文章编号：1003-0077(2020)07-0050-10

融合实体描述及类型的知识图谱表示学习方法

杜文倩，李弼程，王 瑞

（华侨大学 计算机科学与技术学院，福建 厦门 361021）

摘 要：知识图谱在很多人工智能领域发挥着越来越重要的作用。知识图谱表示学习旨在将三元组中的实体和关系映射到低维稠密的向量空间。TransE、TransH 和 TransR 等基于翻译操作的表示学习方法，只考虑了知识图谱的三元组信息孤立的学习表示，未能有效利用实体描述、实体类型等重要信息，从而不能很好地处理一对多、多对多等复杂关系。针对这些问题，该文提出了一种融合实体描述及类型的知识图谱表示学习方法。首先，利用 Doc2Vec 模型得到全部实体描述信息的嵌入；其次，对实体的层次类型信息进行表示，得到类型的映射矩阵，结合 Trans 模型的三元组嵌入，得到实体类型信息的表示；最后，对三元组嵌入、实体描述嵌入及实体类型嵌入进行连接操作，得到最终实体嵌入的表示，通过优化损失函数训练模型，在真实数据集上分别通过链接预测和三元组分类两个评测任务进行效果评估，实验结果表明新方法优于 TransE、TransR、DKRL、SimplE 等主流模型。

关键词：人工智能；知识图谱；表示学习；链接预测；三元组分类

中图分类号：TP391 **文献标识码**：A

Representation Learning of Knowledge Graph Integrating Entity Description and Entity Type

DU Wenqian, LI Bicheng, WANG Rui

(School of Computer Science and Technology, Huaqiao University, Xiamen, Fujian 361021, China)

Abstract: Representation learning of knowledge graph aims to project entities and relations into continuous low-dimensional vector space. Most existing translation-based representation methods, such as TransE、TransH and TransR, usually utilize only triples of knowledge graph, and fail to deal with complex relationships such as one-to-many, many-to-one, and many-to-many. To address this issue, this paper proposes a representation learning model of knowledge graph integrating entity description and type, which is called TDT model. Firstly, the Doc2Vec model is used to obtain the embedding of all entity descriptions. Secondly, treating the hierarchical types as projection matrices for entities, the embedding of entity type information can be obtained via multiplying the projection matrix with triple embedding. Finally, TDT model integrates the information of triple(T), entity description(D), and entity type information(T) in a low-dimensional vector space. This paper evaluates TDT model via the experiments of link prediction and triple classification on the real-world datasets. The results show that new method significantly outperforms other baselines, such as TransE, TransR, DKRL and SimplE etc.

Keywords: artificial intelligence; knowledge graph; representation learning; link prediction; triple classification

0 引言

2012 年，谷歌提出了知识图谱的概念，并将其运用到搜索引擎中。之后，大规模知识图谱的构建取得了巨大的进展，涌现出一大批知识图谱，具有代表性的有 YAGO[1-3]、DBpedia[4]、FreeBase[5] 等。目前，知识图谱在很多人工智能应用上发挥着重要的作用，例如，智能问答、信息推荐、网页搜索[6] 等。知识图谱是一个结构化的语义网络，存储着大量的事实三元组

收稿日期：2019-11-18 **定稿日期**：2019-12-30

基金项目：国家社会科学基金(19BXW110)

(头实体,关系,尾实体),通常简化为(h,r,t)。

但是随着知识图谱规模的逐渐扩大,数据类型逐渐多样化,实体与实体之间的关系越来越复杂,传统基于符号和逻辑的方法,由于其计算低效性,使得知识图谱应用面临挑战。

为了解决这个问题,表示学习被提出并得到蓬勃发展。表示学习的目的是将知识图谱三元组中的实体和关系映射到低维稠密的向量空间,将传统基于逻辑和符号的运算转变为基于数值的向量计算。基于能量函数的表示学习模型由于其简单高效性,在链接预测、三元组分类等任务上取得了较好的结果,被广泛应用于知识图谱补全、实体对齐等领域。然而,这些模型大都只考虑了知识图谱的三元组信息,对知识图谱中丰富的文本信息、类型信息融合程度低,融合方式单一,而这些未被充分融合的信息对于降低实体和关系的模糊程度、提高推理预测的准确度至关重要。图 1 列举了 FreeBase 事实三元组中的实体描述和实体类型,上层实线代表此三元组中最重要的层次类型,下层实线代表其对应的实体描述。

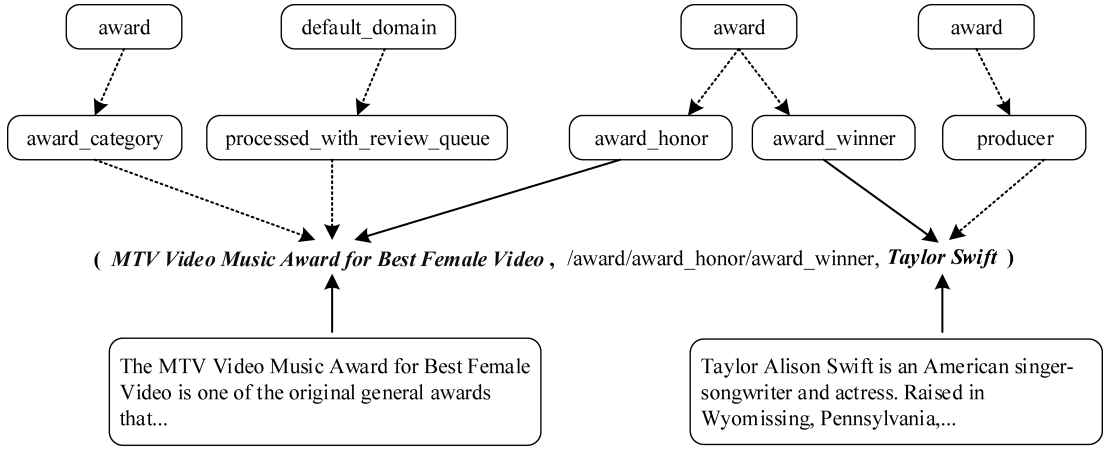

图 1　FreeBase 三元组中的实体描述以及类型举例

因此,本文提出一种融合实体描述及类型的知识图谱表示学习方法,对三元组信息(triple vector)、实体描述信息(description embedding)及实体类型信息(type information)进行融合,简称为 TDT 模型。

