来表示。单层神经网络模型是距离模型的进一步改进，采用了非线性处理来缓解距离模型刻画语义联系时精确度不足的问题，但是该非线性处理只描述实体与关系之间的微弱联系。最近，基于翻译的知识图谱补全方法通过将三元组信息嵌入低维向量空间以实现知识的可计算性[6-8]，降低知识图谱的计算复杂度，从而提高知识图谱补全的性能。TransE[6]模型是最典型的翻译模型之一，其思想是将一个三元组中的关系r当作将头实体h转换到尾实体t的一个翻译(Translation)操作，即当三元组(h,r,t)成立时，TransE认为在低维空间中$\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$。TransE模型适合对一对一的关系建模，但在一对一、多对一和多对多等复杂关系上的表现较差。TransH模型[7]改善了TransE在复杂关系建模上的缺点，通过将关系建模为一个超平面，使得同样的实体在特定的关系下会产生特定的表示。对于关系r，TransH模型同时使用平移向量$\boldsymbol{r}$以及超平面的法向量$\boldsymbol{w}_r$来表示。头实体和尾实体被投影到关系的超平面得到向量$\boldsymbol{h}_\perp=\boldsymbol{h}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{h}\boldsymbol{w}_r$以及$\boldsymbol{t}_\perp=\boldsymbol{t}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{h}\boldsymbol{w}_r$。对于一个正确的三元组$(h,r,t)$，TransH模型认为在向量空间中$\boldsymbol{h}_\perp+\boldsymbol{r}\approx\boldsymbol{t}_\perp$。TransE和TransH的共同假设是实体和关系处于同样的向量空间。而TransR[8]模型则考虑实体和关系是不同类型的对象，通过映射矩阵$\boldsymbol{M}_r$将表示头实体的向量$\boldsymbol{h}$和表示尾实体的向量$\boldsymbol{t}$投影到关系r所在的向量空间，即TransR模型认为在向量空间中$\boldsymbol{M}_r\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{M}_r\boldsymbol{t}$。然而，这些方法都没有考虑三元组中的重要性信息。不同的实体或者三元组的重要程度各不相同，有些实体的信息相比其他实体来说更加重要[9]。例如，根据图1展示的百度指数的搜索日均值(引自百度指数)[1]可知，对于马云和脸萌的创始人郭列这两位企业家，无论是整体日均值还是移动日均值，马云均远远高于郭列，由此可知马云相对郭列来说受关注程度更高，对于知识图谱来说更加重要。

关键词	整体日均值	移动日均值
马云	14,748	12,645
郭列	105	55

图1 百度指数对比：马云和郭列

Fig. 1 Baidu index comparison: Ma Yun and Guo Lie

模型在学习知识图谱信息的过程中应当对不同的三元组给予不同的关注度，由此我们提出了结合三元组重要性的知识图谱补全模型ImpTransE，通过在学习知识图谱三元组的过程中充分考虑重要性信息，来提升补全性能。

本文的贡献有以下3点：1)提出了一个结合三元组重要性的新型知识图谱补全模型ImpTransE，将重要性信息融入知识图谱补全。2)针对实体和关系的不同性质，提出了两个方法，KGNodeRank和MG-RIE，分别用于估计实体重要性和关系重要性，同时对关系重要性从不同粒度进行了考虑，包括一阶重要性和高阶重要性，充分挖掘了重要性信息。相比于TransE，ImpTransE在学习三元组信息时对不同的三元组给予了不同的关注度。3)在真实数据集上，通过实验验证了ImpTransE的有效性；在链接预测任务上，该模型优于TransE和其他的知识表示模型。

2 相关工作

首先定义相关的数学符号。(h,r,t)表示一个三元组，$\boldsymbol{h}$，$\boldsymbol{r}$，$\boldsymbol{t}$分别表示它们的列向量，$\boldsymbol{M}$表示从实体向量空间映射到关系向量空间的矩阵。$f_r(\boldsymbol{h},\boldsymbol{t})$表示一个三元组的得分函数。对于一个正确的三元组，得分函数的打分较低，而对于一个不正确的三元组，打分较高。

