

DOI: 10.13718/j.cnki.xdzk.2020.11.004

“大数据分析 with 智能计算方法”专题

海量数据价值的挖掘,取决于相应的智能算法工具.知识图谱被广泛应用在智能推荐和智能问答等场景,而利用领域数据构建和补全知识图谱则是其应用的先决条件.聚类是数据挖掘中的基础问题,数据复杂化带来的超多类簇进一步提高了传统聚类问题的难度.空间并置挖掘是分析时空数据中的重要手段.本期“大数据分析 with 智能计算方法”专题,刊载了 2020 年第八届中国计算机学会大数据学术会议(CCF BIGDATA2020)推荐的 5 篇论文,对以上问题进行了探讨,并基于不同的数据类型和应用场景提出了性能优良的算法工具.

航空安全事件知识图谱补全方法

王 红, 卢林燕, 王 童

中国民航大学 计算机科学与技术学院, 天津 300300

摘要: 针对知识图谱稀疏实体间关系缺失和学习效率低的问题,提出了一种改进的知识图谱补全方法 I_IterE (improved-iteratively learning embeddings and rules). 首先,在 IterE 模型的表示学习层与规则学习层之间增加了特征层,采用 add 方法将表示学习层学到的关系特征进行增强,并生成新关系特征;然后,在规则学习层将新的关系特征与改进的评分机制相结合进行公理的分值计算,该分值用于判断公理的可信度;最后,利用高分公理进行演绎推理从而获得与稀疏实体相关的新三元组.采用公共数据集 WN18-sparse, WN18RR-sparse, FB15k-sparse 和 FB15k-237-sparse 进行链接预测和规则评估.实验结果表明,该方法在 MRR、Hits@n 和规则评估指标上均取得较好的提升,将该方法应用到航空安全事件数据集上,在显著提升链接预测效果的同时,为航空安全事件知识图谱的补全提供了方法支撑.

关 键 词: 知识图谱;关系特征;评分机制;链接预测;规则评估;航空安全事件

中图分类号: TP391

文献标志码: A

文章编号: 1673-9868(2020)11-0031-12

知识图谱(knowledge graph)是一种揭示实体之间关系的语义网络,可以对现实世界的事物及其相互关系进行形式化地描述^[1],一般表示为三元组(头实体、关系、尾实体)的形式^[2].如 WordNet^[3], Freebase^[4]和 YAGO^[5]等将知识图谱广泛应用于问答系统、智能搜索和个性推荐等领域,但是很多大型开放知识图谱都是由人工或半自动的方式构建的,这些图谱通常比较稀疏,大量实体之间隐含的关系未被充分挖掘出来^[6],因此知识图谱的补全方法成为新的研究热点.

表示学习是实现知识图谱补全的有效方法,它将知识图谱的实体和关系映射到连续的向量空间,利用

收稿日期: 2020-10-14

基金项目: 国家重点实验室开放基金资助项目(SKLATM201902);国家自然科学基金项目(U1633110).

作者简介: 王 红(1963—),女,教授,主要从事研究语义数据挖掘与知识图谱研究.

向量计算方式推出实体之间潜在的语义关系,使得计算复杂度降低^[7].表示学习中常用的两个模型是翻译模型和线性模型^[8],其中翻译模型包括 TransE 模型^[9]和基于 TransE 的一系列改进模型,如 TransH 模型^[10]、TransR 模型^[11]和 TransD 模型^[12]等.线性模型包括 DistMult 模型^[13]、ComplEx 模型^[14]和 ANALOGY 模型^[15]等,线性模型还具有良好的规则学习属性.但这两种模型对具有较少关系的实体(稀疏实体)表示效果较差^[16],因为好的表示依赖于丰富的知识.规则学习也是知识图谱补全的重要手段,主要利用规则的演绎能力进行关系推演,可以保证结果的可靠性和可解释性.规则学习包括基于随机行走的规则挖掘^[17-18]、知识图谱关联规则挖掘算法 AMIE^[19]及改进的 AMIE+^[20]等基于图结构的规则学习方法.这些方法虽然在结构上加快了推理速度^[21],一定程度上缓解了规则学习的高复杂问题,但是仍存在规则搜索空间大、无法较好评估规则和仅考虑知识图谱结构等问题.

2019 年, Zhang 等人^[22]提出了 IterE 模型,首次利用具有丰富语义 OWL2 中的对象属性表达公理进行演绎推理,有效提升了稀疏实体的表示效果和规则学习的效率.但该模型仅包含表示学习层和规则学习层,其泛化能力较弱,且模型中所采用的 Frobenius 范数计算方式复杂、无法显著区分高分公理.在航空安全事件知识图谱的研究方面,王红等人^[23-24]研究了民航突发事件实体与关系的自动抽取方法,初步解决了航空安全事件知识图谱的构建问题,但未充分考虑实例层和概念层(本体)内稀疏实体之间隐含的语义关系.

为此,本研究提出一种改进的知识图谱补全方法(I_IterE),通过增加特征层和改进的评分机制获得高分公理进行演绎推理,旨在更好地利用高分公理进行演绎推理和链接预测,实现航空安全事件知识图谱的进一步补全,为基于知识图谱的航空安全事件语义分析提供更好的数据支持.

1 I_IterE 模型

IterE 模型包括表示学习层和规则学习层,通过表示学习层获得的关系特征,结合规则学习层挖掘的公理集,采用相应的评分机制获得高分公理进行演绎推理.为了提升模型的泛化能力和规则学习的效率,本研究提出的改进模型 I_IterE(图 1).图 1 中:

- (1) 表示学习层 利用线性模型将知识图谱中的实体和关系映射到向量空间中.
- (2) 特征增强层 对表示学习层的关系特征利用 add^[25]方法对关系特征叠加从而增强特征.
- (3) 规则学习层 将生成的公理集与增强后的关系特征采用改进的评分机制对公理进行评分,获得高分公理进行演绎推理,从而实现知识图谱的更新.