1　相关工作

目前表示学习模型主要分为三类:语义匹配模型、距离变换模型、融合多源信息的表示学习模型。

1.1　语义匹配模型

语义匹配模型以 RESCAL 模型[7]为代表,它将知识图谱编码为一个张量,若三元组存在于知识图谱中,则对应张量中的值设置为 1,反之为 0。但是 RESCAL 模型需要大量的参数,计算效率低。Kazemi 等人[8]提出 SimplE 模型,独立学习每个实体的两个嵌入,并且其复杂度随着嵌入的维度线性增长。Yang 等人[9]提出 DISTMULT 模型,通过双线性对角模型学习实体和关系的向量表示。Trouillon 等人[10]提出 ComplEx 模型,通过使用元素点积使 DISTMULT 模型通用化。Liu 等人[11]提出 Analogy 模型,通过实体和关系嵌入的隐含表示,建立类比关系,以可微的方式优化目标,实现计算上的可扩展性。Zhang 等人[12]提出一种贪婪算法,经过过滤器和预测器的增强,在空间中高效地进行搜索,使得实体之间的关系链接更加合理。

1.2　距离变换模型

TransE 模型[13]是这一类方法的典型代表,它将三元组中的关系看作头实体和尾实体之间的翻译操作,基本假设是成立的事实三元组(h,r,t)应该满足等式 $\boldsymbol{h}+\boldsymbol{r}=\boldsymbol{t}$。对于每个三元组$(h,r,t)$,定义能量函数,如式(1)所示 。

$$E(h,r,t)=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\| \tag{1}$$

其中 $\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}$ 分别代表头实体、关系、尾实体的嵌入。

TransE 模型在处理一对一类型的关系上很有效,但是在处理一对多、多对一以及多对多关系时存在一定的问题。为了解决 TransE 模型中的问题,TransH 模型[14]和 TransR 模型[15]在计算事实得分的时候,将同样的实体在不同的关系上使用不同的

表示参与计算,这样就避免了直接计算所带来的表示趋向一致的问题。Wen 等人[16]提出了 m-TransH 模型,直接建模多元关系,将多元关系中的每一元看成一个角色,建模为角色函数的累加。CTransR 模型[15]是 TransR 模型的一个改进方法,它考虑同一种关系在不同实体对之间表示不同的含义,因此它将同一个关系的实体聚类成若干子类别,在每一个实体中,单独学习一个关系的向量表示。

TransR 和 TransH 模型也只是部分地解决了 TransE 模型存在的问题。Ji 等人[17]提出了 TransD 模型,通过让转换矩阵由相应的实体和关系共同决定,旨在解决实体和关系多样性的问题。随后,他们又提出了 TranSparse 模型[18],使用稀疏矩阵作为关系的转换矩阵,根据关系的复杂程度调整稀疏矩阵的稀疏度,使得不同复杂度的关系由不同自由度的模型进行学习,以此解决数据不均衡的问题。为了克服关系之间的异质性,Ji 等人[18]提出了 TranSparse(share)模型,转移矩阵的稀疏度由关系连接的实体对数目决定,头、尾实体共享转移矩阵。考虑到头尾实体的不均衡问题,Ji 等人[18]提出了 TranSparse(seperate)模型,为每个关系定义两个单独的稀疏转移矩阵,头、尾实体各一个矩阵,稀疏度由头、尾实体数目决定。考虑时间因素,Jia 等人[19]提出了 TransA 模型,将边缘参数分为实体和关系指定的局部边缘的加权和,有效解决了损失函数的边缘参数在一个给定的候选闭集中选择造成的局部有效性问题。Fan 等人[20]考虑不同关系映射属性三元组的贡献差别,提出 TransM 模型,直接根据关系的映射属性预先计算每个训练三元组的权重,用于加权损失函数。TransE 等模型基于平移的方法忽略关系的语义,只为一个关系分配唯一的一个平移向量,不能处理关系多语义问题。对此,Xiao 等人[21]提出贝叶斯非参数无限混合表示模型 TransG,利用特定关系的关系向量的混合来进行知识表示,每个分量代表一个特定的潜在语义。