2.1 基于翻译的模型

基于翻译的模型认为，对于一个三元组(h,r,t)，关系r可以当作是从头实体h到尾实体t的一个翻译(Translation)操作。正如引言部分所说，在TransE[6]模型中，如果(h,r,t)是一个事实三元组，TransE希望在向量空间中有$\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$，如图2所示。

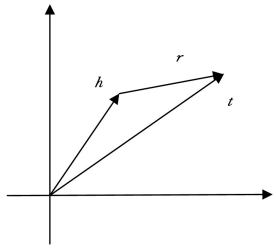


图2 TransE模型

Fig. 2 TransE model

TransE定义得分函数如下：

$$f_r(\boldsymbol{h},\boldsymbol{t})=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_2^2 \tag{1}$$

TransE适合对一对一关系建模，但是在处理一对多、多对一和多对多的情况效果却不够好。

为了解决这些问题，TransH[7]使一个实体在不同关系的超平面下有着不同的表示。对于一个关系r，TransH将关系建模为特定关系的超平面，而不是与实体嵌入在同一个空间。具体地，对于一个三元组(h,r,t)，实体嵌入$\boldsymbol{h}$和$\boldsymbol{r}$首先被投影到超平面$\boldsymbol{w}_r$。投影向量分别用$\boldsymbol{h}_\perp$和$\boldsymbol{t}_\perp$表示。得分函数被定义为：

$$f_r(\boldsymbol{h},\boldsymbol{t})=\|\boldsymbol{h}_\perp+\boldsymbol{r}-\boldsymbol{t}_\perp\|_2^2 \tag{2}$$

TransH通过限制$\|\boldsymbol{w}_r\|=1$确保$\boldsymbol{h}_\perp$和$\boldsymbol{t}_\perp$在关系r的超平面上。对于TransE和TransH来说，它们的共同假设是实体和关系在同一个向量空间。虽然TransH通过利用关系超

[1] http://index.baidu.com

平面提高了建模的灵活性，但没有完全打破该假设的限制。

然而，实体和关系是不同类型的对象，应该位于不同的向量空间。基于该假设，TransR[8]模型被提出，它将实体和关系建模在不同的空间，也就是实体空间以及关系空间。TransR对每个特定关系 r 使用了特定的投影矩阵 $\boldsymbol{M}_r$，从而将实体映射到特定关系的子空间。TransR 的得分函数为：

$$f_r(\boldsymbol{h},\boldsymbol{t})=\| \boldsymbol{M}_r\boldsymbol{h}_\perp+\boldsymbol{r}-\boldsymbol{M}_r\boldsymbol{t}_\perp \|_2^2 \tag{3}$$

2.2 其他模型

除了基于翻译的模型之外，语义匹配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。

DISTMULT[10]模型将实体用向量表示，关系用对角矩阵表示，该关系矩阵对潜在因素之间的成对交互作用进行了建模，它的评分函数是一个双线性函数，通过匹配实体的潜在语义和向量空间表示中包含的关系来度量三元组的可信性。但 DISTMULT 模型只能处理对称的关系，而不适合处理其他非对称关系。

ComplEx[11]模型引入复数扩展 DISTMULT 模型，以便更好地对非对称关系进行建模。此时，实体、关系都在复数空间，来自非对称关系的事实可以根据所涉及实体的顺序得到不同的分数。

3 结合重要性信息的知识图谱补全模型

本节将介绍结合重要性信息的知识图谱补全模型 ImpTransE。该模型主要包含 3 个模块：1）实体重要性模块，考虑到关联结点的重要性及其重要性传递方向的概率，设计了 KGNodeRank 方法用于估计实体重要性；2）关系重要性模块，综合考虑了关系的一阶重要性和高阶重要性，提出了 MG-RIE 方法；3）三元组重要性模块，结合已获得的实体重要性得分和关系重要性得分，设计了三元组的重要性得分公式，从而获得整体的三元组重要性。图 3 是 ImpTransE 的模型结构图。

