DOI:10.3969/j.issn.1671-0673.2021.05.007

基于知识表示学习的知识图谱补全研究综述

杨大伟,周 刚,卢记仓,宁原隆

(信息工程大学,河南 郑州 450001)

摘要:知识图谱补全能够将知识图谱补充完整,是知识图谱领域的一个研究热点。基于知识表示学习的知识图谱补全学习知识的向量表示,利用向量的计算挖掘知识图谱中的隐藏关联,具备更高的计算效率和更强的泛化能力,是知识图谱补全最好的方案之一。首先,介绍知识图谱补全和知识表示学习的概念;其次,按照实体和关系是否固定分别介绍静态知识图谱补全和动态知识图谱补全,对两个不同场景下各类算法的思路及改进过程进行详细说明;最后,总结知识图谱补全研究现状并展望未来研究方向。

关键词:知识图谱;知识图谱补全;知识表示学习

中图分类号:TP311 **文献标识码**:A **文章编号**:1671-0673(2021)05-0558-08

Review of Knowledge Graph Completion Based on Knowledge Representation Learning

YANG Dawei, ZHOU Gang, LU Jicang, NING Yuanlong

(Information Engineering University, Zhengzhou 450001, China)

Abstract: Knowledge graph completion can complete the knowledge graph, which is a research hotspot in the field of knowledge graph. The completion based on knowledge representation learning is to learn the vector representation of knowledge and use vector calculation to mine the hidden associations in the knowledge graph, which has higher computational efficiency and stronger generalization ability, and it is one of the best solutions for the completion of knowledge graph. First, this paper introduces the concepts of knowledge graph completion and knowledge representation learning. Second it introduces static knowledge graph completion and dynamic knowledge graph completion according to whether the entity and relationship are fixed, and the ideas and improvement processes of various algorithms in two different scenarios are explained in detail. Last, the current status and development directions of knowledge graph complementation are summarized.

Key words: knowledge graph; knowledge graph completion; knowledge representation learning

近年来,互联网、物联网、云计算等技术高速发展,催生出大量应用,伴生出海量的数据,这些数据中蕴含着许多有价值的知识。如何有效组织表达这些知识以便进一步进行计算分析成为研究者关注的热点,知识图谱由此产生[1]。2012 年,谷歌公司率先提出知识图谱概念并将其应用于智能化搜索引擎。知识图谱本质上是一种由节点和边组成的语义网络,其中每个节点表示现实世界中存在的“实体”,而每条边为实体间的“关系”。知识图谱通常使用三元组(头实体,关系,尾实体)作为形式化的表达,如(河南,省会,郑州),其中“河南”和“郑州”分别为头实体和尾实体,“省会”是两个实

收稿日期:2021-05-13;**修回日期**:2021-06-02

基金项目:河南省科技攻关资助项目(192102210129)

作者简介:杨大伟(1996-),男,硕士生,主要研究方向为知识图谱表示学习。

体之间的关系,该三元组描述了“河南省会是郑州”这个事实。

目前已经出现许多大规模知识图谱,其中典型的有 YAGO[2]、DBpedia[3]、WordNet[4]、Freebase[5]等,这些知识图谱都是由人工或半自动的方式构建,图谱具有稀疏性,大量实体间隐含的关系尚未被充分挖掘。以 Freebase 为例,作为一个开放式的知识图谱,由所有用户创作共享,其规模是目前已有的知识图谱中最大的,然而 Freebase 中信息缺失的问题十分严重,就“国籍”、“学历”、“父母”信息缺失情况来说,分别占据了图谱中总人物实体数目的 75%、91%和 94%。这些生活中日常可见的关系已是如此,一些特殊的关系相比之下缺失情况更加严重[6]。知识图谱的不完整使得下游应用无法发挥出应当具备的效能,相关领域的研究者们开始思考如何将知识图谱补充完整。

为了将知识图谱补充完整,知识图谱补全技术应运而生。知识图谱补全的方法各不相同,主要的思路有路径查找、规则推理以及知识表示学习,其中基于知识表示学习的补全通过学习实体和关系的向量表示来描述三元组所对应语义关系,在表示空间中利用向量的计算进行链接预测等任务,进而挖掘知识图谱中的隐藏关联,将知识图谱变得更加完整[7]。起初,知识表示学习只是研究如何能够更好地表示知识来构建知识图谱,其本身并不是作为一个专门为知识图谱补全服务而产生的算法,然而由于知识表示学习在知识图谱整体技术中处于上游位置,自然地被人们应用于图谱补全工作。此外,由于向量计算的可并行性,以知识表示学习为基础的补全相比其他方法在面对大规模知识图谱时能够有效提高工作效率;同时,对比独热(one-hot)表示,知识表示学习所得到的向量表示能在维度更低的情况下包含更多的语义信息,这也间接缓解了数据的稀疏性。本文在介绍知识图谱补全和知识表示学习概念的基础上,详细说明知识图谱补全不同场景下各类算法的思路及改进过程,总结其研究现状并展望未来研究方向。