1.1 表示学习层

知识图谱可以描述为 $G = \{E, R, T\}$, 其中 E 为实体的集合, R 为关系的集合, $T = \{(h, r, t) | h, t \in E; r \in R\}$ 为三元组的集合.在三元组 (h, r, t) 中, h, r 和 t 分别为头实体, 关系和尾实体.表示学习层的输入集 I 表示为:

$$I = \{((h, r, t), l_{hrt}) | (h, r, t) \in T \wedge T_{\text{axiom}} \wedge T_{\text{negative}}\} \quad (1)$$

其中: l_{hrt} 是三元组的标签,用以评估三元组的可靠性; T_{axiom} 是规则学习层生成的三元组; T_{negative} 是通过将 $(h, r, t) \in T$ 的 h 或 t 随机替换为 $e \in E$ 或者将 r 替换为 $r' \in R$ 生成的负样本集.

若 $(h, r, t) \in T$, 则 $l_{hrt} = 1$; 若 $(h, r, t) \in T_{\text{negative}}$, 则 $l_{hrt} = 0$; 若 $(h, r, t) \in T_{\text{axiom}}$, 则 $l_{hrt} = \pi(h, r, t)$. 其中 $\pi(h, r, t)$ 是由公理演绎预测的标签值.

对于输入集 I , 采用交叉熵作为损失函数, 即如

$$\min L_{\text{representation}} = \frac{1}{n} \sum_{((h, r, t), l_{hrt}) \in I} [-l_{hrt} \log(\varphi(h, r, t)) - (1 - l_{hrt}) \log(1 - \varphi(h, r, t))] \quad (2)$$

其中: n 表示输入的三元组数量, $\varphi(h, r, t)$ 采用线性模型中表现较好的 ANALOGY 模型的得分函数, 表示为

$$\varphi(h, r, t) = \sigma(\mathbf{V}_h^T \mathbf{M}_r \mathbf{V}_t) \quad (3)$$

其中: $\mathbf{V}_h \in \mathbb{R}^{1 \times d}$, $\mathbf{V}_t \in \mathbb{R}^{1 \times d}$ 是 h 和 t 的向量表示, $\mathbf{M}_r \in \mathbb{R}^{d \times d}$ 是 r 的矩阵表示, d 是表示维度, σ 是 sigmoid 函数。

经过表示学习层之后, 可得到实体特征的集合 ε 和关系特征集合 $\mathbb{R}$ 。

1.2 特征增强层

由于公理的评估以关系特征为基础, 而演绎推理主要依据高分公理, 因此关系特征具有重要意义。

泛化能力是指模型对于新鲜样本(测试集)的预测能力, 提升模型泛化能力的方法包括使用更多的数据、数据过采样、修改损失函数、改进模型深度和权重惩罚等, 本研究采用改进模型深度的方法, 通过增加特征层来提升模型泛化能力。

特征增强的具体过程: 以某一公理包括 M_1 和 M_2 两个关系特征的公理为例, 为避免不同语义信息的关系特征混杂

产生噪音影响公理的评估, 采用 add 方法^[25]对关系特征进行特征增强。如图 2 所示, add 方法将关系特征中元素对应相加, 并且相加后特征维度不变, 只是每一维下的信息量在增加, 进而增强语义信息较强的区域, 通过特征重用提升最终的链接预测结果。其中, 关系特征层数的选择通过链接预测的结果来确定, 层数为 3 层时可使模型链接预测结果达到最佳。由于 ANALOGY 模型获得的关系特征对角相等, 经特征增强后仍相等, 因而仅根据关系特征的某一对角的相关值进行计算即可, 利用 Python 中的 split 函数对增强后的关系特征进行分块(分片)处理, 可以获得代表整个特征的块对角特征, 极大降低计算复杂度。

1.3 规则学习层

规则学习层包括公理集生成、改进的评分机制和演绎推理三部分。首先依据 OWL2 的 7 种对象属性表达公理挖掘出知识图谱中存在的公理集, 之后结合特征增强层的关系特征和改进的评分机制对各公理进行评分计算, 最后根据高分公理进行演绎推理。

