软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(10):4533–4547 [doi: 10.13328/j.cnki.jos.006888]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向时间感知的知识超图链接预测*

陈子睿[1,2], 王 鑫[1,2], 王晨旭[1,2], 张少伟[1,2], 闫浩宇[1,2]

[1](天津大学 智能与计算学部, 天津 300350)

[2](天津市认知计算与应用重点实验室, 天津 300350)

通信作者: 王鑫, E-mail: wangx@tju.edu.cn

摘 要: 知识超图是一种使用多元关系表示现实世界的异构图, 但无论在通用领域还是垂直领域, 现有的知识超图普遍存在不完整的情况. 因此, 如何通过知识超图中已有的链接推理缺失的链接, 是一个具有挑战性的问题. 目前, 大多数研究使用基于多元关系的知识表示学习方法完成知识超图的链接预测任务, 但这些方法仅从时间未知的超边中学习实体与关系的嵌入向量, 没有考虑时间因素对事实动态演变的影响, 导致在动态环境中的预测性能较差. 首先, 根据首次所提出的时序知识超图定义, 提出时序知识超图链接预测模型, 同时从实体角色、位置和时序超边的时间戳中学习实体的静态表征和动态表征, 以一定比例融合后作为实体嵌入向量用于链接预测任务, 实现对超边时序信息的充分利用. 同时, 从理论上证明模型具有完全表达性和线性空间复杂度. 此外, 通过上市公司的公开经营数据构建时序知识超图数据集 CB67, 并在该数据集上进行了大量实验评估. 实验结果表明, 模型能够在时序知识超图数据集上有效地执行链接预测任务.

关键词: 时序知识超图; 链接预测; 知识表示; 嵌入学习; 时序信息

中图法分类号: TP18

中文引用格式: 陈子睿, 王鑫, 王晨旭, 张少伟, 闫浩宇. 面向时间感知的知识超图链接预测. 软件学报, 2023, 34(10): 4533–4547. http://www.jos.org.cn/1000-9825/6888.htm

英文引用格式: Chen ZR, Wang X, Wang CX, Zhang SW, Yan HY. Towards Time-aware Knowledge Hypergraph Link Prediction. Ruan Jian Xue Bao/Journal of Software, 2023, 34(10): 4533–4547 (in Chinese). http://www.jos.org.cn/1000-9825/6888.htm

Towards Time-aware Knowledge Hypergraph Link Prediction

CHEN Zi-Rui[1,2], WANG Xin[1,2], WANG Chen-Xu[1,2], ZHANG Shao-Wei[1,2], YAN Hao-Yu[1,2]

[1](College of Intelligence and Computing, Tianjin University, Tianjin 300350, China)

[2](Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300350, China)

Abstract: A knowledge hypergraph is a form of a heterogeneous graph that represents the real world through *n*-ary relations. However, both in general and specific domains, existing knowledge hypergraphs often suffer from incompleteness. Therefore, it is a challenging task to reason the missing links through the existing links in the knowledge hypergraph. Currently, most research employs knowledge representation learning methods based on *n*-ary relations to carry out link prediction tasks in knowledge hypergraphs. However, these methods only learn embedding vectors of entities and relations from hyperedges with unknown temporal information, neglecting the impact of temporal factors on the dynamic evolution of facts, resulting in poor predictive performance in dynamic environments. Firstly, based on the definition of temporal knowledge hypergraph that proposed by this study for the first time, a link prediction model is proposed for temporal knowledge hypergraphs. Simultaneously, static and dynamic representations of entities are learnt from their roles, positions, and timestamps of temporal hyperedges, which are merged in a certain proportion and utilized as final entity embedding vectors for link prediction tasks to realize the full exploitation of hyperedge temporal information. At the same time, it is theoretically proved that the proposed model is fully expressive and has linear space complexity. In addition, a temporal knowledge hypergraph dataset CB67 is

* 基金项目: 科技创新 2030“新一代人工智能”重大专项(2020AAA0108504); 国家自然科学基金(61972275)
本文由“知识赋能的信息系统”专题特约编辑高宏教授、陈华钧教授、赵翔教授、李瑞轩教授推荐.
收稿时间: 2022-07-05; 修改时间: 2022-08-18, 2022-12-14; 采用时间: 2022-12-28; jos 在线出版时间: 2023-01-13

constructed from the public business data of listed companies, and a large number of experimental evaluations are conducted on this dataset. The experimental results show that the proposed model can effectively perform the link prediction task on the temporal knowledge hypergraph dataset.

Key words: temporal knowledge hypergraph; link prediction; knowledge representation; embedding learning; temporal information

知识超图使用多元关系表示现实世界中的事实, 每个超边由一个 n 元关系(n≥2)及相应有序的 n 个实体组成; 相比之下, 知识图谱[1]是知识超图的一种特殊情况, 即每个关系元数皆为 2 的知识超图. 本质上, 知识超图是一种比知识图谱更具表现力的知识表示形式, 是有效组织人类知识的一种重要方式. 相关工作显示: 传统的基于二元关系的知识图谱对现实世界的高阶语义信息表达存在缺失, Freebase[2]中有超过 1/3 的实体参与至多元关系的表示[3], 有超过 61%的关系以多元关系形式展现[4]. 这证明知识超图更加贴近现实世界的事实表现形式, 其促进了一系列以知识为基础的下游应用发展[5], 例如推荐系统[6]和问答系统[7]等. 尽管现有的知识超图具有较大的规模, 但它们还远远不够完整, 因此, 知识超图链接预测, 即自动推断知识超图中实体间缺失的事实, 已逐渐成为一项重要的工作.

目前, 知识超图链接预测[8,9]的研究工作主要基于时间未知的超边完成实体关系嵌入学习, 并根据学习方法的不同分为 4 类.

- 基于软规则的方法具备可解释性, 可使用规则解释推理结果的来由, 例如较为流行的马尔可夫逻辑网络(Markov logic network, MLN)[10]模型;
- 基于平移的方法主要来自于知识图谱嵌入方法的多元泛化, 通过将实体和关系嵌入到同一向量空间, 学习实体和关系之间的联系. 该类方法的首个模型是基于 TransH[11]泛化而来的 m-TransH[3]模型, 但该模型不具备完全表达性, 即: 对于任何给定不相交的真假超边集合, 模型无法为每个超边都提供一种参数表示以准确表达该超边的所属类型(真或假);
- 基于张量分解的方法通过将高阶张量分解为数个低阶张量, 实现对嵌入向量的学习. 该类方法目前是知识超图链接预测中性能较高的一种, 最新的 SOTA 模型 RAM[12]通过挖掘实体角色的关联性获得最优的性能指标;
- 基于神经网络的方法目前可应用至知识超图结构的模型较少, 目前最新的工作是 G-MPNN[13]模型, 通过将信息传递神经网络(message passing neural network, MPNN)扩展应用至知识超图结构, 可以解决多元关系的链接预测问题.

尽管这些现有的嵌入技术取得了一定成功, 包括超图[14]和知识超图两个领域, 但都是基于超边不含时间属性的假设, 即学习到的实体嵌入向量是一个静态表征, 或是一个时间段内实体特征的聚合. 然而如图 1 所示: 在以小米集团为例的时序知识超图中, 三元关系“分支机构”表示位置 1–位置 3 的实体角色分别为子公司、核心公司和子公司的时序超边, 以(分支机构,金星创业,小米,乐渊网络,2013-12-26)为例, 实体 1 金星创业和实体 3 乐渊网络的角色皆为子公司, 实体 2 小米的角色为核心公司, 该关系发生在 2013 年 12 月 26 日. 可知在现实世界中, 许多事实不是静态的, 而是高度短暂的, 例如(分支机构,金星创业,小米,英鹏互娱)和(股东,小米,捷付睿通,小米信用)都发生在 2013-12-26; (分支机构,金星创业,小米,捷付睿通)应只可能在 2018-07-10 后才是真实的. 直观地说, 当执行知识超图链接预测时, 事实的时间属性同样应该发挥重要作用, 仅为每个实体学习一个静态表征的模式可能是次优的, 这要求对包含时间信息的知识超图嵌入技术进行深入研究, 从而能够为实体提供其在任何特定时间下的特征向量.