1　知识图谱补全与知识表示学习

知识图谱补全的目的是预测出缺失三元组所缺失的部分,从而使得知识图谱变得更加完整。根据三元组中具体预测对象的不同,知识图谱补全可以分为 3 个子任务:头实体预测、尾实体预测和关系预测。对于头实体预测,需要给定三元组的关系和尾实体,预测可以组成正确三元组的头实体,如(?,省会,郑州);对于尾实体预测和关系预测同理,如(河南,省会,?)和(河南,?,郑州),预测其中的尾实体和关系。

知识表示学习是在知识图谱上学习指代实体和关系的向量表示。具体来讲,对于知识图谱中的三元组(头实体,关系,尾实体),用一个 k 维的向量 $\boldsymbol{h}$ 来表示头实体,用一个同维度的向量 $\boldsymbol{t}$ 来表示尾实体,实体间的关系则用一个转换向量 $\boldsymbol{r}$ 或转换矩阵 $\boldsymbol{M}_r$ 来表示;同时,在知识表示学习的过程中,需要定义一个得分函数 $f(h,r,t)$ 计算向量 $\boldsymbol{h},\boldsymbol{r},\boldsymbol{t}$ 的得分,以此来评估这 3 个向量在语义上构成的三元组是正确的还是错误的。倘若得分超过了某个阈值就被认为是错误的,倘若得分低于某个阈值就被认为是正确的,该阈值也叫作划分正确三元组和错误三元组的边界距离。此外,衡量学习模型的好坏是建立在考虑所有三元组得分情况的基础上,因此定义一个损失函数 $L(D^+,D^-,\gamma)$ 来考察所有三元组得分。其中:D^+是正确三元组集合;D^-是错误三元组集合,用来加速模型的训练;γ 是边界距离。损失函数的值越小,意味着所有三元组的得分越低,向量在语义上构成的三元组的正确率也越高,表示学习算法的性能也就越好。

2　静态知识图谱补全

按照补全思想的不同,目前静态知识图谱补全主要划分为 3 种,即平移距离模型、语义匹配模型和神经网络模型。

2.1　平移距离模型

文献[8]首次提出了 Word2vec 词向量工具并发现在词向量空间中存在着平移不变现象,即两个语义相似的词在向量空间上也会存在着一定的关系,如 $\boldsymbol{v}(King)-\boldsymbol{v}(Queen)\approx\boldsymbol{v}(Man)-\boldsymbol{v}(Woman)$,其中 $\boldsymbol{v}(x)$ 代表通过 Word2vec 学习得到的 x 单词的词向量。同时,多次反复实验表明平移不变的现象对于词汇的结构、语义之间的关系具有普遍性。

上述研究成果为后来的研究者们提供了新的思路,文献[9]受该研究的启发,将这种空间中平移不变现象成功应用到了知识表示学习中并提出了 TransE 模型,如图 1 所示。该模型认为,知识图谱中的关系可以被看作是头实体向尾实体的一种平移变化,这种变化不会改变实体和关系的性质,并且能够反映出其所构成的三元组的语义,具体对应到每个三元组$(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$,其应该满足 $\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$;对应

到得分函数就是向量 $\boldsymbol{h}+\boldsymbol{r}$ 与 $\boldsymbol{t}$ 之间的 $L1$ 或 $L2$ 距离：

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

相比以往基于符号学习推理的模型，TransE 模型参数更少，计算复杂度更低，在大规模知识图谱上具备更加明显的优势。

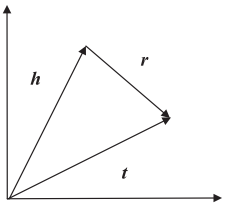

图 1 TransE 模型示意图

尽管 TransE 模型简单高效，但其在复杂关系下无法进行有效的表示学习，如“国籍”这种多对多的关系。当知识图谱中存在多个具有相同“国籍”关系的三元组时，如(小明，国籍，中国)和(小张，国籍，中国)这两个三元组，人物实体“小明”和“小张”在表示空间中的实体向量几乎相同，然而现实生活中这是两个不同的人，在年龄、出生地、学历等方面可能都会有差异，TransE 模型无法对这些差异进行区分，因此在复杂关系下表现欠佳。为了解决 TransE 模型在复杂关系上的局限性，文献[10]引入了映射的思想并提出了基于超平面的模型 TransH。借助超平面映射，TransH 模型可以实现某一个固定的实体在多个不同的关系下拥有多个不同的向量表示。具体的，如图 2 所示，对于某个固定的关系 r，TransH 模型综合使用平移向量 $\boldsymbol{r}$ 及其所在超平面的法向量 $\boldsymbol{w}_r$ 进行表示；对于含有该关系 r 的图谱中三元组 $(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$，TransH 模型考虑将两个实体向量分别沿着关系向量中的法向量映射至该关系所对应的超平面上，映射后的两个向量分别记为 $\boldsymbol{h}_r$ 和 $\boldsymbol{t}_r$。根据映射法则，其计算公式如下：

