基于双层随机游走的关系推理算法

刘 峤 韩明皓 江浏祎 刘 瑶 耿 技
(电子科技大学信息与软件工程学院 成都 610054)

摘 要 关系推理是知识库构建的关键技术之一,典型应用场景包括关系预测和实体链接等.关系推理研究的问题是如何利用知识库中已有的知识推理得到新的知识.当前主流知识库采用的推理模型包括潜在因子模型和随机游走模型.前者将实体和关系映射到一个低维实数向量空间,通过向量相似度计算实现推理.后者基于一阶谓词逻辑进行实体间的关系推理,通过随机算法降低算法复杂度.比较而言,前者由于需要进行大规模矩阵运算而计算复杂度较高,后者则因为采用了随机采样方法,难以完全利用知识库中已有的结构化信息,而导致召回率较低.通过研究现有随机游走模型基本假设存在的问题,提出了两项新的推理建模假设.首先,以PRA为代表的随机游走模型采用关系单向性假设,将知识库中的实体关系三元组视为一阶Horn子句,将关系处理为主语和宾语间的偏序关系,该文提出的假设是,尽管实体间的关系从字面和句法上具有方向性,但关系所包含的信息对两侧实体而言具有语义上的双向性,允许关系推理算法利用从宾语到主语的逆向关系语义进行知识推理;其次,PRA算法采用一阶谓词逻辑进行推理,并通过引入一个随机采样机制来避免穷举搜索和提高计算速度,该文认为这是导致PRA算法及类似算法无法完全利用知识库中已有信息的一个主要原因,据此提出了一个新的假设,即知识库中特定关系子网的拓扑结构所包含的信息可以被利用来改善随机游走模型的关系推理结果,为验证上述假设的有效性,提出了一种基于双层随机游走策略的关系推理新算法.在WN18、FB15K和FB40K等公开数据集上的实验结果表明,该算法能够有效地提高基于随机游走的关系推理模型的准确性和召回率,性能显著优于当前主流的基于潜在因子模型的关系推理算法.

关键词 关系推理;统计关系学习;知识库扩容;随机游走;路径排序算法;人工智能
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2017.01275

Two-Tier Random Walk Based Relational Inference Algorithm

LIU Qiao HAN Ming-Hao JIANG Liu-Yi LIU Yao GENG Ji
(*School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu* 610054)

Abstract Relational inference is one of the crucial techniques for knowledge base population tasks, typical application scenarios include relationship prediction and entity linking. The challenging problem of relational inference is how to infer new relations between entities from the facts existed in the knowledge bases. The reasoning models adopted in current mainstream knowledge bases can be divided into two categories: the latent factor models and the random walk models. The latent factor models realize the reasoning by mapping the entities and relations into a low dimensional real-valued vector space, and then computing with corresponding vector similarity measures. The random walk models, however, are based on the first-order predicate logic to deduce the reasoning between the entities and reduce the algorithm complexity through stochastic algorithm. In comparison, the efficiency of the latent factor models usually suffer from their

收稿日期:2015-12-30;在线出版日期:2016-11-25.本课题得到国家自然科学基金重点项目(61133016,U1401257)、国家自然科学基金青年项目(61502087)、四川省高新技术及产业化面上项目(2017GZ0308)资助.**刘 峤**,男,1974年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为机器学习与数据挖掘、专家系统与推荐系统.E-mail:qliu@uestc.edu.cn.**韩明皓**,男,1992年生,硕士,主要研究方向为机器学习与数据挖掘.**江浏祎**,女,1990年生,硕士,中国计算机学会(CCF)会员,主要研究方向为机器学习与数据挖掘.**刘 瑶**,女,1978年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为机器学习算法、社会网络分析.**耿 技**,男,1963年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为大数据分析、信息安全.

computational complexity caused by large-scale matrix computation operations. While the random walk models usually suffer from their low recall rates, due to the fact that it is difficult to fully utilize all of the available structure information provided by the knowledge bases with any random sampling design. This work studied the potential problems of the basic assumptions adopted by the existing random walk models, and proposed two new inference modeling assumptions thereby. Firstly, the random walk models represented by the Path Ranking Algorithm (PRA) adopt the unidimensionality assumption of the relationships in between the entities. In typical random walk models, the entity-relation-entity tuples that existed in the knowledge base are regarded as first-order Horn clauses, in which the relationships are treated as partial ordering relations between the subjects and the objects. Our hypothesis is that although the relation between two entities is literally, syntactically directional, the information conveyed by this relation is equally shared between the connected entities on both side, thus all of the relations are semantically bidirectional, which allows the relational reasoning algorithm to use the inverse relation semantics from the object to the subject for reasoning. Secondly, the PRA algorithm makes use of the first-order predicate logic for relational reasoning, it also introduces a random sampling scheme in order to avoid exhaustive search in the path space and to speed up the calculation process. However, we argue in this paper that this maybe one of the major reasons that explains why the PRA and alike algorithms can not make full use of the existing information in the knowledge base for relational inference tasks, and then we propose an alternative assumption for remedy, which claims that the topology structures of the relation-specific sub-graphs in knowledge bases can be exploited to improve the performance of the random-walk based relational inference algorithms. In order to verify the validity of the above assumptions and algorithm, we propose a novel relational inference algorithm based on a two-tier random walk strategy. Experimental results on benchmark datasets include WN18, FB15K, and FB40K, show that the proposed algorithm can be very effective in promoting the accuracy and recall rate of the random walk models. The proposed algorithm also outperforms other prevalent latent factor models on each data sets.

Keywords relational inference; statistical relational learning; knowledge base population; random walk; path ranking algorithm; artificial intelligence

1 引 言

关系推理是统计关系学习方法学研究关注的基本问题,也是当前知识图谱领域研究的热点问题.知识图谱是图状结构的知识库,其中的知识(事实)以"实体-关系-实体"三元组的形式表达和存储.本文的研究对象为知识图谱上的关系推理问题,非图结构知识库的关系推理不在讨论范围内,文中提到的知识库均特指图结构的知识库,即知识图谱.关系推理的任务目标是利用知识图谱中已有的知识,采用统计机器学习方法推理实体间可能存在的关系[1,2].

受现阶段信息抽取技术水平的限制,知识的不完备性是制约知识图谱应用的一个主要因素[3].据统计,仅在 Freebase 知识库的人物资料部分,就有约 71%的人缺少"出生地"信息、约 97%的人缺少"双亲"信息[4].信息的缺失极大地制约了各种知识库应用的性能.然而研究表明,知识库缺失的信息中有一部分可以通过已有知识推理得到[5].例如,已知拜登在奥巴马政府中任职副总统,且已知奥巴马是民主党人,则可以通过推理得到拜登是民主党人.

与通过开放域信息抽取获得知识的方式相比,基于知识库的关系推理能够获得高质量的知识,且不存在实体歧义和共指等问题,因此业界和学术界均十分重视对关系推理算法的研究.随着算法研究近期不断取得新进展,相关成果在越来越多的主流知识库产品中取得成功应用,该研究方向也逐渐成为人工智能、机器学习和数据挖掘等领域近期共同关注的热点[2].

当前主流的关系推理方法按照其基本理论模型

的不同可以分为两大类：潜在因子模型和随机游走模型[1]. 这两类模型都已成功应用在主流的知识库产品中. 例如，卡内基梅隆大学发布的 NELL 知识库采用了随机游走模型构建系统的关系推理模块，谷歌公司的 Knowledge Vault 项目则采用了潜在因子与随机游走模型相结合的混合模型用于知识评估[6-8]. 然而，这两类模型各自存在一些局限性：与潜在因子模型相比，随机游走模型的计算效率更高，模型的可解释性较好，但算法准确性和召回率则显著低于前者. 为解决随机游走模型准确性和召回率低的问题，本文首先研究并揭示了问题产生的原因，然后提出了针对性的解决方案，论文的主要贡献包括：

(1) 首次在相同的实验条件下对两类模型中的代表性算法进行性能比对，并深入分析了性能差异产生的原因，找到了现有随机游走模型假设的不合理性，并据此提出了两项新的关系推理建模假设.

(2) 提出了一种新型关系推理算法，称为双层随机游走算法(Two-tier Random Walk Algorithm，TRWA)，该算法采用新的特征模式建模思路，将图中节点链接模式特征区分为全局(跨关系)模式和局部(关系内)模式，分别采用随机游走方式进行特征建模. 在此基础上，提出了一个完整的关系推理算法原型，并通过公开基准数据集验证了算法的有效性，通过与当前主流算法进行性能比较，TRWA 算法在关系推理准确性和召回率等关键指标上优于相关工作.

2　相关工作

在统计关系学习方法发展的早期阶段，主要的关系推理方法是基于一阶谓词逻辑规则进行推理[9,10]，代表性工作是 Richardson 等人[11]提出的马尔科夫逻辑网络(Markov Logic Networks，MLN)模型，该模型将马尔科夫随机场与一阶谓词逻辑规则结合起来，通过构建逻辑网络实现对实体间关系的建模和推理. 其主要优点是关系推理准确性高，缺点是需依赖专家构建一阶谓词逻辑规则，而且模型计算复杂度高，难以适应大规模复杂知识库上的关系推理任务要求.

近年来，随着开放域信息抽取技术的兴起，知识库扩容相关技术成为多方关注的焦点，学术界开始致力于研究更加准确、高效和自动化程度更高的关系推理方法，其中引起广泛关注的方法模型是以 RESCAL 算法[12]为代表的潜在因子模型和以 PRA (Path Ranking Algorithm)算法[13]为代表的随机游走模型.

2.1　潜在因子模型

潜在因子模型(Latent Factor Model)的基本思想是将知识库中的实体及其关系映射到特定维度的向量空间中，得到实体与关系的潜在因子表达式，并按照一定规则计算出任意实体对之间存在某种关系的可能性[14]. 按照映射方式的不同，潜在因子模型可以进一步分为两类：张量分解模型(Tensor Factorization Model)和嵌入式模型(Embedding Model).

