DOI:10.19651/j.cnki.emt.2105695

知识图谱中基于多关系路径的链路预测方法*

袁华兵[1] 刘 敏[2] 杨延庆[1]
(1.西安医学院 信息技术处 西安 710021;2.陕西理工大学 数学与计算机科学学院 汉中 723001)

摘 要:针对基于单一关系路径的链路预测方法无法挖掘知识图谱中不同路径之间影响的问题,提出了一种基于多关系路径的链路预测方法。首先,采用基于路径信息的相似性指标来计算所有关系路径之间的相似度。然后,将不同路径之间的关系投影延伸至新的路径投影和路径约束,并采用随机梯度下降执行训练过程,从而能够在隐式空间中通过低维表示学习筛选出不同路径之间的显式特征。在安然邮件数据集和美国国家自然科学基金会数据集上进行了验证分析。实验结果表明,相比于其他多种路径链路预测算法,该算法在 MAP 和 AUC 指标上的最大提升幅度约 20%,表现出更高的预测精度。

关键词:链路预测;知识图谱;多关系路径;路径投影
中图分类号:TP311 **文献标识码**:A **国家标准学科分类代码**:510.50

Link prediction method based on multi relation path in knowledge map

Yuan Huabing[1] Liu Min[2] Yang Yanqing[1]
(1. Information Technology Department, Xi'an Medical College, Xi'an 710021, China;
2. School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong 723001, China)

Abstract: In order to solve the problem that the link prediction method based on single relational path cannot mine the influence of different paths in the knowledge map, a link prediction method based on multi relational path is proposed. Firstly, the similarity index based on path information is used to calculate the similarity between all relational paths. Then, the relationship projection between different paths is extended to the new path projection and path constraints, and the training process is performed by using random gradient descent, so that the explicit features between different paths can be screened out through low dimensional representation learning in implicit space. The validation analysis is carried out on Enron email data set and National Natural Science Foundation data set. Experimental results show that, compared with other path link prediction algorithms, the maximum improvement of MAP and AUC is about 20%, showing higher prediction accuracy.

Keywords: link prediction; knowledge mapping; multi relational path; path mapping

0 引 言

近年来随着大规模 Web 网络数据的激增,知识图谱等[1-3]智能自动化应用成了当前自然语言处理领域非常热门的研究方向。以图结构作为表示框架的知识图谱是一种特殊类型的信息网络,能够自动从海量的各类型结构化数据中提取结构化信息,例如文本自动生成,从而节省了大量的人工成本。

链路预测作为知识推理技术的一个重要分支,是知识图谱算法应用的主要场景之一,具有广泛的实际应用,例如复杂网络结构推演分析等。目前,相关研究一直是知识图谱领域的热门方向。例如,方阳等[4]提出一种改进的基于翻译的知识图谱表示方法,通过加入了自适应对角权重矩阵来提高链路预测性能,具有良好的知识表达能力。王文涛等[5]提出了一种基于改进随机游走的网络表示学习算法,解决了随机游走倾向于较大度的节点的偏向问题,从而更好地反映知关系路径中节点的结构化信息。但是上述方法均是针对单一关系路径开展的研究,无法处理一对多,多对一和多对多等复杂关系路径,这是因为它们忽视了知识图谱中不同路径之间影响。

为了解决上述问题,本文提出了一种基于多关系路径的链路预测方法,主要创新之处在于将不同路径之间的关

收稿日期:2021-01-14
*基金项目:陕西省自然科学基金(2019JQ-927)项目资助

系投影延伸至新的路径投影，并采用随机梯度下降执行训练过程，从而能够在隐式空间中通过低维表示学习筛选出不同路径之间的显式特征。两种经典数据集上的实验结果验证了所提链路预测方法的有效性和先进性。