$$\begin{cases}\boldsymbol{h}_r=\boldsymbol{h}-\boldsymbol{w}_r^T\boldsymbol{h}\boldsymbol{w}_r\\ \boldsymbol{t}_r=\boldsymbol{t}-\boldsymbol{w}_r^T\boldsymbol{t}\boldsymbol{w}_r\end{cases} \tag{2}$$

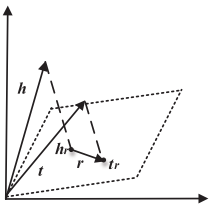

图 2 TransH 模型示意图

TransH 模型能够实现实体在复杂关系下的有效学习建模，但它和 TransE 模型有一个相同的假设前提，即实体和关系都是在一个固定的语义向量空间中进行计算。而实体作为拥有多个关系的结合体，这样的处理显然是考虑不周的，因此文献[11]提出了 TransR 模型。如图 3 所示，TransR 模型对图谱中的实体同样进行映射处理，只是该模型下映射的目标是各个不同关系空间，使现实世界中具有某些相似属性标签的实体在关系所对应的空间中能够被合理地区分开。具体的，每一个关系 r 都能够经过学习得到一个映射矩阵 $\boldsymbol{M}_r\in R^{d\times k}$，通过该映射矩阵后的头尾实体表示为 $\boldsymbol{h}_r$ 和 $\boldsymbol{t}_r$，其计算方式如下：

$$\begin{cases}\boldsymbol{h}_r=\boldsymbol{h}\boldsymbol{M}_r\\ \boldsymbol{t}_r=\boldsymbol{t}\boldsymbol{M}_r\end{cases} \tag{3}$$

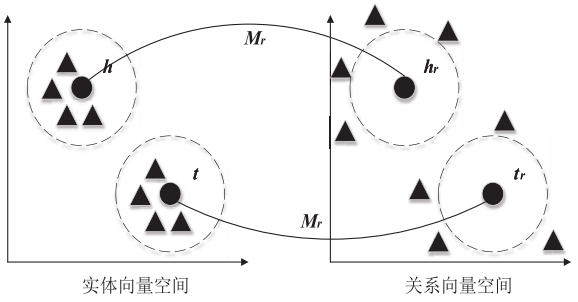

图 3 TransR 模型示意图

TransR 模型中空间级的映射虽然有效提高了图谱中知识的表达，但不可避免地扩大了参数的规模，延长了模型整体的计算周期；同时，TransR 模型中的映射矩阵仅仅是基于关系的，并没有形成一种交互的过程，因此文献[12]提出了 TransD 模型，该模型中的映射矩阵是动态的，对头尾实体进行了区分并构造了相互独立的两个动态映射矩阵 $\boldsymbol{M}_{hr}$ 和 $\boldsymbol{M}_{tr}$，如图 4 所示。动态映射矩阵的定义如下：

$$\begin{cases}\boldsymbol{M}_{rh}=\boldsymbol{r}_p\boldsymbol{h}_p^T+\boldsymbol{I}\\ \boldsymbol{M}_{rt}=\boldsymbol{r}_p\boldsymbol{t}_p^T+\boldsymbol{I}\end{cases} \tag{4}$$

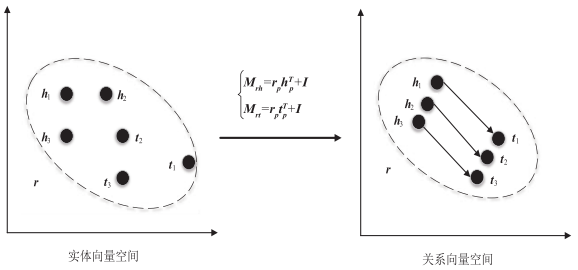

图 4 TransD 模型示意图

由于 TransD 的映射矩阵是由映射向量的乘积得到，因此能够有效缓解 TransR 模型中的参数压力。

上述 4 种模型都是平移距离模型中的代表，此

后的平移距离模型虽然各不相同,但大多都是在这些模型的基础上进行改进。平移距离模型中,通常使用 Margin Ranking Loss 作为损失函数,其定义如下:

$$L=\sum_{x^+\in D^+}\sum_{x^-\in D^-}\max(f(x^+)-f(x^-)+\gamma,0) \quad (5)$$

2.2　语义匹配模型

语义匹配模型又名为“双线性模型”,其核心思想是将实体中隐藏语义同表示空间中已有的关系进行匹配,从而判断事实是否成立。

文献[13]提出了首个进行语义匹配建模的模型 RESCAL 模型,该模型下图谱中的头尾实体用向量 $\boldsymbol{h},\boldsymbol{t}$ 进行表示,关系用矩阵 $\boldsymbol{M}_r$ 进行表示,实体向量和关系矩阵一起构成一个张量 $\boldsymbol{X}$,倘若 $(\boldsymbol{h},\boldsymbol{r},\boldsymbol{t})$ 成立,则对应的张量 $\boldsymbol{X}_{hrt}=1$,否则该张量为 0,其中关系矩阵 $\boldsymbol{M}_r$ 能够实现对潜在因子之间的成对相互作用进行建模。

文献[14]将上述模型中的关系矩阵 $\boldsymbol{M}_r$ 从一般矩阵转变对角矩阵,并由此提出改进后的算法 DistMult,通过对角矩阵的引入,可以显著提高 RESCAL 模型对知识的学习建模能力。

文献[15]提出的 ComplEx 模型考虑在复数空间中学习实体和关系的向量表示,从而有效处理对称与反对称的关系。

文献[16]在上述研究工作的铺垫下提出了一个统一的理论框架 Analogy,对以往语义匹配模型进行了总结,并提出使用了一个近乎对角的二乘二的分块矩阵 $\boldsymbol{B}_r$ 来描述知识图谱中的复杂关系,使得模型对知识有效学习的同时能够显著提高效率。

文献[17]创造性地将任意一个关系 r 分为已有的关系 r 本身和其所对应的逆关系 r^{-1} 两个向量表示进行学习,并以此提出 Simple 模型。

不同于平移距离模型,语义匹配模型中大多使用带正则化的 Logistic Loss 作为损失函数,具体如下:

$$L=\sum_{x\in D^+\cup D^-}\log(1+\exp(-yf(x)))+\lambda\|\theta\| \quad (6)$$

其中 y 是标签符,记录了该三元组是正样本还是负样本,取值为 1 或-1。

2.3　神经网络模型

神经网络模型主要是通过神经网络的学习能力和泛化能力来建模知识图谱中的事实元组,此类模型对实体和关系的维度没有统一的要求。

文献[18]提出一种神经张量网络模型 NTN,该模型中关系被表示成一个三阶的张量,使得每个切片对应的语义类型各不相同,有利于描述和该关系有关联的不同实体之间存在的各种语义联系。不同以往的表示方式,模型中的实体向量是通过对文本信息中包含的词所对应词向量取均值得到,这种表示方式可以充分保留长尾的实体名称中所隐含的实体语义,在没有任何实体描述信息的情况下,通过对实体名称的拆解也能获得一部分对实体的描述。

NTN 模型引入了张量计算,这种计算相对于向量计算来说具有更高的复杂程度,同时该模型依赖大量学习样本才能训练出理想的效果,因此文献[19]在其基础上提出了 ER-MLP 模型,这是一种简化版的 NTN 模型,可以在保持模型性能的同时缩小参数的规模。

除了张量模型外,还有一类卷积模型,其最早是由文献[20]提出的多层卷积网络模型 ConvE,该模型首先对图谱中的实体和关系之间进行一个 2D 的卷积操作,其次通过全连接网络输出三元组的得分,将向量空间中的计算问题转化为图像上的特征提取问题。

文献[21]直接使用 1D 卷积操作,通过这种方式来维持知识图谱中平移不变的特性,从而在学习三元组局部特征的同时也能学习到三元组之间的整体特征,并由此提出了 ConvKB 模型。

文献[22]开创性地提出了 CapsE 模型,这是一种胶囊网络模型,用一个三列的矩阵来对单个三元组进行学习建模,矩阵中的列与头实体、关系和尾实体一一对应;同时,对这个三列矩阵作卷积操作,通过生成不同的特征映射来构建胶囊,胶囊之间会进行路由,从而得到一个连续的向量,最终对该向量取模来计算得分。