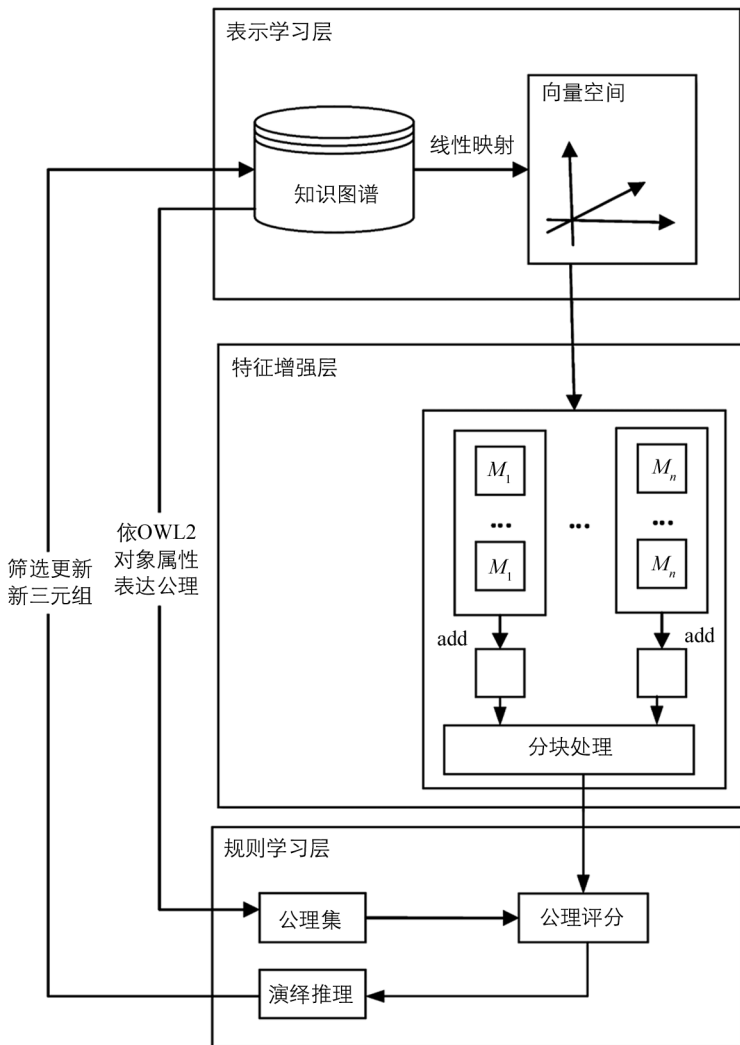


图 1 I_IterE 模型框架

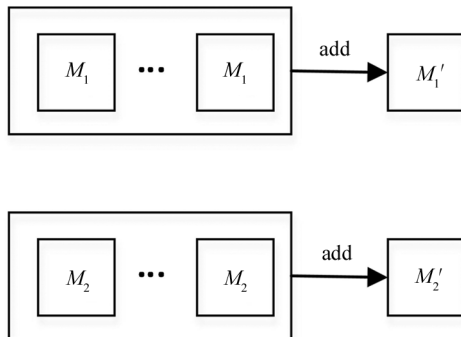


图 2 特征增强过程

1.3.1 公理集生成

OWL2 中 7 种对象属性表达公理见表 1 所述.

表 1 OWL2 对象属性表达公理及规则

对象属性表达公理	规则形式
ReflexiveOP(r)	(x, r, x)
SymmetricOP(r)	$(x, r, y) \rightarrow (y, r, x)$
TransitiveOP(r)	$(x, r, y), (y, r, z) \rightarrow (x, r, z)$
EquivalentOP(r, r_1)	$(x, r, y) \rightarrow (x, r_1, y)$
SubOP(r, r_1)	$(x, r, y) \rightarrow (x, r_1, y)$
InverseOP(r, r_1)	$(y, r, x) \rightarrow (x, r_1, y)$
SubOP(OPChain(r_1, r_2), r))	$(x_0, r_1, x_1), (x_1, r_2, x_2) \rightarrow (x_0, r, x_2)$

表 1 中, r, r_1, r_2 表示关系变量, x, x_0, x_1, x_2, y, z 表示实体变量.

将知识图谱每个关系 $\gamma \in R$ 生成图谱中可能存在的公理主要分为 3 个步骤.

步骤 1: 生成带关系变量的公理 ReflexiveOP(γ), SymmetricOP(γ), TransitiveOP(γ), EquivalentOP(γ, r'), SubOP(γ, r'), InverseOP(γ, r')和 SubOP(OPChain(r', r''), γ), 其中 $\{r', r''\}$ 需用知识图谱中实际关系来替换的关系变量.

步骤 2: 知识图谱中随机选择与 γ 有关的 k 个三元组 $(e', \gamma, e'') \in T$, 将直接链接到 e' 或者 e'' 的关系替换步骤 1 公理中的 r' 或 r'' , 从而具体化步骤 1 的公理.

其中 k 应尽可能小, 并使公理集能覆盖所有可能公理的概率高于公理覆盖阈值, k 的约束条件为

$$k > N - N(1 - t)^{\frac{1}{p^N}}$$

(4)

其中: N 是三元组 $(e', \gamma, e'') \in T$ 的数目, p 为最小公理概率, 存在概率大于 p 的公理为高可能性公理, t 为公理覆盖阈值. 因此在固定的 p, t 的情况下, k 的最佳选择应该是方程 $f(N) = N - N(1 - t)^{\frac{1}{p^N}}$ 的上限.

步骤 3: 在知识图谱中查找并统计每个公理的支持数, 将支持数大于等于 1 的公理添加到公理集 Ω .

由于输入的知识图谱是固定的, 因此公理集 Ω 仅需生成一次.

1.3.2 改进的评分机制

在评分处理之前, 首先将表 1 中的规则用表示学习层学到的实体特征和关系特征进行表示, 并归纳得出相关规则结论(表 2).

表 2 对象属性表达公理向量空间中的表示及结论

对象属性表达公理	规则形式	线性映射	规则结论
ReflexiveOP(r)	(x, r, x)	$V_x M_r = V_x$	$M_r = I$
SymmetricOP(r)	$(x, r, y) \rightarrow (y, r, x)$	$V_x M_r = V_y, V_y M_r = V_x$	$M_r M_r = I$
TransitiveOP(r)	$(x, r, y), (y, r, z) \rightarrow (x, r, z)$	$V_x M_r = V_y, V_y M_r = V_z, V_x M_r = V_z$	$M_r M_r = M_r$
EquivalentOP(r, r_1)	$(x, r, y) \rightarrow (x, r_1, y)$	$V_x M_r = V_y, V_x M_{r_1} = V_y$	$M_r = M_{r_1}$
SubOP(r, r_1)	$(x, r, y) \rightarrow (x, r_1, y)$	$V_x M_r = V_y, V_x M_{r_1} = V_y$	$M_r = M_{r_1}$
InverseOP(r, r_1)	$(y, r, x) \rightarrow (x, r_1, y)$	$V_y M_r = V_x, V_x M_{r_1} = V_y$	$M_r M_{r_1} = I$
SubOP(OPChain(r_1, r_2), r))	$(x_0, r_1, x_1), (x_1, r_2, x_2) \rightarrow (x_0, r, x_2)$	$V_{x_0} M_{r_1} = V_{x_1}, V_{x_1} M_{r_2} = V_{x_2}, V_{x_0} M_r = V_{x_2}$	$M_{r_1} M_{r_2} = M_r$