首先明确由本文首次提出的时序知识超图概念及其链接预测任务的定义, 然后基于知识图谱嵌入模型 DistMult[15]扩展至多元关系, 从静态嵌入与动态嵌入两个角度提取实体信息并作融合, 提出用于时序知识超图链接预测的嵌入模型 THM (temporal knowledge hypergraph model). 该模型先后以获取时序知识超图的静态结构信息和动态时序信息为导向, 首先利用实体在时序超边中的角色和位置差异获取实体与关系的静态嵌入向量, 然后利用实体所在时序超边的时序信息差异获取实体在特定时间戳下的动态嵌入向量, 最后以一定比例融合两类嵌入向量, 用于下游的链接预测任务. 通过理论证明可知, THM 是一个具有完全表达性及线性空

间复杂度的模型. 在真实时序知识超图数据集 CB67 上与知识超图嵌入模型对比, 验证了 THM 的有效性、合理性和鲁棒性. 据我们所知, 这是第一个具有完全表达性的时序知识超图链接预测模型.

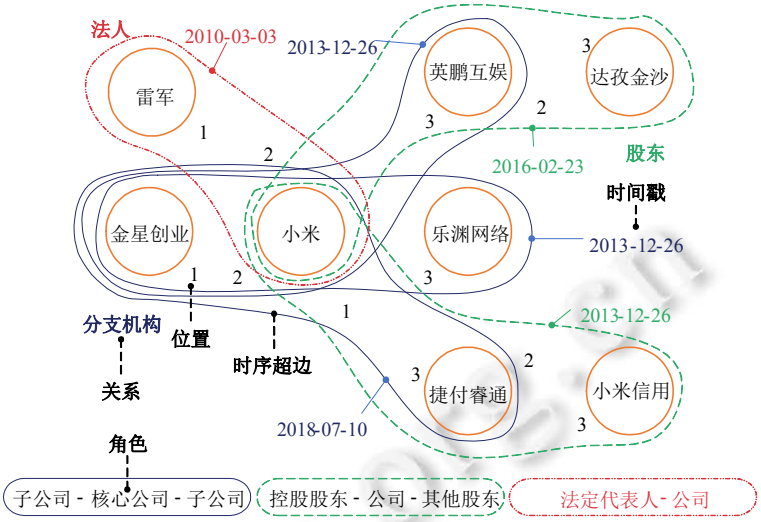


图 1 时序知识超图示例

本文的主要贡献如下:

(1) 基于本文首次提出的时序知识超图定义, 提出用于时序知识超图链接预测的嵌入模型 THM. 该模型同时从实体角色及位置等信息中学习静态嵌入向量, 从时序超边时间戳中学习动态嵌入向量, 将两种嵌入向量以一定比例融合后作为最终实体表征, 实现对超边时序信息的充分利用, 提升链接预测的性能表现;

(2) 从理论上证明 THM 模型具有完全表达性及线性空间复杂度, 能够为任意一个时序知识超图数据集提供一种嵌入表示使模型精确地划分事实与非事实, 在处理大规模时序知识超图数据的情况下具备可扩展性;

(3) 通过上市公司的公开经营数据构建了首个时序知识超图数据集 CB67, 该数据集包含最高元数为 7 元的关系且超边具有时间戳属性, 可作为时序知识超图链接预测任务的基准数据集. 本文模型在 CB67 上进行了大量实验, 通过与知识超图嵌入模型的性能比较, 验证了该模型的有效性、合理性和鲁棒性.

本文第 1 节介绍相关工作. 第 2 节介绍时序知识超图和时序知识超图链接预测的预备知识. 第 3 节提出时序知识超图链接预测模型, 分别从结构静态嵌入、时序动态嵌入和模型训练这 3 个方面进行表述. 第 4 节通过理论证明模型具有完全表达性及线性空间复杂度. 第 5 节对在真实数据集 CB67 上进行实验加以介绍. 第 6 节总结全文.

1 相关工作

近年来, 学术界与工业界逐渐意识到图数据应用的重要性, 链接预测已成为该方向的研究热点之一. 目前, 学术界尚无基于时序知识超图结构的链接预测研究工作. 本节将对目前与时序知识超图链接预测相关的研究工作进行介绍, 主要分为时序知识图谱和知识超图两类.

1.1 时序知识图谱链接预测

目前已有多项工作提出了解决时序知识图谱链接预测的方法, 这些方法通常使用时间嵌入来编码实体和关系随时间的演变. Know-Evolve[16]通过建模时间戳的演变过程对实体表征进行建模, 以表示时间属性对实

体的影响. TA-TransE[17]采用嵌入时间信息的建模方式, 将文本形式的时间通过循环神经网络(RNN)嵌入到关系表征中, 利用 TransE[18]的评分函数进行实体预测. Temporal TransE[19]在同一向量空间中表示实体和关系的嵌入向量, 并使用与 TransE 相似的评分函数进行链接预测. ChronoR[20]同样通过转换将实体和关系投射到其他空间, 根据时间对空间进行旋转, 采用一个新式评分函数进行预测. 然而, 这些动态推理模型不能捕捉到时间的相关性, 且不能将时序知识图谱的结构信息推广到未来的时间中.

随后一些研究分析了时序知识图谱的结构, 并将图神经网络(GNN)和循环神经网络模块的组合应用至时序知识图谱链接预测任务中. 递归事件网络(RE-Net)可预测时序知识图谱中多个时间点的并发事实, 并建立时间相关性模型. 虽然其采用的 RGCN 聚合器[21]可获得某一时间整体下的关系邻接信息, 但在聚合过程中会加入非目标实体的邻接信息, 这将导致模型的预测能力下降. 随后的工作对 RE-Net 模型进行改进, 提出了 RE-GCN 模型[22], 通过对 RGCN 聚合进行优化, 缩短了模型的训练时间; 同时, 在模型中加入实体静态约束条件进行链接预测, 提升了模型的性能表现.

1.2 知识超图链接预测

知识超图链接预测主要通过将超边及实体表示为低维向量空间中的嵌入向量, 以完成对未知超边的预测. 该任务的解决方法主要包括基于软规则、基于平移、基于张量分解以及基于神经网络的方法.

