第 45 卷 第 12 期 计 算 机 工 程 2019 年 12 月
Vol. 45 No. 12 Computer Engineering December 2019

·人工智能及识别技术· 文章编号：1000-3428(2019)12-0160-06 文献标志码：A 中图分类号：TP311

基于网络表示学习的论文影响力预测算法

樊 玮，韩佳宁，张宇翔

（中国民航大学 计算机科学与技术学院，天津 300300）

摘 要： 基于图的随机游走算法在预测论文影响力时，仅利用学术网络的全局结构信息而未考虑局部结构信息，对预测准确率造成影响。针对该问题，提出一种基于异构学术网络表示学习和多变量随机游走的论文影响力预测算法。通过构建异构学术网络表示模型，将网络中的论文、作者和期刊/会议等不同类型的节点表征到同一个低维向量空间中，同时保留网络的局部结构信息，将节点的向量相似度应用于多变量随机游走方法，实现对论文影响力的准确预测。在 AMiner 网站公开数据集上的实验结果表明，相比于 PageRank、FutureRank 等算法，该算法的预测准确性较高。

关键词： 网络表示学习；影响力预测；异构学术网络；多变量随机游走；局部结构信息

开放科学（资源服务）标志码（OSID）：

中文引用格式： 樊玮，韩佳宁，张宇翔. 基于网络表示学习的论文影响力预测算法[J]. 计算机工程，2019，45(12)：160-165，170.

英文引用格式： FAN Wei，HAN Jianing，ZHANG Yuxiang. Paper influence prediction algorithm based on network representation learning[J]. Computer Engineering，2019，45(12)：160-165，170.

Paper Influence Prediction Algorithm Based on Network Representation Learning

FAN Wei，HAN Jianing，ZHANG Yuxiang

（College of Computer Science and Technology，Civil Aviation University of China，Tianjin 300300，China）

【Abstract】 The graph-based random walk algorithm for paper influence prediction exploits only global structural information of academic network，and local structural information is usually ignored，which influence the prediction accuracy. To address the problem，this paper proposes a paper influence prediction algorithm based on heterogeneous academic network representation learning and multivariate random walk. By constructing a heterogeneous academic network representation model，different kinds of nodes of paper，authors，and journals/conferences in the network are represented into the same low-dimensional vector space，and the local structural information of network is kept. Similarity between vectors is applied to multivariate random walk to implement the accurate paper influence prediction. Experimental results on public datasets of AMiner Website show that the proposed method is more accurate in prediction than PageRank，FutureRank and other algorithms.

【Key words】 network representation learning；influence prediction；heterogeneous academic network；multivariate random walk；local structural information

DOI： 10.19678/j.issn.1000-3428.0053395

0 概述

学术论文作为研究人员展示其科研成果的重要载体，其数量随着科学技术的发展呈现快速增长。如何从海量的学术论文中准确识别出具有参考价值的文献变得越来越具有挑战性。论文影响力评估作为科学计量学中的一项重要研究，也引起了广泛的关注[1]。

论文影响力评估主要包括对论文当前影响力的评估和对论文未来影响力的预测两个方面。论文影响力预测主要利用论文的被引用次数、出版论文的期刊或会议的影响力以及作者的影响力等相关信息构建评估模型，从而预测论文的未来影响力[2]。相

基金项目： 国家自然科学基金（U1333109，U1533104）；中央高校基本科研业务费专项资金（ZXH2012P009）。

作者简介： 樊 玮（1968—），男，教授、博士，主研方向为智能信息处理、决策支持系统开发与应用；韩佳宁，硕士研究生；张宇翔，副教授、博士。

收稿日期： 2018-12-13 **修回日期：** 2019-01-24 **E-mail：** wfancauc@163.com

比于论文当前影响力评估,论文影响力预测对于研究者而言具有更重要的实际应用价值。准确评估论文影响力有助于研究人员快速识别具有潜在高影响力的论文,了解最新的研究动态,掌握最新的研究方法,从而在已有研究的基础上进一步改进或创新,以取得更先进的科研成果。然而,目前对于如何准确地识别有价值的论文并预测其未来影响力的研究相对较少。

