DOI:10.19651/j.cnki.emt.2312956

# TransREF:一种改进的基于邻域信息的知识表示模型*

王永康[1] 艾山·吾买尔[2] 顾亚东[2] 何江涛[2]
(1.新疆大学软件学院 乌鲁木齐 830091;2.新疆大学信息科学工程学院 乌鲁木齐 830046)

**摘 要**:近年来,知识表示学习在智能推荐、智能问答,以及智能检索方面发挥了关键性作用,受到了广泛关注。知识表示学习旨在借助实体与关系的低维嵌入,将语义信息向量化,通过数学公式进行知识的推理。在众多知识表示学习模型中,TransE 由于评分函数参数较少、计算复杂度低、计算效率高,被认为是最有前途的模型。然而,TransE 在处理除一对一以外的复杂关系时,存在一定的局限性。为了解决这个问题所带来的困扰,提高知识嵌入的质量,本文提出了一种改进的基于翻译模型的知识表示模型 TransREF。首先,借助关系矩阵投影,实现对实体和关系的嵌入;其次在原有向量的基础上加入关系邻域,增强模型的学习能力。在模型被训练期间,对于语义相似度高的实体,通过概率法实现对头实体与尾实体的替换,进而生成的较高质量的负例三元组,并且在选择关系邻域节点时采用五点随机法。最后,选择英文词典 WordNet 的子集 WN18 和 Freebase 子集 FB15K 上进行相关链接预测实验,之后在 3 个公开数据集 WN11、FB13、FB15K 开展三元组分类的实验。结果表明,相较于 TransE、TransH,TransREF 在 MeanRank、Hits@10,以及 ACC 指标上都有较好的性能改善,证明了 TransREF 的有效性。

**关键词**:知识表示;关系矩阵投影;关系邻域;链接预测;三元组分类
**中图分类号**:TP391 **文献标识码**:A **国家标准学科分类代码**:510

# TransREF: An improved knowledge representation model based on neighborhood information

Wang Yongkang[1] Aishan·Wumaier[2] Gu Yadong[2] He Jiangtao[2]
(1. School of Software, Xinjiang University, Urumqi 830091, China; 2. College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

**Abstract**: In recent years, knowledge representation learning has played a crucial role in intelligent recommendation, intelligent question answering, and intelligent retrieval, and has received widespread attention. Knowledge representation learning aims to vectorize semantic information and infer knowledge through mathematical formulas by means of low-dimensional embedding of entities and relationships. Among many knowledge representation learning models, TransE is considered to be the most promising model due to its fewer scoring function parameters, low computational complexity and high computational efficiency. However, TransE has some limitations in dealing with complex relationships other than one-to-one. In order to solve this problem and improve the quality of knowledge embedding, this paper proposes an improved knowledge representation model TransREF based on translation model. Firstly, the embedding of entities and relations is realized by means of relation matrix projection. Secondly, on the basis of the original vector, the relational neighborhood is added to enhance the learning ability of the model. During the training of the model and for entities with high semantic similarity, the replacement of the head entity and the tail entity is realized by the probability method, and then the high-quality negative example triples are generated, and the five-point random method is used to select the relationship neighborhood nodes. Finally, the relevant link prediction experiment is carried out on the subset WN18 of WordNet and the subset FB15K of Freebase, and then the triplet classification experiment is carried out on the three public datasets WN11, FB13 and FB15K. The results show that compared with TransE and TransH, TransREF has better performance improvement in MeanRank, Hits@10, and ACC indicators, which proves the effectiveness of TransREF.

收稿日期:2023-02-28
* 基金项目:新疆维吾尔自治区自然科学基金(2021D01C081)项目资助

**Keywords**: knowledge representation; relation matrix projection; relational neighborhood; link prediction; triplet classification

## 0　引　　言

知识图谱[1](knowledge graph)是大型的多关系图,在实体、关系空间领域被广泛使用,其中节点对应于实体,而实体之间的关系则通过类型化的边进行描述,并且事实常常被编码成三元组,这样的三元组是按照(实体,关系,实体)的形式进行呈现的。从知识图谱被首次提出到现在已经整整十一年,在长期的发展演化过程中,已经有大量的知识图谱被构建出来,例如 DBPedia、知立方、YAGO[2]等。这些结构化大型的知识库由于涵盖了众多的领域,涉及到众多的知识,在应用方面不仅极大的便利了人们对于知识的查找,而且在表示学习方面带动了知识图谱的发展。然而由于知识规模的不断扩大,数据的更新周期缩短,知识的不完整现象正变得越来越明显。为了解决知识不完整的问题,知识表示学习[3](knowledge representation learning)被提了出来,它的主要思想是改变利用三元组$(h, r, t)$表示知识的方式,利用机器学习的知识,将三元组$(h, r, t)$语义信息,在保证知识结构不变的同时,实现实体与关系的向量化,再利用大量的数学公式展开计算,从而完成知识图谱的补全工作。

由于知识表示学习的计算效率较高,目前受到学者们的广泛关注,使得一大批的知识表示学习模型被建立了起来,例如 TransE[4]、TransH[5]、HyTE[6]、IPTransE[7]、PTransE-RNN[8]、puTransE[9]等。在上述这些比较典型的模型中,TransE 由于得分函数参数少、易操作,以及可以缓解数据稀疏问题,被认为是最有发展前景的模型。然而,TransE 在处理除一对一以外的复杂关系时存在一定的局限性,具体地,在面对一对多、多对一、多对多以及自反等复杂关系时,TransE 对有相同关系的不同实体,容易出现错误区分的问题,并且模型表达能力也受到了限制。例如假设知识库中有两个三元组,分别为(拜登、是、总统)和(特朗普、是、总统),这属于多对一的关系,TransE 不能正确区分"拜登"和"特朗普"这两个实体,很容易得出"拜登"实体等于"特朗普"实体的情况,事实上这两个实体的属性、角色、地位是各不相同的,因此这两个实体是不相同的。