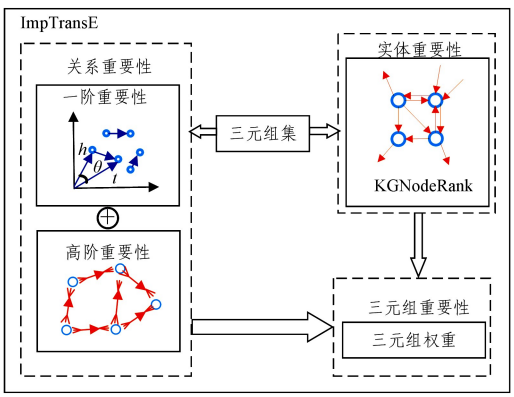


图 3 ImpTransE 模型结构图

Fig. 3 Model structure of ImpTransE

3.1 实体重要性估计

对于三元组中的实体重要性估计，考虑到不同实体受关注的程度不同，引入了 PageRank[12-13]方法的思想，设计了 KGNodeRank 方法。传统的 PageRank 方法在估计网页的重要性时，只考虑了出入度的链接关系。在知识图谱中，考虑到实体结点的重要性不仅与其关联关系的数目相关，还与关联结点的重要性以及重要性传递的方向相关，本文提出了 KGNodeRank 方法，用于估计知识图谱中实体的重要性排名。在 KGNodeRank 方法中，计算实体 e_i 的重要性排名的公式如下：

$$E_{imp}(e_i)=\frac{1-d}{N}+d\sum_{e_j\in M(e_i)}P_{ij}E_{imp}(e_j) \tag{4}$$

其中，e_i 表示实体，i 和 j 分别表示实体索引，$e_j\in M(e_i)$，$M(e_i)$是与e_i关联的入度结点集合；N 是知识图谱中的实体总数；d 是阻尼系数；P_{ij} 代表入度结点e_j传递重要性给结点e_i的概率，该概率由结点e_i到图谱中其他结点的最短距离即紧密度中心性决定；$E_{imp}(e_j)$表示实体e_j的重要性排名。紧密度中心性反映网络中某一结点与其他结点的接近程度。如果一个结点离其他结点越近，则该结点在网络中往往具有更大的影响力。结点e_i的紧密度中心性C_{e_i}是指结点e_i到其余所有结点$e_j(j\neq i)$的最短距离 $d(e_i,e_j)$的平均值的倒数：

$$C_{e_i}=\left[\frac{1}{N-1}\sum_{i\neq j}^{N}d(e_i,e_j)\right]^{-1}=\frac{N-1}{\sum\limits_{i\neq j}^{N}d(e_i,e_j)} \tag{5}$$

其中，C_{e_i} 表示结点e_i的紧密度中心性，N 是知识图谱中的实体结点的总数，$d(e_i,e_j)$表示两个结点e_i和e_j的最短距离。

因此，在知识图谱中，入度结点e_j传递重要性给结点e_i的概率公式如下：

$$P_{ij}=\frac{C_i}{\sum\limits_{m=1}^{n}C_m} \tag{6}$$

其中，P_{ij} 代表入度结点e_j传递重要性给结点e_i的概率，n 是结点e_j指向的某个结点数量，C_i是结点e_i的紧密度中心性，C_m代表结点e_j指向的某个结点e_m的紧密度中心性。

3.2 关系重要性估计

知识图谱的某一个结点会受到除其自身以外的其他结点的影响[14-16]。根据距离的远近，我们将其划分为两类：一阶结点和高阶结点。一阶结点指距离自身结点只有一跳距离的结点，即邻居结点；高阶结点指距离自身结点有两跳及以上距离的结点。针对这两类情况，本文提出了多粒度关系重要性估计方法 MG-RIE，该方法分别估计了关系的一阶重要性和高阶重要性。对于关系r_k的一阶重要性，定义其由该关系两端的实体决定，若在向量空间中，两端实体向量的余弦相似度越大，那么认为这两个实体之间的关联性越强，则一阶重要性越高。一阶关系重要性的计算公式如下：

$$R_{imp_1}(r_k)=\frac{\boldsymbol{h}_k\cdot\boldsymbol{t}_k}{\|\boldsymbol{h}_k\|\ \|\boldsymbol{t}_k\|} \tag{7}$$

其中，$R_{imp_1}(r_k)$表示三元组中关系r_k的一阶重要性，$\boldsymbol{h}_k$和$\boldsymbol{t}_k$分别表示第 k 个三元组的头实体向量和尾实体向量，$\|\cdot\|$表示 L2 范数。