例如, 对于公理 InverseOP(hasWife, hasHusband), 如果知识图谱中存在(Carl, hasWife, Mary), 根据表 1 的规则, 可以推出另一个三元组(Mary, hasHusband, Carl). 假设 Mary, Carl, hasWife 和 hasHus-

band 的表示分别为 $\mathbf{V}_{\text{Mary}}$, $\mathbf{V}_{\text{Carl}}$, $\mathbf{M}_{\text{hasWife}}$ 和 $\mathbf{M}_{\text{hasHusband}}$, 对于上述两个三元组, 根据线性映射假设, 可得到两个等式: $\mathbf{V}_{\text{Carl}}\mathbf{M}_{\text{hasWife}} = \mathbf{V}_{\text{Mary}}$, $\mathbf{V}_{\text{Mary}}\mathbf{M}_{\text{hasHusband}} = \mathbf{V}_{\text{Carl}}$. 依据上述等式可推出: $\mathbf{M}_{\text{hasWife}}\mathbf{M}_{\text{hasHusband}} = \mathbf{I}$, 其中 $\mathbf{I}$ 表示单位矩阵, 该等式表明仅与两个对应的关系特征有关, 与具体的实体特征无关, 因此可将其视为公理 $\text{InverseOP}(\text{hasWife}, \text{hasHusband})$ 的一般结论. 表 2 给出了每个公理规则结论的详细信息.

表 2 中 $\mathbf{V}$ 和 $\mathbf{M}$ 分别表示实体特征和关系特征, $\mathbf{I}$ 是单位矩阵.

给定关系特征的集合 $\mathbb{R}$ 和公理集 Ω , 评分机制是基于每个公理类型的规则结论为公理 $a \in \Omega$ 计算得分. 以 $\mathbf{M}_1^a = \mathbf{M}_2^a$ 的形式来表示规则结论, 其中 $\mathbf{M}_1^a$ 和 $\mathbf{M}_2^a$ 是单个关系特征或两个关系特征相乘后的特征. 由于规则结论是从理想的线性映射中得到的, 但表示学习层学到的关系特征各不相同. 因此, 通过比较 $\mathbf{M}_1^a$ 和 $\mathbf{M}_2^a$ 之间的相似性对公理 a 进行评分.

IterE 模型采用 Frobenius 范数进行评分, 其计算公式为

$$s_a(F) = \|\mathbf{M}_1^a - \mathbf{M}_2^a\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^d |\mathbf{M}_{1ij}^a - \mathbf{M}_{2ij}^a|^2} \quad (5)$$

其中: d 表示关系特征 $\mathbf{M}_1^a$ 和 $\mathbf{M}_2^a$ 的维度. 表示学习层经 ANALOGY 方法学习后关系特征对角块矩阵相同, 因此在(5)式中重复计算对角块特征, 目的是增大关系特征之间相似与不相似的差距. 但关系特征中各值均在 $[-0.01, 0.01]$ 范围内, 关系特征之间的差距经平方后会被大幅度缩小, 从而无法较好过滤低分公理. 为此, 改进评分公式为

$$s'_a(F') = \|\mathbf{M}_1^a - \mathbf{M}_2^a\|_{F'} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d |\mathbf{M}_{1ij}^a - \mathbf{M}_{2ij}^a| \times 2} \quad (6)$$

其中: F' 表示改进的 Frobenius 范数, $s'_a(F')$ 表示通过改进的 Frobenius 范数计算出公理 a 的得分. (6) 式将(5)式中的平方换成常量线性乘法, 因为在 $[-0.01, 0.01]$ 范围内, 常量线性乘法比平方更能较好地增大高分公理与低分公理的差距, 利于筛选, 进而有效进行演绎推理. 但是因不同类型的公理 $s'_a(F')$ 值变化很大, 因此需将 $s'_a(F')$ 归一化, 等到

$$s_a = \frac{s'_{\max}(t) - s'_a(F')}{s'_{\max}(t) - s'_{\min}(t)} \quad (7)$$

其中: t 是公理 a 所属的公理类别, $s'_{\max}(t)$ 和 $s'_{\min}(t)$ 是公理集 Ω 中所有类型 t 的最大和最小得分. $s'_a \in [0, 1]$ 是公理 a 的最终得分, s'_a 越高, 公理 a 的可信度越高.

1.3.3 公理演绎

给定知识图谱 G 和公理集 Ω , 首先利用公理的演绎能力推出新三元组, 并预测其标签, 然后过滤掉与稀疏实体无关的三元组, 最后将 T_{axiom} 更新到知识图谱中.

利用公理的基本形式进行演绎推理, 如公式(8)所示.

$$(h_1, r_1, t_1), (h_2, r_2, t_2), \dots, (h_n, r_n, t_n) \rightarrow (h^a, r^a, t^a) \quad (8)$$

其中 $(h_i, r_i, t_i) \in T$ 且 $i \in [1, n]$ 而 $(h^a, r^a, t^a) \notin T$ 需添加到知识图谱中. 为了预测 (h^a, r^a, t^a) 的标签, 首先将(8)式转换为

$$(h_1, r_1, t_1) \wedge (h_2, r_2, t_2) \wedge \dots \wedge (h_n, r_n, t_n) \Rightarrow (h^a, r^a, t^a) \quad (9)$$

基于 t 范数的模糊逻辑^[26]对公式(9)进行计算, 例如 $(h_1, r_1, t_1) \Rightarrow (h_2, r_2, t_2)$ 的标签由三元组 (h_1, r_1, t_1) 和 (h_2, r_2, t_2) 的标签确定, 这是通过逻辑表示符号 $\Rightarrow$ 定义, 且遵循逻辑与、或和非的逻辑定义:

$$\pi(b \wedge c) = \pi(b) \cdot \pi(c) \quad (10)$$

$$\pi(b \vee c) = \pi(b) + \pi(c) - \pi(b) \cdot \pi(c) \quad (11)$$

$$\pi(\neg b) = 1 - \pi(b) \quad (12)$$

$$\pi(b \Rightarrow c) = \pi(\neg b \vee c) \quad (13)$$

其中: b 和 c 是逻辑表达, $\pi(x)$ 是 x 的标签. 基于上述条件, 可通过(10)–(13)式计算任何规则的标签. 例如 $g_a: (h_1, r_1, t_1) \wedge (h_2, r_2, t_2) \Rightarrow (h^a, r^a, t^a)$:

$$\pi(g_a) = 1 - \pi(h_1, r_1, t_1)\pi(h_2, r_2, t_2) + \pi(h_1, r_1, t_1)\pi(h_2, r_2, t_2)\pi(h^a, r^a, t^a)$$

