

一种基于实体时间敏感度的知识表示方法

田满鑫¹, 寿黎但^{1, 2}, 陈珂^{1, 2}, 江大伟^{1, 2}, 陈刚^{1, 2}

(1.浙江大学计算机科学与技术学院, 浙江 杭州 310027;

2.浙江省大数据智能计算重点实验室, 浙江 杭州 310027)

摘要:在知识图谱(Knowledge Graph)中, 知识表示方法旨在通过一种低维稠密的向量表示方法来高效地挖掘不同实体、关系之间复杂语义关系, 在知识问答、信息检索等领域有着重要意义。然而, 现有的绝大多数的知识表示方法忽略了时间因素, 无法表示应用中随时间变化的动态知识。针对该问题, 本文提出一种基于实体时间敏感度的知识表示方法。该方法将时间信息以不同程度融入不同类型的实体向量表示中, 然后进行实体和关系之间语义挖掘。实验结果表明, 这种基于实体时间敏感度的表示方法能够明显提高知识图谱的时态知识补全和预测任务性能。

关键词: 知识图谱; 表示学习; 时态知识; 复杂关系; 知识补全

中图分类号: TP391.1 **文献标识码:** A

A Knowledge Representation Method Based on Entity Time Sensitivity

TIAN Manxin¹, SHOU Lidan^{1,2}, CHEN Ke^{1,2}, JIANG Dawei^{1,2}, CHEN Gang^{1,2}

(1.College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China;

2.Key Laboratory of Big Data Intelligent Computing of Zhejiang Province, Hangzhou 310027, China)

Abstract:Representation learning in knowledge graph aims to explore the complex semantic relationship between entities and their relations with a low-dimensional, dense vectors representation method, which is of great significance in the fields of knowledge question and answer and information retrieval. However, most of the existing knowledge representation methods ignore the time factor and cannot express the dynamic knowledge of the application over time. For this problem, this paper proposes a knowledge representation method based on entity time sensitivity. This method integrates the time information into different types of the entity vector representation with different degrees, and then performs semantic mining between entities and their relationships. Experimental results show that this entity time sensitivity based representation method can obviously improve the temporal knowledge completion and prediction task performance of the knowledge graph.

Keywords: knowledge graph; representation learning; temporal knowledge; complex relationship; knowledge complement

1 引言(Introduction)

自Google提出知识图谱概念以来, 知识图谱凭借其强大的语义处理能力而得到学术界和工业界广泛关注。它将知识库信息以一种三元组(实体1, 关系, 实体2)的形式进行表达, 构成知识网络。这种高效、简洁的知识组织方法能够清楚地描述实体间的关系、概念和属性, 使得知识的表现形式更接近人类的思维表达, 因而被广泛应用在信息检索、智能问答等方面。

传统的表示方法采用语义符号进行知识表示, 比如

万维网联盟采用资源描述框架(Resolution Description Framework, RDF)来表示知识三元组。这种表示方法虽然简洁直白, 但由于表示稀疏、计算效率低下等问题, 无法进行大规模的知识查询和推理。近年来, 随着深度学习在人工智能领域的不断发展, 分布式的知识表示方法成为学术研究的热点。这种分布式知识表示方法将稀疏的符号知识转换成稠密的向量表达形式, 能够存储更丰富的语义信息和进行高效的计算, 具有重要的学术价值和应用意义。鉴于这些优点, 研究者们提出多种知识表示模型, 如神经网络张量模

型、线性模型和翻译模型等。这些方法考虑知识表示在向量空间上的语义计算,通过对知识图谱实体节点、关系边的学习,使得语义相似的实体之间距离相近。

然而,上述表示方法探究的是知识库的静态知识表示。这种知识表示局限于将实体之间的关系、属性看成静态联系,忽略了时间对知识演变的影响。比如,在新型供电列车储能系统中,(输电母线、引发…故障、车厢子线)这一知识的有效时间为2018/12/20 00:00:00,而其他时间下发生的车厢供电子线故障并非由输电母线引起。如果采用传统的静态知识表示方法进行时态性知识查询或推理,那么相关结果准确度、可信度偏低,无法满足时态知识的表示需求。

