

规则增强的知识图谱表示学习方法

浙江大学计算机科学与技术系 杭州 310027

陈曦 陈华钧 张文

摘要 知识图谱 (Knowledge Graph, 简称 KG) 的表示学习方法旨在将知识图谱的实体和关系表示为稠密低维实值向量, 进而在低维向量空间中高效计算实体、关系及其之间的复杂语义关联, 在知识图谱的构建、推理、融合、挖掘以及应用中具有重要作用。已有的知识图谱表示方法仅仅考虑了知识图谱中的直接事实, 忽略了知识图谱中一些隐藏的语义信息, 这些语义信息对于知识图谱关系和实体的嵌入表示有着重要的影响。本文提出了一种规则增强的知识图谱表示学习方法, 该方法首先通过知识图谱规则挖掘的方法提取一组可代表知识图谱语义信息的 Horn 逻辑规则, 随后通过基于规则的物化推理方法将相应的隐藏语义信息注入到知识图谱表示学习模型中。实验结果表明, 基于规则增强的方法可以显著提升已有知识图谱表示学习模型在链接预测和定理预测上的效果和性能。

关键词: 知识图谱, 表示学习, 规则挖掘, 推理

中图分类号: TP391

Rule-Augmented Representation Learning Approach in Knowledge Graph

Department of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

CHEN Xi CHEN HuaJun ZHANG Wen

Abstract Representation learning in Knowledge Graph (KG) aims to project the entities and relations into a dense, real-valued and low-dimensional vectors, so as to efficiently measure complex semantic correlations between entities and relations, and plays a important role in knowledge acquisition, inference, fusion, mining and applications of KG. Existing methods of representation learning in KG only concern direct facts in KG, ignoring some implicit semantic information. The paper proposes proposed a rule-augmented representation learning approach in KG. This method firstly implemented a rule mining algorithm to extract Horn rules from

作者简介: 陈曦 (1990-), 博士, 研究方向: 知识图谱、知识推理、表示学习, email: xichen@zju.edu.cn; 陈华钧 (1978-), 博士, 教授, 研究方向: 知识图谱、大数据、语义技术等; 张文 (1992-), 博士, 研究方向: 知识图谱、表示学习、知识挖掘。

the KB. Then, a rule-based materialization reasoning was used to encode logical rules into our learning models. Experimental results on real-world datasets showed that, this approach achieved significant and consistent improvements compared with these baselines in the link and axiom prediction tasks.

Keywords: Knowledge graph, representation learning, rule mining, reasoning

1 研究背景

目前,知识图谱(Knowledge Graph)以其强大的语义处理能力与开放互联能力得到了学术界和工业界的极大关注。它为互联网上海量、异构、动态的大数据表达、组织、管理以及利用提供了一种更为有效的方式,使得网络的智能化水平更高,更加接近于人类的认知思维,同时也支持相应的知识推理。包括谷歌、微软、百度、搜狗等在内的各个互联网大公司争先构建相应的知识图谱,并在知识管理、智能搜索、自动问答等方面得到了广泛的应用。当前互联网中已经有一些可以获取的大规模知识库,例如 LinkedData^[1]、DBpedia^[2]、Freebase^[3]、YAGO^[4]等。

这些知识库多采用万维网联盟(W3C)发布的资源描述框架(Resource Description Framework, RDF)^[5]三元组(实体1,关系,实体2)为基本的存储格式。但这种存储方式面临着计算效率低下、数据稀疏、可重用性差等问题。近年来,以深度学习为代表的表示学习技术在自然语言处理、文本分析、图像分析、语音识别等领域取得了重要成果,这为知识图谱的表示学习发展提供了基础。知识图谱的表示学习技术

是通过机器学习将知识图谱中的实体和关系表示为稠密低维实值向量,这不仅大大提升了知识图谱的应用范围,使其在知识融合、推理、挖掘等发挥重要作用,并且有效解决了数据稀疏性等相关问题。