1　多关系路径的图表示

首先，多关系路径采用的图表示方式如下[6-7]：

$$G=(V,E) \tag{1}$$

式中：$E=E_1\cup E_1\cup\cdots\cup E_d$ 为各种类型边缘的集合；d 为整个网络结构中边缘的类型数；V 为全部节点的集合。那么，图 G 的子图 G_r 可以表示为：

$$G_r=(V,E_r) \tag{2}$$

式中：E_r 为第 r 种边缘组成的集合。

子图 G_r 的邻接矩阵表示为 $\boldsymbol{A}^r=[a_{ij}^r]$，其中 a_{ij}^r 为节点 v_i 和 v_j 对应的矩阵元素。当节点 v_i 和 v_j 间存在第 i 类型的边时 $a_{ij}^r=1$，否则 $a_{ij}^r=0$。在对于 $t(1\leqslant t\leqslant d)$，通过相邻矩阵 $\boldsymbol{A}^t$ 表示的各类型链路的拓扑结构，来预测网络中可能的第 i 类链路 $i=1,2,\cdots,d$，获得节点相似度矩阵 $\boldsymbol{S}$ 从而完成链路预测，因此，第 i 类链路就是预测算法的目标维度。

2　基于多关系路径的链路预测方法

2.1　基于路径信息的相似性计算

不同于对单关系路径的链路预测方法，多关系路径的链路预测方法面对的链路类型更多和网络结构更复杂，因此采用了陈春谋[8]提出的概率加权模型，具体相似性计算方式如下：

$$S_z(x,y)\sum_{n\in N_{xy}}\frac{\sigma P_z(x,y)}{P(x,n)P(y,n)} \tag{3}$$

式中：$S_z(x,y)$ 为关系 z 上的连接边 (x,y) 的相似度；N_{xy} 为连接边 (x,y) 上的采样总数；σ 为调节权重；$P_z(x,y)$ 为关系 z 上的连接边 (x,y) 上所有类型的概率；$P(x,n)$ 为连接边 (x,n) 上所有类型的概率；$P(y,n)$ 为连接边 (y,n) 上所有类型的概率。通过式(3)解决多关系路径中同类关系极稀疏问题。

2.2　多关系路径投影

为了挖掘知识图谱中不同路径之间影响，就需要更好的知识表示学习，从而最大化的利用多关系路径之间的语义信息。本文采用组合的路径投影，在路径空间和关系空间(隐式空间)同时映射实体，以便构建多匹配属性关系。

设定图 G 包含的某一个三元组实例 (h,r,t)，其中 h 为头实体，t 为尾实体，r 为头尾实体之间存在的二元关系。通过投影矩阵 $\boldsymbol{M}_r,\boldsymbol{M}_p\in\boldsymbol{R}^{m\times n}$ 把实体空间向量 $\boldsymbol{h},\boldsymbol{t}\in\boldsymbol{R}^n$ 映射到相关的路径空间和关系空间，其中 m 为关系映射维度，n 为实体映射维度。对应映射空间的投影向量 $(\boldsymbol{h}_r,\boldsymbol{h}_p,\boldsymbol{t}_r,\boldsymbol{t}_p)$ 的定义为：

$$\boldsymbol{h}_r=\boldsymbol{M}_r\boldsymbol{h},\quad \boldsymbol{h}_p=\boldsymbol{M}_p\boldsymbol{h} \tag{4}$$

$$\boldsymbol{t}_r=\boldsymbol{M}_r\boldsymbol{t},\quad \boldsymbol{t}_p=\boldsymbol{M}_p\boldsymbol{t} \tag{5}$$

设一个多关系路径的第 k 类子图的邻接矩阵为 $\boldsymbol{A}$，$\boldsymbol{A}$ 的每一行都是一个向量 $\boldsymbol{p}=(r_1,r_2,\cdots,r_m)$，可以被视为关系的序列形式。所以，本文提出采用投影矩阵 $\boldsymbol{M}_r$ 进行路径投影矩阵 $\boldsymbol{M}_p$ 的动态构建，以便减少涉及的参数数量，避免多关系路径投影计算过于复杂。具体采用了元素累加操作方法，定义如下：

$$\boldsymbol{M}_p=\boldsymbol{M}_{r_1}+\boldsymbol{M}_{r_2}+\cdots+\boldsymbol{M}_{r_m} \tag{6}$$