针对以上问题,本文提出一种基于实体时间敏感度的知识表示方法ETA-TransE(Entity-Time-Aware TransE)。这种方法不仅考虑了知识图谱的时态信息和时间对知识表示的影响,还考虑到不同类型实体对时间敏感程度的一致性,联合实体类型和时间信息进行时态领域的知识表示和推理学习。具体地,首先对知识库中带绝对时间信息的知识三元组(头实体,关系,尾实体,绝对时间)相对于实体的起始时间进行相对时间计算,使得处在不同时间维度的知识三元组能在相同的时间平面聚焦;然后,针对不同类型实体构建不同的时态转移矩阵,并通过实体与时态矩阵相乘得到当前时间状态下的实体表示,使得不同类型实体受同一时间影响的程度不同;最后,在融合时间信息情况下进行知识表示学习,得到实体、关系的语义向量。

本文的主要贡献如下:

(1)提出一种基于实体时间敏感度的知识表示模型。该模型联合了实体类型和时间信息,考虑实体在不同时间下表示,解决实体之间的复杂关系。

(2)提出一种基于相对时间粒度的知识处理方法。该方法主要考虑在历史、人物事件等应用场合下,相对时间粒度能够更好地挖掘不同知识之间的内在联系,有效缓解绝对时间所带来的特征稀疏性问题。

(3)在YAGO和Wikidata带时间信息的数据集上,与已有的知识表示方法进行了对比工作,验证了本文所提出的知识表示模型有效性。

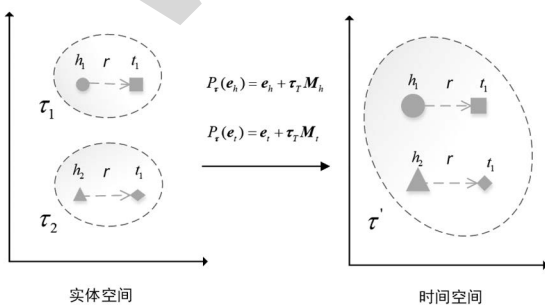


图1 基于实体时间敏感度的知识表示方法

Fig.1 Knowledge representation method based on entity time sensitivity

2 相关工作(Related work)

2.1 知识表示学习

近年来,学术研究者针对知识表示学习提出了许多方法和模型,不断提升知识表示的语义准确度和增强知识补全能力。

Jason等人最早采用使用分布式向量进行知识表示^[1],通过头实体、尾实体在关系上的投影矩阵来计算不同实体的语义相似度。随后,Rodolphe^[2]提出隐变量模型(latent factor model),基于关系的双线性变换来探究实体和关系的二阶联系。这些线性模型简单,但由于矩阵投影方法的协同性较差,无法准确刻画不同实体、关系之间的语义联系。

针对线性模型的缺点,Bordes等人提出TransE^[3]模型,将关系 r 看作为头实体 h 和尾实体 t 之间的翻译过程。这种翻译模型简单高效,得到广泛地应用。TransH^[4]模型引入关系超平面进行知识表示,使同一实体能够在不同关系中扮演不同的角色。TransD^[5]、TransR^[6]、TransM^[7]等采用不同维度的语义空间来表示实体和关系,使得实体和关系支持多元语义关系表达。此外,还有一些方法通过加入外部信息^[8-10],增强知识表示的迁移能力。

然而,以上方法都忽略了知识库的时间因素,没有考虑知识的时效性问题。当采用传统的知识表示方法进行带时间戳的知识图谱补全、知识查询或推理时,由于这些模型对时间不敏感,无法有效处理时间信息,使得预测出来的结果严重偏离实际情况。

2.2 引入时间信息的知识表示学习

时间作为一种信息因子,是构建包含事件等动态知识的信息图谱必要组成内容。目前,已经有相关研究工作针对带时间信息的信息图谱进行了时态性知识表示学习探究。

Jiang^[11]等人在16年首次提出一种基于时态信息的信息图谱表示方法,采用关系的时间相对性(如出生—死亡)对实体表示进行约束和学习。随后,Trivedi^[12]等人采用循环神经网络对蕴含时态的知识进行知识表示。虽然这些方法考虑了时间对知识表示的影响,但是没有直接利用时间信息,只能在模糊的时间范围内进行知识表示和推理。针对以上缺陷,Shib^[13]提出了HyTE模型,该模型是一种基于时间维度的知识表示方法。它构建了不同的时间超平面,然后将三元组的知识信息投影到时间平面上,进行时间维度的知识表示与推理。