近些来,以TransE^[6]为代表的知识图谱表示学习模型陆续提出,然而绝大部分的这些模型在进行向量投影时仅仅考虑了知识图谱中的直接事实,忽略了知识图谱中一些隐藏的语义信息,这会使得学习到的向量不能准确的表达原知识图谱中的语义关系。事实上,一个完整的知识图谱可看作是由两部分组成即TBox(terminological component)和ABox(assertion component),前者包含了术语描述本文简称为定理(Axiom),后者包含了实例描述简称为事实(Fact)。由于当前的大部分公开知识图谱都是自底向上构建的,所以包含很少的术语信息,这使得定理的获取比较困难。此外,在大规模知识图谱中,定理的数量一般比事实少得多,显然直接将定理放入学习模型中也是不可取的。这篇文章提出了一种规则增强的知识图谱表示学习方法,该方法首先自动挖掘知识图谱中的Horn规则,随后通过基于规则的物化推理(Materialization)得出新的事实,这部分事实可以间接代表知识图谱的语义信息,最后通过

一种语义集成的方法注入到知识图谱表示学习模型中。实验结果表明,基于规则增强的方法可以显著提升已有知识图谱表示学习模型的效果和性能。

2 研究现状分析

相关工作主要包括已有的知识图谱表示学习模型以及相应的改进方法。

2.1 基本的知识图谱表示学习模型

TransE模型将知识库中实体之间的关系看成是从实体间的某种翻译过程。关系 r 可以看作是从头实体向量 h 到尾实体向量 l 的翻译。模型对于知识库中的每个三元组 (h,r,t) ,都希望尽量满足以下关系: $h+r \approx t$,其损失函数为向量 $h+r$ 和 t 的L1或L2距离。该模型简单有效,是目前综合效果最好和应用最广泛的一个模型。

TransH^[7]通过引入关系超平面来改进TransE中的翻译过程。它首先将头部、尾部实体向量投影到关系所在超平面,然后在超平面上完成翻译过程。该模型能够使同一个实体在不同的关系中扮演不同的角色,这样可以在复杂关系的链接预测上取得更好的效果。

此外,TransR^[8]模型引入关系矩阵进行投影映射,TransD^[9]模型考虑实体和关系之间的交互,分别定义了头实体和尾实体并在对应关系空间上的投影矩阵,Hole^[10]模型提出头尾实体向量循环相关的张量乘法来表示实体对以获取三元组“全息表示”。

以上这些模型主要从关系和实体的复杂度方向进行模型的扩展,以适应更复杂的建模需

求,但其在进行建模时仅考虑知识图谱中的直接事实,忽略了知识图谱中潜在的定理信息,进而使得学习到的表示向量不能准确表达知识图谱的语义信息。

2.2 增强的知识图谱表示学习方法

文章^[11]将关系定义域和值域的限制作为先验知识加入到已有的表示学习模型中去,在进行样本选取时,仅考虑那些满足关系类型的三元组进行建模,这样可以有效过滤一些不相关的错误样本,可以一定程度上提升表示性能。文献[12]则考虑添加实体的层次信息,并通过编码器加入到表示学习模型中去,这可以为不同层次的实体创建融合的语义向量从而提高表示效果。文献[13]则引入额外的文本信息,通过联合知识图谱中的三元组事实以及文本中的实体描述进行联合训练,从而提升其预测复杂关系的实体链接能力。

以上这些模型主要从添加部分特定信息进行模型的扩展,但其主要有两个缺点,一个是额外信息的获取必须依靠人工构建,使得其扩展能力很差,往往只能应用在特定数据集上;其次额外信息只是知识图谱一小部分定理信息,未能完全捕捉知识图谱中大部分的定理信息,这样学习到的语义向量不能完整表达知识图谱的语义信息。

3 规则增强的表示学习方法

3.1 原理分析

我们用 G 表示知识图谱,即 $G=(E,R,F)$,其中 E 是知识库中的实体集合, R 是知识库中

的关系集合, $F \subseteq E \times R \times E$ 代表知识库中的事实, 每个事实由三元组 (实体1, 关系, 实体2) 构成, 黑体表示对应实体和关系的向量。Horn规则由规则头和规则主体构成, 表示为 $B_1 \wedge B_2 \wedge B_3 \wedge \dots \wedge B_n \Rightarrow r(x,y)$ 简写为 $B \Rightarrow r(x,y)$, Horn规则具有很强的语义表达能力, 能够表达知识图谱大多数语义定理包括属性的近义定理 (Synonym)、逆向定理 (Inverse)、属性链定义 (Property Chain) 等等, 例如规则 $\gamma (?x, \text{出生地}, ?y) \wedge (?y, \text{所属国家}, ?z) \Rightarrow (?x, \text{国籍}, ?z)$ 间接刻画了知识图谱一个属性链定理 α : 出生地 \circ 所属国家 \subseteq 国籍。因此如果能够将这些规则所表达的语义信息嵌入到关系所对应的向量中去, 将会使得我们学习的模型具有更加强的链接预测 (Link Prediction) 和定理预测 (Axiom Prediction) 能力。