在上述步骤后，对路径投影矩阵执行正则化操作，其能量函数的定义为：

$$G(h,r,t)=E(h,r,t)+\lambda E(s,p,t)=$$
$$\|\boldsymbol{h}_r+\boldsymbol{r}-\boldsymbol{t}_r\|+Z\|\boldsymbol{h}_p+\boldsymbol{p}_i^*-\boldsymbol{t}_p\| \tag{7}$$

$$\boldsymbol{p}_i^*=\boldsymbol{r}_1+\boldsymbol{r}_2+\cdots+\boldsymbol{r}_m \tag{8}$$

$$Z=\frac{\lambda}{2}\sum_{p_i\in P(h,t)}P(t\mid h,p_i) \tag{9}$$

式中：λ 为调节两个能量函数的超参数；Z 为正则化因子；$P(t\mid h,p_i)$ 为路径受限的随机游走概率。

通过在路径空间和关系空间同时映射实体，获取了更多的路径语义信息，从而有助于提高知识表示学习模型在隐式低维空间中对相似映射的区分能力。多关系路径投影原理如图 1 所示。

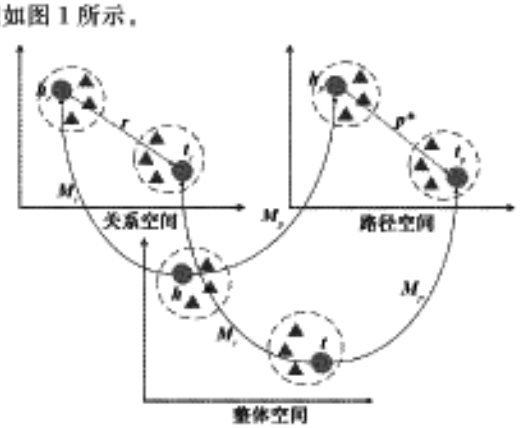

图 1　多关系路径投影原理

采用随机梯度下降执行训练过程，在实际训练中，每次迭代计算的优化目标函数为式(7)。基于多关系路径的链路预测流程如图 2 所示。

3　实验结果与分析

3.1　实验数据集

选择的实验数据集是来自真实世界的两个经典大型知识图谱，分别为安然公司邮件数据集和美国国家自然科学基金会(National Science Foundation，NSF)数据集。上述两个数据集的具体参数分别如表 1 和 2 所示。

3.2　实验评估指标

1)采用机器学习领域中的一种模型评估指标：ROC 曲线下的面积(area under the ROC curve，AUC)，来评估链路预测算法的准确度。AUC 的计算方式如下：

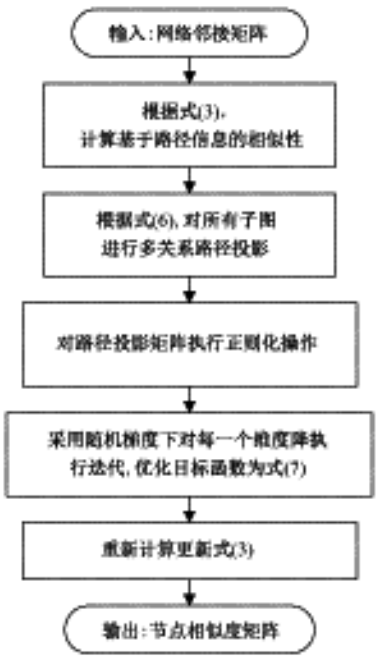

图 2　基于多关系路径的链路预测流程

表 1　安然邮件数据集的具体参数

关系路径类型	节点数量	链路数量
双方存在邮件往来 EA (email adjacency)	351	478
双方均给同一个账号发邮件 CR(co-receiver)	351	5 065
双方收到同一个账号的邮件 CS(co-sender)	351	4 232
双方的职务类似 EP(equal position)	351	5 594