目前基于翻译操作的表示学习方法仅考虑知识图谱中的三元组结构信息,尚有大量与实体和关系有关的其他信息没有得到有效应用,造成实体、关系的语义表示不明确。

1.3 融合多源信息的表示学习模型

在融合多源信息进行知识表示学习方面,已经有一些研究工作。Wang 等人[22]提出在表示学习中考虑文本数据,利用 Word2Vec 学习维基百科正文中的词表示,利用 TransE 模型学习知识图谱中的知识表示。然后利用维基百科正文中的链接信息,让文本中实体对应的词表示与知识图谱中的实体尽可能接近,从而实现文本与知识图谱融合的表示学习。Zhong 等人[23]还将类似的思想用于融合实体描述信息。Guo 等人[24]提出了 SSE(semantically smooth embedding),学习知识图谱实体和关系的表示。SSE 将实体语义类信息作为损失函数的正则项,强制表示空间几何结构语义平滑。Xie 等人[25]提出基于实体描述的表示学习模型(description-embodied knowledge representation learning, DKRL)。DKRL 模型提出在知识表示学习中考虑 FreeBase 等知识图谱中提供的实体描述文本信息,并对其通过卷积神经网络或者连续词袋模型进行编码,然后与 Trans 模型嵌入得到的向量进行拼接,得到最终的实体嵌入。此外,Xu 等人[26]提出基于 Bi-LSTM 编码的 A-LSTM 模型,对实体文本信息进行表示。Nguyen 等人[27]提出了 TransE-NMM 模型,在 TransE 模型的基础上,引入邻居实体信息进行实体和关系的表示学习。对于每个三元组(h,r,t),增加逆关系三元组(t,r^{-1},h),由此得到实体的邻居实体为以该实体为尾实体的所有头实体-关系对,邻居向量则表示为头实体向量加上关系向量的带权和。通过邻居向量与实体向量相加则得到最终实体的表示,从而进行优化。然而,在实际知识图谱中,很多实体缺失实体描述。因此,Wang 等人[28]提出了 TEKE(text-enhanced knowledge embedding)模型,引入文本语料中的实体语义结构,学习实体的表示。Wu 等人[29]扩展了 TransE 模型,融合三元组属性信息,提出 TransEA 模型。An 等人[30]针对实体关系表示的歧义问题,提出一种精确的文本增强(accurate text-enhanced, ATE)方法,通过对关系提及与实体描述使用注意力机制获取增强的文本信息表示,从而使得不同的三元组有不同表示。除了文本信息,Krompaß 等人[31]认为,实体类型信息对于知识图谱来说是一个隐藏的限制变量。Xie 等人[32]提出了融合层次类型的表示学习方法(type-embodied knowledge representation learning, TKRL),学习知识图谱实体和关系的表示,将层级类型信息用于映射矩阵,结合翻译模型得到的嵌入,在知识图谱补全、三元组分类等任务上取得了较以往更好的结果。Xie 等人[33]提出融合图像信息的表示学习方法(image-embodied knowledge representation learning, IKRL),将图像

信息映射到向量空间，提高知识图谱补全的精度。在跨知识图谱表示方面，Cai 等人[34]提出了跨知识图谱表示方法 Cross-KG，同时学习两个不同知识图谱的表示，映射语义相关的两个知识图谱中的实体和关系到统一的向量空间，提升链接效果。也有一些工作将知识图谱看作图数据结构，融合结构信息，丰富实体和关系的语义信息。Feng 等人[35]提出了图感知的知识表示方法（graph aware knowledge embedding，GAKE），利用知识图谱的图结构信息，即邻居上下文、路径上下文和边上下文学习实体和关系的向量表示。

在多源信息融合的表示学习方面，主要是考虑实体描述的知识表示学习模型，以及文本与知识库融合的表示学习，这些模型的信息来源及融合手段都非常有限。此外，知识图谱中的实体分布呈现长尾分布现象，部分实体在异构的数据源中不具有相应的描述文本。而实体类型作为隐藏变量，可以作为文本的补充信息，丰富实体和关系的语义。因此，本文将实体描述文本信息以及实体类型信息进行融合，构建知识图谱表示学习模型。

本文的主要贡献为：

（1）提出一个整合的知识图谱表示学习模型，融合知识图谱的三元组信息、实体描述信息及实体类型信息，从而降低实体和关系的模糊程度。

（2）考虑实体描述的全部语义信息，利用 Doc2Vec 模型进行描述信息的表示；三元组实体存在多种类型，并且类型存在层次性，将层次类型信息进行表示，结合翻译模型嵌入进行拼接，训练一个表示学习模型以提高知识图谱应用的性能。

（3）本文在真实数据集上进行实验，评估链接预测和三元组分类任务，取得了较好的效果。

2　融合实体描述及类型的表示学习方法

2.1　模型框架

TDT 模型融合三部分的信息：三元组信息、实体描述信息以及实体类型信息，将三者结合进行知识图谱的嵌入。最终实体嵌入如式(2)所示。

$$\boldsymbol{e}=\boldsymbol{e}_s \oplus \boldsymbol{e}_d \oplus \boldsymbol{e}_t \tag{2}$$

其中，$\boldsymbol{e}_s$、$\boldsymbol{e}_d$ 和 $\boldsymbol{e}_t$ 分别为三元组向量、实体描述向量及实体类型向量表示，对应三元组信息的嵌入、实体描述的嵌入以及实体类型的嵌入。⊕为连接操作符，$\boldsymbol{e}=\boldsymbol{e}_s \oplus \boldsymbol{e}_d \oplus \boldsymbol{e}_t$ 即为 $\boldsymbol{e}=[\boldsymbol{e}_s||\boldsymbol{e}_d||\boldsymbol{e}_t]$。

图 2 为 TDT 的整体框架，$\boldsymbol{e}_s$、$\boldsymbol{e}_d$ 和 $\boldsymbol{e}_t$ 需要依次计算。