- 基于软规则的方法具备可解释性, 分别将知识超图中的实体和关系作为变量和谓词, 通过设置关系推理的逻辑约束条件进行未知超边的推理. MLN 首次完成一阶谓词逻辑与概率图模型的结合, 为逻辑规则赋予不同的权重, 实现对知识超图数据及规则不确定性的有效处理. 关系逻辑回归(relational logistic regression, RLR)[23]将逻辑回归算法应用至关系模型中, 提升了 MLN 模型的预测性能;
- 基于平移的方法将实体和关系嵌入到相同向量空间, 基于关系嵌入对实体向量进行平移, 从而使模型学习到实体和关系在嵌入空间中的表征, 进一步利用学习到的嵌入向量完成链接预测任务. *m*-TransH 是首个知识超图嵌入模型, 该方法将基于二元关系的 TransH 扩展至多元关系, 从而支持对未知超边进行推理. 接着, RAE[24]以 *m*-TransH 模型为基础, 通过在损失函数中加入两个表示共同参与至同一个超边的可能性数值, 实现对 *m*-TransH 方法的性能扩展. 上述是知识超图链接预测的两个主要前期工作, 它们不具备完全表达性, 在关系建模方面存在局限性. 随后, 基于空间平移的 BoxE 模型[25]被提出, 该模型将实体和关系分别表示为嵌入空间中的点和超矩形, 通过计算实体点到超边关系对应超矩形中心的距离, 预测超边存在的概率值. 该方法是目前基于平移的方法中, 唯一一个具有完全表达性的模型;
- 基于张量分解的方法通常将高阶张量分解为多个低阶张量的和. 该类方法的首个工作是 GETD 模型[26], 该模型是对知识图谱 TuckER[27]模型的扩展, 该模型具有完全表达性但仅能处理 *k*-均匀知识超图, 不能同时处理具有多种元数关系的知识超图, 训练前需要先按照关系元数对数据集进行划分再分别训练出相应关系元数对应的模型. 随后, 受到基于张量分解的知识图谱嵌入方法的启发, 基于 CP[28]和 DistMult 算法的扩展模型 *m*-CP[29]和 *m*-DistMult[29]被提了出来; 接着, 基于 SimplE[30]的可应用至多元关系的 HSimplE[29]模型也被提出. 同时, 为了考量实体位置语义信息在知识超图链接预测上的作用, 基于位置为不同实体学习不同嵌入表示的 HypE[29]模型被提了出来. 这两个模型都具备完全表达性. 随后, 为了解决基于张量分解方法过度参数化的问题, S2S[31]模型通过稀疏化核心张量来减少模型参数, 同时使用神经架构搜索技术保留其表达能力, 实现对该类方法的性能改进. RAM 模型发现, 当前工作都忽略了实体角色这一重要语义属性, 因此从角色层面提出了“角色意识建模”, 鼓励语义相关的角色拥有相近的嵌入表示. 该模型为目前知识超图链接预测的 SOTA 方法, 同时具备完全表达性;
- 基于神经网络的模型是一种有效解决知识超图链接预测问题的方法. 该类方法的首个工作 NaLP[32]引入多元关系的角色-实体形式, 使用 CNN 和 FCN 模块衡量实体与其角色的兼容性. 后续工作 HINGE[33]和 NeuInfer[34]将多元关系超边分解为一个知识图谱三元组和几个角色-实体对, 两者分别主

要依靠 CNN 和 FCN 模型完成兼容性判断. 最近的一项工作 StarE[35]应用 CompGCN 模块对分解的知识图谱三元组进行建模. G-MPNN 模型通过将 MPNN 推广到知识超图结构, 解决多元关系的链接预测问题. 然而, 神经网络模型利用较多的参数以表示知识超图结构, 在训练过程中容易出现过拟合现象, 使训练难以进行.

不同于以上所有方法, 本文所提出的时序知识超图结构同时包含时序知识图谱中的时序信息及知识超图中的多元关系, 由于时序知识超图概念由本文首次提出, 当前尚无相应方法可直接应用于时序知识超图结构以完成链接预测任务. 通过结合上述两个相关工作的主要特点, 本文从知识超图的静态结构与时序知识图谱的动态时序中同时获取实体嵌入, 将两者结合作为最终嵌入表征并应用至下游的链接预测任务中.

2 预备知识

本节将详细介绍相关背景知识, 包括本文首次提出的时序知识超图结构和时序知识超图链接预测任务的定义, 为理解本文模型架构奠定基础. 表 1 给出了本文使用的主要符号及其含义. 其中, 斜体表示变量, 粗体表示向量和矩阵.

表 1 符号列表

符号	含义	符号	含义
e, r, p, t	实体、关系、角色、时间戳	$\boldsymbol{F}^t, \boldsymbol{D}^t$	时间频率、权重矩阵
n	关系元数	ϕ	打分函数
m, d	实体嵌入层数及维度	$\odot$	元素乘积函数
L, l	角色集合及个数	σ	归一化函数
$\boldsymbol{z}, \boldsymbol{e}, \boldsymbol{c}, \boldsymbol{e}', \boldsymbol{e}^t$	实体、实体基础、角色、实体角色、实体时序嵌入向量	cat	连接函数
$\boldsymbol{w}$	角色权重	φ	激活函数
$\boldsymbol{B}, \boldsymbol{B}_a, \boldsymbol{R}$	角色、关系基、角色关系矩阵	η	损失函数

2.1 时序知识超图

据我们所知, 目前尚无任何工作给出时序知识超图结构的具体定义. 在给定时序知识超图链接预测任务的定义之前, 首先给出时序知识超图的定义.

定义 1(时序知识超图). 时序知识超图(temporal knowledge hypergraph)$\mathcal{H}$是一种超边带时间戳的异构图, 表示为 $\mathcal{H}=\{(r,p_1^r:e_1,p_2^r:e_2,...,t)\,|\,r\in R,e_i\in E,t\in T\}$, 其中, R 为关系集, E 为实体集, T 为所有可能的时间戳集. 一条时序超边表示为 m 元组 $h=(r,p_1^r:e_1,p_2^r:e_2,...,t)$, 由 n 元关系 r、关系对应每个位置上实体的 n 个角色 p_i^r 及实体 e_i 和 1 个时间戳 t 组成(m=n+2), 表示一个在 t 时刻发生的事件. 在一个时序知识超图中, 相同实体间可以存在多条时间戳不同但关系类型相同的时序超边, 例如, (供货,A 公司,B 公司,C 公司,2022-01-01)和(供货,A 公司,B 公司,C 公司,2022-02-01)两条时序超边可同时存在于同一个时序知识超图中.

2.2 时序知识超图链接预测

在明确了时序知识超图的定义后, 进一步给出在时序知识超图结构上执行链接预测任务的定义.

定义 2(时序知识超图链接预测). 给定包含所有可观测时序超边的时序知识超图$\mathcal{H}$, 时序知识超图链接预测的目标是: 通过已有的可观测时序超边, 预测在 n 元关系 r 和时刻 t 下, 由替换某一实体的实体序列组成的时序超边 $(r,p_1^r:e_1,p_2^r:e_2,...,t)$ 是否存在, 被替换的实体所在位置 i 可以是 n 下的任意一个.

3 模 型

本节主要介绍基于时序知识超图的链接预测模型 THM, 该模型同时利用时序知识超图的静态结构信息与动态时序信息生成两类实体嵌入, 以一定比例混合生成最终的实体嵌入向量, 该比例值作为超参数调整两类嵌入向量在最终实体表征中的占比. 图 2 给出了 THM 模型的整体架构.

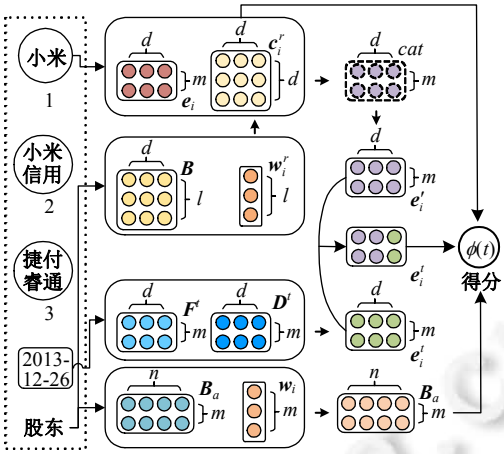


图 2 THM 总体架构图

3.1 结构静态嵌入

本节以获取时序知识超图的静态结构信息为导向, 充分利用实体在不同时序超边中的角色和位置差异, 获取实体与关系的静态嵌入向量. 以知识图谱的 DistMult 模型为基础, 首先泛化至支持多元关系的表示形式; 其次, 对于实体嵌入, 突出实体间的角色差异与位置差异; 最后, 对于关系嵌入, 突出不同实体位置上的角色位置差异, 通过两类信息充分挖掘时序知识超图在静态结构上的信息.

DistMult 作为知识图谱表示学习的一个经典方法, 通过同时学习关系、头实体和尾实体的嵌入向量并将三者线性相乘的方式获得三元组为真的概率值, 将该模型泛化至知识超图结构, 可自然地通过依次将关系嵌入与实体列表中的每个实体嵌入相乘来实现:

$$\phi(r(e_1,e_2,\ldots,e_i))=\odot(\boldsymbol{r},\boldsymbol{e}_1,\boldsymbol{e}_2,\ldots,\boldsymbol{e}_i) \tag{1}$$

针对公式(1)中的实体嵌入部分, 依次获取实体的角色差异与位置差异.