现有方法通常是利用基于图的模型实现论文未来影响力的预测和排序,PageRank 算法[3]是其中具有代表性的算法之一,其本质上是一种基于单变量马尔可夫链的随机游走算法。在早期研究中,基于论文间引用关系的 PageRank 算法被广泛用于论文影响力的评估,但是这种基于单变量的算法只适用于同构网络,其因信息单一而难以获得较好的预测结果。

为了解决这一问题,研究人员在论文引用网络的基础上,加入作者、期刊等实体构建异构学术网络[4-5],提出基于多变量马尔可夫链的随机游走算法,并通过加入时间信息来提高预测结果的准确性。该算法的关键是假设各实体的影响力是相互增强的,例如,如果一篇学术论文是由具有高影响力的作者撰写的,那么该论文可能会具有较高的影响力,如果一位作者发表过高影响力的论文,那么该作者的影响力也更高。该方法的预测效果比单变量随机游走更好,其被认为是目前最先进的论文影响力预测方法之一。但是,基于图的随机游走算法只利用网络的全局结构信息,忽略了局部结构信息,而这种局部结构信息对论文影响力的评估非常重要,因此,可通过保留邻居节点间的影响力依赖关系[6]更全面地评估论文节点的影响力。

近年来,随着网络表示学习技术的发展[7-8],网络局部结构信息可以通过表示学习方法得到有效保留。受此启发,本文基于网络表示学习方法,提出一种新的论文影响力预测算法 NERank。设计融合论文引用网络、论文-作者网络以及论文-期刊/会议网络的异构学术网络,根据指示论文未来影响力的时间信息设置网络中边的权重。在此基础上,构建面向论文影响力预测的异构学术网络表示模型,将所有节点映射到一个统一的低维向量空间中,通过刻画节点间的分布关系保留网络的局部结构信息。利用低维向量空间中节点之间的相似性重新定义论文引用网络、论文-作者网络以及论文-期刊网络,并采用基于多变量随机游走的算法计算论文的未来影响力。

1　相关工作

1.1　论文影响力评估与预测

现有的论文影响力评估与预测方法大多通过构建基于图的模型,即在论文引用网络或相关扩展网络上直接利用随机游走方法实现。文献[9]结合论文引用网络与作者合作网络的网内随机游走和网间随机游走来对论文和作者影响力共同排序。文献[4]通过在包含论文、作者、期刊等实体的异构网络上随机游走的方法同时对论文、作者和期刊进行排名。文献[10]基于不同类型实体间的多维关系提出异构学术超网的概念,利用论文及作者之间的相互强化对论文进行排序。文献[11]基于新论文较旧论文可能获得更多引用的假设,利用论文引用关系、作者和论文发表时间等信息计算论文的未来影响力。文献[12]提出的 MRCoRank 算法是目前较先进的论文影响力预测算法,该算法将时间及文本信息加入异构学术网络中,基于论文、作者、期刊和文本特征影响力之间的互增强关系提出影响力排序模型。然而,这些基于图的方法直接利用边的权重构造实体间的关系矩阵,仅保留了网络的全局结构信息,而忽略了局部结构信息对节点影响力评估的重要作用。

1.2　网络表示学习

网络表示学习旨在获得网络中每个节点的向量表示,并有效地保留网络结构和属性[13-14]。近年来,网络表示学习引起了研究人员的极大关注,DeepWalk 算法[15]将深度学习技术应用于网络表示学习,文献[7]提出适用于大规模有向带权图的 Line 算法。此后,PTE[16]、TriDNR[17]、EOE[18]、methpath2vec[19]等模型相继被提出,并应用于各种数据挖掘和机器学习任务,如节点分类、节点聚类、链路预测、社区发现和可视化等。但是,尚未出现将上述方法应用于论文影响力预测的研究。