虽然HyTE模型直接使用了时间信息,对同一时间范围内的所有实体、关系进行了时间投影和表示学习,但是没有考虑到不同类型的实体受时间的影响不同。因此,这种知识表示方法仍然不能准确地表达时态语义信息。针对该问题,本文提出的模型算法能够较好地缓解该问题,有效地提高时态知识表示的性能。

3 问题定义(Problem definition)

给定静态知识库,我们将其表示为:

$$G = (E, R, S, T) \quad (1)$$

其中, $E=\{e_1, e_2, \dots, e_{|E|}\}$ 代表知识库的实体集合; $R=\{r_1, r_2, \dots, r_{|R|}\}$ 代表知识库的关系集合; $T=\{t_1, t_2, \dots, t_{|T|}\}$ 代表知识库的时间集合; $S \subseteq E \times R \times E \times T$ 代表知识库中的带时间信息的三元组集合, 一般表示为 (h, r, t, τ) , 其中, h 和 t 代表头实体和尾实体, r 代表实体之间的关系, t 代表该事实发生的时间。

不同的头实体和尾实体类型由关系类型确定。知识库中的关系带有方向性, 一般不可逆, 而知识三元组的头实体和尾实体为关系的主体和客体。由于不同的类型实体作用的关系范围不一致, 因此可以对关系类型进行聚类, 划分对应的头实体和尾实体类型。对于关系类型, 我们给出如下定义:

$$C = \{c_1, c_2, \dots, c_k\} \quad (2)$$

其中, $c_k = \{r_1, r_2, \dots, r_n\}$ 代表不同的关系类型, 而 k 代表知识库的关系类型总数。另外, 我们将不同关系类型对应的头实体、尾实体的实体类型表示为 h_c 、 s_c 。

实体时间敏感度是指在相同时间下, 不同类型的实体所发生的信息变化程度不相同。具体为, 实体被向量唯一确定后, 在不同时间下得到的不同的表达, 然后依据如下计算公式得到信息变化量:

$$d = \|e_{s1} - e_{s2}\|_{L_1/L_2} \quad (3)$$

其中, e_{s1} 和 e_{s2} 代表同一实体在不同时间下的表示, L_1 、 L_2 代表不同范数。

该信息量揭示不同类型实体的变化程度, 能解决实体之间复杂的关系, 帮助模型更准确地捕捉实体与关系之间的语义联系。

考虑知识库的两个时态知识(奥巴马, 出生于, 美国, 1961)和(奥巴马, 担任...总统, 美国, 2008), 由于“奥巴马”是人物类型实体, 而“美国”是地理类型实体, 因此, 它们在这两个时间状态的表示差异也不相同。采用 (h, t) 和 (h', t') 分别表示不同时间下的“奥巴马”和“美国”的实体表示, r_i 代表不同关系, θ_i 代表向量变化程度。该实体向量受时间影响的变化以及和关系的作用如图2表示。

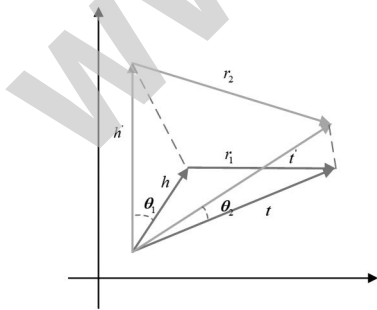


图2 不同实体向量受时间变化的影响

Fig.2 The variation trends of different vectors towards time

4 方法介绍(Method introduce)

4.1 时态知识图谱

传统的知识图谱没有涉及时间信息, 一般对静态的知识

采用 (h, r, t) 三元组方式进行表示。而时态性知识图谱作为一种动态的知识图谱, 会将时间作为一种独立的维度信息加入知识三元组中, 组成 (h, r, t, τ) 四元组形式来表示知识库中的知识, 其中 τ 代表四元组的发生时间。这种表示方法能够较准确地描述事件等知识元素, 完整地记录图谱中人物履历或历史事件发生的过程, 体现知识的时态性。

HyTE模型最早将时间信息作为单独的一个维度融合到知识表示方法中, 在绝对时间平面中对各知识进行向量投影, 建立知识联系。考虑到在绝对时间平面中, 可能出现知识三元组较少的情况, 导致模型难以挖掘不同空间内的知识共性和语义联系, 本文提出一种针对绝对时间信息稀疏问题的处理方案。