对于高阶关系重要性，模型将三元组中的头实体、尾实体和关系之间的关联看作对关系的“投票”，再结合实体重要性，获得关系的高阶重要性。举例来说，如图 4 所示，假设 r 是一个 3-2 的关系，但是对于三元组(h_2,r,t_2)的形成来说，该头实体对关系的关联可看作一次“投票”。同理，尾实体也是如此，将两端实体的重要性得分分别除以各自的可选路径数目，将其相加获得该三元组中关系的高阶重要性。

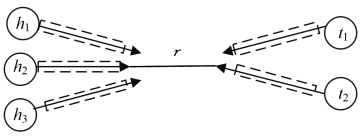


图 4 高阶关系重要性示意图

Fig. 4 Diagram of high-order relation importance

计算关系的高阶重要性的公式如下：

$$R_{imp_2}(r_k)=\frac{E_{imp}(h_k)}{h_{r_k}pt_{r_k}}+\frac{E_{imp}(t_k)}{t_{r_k}ph_{r_k}} \tag{8}$$

其中，$R_{imp_2}(r_k)$表示关系r_i的高阶重要性，$E_{imp}(h_k)$和$E_{imp}(t_k)$分别表示实体h_i和t_i的实体重要性，$h_{r_k}p\,t_{r_k}=\frac{n(\Delta_{r_k})}{n(t_{r_k})}$，$t_{r_k}$表示属于关系$r_k$的尾实体[17]，$\Delta_{r_k}$表示包含第 k 个关系r_k的训练元组，则$h_{r_k}p\,t_{r_k}$表示在关系r_k中，每个尾实体对应的平均三元组数；类似地，$t_{r_k}p\,h_{r_k}=\frac{n(\Delta_{r_k})}{n(h_{r_k})}$，表示在关系$r_k$中，每个头实体对应的平均三元组数。

在获得关系的一阶重要性和高阶重要性之后，将两者结合，可获得总体的关系重要性得分$R_{imp}(r_k)$。结合方式如下：

$$R_{imp}(r_k)=\alpha R_{imp_1}(r_k)+(1-\alpha)R_{imp_2}(r_k) \tag{9}$$

其中，α 是一个超参数，用于权衡关系的一阶重要性和高阶重要性。

通过以上方式，我们分别获得了关系的一阶重要性得分和高阶重要性得分，将两者结合后可获得总体的关系重要性得分。

3.3 三元组重要性估计

在获得实体重要性和关系重要性后，对于一个三元组tri_k，如(h_k,r_k,t_k)，可获得该三元组的重要性分数 $w(tri_i)$，其计算公式如下：

$$w(tri_k)=\frac{E_{imp}(h_k)+E_{imp}(t_k)+R_{imp}(r_k)}{count(fac)} \tag{10}$$

其中，$w(tri_k)$表示三元组tri_i的重要性，$E_{imp}(h_k)$表示实体h_k的重要性，$E_{imp}(t_k)$表示实体t_k的重要性，$R_{imp}(r_k)$表示关系r_k的重要性，$count(fac)$表示因子集合的元素数目，fac 表示用于计算三元组重要性的因子集合。算法 1 给出了 ImpTransE 的学习算法。

算法 1 Learning ImpTransE

input: Training set $S=\{(h,r,t)\}$, $S'=\{(h',r,t')\}$, entities set E and relations set R, S' is produced from S by entity replacement

1. initialize pt←for each triplets∈S, nt←for each triplet $s'\in S'$
2. loop
3. for $(h_k,r_k,t_k)\in S$(resp. ,$(h_k,r_k,t_k')\in S'$) do
4. $E_{imp}(e_i)=\frac{1-d}{N}+d\sum_{e_j\in M(e_i)}\frac{E_{imp}(e_j)}{L(e_j)}$ for S and S'
5. $R_{imp_1}(r_k)=\frac{h_k\cdot t_k}{\|h_k\|\,\|t_k\|}$
6. $R_{imp_2}(r_k)\leftarrow\frac{E_{imp}(h_k)}{h_{r_k}p\,t_{r_k}}+\frac{E_{imp}(t_k)}{t_{r_k}p\,h_{r_k}}$
7. $R_{imp}(r_k)\leftarrow\alpha R_{imp_1}(r_k)+(1-\alpha)R_{imp_2}(r_k)$
8. $w(tri_k)\leftarrow\frac{E_{imp}(h_k)+E_{imp}(t_k)+R_{imp}(r_k)}{count(fac)}$
9. end for
10. Update embeddings w. r. t.
11. $\sum_{(h,r,t)\in S}\sum_{(h',r,t')\in S'}[\gamma+w(tri_k)(dis(h+r,t)-dis(h'+r,t'))]_+$
12. end loop