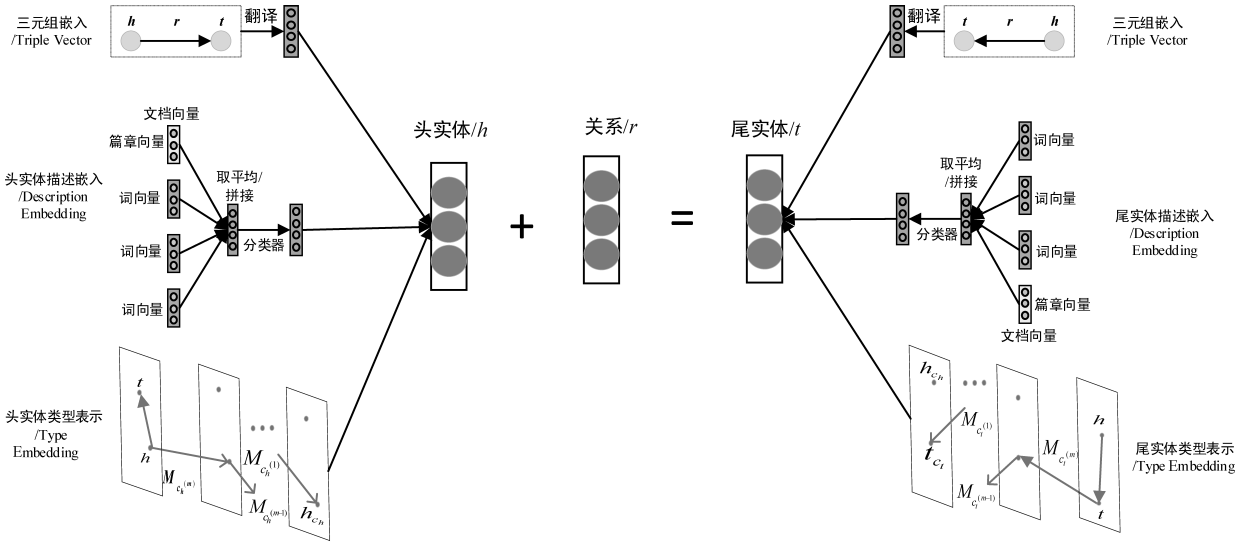

图 2　TDT 模型整体框架

首先，利用 Trans 模型获得三元组实体的嵌入，将三元组中的关系当作头实体和尾实体间翻译操作，得到每个三元组实体和关系的数值向量表示。

其次，通过 Trans 模型得到的实体嵌入，与实体层次类型映射矩阵结合，得到三元组实体类型的嵌入。

然后，考虑知识图谱中三元组实体描述的全部文本信息，采用 Doc2Vec 模型，对实体描述的文本信息进行嵌入。

最后，将所有的表示向量进行连接，得到最终的

三元组实体向量，按照式(1)来优化训练TDT模型。

三元组嵌入主要通过翻译操作得到。接下来详细介绍实体描述的嵌入、类型的表示及模型训练。

2.2 实体描述的嵌入

在这一部分主要讨论实体描述嵌入的实现。DKRL模型是经典的融合实体描述信息进行表示学习的一个模型。但是DKRL模型对实体描述信息进行关键词抽取，得到这些关键词的嵌入，训练模型，造成了一定语义信息的损失。因此，本文利用Doc2Vec模型，对实体描述进行表示。

Doc2Vec模型可以训练出一个文档在低维向量空间中的分布式表示，得到的文档向量包含整个文档的全部语义信息。Doc2Vec模型在词向量表示模型Word2Vec的输入中加入预定义的文档向量，对文档向量的添加使得模型在训练时考虑整个文档的语义信息。

首先，随机生成N维的文档向量$\boldsymbol{x}^{\text{paragraph-id}}$和$N$维文档中每个词语的独热(one-hot)形式的词向量$\boldsymbol{x}^{i-m,\cdots,i+m}$，其中$i$是指由上下文预测的当前中心词的标号，$m$是指窗口大小。

然后，对N维的文档向量和词向量进行降维，如式(3)所示。

$$\boldsymbol{v}_{i-m}=\boldsymbol{V}\boldsymbol{x}^{i-m},\boldsymbol{v}_{i-m+1}=\boldsymbol{V}\boldsymbol{x}^{i-m+1},\cdots,\boldsymbol{v}_{i+m}=\boldsymbol{V}\boldsymbol{x}^{i+m},\ \boldsymbol{v}^{\text{paragraph-id}}=\boldsymbol{V}\boldsymbol{x}^{\text{paragraph-id}} \tag{3}$$

其中，$\boldsymbol{V}$是一个n行N列的单位矩阵，n远小于N。文档向量和词向量降为n维。

通过词向量和文档向量可以得到中心词向量$\boldsymbol{y}_i$，如式(4)所示。

$$\boldsymbol{y}_i=\boldsymbol{U}\times\frac{\boldsymbol{v}_{i-m}+\boldsymbol{v}_{i-m+1}+\cdots+\boldsymbol{v}_{i+m}+\boldsymbol{v}^{\text{paragraph-id}}}{2m+1} \tag{4}$$

其中，$\boldsymbol{U}$为一个N行n列的单位矩阵，进一步将中心词向量通过softmax函数进行归一化，如式(5)所示。

$$\hat{y}_i=\text{softmax}(y_i)=\frac{e^{y_i}}{\sum_i e^{y_i}} \tag{5}$$

最后，优化目标函数，如式(6)所示。

$$\text{Loss}=-\sum_{i=1}^{n}y_i\log\hat{y}_i \tag{6}$$

使用随机梯度下降的优化方法，最小化目标函数，更新并输出向量，即得到实体描述的嵌入。

2.3 实体类型的表示

知识图谱中的大部分实体都有不止一种类型，这对于实体表示是一种丰富的信息。而且，不同的实体在不同的类型下有着不同的表示。本文对知识图谱中的实体类型信息进行表示，从而提高知识表示的语义。

实体类型具有层次化，图1简单列举了一个有层次类型的例子。因此，首先需要对实体类型下的实体进行映射；然后，在1-N、N-1以及N-N的复杂关系模式下，实体具有不同表示，为了更好地进行复杂关系预测，需要对特定关系下的实体进行映射，最后得到融合层次类型的知识表示。

