中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

融合概念和属性信息的领域知识图谱补全方法

陈伯谦, 王坚

引用本文:
陈伯谦,王坚. 融合概念和属性信息的领域知识图谱补全方法[J]. 控制与决策, 2024, 39(7): 2325−2333.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1994

融合概念和属性信息的领域知识图谱补全方法

陈伯谦, 王 坚†

(同济大学 电子与信息工程学院，上海 201804)

摘 要: 针对领域知识图谱具有严格的模式层和丰富的属性信息的特点,提出一种融合概念和属性信息的领域知识图谱补全方法. 首先对领域知识图谱模式层中的概念使用可建模语义分层结构的HAKE模型进行嵌入表示,建立基于概念的实例向量表示;其次对数据层的实例三元组和属性三元组进行区分,通过注意力机制对实例的属性和概念进行融合,建立基于属性的实例向量表示;最后对基于概念和基于属性的实例向量表示进行联合训练以实现对实例三元组的评分. 使用基于DWY100K数据集构建的知识图谱、MED-BBK-9K 医疗知识图谱和根据某钢铁企业设备故障诊断数据构建的知识图谱进行实验,结果表明所提出方法在领域知识图谱补全中的性能优于现有知识图谱补全方法.

关键词: 领域知识图谱；知识图谱嵌入；知识图谱补全；模式层；数据层；注意力机制

中图分类号: TP182 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1994

引用格式: 陈伯谦,王坚. 融合概念和属性信息的领域知识图谱补全方法[J]. 控制与决策, 2024, 39(7): 2325-2333.

Domain knowledge graph completion method incorporating concept and attribute information

CHEN Bo-qian, WANG Jian†

(College of Electronic and Information Engineering，Tongji University，Shanghai 201804，China)

Abstract: Aiming at the characteristics of domain knowledge graphs with strict schema layers and rich attribute information, a method of domain knowledge graph completion incorporating concept and attribute information is proposed. Firstly, the concepts in the schema layer of the domain knowledge graph are represented by embedding using the HAKE model which can model semantic hierarchical structures to build a concept-based instance vector representation. Then, a distinction is made between instance triples and attribute triples for the data layer, and an attribute-based instance vector representation is obtained by incorporating the attributes and concepts of the instance through the attention mechanism. Finally, the concept-based and attribute-based instance vector representations are jointly trained to achieve scoring of the instance triples. Experiments are conducted using the knowledge graph constructed based on the DWY100K dataset, the medical knowledge graph MED-BBK-9K and the knowledge graph constructed based on equipment fault diagnosis data of a steel enterprise, and the experimental results show that the performance of the proposed method in domain knowledge graph completion is better than the existing knowledge graph completion methods.

Keywords: domain knowledge graph；knowledge graph embedding；knowledge graph completion；schema layer；data layer；attention mechanism

0 引 言

知识图谱是Google公司于2012年提出的概念[1],旨在描述真实世界中存在的各种实体或概念及其之间的关系,它可以将多源异构的信息组织为结构化的网状知识库,为人工智能应用提供高质量的结构化知识. 根据知识领域和范围的不同,知识图谱可以分为通用知识图谱和领域知识图谱. 通用知识图谱涉及的知识范围广,通常包含大量现实世界中的常识性知识;而领域知识图谱面向某一特定领域,对知识的深度和准确度具有更高的要求. 知识图谱的应用对于提高垂直领域的智能化水平具有重要意义. 以工业领域为例,传统的工业场景面临着设备信息复

收稿日期: 2022-11-17；**录用日期:** 2023-04-24.

基金项目: 科技创新2030新一代人工智能重大项目(2018AAA0101800)；国家自然科学基金项目(72271188).

责任编委: 侯忠生.

†通讯作者. E-mail: jwang@tongji.edu.cn.

杂、知识处理手段有限等问题,而工业知识图谱可以对设备资料、产线安排等知识进行有效的管理,并利用相关知识进行推理与决策,从而提高知识运营能力和产线的智能化水平.

由于知识图谱通常通过人工或半自动化的方式进行构建,且构建时所使用的语料无法包含对于常识性知识等信息的明显表述[2],知识图谱通常不可避免地存在不完备的问题. 在工业等垂直领域,由于应用场景复杂、语料信息相对缺乏以及存在未被挖掘的知识[3],领域知识图谱不完备的情况通常更加严重,对基于知识图谱的推理和决策等后续应用的效果造成影响. 因此,对领域知识图谱中隐含的关系进行挖掘和补全,对于提升领域知识图谱的应用效果具有重要意义.

目前,基于知识图谱嵌入的方法是知识图谱补全的主流方法. 知识图谱嵌入(knowledge graph embedding, KGE)通过学习知识图谱的内在结构和相关语义信息,将知识图谱中的实体和关系嵌入到低维连续的向量空间中,并通过特定的评分函数衡量三元组的合理性,从而实现对隐含关系的预测和补全. 知识图谱嵌入方法由于能够高效计算实体与关系的语义联系,在学术界和工业界均受到了高度关注[4].

与通用知识图谱相比,领域知识图谱具有对知识准确性要求高、实体属性较多等特点,因此,用于领域知识图谱补全的知识图谱嵌入方法也应结合这些特点进行设计. 领域知识图谱通常包含模式层和数据层两个层次[5],某钢铁企业设备故障诊断知识图谱的模式层和数据层的部分结构如图1所示. 模式层是知识图谱的核心,包含数据层内容的抽象概念及其之间可能存在的关系,对加入数据层的实体和关系起到约束作用,能够有效提高领域知识图谱的准确性,因此,领域知识图谱补全算法也应充分利用模式层的概念信息;数据层是知识图谱的主要部分,包含以实例三元组或属性三元组的形式表示的具体知识数据,属性三元组可以对相关实例起到描述作用,且部分属性对于实例信息的表示很可能具有重要意义,因此,领域知识图谱补全算法还需以不同的层次对实例三元组和属性三元组进行表示,并充分利用属性信息提升实例的表示效果.