在时序知识超图数据集中, 同一个实体可能同时对应多个位置和角色, 如图 1 中的小米实体分别在关系为股东和法人的两种时序超边中对应位置 1 和位置 2 及控股公司和法定代表人角色. 为了充分表示单一实体的角色信息, 将每个实体 $e_i\in E$ 映射到一个实体嵌入矩阵中, 令 $\boldsymbol{e}_i\in\mathbb{R}^{m\times d}$ 表示实体嵌入, m 是嵌入的层数, d 是嵌入的维度. 为了充分利用实体间角色语义的差异, 为角色建立一个 $l\times d$ 维的向量空间 $\boldsymbol{B}$, l 为角色总个数, $\boldsymbol{w}_i^r\in\mathbb{R}^l$ 是角色权重向量, 语义关联性通过角色权重隐式地进行参数化:

$$\boldsymbol{c}_i^r=\sum_{l=1}^{L}\boldsymbol{B}[l]\cdot\sigma(\boldsymbol{w}_i^r)[l] \tag{2}$$

得到实体角色嵌入向量后, 通过实体基础嵌入向量与角色嵌入向量相乘, 获得包含角色语义的实体表征 $\boldsymbol{e}_i'=\boldsymbol{e}_i\cdot\boldsymbol{c}_i^r$. 受知识图谱嵌入方法 SimplE 的启发, 位置 i 的语义信息通过连接函数 cat 于单个时序超边内合并, $cat(\boldsymbol{v},x)$函数将向量 $\boldsymbol{v}$ 向左移动 x 步:

$$\boldsymbol{e}_i'=(\boldsymbol{e}_i^1,cat(\boldsymbol{e}_i^2,m\cdot d/n),\ldots,cat(\boldsymbol{e}_i^{\alpha},m\cdot d\cdot(n-1)/n)) \tag{3}$$

针对公式(1)中的关系嵌入部分, 生成衡量关系与实体兼容度的关系矩阵.

为了衡量在某个关系下, 位置、角色和所有参与实体之间的兼容度, 关系中每个位置上的角色都使用关系矩阵加以表示. 对于关系 $r\in R$, 其第 i 个位置的角色关系矩阵由 $\boldsymbol{R}_i^r\in\mathbb{R}^{n\times m}$ 表示, 其中, 第 j 行 $\boldsymbol{R}_i^r[j,:]$ 表示该角色与第 j 个位置的实体嵌入兼容性, 即当前实体位于当前位置所扮演的角色与时序超边语义信息的适配程度. 令 $\boldsymbol{B}_a\in\mathbb{R}^{n\times m}$ 为表示相同关系内角色间语义关联性的角色基础向量 $\boldsymbol{B}[l]$相关联的关系基矩阵, 该矩阵由关系所含角色基础向量拼接而成, 与角色基础向量 $\boldsymbol{B}[l]$对齐且被σ函数归一化, 用于计算角色关系矩阵[12]:

$$\boldsymbol{R}_i^r = \sum_{l=1}^{L} \sigma(\boldsymbol{w}_i^r)[l] \cdot \sigma(\boldsymbol{B}_a[l]) \tag{4}$$

3.2 时序动态嵌入

除了充分挖掘时序知识超图的结构信息外, 对于时序信息的利用同样很关键. 本节以获取时序知识超图的动态时序信息为导向, 充分利用实体所在不同时序超边的时序信息差异, 获取实体在特定时间戳下的动态嵌入向量. 为时序超边的时间戳设定2个时间特征矩阵, 分别表示当前时间戳出现的频率及其权重. 基于时间特征矩阵获得的实体时序动态嵌入向量以设定超参数的方式确定与结构静态嵌入的融合比例, 获得最终的实体表征. 将角色嵌入、实体最终表征和角色关系矩阵输入至打分函数, 获得所预测时序超边存在的概率值.

直观地说, 通过学习时间特征矩阵 $\boldsymbol{F}^t$ 和 $\boldsymbol{D}^t$, 使得模型明确如何在不同时间点上开启和关闭实体特征, 这样可以在任意时间戳下控制特征的权重并对未知时序超边进行准确的预测. 使用正弦函数作为下述公式的激活函数, 因其可以模拟多个开启和关闭状态, 实体在具体时间戳下的时序动态嵌入通过如下公式获得:

$$\boldsymbol{e}_i^t = \varphi(\boldsymbol{F}^t \cdot t + \boldsymbol{D}^t) \tag{5}$$

在分别获取到实体的静态嵌入向量和动态嵌入向量后, 该模型以一定比例混合两种向量类型并生成最终的实体嵌入向量. 这里提出一个动态嵌入函数, 该函数的输出是混合后的实体表征, 对最终的实体嵌入向量 $\boldsymbol{z}_i$ 的定义如下:

$$\boldsymbol{z}_i[k] = \begin{cases} \boldsymbol{e}_i'[k] \cdot \varphi(\boldsymbol{F}^t[k] \cdot t + \boldsymbol{D}^t[k]), & 0 \leqslant k \leqslant \lambda d \\ \boldsymbol{e}_i'[k], & \lambda d < k \leqslant d \end{cases} \tag{6}$$

其中, 向量前λd的元素捕捉实体的时序动态特征, 剩余$(1-\lambda)d$的元素捕捉实体的结构静态特征, $0 \leqslant \lambda \leqslant 1$是控制特征类型占比的超参数.

由于获取时序知识超图结构静态嵌入的策略是从实体的位置及角色角度出发, 且为多元关系中不同位置上的实体角色生成独有表征, 故可将公式(1)打分函数中的关系嵌入由角色嵌入加以替代. 在获得最终的实体嵌入向量后, 分别将角色嵌入、实体最终嵌入和角色关系矩阵输入值打分函数, 即可获得时序超边成立的打分值:

$$\phi(h) = \sum_{i=1}^{n} \langle \boldsymbol{c}_i^r, \boldsymbol{R}_i^r[1,:]\boldsymbol{z}_1, ..., \boldsymbol{R}_i^r[n,:]\boldsymbol{z}_n \rangle \tag{7}$$

3.3 模型训练

时序知识超图中的时序超边被分成训练、验证和测试集, 使用结合小批次采样技术的随机梯度下降法学习模型参数. 令 S 为训练集的一个小批次采样, 对于 S 中的每个时序超边 $h = (r, p_1^r : e_1, p_2^r : e_2, ..., t) \in S$, 共生成关系元数 n 个查询: $(r, p_1^r : e_1, ..., p_i^r : ?, ..., t)$. 针对生成的每个查询, 给出相应的一个候选答案集 C_e, 其中包含从 E 中随机选取的 q 个不同于 e_i 的实体替换至该实体位置. 然后使用最小化交叉熵作为损失函数, 该损失函数目前已被用于知识图谱和时序知识图谱的链接预测任务并显示出良好的效果:

$$\eta = -\left(\sum_{h \in S} \sum_{i=1}^{\alpha} \frac{\exp(\phi(h))}{\sum_{e_i \in C_e} \exp(\phi(r, p_1^r : e_1, ..., p_i^r : e_i, ..., t))} \right) \tag{8}$$

4 完全表达性与空间复杂度分析

完全表达性是模型的一个重要属性, 也是知识超图工作中的一个研究内容. 模型的理想属性是具有完全表达性. 如果给定任意一个时序知识超图数据集的正确时序超边集和错误时序超边集, 模型都存在一个能正确分类两集合中时序超边正误的参数表示, 那么该模型就具有完全表达性. 对于知识超图嵌入模型, 目前已有几个模型已被证明具有完全表达性; 然而对于时序知识超图嵌入模型来说, 目前还不存在具有完全表达性的模型. 如下定理 1 确立了 THM 具有完全表达性.

定理 1. THM 对时序知识超图具有完全表达性.

证明: 根据公式(6)的定义, THM 的嵌入函数将单个实体位于不同位置上的表示映射为一个向量, 通过一个 THM 的一个特殊情况证明该定理成立, 即实体嵌入为纯时序动态嵌入向量和纯结构静态嵌入向量. 这个特殊情况可通过令λ=1 或λ=0 使所有实体 $e\in E$ 在 $0\leqslant k\leqslant d$ 时实现.