此外,还有一些考虑传递的路径关系并以此为学习建模对象的模型,这类模型所涉及的主体不再是单个三元组。文献[23]最早进行了这类工作,在图谱原有的三元组基础上加入了路径三元组 (h,p,t),$p=(r_1,r_2,\cdots,r_n)$ 被定义为一条 n 长度的路径,该路径会辅助 TransE 算法进行学习,由此提出 PTransE 模型;文献[24]提出用图卷积网络来学习关系路径的 R-GCN 模型;文献[25]提出用递归跳跃网络模型 RSN 来学习长期关系依赖信息,加强了实体间语义的传播。

3　动态知识图谱补全

上一节所介绍的基于知识表示学习的知识图谱补全模型都有一个大前提,即实体和关系是固定

不变的,这也是静态知识图谱补全含义的由来。然而一切事物总是在不断发展变化之中,实体会随着时间改变,新实体也会随着时间产生,采用静态知识图谱补全模型对上述情况进行补全工作需要重新训练知识图谱中的所有数据集,这在知识图谱规模不断扩大的背景下显得愈来愈乏力,成本也越来越高。因此,研究者们考虑如何学习能反映时间信息的知识表示以及如何学习未知实体的知识表示,从而进行动态知识图谱补全工作。

在动态知识图谱补全中,根据实体是否已知可以分成两类场景:①动态变化,实体已经在知识图谱中,但其自身会随着时间变化,需要补全不同时间下的隐含关联;②动态添加,实体不在知识图谱中,是新实体,需要补全到知识图谱的实体集中[26]。针对第一类场景,研究者们考虑将时间维度引入到图谱的知识表示中,通过建模相同实体在不同时间下的不同表示来挖掘知识图谱中的隐含三元组,实现知识图谱的补全。针对第二类场景,为了动态添加新实体,需要建立新实体与现有知识图谱的关联,这种关联需要结合额外信息。按照新实体的额外信息属性,第二类场景又可以分为两类子场景:新实体含有丰富的文本信息,如实体名称、实体描述、实体类型等,通过已知实体和未知实体在文本表达上的相似性可以建立两者之间的关联,进而学习到未知实体的知识表示;新实体含有丰富的结构信息,即存在大量三元组既包含新实体,也包含知识图谱中已知实体,通过传播已知实体的知识表示可以输出新实体的知识表示。

根据动态知识图谱补全应用场景的不同,本节依次以时间信息、文本信息和结构信息为切入点,介绍当前基于知识表示学习的动态知识图谱补全取得的相关工作成果。

3.1 依据时间信息的动态知识图谱补全

文献[27]首次将时间信息用于动态知识图谱补全并提出了 TAE 模型。首先,TAE 模型利用含有时间标注的四元组(e_i,r,e_j,t)来表示事实,其中$t=[t_b,t_e]$是实体e_i和e_j具有关系r的时间区间(t_b为开始时间,t_e为结束时间),当某些事实是非持续瞬时发生时$t_b=t_e$;当某些事实目前尚未结束时$t_e=+\infty$。其次,为了模拟知识的动态变化,TAE 模型规定有时间顺序的关系彼此相关并在时间维度上依次发展,例如对于某一个固定的人来说,“出生->工作->死亡”是依次发生且无法逆行的。当两个关系共享同一个头实体时,考虑将其配对成时序关系对,如<出生,死亡>,其中较早发生的为“先验关系”,较晚发生为的“后继关系”,时序关系对中先验关系在前为正时序对,否则为负时序对。在计算上,通过定义一个矩阵$\boldsymbol{T}$来建模关系演化,即先验关系r_1在矩阵$\boldsymbol{T}$映射下会靠近后继关系r_2,即$r_1*\boldsymbol{T}\approx r_2$,而$r_2$经过$\boldsymbol{T}$的映射后会远离关系$r_1$,以此在训练中自动分离关系的先后。

文献[28]以 TransH 模型为参考提出了 HyTE 模型。在 HyTE 模型中,时间被认为是引起大多数多对一或一对多关系的主要原因,通过面向时间超平面的映射,可以直接利用时间信息,有效解决实体关系歧义,使得模型在复杂关系下有较好表现。为了构建时间超平面,考虑选取离散时间点序列$\{\tau_i\mid i=1,2,\cdots,T\}$对知识图谱进行分割,每个离散时间点对应一个时间超平面,每个超平面用其法向量$\{\boldsymbol{n}_{\tau_i}\}_{i=1}^{T}$来标记,便于表示空间中实体和关系的映射。