基于上述思想,本文提出一种融合概念和属性信息的领域知识图谱补全方法,区别于传统知识图谱嵌入方法对所有三元组平等地进行向量编码的方式,所提出方法充分利用领域知识图谱模式层的概念信息,并对数据层的实例和属性节点进行区分,将实例的概念与属性进行融合. 实验结果显示,所提出方法对于领域知识图谱的补全是行之有效的.

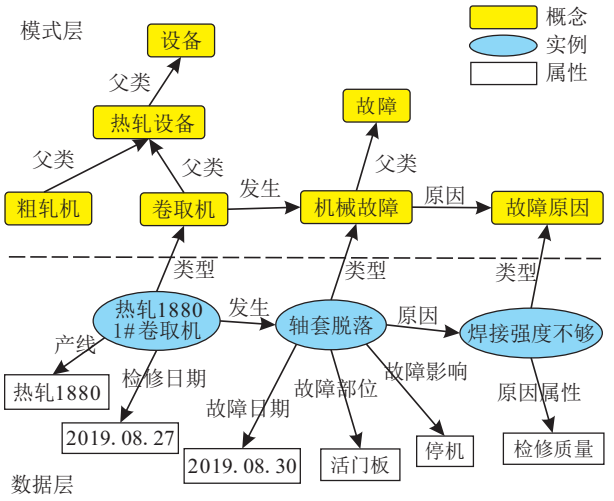


图 1 知识图谱的模式层和数据层结构

1 相关工作

目前知识图谱嵌入的代表模型可以分为平移距离模型、几何模型、语义匹配模型和神经网络模型等类型:

1) 平移距离模型将关系建模为头实体到尾实体的平移向量. TransE[6] 是最具代表性的平移距离模型,该模型将尾实体向量建模为头实体向量与关系向量之和,具有简单高效的优势. TransH[7] 模型将实体向量投影至关系对应的超平面再进行平移操作,使实体在涉及不同关系时具有不同的表示形式. TransR[8] 模型在实体空间和多个关系空间中对实体和关系进行建模,能够区分实体包含的多层面信息.

2) 几何模型将关系建模为语义空间中除平移变换以外的更加复杂的几何变换. RotatE[9] 模型将实体和关系映射到复数空间,并将关系建模为由头实体到尾实体的旋转. HAKE[10] 模型在极坐标系中表示实体和关系,能够对实体之间的语义层次进行建模.

3) 语义匹配模型利用基于相似性的评分函数,通过匹配实体的潜在语义和向量空间表示中体现的关系衡量事实的合理性. RESCAL[11] 模型将知识图谱中的头尾实体建模为向量,将关系建模为矩阵,能够实现对潜在因子之间成对相互作用的建模. DistMult[12] 模型在 RESCAL 模型的基础上将关系矩阵限制为对角矩阵,解决了 RESCAL 容易过拟合的问题.

4) 神经网络由于具有极高的特征学习能力,在知识图谱补全中也得到了广泛应用. ConvE[13] 模型对知识图谱中的实体和关系进行2D卷积操作,并通过全连接网络输出三元组的得分. KBGAT[14] 模型通过

图注意力网络对实体邻域信息的影响力差异进行考虑,从而可以为每个邻居实体分配不同的权重并实现对邻居实体特征信息的聚合.

上述知识图谱补全模型可以有效解决通用知识图谱的补全问题,但将其应用于领域知识图谱时,模型只能对知识图谱内的概念三元组、实例三元组和属性三元组平等地进行向量编码,无法使模式层发挥对数据层内容应有的规范和约束作用. 同时,数据层大量的属性信息加剧了多对一、多对多关系等复杂现象,进一步对模型性能造成影响. 近年来,部分研究工作对概念、属性等不同类型实体之间的语义差异进行了考虑. Lv等[15]提出了一种区分概念和实例的知识图谱嵌入模型TransC,将知识图谱中的概念编码为球体,将实例编码为同一语义空间中的向量,能够有效建模instanceOf与subClassOf关系. Guan等[16]提出了一种考虑概念的知识图谱嵌入模型KEC,通过将损失向量投影到概念子空间衡量三元组成立的可能性,提高了知识图谱补全和实体分类的性能. Li等[17]提出了一种本体信息约束的知识图谱嵌入模型TransO,对本体中的关系和类型约束以及层次结构约束进行考虑,实现了本体信息的无缝融合. Zhang等[18]提出了一种属性嵌入的知识表示学习方法AKRL,利用深度卷积神经网络对实体的属性信息进行编码,将属性信息与使用基于结构的表示方法获得的实体信息统一到同一向量空间,提高了知识图谱补全的性能. Lin等[19]提出了分离结构三元组和属性三元组的KR-EAR模型,缓解了一对多、多对一、多对多现象对模型性能造成的影响. Su等[20]针对生物医学知识图谱,使用编码器-解码器层从药物属性中学习嵌入并将其作为药物节点的初始化表示,通过邻域节点嵌入和三元组事实计算注意力权重并聚合一阶邻域信息,提高了药物相互作用预测的准确性. 上述方法对于领域知识图谱的补全具有一定的指导意义,但大多仅考虑了概念信息或属性信息,而本文的主要创新点在于综合考虑了领域知识图谱中的概念和属性信息,通过注意力机制对实例的概念和属性进行融合,并在评分函数中对基于概念和基于属性的实例向量表示进行加权结合,通过融合概念和属性信息提升相关实例的表示效果,完成领域知识图谱的补全任务.