考虑一个知识图谱G, 包含两个事实F1 (奥巴马, 出生地, 夏威夷)、F2 (夏威夷, 所属国家, 美国) 和一个属性链定理 α 。如图1

左所示TransE模型在训练的时候只会优化F1和F2所对应的两个样本, 使它们头尾实体向量之间的距离差近似等于关系的向量值, 但是由于忽略了潜在的语义定理 α , 使得其模型不能捕捉关系出生地 (r_1)、所属国家 (r_2) 和国籍 (r_3) 之间的语义关系, 显然这将会限制模型定理预测的能力, 进一步这也会影响其预测新事实的能力, 例如如果向原知识图谱中加入两个新的事实F4 (姚明, 出生地, 上海) 和F5 (上海, 所属国家, 中国), TransE将不能通过学习到的向量来推理出新的事实F6 (姚明, 国籍, 中国)。

如图1右所示, 规则增强的方法首先从原知识图谱中挖掘规则 γ , 随后通过规则推理器的推理得出新的事实F3 (奥巴马, 国籍, 美国), 通过添加新的这个事实到原有学习模型中去可以帮助原始模型的训练间接得到属性链定理 α ($r_1+r_2=r_3$), 从而使得这个模型具有定理预测的能力。进一步, 它也能借助这个定理帮助预测推理新的事实F6。

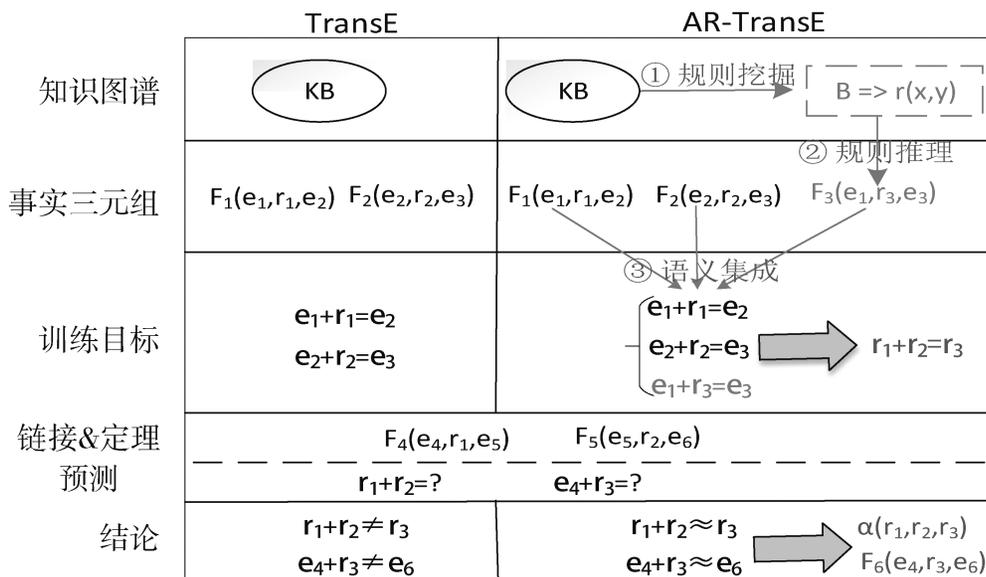


图1 TransE VS AR-TransE (规则增强的TransE)

3.2 框架设计

3.2.1 规则挖掘与推理

对于知识图谱的规则挖掘，基于文献[14]我们开发了一个改进版本的基于归纳逻辑编程（ILP）的模糊Horn规则挖掘算法，该算法克服了传统规则挖掘算法对基于开放世界假设（OWA, Open World Assumption）负证据生成的困难，并提出了基于部分完整假设（PCA, Partial Completeness Assumption）的支持度计算方法，规则的支持度和置信度如公式1和公式2所示。更多的算法细节和源码可参照^①

$$\text{sup}(B \Rightarrow r(x,y)) := \#(x,y) : \exists z_1, \dots, z_m : B \wedge r(x,y) \quad (1)$$