因此, 可得到

$$\pi(h^a, r^a, t^a) = \frac{\pi(g_a) - 1 + \pi(h_1, r_1, t_1)\pi(h_2, r_2, t_2)}{\pi(h_1, r_1, t_1)\pi(h_2, r_2, t_2)}$$

将(7)式计算的公理 a 的得分 s'_a 作为 g_a 的标签, 即 $\pi(g_a) = s'_a$. 知识图谱中存在的三元组标签依据表示学习层的标签进行标记. 若 $(h, r, t) \in T$, 则 $\pi(h, r, t) = 1$. 因此, 对于任何类型的公理 a , 可知三元组 (h^a, r^a, t^a) 的标签为:

$$\pi(h^a, r^a, t^a) = s'_a \tag{14}$$

通过高分公理推出一组新三元组的集合 $T_{\text{axiom}} = \{(h^a, r^a, t^a) | h^a \in E_{\text{sparse}} \text{ or } t^a \in E_{\text{sparse}}\}$, 其中 E_{sparse} 表示知识图谱的稀疏实体的集合. 实体的稀疏性为

$$r(e) = 1 - \frac{f(e) - f_{\min}}{f_{\max} - f_{\min}} \tag{15}$$

其中: $f(e)$ 是实体作为头实体或尾实体出现在图谱中的频率, $f_{\max}$ 和 $f_{\min}$ 是所有实体里出现在图谱中的最大和最小频率. $r(e) \in [0, 1]$, 若 $r(e) = 0$, 则表示 e 是图谱中出现最频繁的实体; 若 $r(e) = 1$, 则表示 e 是图谱中出现频率最小的实体. 当 $r(e) > \theta_{\text{sparsity}}$, 将实体 e 作为稀疏实体, 其中 θ_{sparsity} 是稀疏阈值. 以此推出所有可能的新三元组后, 并过滤掉与稀疏实体无关的三元组, 可以实现知识图谱的更新.

1.4 实验效果分析

1.4.1 实验数据与参数设置

实验采用 WN18-sparse, WN18RR-sparse, FB15k-sparse 和 FB15k-237-sparse 数据集^[22], 它们是由知识图谱补全常用的公开数据集 WN18 和 FB15k 生成的^[17]. 表 3 中给出了各数据集的实体、关系、训练集、验证集和测试集的数量.

表 3 数据集统计信息

Dataset	WN18-sparse	WN18RR-sparse	FB15k-sparse	FB15k-237-sparse
# ent	40 943	40 943	14 951	14 541
# rel	18	11	1 345	237
# Train	141 442	86 835	483 142	272 115
# Valid	3 624	1 609	18 544	10 671
# Test	3 590	1 661	22 013	12 454

表示学习层采用均匀分布 $U(-0.01, 0.01)$ 初始化实体和关系. 规则学习层中的最小公理概率一般取 $p = 0.5$, 公理覆盖阈值 $t = 0.95$, 由(4)式可知每个关系选择的相关三元组的数量设置为 $k = 6$. 对于公理演绎, 评分高的公理更可靠且不易产生噪声, 因此为每个数据集设置阈值 θ 为 0.9, 并将 $s'_{\text{axiom}} \geq \theta$ 的公理作为高分公理.

对于 WN18-sparse 和 FB15k-sparse, 最大训练迭代次数设置为 100, 而 WN18RR-sparse 和 FB15k-237-sparse 的最大训练迭代次数设置为 150. 采用 Adam 优化器^[27], 其优势能够在训练过程中自动调整恰当的学习率, 使得模型能够较快收敛, 并适用于解决包含高噪声或稀疏梯度的问题^[28]. 结合表示维度 $d \in (100, 200)$ 和 L1 正则化权重 $\lambda \in \{10^{-4}, 10^{-5}, 10^{-6}\}$, 最终在验证集中表现最好的实验参数见表 4.

表 4 实验参数设置

实验参数	WN18-sparse	WN18RR-sparse	FB15k-sparse	FB15k-237-sparse
d	$d=200$	$d=200$	$d=200$	$d=200$
λ	$\lambda=10^{-5}$	$\lambda=10^{-4}$	$\lambda=10^{-5}$	$\lambda=10^{-4}$
θ	$\theta=0.9$	$\theta=0.9$	$\theta=0.9$	$\theta=0.9$
lr	$lr=0.001$	$lr=0.001$	$lr=0.001$	$lr=0.001$
max_epoch	$max_epoch=100$	$max_epoch=150$	$max_epoch=100$	$max_epoch=150$

1.4.2 链接预测

链接预测是对知识图谱中的实体或关系进行预测. 链接预测实验采用平均倒数排名(MRR)的原始设置和过滤设置, Hits@1, Hits@3 和 Hits@10 在过滤设置情况下的得分作为模型性能评价指标, 通常较高的 MRR 或 Hits@n 表示具有较好的预测效果.

对比模型考虑了表示学习方法中翻译模型的经典模型 TransE 和线性模型 DistMult, ComplEx, ANALOGY 以及基础模型 IterE. WN18-sparse, WN18RR-sparse, FB15k-sparse 和 FB15k-237-sparse 在不同方法上的链接预测结果见表 5 和表 6.