如果 THM 在该特殊情况下可以实现完全表达性, 那么这两种情况下实体最终嵌入可分别写为

$$\boldsymbol{z}_i[k]=\boldsymbol{e}_i'[k]\cdot\sin(\boldsymbol{F}^t[k]\cdot t+\boldsymbol{D}^t[k]) \tag{9}$$

$$\boldsymbol{z}_i[k]=\boldsymbol{e}_i'[k] \tag{10}$$

为进一步简化证明, 依据文献[30], 可通过说明为何当 $(r,p_1^r:e_1,p_2^r:e_2,...,t)\in S$ 或 $(r,p_1^r:e_1,p_2^r:e_2,...,t)\notin S$ 时, 时序超边的得分可分别为一个正数或负数, 来证明此定理.

假设 $d=|R|\cdot|E|\cdot|T|\cdot L$, 其中, L 是一个自然数. 这些实体嵌入向量可视为$|R|$个长度为$|E|\cdot|T|\cdot L$ 的块, 对于第 j 个关系 r_j, 令实体向量 z 中除第 j 块为 1 外, 其他所有元素皆为 0. 通过这种赋值方式, 时序超边的打分仅与嵌入向量的第 j 块相关, 现将重点移至向量的第 j 块.

第 j 个块的长度(与其他所有块相似)为$|E|\cdot|T|\cdot L$, 它可以被看作是$|E|$个长度为$|T|\cdot L$ 的子块. 对于第 i 个实体 e_i, 令 z 在除第 i 个子块外的所有子块的值为 0. 有了这样的值分配, 要获得一个时序超边的得分, 只有第 j 个块的第 i 个子块是重要的. 注意, 这个子块对每个包含实体 e_i 和关系 r_j 的时序超边都是唯一的. 现将重点放到第 j 个块的第 i 个子块.

第 j 个块的第 i 个子块的长度为$|T|\cdot L$, 也可被视为$|T|$个长度为 L 的子块. 根据傅里叶正弦数列, 在 L 足够大的情况下, 可通过设置 z、F^t 和 D^t 的值, 当 $t=t_p$ 时, 第 p 个子块的 z 元素之和为 1; 当 t 是 t_p 之外的时间戳时, 为 0. 注意, 这个子块对每个包含实体 e_i、关系 r_j 和时间戳 t 的时序超边都是唯一的.

有了上述值的分配方式, 可实现当 $(r,p_1^r:e_1,p_2^r:e_2,...,t)\in S$ 或 $(r,p_1^r:e_1,p_2^r:e_2,...,t)\notin S$ 时, 时序超边的得分可为一个正数或负数.

除了完全表达性外, 本模型在面对大规模时序知识超图数据的情况下同样具备可扩展性. 如下定理 2 确立了 THM 的线性空间复杂度.

定理 2. THM 模型的空间复杂度为 $O(m_e d+Lm_r m_\alpha)$, 其中, m_e 是实体总数, m_r 是关系总数, m_α是时序知识超图中的最大关系元数.

证明: 由于时序知识超图中的关系数很少高于 7(见表 2), 所以参数 m 的赋值不会超过 3, 进而在表示角色嵌入向量上的参数最多为 $m_e d$, 在表示关系基础矩阵上的参数最多为 Lmm_α, 在表示角色权重向量上的参数最多为 $Lm_r m_\alpha+Ld$, 模型总参数最多为 $O(m_e d+Lm_r m_\alpha+Ld+Lmm_\alpha)=O(m_e d+Lm_r m_\alpha)$. 因此, THM 模型的空间复杂度为 $O(m_e d+Lm_r m_\alpha)$.

表 2 数据集统计信息

数据集	实体数	关系数	2 元数	3 元数	4 元数	5 元数	6 元数	≥7 元数	时间类型	区间
JF17K	29 177	322	56 332	34 550	9 509	2 230	37	0	无	–
FB-AUTO	3 388	8	3 786	0	215	7 212	0	0	无	–
M-FB15k	10 314	71	400 027	26	11 220	0	0	0	无	–
WikiPeople	47 765	707	337 914	25 820	15 188	2 514	718	75	无	–
CB67	7 840	67	6 200	957	476	300	205	138	时间点	2014.10.1–2021.8.4

5 实 验

本节实验在真实数据集上验证 THM 的有效性、合理性和鲁棒性. 设计思路为: (1) 在真实的时序知识超图数据集上与之前的知识超图方法分别进行宏观及细分性能对比, 以验证模型设计的有效性; (2) 执行消融实验, 以验证模型设计的合理性; (3) 通过参数敏感性实验, 以验证模型设计的鲁棒性. 在分析实验结果之前, 首先说明时序知识超图真实数据集的构建方式; 其次介绍所比较的知识超图嵌入模型及采取的实验设置; 然后给定实验评价指标.

5.1 数据集

本文通过上市公司公布的年报数据, 以公司集团为单位构建了 CB67 数据集(company business, CB), 并对数据进行了脱敏处理. 实体包括公司类型和人物类型共 7 840 个. 关系涵盖公司内的职务关系与公司间的供应和隶属关系共 67 种; 关系元数涵盖 2 元至 7 元; 规定关系对应实体列表首个实体的角色为该时序超边所描述的公司, 关系名由“关系名称”和数字“关系元数−1”组合而成(减 1 即实体列表中表示所描述公司的首个元素), 以区别由不同个数的实体列表所组成的相同关系事实, 例如“监事 3”和“监事 4”分别表示某公司由 3 个和 4 个人物组成的监事; 单个时序超边内的实体按照加入时间的递增顺序排列, 例如公司 D 和公司 E 分别于 2020-03-14 和 2020-07-21 加入公司 A 的供应链, 公司 B 于 2020-12-02 退出公司 A 的供应链, 则公司 A 于 2020 年和 2021 年的供应关系时序超边分别表示为(供应 2,A,B,C,2020-01-01)和(供应 3,A,C,D,E,2021-01-01). 整个数据集的时序超边共 8 276 个, 时间戳跨度为 2014-10-01 至 2021-08-04. 通过打乱时序超边集随机抽取 89%的数据作为训练集, 剩余数据作为验证集和测试集. 表 2 展示了该时序知识超图数据集与当前公开的知识超图数据集的统计数据, 表中“X 元数”表示 X 元关系(时序)超边数.

目前, CB67 数据集已开源至 https://github.com/zirui-chen/CB67.

5.2 基线模型和实验设置

由于目前没有可直接应用于时序知识超图结构的链接预测模型, 为了公平比较, 本文只选取同样基于张量分解的知识超图链接预测工作中, 代码开源且提供实验最优超参数的模型, 进一步筛选出能够处理具有不同元数关系数据的模型进行比较.

(1) m-TransH[3]: 将 TransH 算法由二元扩展至多元关系, 通过在超边层面对多元关系进行建模, 使得打分函数能够对未知超边是否存在进行打分;

(2) m-CP[29]: CP (canonical polyadic)分解是一种仅能处理二元关系的基于张量分解的方法, m-CP 模型则可以容纳任何数量级的关系, 实现对知识超图的关系建模;

(3) m-DistMult[29]: 对 DistMult 算法进行扩展, 将双线性得分函数泛化至多元关系, 以实现对未知超边的打分;

(4) HSimplE[29]: 受 SimplE 模型的启发, 该算法根据实体在不同关系对应超边中的不同位置, 为实体提供不同的嵌入表征;

(5) HypE[29]: 为实体在知识超图数据集中每个可能的位置学习一个相应的位置卷积权重转换器, 为不同位置的实体使用相应转换器获得实体嵌入, 随后与关系嵌入结合并输入打分函数, 产生未知超边存在的概率;

(6) RAM[12]: 这是目前知识超图链接预测上的 SOTA 模型, 该模型从实体的角色层面入手, 为知识超图中的超边提出了“角色意识建模”, 鼓励语义上相关的角色有接近的表示.