张量分解模型的代表性的工作是 Nickel 等人[12]提出的 RESCAL 模型，该模型采用三维张量对知识图谱进行建模，推理模型如式(1)所示：

$$f_r(h,t)=\boldsymbol{h}^{\mathrm{T}}\boldsymbol{M}_r\boldsymbol{t} \tag{1}$$

其中：$f_r(h,t)$ 表示三元组 (h,r,t) 中首实体 h 和尾实体 t 之间存在关系 r 的可能性. 黑体符号表示二阶张量，其中 $\boldsymbol{h}$ 和 $\boldsymbol{t}$ 表示三元组中的实体投影到 k 维向量空间得到的向量，$\boldsymbol{M}_r$ 表示 $k\times k$ 矩阵，对知识图谱中的每种关系 r，均有唯一的 $\boldsymbol{M}_r$ 与之相对应.

受张量分解模型的启发，Bordes 等人[15]提出了一种新的潜在因子模型，称为结构化嵌入式(Structured Embedding，SE)模型. 其基本假设是：若实体 h 与 t 之间存在关系 r，则这两个实体映射到潜在因子空间后(通过关系 r 做线性变换实现)，二者的欧氏距离应趋近于零. 其推理模型如式(2)所示：

$$f_r(h,t)=\|\boldsymbol{M}_{rh}\boldsymbol{h}-\boldsymbol{M}_{rt}\boldsymbol{t}\|_2 \tag{2}$$

其中：矩阵 $\boldsymbol{M}_{rh}$ 和 $\boldsymbol{M}_{rt}$ 分别用于实现关系三元组两侧实体关于关系 r 的线性变换. 由于该模型涉及对实体关系的两次分解，因此计算复杂度较高，Bordes 等人[16]进一步提出了更为简化的 TransE 模型. 与 SE 模型的区别在于，TransE 模型不再将关系视为对实体的线性变换，而是将关系与实体视为同一类对象，共同映射到相同维度的潜在因子空间中，并假设：$\boldsymbol{h}+\boldsymbol{r}\approx\boldsymbol{t}$. TransE 的推理模型如式(3)所示：

$$f_r(h,t)=\|\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\|_2 \tag{3}$$

研究表明，TransE 模型在处理反身关系和复杂关系时的性能表现不佳[17]. Lin 等人[18]认为实体和关系是两种不同类型的对象，将两者映射到同一向量空间的假设并不合理，并据此提出了 TransR 混合模型. TransR 模型将实体和关系分别投影到不同的向量空间中. TransR 的推理模型如式(4)所示：

$$f_r(h,t)=\|\boldsymbol{M}_r\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{M}_r\boldsymbol{t}\|_2 \tag{4}$$

从式(4)可以看出，TransR 是对 TransE 和 SE 模型的融合，实验表明 TransR 模型在处理复杂知

识库推理任务时的性能优于另外两者[18].

潜在因子模型的主要缺点是建模时需要对全局知识库进行分解,分别得到实体和关系的潜在因子表达式,对于知识更新频繁的大规模知识库而言,采用该模型的计算代价较高.另一个缺点是模型的可解释性差,潜在因子空间的维度和因子的含义均难以确定,对预测结果难以提供有效解释.上述问题限制了潜在因子模型在大规模知识图谱中的推广应用,而随机游走模型则在近年来吸引了业界和学术界的广泛关注.

2.2 随机游走模型

将随机游走用于关系推理的思路主要受 FOIL (First Order Inductive Learner)算法的启发[19].该算法的基本思想是对每种关系类型在知识图谱上进行穷举搜索,获取指定长度的 Horn 子句集合作为预测该关系存在的特征模式,并通过机器学习得到关系判别模型.该算法在小规模数据集上获得了较高的预测准确率,表明知识图谱中的"实体-关系"关联模式可用于关系推理建模.然而由于穷举搜索的计算开销过大,使得该算法难以胜任大规模关系推理任务.

为解决大规模知识库关系推理的效率问题,Lao 等人在 FOIL 算法的基础上提出了 PRA(Path Ranking Algorithm)算法[13].与 FOIL 算法的区别在于,PRA 算法采用抽象的实体间关系路径取代了具象的 Horn 子句作为关系推理的依据,并采用随机游走的方式替代 FOIL 算法采用的穷举搜索策略以提高计算效率.实验表明,PRA 算法不仅保持了 FOIL 算法的关系推理准确性,而且大幅提高了模型的计算效率[20,21].

PRA 算法为解决大规模知识库的复杂关系推理问题提供了一个可行的解决方案,并很快被应用于 NELL 和 Knowledge Vault 等大型知识库项目中[6-8].但由于该算法与当前性能表现最好的潜在因子模型(如 TransE)相比,在关系推理准确率和召回率方面仍有较大差距,学术界随后开展了一系列相关研究.

对 PRA 算法的改进思路主要有两种:一种思路是通过改进特征建模方法提高算法性能,主要的手段是设计新的机器学习算法或特征选择方法[22].例如,Gardner 等人[23]发现路径特征的稀疏性是影响 PRA 算法性能的原因之一,因此考虑采用关系合并的方式进行特征降维,即首先将知识库中的关系投影到连续的潜在因子空间中,然后通过对关系向量进行相似度计算实现关系合并,从而降低路径特征的维度.实验表明,通过特征降维能够显著提升 PRA 算法的性能.

另一种思路是利用随机游走模型与潜在因子模型在性能上的互补性,设计新的混合模型.例如,Nickel 等人[24]通过实验发现,当知识图谱中包含大量强连通分量时,PRA 算法的召回率会受到负面影响.据此提出了一个混合模型,针对知识库中 PRA 不能很好建模的部分关系,采用基于张量分解的 RESCAL 模型做补充推理,由此既保留了 PRA 算法计算效率高的特点,同时也提高了算法整体关系推理性能.

2.3 本文工作与相关工作的关系

本文提出的 TRWA 算法采用随机游走模型进行推理建模,目标是解决 PRA 算法在实际关系推理任务中准确性和召回率低的问题.与其他基于随机游走模型的相关工作相比,本文的研究采用了不同的思路,即从 PRA 算法基本假设中存在的不合理因素出发对算法进行改造.最终得到的算法模型与 PRA 算法相比,不仅在准确性和召回率指标上有了显著提高,而且保持了 PRA 算法的计算效率,使之能够适用于大规模知识库的复杂关系推理任务.TRWA 算法与 PRA 算法的主要区别在于如下两方面.首先,TRWA 算法将知识库中的关系视为双向联系,采用无向图来表达知识图谱,区别于 PRA 算法采用有向图的建模方式,从而增加了实体间关系路径特征模式(以下简称路径特征模式)的数量,从而提高了算法召回率.其次,TRWA 算法将路径特征模式区分为全局模式和局部模式,引入双层随机游走机制对路径特征进行评估,进一步提高了算法的准确性和召回率.

在对 TRWA 算法的性能进行实验评估时,除了基于随机游走的 PRA 算法外,本文还选择了潜在因子模型中 3 种代表性算法进行性能对比.原因是潜在因子模型在社会网络分析和推荐系统等领域取得了广泛应用,是当前的研究热点,但调研表明,这两种模型的研究工作很少进行相互比较.我们通过实验发现,潜在因子模型在中小型数据集上的综合性能表现显著优于现有的随机游走模型.因此本文将其放在一起比较,希望通过实验证明:随机游走模型不仅在计算效率上优于计算密集型的潜在因子模型,而且可以在关系预测的准确率和召回率等指标上取得优势.

3　基于双层随机游走的关系推理算法

首先介绍本文的符号体系. 知识图谱中的知识元素以“实体-关系-实体”三元组的形式存储，以符号 $\boldsymbol{G}$ 表示知识图谱，n 为 $\boldsymbol{G}$ 中包含的三元组个数，即图 $\boldsymbol{G}$ 中的边数；$\boldsymbol{e}$ 表示知识图谱中包含的实体集合，m 为 $\boldsymbol{e}$ 中包含的实体个数，即图 $\boldsymbol{G}$ 中的顶点总数；$\boldsymbol{R}$ 表示知识图谱中包含的关系种类集合，n' 为 $\boldsymbol{R}$ 中包含的关系种类个数；r_i 表示 $\boldsymbol{R}$ 中的第 i 个关系，$\boldsymbol{G}_i$ 表示满足第 i 个关系的实体关系三元组所构成的子图.

本文以 (h,r,t) 的形式表示图 $\boldsymbol{G}$ 中的实体关系三元组，其中 r 为关系，$h,t\in\boldsymbol{e}$ 为关系两侧的实体，分别称为 $head$ 实体和 $tail$ 实体. 知识图谱中可以包含多种关系，如配偶关系、隶属关系等，因此对于任意关系 $r_i\in\boldsymbol{R}$，可以定义两组均值：平均头尾比 (hpt_i) 和平均尾头比 (tph_i)，如式(5)所示：

$$hpt_i=\frac{\sharp tuple_i}{\sharp tail_i};\ tph_i=\frac{\sharp tuple_i}{\sharp head_i} \tag{5}$$

其中：$\sharp tuple_i$ 表示 $\boldsymbol{G}_i$ 中包含的实体关系三元组个数，$\sharp head_i$ 和 $\sharp tail_i$ 分别表示 $\boldsymbol{G}_i$ 中包含的 $head$ 实体和 $tail$ 实体的总数. 根据式(5)，可以进一步将知识图谱中的所有关系归纳为 4 种类型：一对一类型(1∶1)、一对多类型(1∶M)、多对一类型(M∶1)和多对多类型(M∶M). 如式(6)所示：

$$\begin{cases} hpt_i<\delta \text{ 且 } tph_i<\delta \Rightarrow 1:1 \\ hpt_i<\delta \text{ 且 } tph_i\geq\delta \Rightarrow 1:\mathrm{M} \\ hpt_i\geq\delta \text{ 且 } tph_i<\delta \Rightarrow \mathrm{M}:1 \\ hpt_i\geq\delta \text{ 且 } tph_i\geq\delta \Rightarrow \mathrm{M}:\mathrm{M} \end{cases} \tag{6}$$

其中：参数 δ 的取值范围是 $(1,+\infty)$，研究者通常将其取为 1.5[13-15]. 为了保持实验条件的一致性，本文也采用 $\delta=1.5$ 作为关系类型划分标准.