2 本文方法

2.1 问题描述

领域知识图谱包括模式层和数据层两部分. 模式层包含知识图谱中实例数据的抽象概念信息,表示为$G_c = \{E_c, R_c, T_c\}$. 其中:E_c为模式层中的概念集合,R_c为模式层中的关系集合,T_c为模式层中的三元组集合. 数据层包含模式层中概念的实例、实例之间的关系以及实例所具有的属性等信息,表示为$G = \{E, A, R, R_a, T, T_a\}$. 其中:$E$为实例集合,$R$为实例间关系的集合,$R_a$为属性类型集合,$A$为属性值集合,$T$为实例三元组集合,$T_a$为属性三元组集合. 模式层和数据层的知识均以三元组的形式表示,模式层三元组表示为$(h_c, r_c, t_c) \in T_c$,其中$h_c, t_c \in E_c, r_c \in R_c$;数据层的实例三元组表示为$(h, r, t) \in T$,其中$h, t \in E, r \in R$;属性三元组表示为$(e, r_a, a) \in T_a$,其中$e \in E, a \in A, r_a \in R_a$. 同时,数据层的实例与模式层的概念、数据层的关系与模式层的关系均有对应关系,即对于任意数据层的实例$e \in E$,在模式层均存在一个概念$e_c \in E_c$与之对应,表示e是e_c的一个实例;对于任意数据层实例之间的关系$r \in R$,其所连接的头尾实体所对应的模式层中的概念之间也存在相应的概念关系$r_c \in R_c$.

知识图谱补全任务可以抽象为三元组预测任务,即对三元组中缺失的部分进行预测,具体分为头实体预测、尾实体预测和关系预测. 本文仅考虑数据层实例三元组的头、尾实体预测问题,即对于?部分未知的实例三元组$(?, r, t)$或$(h, r, ?)$,预测出?所对应的实例. 在知识图谱补全模型中,通常使用基于嵌入表示的评分函数$f(h, r, t)$评估三元组(h, r, t)的合理性,通过模型的训练使得真实三元组的得分高于非真实三元组,从而完成三元组预测任务.

2.2 算法架构

根据领域知识图谱具有严格的模式层和丰富的属性的特点,本文提出一种融合概念和属性信息的领域知识图谱补全方法. 首先对知识图谱的模式层和数据层分别使用现有的知识图谱嵌入模型进行预训练,获取模式层概念的向量表示以及数据层实例和属性的初始向量,并基于模式层的向量表示构建数据层实例的基于概念的向量表示;其次通过注意力机制融合实例的属性表示和概念表示,获得各个属性的注意力系数,从而聚合实例的各项属性信息,构建基于属性的实例向量表示;最终通过由基于概念和基于属性的实例向量表示加权结合的评分函数实现对三元组合理性的评估,完成三元组预测任务. 算法基本框架如图2所示.

2.3 模式层表示

领域知识图谱的模式层包含知识实例的抽象概念及其之间的关系,用于对数据层实体和关系的构

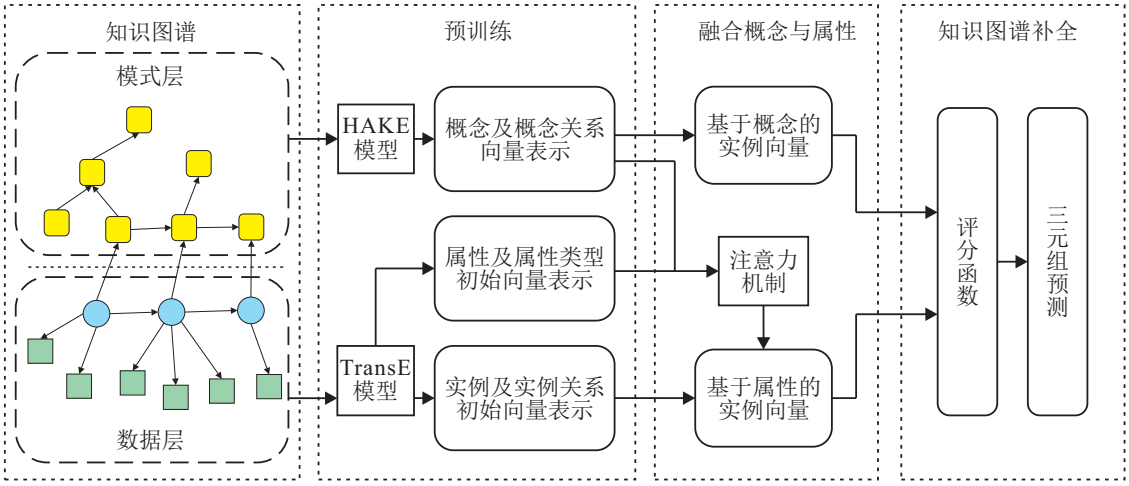


图 2 算法框架

建进行规范和指导. 在知识图谱补全任务中,要实现对模式层信息的利用,首先需要对其中的概念和关系进行嵌入表示. 由于概念之间广泛存在父子类和包含等关系,模式层通常具有较为明显的语义分层现象. 例如,图1概念层中“热轧设备”比“卷取机”的概念更加抽象,具有更高的语义层次. 基于这一特点,本文使用能够建模语义分层现象的HAKE模型[10]对模式层中的概念和关系进行嵌入表示.