HyTE 模型对实体和关系链接预测的准确度有了显著提升,受到广泛关注,但其利用离散时间点对图谱切割的前提决定了该模型不能有效学习到时间长度的知识。针对这个问题,文献[29]提出了一种融合超平面和持续时间建模的模型 Duration-HyTE。该模型将事实中的关系分成持续型关系和瞬时关系,并参考文献[30]提出的持续时间模型对知识的有效持续时间建模,计算某类事件在特定有效时间点的可信度。该模型定义,若$\boldsymbol{T}$是连续型随机变量,且服从累计分布函数F,则在$t\in[0,+\infty)$上,可靠函数的表达为

$$R(t)=P(\boldsymbol{T}>t)=\int_t^{+\infty}f(u)\,\mathrm{d}u=1-F(t) \tag{7}$$

关于累计分布函数的选择,采用了能够快速有效模拟大多数有效持续时长分布的高斯函数作为关系密度函数,得到累计分布函数F为

$$F(r,\tau_s,\tau_e,\tau_p)=P(\{\boldsymbol{T}\leqslant\tau_p\})=\int_{\tau_s}^{\tau_p}\frac{1}{\sigma_r\sqrt{2\pi}}e^{-\frac{(y-\tau_e)^2}{2\sigma_r^2}}\,\mathrm{d}y \tag{8}$$

其中,τ_p是四元组$(h,r,t,[\tau_s,\tau_e])$的当前持续时间,σ_r是包含关系r的元事件持续时长的标准差。将式(7)和式(8)相结合,可以推导出有效可信度c:

$$c(r,\tau_s,\tau_e,\tau_p)=1-P(\{\boldsymbol{T}\leqslant\tau_p\})=1-\int_{\tau_s}^{\tau_p}\frac{1}{\sigma_r\sqrt{2\pi}}e^{-\frac{(y-\tau_e)^2}{2\sigma_r^2}}\,\mathrm{d}y \tag{9}$$

对于瞬时关系来说,其持续时间为 0,因此包含瞬时型关系的事实四元组的有效可信度在有效时间点上为 1,其余时间为 0;对于负样本来说,由于其在任何时间不正确的有效性都成立,所以在任何时间其有效可信度都为 1。由于正样本的可信度始终不会超过负样本,将这样的可信度直接训练会降低正样本的作用,因此在正样本的可信度基础上增加了超参数平衡因子 q,保持了正负样本在训练中的平衡。最后,将持续时间模型和 HyTE 模型相结合,因此模型得分函数为

$$f_{\tau_i}(h,r,t)=c\times\| h_{\tau_i}^{\perp}+r_{\tau_i}^{\perp}-t_{\tau_i}^{\perp}\|_{l_1/l_2} \tag{10}$$

在现实世界中,关系除了有时间上的瞬时性或持续性外,还有结构上的对称性、自反性,如“同学”“配偶”关系具有对称性,“等价”“子集”关系具有自反性。对于有自反性的关系来说,TAE、HyTE 等模型会出现如下现象:

$f(e_1,r_1,e_1,\tau)=f(e_2,r_2,e_2,\tau)=0\Rightarrow r_1=r_2=0$,

无法进行有效的表示学习。针对此问题,文献[31]提出在复平面上通过时间旋转建模知识表示的模型 TeRo。TeRo 模型将实体和关系映射到各自的复平面中,即令 $h,r,t\in C^k$。对于每个三元组发生的时间都会对应一个时间步 T,由 T 引导的函数将时间独立的实体知识表示向时间相关的实体知识表示逐元素旋转,映射函数定义为

$$h_\tau=h\circ T,t_\tau=t\circ T \tag{11}$$

其中∘是复平面上的厄米点乘。在复平面中,一个单位复数可以被认为是复平面上的一种旋转,因此限制 T 中每个元素的模的大小为 1,使得 τ 只是改变实体知识表示在复平面上的相位。其次,在复平面空间中设计关系为头实体知识表示 h_τ 到尾实体共轭知识表示 $\bar{t}_\tau$ 的转变,以便对复杂关系进行建模。同时,将具有时间跨度的事实按起始时间和结束时间分为(h,r_s,t,τ_s)和(h,r_e,t,τ_e)两个部分计算得分函数,避免了时间跨度大,涉及时间间隔多的事实因多次反复训练放大数据规模,节省了模型的训练时长。

3.2　依据文本信息的动态知识图谱补全

文献[32]最先提出了实体描述知识表示学习模型 DKRL,考虑将丰富的实体描述信息融入 TransE 等表示学习算法中,以此增强实体的表示能力。同时,文献中增加了泛化实验,表明借助实体描述信息可以学习到新实体的表示,引起了研究者的广泛关注。关于实体描述表示学习,提出了连续词袋模型(Continuous Bag of Words,CBOW)和卷积神经网络(Convolutional Neural Network, CNN)模型。在 CBOW 模型中,选取每个实体描述中前 n 个关键词后对其词向量进行求和来得到实体描述的知识表示;在 CNN 模型中,将词向量输入一个双卷积层神经网络,以此捕获丰富的特征信息来得到实体描述的知识表示,并且相较于前者,CNN 模型训练出的知识表示具备更为突出的性能。