4 实验和分析

4.1 数据集和评测指标

为了评估本文提出的方法，我们使用了常用的标准数据集 FB15K[18] 和 WN18[18]。数据集的统计情况如表 1 所列。

表 1 数据集统计信息

Table 1 Statistics of datasets

Dataset	# Rel	# Ent	# Train	# Valid	# Test
FB15K	1345	14951	483142	50000	59071
WN18	18	40943	141442	5000	5000

在实体链接预测中，已有的研究工作通常将平均排序 MR(mean rank)和 HITS@10 作为评测指标。平均排序指标用来衡量正确的实体在所有候选实体中的平均排名，该指标数值越低，表明正确实体在候选实体列表中的排名越靠前，模型预测越准确。HITS@10 指标用来衡量正确的实体排名前 10 位的概率，该指标数值越大，表明效果越好。

4.2 实验设置

首先描述数学符号。假设在训练集中有n_t个三元组，$(h_k,r_k,t_k)(k=1,2,\cdots,n_t)$表示第 k 个三元组。每个三元组有一个标签y_i，表明该三元组是正的($y_k=1$)或者是负的($y_k=0$)，即正三元组集 $S=\{(h_k,r_k,t_k)\mid y_k=1\}$，负三元组集 $S'=\{(h_k,r_k,t_k)\mid y_k=0\}$。然而，知识图谱本身只包含正三元组，没有负三元组，因此，我们从知识图谱获得正三元组集 S，然后根据相应的负三元组生成规则：$S'=\{(h_j,r_k,t_k)\mid h_j\neq h_k\wedge y_k=1\}\cup\{(h_k,r_k,t_j)\mid t_j\neq t_k\wedge y_k=1\}$，通过将正三元组中正确的头实体或尾实体替换成错误的实体来生成对应的负三元组。另外，根据以往的研究工作，通常有两种替换策略：Unif 和 Bern[18]。Unif 指随机替换正三元组包含的头实体或者尾实体，但是知识图谱本身是相当不完善的，随机抽样策略可能会在训练中引入错误的负三元组。Bern 方法更多地考虑了三元组中关系的映射性质，基于不同的概率来选择替换头实体或尾实体。例如，知识图谱中一般有 4 种类型的三元组关系：一对一、一对多、多对一和多对多。如果一个三元组中的关系是一对多的，Bern 方法倾向于以更大的概率替换头实体；如果一个三元组中的关系是多对一的，则其倾向于以更大的概率替换尾实体。

训练目标函数如下：

$$L=\sum_{(h,r,t)\in S}\sum_{(h',r,t')\in S'}[\gamma+w(tri_k)(dis(h+r,t)-dis(h'+r,t'))]_+ \tag{11}$$

其中，$[\cdot]_+$表示取正部运算，$\gamma>0$ 是一个 margin 超参数，S 是一个正三元组的集合，S'是一个负三元组的集合，$d(h+r,t)$表示一个三元组的距离得分(或者能量得分)。例如，对于一个正三元组(h,r,t)，$h+r\approx t$，它的距离得分 $d(h+r,t)$往往较小，而对于一个负三元组，它的距离得分往往较大，这里使用 $L1$ 范数来计算距离得分。最小化该目标函数的过程可利用随机梯度下降来执行。

实验中使用了两个数据集：FB15K 和 WN18。并且设置了超参数：$margin\gamma=\{1,2,2.5,3\}$，向量的嵌入维度 $m=\{10,$

30，50，100}，批大小 $B=\{100,200,500,1000\}$，在验证集上的最好设置是 $\gamma=\{2\}$，$m=\{50\}$，$B=\{1000\}$。