HAKE模型将实体建模到极坐标系,通过模长区分不同语义层级的实体,通过相角区分位于同一语义层级的不同实体. 因此,对于模式层中的概念和关系,其向量均由模长和相角两部分组成. 对于给定的模式层三元组(h_c, r_c, t_c),使用$\boldsymbol{h}_{cm}$、$\boldsymbol{r}_{cm}$和$\boldsymbol{t}_{cm}$表示h_c、r_c、t_c的模长向量,则$\boldsymbol{h}_{cm}$、$\boldsymbol{r}_{cm}$和$\boldsymbol{t}_{cm}$期望满足

$$\boldsymbol{h}_{cm} \circ \boldsymbol{r}_{cm} = \boldsymbol{t}_{cm},\ \boldsymbol{h}_{cm}, \boldsymbol{t}_{cm} \in \mathbb{R}^k,\ \boldsymbol{r}_{cm} \in \mathbb{R}_+^k. \tag{1}$$

其中:k为向量的维度,$\circ$为Hadamard积. 即对于$\boldsymbol{h}_{cm}$、$\boldsymbol{r}_{cm}$和$\boldsymbol{t}_{cm}$的每个维度$[\boldsymbol{h}_{cm}]_i$、$[\boldsymbol{r}_{cm}]_i$和$[\boldsymbol{t}_{cm}]_i$,满足$[\boldsymbol{h}_{cm}]_i[\boldsymbol{r}_{cm}]_i = [\boldsymbol{t}_{cm}]_i$. 在此基础上,模长部分的评分函数定义为

$$f_{cm}(\boldsymbol{h}_{cm}, \boldsymbol{r}_{cm}, \boldsymbol{t}_{cm}) = -\|\boldsymbol{h}_{cm} \circ \boldsymbol{r}_{cm} - \boldsymbol{t}_{cm}\|_2. \tag{2}$$

使用$\boldsymbol{h}_{cp}$、$\boldsymbol{r}_{cp}$和$\boldsymbol{t}_{cp}$表示h_c、r_c、t_c的相角向量,则$\boldsymbol{h}_{cp}$、$\boldsymbol{r}_{cp}$和$\boldsymbol{t}_{cp}$期望满足

$$(\boldsymbol{h}_{cp} + \boldsymbol{r}_{cp}) \bmod 2\pi = \boldsymbol{t}_{cp},\ \boldsymbol{h}_{cp}, \boldsymbol{r}_{cp}, \boldsymbol{t}_{cp} \in [0, 2\pi)^k. \tag{3}$$

相角部分的评分函数可以定义为

$$f_{cp}(\boldsymbol{h}_{cp}, \boldsymbol{r}_{cp}, \boldsymbol{t}_{cp}) = -\|\sin((\boldsymbol{h}_{cp} + \boldsymbol{r}_{cp} - \boldsymbol{t}_{cp})/2)\|_1. \tag{4}$$

模式层三元组的最终评分函数由模长部分和相角部分的评分函数加权求和得到,即

$$f_c(\boldsymbol{h}_c, \boldsymbol{r}_c, \boldsymbol{t}_c) = \lambda_{cm} f_{cm}(\boldsymbol{h}_{cm}, \boldsymbol{r}_{cm}, \boldsymbol{t}_{cm}) + \lambda_{cp} f_{cp}(\boldsymbol{h}_{cp}, \boldsymbol{r}_{cp}, \boldsymbol{t}_{cp}), \tag{5}$$

其中λ_{cm}、$\lambda_{cp} \in \mathbb{R}$为由模型学习到的权重参数.

2.4 数据层表示

数据层是知识图谱的主要部分,包含以实例三元组和属性三元组的形式存储的具体知识. 为了充分利用知识图谱的模式层信息和数据层中的属性三元组信息,将数据层的实例和实例之间关系的嵌入分为基于概念的向量表示和基于属性的向量表示两部分.

2.4.1 基于概念的实例向量表示

对于数据层中的实例$e \in E$及实例之间的关系$r \in R$,仍使用第2.3节所述的概念层表示方法来建模基于概念的实例表示. 对于给定的实例三元组(h, r, t),使用$\boldsymbol{h}_m$、$\boldsymbol{r}_m$和$\boldsymbol{t}_m$表示h、r、t的模长向量,使用$\boldsymbol{h}_p$、$\boldsymbol{r}_p$和$\boldsymbol{t}_p$表示h、r、t的相角向量,则模长部分和相角部分的评分函数分别为

$$f_m(\boldsymbol{h}_m, \boldsymbol{r}_m, \boldsymbol{t}_m) = -\|\boldsymbol{h}_m \circ \boldsymbol{r}_m - \boldsymbol{t}_m\|_2, \tag{6}$$

$$f_p(\boldsymbol{h}_p, \boldsymbol{r}_p, \boldsymbol{t}_p) = -\|\sin((\boldsymbol{h}_p + \boldsymbol{r}_p - \boldsymbol{t}_p)/2)\|_1. \tag{7}$$

2.4.2 基于属性的实例向量表示

领域知识图谱的数据层除实例三元组外,还包含大量属性三元组,且一些属性很可能是相关实例的重要信息. 为充分利用这些信息,使用注意力机制对实例的属性进行融合,获取实例基于属性的向量表示.

注意力机制(attention mechanism)的本质是关注重要信息、抑制无用信息. Veličković 等[21] 提出的图注意力网络(GAT)模型将注意力机制引入图结构的建模中,实现了对不同邻居节点影响力差异的考虑,能够为每个邻居节点分配不同的权重; Nathani 等[14] 将图注意力网络应用于知识图谱,提出了KBGAT模型,在图注意力网络的基础上引入对关系的考虑. 而本文认为属性的不同是区分同一概念不同实例的重

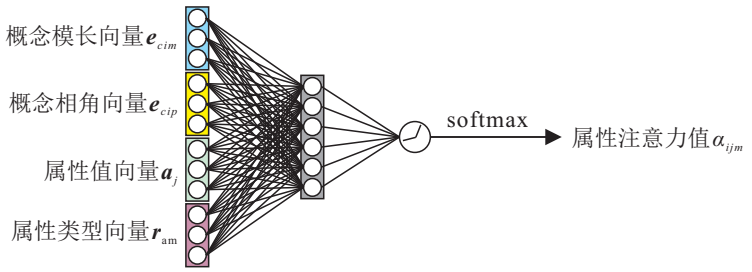


图 3　注意力机制结构

要方式,因此在KBGAT模型的基础上引入概念和属性信息,通过将实例的属性信息与概念信息进行融合获取该实例各个属性的注意力权重,具体结构如图3所示.