给定知识图谱中的时态知识四元组(头实体, 关系, 尾实体, 发生时间), 我们将它表示为 (h, r, t, τ) , 其中 $h, t \in E, r \in R, \tau \in T$, E 、 R 和 T 分别表示实体集合、关系集合以及时间集合。对于每个实体, 统计他们在知识库中出现的起始时间, 记为 (e, τ_s) ; 然后针对每条时态知识 (h, r, t, τ) , 按照如公式(4):

$$\tau' = \tau - \tau_s \quad (4)$$

得到相对时间 τ' , 并记录到时态知识表示中, 最终得到 (h, r, t, τ, τ') 表达形式。这样的处理方法, 允许我们在相对时间平面聚集较多的知识四元组, 容易建立不同时间平面的语义联系。

4.2 时态知识表示

时态知识表示方法基于TransE模型, 本部分先介绍TransE模型相关原理。

TransE模型将知识库中的实体和关系映射到同一个低维向量平面, 然后将关系当作实体和实体之间的向量平移, 以此满足实体和关系之间的语义计算。给定一个知识三元组(头实体, 关系, 尾实体), TransE模型将它们转换成向量的形式 (e_h, r, e_t) , 并使合法三元组尽量满足如下计算公式: $e_h + r \approx e_t$, 而非法三元组则无法满足。因此, TransE模型设计了如下得分函数来衡量不同的三元组质量高低:

$$f(h, r, t) = \|e_h + r - e_t\|_{L_1/L_2} \quad (5)$$

式(5)表示向量 $e_h + r$ 和 e_t 之间的 L_1 或 L_2 距离。该函数通过实体间的距离大小来衡量不同实体、关系之间的语义相似度, 使得正确三元组和错误三元组能较好地被区分出来。在实际运用时, TransE采用如下最大间隔损失函数来优化模型的评分过程:

$$L = \sum_{(h, r, t) \in D^+} \sum_{(h', r, t') \in D^-} \max(0, f(h, r, t) - f(h', r, t') + \gamma) \quad (6)$$

其中, D^+ 和 D^- 分别代表正确三元组集合和错误三元组集合, γ 是用于区分正确三元组和错误三元组的间隔距离。 (h', r, t') 代表错误三元组, 是由正确三元组随机替换头实体或尾实体, 构成不符合实际的三元组而产生。该损失函数能在最小化正确三元组得分的同时最大化了错误三元组的得分, 使正确三元组之间语义联系更紧密, 而尽可能让错误三元组中

的实体、关系存在较大的语义差异。

TransE模型虽然简单,涉及的参数较少,计算复杂度低,但能很好地构建实体和关系之间的语义联系,非常适合用于大规模知识图谱的知识表示。本文沿用TransE模型的思想,将关系关系看作受时间影响的实体之间的平移向量,对加入时间维度信息的知识图谱进行翻译模型的表示学习。

4.3 融合实体类型的时态知识表示

TransE模型将实体和关系投影到相同维度的低维向量空间,无法解决知识图谱中实体之间的一对多、多对一和多对多等复杂关系问题。比如知识库存在(奥巴马,是…总统,美国)和(特朗普,是…总统,美国)两个知识。如果仅在总统这种关系下,奥巴马和特朗普的知识表示结果非常相近,但是在其他属性领域,该两者具有较大的差异性,TransE模型无法解决该类问题。在时态性知识图谱中,我们发现对时间信息特征的利用能够解决实体之间的复杂关系。作用于同一实体的不同关系受到同一时间约束不一致,而存在相同关系的不同实体对也存在时间差异,比如“出生”发生的时间要早于“死亡”和奥巴马比特朗普更早就职美国总统。因此,本文所提出的时态知识表示模型能够较好地地区分不同类型关系对不同实体影响。

由于不同类型实体对时间的敏感程度不一致,使得同一时间对不同类型实体的信息变化影响不相同,如人物类型实体存在年轻、年老等状态,不同状态下实体的活动范围和行动能力等不尽相同,而地理类型实体随着时间演变而几乎不发生变化。这种实体类型对时间的敏感性特点,能够影响不同实体的关系作用域,以及不同关系的作用范围。如果将实体类型考虑到时态知识表示学习中,那么模型能提取更准确的语义特征,以增强时态性知识补全和推理能力。