其中支持度sup表示这个规则在图谱中命中三元组事实的个数，z是自由变元，注意这里只考虑规则头主语和宾语所对应的事实个数。

$$\text{conf}(B \Rightarrow r(x,y)) := \frac{\text{sup}(B \Rightarrow r(x,y))}{\#(x,y) : \exists z_1, \dots, z_m, y' : B \wedge r(x,y')} \quad (2)$$

其中置信度表示这个规则的可信程度（模糊度），是图谱中命中三元组正确事实的个数与总命中个数的比值，分子表示支持度，分母中z表示自由变元，y'表示通过PCA假设所计算得到的规则头所有可能宾语。

随后，通过实施一个基于模糊规则的物化推理算法可以得到一组新的事实，可表示为 $\delta:G \vdash G'$ ， δ 表示推理器， τ 表示第一步挖掘得出的规则， G' 表示推理得到的新的事实。

3.2.2 语义集成

在得出能间接表达知识图谱语义的事实之

后，我们提出了一种语义集成的方案将其语义信息嵌入进模型。我们在原有知识图谱之上融合了这些新的事实，并添加一个事实的可信度（Re），事实的可信度是由它对应的推理规则所决定的，如果其对应多个不同推理规则，则取最大值。这样原知识图谱G扩展为四元组形式 G^* ，公式3表示语义集成后的知识图谱。事实的可信度通过公式4计算得到，具体说来，如果推理得出的事实已经存在，则可信度为1；如果推理得出的事实不存在并且它的可信度大于一个阈值 σ （称为截断因子），则取其对应推理规则置信度最大的值；如果推理得出的事实不存在并且它的可信度小于 σ 则可信度为0。设置截断因子 σ 主要是想过滤一些噪声，因为挖掘出来的规则不是完全正确的，置信度低的规则所引入的语义信息很可能是不可靠的，因此可以直接截断这些样本，以免影响模型训练的效果。

$$G^* = \{(h,r,t,w) | (h,r,t) \in G \cup G', w = \text{Re}(h,r,t)\} \quad (3)$$

其中Re表示事实的可信度

$$\text{Re}(F) = \begin{cases} 1, & F \in G \cap G' \\ w, & F \in G' - G \text{ 并且 } w \geq \sigma \\ 0, & F \in G' - G \text{ 并且 } -w < \sigma \end{cases} \quad (4)$$

其中 $w = \text{Max}(\text{conf}(R^F))$ 表示事实F所对应的推理规则置信度的最大值， R^F 表示能推理出事实F的所有规则集合。

3.2.3 模型融合

我们所提出的规则增强的方法是一个通用的方法，可以适用于多个表示学习模型，从而增强原始模型的表示能力。以TransE为例，公

^①www.github.com/hualichenxi/AR-TransE

式5描述了应用规则增强方法后的TransE模型（简称为AR-TransE）的损失函数，该损失函数是一个基于margin的排序得分函数，主要的思想是想让那些正样本的距离函数值远远小于负样本的距离函数值，这里正样本的距离值增加了一个表示样本可信度的权重，对于可信度低的样本，其损失值相对增大，公式6是对应的距离假设函数。

$$\text{loss}_{\text{AR-TransE}} = \sum_{(h,r,t,w) \in G^*} \sum_{(h',r',t') \in \bar{G}} [\Upsilon + w * d(h,r,t) - d(h',r',t')]_+ \quad (5)$$

其中 G^* 表示规则增强后的新样本数据， \bar{G} 表示通过负抽样（Negative Sampling）的方法生成的负样本， Υ 是一个表示margin的超参数， w 表示样本可信度。

$$d(h,r,t) = |h+r-t| \quad (6)$$

h 、 t 、 r 分别表示三元组样本头、尾实体和关系的向量。

4 实验

实验包含两个任务，第一个任务是链接预测：通过学习到的向量预测一个正确三元组样本的尾实体；第二个任务是定理预测：通过学习到的向量预测一个正确语义定理，实验包含三类定理即同义定理（Synonym Axiom）、反向定理（Inverse Axiom）和属性链定理（Property Chain Axiom）。