对于数据层中的一个实例$e_i \in E$,为了融合e_i具有的属性信息,对其所对应的概念$e_{ci} \in E_c$的向量和以e_i为头实体的属性三元组$x_{ij}^m = (e_i, r_{am}, a_j)$中的属性类型和属性值向量进行拼接和线性变换,有

$$\boldsymbol{c}_{ijm} = \boldsymbol{W}_1[\boldsymbol{e}_{cim} \| \boldsymbol{e}_{cip} \| \boldsymbol{a}_j \| \boldsymbol{r}_{am}]. \tag{8}$$

其中:$\boldsymbol{c}_{ijm}$为e_i属性三元组x_{ij}^m融合了实例概念信息的向量表示,$\boldsymbol{W}_1$为线性变换矩阵,$\boldsymbol{e}_{cim}$和$\boldsymbol{e}_{cip}$分别为e_i所对应概念e_{ci}的模长向量和相角向量,$\boldsymbol{a}_j$为属性值a_j的向量表示,$\boldsymbol{r}_{am}$为属性类型r_{am}的向量表示,$\|$为向量的拼接操作. 在此基础上,对$\boldsymbol{c}_{ijm}$进行权重矩阵为$\boldsymbol{W}_2$的线性变换,并通过PReLU激活函数获得属性三元组x_{ij}^m的绝对注意力值b_{ijm},有

$$b_{ijm} = \mathrm{PReLU}(\boldsymbol{W}_2 \boldsymbol{c}_{ijm}). \tag{9}$$

绝对注意力值b_{ijm}表示属性三元组x_{ij}^m对于实例e_i的重要程度. 对e_i所有属性三元组的绝对注意力值进行softmax操作,得到各个属性三元组的相对注意力值α_{ijm},有

$$\alpha_{ijm} = \mathrm{softmax}(b_{ijm}) = \frac{\exp(b_{ijm})}{\sum\limits_{n \in A_i} \sum\limits_{r \in R_{ai}} \exp(b_{inr})}. \tag{10}$$

其中:A_i为实例e_i所有属性三元组的属性值集合,R_{ai}为实例e_i所有属性三元组的属性类型集合. 在此基础上,实例e_i基于属性的向量表示由属性三元组的向量根据其各自的相对注意力值加权求和得到,有

$$\boldsymbol{e}_{ai} = \sigma\Big(\sum_{j \in A_i} \sum_{m \in R_{ai}} \alpha_{ijm} \boldsymbol{c}_{ijm} \Big), \tag{11}$$

其中σ为非线性的激活函数.

为了捕获更多的实例属性信息,引入多头注意力机制,使用S个相互独立的注意力机制分别计算基于属性的实例向量表示,并对每个注意力机制得到的向量表示求平均值得到实例的基于属性的向量表示,有

$$\boldsymbol{e}_{ai} = \frac{1}{S} \sum_{s=1}^{S} \sigma\Big(\sum_{j \in A_i} \sum_{m \in R_{ai}} \alpha_{ijm}^s \boldsymbol{c}_{ijm}^s \Big). \tag{12}$$

由于上述属性融合过程中缺失了实例e_i的初始向量信息,本文对实例的初始向量进行权重矩阵为$\boldsymbol{W}_e$的线性变换,使其维度与通过注意力机制融合属性和概念信息得到的向量维度一致,并将二者相加,得到实例e_i的最终向量,有

$$\boldsymbol{e}'_{ai} = \boldsymbol{W}_e \boldsymbol{e}_i + \boldsymbol{e}_{ai}, \tag{13}$$

其中$\boldsymbol{e}_i$为实例e_i的初始向量表示.

对于实例之间的关系$r_i \in R$,由于关系的不同头尾实例具有不同的属性,难以在关系中融入特定的属性信息,本文仅考虑关系的概念信息,通过权重矩阵为$\boldsymbol{W}_r$的线性变换使其概念向量的维度与实例向量一致,即关系r_i的最终向量为

$$\boldsymbol{r}'_{ai} = \boldsymbol{W}_r[\boldsymbol{r}_{cim} \| \boldsymbol{r}_{cip}], \tag{14}$$

其中$\boldsymbol{r}_{cim}$和$\boldsymbol{r}_{cip}$分别为r_i所对应概念r_{ci}的模长向量和相角向量.

在此基础上,对于给定的实例三元组(h, r, t),使用$\boldsymbol{h}_a$、$\boldsymbol{r}_a$和$\boldsymbol{t}_a$表示h、r、t的基于属性的向量,则基于属性的评分函数为

$$f_a(\boldsymbol{h}_a, \boldsymbol{r}_a, \boldsymbol{t}_a) = -\|\boldsymbol{h}_a + \boldsymbol{r}_a - \boldsymbol{t}_a\|_1. \tag{15}$$

2.4.3　训练方式

对于实例三元组(h, r, t),最终的评分函数由基于概念的向量表示中模长部分和相角部分的评分函数以及基于属性的评分函数加权求和得到,即

$$\begin{aligned} f(h, r, t) = {} & \lambda_m f_m(\boldsymbol{h}_m, \boldsymbol{r}_m, \boldsymbol{t}_m) + \\ & \lambda_p f_p(\boldsymbol{h}_p, \boldsymbol{r}_p, \boldsymbol{t}_p) + \lambda_a f_a(\boldsymbol{h}_a, \boldsymbol{r}_a, \boldsymbol{t}_a), \end{aligned} \tag{16}$$

其中$\lambda_m, \lambda_p, \lambda_a \in \mathbb{R}$为由模型学习到的权重参数.