设n为实体e的所有实体类型数，对于每一个实体类型c，c_i代表实体e属于的第i个类型，$\boldsymbol{M}_{c_i}$为c_i的映射矩阵，α_i为c_i对应的权重。α_i可通过实体e属于c_i的频率得到。本文设置α_i的值相同。对于特定的三元组(h,r,t)，头实体映射矩阵的计算公式如式(7)所示。

$$\boldsymbol{M}_{rh}=\frac{\sum_{i=1}^{n}\alpha_i\boldsymbol{M}_{c_i}}{\sum_{i=1}^{n}\alpha_i},\quad \alpha_i=\begin{cases}1, & c_i\in C_{rh}\\ 0, & c_i\notin C_{rh}\end{cases} \tag{7}$$

其中，C_{rh}代表给定的关系r下，头实体的关系类型集合。

$$\boldsymbol{M}_{rt}=\frac{\sum_{i=1}^{n}\alpha_i\boldsymbol{M}_{c_i}}{\sum_{i=1}^{n}\alpha_i},\quad \alpha_i=\begin{cases}1, & c_i\in C_{rt}\\ 0, & c_i\notin C_{rt}\end{cases} \tag{8}$$

同理，C_{rt}为给定关系r下，尾实体的关系类型集合。$\boldsymbol{M}_c$是类型c的投影矩阵。

然后，在投影过程中，实体(比如Taylor Swift)首先被映射到更一般的子类型空间(比如award)，然后被映射到更精确的子类型空间(比如award/award_winner)。$\boldsymbol{M}_c$被定义如式(9)所示。

$$\boldsymbol{M}_c=\prod_{i=1}^{m}\boldsymbol{M}_{c^{(i)}}=\boldsymbol{M}_{c^{(1)}}\boldsymbol{M}_{c^{(2)}}\cdots\boldsymbol{M}_{c^{(m)}} \tag{9}$$

其中，m是层次类型的层数，$\boldsymbol{M}_{c^{(i)}}$表示第i个子类型$c^{(i)}$的映射矩阵。

最后，将$\boldsymbol{M}_{rh}$、$\boldsymbol{M}_{rt}$与TransE模型得到的三元组实体嵌入相乘得到实体类型的嵌入。

2.4 模型训练

给定知识图谱(knowledge graph，KG)，KG中存在的三元组集合为$T=\{(h,r,t)\mid h,t\in E,r\in$

$R\}$，其中 E 为实体的集合，R 为关系的集合。设置 TDT 模型参数为 $\theta=\{E,R,\boldsymbol{X},\boldsymbol{M}\}$，其中，$\boldsymbol{X}$ 为实体描述的嵌入，$\boldsymbol{M}$ 为所有子类型的映射矩阵。

定义基于间隔的排序损失函数 L，通过最小化损失函数优化模型，如式(10)所示。

$$L=\sum_{(h,r,t)\in T}\sum_{(h',r',t')\in T'}\max(\gamma+d(h+r,t)-d(h'+r',t'),0) \tag{10}$$

其中，γ 为超参数，衡量正确三元组和错误三元组的边界。

$$T'=\{(h',r,t)\mid h'\in E\}\cup\{(h,r',t)\mid r'\in R\}\cup\{(h,r,t')\mid t'\in E\} \tag{11}$$

其中，T 为正例三元组集合，T' 为负例三元组集合，通过随机替换正例三元组的头实体或者尾实体或者关系得到，如式(11)所示。$d(h+r,t)$ 为 $\boldsymbol{h}+\boldsymbol{r}$ 和 $\boldsymbol{t}$ 的距离度量，如式(12)所示。

$$d(h+r,t)=||\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}|| \tag{12}$$

由于 TDT 模型包括三个部分的内容，因此需要从三个方面进行初始化：①三元组的实体描述的表示 $\boldsymbol{X}$ 可以通过 Doc2Vec 模型得到描述的嵌入；②知识图谱三元组的嵌入 $\boldsymbol{E},\boldsymbol{R}$ 可以通过 TransE 模型得到；③可以通过三元组的嵌入 $\boldsymbol{E}$ 与映射矩阵 $\boldsymbol{M}$ 得到实体类型信息的嵌入。所有的初始化向量通过式(2)组合成最终的模型的初始向量。

初始化之后，使用随机梯度下降方法进行优化。TDT 模型在训练的过程中按式(10)最小化损失函数。

3　实验结果与分析

3.1　实验数据集

本文中运用 FB15K 标准数据集去评估 TDT 模型的性能，FB15K[13] 是从大规模知识图谱 Free-Base[5] 抽取得到的，实验中将其分为训练数据集、测试数据集及验证数据集。

在 FB15K 中，每个实体描述平均包括 69 个词，最多包括 343 个词；在实体类型方面，每个实体最少包含一个类型，平均实体类型数为 8。训练数据集有 472 860 个三元组及 1 341 种关系；有效数据集有 48 991 个三元组，测试数据集有 57 803 个三元组。表 1 为 FB15K 的数据统计。

表 1　FB15K 数据统计

DataSet	# Rel	# Train	# Valid	# Test
FB15K	1 341	472 860	48 991	57 803

3.2　测试模型和参数设置

本文定义：基于 TDT 模型并严格遵守式(2)的嵌入形式，称为完全 TDT 模型。但是，TDT 模型有很多可选变量。当 $\boldsymbol{e}_d=0$ 时，$\boldsymbol{e}=\boldsymbol{e}_s\oplus\boldsymbol{e}_t$；当 $\boldsymbol{e}_t=0$ 时，$\boldsymbol{e}=\boldsymbol{e}_s\oplus\boldsymbol{e}_d$，为基于 TDT 的非完全嵌入模型，被称为非完全 TDT 模型。实验设置如下：

(1) TDTE(TransE＋Description＋Type)：使用 TransE 模型得到三元组实体嵌入，与映射矩阵结合，得到融合实体类型的表示，通过 Doc2Vec 得到实体描述向量表示。