这些模型将仅使用时序超边中的关系和实体数据进行参数调优, 默认采用各模型在 JF17K 数据集上的最优超参数完成模型训练并预测时序超边, 所得实验数据用于模型性能对比. 本模型将额外使用时序超边中的时间戳信息进行模型训练.

5.3 评价指标

在 CB67 数据集上, 使用平均倒数排名 MRR 和命中率 Hits@k (k=1,3,10)这两个评价指标对模型性能进行评估. 负采样方法如下: 给定一个时序超边集合, 用 h 表示测试集中的任意一个时序超边正例. 对于 h 中位置 i 上的实体 e_i, 利用 $E-\{e_i\}$中的实体进行替换, 从而构造出超参数 q 个时序超边, 这样就构造出了一个对应于时序超边正例 h 位置 i 的时序超边负例集合 $N_a(h)$. 令 $H_a(h)=\{h\}\cup N_a(h)$, $rank_a(h)$表示基于评分函数$\phi(\cdot)$时序超边正例 h 在 $H_a(h)$中的排名. 令 r 为时序超边正例 h 对应的关系, $cond(\cdot)$为条件函数, 当条件成立时值为 1, 否则值为 0. MRR 和 Hits@k 的具体计算公式分别为

$$MRR=\frac{1}{\sum\limits_{h\in H_{test}}|r|}\sum_{h\in H_{test}}\sum_{p=1}^{|r|}\frac{1}{rank_p(h)},$$

$$Hits@k=\frac{\sum\limits_{h\in H_{test}}\sum\limits_{p=1}^{|r|}cond(rank_p(h)\leqslant k)}{\sum\limits_{h\in H_{test}}|r|}.$$

5.4 实验结果

为了评估基于时序知识超图的链接预测模型 THM 的有效性, 我们研究了以下 5 个问题.

(1) Q1: 相比于仅使用时序知识超图结构信息的知识超图嵌入模型, 额外使用到数据集中时序信息的 THM 模型能否获得更优的预测效果?

(2) Q2: 细化至不同元数关系下的链接预测, 时序知识超图嵌入模型 THM 相比知识超图 SOTA 模型 RAM 的性能表现有何差异?

(3) Q3: THM 中的静态结构模块、动态频率模块和动态权重模块的组合是否为链接预测效果最优的搭配方式?

(4) Q4: THM 模型的参数规模是否会产生训练过拟合现象, 并进一步影响模型的泛化能力?

(5) Q5: 面对不同超参数搭配下的模型训练, THM 模型是否具有鲁棒性?

(1) Q1: THM 的有效性.

表 3 给出了 THM 模型和实验所选的其他知识超图基线模型在数据集 CB67 上对比的实验结果.

表 3 链接预测结果

模型	MRR	Hit@1	Hit@3	Hit@10
m-TransH	0.063	0.053	0.065	0.076
m-CP	0.196	0.171	0.215	0.234
m-DistMult	0.238	0.216	0.256	0.268
HSimplE	0.222	0.199	0.236	0.257
HypE	0.214	0.193	0.228	0.246
RAM	0.221	0.192	0.237	0.267
THM	**0.257**	**0.235**	**0.281**	**0.289**

实验结果表明, 本文提出的 THM 模型在时序知识超图数据集上的所有评价指标均超过了基线模型. 就 MRR 指标而言, THM 模型相较于知识超图嵌入的 SOTA 模型 RAM 提升了 14.01%, 表明该模型能够有效地利用时序超边中的时序信息完成链接预测任务, 实现对数据集中时序信息的有效利用. 根据表 3 实验结果表明: 这些知识超图算法在时序知识超图数据集上的实验效果并不理想, 不仅需要将实体角色和位置等静态特征作为嵌入学习的内容, 还要充分利用时序超边中特有的时序动态信息以完成基于时序知识超图结构的链接预测. 从静态结构信息的角度分析, RAM 是目前的 SOTA 模型, 但该模型在本文所构建的时序知识超图数据集 CB67 上的实验结果显示, 其在时序知识超图数据集下所取得的效果并非最优. 相反, THM 可取得最优结果. 考虑到知识超图嵌入模型的设计与 CB67 数据集的差异性, 出现该现象的主要原因是: CB67 数据集中单个时序超边内的实体按照实体加入时间的递增顺序排序, 而非绝对按照实体角色的不同而分配实体所在的位置, 同一时序超边中的不同实体可能具有相同的角色. 例如, 时序超边(供应 3,*A*,*C*,*D*,*E*,2021-01-01), 除了公司 *A* 表示供应链中作为客户角色的公司外, 后 3 个公司都是供应链中作为供应商角色的公司, 其先后顺序的差异仅在于加入该供应链的时间先后不同. 在此背景下, 相较于 RAM 更多考虑实体角色信息的不同, THM 按实体顺序完成嵌入乘积的得分函数设计更能捕获 CB67 数据集中时序超边实体顺序在时间上的差异. 该实验现象进一步体现出为时序知识超图数据设计符合其结构特性的嵌入模型的必要性. 从动态结构信息的角度分析, RAM 结果低于 THM 的主要原因在于两点: RAM 相比本文的静态嵌入方面没有进一步增强实体位置信息对超边的语义影响; 相比本文的动态嵌入方面, 则完全没有考虑超边的动态变化趋势. 综合分析实验结果, THM 模型获得最优性能的主要原因不仅在于对超边静态结构信息的考虑, 还融入了时序知识超图中特有的时序信

息作为动态嵌入, 并通过一定比例的混合将两类信息加以结合, 充分利用知识超图的结构信息和时序信息, 以增强链接预测的性能效果.

(2) Q2: THM 在不同元数关系下的性能表现.

图 3 显示了 THM 在不同关系元数下的细分性能. 其在 CB67 不同关系元数上的性能表现一致, 超越了 RAM 模型; 在较高元数时序超边上的表现相对较弱, 这是由于数据集中的数据元数分布不均所致, 具体可参考表 2 中不同元数关系下的时序超边统计数据; 此外, THM 在二元关系数据上获得了显著的结果, 这验证了其在二元关系上的泛化能力.

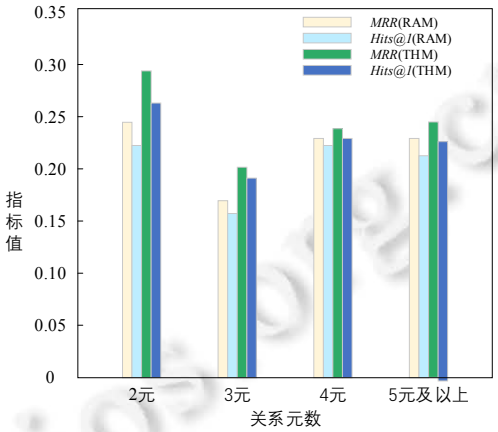


图 3 不同关系元数下的细分性能

(3) Q3: 消融实验.

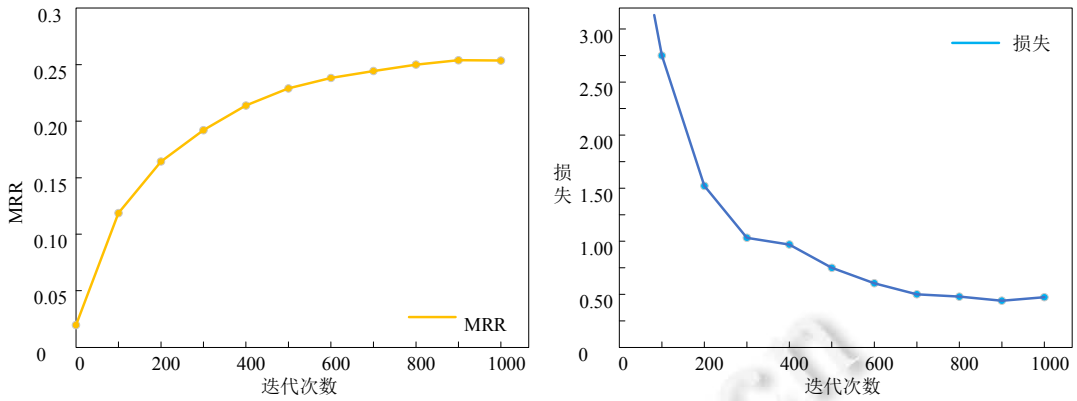
为了验证各模块对模型性能的影响, 本文基于 THM 模型设计了 6 个变体, 表 4 中自上而下的模型变体分别为纯结构、纯频率、纯权重、结构+频率、结构+权重、频率+权重. 由表 4 可见, THM 模型在各项指标上均优于其 6 个变体. 具体地, 就 Hits@1 指标而言, THM 模型相较于变体分别提升了 8.93%、94.47%、95.32%、1.70%、2.12%、93.61%. 显然, 缺少任何一个模块都会使得模型效果变差, 且相比于纯粹采用动态时序信息进行链接预测, 仅采用静态嵌入的效果更好. 这表明静态嵌入依旧是模型进行链接预测所主要使用到的信息, 而时序信息的加入可从动态角度丰富信息度, 提升模型预测性能, THM 模型的静态嵌入、时间频率及时间权重能够很好地捕捉到实体与关系之间的交互及事实随时间发展的变化趋势.