在学习实体文本信息时,文本中的噪声会直接影响模型的性能,因此需要在学习的过程中自动过滤文本噪声。文献[33]提出了 Conmask 模型,该模型通过依赖关系的内容掩码机制与注意力机制结合,使得文本中的词可以根据其与给定关系的相似度分配到不同权重,通过这些权重的分配可以做到对文本信息的筛选,过滤了文本中的噪声。关于权重的计算,分别考虑最大词关系权重和最大上下文关系权重两种方式,后者放宽了权重求解的目标选择范围,可以减少目标预测实体被附近高权重词掩盖情况的发生。

此外,文献[34]提出了基于映射思想的 OWE 模型,考虑将文本描述映射到图的结构空间中去,实现文本信息和结构信息在表示空间上的融合。在该模型的学习过程中,文本特征和结构特征相互分开,独立学习,因此模型整体能够保留完整的图谱结构特征,并且对文本信息并不丰富的新实体也能进行有效的知识表示学习。

3.3　依据结构信息的动态知识图谱补全

文献[35]最先提出依据结构信息进行动态知识图谱补全工作,其想法来源于图神经网络对邻居节点特征进行聚合能够得到节点特征表示,所示考虑将图神经网络应用到知识图谱中新实体的知识表示学习上来,以类似的手法聚合实体周围有链接的实体来得到新实体的表示,并通过 TransE 中的得分函数对得到的新实体表示进行训练,可以有效实现动态知识图谱补全工作。

在对实体周围有链接的实体进行聚合时,不同实体间的关系可能存在一定的冗余,例如对于一名足球运动员来说,其中关系“*play_for*”和“*work_for*”两个关系其实表达的是相同的含义,同时,聚合的关系对当前所要预测的关系也应该具有不同的影响。因此,文献[36]提出了 LAN 模型,该模型在聚合之前考虑利用关系间的逻辑约束和注意力机制对不同的邻接实体关系分配了不同的权重,使得模型具备更好的预测能力。

不同于上述两篇文章在依据结构信息的动态知识补全工作中利用神经网络对实体周围有链接的实体关系进行聚合,文献[37]提出的 InvTransE

模型考虑利用 TransE 假设来预训练邻接实体的表示,再对这些表示直接计算,就能快速得到新实体的知识表示。在 InvTransE 模型中,如果新实体为头实体,则对邻接尾实体和关系的表示进行相减得到;如果新实体为尾实体,则对邻接头实体和关系的表示进行相加得到。同样的思想也可以将 TransR 等其他静态知识图谱补全模型引入来学习新实体的知识表示。

4 结束语

在新一代信息技术支持下,互联网大数据已经呈现出了爆炸式的增长,以这些数据为基础的知识图谱,其规模也变得越来越大。但知识图谱是由人工设计或半自动地从数据资源中抽取得到,规模的增大对于知识图谱来说,无论是图谱的完整程度,还是图谱中数据的质量,都没有相应的提升,因此知识图谱补全仍将是研究的重点内容。

动态知识图谱补全相比静态知识图谱补全考虑了图谱中实体关系随着时间的动态变化以及新实体的动态添加,具有很好的现实意义,目前已经取得了一定的研究成果,但仍有很多问题尚未考虑完全。①依据时间信息的动态知识图谱补全中,现有的模型在引入时间信息时忽视了结构化信息的历史变化趋势,这种变化趋势通常会影响实体及关系的表示;②依据文本信息的动态知识图谱补全中,现有的模型主要依赖实体描述来建立已知实体和未知实体的关联,然而实体的文本信息除了实体描述以外还有实体名称、实体类型等,这些都是丰富的文本资源,在后续的工作中可以研究如何将这些实体描述以外的文本信息进行统一考虑,以此来优化模型的训练效果,同时提高模型对于部分文本信息稀缺的容忍度;③依据结构信息的动态知识图谱补全中,现有的模型为了简化分析在学习过程中只利用了单步关联实体,未利用多跳关联实体,考虑到这些多跳关联实体有可能比单步关联实体具有更高的影响力,后续的工作中可以将多跳关联实体引入到模型的学习建模过程中,以此来优化未知实体的知识表示。

参考文献:

[1] 官赛萍,靳小龙,贾岩涛,等. 面向知识图谱的知识推理研究进展[J]. 软件学报,2018,29(10):2966-2994.

[2] LEHMANN J, ISELE R, JAKOB M, et al. DBpedia-a large-scale, multilingual knowledge base extracted from Wikipedia [J]. Semantic Web, 2015, 6(2): 167-195.

[3] 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. Vancouver, Canada, 2008. 1247-1250.