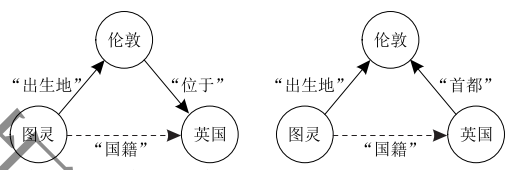
3.1　算法设计思想概述

TRWA 算法的设计思想源于 FOIL 算法和 PRA 算法. 具体说来，是将知识库中的实体关系三元组视为一阶 Horn 子句，然后基于一阶谓词逻辑进行实体间的关系推理[9,21]. 例如，已知如下的两组实体关系三元组：(图灵，出生地，伦敦)，(伦敦，位于，英国)，则如下的一阶谓词逻辑推理规则成立：

$$\begin{aligned} &(\text{图灵，出生地，伦敦}) \\ \wedge\ &(\text{伦敦，位于，英国}) \\ \Rightarrow\ &(\text{图灵，国籍，英国}) \end{aligned} \tag{7}$$

基于谓词逻辑规则的推理方法是知识推理研究领域普遍采用的经典常规方法，其有效性经过了长期的实践验证，在大规模知识图谱领域也有成功的应用案例，例如 YAGO 和 Stanford Elementary 等知识库均采用基于谓词逻辑规则的方法进行推理. 然而，对于面向开放域的知识库而言，由于关系种类繁多且复杂，依靠人工编写逻辑规则或采用穷举搜索实现规则提取都是不可行的. 因此 PRA 算法提出采用随机游走的方式进行规则采样，从而将规则构造问题转化为路径特征模式的搜索问题，并且通过 NELL 项目的实践验证了该方法的有效性[7]. TRWA 算法继承了 PRA 算法的路径特征建模思想，但是建模方法上提出了两个完全不同的基本假设，分别介绍如下.

首先，PRA 算法的基本假设是知识库中关系的有向性，即依据知识图谱中实体间的有向关系进行关系推理. 例如，在图 1(a)中，从节点“图灵”出发，可以根据关系路径“出生地→位于”推理出在“图灵”与“英国”之间可能存在“国籍”关系. 该假设的合理性在于，实体关系三元组从语法角度可以视为“主谓宾”关系，按照关系的方向进行推理，某种程度上顺应了语义的连贯性，是有效的谓词逻辑推理规则.



(a) 适用一阶谓词逻辑规则推理的情况　　(b) 不适用一阶谓词逻辑规则推理的情况

图 1　关系推理的简单案例

然而，该假设的局限性在于忽视了关系的多样性和语法的复杂性. 例如图 1(b)中的两条实线关系，对于“国籍”关系的推理而言，是比图 1(a)中的实线关系更强的证据. 然而在有向图假设下，模型只能沿着有向边进行推理，由于图 1(b)中不存在从“英国”到“伦敦”的有向边，因此 PRA 无法通过图 1(b)得到从“图灵”到“英国”之间的路径特征模式，因而无法形成对两个实体之间存在“国籍”关系的直接推理.

由此可见，实体关系三元组中的关系虽然具有语法上的方向性(定义了两侧实体间的一种偏序关系)，然而它所包含的语义信息对于关系两侧的实体而言是双向的. 例如在图 1(b)中，(英国，首都，伦敦)从语法上可解读为“英国的首都是伦敦”，显然其中的实体之间存在严格的偏序关系，即“英国”和“伦敦”之间的顺序不可交换. 然而从该事实所包含的语义来看，我们完全可以直接从中获得如下的信息“伦敦是英国的首都”. 这种情况可以推广到其他任意

的关系类型,例如“夫妻(CoupleOf)”关系两侧的实体本身就是对等的,因此关系本身并不包含偏序关系;而“父子(SonOf)”关系两侧的实体虽不对等(不能互换),但包含的语义关系对双方而言是等同的.因此,如果生硬地将实体间的关系规定为有向边,也就意味着在推理时放弃了逆向关系包含的有用信息,从而在很大程度上限制了随机游走模型能够发现的特征路径数量,降低了模型的预测性能.

因此本文认为,尽管关系在语法上是有向的,但从实体关系三元组包含的语义信息来看,该信息对于关系两侧的实体而言是对等的,对于关系推理任务而言,逆向的关系也包含着不应被忽视的信息.如果将关系视为无向边,则相当于放宽了 PRA 对随机漫步者的采样约束,由此可以得到更多的关系路径特征模式(即逻辑推理规则),提高“有效模式”的发现率.

据此提出 TRWA 算法的第 1 项假设如下:关系具有语义双向性,可以采用无向图来表达知识图谱,并依据路径特征模式进行关系推理.

需要说明的是,TRWA 算法的设计思想是基于无向图假设的(即不区分关系的方向性).然而,在描述和实现算法时,本文采用的方法是为知识图谱中的每个关系 r_i 增加一个逆向关系 r_i^{-1}.这样处理的目的有两个:一是与人们对于关系的直觉(偏序性)保持一致,便于阐述和分析算法的设计原理;二是对图 G 中每条有向边增加相应的逆向边后,对随机游走模型而言,从逻辑上仍是将图 G 作为无向图看待的.

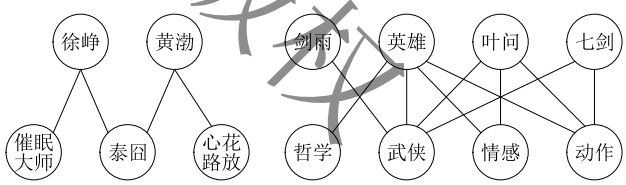
除了有向图假设的缺陷外,我们还发现了 PRA 算法基本假设中存在的另外一个问题:对一对一(1∶1)、一对多(1∶M)、多对一(M∶1)及多对多(M∶M)这 4 种类型的关系均采用同一方式建模,忽视了不同关系类型中所包含的信息.通过实验发现,PRA 算法不擅长处理 1∶M 和 M∶M 类型的关系推理任务.表 1 给出了 PRA 算法和 TransE 算法在 FB15k 数据集中 4 种关系类型上的推理结果命中率情况(*hit*@10 的意思是正确的结果出现在推理结果列表前 10 名的百分比,有关实验细节详见本文第 4 节).可见与潜在因子模型 TransE 相比,PRA 算法在 1∶M 和 M∶M 类型关系上的推理性能处于明显劣势.

表 1 PRA 和 TransE 算法在 FB15k 数据集上的 *hit*@10

算法/关系类型	TransE/%	PRA/%
1∶1	71.5	63.3
1∶M	49.0	20.4
M∶1	85.0	81.4
M∶M	72.9	32.6

通过对 PRA 算法进行分析,我们进一步发现该问题与算法的随机游走策略有关.为了避免特征空间维度过高导致算法性能恶化,PRA 算法限制路径模式特征的长度最大值为 3(与 FOIL 算法限制 Horn 子句的长度情况类似),同时不允许路径模式中出现环路.如果对 4 种类型的关系均采用同样的随机游走策略,上述限制将造成算法对关系密集型网络(特别是 M∶M 类型)的关系路径特征模式的采样不充分,从而导致算法对关系密集型网络的推理性能不良.

以图 2 为例,其中图 2(a)仅包含“参演”关系,为一对多类型,图 2(b)仅包含“电影类型”关系,为多对多类型.将其分别表示为二部图,位于上层的是 *head* 实体,下层节点为 *tail* 实体.可以看出这两种类型的关系本身均包含可用于推理的信息量.例如从图 2(a)的“徐峥”出发,经 3 跳之后,有较大的可能性推导出(徐峥,参演,心花路放)的结论,该结果具有较强的合理性.类似地,从图 2(b)的“剑雨”出发,也易于推导出(剑雨,类型,动作)这样的合理结果.以上这两种情况均可以抽象为形如“$r_i \rightarrow r_i^{-1} \rightarrow r_i$”的路径特征模式(其中 r_i^{-1} 表示关系 r_i 的逆).虽然在采用无向图假设进行建模后,使用 PRA 算法同样可以发现这样的路径模式,但受其随机游走策略中关于跳数和环路的限制,PRA 算法并不能充分利用这种关系模式所包含的信息进行推理.在本文 4.3 节将结合实验对上述分析结果进行进一步验证.



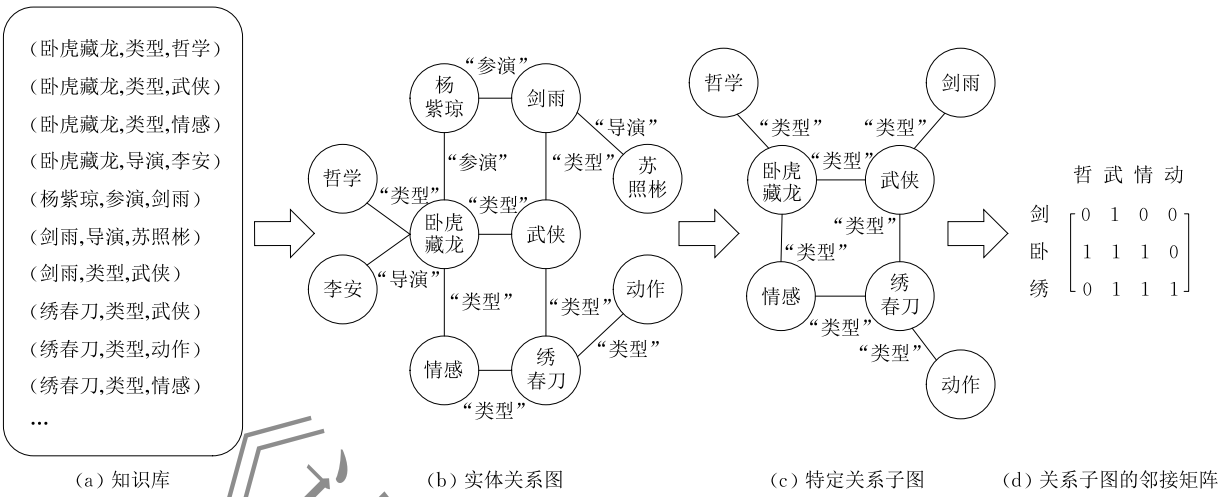
(a) 1:M类型 (b) M:M类型

图 2 两种典型的关系类型示例

通过以上分析,提出 TRWA 算法的第 2 项假设:某些类型的关系本身具有良好的可推理性,关系两侧的实体间具有关系传递性.通过对此类关系构成的局部子图进行推理建模,可以改善 PRA 模型对于一对多(1∶M)和多对多(M∶M)类型关系的推理能力.