为此,本文提出了一种改进的基于邻域信息的知识表示模型,即 TransREF。具体地,首先引入关系矩阵投影的思想,进行实体和关系的向量化表示;其次,根据实体的稀疏程度确定邻节点数;然后,根据实体与邻节点关系的距离确定最佳的关系作为邻域信息,并进行邻节点的实体表示。此外,在选择关系邻域节点时采用五点随机法。在模型被训练时,将改进的抽样策略施加在生成的负例三元组[10]上面,即根据头实体对应尾实体的多少或者是尾实体对应头实体的多少这两种不同的映射关系,实现对头实体与尾实体的替代作用,力求在一定范围内使实体达到充分训练的目的。替换实体时,选择语义相似度高的实体进行替换,进而生成的负例三元组的品质得到了提升。在 WN18、FB15K、WN11、FB13 这 4 个公开的数据集分别实施链接预测和三元组分类两项任务,然后根据评价指标得到的结果进行评估与测试。实验结果表明,TransREF 在 *MeanRank*、*Hits*@10,以及 ACC 三个指标上均有一定幅度改善,从而验证了模型的有效性与准确性。

本文的贡献在于:1)提出了一种全新的模型 TransREF,该模型引入了关系矩阵投影的思想,对实体和关系进行嵌入,较好解决了 TransE 处理复杂关系效率不高的问题,可以对复杂关系进行高效的处理;2)在原有向量的基础上加入关系的邻域信息,使得模型的学习能力增强;3)在选择关系邻域节点时采用五点随机法;4)在实验中,本文的方法在链接预测和三元组分类任务方面性能有明显的改善,优于经典的 TransE 和 TransH 模型。

## 1　相关工作

本文收集了当前已有的知识表示学习模型,对每一种模型的核心思想进行了详细的对比研究。接下来,我们将会从 TransE 和 TransH、其他的方法两个方面进行展开。在这一部分,介绍了当前典型的知识表示学习模型,并分析了其算法思想、评分函数和时间复杂度。在表 1 中,本文使用 $N_e$ 表示时间复杂度,使用小写字母 $d$ 表示实体嵌入的维数,使用小写字母 $k$ 表示关系嵌入的维数,使用小写字母 $j$ 表示神经网络的节点数,使用小写字母 $s$ 表示张量数。小写的 $x$ 表示调整超参数的次数。$W_r$ 表示三维张量。$\boldsymbol{M}$ 代表矩阵。小写字母 $u$ 表示超参数,小写字母 $g$ 表示函数。

### 1.1　TransE 与 TransH 模型

在 TransE 模型中,为了从计算方面达到实体和关系方便运算的目的,采用的方式是将实体和关系映射到向量空间中,这样以来就替换成了对应向量之间的运算,即实体与关系之间的计算采用翻译的规则,将关系视为从实体到实体的翻译过程,通过连续调整 $\boldsymbol{h}$(头实体)、$\boldsymbol{r}$(关系)和 $\boldsymbol{t}$(尾实体)的向量,使$(\boldsymbol{h}+\boldsymbol{r})$尽可能地等于 $\boldsymbol{t}$,即 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$。TransE 的得分函数为:

$$f_r(\boldsymbol{h},\boldsymbol{t}) = \| \boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t} \|_{l_1/l_2} \tag{1}$$

TransH 模型主要是为了解决 TransE 建模在一对多、多对一、多对多等复杂关系时性能效率不佳的问题。此模型采用了超平面投影的方式,关系映射矩阵 $\boldsymbol{w}_r$ 在头实体与尾实体投影到关系超平面上充当了重要角色,从而让不同的实体具有不同的角色表示,再结合法向量经过一定的

**表1 知识表示模型的评分函数与时间复杂度**

| 模型 | 评分函数 | 时间复杂度 |
|---|---|---|
| Unstructured[11] | $-\|\boldsymbol{h}-\boldsymbol{t}\|_2^2$ | $O(N_C)$ |
| SE[12] | $-\|\boldsymbol{M}_{rh}\boldsymbol{h}-\boldsymbol{M}_{rt}\boldsymbol{t}\|_1$ | $O(2\ d^2\ N_c)$ |
| SME(linear)[13] | $g_{left}^{\mathrm{T}}g_{right}$ | $O(4dj\ N_c)$ |
| SME(bilinear)[13] | $g_{left}^{\mathrm{T}}g_{right}$ | $O(4dj\ s\ N_c)$ |
| NTN[14] | $u_r^{\mathrm{T}}f(\boldsymbol{h}^{\mathrm{T}}W_rt+\boldsymbol{M}_{r,1}\boldsymbol{h}+\boldsymbol{M}_{r,2}\boldsymbol{t}+b_r)$ | $O(((d^2+d)s+2dj+j)N_c)$ |
| LFM[15] | $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{M}_r\boldsymbol{t}$ | $O((d^2+d)N_c)$ |
| TransE | $\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_{l_1/l_2}$ | $O(N_c)$ |
| TransH | $\|(\boldsymbol{h}-W_r^{\mathrm{T}}hW_r)+d_r-(\boldsymbol{t}-W_r^{\mathrm{T}}\boldsymbol{t}W_r)\|_{l_1/l_2}$ | $O(2\ dN_c)$ |
| TransR[16] | $\|\boldsymbol{h}_r+\boldsymbol{r}-\boldsymbol{t}_r\|_{l_1/l_2}$ | $O(2\ dkN_c)$ |
| TRKRL[17] | $\|W_{rh}\boldsymbol{h}+\boldsymbol{r}-W_{rt}\boldsymbol{t}\|$ | $O(2\ k^2N_c)$ |
| TransR*[18] | $\|\boldsymbol{h}_r+\boldsymbol{r}-\boldsymbol{t}_r+u\|_{l_1/l_2}$ | $O(2\ dkN_c+x)$ |
| TransREF(本文) | $u^*\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}+\lambda\|_{l_1/l_2}$ | $O(u(2\ dkN_c+x))$ |