2　NERank 方法

本文提出一种基于网络表示学习的论文影响力预测方法,其主要包含以下 3 个关键步骤:1)构建一个包含论文、作者、期刊/会议 3 种类型实体的异构学术网络,根据节点间关系的建立时间来定义边的权重,以指示节点的未来影响力;2)提出一种面向论文影响力预测的异构学术网络表示模型,将不同类型的节点表征在同一个低维向量空间中,以指示节点在网络中的潜在影响力;3)将学习到的节点向量应用于多变量随机游走方法中,实现对论文未来影响力的预测。

2.1　异构学术网络构建

定义 1(论文引用网络)　令 P 为论文节点的集合,E_{pp} 为论文之间引用关系的边集合,W_{pp} 为 E_{pp} 中边的权重集合,则论文引用网络定义为:$G_{pp}=(P,E_{pp},W_{pp})$。

定义 2(论文-作者网络)　令 A 为作者节点的集合,E_{pa} 为论文与作者间写作关系的边集合,W_{pa} 为 E_{pa} 中边的权重集合,则论文-作者网络定义为:$G_{pa}=(P\cup A,E_{pa},W_{pa})$。

定义 3(论文-期刊/会议网络) 令 V 为期刊/会议节点的集合,E_{pv} 表示论文与期刊/会议间发表关系的边集合,W_{pv} 为 E_{pv} 中边的权重集合,则论文-期刊/会议网络定义为:$G_{pv}=(P\cup V,E_{pv},W_{pv})$。

将上述 3 个网络联结起来,即可得到如图 1 所示的异构学术网络,其定义如下:

定义 4(异构学术网络) 令 $N=\{P\cup A\cup V\}$、$E=\{E_{pp}\cup E_{pa}\cup E_{pv}\}$、$W=\{W_{pp}\cup W_{pa}\cup W_{pv}\}$ 分别表示论文引用网络、论文-作者网络、论文-期刊/会议网络的节点、边以及边的权重的集合,则异构学术网络定义为 $G=(N,E,W)$。

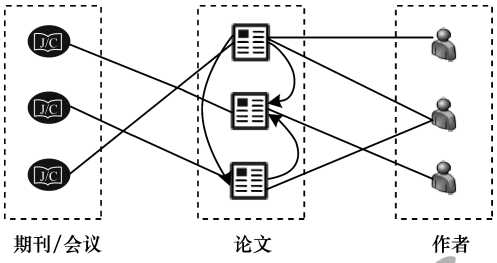

图 1 异构学术网络

2.2 异构学术网络表示学习

2.2.1 模型构建

异构学术网络以论文节点为纽带,由论文引用网络 G_{pp}、论文-作者网络 G_{pa} 和论文-期刊/会议网络 G_{pv} 这 3 个不同类型的网络组成。异构学术网络表示模型旨在将学术网络中 3 种不同类型的节点表征在同一个向量空间中,并为每个节点学习一个实数向量,该向量的维度远小于节点个数,最直观的方式就是将 3 个网络联合起来学习。通过刻画节点在向量空间中的分布与节点在网络中的分布之间的关系,保留网络的局部结构信息。令 $\hat{\mathbb{P}}(p\cdot|p_i)$、$\hat{\mathbb{P}}(a\cdot|p_i)$ 和 $\hat{\mathbb{P}}(v\cdot|p_i)$ 分别表示论文 p_i 在网络 G_{pp}、G_{pa} 和 G_{pv} 中的经验分布,令 $\mathbb{P}(p\cdot|p_i)$、$\mathbb{P}(a\cdot|p_i)$ 和 $\mathbb{P}(v\cdot|p_i)$ 分别表示 p_i 在低维向量空间中相应的条件概率分布。用 KL 散度来刻画概率分布和经验分布间的距离,并使 2 种分布之间的 KL 距离最小化,简化一些常量后的目标函数如式(1)所示。