基于以上讨论,提出 TRWA 算法的设计思路如下:首先,将知识库中所有的关系三元组表示为一

张无向图 $\boldsymbol{G}$，图中节点为知识库中的实体对象，边为实体间的各种关系对象，称为全局实体关系图（简称全局图，如图 3(b)所示），并基于 PRA 算法采用的随机游走策略对其进行特征提取和关系推理建模.



(a) 知识库　(b) 实体关系图　(c) 特定关系子图　(d) 关系子图的邻接矩阵

图 3　TRWA 算法的数据建模示意图

然后，逐一提取每种关系的子图 $\boldsymbol{G}_i$（如图 3(c)所示的“类型”关系），并根据 $\boldsymbol{G}_i$ 中节点的拓扑关系构造邻接矩阵 $\boldsymbol{A}_i$（如图 3(d)所示）. $\boldsymbol{A}_i$ 的行对应于 $\boldsymbol{G}_i$ 中的 *head* 节点，列对应于 $\boldsymbol{G}_i$ 中的 *tail* 节点. 对 $\boldsymbol{G}_i$ 采用 k 步随机游走算法，可以得到 $\boldsymbol{G}_i$ 中 *head* 节点到 *tail* 节点的 k 阶转移概率，以此作为评判节点间是否存在关系 r_i 的辅助判据.

最后，对通过上述两种关系推理模型得到的结果进行融合，得到最终的关系推理结果.

综上，TRWA 算法由 3 个子算法模块组成，分别是：全局关系推理算法模块、局部关系推理算法模块和推理结果融合模块（算法流程如图 4 所示）. 接下来分别对这 3 个模块所采用的算法进行介绍和分析.

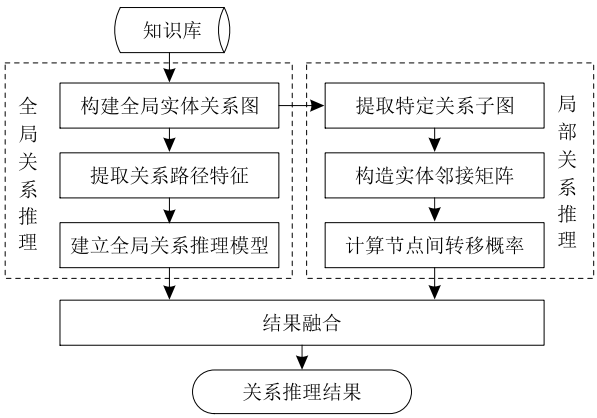


图 4　TRWA 算法流程图

3.2　全局关系推理算法

TRWA 算法采用的全局关系推理算法与 PRA 算法类似. 二者的主要区别在于 TRWA 算法在全局关系推理时采用无向图假设，即认为关系具有自反性，因此二者得到的路径特征模式有较大区别. 另一个区别在于 TRWA 所采用的全局随机游走策略比 PRA 更为自由（仅限制随机漫步者每一步均不能在原地停留，且设定最大跳数为 3）. 算法实现细节描述如下.

TRWA 的全局算法是一个有监督学习算法，其基本设计思想是：针对知识图谱中包含的每种关系，采用随机游走策略在全局图上提取关系的路径特征，然后利用得到的路径特征建立每种关系的推理模型，并利用已知的关系数据对模型进行训练得到特征参数.

以关系 $r_i \in \boldsymbol{R}$ 的建模为例，首先采集知识库中所有包含关系 r_i 的关系三元组构造实体对集合 $\boldsymbol{P}_i$，$\boldsymbol{P}_i$ 中的元素为实体对 (h_j, t_j)，表示实体 h_j 和 t_j 之间存在关系 r_i. 对于 $\boldsymbol{G}_i$ 中的每个 *head* 实体 h_j，从 $\boldsymbol{G}_i$ 中选出所有以 h_j 为 *head* 节点的关系三元组，以其中的 *tail* 节点构成集合 $\boldsymbol{T}_j$.

然后，采用随机游走的方式在全局图 $\boldsymbol{G}$ 上搜索识别关系 r_i 的路径特征模式. 方法是对于 $\boldsymbol{P}_i$ 中的每个实体对，派遣随机漫步者从实体 h_j 出发沿着关系构成的路径进行漫游，若该随机漫步者在 3 跳之内落到 $\boldsymbol{T}_j$ 中的任意实体节点上，则认为识别出一个可行的关系路径特征模式（记为 π），将其记录到关系 r_i 的路径特征集合中. 在随机游走过程结束后，以路径特征集合中的元素为特征构造关于关系 r_i 的特征向量 $\boldsymbol{\Pi}_i$，以符号 $|\boldsymbol{\Pi}_i|$ 表示 $\boldsymbol{\Pi}_i$ 的元素总数，则元素 $\pi_k \in \boldsymbol{\Pi}_i (k=1,2,\cdots,|\boldsymbol{\Pi}_i|)$ 表示在图 $\boldsymbol{G}$ 上出现的某种关于关系 r_i 的关系路径模式.

据此,可以如下所示基于图 $\boldsymbol{G}$ 以及特征向量集合 $\boldsymbol{\Pi}_i$ 建立关系 r_i 的推理模型.对于给定的实体对 (h,t),以符号 $\boldsymbol{x}$ 表示 (h,t) 关于 $\boldsymbol{\Pi}_i$ 的特征值向量,$x_k \in \boldsymbol{x}$ 表示从图 $\boldsymbol{G}$ 上的节点 h 出发,经过关系路径 π_k 到达节点 t 的概率.注意到 π_k 是关系路径,因此对于给定的实体对 (h,t),满足条件 π_k 的实体节点构成的实际路径可能有多条,所以将 x_k 定义为所有满足条件的节点路径的概率值之和.其中,每条节点路径的概率值定义为路径中经过的实体节点的度的乘积的倒数.设路径特征向量 $\boldsymbol{\Pi}_i$ 对应的权重向量为 $\boldsymbol{\theta}^i \in \mathbb{R}^k$($\mathbb{R}$ 表示实数域),以符号 $f(h,r_i,t)$ 表示实体对 (h,t) 之间存在关系 r_i 的可能性,全局关系推理的数学模型如式(8)所示:

$$f(h,r_i,t)=\boldsymbol{x}^{\mathrm{T}}\theta^i=\sum_{k=1}^{|\Pi|}x_k\theta_k^i \tag{8}$$

本文采用逻辑斯蒂回归算法对模型参数 θ^i 进行估计,首先利用 sigmoid 函数将模型预测结果映射到 $(0,1]$ 区间,以 0.5 作为决策边界,如式(9)所示:

$$p=1/(1+\mathrm{e}^{-x^{\mathrm{T}}\theta^i}) \tag{9}$$

设训练样本的容量为 N,$y_s=\{0,1\}$ 表示样本 $\boldsymbol{x}_s$ 的属性值,$y_s=1$ 表示 $\boldsymbol{x}_s$ 对应的实体对 (h_s,t_s) 之间存在关系 r_i,否则 $y_s=0$.则模型(8)在训练集上预测结果正确程度的对数似然函数 $L(x,\theta)$ 可表示为

$$L(x,\theta)=\sum_{s=0}^{N}(y_s\ln p_s+(1-y_s)\ln(1-p_s)) \tag{10}$$

由此可以定义优化目标函数为

$$\arg\max_{\theta}(L(x)-\lambda_1\|\theta\|_1-\lambda_2\|\theta\|_2) \tag{11}$$

其中:$\|\theta\|_1$ 和 $\|\theta\|_2$ 分别表示对模型复杂度的 L1-范数和 L2-范数约束条件,其中,L1-范数约束用于实现对路径模式特征的特征选择;L2-范数约束用于控制模型参数 θ 的向量长度.综合采用这两种模型复杂度约束条件的目的是为了减少模型的过拟合现象[19].$\lambda_1 \geq 0$ 和 $\lambda_2 \geq 0$ 为相应的模型复杂度惩罚因子,用于控制 L1-范数和 L2-范数对模型的影响程度.

在构造训练样本时,采用知识库中已有知识作为正样本,负样本的构造则采用封闭世界假设[1],即针对给定关系 r_i 的正样本 (h,t),采用随机替换头尾实体的方式构造候选负样本集合,去除其中已知的正样本,得到关于关系 r_i 的一组负样本集合.

3.3 局部关系推理算法

局部关系推理针对知识图谱中给定关系 r_i 所对应的子图 $\boldsymbol{G}_i$(如图 3(c)所示),推理的基本思路是借助 $\boldsymbol{G}_i$ 上的随机游走过程,估计图中实体对之间的转移概率,据此评估实体对之间存在关系 r_i 的可能性.

考虑到关系在语法上的有向性,建模时将 $\boldsymbol{G}_i$ 视为二部图,上层节点为 $\boldsymbol{G}_i$ 中的 *head* 实体,下层为 *tail* 实体.在计算节点对之间的转移概率时,不考虑同层节点间的组合情况,即仅采用奇数跳数的随机漫步策略进行节点间的转移概率估计.