转化。最后,再次利用翻译原则进行计算。投影后的分量分别为:

$$\boldsymbol{h}_\perp=\boldsymbol{h}-W_r^{\mathrm{T}}\boldsymbol{h}W_r \tag{2}$$

$$\boldsymbol{t}_\perp=\boldsymbol{t}-W_r^{\mathrm{T}}\boldsymbol{t}W_r \tag{3}$$

关系 $\boldsymbol{r}$ 在超平面上的向量关系为 $d_r$。因此,TransH的损失函数是:

$$f_r(\boldsymbol{h},\boldsymbol{t})=\|(\boldsymbol{h}-W_r^{\mathrm{T}}\boldsymbol{h}W_r)+d_r-(\boldsymbol{t}-W_r^{\mathrm{T}}\boldsymbol{t}W_r)\|_{l_1/l_2} \tag{4}$$

### 1.2 其他模型

彭敏等[19]提出了TransE-NA,该模型利用TransE的翻译规则,以及领域信息的确定是依据实体的领边关系的选取,在邻节点上最相关的属性被选取了,很好解决了已有的模型在建模知识库中的三元组时,或是忽略三元组的邻域信息,导致无法处理关联知识较少的罕见实体的问题。Wang等[20]提出了一种全新的知识表示学习模型,即TransAE,这种模型的关键思想是:首先,借助TransE进行实体与关系的嵌入;其次,加入多模态的自编码器,它不仅将结构知识编码,还将视觉和纹理等知识编码到最终的表示中。王会勇等[21]提出了一种全新的知识表示学习模型,即ITMEA,这种模型能够实现文本与图像多模态的有用融合。模型的核心思想为:首先,获取图像与文本的数据;其次,借助TransE与TransD模型完成实体与关系的嵌入,通过不断迭代的方式不断学习多模态的实体与关系,多模态数据实体对齐的任务被圆满完成了。周泽华等[22]首次提出了一种焕然一新的知识表示学习模型,即Context_RL。模型的关键核心思想为:在进行实体与关系的嵌入时,实体和关系周围包含丰富的上下文信息,为了使这些信息得到了充分的利用,向量是由这些实体上下文信息转换而得到的,然后,输入到初始的模型内,得到嵌入后的实体与关系,最后通过上下文信息来加强实体与关系的语义表示。熊盛武等[23]提出了一种全新的知识表示学习模型,即TransV。该模型的关键思想为:加入了不同维度的元素值,从而达到强化对属性的控制,核心是让不同的关系能够与不同的实体属性有较强的联系。此外,本文提出了一种全新的模型TransREF,首先,该模型引入了关系矩阵投影的思想,对实体和关系进行嵌入,较好解决了TransE处理复杂关系效率不高的问题;其次,在原有向量的基础上加入关系的邻域信息,使得模型的学习能力增强;接着,在选择关系邻域节点时采用五点随机法;最后,在实验中,TransREF在链接预测和三元组分类任务方面性能有明显的改善,优于经典的TransE和TransH模型。

## 2 TransREF

首先,本文对TransREF模型的研究动机进行了介绍,主要是从TransE模型处理复杂关系方面存在的问题入手;其次,分析了TransREF模型的算法思想;最后,介绍了模型的训练过程。

### 2.1 研究动机

虽然现有的基于翻译的方法在链接预测和三重分类方面取得了显著的效果,但它们在处理包含不同实体和关系的大规模知识库方面仍存在不足。主要有以下3点:

1)知识库中的实体是复杂多样的,实体间的联系极其复杂。每一个头实体可能与多个尾实体相连接。同样地,每一个尾实体可能与多个头实体相连接。此外,对于每一对头实体与尾实体,它们之间的关系也是多种多样的,表现出复杂的特征。

2)知识库中的关系比较复杂,在常用的公共数据集FB15K中,只有不到30%的关系是1—1,其他的都是复杂关系。此外,还有其他的一些关系,例如自反、传递等。

3)当三元组$(\boldsymbol{h},\boldsymbol{r}_1,\boldsymbol{t})\in S,\cdots(\boldsymbol{h},\boldsymbol{r}_n,\boldsymbol{t})\in S$,如果使用TransE的翻译规则$h+r\approx t$,那么,容易获得$\boldsymbol{r}_1=\boldsymbol{r}_2=\boldsymbol{r}_3=\cdots=\boldsymbol{r}_n$的奇怪结果。

正是受到这些问题的影响,使得TransE模型在处理复杂关系条件时存在一定的局限性,获得的评分不是很

高，预测的效果不佳。因此，需要对 TransE 的翻译规则进行改进。

### 2.2　TransREF 模型

从前面的分析可以看出，TransE 模型采用了 $h+r\approx t$ 的翻译原则，它在处理一对一和不可逆关系时表现很好，但是，当处理一对多，多对一，多对多，以及自反关系时性能与效率很低。因此，提出了 TransREF 模型。模型的算法思想是：首先引入关系矩阵投影的思想，进行实体和关系的向量化表示；其次，根据五点随机法选择邻节点数，具体地，如果目标关系周围的其他关系多于 5 个，选择最近的 5 个关系，计算每个关系与目标关系之间的距离，并以距离的倒数最为每个关系对于目标关系的重要程度 $u$，如果目标关系周围的其他关系少于 5 个，那么，有多少个关系就选择多少个关系，计算每个关系与目标关系之间的距离，并以距离的倒数最为每个关系对于目标关系的重要程度 $u$；然后，进行邻节点的实体表示，$\lambda$ 为超参数。在模型训练期间，样本中负例三元组的抽样策略的品质得到了更新，即利用一对多和多对一的两种截然不同的映射关系选择替代的实体，使实体中的属性得到充分的训练。替换实体时，如果遇到语义相似度高的实体，则替代他们，这样截然不同的实体之间的区分度就得到了提升（如图 1 所示）。TransREF 的得分函数为：