在训练过程中,对于基于概念的向量表示部分,实例e和关系r的向量初始值均设置为其对应的模式层概念e_c和关系r_c的向量,权重参数λ_m、λ_p的初始值设置为模式层预训练得到的λ_{cm}和λ_{cp}值;对于基于属性的向量表示部分,实例与属性的初始向量使用

经TransE模型[6]对实例三元组和属性三元组联合进行预训练得到的向量. 对于训练集中的实例三元组,通过最小化如下损失函数优化模型结果:

$$L=\sum_{(h,r,t)\in T}\sum_{(h',r',t')\in T'}[\gamma+f(h,r,t)-f(h',r',t')]_+. \tag{17}$$

其中: $[x]_+$ 为0与 x 中的较大值, T 为正例三元组集合, T' 为由正例三元组随机替换头实体或尾实体生成的负例三元组集合, γ 为划分正例三元组和负例三元组的边界距离.

3 实验分析

3.1 数据集

知识图谱补全研究中常用的FB15k、WN18[6]等数据集均为通用知识图谱,不具备完整的模式层结构和丰富的属性信息,因此不适用于本文提出的方法. 为了评估本文方法在领域知识图谱补全任务中的有效性,使用DWY100K[22]、 MED-BBK-9K[23]以及根据某钢铁企业设备故障诊断数据构建的知识图谱进行实验. DWY100K数据集包含从3个百科知识库DBpedia、WiKidata和YAGO3中抽取的实例三元组和属性三元组,从该数据集中的DBpedia部分选取属性数量适中的实例数据,并使用由DBpedia的SPARQL查询接口查询到的实例类型信息构建模式层,经预处理后的知识图谱包含人物、地点、组织等10种类型的实例和出生日期、邮编等149种类型的属性. MED-BBK-9K为医学领域知识图谱,在对部分头尾实体相同的三元组进行删除后包含疾病、药物、症状等10种类型的实例和病因、临床表现、治疗原则等21种类型的属性. 钢铁生产设备故障诊断知识图谱根据某钢铁企业故障诊断文本中抽取的信息人工构建,包含电气设备、机械设备、电气故障、机械故障等85种类型的实例和设备所属产线、故障现象、故障影响等16种类型的属性. 各个数据集的数据统计信息如表1所示.

表1 数据集信息

数据集	DWY100K (DBpedia)	MED-BBK-9K	钢铁生产设备故障诊断
概念数量	10	10	85
实例数量	3 296	9 095	270
关系数量	31	18	5
属性类型数量	149	21	16
实例三元组数量	4 569	41 118	259
属性三元组数量	36 994	44 804	1 665

3.2 对比方法

为了验证方法的有效性,选取以下常用的知识图谱补全方法作为基准方法进行对比:

1) TransE[6]:将关系向量建模为头实体向量到尾实体向量的平移.

2) TransH[7]:将实体向量投影至关系对应的超平面再进行平移操作.

3) RotatE[9]:将关系建模为复数空间中头实体到尾实体的旋转.

4) HAKE[10]:将实体建模到极坐标系,通过模长区分不同语义层级的实体,通过相角区分位于同一语义层级的不同实体.

5) ConvE[13]:对实体和关系进行2D卷积操作,通过全连接网络输出三元组的得分.

6) KBGAT[14]:基于图注意力网络对实体的邻域信息进行聚合.

为了确保对比方法与本文方法在训练过程中使用相同信息,对比方法的实验中将模式层的概念三元组和数据层的属性三元组均视为普通三元组并与训练集中的实例三元组合并训练,而验证集和测试集与本文方法的实验相同,仅包含实例三元组.

3.3 评价指标

本文选用知识图谱补全研究中常用的MR (mean rank)、MRR (mean reciprocal rank)和Hits@n作为评价指标.

MR即平均排名,计算公式如下:

$$\mathrm{MR}=\frac{1}{|T|}\sum_{i=1}^{|T|}\mathrm{rank}_i. \tag{18}$$

其中: T 为三元组集合, $|T|$ 为三元组集合的个数, rank_i 为第 i 个三元组的得分在所有候选三元组中的排名. MR指标越小表明模型的性能越好.

MRR即平均倒数排名,计算公式如下:

$$\mathrm{MRR}=\frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{\mathrm{rank}_i}. \tag{19}$$

符号含义与MR相同. MRR指标越大表明模型的性能越好.

Hits@n指排名小于等于 n 的三元组的平均占比,计算公式如下:

$$\mathrm{Hits@}n=\frac{1}{|T|}\sum_{i=1}^{|T|}\mathbb{I}(\mathrm{rank}_i\leqslant n). \tag{20}$$

其中 $\mathbb{I}(\cdot)$ 为indicator函数,若条件为真则函数值为1,否则函数值为0. Hits@n指标越大表明模型的性能越好,本文选取 n 为1、3和10.

3.4 参数设置

本文方法需要设置的主要超参数包括实体和关系向量维度d、正负例三元组边界距离γ、注意力机制头数S以及与评分函数相关的权重参数λ_{cm}、λ_{cp}、λ_a的初始值等. 为了保证对比实验的公平性,各项实验中实体和关系向量维度均设置为100,其他超参数根据方法在验证集上的性能进行调优,调优得到的各项超参数设置如表2所示.

表 2　超参数设置

	DWY100K (DBpedia)	MED-BBK-9K	钢铁生产设备故障诊断
γ	5	5	5
S	2	2	2
λ_{cm} 初始值	1.0	1.0	1.0
λ_{cp} 初始值	1.0	0.5	1.0
λ_a 初始值	0.4	0.5	0.25

3.5 实验结果

各个方法在3个数据集上的实验结果见表3.