由于 $\boldsymbol{G}_i$ 的规模通常远小于全局图 $\boldsymbol{G}$ 的规模,因此可以采用邻接矩阵相乘的方法直接计算转移概率.为此,首先按照图 3(d)所示的方式构造图 $\boldsymbol{G}_i$ 的邻接矩阵 $\boldsymbol{A}_i$,其中每一行对应于二部图上层的一个 *head* 实体,每一列对应于二部图下层的一个 *tail* 实体.对角矩阵 $\boldsymbol{D}_h$ 的维度与 $\boldsymbol{A}_i$ 的行数一致,对角元素为与 $\boldsymbol{A}_i$ 的行相对应的 *head* 实体在 $\boldsymbol{G}_i$ 中的度.类似地可以定义 *tail* 实体的度对角矩阵 $\boldsymbol{D}_t$.计算图 $\boldsymbol{G}_i$ 中上下层节点间的 K 步转移概率矩阵 $\boldsymbol{Q}_i$ 的公式如下:

$$\boldsymbol{Q}_i=(\boldsymbol{D}_h^{-1}\boldsymbol{A}_i\boldsymbol{D}_t^{-1}\boldsymbol{A}_i^{\mathrm{T}})^{\frac{K-1}{2}}(\boldsymbol{D}_h^{-1}\boldsymbol{A}_i) \tag{12}$$

考虑到计算开销,并且与 PRA 算法的全局随机游走步长保持一致,本文步长取 $K=3$.相应地,$\boldsymbol{Q}_i$ 中的元素 $\boldsymbol{Q}_i[a,b]$ 表示随机漫步者从图 $\boldsymbol{G}_i$ 中的节点 h_a 出发,经过 3 步转移之后,到达节点 t_b 的概率.以符号 $g(h_a,r_i,t_b)$ 表示实体对 (h_a,t_b) 之间存在关系 r_i 的可能性,局部关系推理的数学模型如式(13)所示:

$$g(h_a,r_i,t_b)=\boldsymbol{Q}_i[a,b] \tag{13}$$

3.4 推理结果融合算法

经过以上两个建模过程,在知识图谱 $\boldsymbol{G}$ 上对于任意给定的关系 r 和节点对 (h,t),可以依据式(8)和式(13)分别求出 $f(h,r,t)$ 和 $g(h,r,t)$ 的值.根据式(8)可知,$f(h,r,t)$ 是一组概率值的线性组合,因此 $f(h,r,t)$ 和 $g(h,r,t)$ 的值具有可加性,可以采用如下公式对上述两类关系推理结果进行融合:

$$score(h,r,t)=f(h,r,t)+\alpha\cdot g(h,r,t) \tag{14}$$

其中:$score(h,r,t)$ 为 TRWA 算法对于三元组 (h,r,t) 给出的推理评分,$score(h,r,t)$ 的值越大,表明在实体 h 和 t 之间存在关系 r 的可能性越大.$\alpha \geq 0$ 为局部推理结果 $g(h,r,t)$ 在整个评分体系中的权重.

然而,采用上述方式进行结果融合不利于定量评估局部关系推理算法对于算法整体性能的影响,且 α 值的可解释性不好.因此本文将全局推理的结果视为一个整体,将其映射到 $g(h,r,t)$ 的取值范围 $[0,1]$ 区间上,然后再对二者进行融合,计算公式如下:

$$score(h,r,t)=(1+\mathrm{e}^{-f(h,r,t)})^{-1}+\alpha\cdot g(h,r,t) \tag{15}$$

其中 $\alpha \geq 0$ 仍然为权重因子,但其含义更为明确,$\alpha/(1+\alpha)$ 表示局部关系推理结果在整个推理模型中的相对重要性.此外,透过 α 值的变化对算法性能的

影响，有助于更好地观察和分析实验结果. 对 α 的取值将在本文的实验部分进行详细讨论.

3.5　算法复杂度分析

由式(15)可以看出，TRWA 算法的时间复杂度由两部分组成：一是计算 $f(h,r,t)$ 所需的时间开销；二是计算 $g(h,r,t)$ 所需的时间开销. 设知识库中的实体总数为 n，关系总数为 m，三元组总数为 e.

为求解全局关系推理模型 $f(h,r,t)$，由于 TRWA 采用了受限的随机游走策略(限制随机漫步者的步长为 3)，根据 3.2 节给出的算法流程可知，在模型特征发现阶段，对全部三元组进行路径特征采样的计算复杂度为 $\boldsymbol{O}(e\times n)$，后续的模型训练阶段的计算复杂度为 $\boldsymbol{O}(e)$，因此求解 $f(h,r,t)$ 的复杂度为 $\boldsymbol{O}(e\times n)$.

为求解局部关系推理模型 $g(h,r,t)$，需要对每种关系构成的局部子图进行矩阵乘法，由于 $\boldsymbol{D}_h$ 为对角矩阵，且限制随机游走步长 $K=3$，设最大的局部子图中包含的实体个数为 $p\times n(0\leqslant p\leqslant 1)$，因此对 m 种关系求 $g(h,r,t)$ 的计算复杂度上界为 $\boldsymbol{O}(m\times p^3\times n^3)$.

综上，TRWA 算法的计算复杂度上界为 $\boldsymbol{O}(e\times n+m\times p^3\times n^3)$，从理论上看，算法复杂度取决于 $g(h,r,t)$ 的计算复杂度. 然而在实际应用中，由于关系子图通常十分稀疏，因此计算开销主要由该关系对应的事实数量决定，本文对算法复杂度上界值的估计偏高. 实验过程中我们发现 TRWA 算法对上述两个部分的计算时间开销基本相当. 在下一步工作中，我们将进一步研究 TRWA 算法的计算复杂度下界，以进一步提高算法计算效率. 在空间复杂度方面，由于算法需对全局知识图谱执行随机游走操作，因此需将整张图 G 读入内存，算法空间复杂度为 $\boldsymbol{O}(n^2)$.

4　实验结果与分析

4.1　实验数据

为验证基于双重随机游走的关系推理算法的有效性，本文采用关系推理研究领域近期研究工作中普遍采用的 3 个公开数据集进行测试，学术界将其分别称为 WN18、FB15k 和 FB40k 数据集. WN18 数据集是 Wordnet 知识库的子集，其中包含 18 种关系(例如词汇的上下位关系、整体部分关系等). FB15k 和 FB40k 数据集是 Freebase 知识库的子集，其中 FB15k 数据集包含 1345 种关系，FB40k 数据集包含 1318 种关系，但其中所包含的实体数约为 FB15k 的 2.5 倍.

WN18[①] 和 FB15k[①] 数据集均由发布者分割为 3 个部分，分别用于模型训练、参数验证和性能评估. 由于 WN18 包含的关系种类较少，因此本文对于 4 种关系类型对算法性能影响的讨论主要围绕 FB15k 数据集展开. FB15k 的数据来源于真实的 Freebase 知识库，其中的知识范围覆盖现实世界中的多个方面，关系类型的分布以及实体关系三元组之间的关联方式均接近真实知识库的情况. FB40k[②] 数据集同样采集自 Freebase 知识库，与 FB15k 的区别在于它是由不同的学者使用不同的采样策略构造的，因此二者不仅内容完全不同，而且数据呈现出不同的统计特征. 本文将其随机划分为训练集、验证集和测试集这 3 个部分，该数据集主要用于验证算法的有效性和可扩展性. 以上 3 种测试数据集的统计信息如表 2 所示.

表 2　基准数据集的统计信息

数据集名称	实体总数	关系种类个数	训练集三元组数	验证集三元组数	测试集三元组数
WN18	40943	18	141442	5000	5000
FB15k	14951	1345	483142	50000	59071
FB40k	37561	1318	235350	50000	50000

经统计，FB15k 数据集的 1345 种关系中，平均每种关系拥有的三元组数量均值约为 440 个，中位数为 21 个，表明大量的关系仅出现于少数三元组中，而少量关系则拥有大量的三元组，三元组个数关于关系的分布呈幂律分布. 按照本文第 3 节式(6)介绍的分类方法对 FB15k 数据集中的关系进行划分(取 $\delta=1.5$)，得到 FB15k 数据集的关系类型划分情况如表 3 所示. 从表 3 可见，FB15k 数据集的 1345 种关系平均分布在 4 种类型中. 表 4 进一步给出了每种关系类型所包含的三元组数量，可以看出在 FB15k 数据集中，一对多和多对多关系的三元组在整个知识库中合计占比接近 90%，因此关系推理算法在这两类关系上的性能表现，对算法的整体性能影响极大.

表 3　FB15k 数据集的关系类型分布

关系类型	训练集/%	测试集/%
1∶1	27.36	25.50
1∶M	22.97	23.27
M∶1	29.29	28.92
M∶M	20.38	22.31

① https://everest.hds.utc.fr/doku.php?id=en:transe
② https://github.com/Mrlyk423/Relation_Extraction

表 4 FB15k 数据集的 4 种关系类型三元组数量分布

关系类型	训练集/%	测试集/%
1∶1	1.57	1.51
1∶ M	15.88	15.26
M∶1	9.48	9.49
M∶M	73.07	73.74

表 5 和表 6 给出了 FB40k 数据集中 1318 种关系的类型分布情况，从中可以看出该数据集的关系类型分布与 FB15k 数据集存在较大差异. 首先，FB40k 的关系类型分布不如 FB15k 均衡(表 5)，表明这两个数据集所采用的采样策略有本质区别，对两者分别进行实验，有助于全面评估算法在真实的 Freebase 知识库上的实际性能；其次，FB40k 中 4 种类型的关系所拥有的实例(二元组)数量的分布与 FB15k 相比有显著差异，突出表现在该数据集拥有大量的一对多和多对一关系实例，通过比较算法在这两个数据集上的性能表现，可以更完整地揭示出算法的特点.