[4] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada,2007: 697-706.

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

[6] ROBERT W, EVGENIY G, KEVIN M, et al. Knowledge base completion via search-based question answering [P]. World wide web,2014.

[7] 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.

[8] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, 2013: 3111-3119.

[9] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, 2013: 2787-2795.

[10] 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. Quebec City, 2014: 1112-1119.

[11] 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. Austin, Texas, 2015: 2181-2187.

[12] 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: 687-696.

[13] 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 Machine Learning, 2011: 809-816.

[14] YANG B, YIH W, HE X, et al. Embedding entities and relations for learning and inference in knowledge bases [C]//3rd International Conference on Learning Representations. 2015:202-206.

[15] TROUILLON T, WELBL J, RIEDEL S, et al. Complex embeddings for simple link prediction[C]//Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, 48: 2071-2080.

[16] LIU H, WU Y, YANG Y. Analogical inference for multi-relational embeddings [C]//Proceedings of the 34th International Conference on Machine Learning, 2017, 70: 2168-2178.

[17] KAZEMI S M, POOLE D. SimplE embedding for link prediction in knowledge graphs[C]//Annual Conference on Neural Information Processing Systems, 2018: 4289-4300.

[18] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//27th Annual Conference on Neural Information Processing Systems 2013. 2013: 926-934.

[19] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]//The 20th ACM SIGKDD International Conference on knowledge Discovery and Data Mining, 2014: 601-610.

[20] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D knowledge graph embeddings[C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018: 1811-1818.

[21] NGUYEN D Q, NGUYEN T D, NGUYEN D Q, et al. A novel embedding model for knowledge base completion based on convolutional neural network[C]//Association for Computational Linguistics, 2018: 327-333.

[22] NGUYEN D Q, VU T, NGUYEN T D, et al. A capsule network-based embedding model for knowledge graph completion and search personalization[C]//Association for Computational Linguistics, 2019: 2180-2189.

[23] LIN Y, LIU Z, LUAN H-B, et al. Modeling relation paths for representation learning of knowledge bases [C]//The Association for Computational Linguistics. 2015: 705-714.

[24] SCHLICHTKRULL M S, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//The Semantic Web-15th International Conference, 2018, 10843: 593-607.

[25] GUO L, SUN Z, HU W. Learning to exploit long-term relational dependencies in knowledge graphs[C]//Proceedings of the 36th International Conference on Machine Learning, 2019, 97: 2505-2514.

[26] 丁建辉,贾维嘉.知识图谱补全算法综述[J].信息通信技术,2018,12(1):56-62.

[27] JIANG T, LIU T, GE T, et al. Towards time-aware knowledge graph completion [C]//26th International Conference on Computational Linguistics. Osaka, Japan, 2016:1715-1724.

[28] DASGUPTA S S, RAY S N, TALUKDAR P. HyTE: Hyperplane based temporally aware knowledge graph embedding[C]//the 2018 Conf on Empirical Methods in Natural Language Processing, 2018: 2001-2011.

[29] CUI Y N, LI J, SHEN L, et al. Duration-HyTE: A time-aware knowledge representation learning mrthod based on duration modeling[J]. Journal of Computer Research and Development, 57(6): 1239-1251.

[30] 于浏洋,郭志刚,陈刚,等.面向知识图谱构建的知识抽取技术综述[J].信息工程大学学报,2020,21(2):227-235.

[31] CHENG J X, MOJTABA N. TeRo: A time-aware knowledge graph embedding via temporal rotation [C]// Proceedings of the 28th International Conference on Computational Linguistics, 2020: 1583-1593.

[32] XIE R, LIU Z, JIA J, et al. Representation learning of knowledge graphs with entity descriptions [C]//The Thirtieth AAAI Conference on Artificial Intelligence, 2016:2659-2665.

[33] SHI B X and TIM W. Open-world knowledge graph completion [C]// The Thirty-Second AAAI Conference on Artificial Intelligence, 2018:101-105.

[34] HASEEB S, JOHANNES V. An open-world extension to knowledge graph completion models[C]//The Thirty-third AAAI Conference on Artificial Intelligence, 2019: 3044-3051.

[35] HAMAGUCHI T, OIWA H, SHIMBO M, et al. Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach[C]//The Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017:1802-1808.

[36] WANG P, HAN J, LI C, et al. Logic Attention based neighborhood aggregation for inductive [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017:1802-1808.

[37] DAI D, ZHENG H, LUO F, et al. Inductively representing out-of-knowledge-graph entities by optimal estimation under translational assumptions [DB/OL]. (2020-09-27)[2021-06-02]. https://arxiv.org/pdf/2009.12765.pdf.

(编辑:李志豪)