4.3 链接预测

链接预测的任务是预测一个三元组中缺失的头实体或尾实体。在该任务的测试过程中，三元组的头实体或尾实体被移除，然后使用知识图谱的其他实体替换，知识图谱补全模型对用其他实体替换后的三元组整体进行打分。对于每个位置的缺失实体，模型根据打分对来自知识图谱的候选实体进行降序排列，而不是产生一个最好的结果。正确实体的排序会被记录。因为知识图谱中存在一对多、多对一和多对多等复杂类型的关系，所以一个被扰乱的三元组也可能是正确的三元组。因此，在排序之前，过滤包含在训练集、验证集以及测试集中的扰乱的三元组，可以更好地评估模型性能，这种设置被称为 Filter。因此，本文存在两种评估设置：Raw（没有移除被扰乱的三元组）和 Filter（移除了被扰乱的三元组）。同时，为了进一步验证考虑重要性信息的有效性，我们在实验过程中还设计了只考虑实体重要性信息的模型 EnTransE。通过与 5 种对比模型以及 ImpTransE 模型的对比，可以进一步验证考虑实体重要性和三元组重要性的知识图谱补全方法的有效性。

ImpTransE 模型在数据集上的实验效果如表 2 和表 3 所列。从表 2 和表 3 的实验结果可以发现：1）ImpTransE 的效果总体上优于以往的知识图谱补全模型。2）三元组重要性对于知识图谱补全的有效性得到了验证，在 FB15K 和 WN18 两个公开数据集上，ImpTransE 在 Mean Rank 指标上取得了一致且较明显的提升；同时，EnTransE 模型通过考虑实体重要性，相比于对比模型同样取得了不错的性能提升，ImpTransE 在 EnTransE 的基础上进一步考虑了关系重要性，总体上达到了最佳的效果。3）在 FB15K 公开数据集上，ImpTransE 在 Raw 设置下的 HITS@10 性能上要稍弱于 ComplEx，但在 Filter 设置下的 HITS@10 性能上要优于所有对比模型；在 WN18 公开数据集上，ImpTransE 在 Mean Rank 指标上的表现优于所有对比模型，表明 ImpTransE 在进行实体链接预测时，预测正确候选实体的位置更加靠前，精确度更高。

表 2 FG15K 数据集上的链接预测结果

Table 2 Link predicion results on FB15K dataset

Model	FB15K			
	Mean Rank		HITS@10/%	
	Raw	Filter	Raw	Filter
TransE[6]	243	125	34.9	47.1
TransH[7]	212	87	45.7	64.4
TransR[8]	198	77	48.2	68.7
DistMult[10]	264.9	167.7	47.3	61.2
ComplEx[11]	275.6	173.4	**50.4**	67.3
EnTransE(Ours)	*213.7*	*68.8*	*48.1*	***73.2***
ImpTransE(Ours)	**189.4**	**55.2**	49.1	71.3

从表 2 可知，在 FB15K 数据集上，在 Raw 设置下，相比 TransE 模型，ImpTransE 的 Mean Rank 指标提升了约 22.1%，HITS@10 指标提升了约 40.7%；在 Filter 设置下，相比 TransE，ImpTransE 的 Mean Rank 指标提升了约 55.8%，HITS@10 指标提升了 51.4%。TransE 模型及其他对比模型均没有考虑三元组重要性信息，由此验证了 ImpTransE 模型在知识图谱补全中考虑重要性信息的有效性。同时，通过 EnTransE 模型与 TransE 模型的效果对比，验证了实体重要性信息对于性能提升的有效性。在 EnTransE 模型的基础上，ImpTransE 模型进一步考虑了关系重要性，验证了 ImpTransE 模型的有效性。

表 3 WN18 数据集上的链接预测结果

Table 3 Link prediction results on WN18 dataset

Model	WN18			
	Mean Rank		HITS@10/%	
	Raw	Filter	Raw	Filter
TransE[6]	263	251	75.4	89.2
TransH[7]	318	303	75.4	86.7
TransR[8]	238	225	79.8	92
DistMult[10]	250.3	235.9	76.4	87.3
ComplEx[11]	306	291	79.0	89.4
EnTransE(Ours)	*284.5*	*271.7*	***79.9***	***93.3***
ImpTransE(Ours)	**205.7**	**193.4**	77.8	90.3