表 5 FB40k 数据集的关系类型分布

关系类型	训练集/%	测试集/%
1∶1	56.22	56.22
1∶M	20.03	20.03
M∶1	16.31	16.31
M∶M	7.44	7.44

表 6 FB40k 数据集的 4 种关系类型三元组数量分布

关系类型	训练集/%	测试集/%
1∶1	6.46	6.45
1∶M	33.83	32.06
M∶1	39.72	39.85
M∶M	7.44	7.44

4.2 实验方法与评价指标

为客观评价 TRWA 算法的性能，本文选取近期发表的相关工作进行实验比较，包括随机游走模型中的代表性模型 PRA 算法[13]和另外 3 个在 WN18 和 FB15k 数据集上性能表现突出的潜在因子模型，分别为：Nickel 等人[12]在 2011 年提出的 RESCAL 算法、Bordes 等人[16]在 2013 年提出的 TransE 算法以及 Lin 等人[18]在 2015 年提出的 TransR 算法.

对算法性能的测试采用通行的干扰列表法. 方法是对于测试集中给出的每个实体关系三元组 (h,r,t) 构造一个干扰列表 L，列表中的元素为一对实体，但是除给定的 (h,t) 之外，其余实体对之间的关系 r 并不成立，因此称为干扰列表. 算法测试的目标就是对包括 (h,t) 在内的干扰列表中的每个实体对计算出一个置信度，然后据此对表中元素进行降序排序，通过查看正确的实体对 (h,t) 在干扰列表中出现的位置来评估算法的性能. 干扰列表可以采用多种方式进行构造，本文使用的方式与参与比较的相关工作保持一致，即对于给定的三元组 (h,r,t)，枚举知识库中的实体，分别置换三元组中的 *head* 和 *tail* 实体，得到两个备选的三元组集合，去掉其中曾经出现在训练集和验证集中的部分，得到两个干扰列表(一组的 *head* 实体为 h，*tail* 实体为干扰项；另一组的 *tail* 实体为 t，*head* 实体为干扰项). 当推理算法完成对干扰列表的排序后，取正确的 (h,t) 在两张排序表中的位置求平均，作为算法对该三元组的预测结果(排名).

本文采用的算法性能评价指标包括：平均倒数排序指标(Mean Reciprocal Rank，*MRR*)和前 N 命中率指标($hit@N$)，这两个指标也是相关研究工作用于算法性能评估的首选指标[15,21].

以符号 $\boldsymbol{C}$ 表示测试集，$|\boldsymbol{C}|$ 表示测试集中的实体关系三元组个数，$rank(c)$ 表示三元组 c 在相应的干扰列表中的排名，*MRR* 指标的计算公式定义为

$$MRR=\frac{1}{|\boldsymbol{C}|}\sum_{c\in C}\frac{1}{rank(c)} \tag{16}$$

$hit@N$ 指标($N\in Z^+$)的计算公式定义为

$$hit@N=\frac{100}{|\boldsymbol{C}|}\sum_{c\in C}(Ind(rank(c)\leq N))\% \tag{17}$$

其中 $Ind(\cdot)$ 为指示函数，定义式如下：

$$Ind(rank(c)\leq N)=\begin{cases}1, & 当:rank(c)\leq N\\ 0, & 当:rank(c)>N\end{cases} \tag{18}$$

MRR 指标和 $hit@N$ 指标的取值范围均为[0,1]. *MRR* 指标值反映的是关系推理算法在给定测试集上对所有测试样本给出的排序值的倒数的均值，*MRR* 值越大，说明正确结果在关系推理结果列表中出现的平均位置越靠前，表明相应的关系推理算法对该数据集的推理结果适用性越好. $hit@N$ 指标的含义是对测试集中的每个测试样本，取排序后的干扰列表中的前 N 个元素，查看该测试样本是否命中. 对全体测试样本的命中情况求均值，就得到 $hit@N$ 指标值. 显然，$hit@N$ 指标值越大，说明算法在采用前 N 个元素进行结果推荐时的关系推理召回率越高. 通常在实际应用中比较关注的是 $hit@1$ 和 $hit@10$ 指标，原因是 $hit@1$ 值在一定程度上反映出推理算法的准确性，即在多大程度上可以接受第一个推理结果；而 $hit@10$ 值则反映了在可接受的推荐数量范围内关系推理算法的召回率，即算法的有效性(在 N 选一的情况下，通常人们可接受的推荐列表长度不超过 10).

4.3　算法性能综合测评

TRWA 算法包含 4 个实验参数，分别为模型复杂度惩罚因子 λ_1 和 λ_2，关系推理时的随机游走步长限制因子 K 以及推理结果融合算法中的权重因子 α.

通过在 WN18、FB15k 和 FB40k 等数据集的验证集上进行参数选择实验，我们发现当 λ_1 和 λ_2 的取值范围在(0.0005，0.002)区间变化时，TRWA 算法的性能表现较好且受这两个参数变化的影响不大. 由于 PRA 及其后续工作也大多提供 WN18 和 FB15k 数据集上的实验结果，为便于参照和比较本文实验与相关工作所报道的实验结果，本文选择与之相同的实验参数设置，即 $\lambda_1=\lambda_2=0.001$.

同时，限制随机游走的步长为 $K=3$，一方面与 PRA 算法保持一致，另一方面也符合逻辑上的直觉. 以图 3(c)给出的电影类型关系为例，设某个随机漫步者从某部电影出发沿关系路径进行漫游，经一跳后到达一种特定“类型”，两跳后回到电影名称的集合中，3 跳后再次到达“类型”集合，由此可以得到一个有推理价值的关系路径模式“$r_i \rightarrow r_i^{-1} \rightarrow r_i$”.

推理结果融合算法中的参数 α 用于控制和调节全局推理与局部推理结果对最终推理结果的贡献，本文的实验取 $\alpha=0.5$. 由于该参数对算法性能影响较大，在 4.4 节将专门对 α 的取值结合实验进行讨论.

表 7 和表 8 给出了 TRWA 算法在 WN18 和 FB15k 两组数据集上与相关工作的性能对比情况. 其中，PRA 算法是随机游走模型的代表性工作[13]，RESCAL、TransE 和 TransR 等 3 个算法则代表了近期潜在因子模型的前沿水平[12,16,18]. 为了更好地评估本文提出的两项假设的有效性，采用 TRWA 算法的全局关系推理子模块作为 Baseline，以模拟 PRA 算法在无向图假设下的实验性能.

表 7　WN18 数据集上的实验结果

算法/指标	*MRR*	*hit*@1	*hit*@10
TRWA	**0.691**	**79.1**	90.8
Baseline	0.667	65.4	67.9
PRA	0.458	42.2	48.1
Rescal	0.431	10.2	52.8
TransE	0.495	11.3	89.2
TransR	0.605	33.5	**91.7**

表 8　FB15k 数据集上的实验结果

算法/指标	*MRR*	*hit*@1	*hit*@10
TRWA	**0.603**	**54.7**	70.3
Baseline	0.515	49.7	54.3
PRA	0.336	30.3	39.2
Rescal	0.354	23.5	44.1
TransE	0.463	29.7	**73.4**
TransR	0.346	21.8	65.5

为进一步验证 TRWA 算法提出的两个基本假设的合理性，同时客观评估算法的实际性能，表 9 给出了 TRWA 算法在 FB40k 数据集上与相关工作的性能对比情况. 根据相关工作在 FB15k 数据集上的表现，选择在 Freebase 知识库上性能表现最好的 TransE 算法作为性能比较的参照对象，同时选择 PRA 和 Baseline 作为验证模型基本假设的参照对象.

表 9　FB40k 数据集上的实验结果

算法\指标	*MRR*	*hit*@1	*hit*@10
TRWA	**0.539**	**51.6**	58.1
Baseline	0.514	49.9	54.0
PRA	0.362	33.8	40.2
TransE	0.514	40.65	**70.4**

首先观察 TRWA 算法与 4 种相关算法间的性能对比情况. 与同样采用随机游走机制相比，能够有效提高随机游走模型的关系推理性能，与 PRA 相比，TRWA 算法给出的正确结果排名更为靠前，算法的适用性更强. 从 *hit*@1 指标来看，TRWA 算法对测试集中的每个三元组给出的第 1 项推理结果的正确性在 WN18 测试集上高达 79.1%，在 FB15k 和 FB40k 测试集上也达到了 54.7%和 51.6%，远高于相关算法，这意味着用户能够分别以 79.1%，54.7%和 51.6%的信心接受 TRWA 算法的第 1 项推理结果. 与之相比，PRA 算法给出的第 1 项的推理结果的准确性相对较差. *hit*@10 指标反映了算法的召回率，从数据对比情况看，TRWA 算法在排名前 10 的推理结果中命中给定事实的概率比 PRA 算法有大幅提高，特别是该算法在接近真实知识库情况的 FB15k 数据集上取得了高达 70%以上的 *hit*@10 命中率，表明 TRWA 算法具有良好的实际应用前景.

观察表 7、表 8 和表 9 中 3 种潜在因子模型的性能表现，可以看出 TransR 算法在语法型知识库(WN18)上的性能表现较好，而 TransE 算法在 FB15k，FB40k 这样的事实型知识库上的性能表现较好. 与 TRWA 算法相比，二者在 *hit*@10 指标上的表现略优于前者(差异小于 5%，4.4 节将做进一步讨论)，但在 *MRR* 和 *hit*@1 等指标上，TRWA 算法则显著占优. 该结果表明，TRWA 算法的召回率与性能最好的潜在因子算法相当，但推理结果的准确性更高.

接下来观察 TRWA 与 PRA 和 Baseline 这 3 种算法的性能差异. 实验表明 Baseline 算法在 3 组测试集上的 3 个性能指标值均显著优于 PRA 算法，在

WN18 测试集上 3 个指标分别提高了 45.63%、54.98%和 41.16%;在 FB15k 测试集上 3 个指标分别提高了 53.27%、64.03%和 38.52%;在 FB40k 测试集上 3 个指标分别提高了 15.2%、47.63%和 34.33%.