$$\mathcal{L}=\mathcal{L}_{pp}+\mathcal{L}_{pa}+\mathcal{L}_{pv} \tag{1}$$

其中:

$$\mathcal{L}_{pp}=-\lambda_{pp}^{i}\hat{\mathbb{P}}(p_j|p_i)\ln\mathbb{P}(p_j|p_i) \tag{2}$$

$$\mathcal{L}_{pa}=-\lambda_{pa}^{i}\hat{\mathbb{P}}(a_j|p_i)\ln\mathbb{P}(a_j|p_i) \tag{3}$$

$$\mathcal{L}_{pv}=-\lambda_{pv}^{i}\hat{\mathbb{P}}(v_j|p_i)\ln\mathbb{P}(v_j|p_i) \tag{4}$$

在式(2)中引入 λ_{pp}^{i},表示论文 p_i 在 G_{pp} 中的影响力,本文定义 $\lambda_{pp}^{i}=\sum_{k\in R_{\text{in}}(p_i)}\omega_{k_i}$,即节点 p_i 的带权入度,$R_{\text{in}}(p_i)$ 表示引用了 p_i 的论文集合。$\omega_{k_i}\in E_{pp}$ 表示边 $p_k\rightarrow p_i$ 的权重。本文定义经验分布 $\hat{\mathbb{P}}(p_j|p_i)=\frac{\omega_{ij}}{\sum_{k\in R_{\text{out}}(p_i)}\omega_{ik}}$,$R_{\text{out}}(p_i)$ 表示 p_i 的参考论文集合。考虑到不同时期的引用对论文未来影响力的作用不同,本文基于论文间的引用关系建立时间来分配不同的权重,利用时间衰减函数对新的引用关系赋予更高的权重。定义 $\omega_{ij}=\mathrm{e}^{-\rho(T_c-T_{i\rightarrow j})}$,其中,$\rho$ 是衰减参数,本文设置为 2,T_c 表示当前时间,$T_{i\rightarrow j}$ 表示论文 p_i 与论文 p_j 的引用关系发生时间,即论文 p_i 的发表时间。$\mathbb{P}(p_j|p_i)$ 可由式(5)所示的 softmax 函数得到。

$$\mathbb{P}(p_j|p_i)=\frac{\exp(\boldsymbol{p}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i)}{\sum_{j\in V_B}\exp(\boldsymbol{p}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i)} \tag{5}$$

其中,$\boldsymbol{p}_i\in\mathbb{R}^d$ 是节点 p_i 的 d 维向量表示,$\boldsymbol{p}_j\in\mathbb{R}^d$ 是节点 p_j 的 d 维向量表示。

在式(3)中,λ_{pa}^{i} 表示论文 p_i 在 G_{pa} 中的影响力,$\lambda_{pa}^{i}=\sum_{k\in R(p_i)}\omega_{ik}$,$R(p_i)$ 表示论文 p_i 的作者集合,$\omega_{ik}\in E_{pa}$ 表示边 $p_i\rightarrow a_k$ 的权重,本文设置为 1。定义经验分布 $\hat{\mathbb{P}}(a_j|p_i)=\frac{\omega_{ij}}{\sum_{k\in R(p_i)}\omega_{ik}}$,而 $\mathbb{P}(a_j|p_i)$ 由式(5)得到。

在式(4)中,λ_{pv}^{i} 表示论文 p_i 在 G_{pv} 中的影响力,$\lambda_{pv}^{i}=\omega_{ik}$,$\omega_{ik}\in E_{pv}$ 表示边 $p_i\rightarrow v_k$ 的权重。本文定义 $\omega_{ik}=\mathrm{e}^{-\rho(T_c-T_{\text{pub}})}$,$T_{\text{pub}}$ 表示论文发表时间。因为版权问题,一篇论文只能被一个期刊或会议发表,所以设置经验分布 $\hat{\mathbb{P}}(v_j|p_i)=1$。$\mathbb{P}(v_j|p_i)$ 同样可利用如式(5)所示的 softmax 函数得到。