表 5 WN18-sparse 和 FB15k-sparse 上链接预测实验结果

模型	WN18-sparse					FB15k-sparse				
	MRR		Hits@			MRR		Hits@		
	Raw	Filter	1	3	10	Raw	Filter	1	3	10
TransE	33.5	41.8	10.2	71.1	84.7	25.5	39.8	25.8	48.6	64.5
DistMult	55.8	73.8	59.3	87.5	93.1	32.4	60.0	53.8	65.1	75.9
ComplEx	67.7	91.1	89.0	93.3	94.4	32.7	61.6	54.0	65.7	76.1
ANALOGY	67.5	91.3	89.0	93.4	94.4	33.1	62.0	54.3	66.1	76.3
IterE	78.9	91.3	89.1	93.5	94.8	38.8	62.8	55.1	67.3	77.1
I_IterE	80.0	92.5	91.1	95.0	96.7	40.1	63.8	56.5	68.8	79.1

表 6 WN18RR-sparse 和 FB15k-237-sparse 上链接预测实验结果

模型	WN18RR-sparse					FB15k-237-sparse				
	MRR		Hits@			MRR		Hits@		
	Raw	Filter	1	3	10	Raw	Filter	1	3	10
TransE	12.4	14.6	3.4	24.7	28.8	15.6	23.8	16.4	26.1	38.5
DistMult	20.8	25.5	23.8	26.0	22.5	12.9	20.4	12.8	22.6	36.2
ComplEx	21.4	25.9	24.6	26.2	28.6	13.3	19.7	12.0	21.7	35.4
ANALOGY	13.3	19.8	24.6	27.5	28.7	13.9	19.8	12.3	21.4	34.9
IterE	25.7	27.4	25.4	28.1	31.4	18.6	24.7	17.9	26.2	39.2
I_IterE	27.0	28.9	27.1	30.5	34.5	20.0	26.6	20.1	28.6	42.8

由表 5 和表 6 可知, 在各个数据集上, I_IterE 的链接预测结果均高于其他模型, 并且与模型 IterE 相比, 在 Hits@1, Hits@3 和 Hits@10 指标上, 最低高出了 1.0%, 最高高出了 3.4%, 达到了较好的实验效果. 表明增加特征层和改进的评分机制, 有效提升了模型的链接预测效果.

1.4.3 规则评估

规则学习的效率通过学习的时间评估, 而规则学习的质量通过高质量规则(HQr)的数量及其所占挖掘出的所有规则的比例(%). 规则的质量通常使用头覆盖率(HC)进行评估. 对于规则 rule 的头覆盖率定义如公式(16)所示.

$$HC(rule) = \frac{\#(e, e') : support(rule) \wedge head(rule)(e, e')}{\#(e, e') : head(rule)(e, e')}$$

(16)

其中 $head(rule)(e, e') = \{(e, r, e') \in T, support(rule)\}$ 是 $rule$ 的支持度, 并且将 $HC > 0.7$ 的规则视为高质量规则.

对比模型为规则学习中效果较好的 AMIE+^[20] 和基础模型 IterE^[22], 规则评估结果见表 7.

表 7 规则评估实验结果

模型	WN18-sparse			WN18RR-sparse			FB15k-sparse			FB15k-237-sparse		
	Time/s	HQr	比例/%	Time/s	HQr	比例/%	Time/s	HQr	比例/%	Time/s	HQr	比例/%
AMIE+	4.98	16	11.4	3.0	2	5.71	428	1 820	4.4	66.00	470	1.9
IterE	1.63	20	20.2	0.75	6	19.30	26.49	11 375	17.6	4.72	653	11.8
I_IterE	1.17	25	23.6	0.52	8	27.60	14.03	18 876	23.2	2.35	831	17.2

由表 7 可知, I_IterE 模型的规则学习效率均优于 IterE 模型和 AMIE+ 模型, 同时高质量的规则数量增加. 实验结果表明, 评分公式中的平方替换为常量线性乘法, 降低了计算复杂度, 筛选出较多的高质量规则.

2 基于 I_IterE 的航空安全事件知识图谱补全

航空安全事件知识图谱的数据集来自航空安全事故调查报告 2007-2020 年的 1 879 个航空安全事件, 共提取出 8 830 个实体、32 种关系和 20 000 个三元组. 将三元组按 6:2:2 比例分别抽取到训练集、测试集和验证集中. 参照公共数据集的实验设置进行实验, 选出在验证集中表现最好的链接预测结果.

2.1 实验过程

步骤 1: 采用均匀分布 $U(-0.01, 0.01)$ 向量初始化表示航空安全事件知识图谱的实体和关系, 再经 ANALOGY 模型训练得到较好表示知识图谱的实体特征的集合和关系特征的集合.

步骤 2: 按照 1.2 节所述方法, 对知识图谱的关系特征进行特征增强, 之后利用 Python 的 split 函数对新关系特征进行分块(分片)处理.

步骤 3: 依据 OWL2 的 7 种对象属性表达公理, 按照 1.3.1 节的 3 个步骤挖掘航空安全事件知识图谱中可能存在的公理集.

步骤 4: 结合步骤 2 获得的关系特征和步骤 3 的公理集, 采用 1.3.2 节改进的评分机制对各公理进行评分并归一化.

步骤 5: 从步骤 4 中筛选出大于等于 0.9 的高分公理推演出新三元组并根据公式(14)对三元组进行标记, 并将新三元组中实体属于稀疏实体的保留, 反之过滤. 最后更新航空安全事件知识图谱.

2.2 效果分析

将 I_IterE 模型应用于航空安全事件数据集进行链接预测, 并通过 MRR 和 Hits@n 来反映实验效果, 使用 IterE 模型与 I_IterE 模型进行比较, 实验结果见表 8.

表 8 航空安全事件数据集下的实验结果

模型	MRR	Hits@1	Hits@2	Hits@3
IterE	44.0	37.8	48.5	53.8
I_IterE	50.2	45.7	54.7	57.0

由表 8 可知, 在航空安全事件数据集上, I_IterE 模型的链接预测结果相比于 IterE 模型均有显著提高, 表明 I_IterE 模型利于航空安全事件知识图谱的补全. 针对航空安全事件知识图谱补全效果进行分析, 如下所述.