表 4 CB67 数据集上的消融实验结果

模型	MRR	Hit@1	Hit@3	Hit@10
THM-S	0.233	0.214	0.251	0.259
THM-F	0.014	0.013	0.016	0.017
THM-W	0.010	0.011	0.016	0.016
THM-SF	0.253	0.231	0.273	0.279
THM-SW	0.251	0.230	0.274	0.281
THM-FW	0.016	0.015	0.017	0.018
THM-SFW	0.257	0.235	0.281	0.289

(4) Q4: THM 是否存在训练过拟合现象.

具有大量参数的模型很容易对训练数据产生过拟合现象, 进而损害泛化性能影响测试集上的性能表现. 为了验证 THM 在过拟合方面的表现, 在训练过程中采用早停技术测试该模型是否存在过拟合现象. 图 4(a)和图 4(b)中分别绘制了以 MRR 和损失为单位的训练曲线. 随着训练的迭代进行, MRR 指标在经过快速提升后逐渐达到收敛状态; 对于损失曲线, 在前 300 次及 300–700 次迭代期间先后快速降低及缓慢降低, 并在 700 次迭代后逐渐趋于收敛.



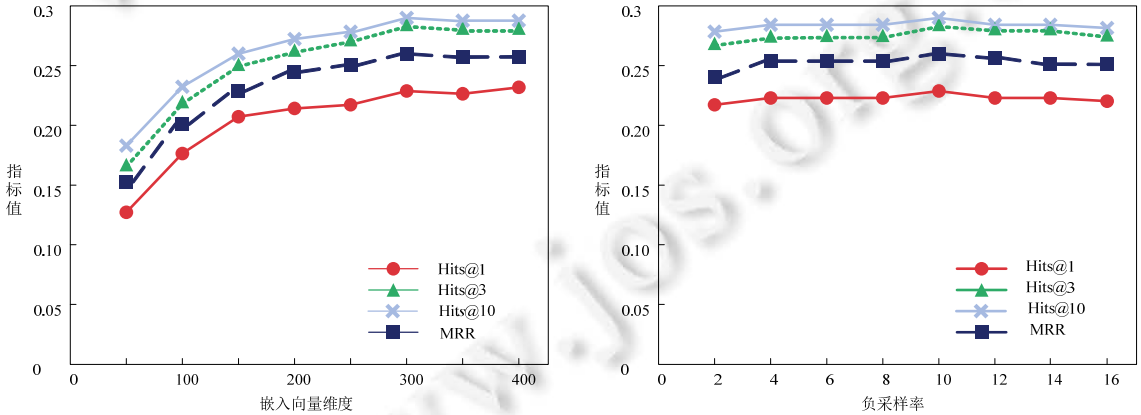
(a) MRR vs. 迭代次数　　(b) 损失 vs. 迭代次数

图 4　从 MRR 和损失指标分析过拟合现象

(5) Q5: THM 的鲁棒性.

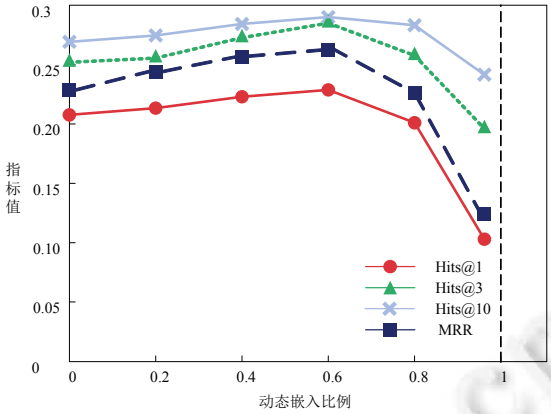
为了验证 THM 模型的鲁棒性, 从模型设计的 3 个超参数——嵌入维度 d、负采样率 q 以及动态嵌入比例 λ共 3 个维度, 在 CB67 数据集上分析超参数设定对算法性能的影响. 令嵌入维度 $d\in\{50,100,150,200,250,300,350,400\}$, 负采样率 $q\in\{2,4,6,8,10,12,14,16\}$以及动态嵌入比例$\lambda\in\{0,0.2,0.4,0.6,0.8,1\}$, 进行实验测试.

- 图 5(a)所示为不同嵌入维度下的模型性能折线图: 随着模型嵌入维度的增高, 模型性能首先获得快速提升, 并在嵌入维度达到 300 后实现了 4 个性能指标的稳定;
- 图 5(b)所示为不同负采样率下的模型性能折线图, 可见: 负采样率对 THM 模型性能的影响甚微. 该超参数对于模型性能而言是稳定的, 不会因负采样率的变动产生较大性能的差异;
- 图 5(c)所示为不同动态嵌入比例下的模型性能折线图, 可见: 动态嵌入比例对 THM 模型是一个较为敏感的超参数, 在动态嵌入比例较低时, 模型主要学习到时序知识超图的静态结构信息, 此时随着动态嵌入比例的升高, 模型的 4 项指标的性能逐渐获得提升; 但当动态嵌入比例大于 0.8 后, 模型性能急剧下滑, 表明以静态结构信息为基础、辅以时序信息可提升模型的预测性能, 而过少考虑实体角色位置的差异可能对模型学习无益. 故该参数需要根据实际应用情况采取不同的取值.



(a) 参数嵌入向量维度对指标的影响　　(b) 参数负采样率对指标的影响

图 5　参数敏感性分析



(c) 参数动态嵌入比例对指标的影响

图 5 参数敏感性分析(续)

6 总 结

本文基于时序知识超图的定义提出一个时序知识超图链接预测模型 THM, 该模型利用时序知识超图的静态结构信息, 包括实体的角色和位置信息以及关系与实体角色的兼容度信息; 同时, 利用时序知识超图的动态时序信息, 包括时间戳频率及权重信息, 以一定比例混合两类嵌入向量生成最终的实体表征, 充分利用时序知识超图中已有的结构时序信息, 提升下游链接预测任务的性能指标. 同时, 从理论上证明了 THM 具有完全表达性及线性空间复杂度. 此外, 通过公开的经营数据构建了首个时序知识超图数据集 CB67, 并在该数据集上进行了大量实验评估. 实验结果表明, THM 能够在时序知识超图数据集上有效地执行链接预测任务.

未来可行的研究方向包括将时序知识超图视为一个完整的动态图进行整体性建模, 而非分而治之地由静态结构与动态时序表征组合而成; 其次, 实现归纳式学习, 拥有对非训练集内时间戳下的实体关系进行嵌入表征的能力.

References:

[1] Wang X, Zou L, Wang CK, Peng P, Feng ZY. Research on knowledge graph data management: A survey. Ruan Jian Xue Bao/Journal of Software, 2019, 30(7): 2139−2174 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5841.htm [doi: 10.13328/j.cnki.jos.005841]

[2] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. New York: Association for Computing Machinery, 2008. 1247−1250.