2.2.2 模型优化

对于大规模网络而言,直接计算 softmax 函数是不切实际的,一般采用负采样加快训练速度。本文利用负采样技术概率函数得到如下公式:

$$\mathbb{P}(p_j|p_i)=\sigma(\boldsymbol{p}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i)+\sum_{m=1}^{K}(1-\sigma(\boldsymbol{p}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i)) \tag{6}$$

其中,$\sigma(x)=1/(1+\exp(-x))$ 为 sigmoid 函数,K 为负采样比率,即对每一条边随机采样 K 条负边,得到 K 对不相连的节点。此外,采用 L2 范式正则项以避免过拟合,由于空间限制,将目标函数 $\mathcal{L}$ 合并同类项后得到如下公式:

$$\begin{aligned}\mathcal{L}=&\sum_{(i,j)\in E}\lambda_{pn}^{i}\hat{\mathbb{P}}(n_j|p_i)\ln(1+\exp\{-\boldsymbol{n}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\})+\\&\sum_{(i,j)\notin E}\lambda_{pn}^{i}\ln(1+\exp\{\boldsymbol{n}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\})+\gamma_p\sum_{k=1}^{|P|}\|\boldsymbol{p}_k\|_2+\\&\gamma_a\sum_{k=1}^{|A|}\|\boldsymbol{a}_k\|_2+\gamma_v\sum_{k=1}^{|V|}\|\boldsymbol{v}_k\|_2\end{aligned} \tag{7}$$

其中,γ_p、γ_a 和 γ_v 是正则化系数,$(i,j)\notin E$ 表示负采样边的集合。对应 E 中 3 种不同类型的边时,n_j 分别代表论文 p_j,作者 a_j 和期刊/会议 v_j 这 3 种类型的节点。

本文使用梯度下降算法最小化目标函数 $\mathcal{L}$,并利用回溯线搜索方法在迭代过程中调整步长。$\mathcal{L}$ 对 $\boldsymbol{p}_i$、$\boldsymbol{a}_j$ 和 $\boldsymbol{v}_j$ 求导的过程如式(8)~式(10)所示。

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{p}_i} = -\sum_{(i,j)\in E}\frac{\lambda^i_{pn}\hat{\mathbb{P}}(n_j\mid p_i)\exp\{-\boldsymbol{n}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\}}{1+\exp\{-\boldsymbol{n}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\}}\boldsymbol{n}_j + \sum_{(i,j)\notin E}\frac{\lambda^i_{pn}}{1+\exp\{-\boldsymbol{n}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\}}\boldsymbol{n}_j + \gamma_p\sum_{d=1}^{D_p}2\boldsymbol{p}_i^d \quad (8)$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}_j} = -\sum_{(i,j)\in E_{pa}}\frac{\lambda^i_{pa}\hat{\mathbb{P}}(a_j\mid p_i)\exp\{-\boldsymbol{a}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\}}{1+\exp\{-\boldsymbol{a}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\}}\boldsymbol{p}_i + \sum_{(i,j)\notin E_{pa}}\frac{\lambda^i_{pa}}{1+\exp\{-\boldsymbol{a}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\}}\boldsymbol{p}_i + \gamma_a\sum_{d=1}^{D_a}2\boldsymbol{a}_j^d \quad (9)$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_j} = -\sum_{(i,j)\in E_{pv}}\frac{\lambda^i_{pv}\hat{\mathbb{P}}(v_j\mid p_i)\exp\{-\boldsymbol{v}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\}}{1+\exp\{-\boldsymbol{v}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\}}\boldsymbol{p}_i + \sum_{(i,j)\notin E_{pv}}\frac{\lambda^i_{pv}}{1+\exp\{-\boldsymbol{v}_j^{\mathrm{T}}\cdot\boldsymbol{p}_i\}}\boldsymbol{p}_i + \gamma_v\sum_{d=1}^{D_v}2\boldsymbol{v}_j^d \quad (10)$$