表 2　NSF 数据集的具体参数

关系路径类型	节点数量	链路数量
两篇论文的作者一致 SA(same-author)	1 696	1 533
两篇论文的出版会议一致 SC(same-conference)	1 696	501
两篇论文的关键词一致 CK(co-keywords)	1 696	75 590

$$AUC = \frac{\sum_{(x_i, x_i \in \text{正样本集}} rank_{x_i} - \frac{M \times (M+1)}{2}}{M \times N} \tag{10}$$

式中：$rank_{x_i}$ 为第 i 条样本的序号；M 为正样本的个数；N 为负样本的个数；$\sum_{(x_i, x_i \in \text{正样本集}}$ 表示只把正样本的序号加起来。

2)采用平均精度值(mean average precision，MAP)，其定义如下：

$$MAP = \int_0^1 P(R)\mathrm{d}R \tag{11}$$

式中：P，R 分别为准确率与召回率。AUC 和 MAP 评估指标均是数值越接近 1，则预测精度越高，而 MAP 更能体现多对多复杂关系路径的链路预测性能。

3.3　结果分析

为了验证所提方法的先进性，将其与现有的 8 种链路预测算法(PA[9]、Sorensen[10]、Jaccard[11]、HPI[12]、CN[13]、LHN_I[14]、ELLPMDA[15]、改进随机游走[17])进行了对比分析。安然邮件数据集和 NSF 数据集上 9 种算法的 AUC 结果对比，分别如表 3 和 4 所示。

表 3　安然邮件数据集上的 AUC 结果对比

链路预测算法	EA	CR	CS	EP
PA	0.797	0.769	0.820	0.803
Sorensen	0.803	0.850	0.830	0.811
Jaccard	0.805	0.851	0.831	0.816
HPI	0.811	0.873	0.826	0.837
CN	0.813	0.864	0.840	0.845
LHN_I	0.814	0.855	0.848	0.862
ELLPMDA	0.815	0.903	0.900	0.876
Katz	0.872	0.944	0.959	0.923
本文	**0.899**	**0.968**	**0.973**	**0.946**

表 4　NSF 数据集上的 AUC 结果对比

链路预测算法	SA	SC	CK
PA	0.591	0.77	0.73
Sorensen	0.925	0.83	0.91
Jaccard	0.391	0.67	0.75
HPI	0.932	0.83	0.97
CN	0.954	0.82	0.96
LHN_I	0.963	0.82	0.97
ELLPMDA	0.957	0.82	**0.98**
Katz	0.942	0.84	**0.98**
本文	**0.976**	**0.90**	**0.98**

安然邮件数据集和 NSF 数据集上 9 种算法的 MAP 结果对比，分别如表 5 和 6 所示。

表 5　安然邮件数据集上的 MAP 结果对比

链路预测算法	EA	CR	CS	EP
PA	0.877	0.849	0.900	0.883
Sorensen	0.883	0.930	0.910	0.891
Jaccard	0.885	0.931	0.911	0.896
HPI	0.891	0.953	0.906	0.917
CN	0.893	0.944	0.920	0.925
LHN_I	0.894	0.935	0.928	0.942
ELLPMDA	0.895	0.963	0.962	0.956
Katz	0.952	0.982	0.971	0.980
本文	**0.989**	**0.984**	**0.983**	**0.991**

表 6　NSF 数据集上的 MAP 结果对比

链路预测算法	SA	SC	CK
PA	0.621	0.820	0.761
Sorensen	0.825	0.866	0.842
Jaccard	0.421	0.720	0.785
HPI	0.842	0.863	0.815
CN	0.854	0.855	0.842
LHN_I	0.873	0.854	0.887
ELLPMDA	0.877	0.853	0.891
Katz	0.862	0.877	0.892
本文	**0.980**	**0.982**	**0.994**