[3] Wen J, Li J, Mao Y, Chen S, Zhang R. On the representation and embedding of knowledge bases beyond binary relations. arXiv:1604.08642, 2016.

[4] Fatemi B, Taslakian P, Vazquez D, Poole D. Knowledge hypergraphs: Prediction beyond binary relations. arXiv:1906.00137, 2019.

[5] Ernst P, Siu A, Weikum G. Highlife: Higher-arity fact harvesting. In: Proc. of the 2018 World Wide Web Conf. Int'l World Wide Web Conf. Steering Committee, 2018. 1013−1022.

[6] Zhang F, Yuan NJ, Lian D, Xie X, Ma WY. Collaborative knowledge base embedding for recommender systems. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2016. 353−362.

[7] Lukovnikov D, Fischer A, Lehmann J, Auer S. Neural network-based question answering over knowledge graphs on word and character level. In: Proc. of the 26th Int'l Conf. on World Wide Web. Int'l World Wide Web Conf. Steering Committee, 2017. 1211−1220.

[8] Hou ZN, Jin XL, Chen JY, Guan SP, Wang YZ, Cheng XQ. Survey of interpretable reasoning on knowledge graphs. Ruan Jian Xue Bao/Journal of Software, 2022, 33(12): 4644−4667 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6522.htm [doi: 10.13328/j.cnki.jos.006522]

[9] Yang DH, He T, Wang HZ, Wang JB. Survey on knowledge graph embedding learning. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3370−3390 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6426.htm [doi: 10.13328/j.cnki.jos.006426]

[10] Richardson M, Domingos P. Markov logic networks. Machine Learning, 2006, 62(1): 107−136.

[11] Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. In: Proc. of the AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2014. 1112−1119.

[12] Liu Y, Yao Q, Li Y. Role-aware modeling for *n*-ary relational knowledge bases. In: Proc. of the Web Conf. 2021. New York: Association for Computing Machinery, 2021. 2660−2671.

[13] Xu F, He F, Xie E, Li L. Fast OBDD reordering using neural message passing on hypergraph. arXiv:1811.02178, 2018.

[14] Hu BD, Wang XG, Wang XY, Song ML, Chen C. Survey on hypergraph learning: Algorithm classification and application analysis. Ruan Jian Xue Bao/Journal of Software, 2022, 33(2): 498−523 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6353.htm [doi: 10.13328/j.cnki.jos.006353]

[15] Yang B, Yih WT, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. arXiv: 1412.6575, 2014.

[16] Trivedi R, Dai H, Wang Y, Song L. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In: Proc. of the Int'l Conf. on Machine Learning. 2017. 3462−3471.

[17] García-Durán A, Dumančić S, Niepert M. Learning sequence encoders for temporal knowledge graph completion. arXiv:1809. 03202, 2018.

[18] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. New York: Curran Associates, Inc., 2013. 2787−2795.

[19] Leblay J, Chekol MW. Deriving validity time in knowledge graph. In: Companion Proc. of the the Web Conf. 2018. Int'l World Wide Web Conf. Steering Committee, 2018. 1771−1776.

[20] Sadeghian A, Armandpour M, Colas A, Wang DZ. ChronoR: Rotation based temporal knowledge graph embedding. In: Proc. of the AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2021. 6471−6479.

[21] Schlichtkrull M, Kipf TN, Bloem P, Berg RV, Titov I, Welling M. Modeling relational data with graph convolutional networks. In: Proc. of the European Semantic Web Conf. Cham: Springer, 2018. 593−607.

[22] Li Z, Jin X, Li W, Guan S, Guo J, Shen H, Wang Y, Cheng X. Temporal knowledge graph reasoning based on evolutional representation learning. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: Association for Computing Machinery, 2021. 408−417.

[23] Kazemi SM, Buchman D, Kersting K, Natarajan S, Poole D. Relational logistic regression. In: Proc. of the 14th Int'l Conf. on the Principles of Knowledge Representation and Reasoning. Vancouver: University of British Columbia, 2014.

[24] Zhang R, Li J, Mei J, Mao Y. Scalable instance reconstruction in knowledge bases via relatedness affiliated embedding. In: Proc. of the 2018 World Wide Web Conf. Int'l World Wide Web Conf. Steering Committee, 2018. 1185−1194.

[25] Abboud R, Ceylan I, Lukasiewicz T, Salvatori T. Boxe: A box embedding model for knowledge base completion. In: Advances in Neural Information Processing Systems. New York: Curran Associates, Inc., 2020. 9649−9661.

[26] Liu Y, Yao Q, Li Y. Generalizing tensor decomposition for *n*-ary relational knowledge bases. In: Proc. of the Web Conf. New York: Association for Computing Machinery, 2020. 1104−1114.

[27] Balažević I, Allen C, Hospedales TM. Tucker: Tensor factorization for knowledge graph completion. arXiv:1901.09590, 2019.

[28] Hitchcock FL. The expression of a tensor or a polyadic as a sum of products. Journal of Mathematics and Physics, 1927, 6(1−4): 164−189.

[29] Fatemi B, Taslakian P, Vazquez D, Poole D. Knowledge hypergraphs: Prediction beyond binary relations. arXiv:1906.00137, 2019.

[30] Kazemi SM, Poole D. Simple embedding for link prediction in knowledge graphs. In: Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2018. 4289−4300.

[31] Di S, Yao Q, Chen L. Searching to sparsify tensor decomposition for *n*-ary relational data. In: Proc. of the Web Conf. New York: Association for Computing Machinery, 2021. 4043−4054.

[32] Guan S, Jin X, Wang Y, Cheng X. Link prediction on *n*-ary relational data. In: Proc. of the World Wide Web Conf. New York: Association for Computing Machinery, 2019. 583−593.

[33] Rosso P, Yang D, Cudré-Mauroux P. Beyond triplets: Hyper-relational knowledge graph embedding for link prediction. In: Proc. of the Web Conf. New York: Association for Computing Machinery, 2020. 1885−1896.

[34] Guan S, Jin X, Guo J, Wang Y, Cheng X. Neuinfer: Knowledge inference on *n*-ary facts. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 6141−6151.

[35] Galkin M, Trivedi P, Maheshwari G, Usbeck R, Lehmann J. Message passing for hyper-relational knowledge graphs. arXiv:2009.10847, 2020.

附中文参考文献:

[1] 王鑫, 邹磊, 王朝坤, 彭鹏, 冯志勇. 知识图谱数据管理研究综述. 软件学报, 2019, 30(7): 2139−2174. http://www.jos.org.cn/1000-9825/5841.htm [doi: 10.13328/j.cnki.jos.005841]

[8] 侯中妮, 靳小龙, 陈剑赟, 官赛萍, 王元卓, 程学旗. 知识图谱可解释推理研究综述. 软件学报, 2022, 33(12): 4644−4667. http://www.jos.org.cn/1000-9825/6522.htm [doi: 10.13328/j.cnki.jos.006522]

[9] 杨东华, 何涛, 王宏志, 王金宝. 面向知识图谱的图嵌入学习研究进展. 软件学报, 2022, 33(9): 3370−3390. http://www.jos.org.cn/1000-9825/6426.htm [doi: 10.13328/j.cnki.jos.006426]

[14] 胡秉德, 王新根, 王新宇, 宋明黎, 陈纯. 超图学习综述: 算法分类与应用分析. 软件学报, 2022, 33(2): 498−523. http://www.jos.org.cn/1000-9825/6353htm [doi: 10.13328/j.cnki.jos.006353]

陈子睿(1998－), 男, 博士, 主要研究领域为知识表示学习, 大型语言模型.

王鑫(1981－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为知识图谱数据管理, 图数据库, 大数据分布式处理.

王晨旭(1998－), 男, 硕士, CCF 学生会员, 主要研究领域为知识表示学习.

张少伟(1996－), 男, 硕士, 主要研究领域为知识表示学习, 知识图谱构建.

闫浩宇(1997－), 男, 硕士, 主要研究领域为知识表示学习.