(2) TDTEWT(TransE＋Description)：通过 TransE 模型得到三元组实体嵌入，通过 Doc2Vec 模型得到实体嵌入的表示。

(3) TDTEWD(TransE＋Type)：通过 TransE 模型得到三元组实体表示，与映射矩阵结合，得到融合实体类型信息的表示。

(4) TDTR(TransR＋Description＋Type)：使用 TransR 模型得到三元组实体嵌入，与映射矩阵结合，得到融合实体类型的表示，通过 Doc2Vec 模型得到实体描述向量表示。

(5) TDTRWT(TransR＋Description)：通过 TransR 模型得到三元组实体嵌入，通过 Doc2Vec 模型得到实体嵌入的表示。

(6) TDTRWD(TransR＋Type)：通过 TransR 模型得到三元组实体表示，与映射矩阵结合，得到融合实体类型信息的表示。

对于这些模型的训练，设置参数：三元组嵌入维度 n_{tr}、实体描述向量维度 n_{ds} 及实体类型向量维度 n_{ty} 的取值集合为 $\{50,100,200\}$。依照大多数基于翻译操作的模型，设置学习率 λ 取值集合为 $\{0.000\,5,0.001,0.002\}$，边界 γ 的取值集合为 $\{1.0,2.0\}$。在以上 6 组实验中，首先设置参数：$\lambda=0.001$，$\gamma=1.0$，$n_{\mathrm{tr}}=100$，$n_{\mathrm{ds}}=100$，$n_{\mathrm{ty}}=100$，即：完全 TDT 模型的向量(包含三元组嵌入、实体描述嵌入、实体类型嵌入维度)$n=n_{\mathrm{tr}}+n_{\mathrm{ds}}+n_{\mathrm{ty}}=300$。对于可选择的向量嵌入，当选择三元组和实体描述进

行融合时，向量维度 $n=n_{tr}+n_{ds}=200$，当选择三元组和实体类型进行融合时，向量维度 $n=n_{tr}+n_{ty}=200$。

3.3 链接预测

链接预测是知识图谱补全的一个子任务，旨在通过最小化得分函数对给定的三元组(h,r,t)预测丢失的 h 或者 t。这个任务对一系列候选实体进行排名，而不是从知识图谱中给出一个最佳答案。在测试中，对于给定的三元组(h,r,t)，本文用知识图谱实体集中的全部实体随机替换三元组中头实体或者尾实体，然后按照得分函数递减进行排序。

在评测任务中，本文选择翻译模型的两个评价标准：①正确三元组或关系的平均排名(MeanRank，MR)；②对于实体来说，正确答案排名在前 10 的概率 Hits@10。较低的 MR 或者较高的 Hits@10 是一个较好的实验评价结果。此外，被随机替换掉头实体或尾实体之后的三元组可能也存在于知识图谱中，在评估过程中可能会被估计不足。因此，本文遵循两个设置：按照是否在排序前过滤掉这些被替换但是正确的三元组，可以将评测任务分为“Raw”和“Filter”。

本文对 TransE[①]、TransR[①]、DKRL[②]、TKRL[③]进行了复现实验，并在此 TransE、TransR 的基础上扩展融合实体描述及类型的表示学习模型。实体链接预测评价结果如表 2 所示，每个指标下最好的两个结果加粗显示。

表 2 链接预测实验结果

方法	Mean Rank		Hits@10/%	
	Raw	Filter	Raw	Filter
TransE	270	171	46.3	54.2
TransR	227	153	48.0	53.8
DKRL	181	90	49.6	60.3
TKRL	203	126	48.5	56.6
TDTEWD	222	137	47.8	55.7
TDTEWT	141	89	54.6	64.1
TDTE	**127**	70	**66.6**	**73.6**
TDTRWD	152	113	49.7	55.1
TDTRWT	146	**68**	54.9	67.2
TDTR	**113**	**61**	**66.9**	**74.5**

从表 2 的实验结果，可以得出以下结论，均以“Filter”为准：

(1) 本文提出的融合实体描述及实体类型的表示学习方法在链接预测上取得了较好的结果。TDTE、TDTR 在 MR 和 Hits@10 指标下，结果均好于 TransE 和 TransR，总体来看，TDTR 是最好的表示学习模型。TDTR 在 MR 下指标更低，相较于 TransR 降低了 92；Hits@10 指标更高，相较于 TransR 提高了 20.7%。这个对比结果表明，实体文本描述以及实体类型均可以在一定程度上提高链接预测的精度。

(2) 在所有基于 TDT 的模型中，完全 TDT 模型显著优于非完全 TDT 模型。以基于 TransE 的 TDTE、TDTEWD、TDTEWT 为例，TDTE 在 MR 指标上分别降低了 67 和 19；在 Hits@10 指标上分别提升了 17.9%和 9.5%，这意味着实体描述和实体类型对于提高知识图谱补全的精度发挥一定的作用。

(3) 在基于 TDT 的非完全表示学习模型中，实体描述相较于实体类型是一个更好的可选因素。TDTEWT 以及 TDTRWT 在两个指标下的表现均优于 TDTEWD 和 TDTRWD，这表明在链接预测任务中，实体描述是主要的作用因素。

(4) 融合实体全部描述信息可提高实体的语义表示。利用 Doc2Vec 模型表示全部实体描述的 TDTEWT 相较于仅用部分关键词做融合的 DKRL 取得了较好的结果，在 Hits@10 指标下提升了 3.8%，说明完善三元组信息对于预测来说是必要的。