同时,TRWA 算法在 WN18 和 FB15k 两个数据集上的综合性能显著优于 Baseline 算法,其中,在 WN18 测试集上 3 个指标分别提高了 3.60%、20.95%和 33.73%,在 FB15k 测试集上 3 个指标分别提高了 17.09%、10.06%和 29.47%.由于 Baseline 算法可以看作是 PRA 算法的无向图版本,因此可以得出如下结论:采用无向图假设能够有效提高 PRA 算法的性能,但仅有该假设是不够的,通过引入局部关系推理机制对全局推理结果进行修正,可以再度有效地提高算法性能.注意到 Baseline 算法与潜在因子模型相比,前者的 *MRR* 和 *hit*@1 等指标显著占优,但后者在 *hit*@10 指标上的表现则显著优于前者.这说明全局的随机漫步模型对于跨关系的推理建模更为适用,而潜在因子模型则能够更好地处理同种关系的推理建模.

此外,注意到在 FB40k 测试集上,TRWA 算法相对于 Baseline 算法的性能优势并不明显,3 个指标分别提高了 4.86%、3.41%和 7.59%,仅在 *hit*@10 指标上取得显著优势.然而与 PRA 算法相比,TRWA 算法的优势十分明显,3 个指标分别提高了 48.89%、52.67%和 44.53%,该结果进一步验证了本文提出的关系的语义双向性假设的合理性.

4.4 局部关系推理机制的影响分析

通过 4.3 节对 Baseline 算法性能的讨论,确认了本文所采用的无向图假设的有效性.接下来重点讨论 TRWA 算法中局部关系推理算法模块对整体性能的影响,以验证本文提出的第 2 项假设对于关系推理任务的有效性.由于 WN18 数据包含的关系数量较少,且 FB15k 的数据分布情况更接近于真实知识库,本节选用 FB15k 数据集作为测试对象.

首先考查推理结果融合算法中的权重因子 α 对算法性能的影响.图 5 给出了 α 取值在[0,12]区间变化时(α 变化的步长为 0.05),TRWA 算法在 FB15k 测试集上的 3 种指标变化情况.图中虚线为 Baseline 算法的性能指标,相当于 TRWA 算法中 $\alpha=0$ 的情况.可以看出图中的 3 条曲线均在 $\alpha=0.5$ 处出现拐点,相应的 *MRR*、*hit*@1 和 *hit*@10 的值与 Baseline 相比分别提高了 17.09%、10.06%和 29.47%,表明 TRWA 算法在引入局部关系推理模块后性能有了显著提升,对正确推理结果在干扰列表中的排名有明显的提升效应,说明测试集中有大量的实体关系是可以通过同种关系两侧的实体间的 K 步转移概率辅助推理得到的.

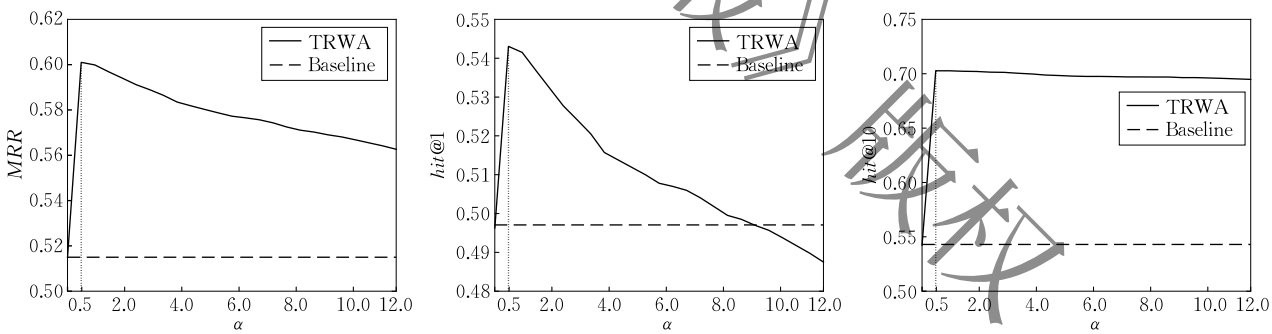


图 5 TRWA 算法在 FB15k 测试集上的 3 种性能指标随 α 的变化情况

3 种指标均在 $\alpha=0.5$ 时达到峰值后随 α 的增加而单调递减.其中,*hit*@1 值的递减最为显著,当 $\alpha>10.7$ 时,TRWA 算法的 *hit*@1 指标甚至低于 Baseline 算法.相应地,*MRR* 值也呈显著递减趋势.该结果表明,当局部推理结果的权重超过合理范围时,会导致算法整体性能下降,导致正确结果在干扰列表中的平均排名下降.同时注意到 *hit*@10 的比例高达 70%且递减效应不明显,这看似与 *MRR* 值的快速递减形成"矛盾",表明当 α 增大时,在其余约 30%的干扰列表中,正确结果的排名发生了较大幅度的后退.

为进一步了解权重因子 α 的取值对算法性能的影响,以揭示出算法性能改进的深层次原因,本文对 FB15k 测试集按 4 种关系类型进行划分后,分别进行了关系推理实验.论文中使用的 3 个性能指标随 α 取值的变化情况如图 6 所示.

首先,TRWA 算法在全部 4 种关系类型上的 *MRR* 指标值与 *hit*@1 指标的值高度正相关,并且,两者在到达峰值后的下降趋势比较明显.这一现象表明,算法 TRWA 的推理准确率同时依赖全局推理和局

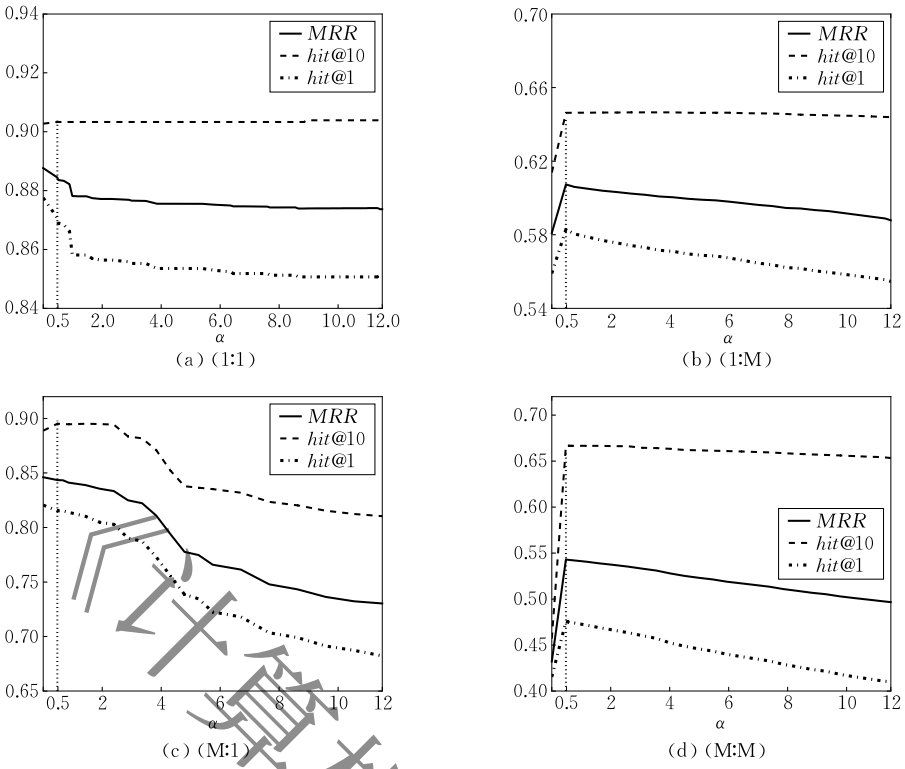


图 6　TRWA 算法在 FB15k 测试集上 4 种关系类型上 3 种性能指标随 α 的变化情况

部推理两个部分的推理结果. 然而，*hit*@10 指标的值在全部 4 种关系类型上，相对于 *MRR* 和 *hit*@1 而言，对权重参数 α 到达拐点后的取值变化并不敏感(多对一(M：1)类型关系在 $\alpha>3.5$ 之后有一个较为明显的下降趋势). *hit*@10 指标的稳定性表明，算法 TRWA 两个推理过程(全局推理和局部推理)的推理结果很大程度上是彼此一致的，而 *MRR* 和 *hit*@1 指标值(以及 *hit*@10 在多对一(M：1)类型关系上的值)的变化反映出，全局推理与局部推理两个推理过程给出的结果列表仍会存在一些差异，表明这两个推理过程从不同的角度揭示了知识库中实体间关系的一般模式，可以通过对 α 值的调节取得更好的最终推理结果. 该实验结果为支持 TRWA 算法的第 2 个假设提供了有力的实验证据，即利用特定关系的连通子结构能够提高关系推理结果的准确率.

其次，图 6(b)和图 6(d)中的实验曲线在 $\alpha\geq 0.5$ 时呈明显上升趋势，表明增加局部推理过程的推理结果，对于一对多和多对多关系类型的推理准确率有较大幅度的提升. 同时可以得出结论，全局随机游走推理模型未能充分利用特定关系子图中的有效信息. 图 6(a)和图 6(c)中 *MRR* 和 *hit*@1 指标值的变化表明，加入局部推理过程的推理结果会小幅度降低 TRWA 算法在一对一(1：1)和多对一(M：1)关系类型上的推理准确率，对这一现象较为合理的解释是，“$r_i\geq r_i^{-1}\geq r_i$”形式的一阶 Horn 子句推理规则并不适用于一对一和多对一关系类型的推理.

第三，我们发现 TRWA 算法在一对一(1：1)和一对多(1：M)关系类型上的 *MRR* 和 *hit*@1 指标的值相对于其它两种关系类型对 α 的取值更不敏感，表明 TRWA 算法的全局推理和局部推理两个过程在一对一(1：1)和一对多(1：M)关系类型上的推理结果具有相对的一致性，而其他两种关系类型则不然. 因此我们认为，随机游走模型在一对一(1：1)和一对多(1：M)关系类型上的预测能力主要来自于特定关系子图中的内部结构，该结构能够利用局部推理算法进行有效建模. 然而，局部推理算法忽略了特定关系子图之间的联系，不能找到除“$r_i\geq r_i^{-1}\geq r_i$”类型之外的推理规则，因此，全局推理算法对 TRWA 算法最终的推理过程仍然是不可或缺的.