首先, 表示学习层需输入航空安全事件知识图谱(图 3).

其次, 依据 OWL2 的 7 种对象属性表达公理生成可能的公理集, 并结合增强后的关系特征与改进的评分机制对公理进行评分(表 9).

最后, 基于得分大于等于 0.9 的公理推演出新三元组见表 10, 加黑的三元组表示新三元组, 并将其更新到知识图谱(图 4), 虚线箭头表示新添加的两实体间存在的语义关系.

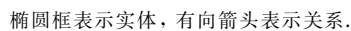


表 9 可能的公理集及得分情况

可能的公理	公理得分
SymmetricOP(subclassof, part_of)	0.9
InverseOP(subclassof, haspart)	0.9
InverseOP(cause_of, result_of)	0.9
SubOP(OPChain(type_of, subclassof), is_a)	0.9
SubOP(OPChain(is_a, haspart), type_of)	0.7
...	...

新三元组

0.9SymmetricOP(subclassof, part_of)	(沙赫乔斯特, part_of , 事故地点)←(沙赫乔斯特, subclassof, 事故地点)
0.9InverseOP(subclassof, haspart)	(航空器紧急事件, haspart , 航空失事)←(航空失事, subclassof, 航空器紧急事件)
0.9InverseOP(cause_of, result_of)	(山毛榉防空导弹击落, result_of , MH17)←(MH17, cause_of, 山毛榉防空导弹击落)
0.9SymmetricOP(subclassof, part_of)	(航空受到非法干扰, part_of , 航空器紧急事件)←(航空受到非法干扰, subclassof, 航空器紧急事件)
0.9SymmetricOP(subclassof, part_of)	(航空器紧急事件, part_of , 事故灾害类)←(航空器紧急事件, subclassof, 事故灾害类)
0.9InverseOP(subclassof, haspart)	(事故灾害类, haspart , 航空器紧急事件)←(航空器紧急事件, subclassof, 事故灾害类)
0.9SubOP(OPChain(type_of, subclassof), is_a)	(MH17, is_a , 航空器紧急事件)←(MH17, type_of, 航空器失事), (航空器失事, subclassof, 航空器紧急事件)
...	...

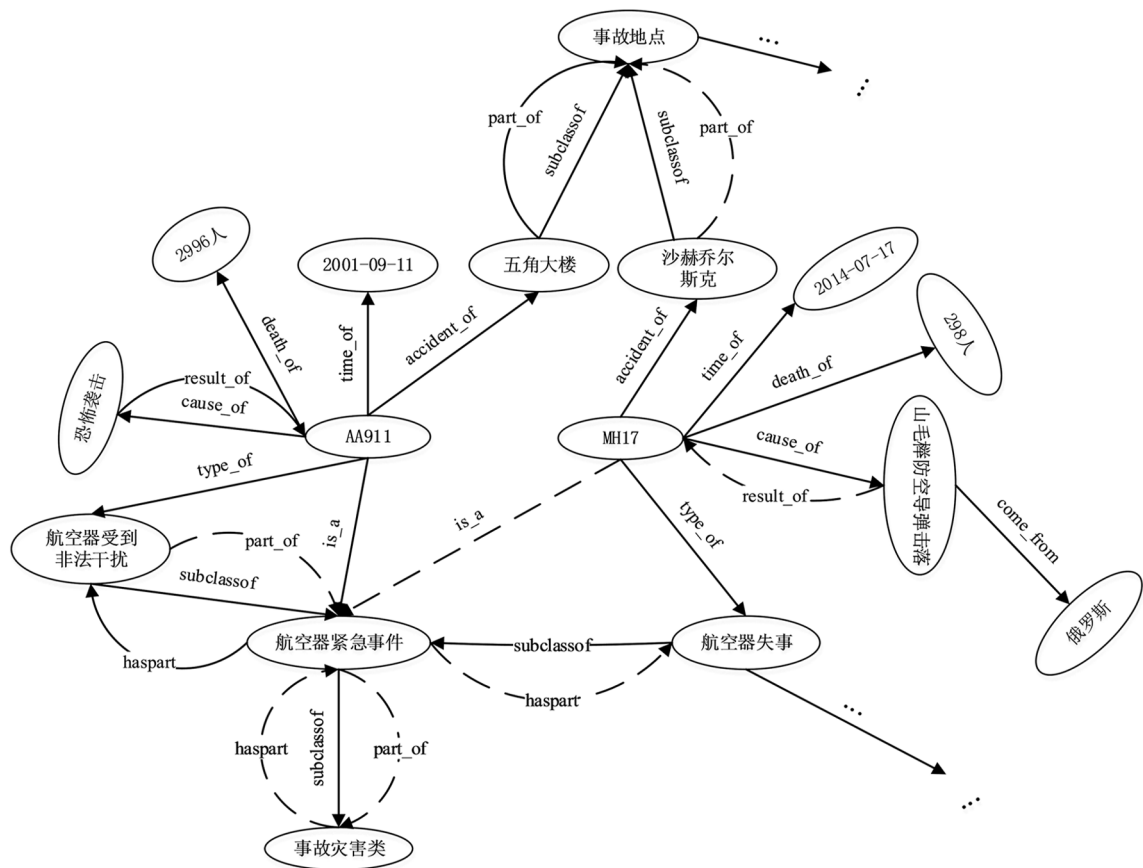


图 4 更新后的知识图谱

3 结束语

本研究提出了一种改进的知识图谱补全方法(I_IterE)，通过增加模型深度且改进评分机制，有效提升了链接预测的准确率和规则学习的效率，在获得较多高质量规则实现推理的同时，进一步提高了模型的泛化能力。将该方法应用于航空安全事件数据集，较好地解决了航空安全事件知识图谱的补全问题。未来围绕 OWL2 的数据属性表达公理和类属性表达公理，可以进一步研究知识图谱补全的优化方法。

参考文献：

[1] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述 [J]. 电子科技大学学报, 2016, 45(4): 589-606.