其中,D_p、D_a 和 D_v 分别表示论文、作者和期刊/会议节点向量表示的维度,本文设置为 $D_p = D_a = D_v = d$。

将本文异构学术网络表示学习算法命名为 HANE,其具体描述如算法 1 所示。

算法 1　HANE 算法

输入　网络 $G(N,E,W)$,向量维度 d,正则项系数 γ_p、γ_a 和 γ_v,负采样比率 K,初始步长 η

输出　论文、作者及期刊/会议的 d 维向量表示

1. 初始化论文、作者及期刊/会议的向量表示。
2. while (目标函数不收敛) do
3. 基于式(8)计算论文的梯度;
4. 利用回溯线搜索更新步长 η;
5. 更新论文的向量表示;
6. 利用式(8)学习作者的向量表示,过程与步骤 2 ~ 步骤 5 类似。

2.3　基于多变量随机游走的论文影响力排序

通过上述的异构学术网络表示学习过程,3 种类型节点的未来影响力被表征在一个统一的低维向量空间中,通过保留网络的局部结构信息使得具有相似潜在影响力的节点在低维向量空间中彼此接近。基于学习到的论文向量表示,节点 p_i 和节点 p_j 之间的相似度可通过式(11)来计算。

$$Sim(p_i,p_j) = \frac{\boldsymbol{p}_i\cdot\boldsymbol{p}_j}{\|\boldsymbol{p}_i\|\times\|\boldsymbol{p}_j\|} \quad (11)$$

同理,可以计算得到论文 p_i 与作者 a_j 间的相似度 $Sim(p_i,a_j)$ 以及论文 p_i 与期刊/会议 v_j 间的相似度 $Sim(p_i,v_j)$。因此,论文引用网络可以重新定义为如下形式:

$$\boldsymbol{M}_{pp}(j,i) = \begin{cases}\dfrac{Sim(p_i,p_j)}{\sum_{p_k\in R(p_i)}Sim(p_i,p_k)}, e_{p_ip_j}\in E_{pp}\\ 0,\text{其他}\end{cases} \quad (12)$$

其中,$e_{p_ip_j}$表示连接 2 个节点的边,$R(p_i)$ 表示 p_i 的邻居节点集合。同理,可以得到表示论文-作者网络 G_{pa}的转移矩阵 $\boldsymbol{M}_{pa}$与 $\boldsymbol{M}_{ap}$,以及表示论文-期刊/会议网络 G_{pv}的转移矩阵 $\boldsymbol{M}_{pv}$与 $\boldsymbol{M}_{vp}$。

得到上述转移矩阵后,进一步提出基于多变量马尔可夫链的随机游走模型,论文、作者和期刊/会议的影响力在式(13) ~ 式(15)的迭代过程中相互影响和增强,最终得到论文的影响力指数。

$$\boldsymbol{P}^{(t+1)} = \alpha\boldsymbol{M}_{pp}\boldsymbol{P}^{(t)} + \beta\boldsymbol{M}_{pa}\boldsymbol{A}^{(t)} + \gamma\boldsymbol{M}_{pv}\boldsymbol{V}^{(t)} \quad (13)$$

$$\boldsymbol{A}^{(t+1)} = \boldsymbol{M}_{ap}\boldsymbol{P}^{(t)} \quad (14)$$

$$\boldsymbol{V}^{(t+1)} = \boldsymbol{M}_{vp}\boldsymbol{P}^{(t)} \quad (15)$$

其中,$\boldsymbol{P}^{(t)}$、$\boldsymbol{A}^{(t)}$和 $\boldsymbol{V}^{(t)}$分别表示论文、作者和期刊/会议在第 t 次迭代后的概率分布向量,向量中的值代表其相应的未来影响力。参数 α、β 和 γ 分别控制论文未来影响力受其他论文、作者和期刊/会议影响的权重,$\alpha+\beta+\gamma=1$。最后,通过迭代执行式(13) ~ 式(15)直至收敛,将收敛后的 $\boldsymbol{P}$ 由高到低排序,即可得到论文的未来影响力排序。