给定知识图谱的时态知识四元组(头实体,关系,尾实体,发生时间),得到相对时间的知识表达(h, r, t, τ, τ')后,我们用向量 $\tau_1, \tau_2, \dots, \tau_r$ 来代表不同的相对时间,使得相对时间能够被唯一表示。为了使不同类型的实体在相同时间下的能量变化幅度不相同,本文定义一种时态转移矩阵 $M \in \mathbb{R}^{m \times n}$ 。对于头实体、尾实体有如下计算公式:

$$P_{\tau}(e_h) = e_h + \tau_r M_h \quad (7)$$

$$P_{\tau}(e_t) = e_t + \tau_r M_t \quad (8)$$

其中, M_h 代表头实体的时态转移矩阵, M_t 代表尾实体的时态转移矩阵。

对于正确三元组以及对应的发生时间,我们希望实体和关系向量之间满足 $P_{\tau}(e_h) + r = P_{\tau}(e_t)$ 。因此,本文为ETA-TransE模型定义如下得分函数:

$$f(h, r, t, \tau) = \|P_{\tau}(e_h) + r - P_{\tau}(e_t)\|_{L_1/L_2} \quad (9)$$

通过以上公式,不同的时态转移矩阵能够被表示出来。基于这种时态转移矩阵,在同一时间下,对于不同类型的实体,通过不同的时态转移矩阵能够产生不同程度的状态变化。以及在不同含义下,对于同一实体,在同一时间下的表

示也不相同。

在实际过程,我们借鉴TransE模型的思想,采用最大间隔法,定义如下损失函数来对模型进行优化:

$$L = \sum_{(h,r,t,\tau) \in D^+} \sum_{(h',r',t',\tau') \in D^-} \max(0, f(h, r, t, \tau) + r - f(h', r', t', \tau')) \quad (11)$$

其中, D^+ 和 D^- 分别代表带时间信息的正确三元组集合和错误三元组集合, γ 是正确三元组和错误三元组之间的间距。在实验过程中,我们对上述公式中的向量等元素做了如下限制:

$$\|e_i\|_2 \leq 1, \|r_i\|_2 \leq 1, \|\tau_i M\|_2 \leq 1$$

5 实验结果及分析(Experimental results and analysis)

实验包含了三个任务,分别为实体链接预测、关系链接预测和三元组分类预测任务。与以往知识表示任务评测对象不同,本文的评测对象为带有时间信息的数据集。

5.1 实验数据

在开源的知识库中,Wikidata和YAGO数据集都包含涉及时间的知识数据。本文针对这两个数据集提取时间信息丰富的数据,得到两个子集Wikidata12K和YAGO11K,作为我们的评测数据,并划分成训练集、验证集和测试集三部分。其中训练集用于模型训练,验证集用于超参数调优,测试集用于评价不同模型的性能。表1展示了数据集的相关属性:

表1 数据集统计结果

Tab.1 Statistics of data sets

数据集	关系	实体	训练集/验证集/测试集
Wikidata12K	24	12,554	32.5k/4k/4k
YAGO11K	10	10,623	16.4k/2k/2k

对于YAGO数据,为了防止出现数据分布稀疏问题,我们对带时间信息的数据集进行了数据筛选,按照出现频率挑选了排名前10的关系数据以及去除只出现一次的实体数据,得到YAGO11K。

针对Wikidata数据,我们做了类似同样的操作,按照出现频率挑选了排名前24名的关系数据,以及相关实体数据,组成Wikidata12K数据集,使得数据规模为YAGO11K的两倍。

5.2 实验对比方法

为了评估本文所提出来的时态知识表示模型,我们对比较了传统领域的知识表示方法TransE、TransH和TransR,以及融合时间信息的知识表示方法HyTE。

5.3 参数设置

模型包含学习率 λ 、向量维度 d 和最大间隔 γ 。其中 $\lambda \in \{0.1, 0.01, 0.05\}$, $d \in \{50, 80, 100\}$, $\gamma \in \{1, 2, 4\}$ 。验证集达到最优效果对应的参数分别为0.01、100、4。训练迭代的次数为1000。

5.4 实验结果

5.4.1 实体链接预测

实体链接预测任务旨在预测知识三元组缺失的头实体

或尾实体。给定正确三元组 (h, r, t, τ) ，我们依次去除其中的头实体和尾实体，即得到 $(?, r, t, \tau)$ 和 $(h, r, ?, \tau)$ 两种形式；然后依次将知识库中的其他实体填充到对应的缺失位置，计算新构成的三元组得分；最后对该得分进行排序，得到不同三元组的排名。