[2] 王 鑫, 邹 磊, 王朝坤, 等. 知识图谱数据管理研究综述 [J]. 软件学报, 2019, 30(7): 2139-2174.

[3] FELLBAUM, C, MILLER, G. WordNet ; an electronic lexical database [M]. WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 1998.

[4] 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 : ACM Press, 2008: 1247-1250.

[5] 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: ACM Press, 2007: 697-706.

[6] 饶子昀, 张 毅, 刘俊涛, 等. 应用知识图谱的推荐方法与系统 [J/OL]. 自动化学报: 1-17 [2020-09-28]. <https://doi.org/10.16383/j.aas.c200128>.

[7] 蔡泽胤. 基于平移模型的知识图谱补全算法研究 [D]. 武汉: 华中科技大学, 2019.

[8] 崔员宁, 李 静, 沈 力, 等. Duration-HyTE: 基于持续时间建模的时间感知知识表示学习方法 [J]. 计算机研究与

- 发展, 2020, 57(6): 1239-1251.
- [9] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating Embeddings for Modeling Multi-Relational Data [C] //NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2. Red Hook: Curran Associates Inc, 2013: 2787-2795.
- [10] WANG Z, ZHANG J W, FENG J L, et al. Knowledge Graph Embedding by Translating on Hyperplanes [C] // AAAI'14: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. Québec: AAAI Press, 2014: 1112-1119.
- [11] LIN Y, LIU Z, SUN M, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion [C] // AAAI'15: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas: AAAI Press, 2015: 2181-2187.
- [12] JI G L, HE S Z, XU L H, 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 (Volume 1: Long Papers), Beijing : Association for Computational Linguistics, 2015: 687-696.
- [13] YANG B S, YIH W T, HE X D, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases [EB/OL]. [2014-10-27]. <https://arxiv.org/abs/1412.6575v2>.
- [14] TROUILLON T, WELBL J, RIEDEL S, et al. Complex Embeddings for Simple Link Prediction [C] //ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48. New York: JMLR org, 2016: 2071-2080.
- [15] LIU H X, WU Y X, YANY Y. Analogical Inference for Multi-Relational Embeddings [C] //ICML'17: Proceedings of the 34th International Conference on Machine Learning-Volume 70. Sydney: JMLR Org, 2017: 2168-2178.
- [16] 胡 卓. 面向知识图谱的表示学习算法的研究与应用 [D]. 成都: 电子科技大学, 2020.
- [17] LAO N, COHEN W W. Relational Retrieval Using a Combination of Path-Constrained Random Walks [J]. Machine Learning, 2010, 81(1): 53-67.
- [18] LAO N, MITCHELL T M, COHEN W W, et al. Random Walk Inference and Learning in A Large Scale Knowledge Base [C] //EMNLP'11: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh: Association for Computational Linguistics, 2011: 529-539.
- [19] GALARRAGA A L, TEFLIOUDI C, HOSE K, et al. AMIE: Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases [C] //WWW'13: Proceedings of the 22nd International Conference on World Wide Web. New York: Association for Computing Machinery, 2013: 413-422.
- [20] GALÁRRAGA L, TEFLIOUDI C, HOSE K, et al. Fast Rule Mining in Ontological Knowledge Bases with AMIE+ [J]. The VLDB Journal, 2015, 24(6): 707-730.
- [21] 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.
- [22] ZHANG W, PAUDEL B, WANG L, et al. Iteratively Learning Embeddings and Rules for Knowledge Graph Reasoning [C] //WWW'19: The World Wide Web Conference. New York: Association for Computing Machinery, 2019: 2366-2377.
- [23] 王 红, 李 晗, 李浩飞. 民航突发事件领域本体关系提取方法的研究 [J]. 计算机科学与探索, 2020, 14(2): 285-293.
- [24] 王 红, 李浩飞, 邸 帅. 民航突发事件实体识别方法研究 [J]. 计算机应用与软件, 2020, 37(3): 166-172.
- [25] XIE S, TU Z. Holistically-Nested Edge Detection [C] //2015 IEEE International Conference on Computer Vision(ICCV). Santiago: IEEE, 2015: 1395-1403.
- [26] HAJEK P. Towards Metamathematics of Weak Arithmetics over Fuzzy Logic [J]. Logic Journal of IGPL, 2011, 19(3): 467-475.

[27] KINGMA D, BA J. Adam: A Method for Stochastic Optimization [EB/OL]. [2014-10-24]. <https://arxiv.org/abs/1412.6980v1>.

[28] 冯兴杰, 曾云泽. 基于评分矩阵与评论文本的深度推荐模型 [J]. 计算机学报, 2020, 43(5): 884-900.

The Method of Completing the Knowledge Graph of Aviation Safety Incidents

WANG Hong, LU Lin-yan, WANG Tong

School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China

Abstract: In view of the lack of relationship between sparse entities and low learning efficiency of knowledge graphs, an improved knowledge graph completion method, I_IterE (Improved- Iteratively learning embeddings and rules), is proposed in this paper. First, a feature layer is added between the representation learning layer and the rule learning layer of the IterE model, which uses the add method to enhance the relationship features learned by the representation learning layer and generates new relationship features. Then, in the rule learning layer, the new relational features are combined with the improved scoring mechanism to calculate the score of axioms, which is used to judge the credibility of the axioms. Finally, the high resolution axioms are used for deductive reasoning to obtain new triples related to sparse entities. The public datasets WN18-sparse, WN18RR-spars, FB15k-sparse and FB15k-237-sparse are used for link prediction and rule evaluation. Experimental results show that the method described in this paper has achieved good improvement in MRR, Hits@n and rule evaluation. This method is applied to the dataset of aviation safety incidents. While significantly improving the effect of link prediction, it also provides a method support for completing the knowledge graph of aviation safety incidents.

Key words: knowledge graph; relationship feature; scoring mechanism; link prediction; rule evaluation; aviation safety incidents

责任编辑 潘春燕