4.1 数据集描述

我们选取了两个标准的知识图谱表示学习benchmark数据集WN18和FB15k见表1，同其他模型一样数据集分为了训练集、验证集和测试

集，训练集用于训练我们的模型，验证集用于确定最优的超参数（学习率、margin和向量维度 k ），测试集用于评价模型的性能。

表1 数据集

数据集	# 关系	# 实体	# 训练集	# 验证集	# 测试集
WN18	18	40,943	141,442	5000	5000
FB15k	1345	14,951	483,142	50,000	59,071

4.2 实验对比方法

我们选取了两个代表性的表示学习模型TransE和TransH，基于这两个模型分别实施了对应的规则增强的方法简称为AR-TransE和AR-TransH，并根据从低到高的三种截断因子（ $\sigma=0.2, 0.5, 0.8$ ）分别设置了三个不同的规则增强模型。

关于模型性能评价我们使用标准的评价方法：1）平均排名（Mean Rank）；2）前十占比（HITS@10）；3）前1占比（HITS@1）。链接预测的评价方法是在测试样本中给出三元组样本头实体和关系的表示向量，随后验证测试集的尾实体所对应的表示向量在实际所有实体中的排名，如果排名前十，则HITS@10指标加1，随后计算相应所有测试样本的平均值。定理预测的评价方法与前面类似，不过这里采用HITS@1指标，因为所有模型的HITS@10都大于95%。除此之外，由于链接预测的正确尾实体可能包含多个，这里采用原始（raw）和过滤（filter）两种形式，filter表示在测试尾实体时，虽然正确的这个实体不是正在测试的这个三元组的尾实体，但如果它所对应的三元组同样也出现在测试集中，我们也认为这个实体是正确的尾实体，可以直接将其过滤，不影响其

排名。

4.3 实验设置

模型包含了几个超参数：学习率、向量维度和margin。其中学习率设置为：0.1、0.01、0.001，向量维度k为：10、50、100，margin设置为：1、2、5。验证集中最优的模型参数分别为0.001、50、1。训练迭代次数为1000。我们通过实施mini-batch的梯度下降算法训练了这些模型。

4.4 实验结果

4.4.1 链接预测

表2显示了模型在WN18和FB15k链接预测任务上的实验结果，结果显示规则增强的方法的性能在所有数据集和评价方法上都优于baseline模型。对于TransE模型，WN18和FB15k上其平均排名效果最高分别提升了14%、15%、37%、61%，前十占比提高了8%、3%、32%、42%；对于TransH模型，WN18和FB15k上其平均排名效果最高分别提升了16%、17%、26%、31%，前十占比提高了3%、6%、8%、20%。

表2 链接预测结果

数据集	WN18								FB15k			
	Mean Rank		HITS@10		Mean Rank		HITS@10		Mean Rank		HITS@10	
	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter
TransE	263	251	75.4	89.2	243	125	34.9	47.1				
AR-TransE ($\sigma=0.2$)	225	213	81.3	92.2	151.9	48.9	42.6	67.0				
AR-TransE ($\sigma=0.5$)	234	225	81.1	91.8	196.7	72.1	45.9	63.5				
AR-TransE ($\sigma=0.8$)	254	242	80.5	91.7	163.5	61.1	46.2	62.6				
TransH	318	303	75.4	86.7	211	84	42.5	58.5				
AR-TransH ($\sigma=0.2$)	264.4	251.2	77.9	91.7	156.7	64	45.6	65.7				
AR-TransH ($\sigma=0.5$)	264.8	252.7	77.6	89.9	186.3	29.5	45.7	70.1				
AR-TransH ($\sigma=0.8$)	277.2	262.3	76.8	88.1	178.1	75.3	44.2	65.1				

结果表明，通过我们的规则增强方法可以有效的向知识图谱表示学习模型嵌入更多语义信息，从而提升了模型链接预测的能力。

4.4.2 定理预测

表3显示了模型在FB15k定理预测任务上的实

验结果，结果显示应用规则增强的方法大大提升了其定理预测的能力，同义定理（Synonym）、反向定理（Inverse）和属性链定理（Inverse）平均排名最高分别提升了47%、41%、82%，前十占比提高了11%、3%、14%。