$$f_r(h,t)=u^*\parallel h+r-t+\lambda\parallel_{l_1/l_2} \tag{5}$$

$$u=\frac{1}{dis} \tag{6}$$

$$dis(r_0,r_i)=\sqrt{|r_0^2-r_i^2|},\quad i=0,1,2,3,4,5 \tag{7}$$

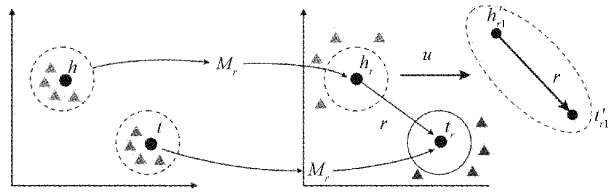


图 1　TransREF 的核心思想

### 2.3　模型训练

无论是进行链接预测，还是三元组分类的实验，在模型建立后都需要进行模型的训练。在其中，尤其需要注意负例三元组的生成方式，因为质量较差的负例三元组会使得模型的训练效果达不到预期的目标，从而影响到实际的得分，在构建方法上主要有两种：随机替换法和概率法。前者是随机打乱正例三元组，优点是：构造的速度比较快，缺点是容易生成过多的假阴性三元组；后者的构造速度比较缓慢，但是相对来说，生成的错的假阴性三元组比较少。因此，在进行模型训练时，选择概率法进行头尾实体的替换，因为这种方式能够让更多的实体属性得到训练，模型能够学习到更多的实体与关系，能够得到更好的参数组合。

1）概率法替换头尾实体

在构建负例三元组，进行头尾实体的替换主要是依据实体间的关系类型。无论实体间是一对多的关系，还是多对一的关系，头实体和尾实体的属性都是不唯一的。当实体间的关系为多对一时，由于尾实体对应头实体数目较多，那么，采用的策略是以更高的概率替换尾实体，因为这样会使尾实体的多个属性得到更好的训练。当实体间的关系为一对多时，由于头实体连接尾实体数目较多，采用的策略是以更高的概率替换头实体，因为可以让头实体的多个属性训练的更加充分。

在进行模型训练时，将会得到每个头实体对应的尾实体的平均数量 $tqh$ 和每个尾实体对应的头实体的平均数量 $hqt$。当采用概率法时，则按照 $m=tqh/(tqh+hqt)$ 的伯努利分布来抽样[24]。本文的负例三元组是从正例三元组构造完成的，用不同的概率替换头尾实体，若以概率 $m$ 替换头实体，则以概率 $1-m$ 替换尾实体，令总的机率为 1。同时本研究采用的抽样法满足伯努利分布的相关要求。之所以选择伯努利分布，因为这种方法能够带来两点好处：(1)可以提升得到正例三元组的概率，(2)可以降低计算复杂度。本文规定，当 $tqh<1.5$ 且 $hqt<1.5$，那么表示关系 $r$ 是一对一的；当 $tqh>1.5$ 且 $hqt>1.5$，那么表示关系 $r$ 是多对多的；当 $tqh\geqslant 1.5$ 且 $hqt<1.5$，那么表示关系 $r$ 是一对多的；当 $tqh<1.5$ 且 $hqt\geqslant 1.5$，那么表示关系 $r$ 是多对一的。

2）选择语义相似度高实体进行替换

当实体具有比较相似的类型时，那么在向量空间中的反映是与实体相对应的向量空间位置通常也较为相近。例如关系为“居住于”对应的头实体通常是人的名字，地点通常是对应于尾实体，人名会聚集分布在一个区域，地点会集中分布在另一个区域，发现聚集在一起实体往往难以区分，进而导致预测结果出现偏差。因此，常常选择语义相似度高的实体进行替换，以提升模型的区分能力。具体情况如图 5 所示。

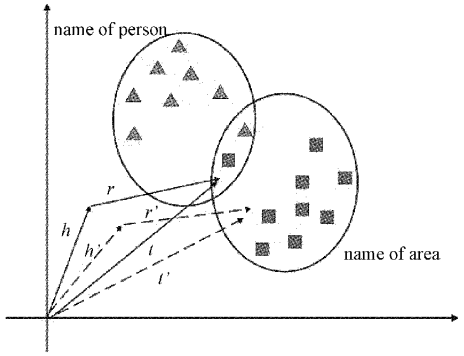


图 2　不易区分的实体举例

在进行实体之间相似度判断时，往往会选择实体或者关系之间语义的相似度进行判断，反映到向量空间也就是计算向量之间的相似度，计算的公式为：

$$gis(h,h')=\sqrt{\sum_{j=1}^{k}(h_j-h_j')} \tag{8}$$

因此,若样本中的一个正例三元组($h$, $r$, $t$)是确定的,那么把头实体替换下来,生成负例三元组($h'$, $r$, $t$)也是确定的,选择$h'$使得$gis(h, h')$最小;另外在替代尾实体生成负例三元组($h$,$r$,$t'$)时,选择$t'$使得$gis(t, t')$最小。

在模型得到训练过程中,为了使正确三元组和错误三元组从样本中相互区分开来,如下基于边际的损失函数将作为训练模型中的优化目标函数:

$$L=\sum_{(h,r,t)\in S}\sum_{(h',r,t')\in S'}\max(f_r(h,t)+\gamma-f_r(h',t'),0) \tag{9}$$

在式(9)中,$S$ 指代的内容为所有的正例三元组所存在的集合,$S'$指代的内容为所有的负例三元组构成的集合。$\max(a,b)$表示的内容是如果$a>b$则返回$a$;反之,返回$b$的数值大小;而$\gamma$指代的内容是通过计算正例三元组损失函数得分与负例三元组损失函数得分之间的距离。所以,该目标函数的优化目标就是最大程度地将样本中正确的三元组和错误的三元组相互分离出来。

## 3 实 验

本实验的运行平台为 pycharm,操作系统为 Linux,编程语言为 python,cpu 为 i7-6700HQ。在框架上,选择的版本为 1.5.0 的 pytorch。此外,为了提高模型的训练速度,本实验选择了内存为 11 G 的 GTX1080ti 的 GPU。

### 3.1 数据集

本组实验选用了 TransE 模型训练过程中使用的 4 个公开数据集:词典知识库 WordNet 由于具有极高的准确率与丰富本体知识,因此选用由 WordNet 抽取出的两个子集 WN18 和 WN11,以及以维基百科为基础构建的 Freebase 知识库,由于该知识库质量高,覆盖面广。因此选用 Freebase 中的两个子集 FB15K 和 FB13。这些数据集的实体数、关系数、训练集、验证集、测试集的具体情况如表 2 所示。