由表 3 可知，在 WN18 数据集上，EnTransE 考虑了实体重要性，相比对比模型取得了不错的性能提升。ImpTransE 在 EnTransE 的基础上进一步考虑了关系重要性，总体上达到了最佳的性能。在 Raw 设置下，相比 TransE，ImpTransE 的 Mean Rank 指标提升了约 21.8%，HITS@10 指标提升了约 3.2%；在 Filter 设置下，相比 TransE，ImpTransE 的 MeanRank 指标提升了约 22.9%，HITS@10 指标提升了约 1.2%，进一步验证了 ImpTransE 模型的有效性。

结束语 本文提出了结合三元组重要性信息的知识图谱补全模型，设计了 KGNodeRank 方法和 MG-RIE 方法来分别估计实体重要性和关系重要性，在 TransE 的基础上结合知识图谱中的三元组重要性信息，用于模型学习时对不同的三元组采取相适宜的重视程度，以提升模型的预测能力。另外，在考虑三元组重要性信息时，对知识图谱中隐藏的重要性信息进行了较充分的挖掘，使得模型在训练过程中对三元组重要性的评估更加合理，进而提高了模型的性能。实验数据表明，与 TransE 和其他一些知识表示模型相比，ImpTransE 总体上取得了最优的性能表现，几乎在各类指标上均取得了一致的性能提升。

下一步将考虑以下研究方向：1）在 ImpTransE 模型的基础上考虑其他信息，如实体类别信息、关系类别信息等；2）考虑知识图谱的自动补全，使得补全后的实体可以被进一步学习；3）在其他的知识图谱表示学习模型上考虑三元组重要性，进一步验证方法的有效性与其他因子的相互影响。

参 考 文 献

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

[2] 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. Association for Computing Machinery，2008：1247-1250.

[3] FABIAN M S，GJERGJI K，GERHARD W. Yago：A core of se-

mantic knowledge unifying wordnet and wikipedia[C]// 16th International World Wide Web Conference, WWW. Association for Computing Machinery, 2007: 697-706.

[4] BORDES A, WESTON J, COLLOBRET R, et al. Learning structured embeddings of knowledge bases[C]// Conference on Artificial Intelligence. 2011 (CONF).

[5] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]// Advances in Neural Information Processing Systems. 2013: 926-934.

[6] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]// Advances in Neural Information Processing Systems. MIT Press, 2013: 2787-2795.

[7] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]// Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI, 2014: 1112-1119.

[8] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]// Twenty-ninth AAAI Conference on Artificial Intelligence. AAAI, 2015: 2181-2187.

[9] PARK N, KAN A, DONG X L, et al. Estimating node importance in knowledge graphs using graph neural networks[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, 2019: 596-606.

[10] YANG B, YIH W, HE X, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv: 1412.6575.

[11] TROUILLON T, WELBL J, RIEDEL S, et al. Complex embeddings for simple link prediction[C]// International Conference on Machine Learning (ICML). 2016.

[12] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: Bringing order to the web[R]. Stanford: StanfordInfoLab, 1999.

[13] ZHANG Z, CAI J, ZHANG Y, et al. Learning Hierarchy-Aware Knowledge Graph Embeddings for Link Prediction[J]. arXiv: 1911.09419.

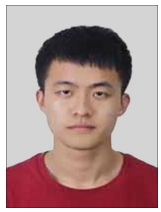
[14] OH B, SEO S, LEE K H. Knowledge graph completion by context-aware convolutional learning with multi-hop neighborhoods [C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 257-266.

[15] ZHU Y, LIU H, WU Z, et al. Representation Learning with Ordered Relation Paths for Knowledge Graph Completion[J]. arXiv: 1909.11864.

[16] WANG C C, CHENG P J. Translating Representations of Knowledge Graphs with Neighbors[C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Association for Computing Machinery, 2018: 917-920.

[17] 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. PACLIC, 2014: 328-337.

[18] HAN X, CAO S, LV X, et al. Openke: An Open Toolkit for Knowledge Embedding[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018: 139-144.



LI Zhong-wen, born in 1996, postgraduate. His main research interests include artificial intelligence and natural language processing.

LIAO Qing, born in 1988, Ph. D, assistant professor. Her research interests include artificial intelligence and data mining.