本文沿用其他知识表示论文的评价指标：(1)Mean Rank：正确三元组的实体得分排名平均值，该指标值越小代表模型性能越好；(2)Hits@10：正确三元组实体排名小于10的占比，该指标值越大则模型表示越准确。除此之外，由于三元组替换头实体或尾实体后，所构成的新三元组可能已经存在于原知识库中，这部分三元组会干扰当前正确三元组的实体排名，因此采用了“原始”(raw)和“过滤”(Filter)两种评估方式。其中，“原始”保留所有三元组再进行排名，“过滤”代表排除已经存在的三元组再进行排名。各模型在数据集Wikidata12K和YAGO11K的测试实验结果如表2和表3所示。

表2 YAGO11K的实体链路预测结果

Tab.2 Evaluation results on entity-linked prediction from YAGO11K

数据集	YAGO11K							
	Mean Rank(head)		Hits@10(head,%)		Mean Rank(tail)		Hits@10(tail,%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
TransE	1182	1160	13.5	20.9	141	138	32.0	36.4
TransH	990	968	13.7	21.7	105	102	35.2	40.1
TransR	1290	1268	11.8	15.5	211	208	28.5	31.9
HyTE	1060	1040	13.0	19.3	107	105	36.5	39.2
ETA-TransE	910	890	17.9	25.7	132	129	36.9	41.1

表3 Wikidata12K的实体链路预测结果

Tab.3 Evaluation results on entity-linked prediction from Wikidata12K

数据集	Wikidata11K							
	Mean Rank(head)		Hits@10(head,%)		Mean Rank(tail)		Hits@10(tail,%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
TransE	349	336	30.8	48.4	255	251	46.0	54.3
TransH	284	271	32.5	49.9	215	210	47.3	56.2
TransR	444	431	27.4	41.3	376	371	40.1	47.0
HyTE	237	225	25.0	40.6	179	174	41.6	49.4
ETA-TransE	227	214	34.3	50.7	198	193	49.2	58.2

从以上结果可以看到，本文所提出的ETA-TransE模型虽然在尾实体的MeanRank指标不是最优，但是在其他指标上均有较大程度的提高。在数据集中，相对头实体，尾实体的实体类型较为丰富，比如头实体类型以“人物”类型为

主，而尾实体除了“人物”类型实体、还包含“地理”“国家”等类型，并且部分实体的类型比较少，尾实体的时态转移矩阵较难准确地捕捉这类实体语义信息随时间变化的变化程度，因此Mean Rank指标没能达到最优。在Hits@10指标上，不管是头实体链接预测还是尾实体链接预测，都达到目前最优的结果，说明在时态知识表示方面，ETA-TransE模型能够较好地捕捉不同时间下实体与关系的语义信息。因此，在带时间信息的实体补全或推理任务中本文所提的模型具有一定的优势。

5.4.2 关系链接预测

关系链接预测是用于评估模型的关系补全能力。给定正确三元组，将关系依次替换成知识库中的其他关系，得到相应的知识三元组得分；然后对得分排序，得到关系排名。类似实体链接任务，该实验采用的指标为Mean Rank和Hits@1，其中Hits@1代表正确三元组关系排名第一的占比。评估方式也采用“Raw”和“Filter”两种形式。

表4 关系链接预测结果

Tab.4 Evaluation results on relation-linked prediction

数据集	YAGO11K				Wikidata11K			
	Mean Rank		Hits@1(%)		Mean Rank		Hits@1(%)	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
TransE	1.78	1.76	69.9	70.6	1.22	1.17	87.3	91.5
TransH	1.50	1.49	76.8	78.0	1.14	1.09	91.1	95.5
TransR	1.29	1.28	81.7	83.1	3.07	3.03	53.3	55.9
HyTE	1.28	1.26	82.3	83.7	1.13	1.09	92.6	96.4
ETA-TransE	1.23	1.21	84.7	86.0	1.10	1.05	92.7	97.3

从上述结果可以看到，我们提出的模型在两份数据集上的各项指标均取得最佳效果。在关系链路预测方面，TransH比TransE预测效果更好，表明实体之间复杂的关系需要复杂的模型进行不同关系下的实体表示学习；而融入时间信息的HyTE模型能够较好地预测实体之间的关系，说明时间对关系的影响比较大；在保持前者的优点情况下，本文提出的ETA-TransE模型在时态信息处理方面，拥有更好的关系预测能力。