**表 2 数据集统计**

| 数据集 | 实体 | 关系 | 训练集 | 验证集 | 测试集 |
|---|---|---|---|---|---|
| WN18 | 40 943 | 18 | 141 442 | 5 000 | 5 000 |
| FB15K | 14 951 | 1 345 | 483 142 | 50 000 | 59 071 |
| WN11 | 38 696 | 11 | 112 581 | 2 609 | 10 544 |
| FB13 | 75 043 | 13 | 316 232 | 5 908 | 23 733 |

### 3.2 链接预测

链接预测的任务是给定一个由实体和关系组成的事实,去猜测三元组中缺失的实体。具体来说,如果($h$,$r$,)中的$t$是缺失的,那么$t$将会被预测,或者在三元组中(,$r$,$t$)中的$h$是缺失的,那么$h$将会被预测。该任务强调的是候选实体依据损失值排名的清单,而不仅是找到一个最匹配的实体。在该任务中,选择的数据集为 WN18 和 FB15K。

1)评价指标。在链接预测的实验任务中,由于要与 TransE 使用的数据集进行数据的分析与对比,因此,本文选择了与 TransE 相同的评价指标。具体来说,该评测任务中包含了两个评价指标:正确实体的平均排名(*MeanRank*)和正确实体排在前 10 名的概率(*Hits*@10)[25]。经过分析可知:*MeanRank* 的数值越低,以及 *Hits*@10 概率越高,那么在实验层面上反映预测的效果越好。对于每一个三元组($h$ , $r$ , $t$),如果知识库中的尾实体被一个 $x$ 实体所替代,同时用评分函数计算替换后的三元组($h$ , $r$ , $x$)的得分,然后,按照从小到大的顺序对得到的分数进行排列。同理,得到替换头实体的三元组的得分,并依据得分从小到大对分数进行排列;最后,就得到了原来正确的三元组的排名。考虑到知识库中存在一些损坏的三元组,可能会对实验的结果造成影响。因此,在实验进行之前,就把存在于训练集、验证集、测试集里的这些有"干扰"的三元组剔除了。把经过这种剔除处理的设置称为"Filt",未经上述处理的实验设置将被定义为"Raw"。

2)实验实现。选择了现有的几种模型进行实验效果的对比,主要包括 Unstructured、距离模型(structured embedding,SE)、语义能量匹配模型(semantic matching energy,SME)、潜变量模型(latent factor model,LFM)、TransE、TransH。考虑到各个模型在参数设置上变化比较大,在复现的时候没有得到文献中最佳的数据。由于所使用的数据集相同,直接选择了每种模型文献里最好的数据,以此作为本文实验对比的依据。此外,因为实验存在随机误差,为了减少这种情况对实验结果造成的不好影响,每组实验都进行 10 次,然后,将记录 10 次的实验结果,再计算平均值作为最终的结果。训练 TransREF 时,在随机梯度下降(stochastic gradient descent,SGD)过程中学习率$\alpha$将会从集合{0.000 1,0.001,0.005,0.01}当中选取、在{0.25,0.5,1.5,2,4,4.5}一组数值中选择边际 $\gamma$、在{50,100,150,200}当中的选择嵌入维度 $k$ 的大小、在{0,0.1,0.5,1,2}的超参数 $\mu$,在(0.000 1,0.001)之间的随机数 $j1$,$j2$,$j3$,以及在{20,75,1 200,4 800,9 600}中的 batch 的大小 $B$。经过实验发现验证集确定了一组最佳参数的组合。

在 unif 设置下,最佳配置为:在 WN18 数据集上,$\mu=0.1$,$\alpha=0.000\ 1$,$\gamma=5$,$k=100$,$B=4\ 800$;在 FB15K 上,

$\mu=0.1,\alpha=0.001,\gamma=4,k=200,B=4\ 800$。在 bern 设置下，最佳配置为：在 WN18 上，$\mu=0.1,\alpha=0.000\ 1,\gamma=5,k=100,B=4\ 800$；在 FB15K 上，$\mu=0.1,\alpha=0.001,\gamma=4,k=200,B=4\ 800$。对于这两个数据集，虽然它们包含的关系数是不同的，但本实验将所有训练三元组迭代训练 500 次。

3)实验结果。在表 3 通过观察发现，在 WN18 数据集上，TransREF(unif)和 TransREF(bern)的 *MeanRank* 比其他方法都要好。这可能是因为 WN18 中存在的关系数量比 FB15K 数据集上的关系少一些，所以将不同类型的关系忽略掉也是合理的。在 FB15K 数据集上，TransREF(unif)和 TransREF(bern)的 *MeanRank* 比其他基线方法都要好。在 *Hits*@10 这一指标上，与 TransE 和 TransH 相比，TransREF 在 WN18 上分别提高了 3.6%和 10.5%，在 FB15K 上分别提高了 22.3%和 5.0%，效果提升明显。