表 3～6 中加黑的数字表示链路预测中性能评估最优。从表 3～6 的结果可以看出，所提链路预测算法在两个测试数据集上均表现出最好的性能，即最高的 AUC 和 MAP，最大提升幅度约 20%。其中，最显著的是两个数据集上的 MAP 结果，4 类关系路径的平均 MAP 达到了 0.988。这说明，相比其他 8 种链路预测算法，该算法通过路径投影能够充分挖掘每一类关系路径所包含的语义信息，从而获得了更精确的结果。

4　结　　论

本文提出了一种基于多关系路径的链路预测方法。通过将不同路径之间的关系投影延伸至新的路径投影，从而能够充分挖掘每一类关系路径所包含的语义信息。通过两种真实数据集的实验结果得出如下结论：1）相比其他链路预测算法，该算法的 AUC 和 MAP 均有提高，获得了更好的预测精度；2）在 MAP 指标方便的提升效果较为明显，说明该算法能更有效完成多对多复杂关系路径的链路预测任务。然而，映射过程中没有考虑路径类型的约束问题，后续将针对开问题开展进一步研究。

参考文献

[1] ZHAO X Y, SHENG L, DIAO T X, et al. Knowledge mapping analysis of Ebola research[J]. Bratisl Lek Listy, 2015, 116(12): 729-734.

[2] LIN Z, WU C, HONG W. Visualization analysis of ecological assets/values research by knowledge mapping[J]. Acta Ecologica Sinica, 2015, 35(5): 142-154.

[3] CHANG H. Synergy of scientometric analysis and knowledge mapping with topic models: Modelling the development trajectories of information security and cyber-security research[J]. Journal of Information & Knowledge Management, 2016, 15(4): 77-84.

[4] 方阳，赵翔，谭真，等. 一种改进的基于翻译的知识图谱表示方法[J]. 计算机研究与发展，2018，55(1)：139-150.

[5] 王文涛，黄焯，吴琳涛，等. 基于改进随机游走的网络表示学习算法[J]. 计算机应用，2019，39(3)：651-655.

[6] 李同琴，王瑾，谭艳丽，等. 基于图像相似度的多权重图谱 DTI 自动分割算法研究[J]. 电子测量技术，2020，43(6)：116-122.

[7] 张宾，高亚斌，孙喜民，等. 知识图谱在计算思维聚类和多尺度分析中的应用及热点预测[J]. 微型电脑应用，2020，36(3)：123-125.

[8] 陈春谋. 基于流量阈值裁决分割机制的 WSN 网络抗 DDoS 算法研究[J]. 国外电子测量技术，2020，39(1)：59-62.

[9] ERMI B, ACAR E, CEMGIAL A T. Link prediction in heterogeneous data via generalized coupled tensor factorization[J]. Data Mining & Knowledge Discovery, 2015, 29(1): 203-236.

[10] BEYZA E, ACAR E, CEMGIL A T. Link prediction in heterogeneous data via generalized coupled tensor factorization[J]. Data Mining & Knowledge Discovery, 2015, 12(3): 110-119.

[11] ZHOU M. Infinite edge partition models for overlapping community detection and link prediction[J]. Computer Science, 2015, 12(23): 89-96.

[12] BURGESS M, ADAR E. Link-prediction enhanced consensus clustering for complex networks[J]. PLoS ONE, 2015, 11(5): 30-41.

[13] LIU M L, ZHU D, JIA D, et al. Link prediction in knowledge graphs: A hierarchy-constrained approach[J]. IEEE Transactions on Big Data, 2018, 10(12): 21-29.

[14] KIM J, DESNIRE J. Formational bounds of link prediction in collaboration networks[J]. Scientometrics, 2019, 119(3): 233-241.

[15] CHEN X, ZHOU Z. ELLPMDA: Ensemble learning and link prediction for miRNA-disease association prediction[J]. RNA Biology, 2018, 23(12): 42-50.

作者简介

袁华兵，本科，工程师，主要研究方向为计算机网络技术等。

E-mail: qe3260@163.com