5.4.3 三元组分类

三元组分类旨在判断给定的知识三元组是否符合客观事实，即是否真实存在。由于YAGO11K和Wikidata12K数据集只包含正样本，本文依照Socher等人提出的方法来构造负样本。对于验证集和测试集的正确三元组，我们随机替换其头实体或尾实体。而替换的实体必须从当前三元组关系对应的头实体集合或尾实体集合中选出，使得产生负样本集合不存在明显不合理的三元组。为了平衡数据，产生的负样本数量与正样本数量相同，即每一条正样本对应唯一的负样本。

实验采用准确率、精确率、召回率和 F_1 值作为带时间信息三元组分类的评估指标。对于每种关系，我们设置不同的

阈值 δ_r , 使得在验证集上达到最佳的准确率。然后针对测试集中带时间的三元组 (h, r, t, τ) 计算距离得分, 如果该分数小于 δ_r , 则表示为正确三元组, 否则为错误三元组。实验最终的评测标准为所有关系评估指标的平均值。各模型的三元组分类结果如表5所示。

表5 三元组分类的准确率(%)

Tab.5 The accuracy of triple classification(%)

数据集	YAGO11K	Wikidata12K
TransE	68.80	80.86
TransH	70.43	83.04
TransR	69.19	81.30
HyTE	—	—
ETA-TransE	74.23	86.47

从上述结果可以看到, 本文的ETA-TransE在带时间信息的三元组分类任务上达到最佳的分类效果。这说明, 融合时间信息的知识表示能够较好地辨别推理得到的三元组是否符合客观事实, 使正确三元组之间的联系更紧密以及错误三元组中实体和关系的存在较大的语义差异。

6 结论(Conclusion)

传统的知识表示方法由于没有考虑知识图谱中的时间信息, 在带有时间信息的知识补全和推理任务中存在较大的性能缺陷。本文针对该缺点提出一种基于实体时间敏感度的时态知识表示方法。该方法在利用时间特征进行知识表示的同时, 还考虑到不同实体类型对时间敏感程度不同的特点, 采用时态转移矩阵更准确地捕捉实体、关系与时间的语义联系。在含有时间信息的YAGO11K和Wikidata12K数据集上的实验结果表明, 本文提出的ETA-TransE方法能够显著提升带时间信息的实体链接、关系链接和三元组分类任务, 能较好地支持时态性的知识补全任务和推理等任务。

参考文献(References)

- [1] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases[C]. Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.
- [2] Jenatton R, Roux N L, Bordes A, et al. A latent factor model for highly multi-relational data[C]. Advances in Neural Information Processing Systems, 2012: 3167–3175.
- [3] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]. Advances in neural information processing systems, 2013: 2787–2795.
- [4] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]. Twenty-Eighth AAAI conference on artificial intelligence, 2014.
- [5] 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.
- [6] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]. Twenty-ninth AAAI conference on artificial intelligence, 2015.
- [7] 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 Computing, 2014.
- [8] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions[C]. Thirtieth AAAI Conference on Artificial Intelligence, 2016.
- [9] Xie R, Liu Z, Luan H, et al. Image-embodied knowledge representation learning[J]. arXiv preprint arXiv:1609.07028, 2016.
- [10] Xie R, Liu Z, Sun M. Representation Learning of Knowledge Graphs with Hierarchical Types[C]. IJCAI, 2016: 2965–2971.
- [11] Jiang T, Liu T, Ge T, et al. Encoding temporal information for time-aware link prediction[C]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 2350–2354.
- [12] Trivedi R, Dai H, Wang Y, et al. Know-evolve: deep temporal reasoning for dynamic knowledge graphs[C]. Proceedings of the 34th International Conference on Machine Learning, 2017.
- [13] Dasgupta S S, Ray S N, Talukdar P. HyTE: Hyperplane-based Temporally aware Knowledge Graph Embedding[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 2001–2011.

作者简介:

田满鑫(1994–), 男, 硕士生. 研究领域: 数据挖掘, 知识图谱.

寿黎但(1974–), 男, 博士, 教授. 研究领域: 空间数据库, 数据挖掘, 数据可视化.

陈珂(1977–), 女, 博士, 副教授. 研究领域: 时空数据库, 数据挖掘, 数据隐私保护.

江大伟(1982–), 男, 博士, 研究员. 研究领域: 数据库技术, 大数据管理技术, 区块链技术.

陈刚(1973–), 男, 博士, 教授. 研究领域: 大数据管理.