**表 3　链接预测实验结果**

| Date Sets | WN18 | | | | FB15K | | | |
|---|---|---|---|---|---|---|---|---|
| Metric | MeanRank | | Hits@10/% | | MeanRank | | Hits@10/% | |
| | Raw | Filt | Raw | Filt | Raw | Filt | Raw | Filt |
| Unstructured | 315 | 304 | 35.3 | 38.2 | 1074 | 979 | 4.5 | 6.3 |
| SE | 1 011 | 985 | 68.5 | 80.5 | 273 | 162 | 28.8 | 39.8 |
| SME | 545 | 533 | 65.1 | 74.1 | 274 | 154 | 30.7 | 40.8 |
| SME | 526 | 509 | 54.7 | 61.3 | 284 | 158 | 31.3 | 41.3 |
| LFM | 469 | 456 | 71.4 | 81.6 | 283 | 164 | 26.0 | 33.1 |
| TransE | 263 | 251 | 75.4 | 89.2 | 243 | 125 | 34.9 | 47.1 |
| TransH | 401 | 388 | 73.0 | 82.3 | 212 | 87 | 45.7 | 64.4 |
| STransH[26] | 347 | 330 | 77.1 | 90.6 | 196 | 68 | 46.6 | 69.4 |
| TransV | 256 | 212 | 72.6 | 89.9 | 182 | 81 | 42.2 | 63.1 |
| TransE-NA | 210 | 198 | 71.6 | 85.3 | 186 | 82 | 43.7 | 68.6 |
| Context_RL | 268 | 176 | 77.6 | 91.0 | 174 | 58 | 48.2 | 65.6 |
| E-CP[27] | 196 | 168 | 78.4 | 91.8 | 166 | 58 | 46.3 | 66.6 |
| TRKRL | — | — | — | — | 178 | 55 | 50.2 | 68.6 |
| TransREF(unif) | 148 | 135 | 79.6 | **92.8** | 157 | **35** | 47.7 | 68.4 |
| TransREF(bern) | **135** | **124** | **80.2** | 92.4 | **142** | 50 | **50.9** | **69.4** |

为了证实 TransREF 能够更好地处理复杂关系，选择了关系比较多的数据集 FB15K，并利用了在 *Hits*@10 上最优的配置参数，分别测试 1－1、1－*n*、*n*－1、*n*－*n* 关系下的得分，并对得分进行了排序。从表 4 的实验结果可以看出，在 Predicting Left 和 Predicting Right 上，TransREF 模型都达到了最好的效果，优于其他的模型。其中，TransREF 在 Predicting Left 的 1－to－N 关系下达到了 95.0%；在 Predicting Right 的 N－to－1 关系下达到了 94.3%。

### 3.3　三元组分类

三元组分类用于判定一个给定三元组($h$，$r$，$t$)是否在知识图谱中是确实存在的，其首要目的是对一个三元组进行“正确”或“错误”的二元分类，这个任务完成的是否准确关系着知识图谱补全的评价效果的好坏。实验在进行的时候，一个阈值 $\sigma_r$ 将会被设定。在对一个三元组($h$，$r$，$t$)进行计算的时候，如果给定的阈值 $\sigma_r$ 小于计算的得分，那么预测结果是错误的，反之则正确。在进行阈值大小设置时主要是依据最大分类精度，在本次评价分析实验中，我们选择了 WN11、FB13、FB15K 三个公开数据集。其中经过对比发现，WN11 和 FB13 里面的关系数比较少，可以认为是稀疏数据集；FB15K 里的关系数比较多，可以被认为是稠密数据集，具体数据如表 2。

1)评价指标

在三元组分类任务中主要是使用准确率(ACC)作为评价模型的性能指标，ACC 的数值与表示模型在三元组分类效果存在正相关的关系。以下公式为其计算方法：

$$ACC=\frac{T_p+T_n}{N_{pos}+N_{neg}}=10 \tag{10}$$

其中，式(10)预测样本中正确的正例三元组个数用 $T_p$ 指示；预测样本中正确的负例三元组个数指的是 $T_n$；$N_{pos}$ 表示训练集中的正例三元组的数量，$N_{neg}$ 表示的是负例三元组的个数。

2)实验实现

在 SGD 过程中，选择了{0,0.1,0.5,1,2}的超参数 $\mu$，(0.000 1,0.001)之间的随机数 $j1,j2,j3$，集合{0.001,0.01,0.1}中选取学习率 $\alpha$、{1,2,4,4.5,5,10}中确定边际

**表4 FB15K各类关系的Hits@10值**

| Method | Predicting Left(Hits@10/%) | | | | Predicting Right(Hits@10/%) | | | |
|---|---|---|---|---|---|---|---|---|
| | 1-to-1 | 1-to-*N* | *N*-to-1 | *N*-to-*N* | 1-to-1 | 1-to-*N* | *N*-to-1 | *N*-to-*N* |
| Unstructured | 34.5 | 2.5 | 6.1 | 6.6 | 34.3 | 4.2 | 1.9 | 6.6 |
| SE | 35.6 | 62.6 | 17.2 | 37.5 | 34.9 | 14.6 | 68.3 | 41.3 |
| SME | 35.1 | 53.7 | 19.0 | 40.3 | 32.7 | 14.9 | 61.6 | 43.3 |
| SME | 30.9 | 69.6 | 19.9 | 38.6 | 28.2 | 13.1 | 76.0 | 41.8 |
| TransE | 43.7 | 65.7 | 18.2 | 47.2 | 43.7 | 19.7 | 66.7 | 50.0 |
| TransH(unif) | 66.7 | 81.7 | 30.2 | 57.4 | 63.7 | 30.1 | 83.2 | 67.2 |
| TransH(bern) | 66.8 | 87.6 | 28.7 | 64.5 | 65.5 | 39.8 | 83.3 | 67.2 |
| TransV | 79.0 | 91.6 | 37.2 | 63.0 | 77.7 | 47.3 | 90.6 | 66.5 |
| TransE-NA | 85.8 | 93.9 | 42.6 | 64.3 | 84.9 | 53.8 | 91.2 | 62.4 |
| ERDERP[28] | 78.1 | 92.3 | 49.5 | 62.3 | 79.1 | 51.7 | 85.6 | 69.3 |
| TransREF(unif) | 89.3 | 92.6 | **54.3** | 64.8 | 87.9 | **57.5** | 91.5 | 67.3 |
| TransREF(bern) | **92.2** | **95.0** | 51.7 | **66.5** | **92.7** | 49.7 | **94.3** | **69.7** |