为了进一步挖掘分析 FB15K 上不同关系不同映射类型的相应结果，在 Hits@10 任务下预测头实体和尾实体，具体数值如图 3 所示，每种关系类别下最好的结果加粗显示。从图 3 可以看出，对于复杂关系 1-N、N-N 关系类别，本文提出的 TDTR 模型在预测头实体实验中的表现优于所对比模型；TDTE 在 1-N 关系类别下，预测尾实体实验中的表现较好。对于复杂关系 N-1 关系类别，TDTR 模型在预测尾实体的实验中，效果最好。这表明融合实体描述及类型可降低实体表示的模糊程度，有助于建模复杂关系。

① https://github.com/thunlp/TensorFlow-TransX
② https://github.com/xrb92/DKRL
③ https://github.com/thunlp/TKRL

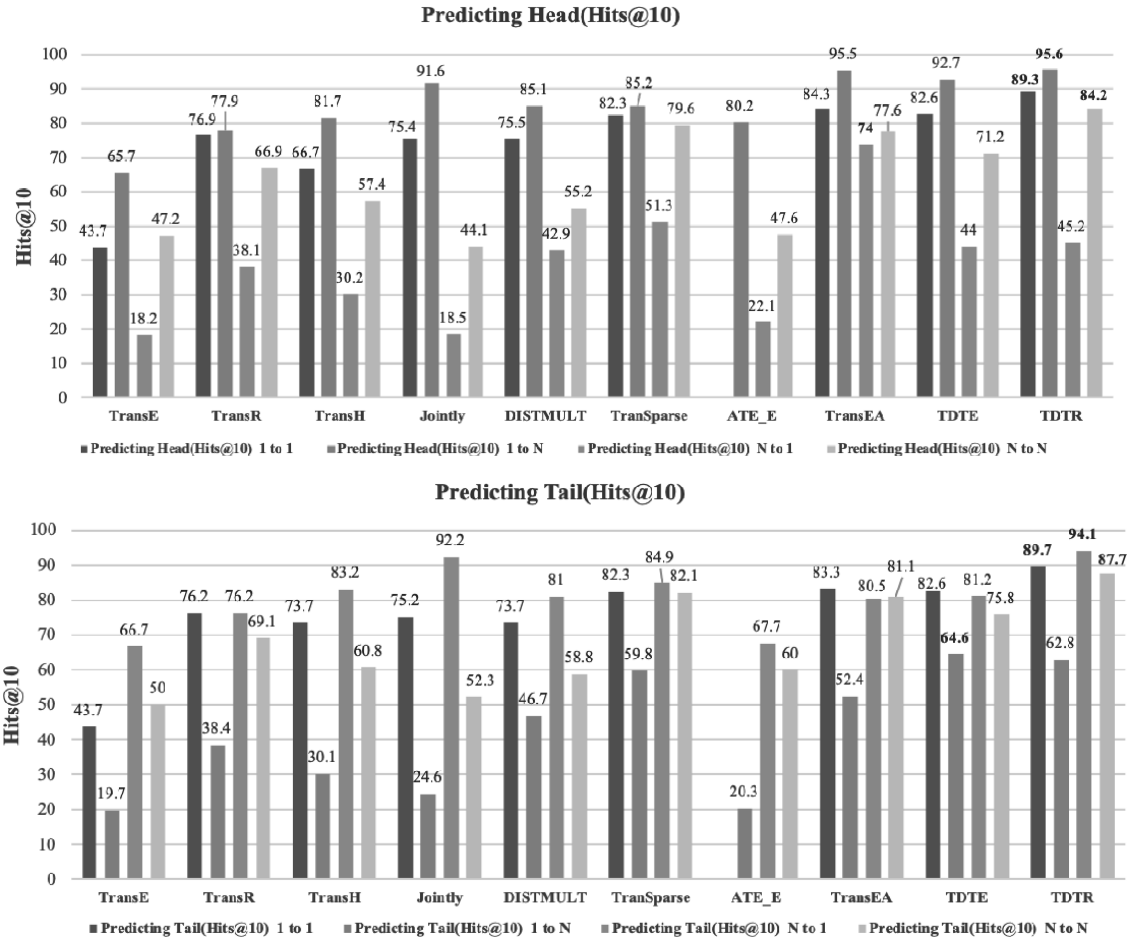

图 3　Hits@10 在不同关系类型下实体映射的实验结果

3.4　三元组分类

三元组分类是判断给定三元组是否正确的一个二分类任务。在实验中使用 FB15K 数据集，采用和链接预测一样的方式构建负例三元组。分类策略如下：为每个关系设置不同的特定关系阈值 σ，对于一个三元组 (h,r,t)，如果 $d(h+r,t)$ 得分小于 σ，则认为这个三元组预测正确。三元组分类的实验结果如表 3 所示，DISMULT、Complex、Analogy、SimplE、AutoKGE 的结果来源于文献[12]。

从表 3 中可以看出，TDTE 和 TDTR 相较于对比实验模型取得了较好的效果。与 TransE 相比，TDTE 提高了 4.6%；TDTR 比 TransR 准确率提高了 5.3%。由此表明，融合实体描述和类型可提高实体表示的语义信息；同时，相较于所对比的其他模型，TDTE 与 TDTR 取得了较好的效果，显示出其在三元组分类任务中的优势。

表 3　三元组分类实验对比结果

方法	准确率/%
TransE	82.6
TransR	83.4
DKRL	86.3
TKRL	85.7
DISMULT	80.8
Complex	81.8
Analogy	82.1
SimplE	81.5
AutoKGE	82.7
TDTE	**87.2**
TDTR	**88.7**

4 结论

传统的基于翻译操作的表示学习模型存在无法有效处理复杂关系和忽略知识图谱实体多源信息等缺陷,导致知识表示语义不完善。针对这个问题,本文提出融合实体描述和类型的表示学习模型,称为TDT模型,同时学习三元组信息、实体描述以及实体类型信息。为了验证方法的有效性,我们在FB15K公开数据集上对链接预测和三元组分类这两项任务进行实验。与现有的表示学习方法进行对比,结果表明TDT模型取得了性能的提升。本文只针对实体进行信息融合,下一步将针对关系描述对模型进行优化。此外,对关系进行组合形成实体之间的路径,建模路径可以提供更加精确的约束信息,实现多步推理,这也将是我们接下来的研究方向。

参考文献

[1] Suchanek F M, Kasneci G, Weikum G. Yago: A core of semantic knowledge[C]//Proceedings of the 16th International Conference on World Wide Web. New York: ACM, 2007: 697-706.