表3 FB15k定理预测结果

评价参数	Mean Rank			HITS@1		
	Synonym	Inverse	Inverse	Synonym	Inverse	Chain
TransE	109.4	70.7	28.6	47.7	64.5	79.5
AR-TransE ($\sigma=0.2$)	57.7	41.5	6.5	51.2	66.2	84.8
AR-TransE ($\sigma=0.5$)	58.4	49.4	4.9	53.1	66.4	90.9
AR-TransE ($\sigma=0.8$)	69.3	49.9	6.1	51.7	66.6	86.3

结果表明，我们所提出的规则增强方法可以有效的将知识图谱潜在的定理信息编码在表示向量中，这是由于通过规则嵌入的新知识可以间接的表达知识图谱的语义定理。

4.4.3 截断因子 σ 的影响

从表1链接预测结果可以看出，截断因子 σ 取值较小时（0.2）模型效果更好，这是由于尽管更小的截断因子将会给模型带来更多噪声，但是与此同时也将会向模型中引入更多的规则和语义信息，而收益于我们在设计损失函数时加入了可信度权重，可以减轻这些噪声对模型的影响，使得在链接预测任务中噪声所带来的影响小于规则引入所带来的益处，因此对于链接预测任务通常设置一个相对较小的截断因子。从表2定理预测结果可以得出，截断因子 σ 取值为0.5时模型效果稍好于取值为0.2的模型，这是因为定理预测任务对噪声的敏感度较强，引入过多可信度不高的规则反而会影响模型的学习效果，因此对于定理预测任务倾向于取一个适中的截断因子即可。

5 小结

传统的知识图谱表示学习模型由于忽略了知识图谱中隐藏的语义信息，使得其表示学习的性能和效果存在一些缺陷，针对这个缺陷本文提出了一种规则增强的知识图谱表示学习方法，该方法首先自动挖掘知识图谱中的Horn规则，这些规则能够作为间接反映图谱语义信息的先验知识，随后通过基于规则的物化推理得出新的事实，并且设计了一种语义集成的方

法将这部分知识注入到知识图谱表示学习模型中。在标准的benchmark数据集上的实验结果表明，基于规则增强的方法可以显著提升已有知识图谱表示学习模型在链接预测和定理预测任务上的性能。

参考文献

- [1] Bizer C, Heath T, Idehen K, et al. Linked Data on the Web[C]//Proceedings of the 17th International Conference on World Wide Web. ACM, 2008: 1265-1266.
- [2] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data[C]// The Semantic Web, International Semantic Web Conference, Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November. 2007:11-15.
- [3] Bollacker K, Evans C, Paritosh P, et al. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge[C]// ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, Bc, Canada, June. DBLP, 2008:1247-1250.
- [4] Suchanek F M, Kasneci G, Weikum G. Yago: A Core of Semantic Knowledge[C]// International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May. OAI, 2007:697-706.
- [5] Klyne, Graham, Carroll, et al. Resource Description Framework (RDF): Concepts and Abstract Syntax[J]. World Wide Web Consortium Recommendation, 2004.
- [6] 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.
- [7] Wang Z, Zhang J, Feng J, et al. Knowledge Graph Embedding by Translating on Hyperplanes[J]. AAAI - Association for the Advancement of Artificial Intelligence, 2014.
- [8] 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. AAAI Press, 2015:2181-2187.

[9] Ji G, He S, Xu L, et al. Knowledge Graph Embedding via Dynamic Mapping Matrix[C]//Proc of ACL. Stroudsburg, PA: ACL, 2015: 687-696.

[10] Nickel M, Rosasco L, Poggio T. Holographic Embeddings of Knowledge Graphs[J]. Computer Science, 2015.

[11] Krompaß D, Baier S, Tresp V. Type-Constrained Representation Learning in Knowledge Graphs[M]// The Semantic Web - ISWC 2015. Springer International Publishing, 2015:A92.

[12] Xie R, Liu Z, Sun M. Representation Learning of Knowledge Graphs with Hierarchical Types[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, New York, USA. 2016:2965-2971.

[13] Wang Z, Li J. Text-enhanced Representation Learning for Knowledge Graph[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, New York, USA. 2016: 1293-1299.

[14] Galárraga L A, Teflioudi C, Hose K, et al. AMIE: Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases[C]//Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 413-422.