$\gamma$的大小,实体向量和关系向量的维度$k$均从{20,50,100,200}中选取,$B$的大小从{50,120,480,960,4 800}中选取。最佳配置的精度由验证集确定。WN11上的最佳配置为:$\mu=0.1$,$\alpha=0.001$,$\gamma=10$,$k=100$,$B=4\ 800$,且在处理过程中可基于$L_1$对相似性度量进行表征;FB13上的最佳配置为:$\mu=0.1$,$\alpha=0.001$,$\gamma=5$,$k=200$,$B=4\ 800$,且在处理过程中可基于$L_1$对相似性度量进行表征;FB15K上的最佳配置为:$\mu=0.1$,$\alpha=0.001$,$\gamma=5$,$k=100$,$B=120$,且在处理过程中可基于$L_1$对相似性度量进行表征。

3)实验结果

表5直观的展示了三元组分类的评估结果。此处结合该表展开分析,不难发现,在WN11上,TransREF最为理想,TransREF模型比TransE和TransH方法好,分别提升了14.1%和11.1%,可见提升效果显著;而在FB13上,TransREF模型的性能表现好于TransE和TransH,分别提高了8.7%和6.4%。在FB15K数据集上,TransREF的性能表现最为优异;相较于TransE和TransH,分别提高了16.1%和8.1%。这表明TransREF无论是在稀疏数据集上,还是在相对稠密的数据集上,TransREF模型都能够适应,并且拥有很好的表现。

为了进一步分析学习率和边际值对实验结果的影响,在这一环节我们引入控制变量法,分别做了学习率($\alpha$)和准确率、边际值($\lambda$)和准确率的实验,具体结果如图3、4所示。

从图3可以看出,在学习率处在[0.1,1]的范围内,TransREF的准确率要高于TransE,说明受学习率的影响,TransREF的准确率上升幅度要大于TransE。从图4可以看出,在边际值处在[0,5]的范围内,TransREF的准确率要高于TransE,并且TransREF更加平稳,说明在边界值影响下,TransREF模型稳定性要高于TransE。

**表5 不同模型的三元组分类精度** %

| Method | WN11 | FB13 | FB15K |
|---|---|---|---|
| SE | 53.0 | 75.2 | — |
| SME | 73.8 | 84.3 | — |
| NTN | 70.4 | 87.1 | 66.5 |
| TransE | 75.8 | 81.5 | 79.7 |
| TransH | 77.7 | 76.5 | 74.2 |
| TransH | 78.8 | 83.8 | 87.7 |
| STransH | 79.6 | 85.2 | 89.6 |
| TransAH[29] | 85.2 | 88.1 | 92.0 |
| TransV | — | — | 91.6 |
| TRKRL | — | — | 88.7 |
| TransE-NA | 87.0 | 86.8 | — |
| Context_RL | — | — | 87.4 |
| TransR* | 86.2 | 81.7 | **97.1** |
| TransREF(unif) | 87.5 | 89.6 | 93.8 |
| TransREF(bern) | **89.9** | **90.2** | 95.8 |

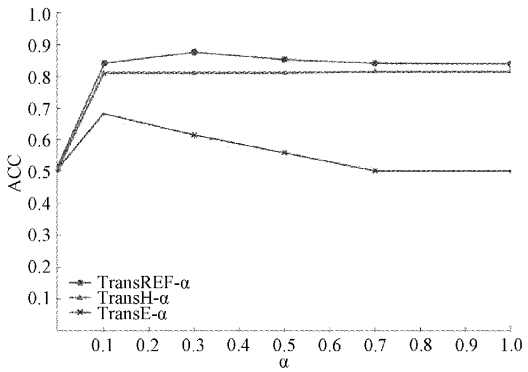


图3 在FB15K上的学习率与准确率实验

最后,为了验证模型的分类能力,本文做了精确率

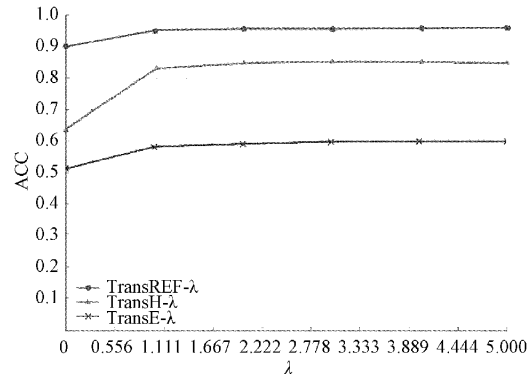


图 4　在 FB15K 上的边际值与准确率实验

(Precision)和召回率(Recall)的实验，选择 FB15K 作为数据集，将最终的结果按照阈值从小到大的顺序进行排列，具体的结果如图 5 所示。

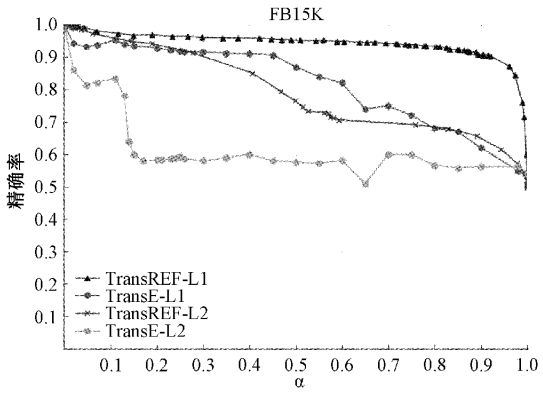


图 5　在 FB15K 上的精确率与召回率的实验

从图 5 可以看出，在采用 $L_1$ 范式后，TransREF 在 Recall 为[0,1]的范围内，Precision 都要高于 TransE，整体下降的幅度非常平缓；在采用了 $L_2$ 范式后，TransREF 在 Recall 为[0,1]的范围内，Precision 都要高于 TransE，总体呈现下降趋势，但是，下降的幅度要小于 TransE。这说明 TransREF 的分类效果要好于 TransE。