[2] Hoffart J, Suchanek F M, Berberich K, et al. Yago2: A spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial Intelligence, 2013, 194: 28-61.

[3] Mahdisoltani F, Biega J, Suchanek F. Yago3: A knowledge base from multilingual Wikipedias[C]//Proceedings of the 7th Biennial Conference on Innovative Data Systems Research. Asilomar: CIDR, 2015. http: //cidrdb. org/cidr2015/Papers/CIDR15 \ _ Paper1.pdf.

[4] Auer S, Bizer C, Kobilarov G,et al. DBpedia: A nucleus for a Web of open data[C]//Proceedings of the 6th International Semantic Web Conference, 2007: 722-735.

[5] 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. New York: ACM Press, 2008: 1247-1250.

[6] Szumlanski S, Gomez F. Automatically acquiring a semantic network of related concepts[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010: 19-28.

[7] Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data[C]//Proceedings of the 28th International Conference on Machine Learning, 2011: 809 - 816.

[8] Kazemi S M, Poole D. Simple embedding for link prediction in knowledge graphs[C]//Proceedings of Advances in Neural Information Processing Systems, 2018: 4284-4295.

[9] Yang B, Yih W, He X, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint arXiv: 1412.6575, 2014.

[10] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]//Proceedings of the International Conference on Machine Learning, 2016: 2071-2080.

[11] Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR.org, 2017: 2168-2178.

[12] Zhang Y, Yao Q, Dai W, et al. AutoKGE: Searchingscoring functions for knowledge graph embedding[J]. arXiv preprint arXiv: 1904.11682, 2019.

[13] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]//Proceedings of Advances in Neural Information Processing Systems, 2013: 2787 –2795.

[14] 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, 2014: 1112-1119.

[15] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015: 2181-2187.

[16] Wen J, Li J, Mao Y, et al. On the representation and embedding of knowledge bases beyond binary relations[J]. arXiv preprint arXiv: 1604.08642, 2016.

[17] Ji G, He S, Xu L, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the International Joint Conference on Natural Language Processing, 2015: 687-696.

[18] Ji G, Liu K, He S, et al. Knowledge graph completion with adaptive sparse transfer matrix[C]//Proceedings of the 30th AAAI, 2016: 985-991.

[19] Jia Y, Wang Y, Lin H, et al. Locally adaptive translation for knowledge graph embedding[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016: 992-998.

[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, 2014: 328-337.

[21] Xiao H, Huang M, Hao Y, et al. TransG: A generative mixture model for knowledge graph embedding [J]. arXiv Preprint arXiv: 1509.05488, 2015.

[22] Zhen Wang, Jianwen Zhang, Jianlin Feng, et al. Knowledge graph and text jointly embedding [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1591 - 1601.

[23] Zhong H, Zhang J, Wang Z, et al. Aligning knowledge and text embeddings by entity descriptions [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 267-272.

[24] Guo S, Wang Q, Wang B, et al. Semantically smooth knowledge graph embedding [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015: 84-94.

[25] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions [C]//Proceedings of the National Conference on Artificial Intelligence, 2016: 2659-2665.

[26] Xu J, Chen K, Qiu X, et al. Knowledge graph representation with jointly structural and textual encoding [J]. arXiv preprint arXiv: 1611.08661, 2016.

[27] Nguyen DQ, Sirts K, Qu L, et al. Neighborhood mixture model for knowledge base completion [J]. arXiv Preprint arXiv: 1606.06461, 2016.

[28] Wang Z, Li J. Text-enhanced representation learning for knowledge graph [C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016: 1293-1299.

[29] Wu Y, Wang Z. Knowledge graph embedding with numeric attributesof entities [C]//Proceedings of the 3rd Workshop on Representation Learning for NLP, 2018: 132-136.

[30] An B, Chen B, Han X, et al. Accurate text-enhanced knowledge graph representation learning [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 745-755.

[31] Denis Krompaß, Stephan Baier, Volker Tresp. Type-constrained representation learning in knowledge graphs [C]//Proceedings of ISWC, 2015: 640-655.

[32] Ruobing Xie, Zhiyuan Liu, Maosong Sun. Representation learning of knowledge graphs with hierarchical types [C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016: 2965-2971.

[33] R. Xie, Z. Liu, H. Lan, et al. Image-embodied knowledge representation learning [J]. arXiv preprint arXiv: 1609.07028, 2016.

[34] Cai P, Li W, Feng Y, et al. Learning knowledge representation across knowledge graphs [C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017: 704-710.

[35] Feng J, Huang M, Yang Y. GAKE: graph aware knowledge embedding [C]//Proceedings of the 26th International Conference on Computational Linguistics, 2016: 641-651.

杜文倩(1996—)，硕士研究生，主要研究领域为知识图谱表示学习、知识推理技术。
E-mail: 1850617161@qq.com

王瑞(1995—)，硕士研究生，主要研究领域为事件抽取、实体消歧、知识的表示与推理技术。
E-mail: 2911525399@qq.com

李弼程(1970—)，通信作者，博士，教授，博士生导师，主要研究领域为文本分析与理解、语音处理与识别、图像/视频处理与识别、信息融合。
E-mail: lbclm@163.com