最后，从图 6(b)和图 6(c)中可以看出，算法 TRWA 在一对多和多对一类型关系上的推理性能随 α 取值的增加，呈现出不同的下降趋势，看起来关系的方向性对于推理的结果产生了影响，这似乎与 TRWA 算法的第 1 个假设(即语义双向性假设)相互矛盾. 为了进一步验证该假设的有效性性，同时为了

深入理解随机游走模型推理能力产生的原因，我们将 FB15k 测试集中的 4 类关系分别与相关工作进行对比实验. 由于 TRWA 算法在 *MRR* 和 *hit*@1 指标上显著地优于其它相关工作，这里仅以 *hit*@10 指标为例对实验结果进行讨论，实验结果如表 10 中所示.

表 10 FB15k 数据集上 *hit*@10 的实验结果

算法/关系类型	TRWA	Baseline	PRA	TransE
1：1	**89.5**	89.5	63.3	71.5
1： M	**60.5**	60.0	20.4	49.0
M：1	**92.3**	91.4	81.4	85.0
M：M	70.6	45.8	32.6	**72.9**

从表 10 中可见，采用无向图假设的 Baseline 算法的性能较采用有向图假设的 PRA 算法在一对一、一对多和多对一等关系类型上的 *hit*@10 性能有较大提升，进一步增加局部推理机制后，性能改进不明显. 这表明 TRWA 算法在这 3 类关系上相对于 PRA 算法的性能提升主要是由无向图假设带来的. 特别是在一对多类型的关系上，采用无向图进行推理导致的性能提升高达近 200%，进一步验证了本文提出的实体间关系在语义上具有自反性的断言.

通过观察 TRWA 算法在多对多关系类型上的性能表现，可以看到引入无向图假设后的 Baseline 算法的 *hit*@10 值比 PRA 算法高 40.49%(Baseline vs. PRA)，进一步引入局部关系推理机制后，TRWA 算法的 *hit*@10 值比 Baseline 算法再度提高了 54.15%(TRWA vs. Baseline)，该实验结果可以有效证明 TRWA 算法第 2 个假设的有效性，即对于某些关系类型(多对多类型)而言，关系内部包含有价值的可推理信息，采用双层随机游走机制可以充分发掘并利用知识库的全局拓扑结构和同种关系内部的拓扑关系中包含的信息，得到更好的关系推理模型.

最后，对表 10 中 TRWA 和 TransE 算法的性能指标进行对比，可以看出 TRWA 算法在前 3 种关系类型上的 *hit*@10 指标显著优于基于潜在因子模型的 TransE 算法，而在最后一项多对多类型关系上的表现略逊于后者(性能相差 3.26%). 该结果表明，在处理关系密集型知识库的推理任务时，两种模型的性能相当，但在处理关系稀疏型数据的推理任务时，基于网络拓扑的随机游走模型具有优势.

5 结束语

本文研究了知识图谱上的关系推理问题，发现了现有随机游走方法模型的基本假设存在的问题，提出了两项新的假设：关系的语义双向性假设和同种关系两侧的实体间具有关系可推理性假设，并据此提出了一种基于双层随机游走机制的关系推理算法(TRWA 算法). 实验表明，TRWA 算法在 WN18、FB15k、FB40k 等公开数据集上的性能表现一致且显著地优于相关工作，实验证据有力地支持了本文提出的两项假设.

该项研究工作为研究和利用随机游走模型实现大规模知识推理提供了新的思路和解决方案，研究成果可应用于知识库扩容和推荐系统等相关领域，具有良好的实用性和可推广性. 在后续的工作中，将进一步深入研究不同的全局关系路径特征模式定义方法和随机游走策略对算法性能的影响，并尝试设计更为高效低耗的算法解决方案.

参 考 文 献

[1] Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs: From multi-relational link prediction to automated knowledge graph construction//Proceedings of the IEEE, 2016, 104(1): 11-33

[2] Wang Yuan-Zhuo, Jia Yan-Tao, Liu Da-Wei, et al. Open web knowledge aided information search and data mining. Journal of Computer Research and Development, 2015, 52(2): 456-474(in Chinese)
(王元卓，贾岩涛，刘大伟等. 基于开放网络知识的信息检索与数据挖掘. 计算机研究与发展，2015，52(2)：456-474)

[3] Choi E, Kwiatkowski T, Zettlemoyer L. Scalable semantic parsing with partial ontologies//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 1311-1320

[4] West R, Gabrilovich E, Murphy K, et al. Knowledge base completion via search-based question answering//Proceedings of the 23rd International World Wide Web Conference. Seoul, Korea, 2014: 515-526

[5] Neelakantan A, Roth B, McCallum A. Compositional vector space models for knowledge base completion//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China, 2015: 156-166

[6] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning//Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, USA, 2010, 5: 3

[7] Mitchell T M, Cohen W W, Talukdar P P, et al. Never-ending learning//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2302-2310

[8] Dong Xin, Gabrilovich E, Heitz G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 601-610

[9] Zhong Xiu-Qin, Liu Zhong, Ding Pan-Ping. Construction of knowledge base on hybrid reasoning and its application. Chinese Journal of Computers, 2012, 35(4): 761-766 (in Chinese)
(钟秀琴，刘忠，丁盘苹. 基于混合推理的知识库的构建及其应用研究. 计算机学报，2012，35(4)：761-766)

[10] Gu Rong, Wang Fang-Fang, Yuan Chun-Feng, et al. YARM: Efficient and scalable semantic reasoning engine based on MapReduce. Chinese Journal of Computers, 2015, 38(1): 74-85(in Chinese)
(顾荣，王芳芳，袁春风等. YARM：基于 MapReduce 的高效可扩展的语义推理引擎. 计算机学报，2015，38(1)：74-85)

[11] Richardson M, Domingos P. Markov logic networks. Machine Learning, 2006, 62(1-2): 107-136

[12] Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data//Proceedings of the 28th International Conference on Machine Learning. Bellevue, USA, 2011: 809-816

[13] Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks. Machine Learning, 2010, 81(1): 53-67

[14] Lin Yan-Kai, Liu Zhi-Yuan, Luan Huan-Bo, et al. Modeling relation paths for representation learning of knowledge bases//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 705-714

[15] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases//Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011: 301-306

[16] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data//Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2787-2795

[17] Wang Zhen, Zhang Jian-Wen, Feng Jian-Lin, et al. Knowledge graph embedding by translating on hyperplanes//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Quebec City, Canada, 2014: 1112-1119

[18] Lin Yan-Kai, Liu Zhi-Yuan, Sun Mao-Song, et al. Learning entity and relation embeddings for knowledge graph completion//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2181-2187

[19] Schoenmackers S, Etzioni O, Weld D S, et al. Learning first-order horn clauses from web text//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts, USA, 2010: 1088-1098

[20] Lao Ni, Cohen W W. Fast query execution for retrieval models based on path-constrained random walks//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC, USA, 2010: 881-888

[21] Lao Ni, Mitchell T, Cohen W W. Random walk inference and learning in a large scale knowledge base//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 529-539

[22] Gardner M, Talukdar P P, Kisiel B, et al. Improving learning and inference in a large knowledge-base using latent syntactic cues//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Washington, USA, 2013: 833-838

[23] Gardner M, Talukdar P P, Krishnamurthy J, et al. Incorporating vector space similarity in random walk inference over knowledge bases//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 397-406

[24] Nickel M, Jiang Xue-Yan, Tresp V. Reducing the rank in relational factorization models by including observable patterns//Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Quebec City, Canada, 2014: 1179-1187

LIU Qiao, born in 1974, Ph. D., associate professor. His research interests include machine learning and data mining, expert system and recommendation system.

HAN Ming-Hao, born in 1992, M. S. candidate. His research interests include machine learning and data mining.

JIANG Liu-Yi, born in 1990, M. S. candidate. Her research interests include machine learning and data mining.

LIU Yao, born in 1978, Ph. D., lecturer. Her research interests include machine learning, social network analysis.

GENG Ji, born in 1963, Ph. D., professor. His research interests include big data analysis and information security.

Background

This paper focuses on the relation inference problem in knowledge bases(KBs). With the progress of the open domain information extraction technology, many large-scale knowledge bases are booming rapidly, typical examples include the Freebase, Knowledge vault, DBpedia and YAGO. Knowledge bases are the building blocks of business intelligence and other intelligent applications. However, in many practical cases, the available KBs are still suffer from information incompleteness. Relation inference techniques provide an efficient solution for knowledge base population tasks by automatically generating new facts from the existing knowledge.

Many research efforts can be found in this research area in the past decade, among which two modeling approaches are impressive to the whole community. The most popular one is the embedding models, also called the latent factor models, is to represent the knowledge (entities and relations) with real value low-dimensional vectors, then compute with it for relation inference tasks. Another one is the random walk models, which tries to model the KBs with directed graphs, and perform inference on it with random walk algorithms.

This research work starts from the problems we found in the basic assumptions of the random walk model, based on experimental proofs, we propose two alternative assumptions for relation inference tasks. The first assumption we claim is that the relations between entities are reciprocal semantically. The second assumption is that the relations are transitive in between the entities on both side of the same relation. Based on above assumptions, we propose a novel two-tier random walk based algorithm for relational inference, experimental results demonstrate the validity of our assumptions, and the effectiveness of the proposed algorithm. The evaluation process and results are systematically analyzed in details, which show that our approach significantly outperform other prevalent algorithms, meantime we also get solid proof for the validity of the above mentioned two modeling assumptions.

This work has been supported in part by the National Natural Science Foundation of China (61133016, U1401257, 61502087), and the Sichuan Hi-Tech Industrialization Program (2017GZ0308).