## 4　结　　论

本文提出了一种改进的基于邻域信息的知识表示模型，即 TransREF，主要用于克服 TransE 处理复杂关系效率不高的问题。首先引入关系矩阵投影的思想，进行实体和关系的向量化表示；其次，根据五点随机法选择邻节点数；然后，根据实体与邻节点关系的距离确定最佳的实体作为邻域信息，并进行邻节点的实体表示。在模型训练时，对负例三元组的抽样策略进行改进，即利用一对多和多对一的映射关系选择替换实体，使尽可能多的实体得到训练。实体替换过程很复杂，为更好地满足实际应用要求，需要选择与其语义最相似的实体，通过对此实体进行替换，以提高实体之间的区分度。最后，在公开的数据集 FB15K 和 WN18 上进行了链接预测实验，以及在 WN11、FB13、FB15K 数据集上进行了三元组分类的实验，分析和验证了所提方法的有效性。经过一系列实验研究后，大量结果表明，相比于 TransE 和 TransH，TransREF 在 *Hits*@10、ACC 上提升明显，并且分类效果要好于 TransE，可以应用到真实大规模知识图谱的完善和推理应用中。

在未来的研究中，本文计划对 TransREF 模型进行改进，不断提高其性能。从文献中注意到在进行模型训练时，可以加入聚类[30]算法，提高模型的分类性能和推理效果，为其推广应用提供支持。在生成负例三元组时，先将需要训练的实体进行大类小类的划分，然后，再利用概率法进行头尾实体的替换，因为在进行链接预测实验时，实验的结果并不算特别好，这与负例三元组的生成方式有一定的关联。此外，我们并不满足于只做链接预测和三元组分类的实验，在未来的研究中，我们还将研究从文本中提取关系事实的任务，分析稀疏数据集上精确率与召回率之间的关系。

## 参考文献

[1] CHEN X, JIA S, XIANG Y. A review: Knowledge reasoning over knowledge graph[J]. Expert Systems with Applications, 2020, 141: 112948.

[2] WANG Q, MAO Z, WANG B, et al. Knowledge graph embedding: A survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.

[3] BALAžEVIC I, ALLEN C, HOSPEDALES T M. Hypernetwork knowledge graph embeddings [C]. Artificial Neural Networks and Machine Learning-ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, 2019: 553-565.

[4] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]. Advances in Neural Information Processing Systems, 2013.

[5] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2014.

[6] DASGUPTA S S, RAY S N, TALUKDAR P P. HyTE: Hyperplane-based temporally aware knowledge graph embedding[C]. EMNLP. 2018: 2001-2011.

[7] PERSHINA M, YAKOUT M, CHAKRABARTI K. Holistic entity matching across knowledge graphs[C]. 2015 IEEE International Conference on Big Data (Big Data), IEEE, 2015: 1585-1590.

[8] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C].

Proceedings of the AAAI Conference on Artificial Intelligence, 2015.

[9] TAY Y, LUU A, HUI S C. Non-parametric estimation of multiple embeddings for link prediction on dynamic knowledge graphs[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017.

[10] WANG X, XU Y, HE X, et al. Reinforced negative sampling over knowledge graph for recommendation[C]. Proceedings of the Web Conference, 2020: 99-109.

[11] BORDES A, GLOROT X, WESTON J, et al. Joint learning of words and meaning representations for open-text semantic parsing[C]. Artificial Intelligence and Statistics. PMLR, 2012: 127-135.

[12] BORDES A, WESTON J, COLLOBERT R, et al. Learning structured embeddings of knowledge bases[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2011, 25(1): 301-306.

[13] 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: 233-259.

[14] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[J]. Advances in Neural Information Processing Systems, 2013, 26.

[15] JENATTON R, ROUX N, BORDES A, et al. A latent factor model for highly multi-relational data[J]. Advances in Neural Information Processing Systems, 2012, 25.

[16] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embeddings for knowledge graph completion[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2015.

[17] MAO Y, CHEN H. Rule-guided compositional representation learning on knowledge graphs with hierarchical types[J]. Mathematics, 2021, 9(16): 1978.

[18] ZHANG Z, JIA J, WAN Y, et al. TransR*: Representation learning model by flexible translation and relation matrix projection[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(5): 10251-10259.

[19] 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.

[20] WANG Z, LI L, LI Q, et al. Multimodal data enhanced representation learning for knowledge graphs[C]. 2019 International Joint Conference on Neural Networks(IJCNN). IEEE, 2019: 1-8.

[21] 王会勇,论兵,张晓明,等.基于联合知识表示学习的多模态实体对齐[J].控制与决策,2020,35(12):2855-2864,DOI:10.13195/j.kzyjc.2019.0331.

[22] 周泽华,陈恒,李冠宇.基于图上下文的知识表示学习[J].计算机应用与软件,2021,38(6):120-125.

[23] 熊盛武,陈振东,段鹏飞,等.基于可信向量的知识图谱上下文感知表示学习[J].武汉大学学报(理学版),2019,65(5):488-494,DOI:10.14188/j.1671-8836.2019.05.010.

[24] LIU H C, XUE L, LI Z W, et al. Linguistic petri nets based on cloud model theory for knowledge representation and reasoning[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(4): 717-728.

[25] NAYYERI M, CIL G M, VAHDATI S, et al. Trans4E: Link prediction on scholarly knowledge graphs[J]. Neurocomputing, 2021, 461: 530-542.

[26] 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.

[27] 赵博,王宇嘉,倪骥.知识图谱的增强CP分解链接预测方法[J/OL].计算机应用研究:1-7[2023-02-27].DOI:10.19734/j.issn.1001-3695.2022.09.0498.

[28] LIN L, LIU J, GUO F, et al. ERDERP: Entity and relation double embedding on relation hyperplanes and relation projection hyperplanes[J]. Mathematics, 2022, 10(22): 4182.

[29] 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.

[30] 姚佳奇,唐波,刘子怡.基于改进K-means聚类算法的海外欠发达城市配电网规划[J].电子测量技术,2021,44(23):54-60,DOI:10.19651/j.cnki.emt.2107828.

## 作者简介

**王永康**,硕士研究生,主要研究方向为知识表示,知识图谱。

**艾山·吾买尔**(通信作者),博士,教授,主要研究方向为为自然语言处理、机器翻译,语音识别。

**顾亚东**,硕士研究生,主要研究方向为多模态谣言检测。

**何江涛**,硕士研究生,主要研究方向为方面级情感分析。

E-mail:Hasan1479@xju.edu.cn