由表3可知, 除MED-BBK-9K数据集实验中Hits@1指标略低于ConvE方法、Hits@3指标略低于TransE方法以及钢铁生产设备故障诊断数据集实验中Hits@10指标略低于RotatE方法外,本文方法的各项性能指标均优于对比方法. 其中,MR指标与对比方法相比具有明显优势,表明本文方法预测得到的正确三元组的平均排名更加靠前,因此更加适用于容错率较低的领域知识图谱补全任务. 各项实验结果表明,本文方法对于具有完整模式层以及丰富属性信息的领域知识图谱的补全具有较好的效果.

3.6 消融实验

为了验证概念和属性信息融合的有效性,在3个数据集上分别进行消融实验. 在仅使用基于概念的向量表示和仅使用基于属性的向量表示的情况下分别进行训练和预测,实验结果如表4所示.

表 3　实验结果

	DWY100K (DBpedia)					MED-BBK-9K					钢铁生产设备故障诊断				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
TransE	101.4	0.569	0.490	0.618	0.693	593.2	0.126	0.064	**0.142**	0.244	20.2	0.483	0.346	0.577	0.750
TransH	100.3	0.574	0.504	0.621	0.696	581.5	0.126	0.066	0.136	0.243	13.5	0.473	0.327	0.538	0.788
RotatE	125.0	0.565	0.478	0.625	0.708	559.7	0.107	0.054	0.113	0.208	22.5	0.360	0.115	0.519	**0.865**
HAKE	198.8	0.546	0.470	0.595	0.667	778.7	0.071	0.026	0.069	0.154	20.7	0.418	0.269	0.500	0.750
ConvE	184.6	0.285	0.188	0.314	0.488	634.9	0.127	**0.073**	0.135	0.228	23.9	0.444	0.308	0.519	0.692
KBGAT	162.1	0.275	0.196	0.307	0.416	743.3	0.059	0.026	0.057	0.116	11.7	0.252	0.100	0.280	0.620
本文方法	**30.5**	**0.732**	**0.651**	**0.791**	**0.854**	**552.1**	**0.129**	0.068	0.140	**0.245**	**5.7**	**0.601**	**0.500**	**0.653**	0.808

表 4　消融实验结果

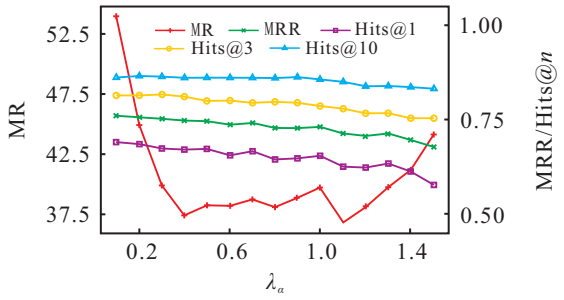
	DWY100K (DBpedia)					MED-BBK-9K					钢铁生产设备故障诊断				
	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10	MR	MRR	Hits@1	Hits@3	Hits@10
仅使用基于概念的向量表示	56.4	0.723	0.641	0.778	**0.856**	606.2	0.114	0.060	0.119	0.224	6.3	**0.624**	**0.500**	**0.730**	**0.808**
仅使用基于属性的向量表示	53.7	0.657	0.587	0.692	0.782	671.2	0.065	0.025	0.061	0.141	12.9	0.497	0.423	0.538	0.615
完整方法	**30.5**	**0.732**	**0.651**	**0.791**	0.854	**552.1**	**0.129**	**0.068**	**0.140**	**0.245**	**5.7**	0.601	**0.500**	0.653	**0.808**

由表4可知,同时使用基于概念和基于属性的向量时,大部分性能指标均优于仅使用基于概念的向量表示和仅使用基于属性的向量表示的情形,表明所提出方法可以有效融合概念和属性信息,并利用所融合的信息提升对相关实例的表示效果,从而提高实例三元组的补全性能.

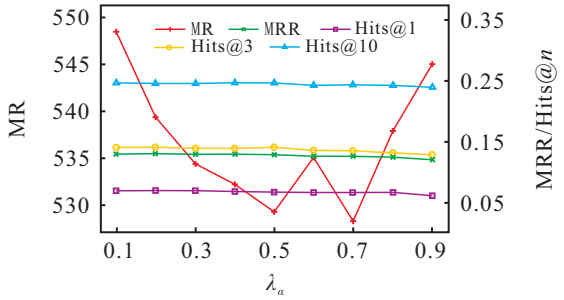
3.7 权重参数分析

式(16)所示的评分函数中权重参数λ_m、λ_p和λ_a的值对模型的性能具有重要影响. 其中,λ_m、λ_p决定了基于概念的向量表示部分的权重大小,初始值根据

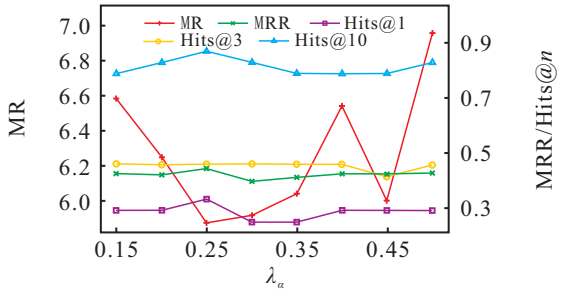
模式层预训练得到的λ_{cm}和λ_{cp}进行设置,由于模式层的规模通常较小,结构较为简单,λ_{cm}和λ_{cp}的初始取值在一定范围内调整时,均可实现对模式层三元组的准确表示,因此无需对其初始值进行严格的调节. 而基于属性的向量表示部分的权重λ_a需要在正式训练时手动设置,其初始值会影响到方法的最终性能. 图4展示了其他参数不变的情况下,λ_a初始值变化时方法在验证集上的性能指标变化情况.



(a) DWY100K(DBpedia)



(b) MED-BBK-9K



(c) 钢铁生产设备故障诊断

图 4 性能指标随λ_a初始值的变化情况

由图4可知,随着λ_a初始值的增大,各个数据集上的MR指标均呈现出先减小再增大的趋势,而MRR和Hits@n指标仅出现了小幅变化,因此可根据验证集上MR指标的较小值选取λ_a的初始值. MR指标的变化规律表明概念和属性信息任何一部分的相对权重过低都会对方法的性能造成影响,进一步表明了方法的性能是概念和属性信息综合作用的结果.

4 结 论

与通用知识图谱相比,领域知识图谱具有知识准确度要求高、实例属性丰富且重要等特点. 本文针对领域知识图谱的特点提出了一种融合概念和属性信息的领域知识图谱补全方法. 首先通过HAKE模型对知识图谱模式层中的概念和关系进行建模,并基于概念向量建立基于概念的实例向量表示;其次通过注意力机制对实例的属性和概念信息进行融合,建立基于属性的实例向量表示;最后对基于概念和基于属性的评分函数加权结合进行训练. 该方法能够充分利用领域知识图谱模式层对数据层的规范作用,同时对数据层的实例和属性进行了区分,通过对概念和属性信息的融合提升相关实例向量的表示效果. 在多个领域知识图谱数据集上的实验结果表明,所提出方法能够有效地对包含模式层和丰富属性信息的领域知识图谱进行补全. 在未来的研究中将进一步探索更加合适的模式层建模方法,并考虑将属性内容的语义信息引入知识图谱补全模型,进一步提升领域知识图谱补全效果.

参考文献(References)

[1] Singhal A. Introducing the knowledge graph: Things, not strings[EB/OL]. (2012-05-16)[2023-04-12]. https://blog.google/products/search/introducing-knowledge-graph-things-not/.

[2] 王硕, 杜志娟, 孟小峰. 大规模知识图谱补全技术的研究进展[J]. 中国科学: 信息科学, 2020, 50(4): 551-575.
(Wang S, Du Z J, Meng X F. Research progress of large-scale knowledge graph completion technology[J]. Scientia Sinica: Informationis, 2020, 50(4): 551-575.)

[3] Han H H, Wang J, Wang X W, et al. Construction and evolution of fault diagnosis knowledge graph in industrial process[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.

[4] 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.
(Wang H Y, Lun B, Zhang X M, et al. Multi-modal entity alignment based on joint knowledge representation learning[J]. Control and Decision, 2020, 35(12): 2855-2864.)

[5] 牟天昊, 李少远. 流程工业控制系统的知识图谱构建[J]. 智能科学与技术学报, 2022(1): 129-141.
(Mou T H, Li S Y. Knowledge graph construction for control systems in process industry[J]. Chinese Journal of Intelligent Science and Technology, 2022(1): 129-141.)

[6] Bordes A, Usunier N, Garcia-Durán A, et al. Translating embeddings for modeling multi-relational data[C]. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, 2013: 2787-2795.

[7] Wang Z, Zhang J W, Feng J L, et al. Knowledge

graph embedding by translating on hyperplanes[C]. Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec City, 2014: 1112-1119.

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

[9] Sun Z Q, Deng Z H, Nie J Y, et al. RotatE: Knowledge graph embedding by relational rotation in complex space[J/OL]. 2019, arXiv: 1902.10197.

[10] Zhang Z Q, Cai J Y, Zhang Y D, et al. Learning hierarchy-aware knowledge graph embeddings for link prediction[C]. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 3065-3072.

[11] 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 International Conference on Machine Learning. Bellevue, 2011: 809-816.

[12] Yang B S, Yih W T, He X D, et al. Embedding entities and relations for learning and inference in knowledge bases[J/OL]. 2014, arXiv: 1412.6575.

[13] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D knowledge graph embeddings[C]. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 1811-1818.

[14] Nathani D, Chauhan J, Sharma C, et al. Learning attention-based embeddings for relation prediction in knowledge graphs[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019: 4710-4723.

[15] Lv X, Hou L, Li J Z, et al. Differentiating concepts and instances for knowledge graph embedding[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, 2018: 1971-1979.

[16] Guan N N, Song D D, Liao L J. Knowledge graph embedding with concepts[J]. Knowledge-Based Systems, 2019, 164: 38-44.

[17] Li Z, Liu X, Wang X, et al. TransO: A knowledge-driven representation learning method with ontology information constraints[J]. World Wide Web, 2023, 26(1): 297-319.

[18] Zhang Z W, Cao L, Chen X L, et al. Representation learning of knowledge graphs with entity attributes[J]. IEEE Access, 2020, 8: 7435-7441.

[19] Lin Y K, Liu Z Y, Sun M S. Knowledge representation learning with entities, attributes and relations[C]. Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York: ACM, 2016: 2866-2872.

[20] Su X R, Hu L, You Z H, et al. Attention-based knowledge graph representation learning for predicting drug-drug interactions[J]. Briefings in Bioinformatics, 2022, 23(3): bbac140.

[21] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J/OL]. 2018, arXiv: 1710.10903.

[22] Sun Z Q, Hu W, Zhang Q H, et al. Bootstrapping entity alignment with knowledge graph embedding[C]. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, 2018: 4396-4402.

[23] Zhang Z H, Liu H L, Chen J Y, et al. An industry evaluation of embedding-based entity alignment[C]. Proceedings of the 28th International Conference on Computational Linguistics: Industry Track. Stroudsburg, 2020: 179-189.

作者简介

陈伯谦 (1999−), 男, 硕士生, 从事工业知识图谱的研究, E-mail: 2133000@tongji.edu.cn;

王坚 (1961−), 男, 教授, 博士, 博士生导师, 从事工业大数据与工业互联网、工业知识图谱与知识服务等研究, E-mail: jwang@tongji.edu.cn.