3　实验结果与分析

3.1　实验数据

本文实验采用 AMiner 公开发表的 Academic Social Network 数据集,包括发表于 2014 年前的 2 092 356 篇论文、1 712 433 位作者以及 8 024 869 条引用关系[20]。对数据集进行如下预处理:1)本文侧重于预测较新论文的未来影响力,因此仅保留发表于 1998 年以后的论文数据;2)去除信息不完整的论文数据,例如缺失作者信息或发表时间的论文;3)去除没有任何参考文献或未被其他论文引用的数据。经过数据预处理后,得到 328 971 篇论文和2 732 340 条论文引用关系。

3.2　评估指标

由于论文影响力预测缺少统一的标准,因此评估实验结果具有一定的挑战性。本文采用文献[11]的方法,将论文的未来引用数(Future Citations,FC)作为论文未来影响力的评估标准。具体来讲,本文将发表于 2009 年之前的论文作为训练数据,计算其未来影响力并进行排序,然后根据这些论文在 2010 年—2014 年间的引用次数对排序结果进行评估。经过划分后的训练数据中包含 127 711 篇论文、558 138 条引用关系、383 276 条论文-作者的写作关系以及 76 432 条论文-期刊/会议的所属关系。

为了评估本文 NERank 算法的性能,本文采用以下 2 个评估指标:1)机器学习中常用的准确率(Precision)指标;2)考虑到返回的论文影响力预测结果存在顺序,采用文献[4,12]使用的推荐强度(Recommendation Intensity,RI)作为评估各算法的指标。令 R 表示任一算法返回的 Top-k 个论文序列,L 表示评估标准 FC 返回的Top-k 个论文序列,论文 p_i 在 R 中的顺序记为 o_r,则 p_i 的 RI 值可定义为:

$$RI(p_i)@k = \begin{cases}1+\dfrac{k-o_r}{k}, p_i\in L\\ 0, p_i\notin L\end{cases} \quad (16)$$

式(16)表示如果论文 p_i 在 L 中且其在 R 中的排名越靠前,它的 RI 值越高。基于 R 中每篇论文的 RI 值,Top-k 个序列 R 的 RI 值可定义为:

$$RI(R)@k=\sum_{p_i\in R}RI(p_i)@k$$

显然,$RI(R)@k$ 值越大表示论文排序结果越好。可以看出,如果将 R 视为无序序列,用 $RI(R)@k$ 除以 k,即可得到准确率指标,为方便表述,下文将 $RI(R)@k$ 简写为 $RI@k$。

3.3 对比方法

为了评估 NERank 算法的预测性能,将以下5 种论文影响力预测算法与本文算法进行对比:

1)PageRank 算法[2]:在论文引用网络中通过 PageRank 算法对论文影响力进行排序,其中边的权重设置同本文的 W_{pp}。

2)FutureRank 算法[10]:该算法是预测未来论文影响力排名的代表性模型,其结合引文、作者和论文发表时间等信息来对论文的未来影响力进行预测。

3)HHGBiRank 算法[9]:利用论文和作者间的多维关系,构建基于异构学术超网的模型对论文影响力进行排序。

4)MRCoRank 算法[11]:将论文、作者、期刊、文本特征集成到统一的框架中,通过一种互增强模型对论文、作者等实体的未来影响力进行排序。

此外,本文还将 NERank 与其变种方法 NERank-NE(不包含网络表示学习过程)进行对比。

3.4 参数设置

本文实验参数设置如下:

1)向量维度设置。在网络表示学习模型中,正则项系数 $\gamma_p=\gamma_a=\gamma_v=1$,负采样比率 $K=5$。在实验中,将向量维度 d 设置不同的取值,以 Top-20 的论文为例,计算 d 对 NERank 算法的准确率和 RI 值的影响,结果如表 1 所示。可以看出,$d=100$ 是一个较合理的选择,因此在以下实验中将 100 作为 d 的默认取值。

表 1 d 取不同值时的检测结果对比

d 的取值	准确率	$RI@20$
20	0.30	9.40
30	0.30	10.15
50	0.30	10.50
64	0.35	11.20
100	0.40	13.45
128	0.40	12.95
150	0.35	12.25
200	0.35	11.35

2)式(13)中的调节参数设置。通过固定 γ 求 α 和 β,固定 α 求 β 和 γ 的方法观察不同取值对 RI 值的影响,其中,参数 α 对 RI 值的影响如图 2 所示。可以看出,$\alpha=0.6$ 时实验效果较好,此时,$\beta=0.25$,$\gamma=0.15$。

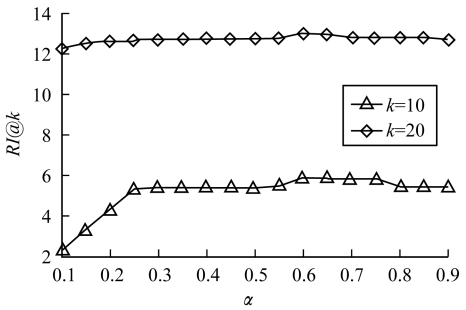

图 2 参数 α 对 RI 值的影响

3.5 结果分析

本文首先根据 6 种算法返回的 Top-100 的论文,计算其准确率和 RI 值,结果如图 3 所示。

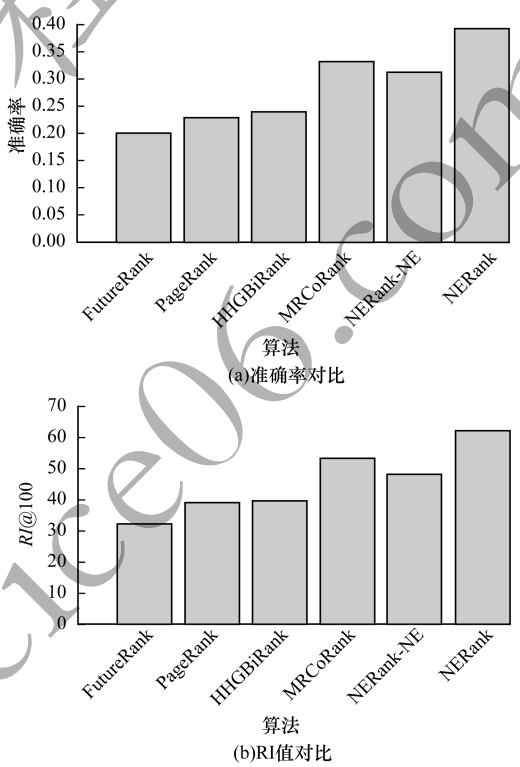

图 3 6 种算法对 Top-100 论文的预测结果对比

由图 3 可知,MRCoRank 算法优于 PageRank、FutureRank、HHGBiRank 和 NERank-NE 这 4 种算法。与 MRCoRank 相比,NERank 算法的准确率提升 6.0%,RI 值提升 16.5%,这是因为 NERank 算法采用网络表示学习结合多变量随机游走的方法,能够更有效地利用学术网络的局部结构和全局结构信息,改善预测性能。

通过数据统计可以发现,不同领域的论文获得的引用数据差别较大,例如,按照未来引用数排序得到的 Top-100 的论文中,39 篇属于人工智能领域,而仅有 3 篇属于网络安全领域。一般情况下,研究人员更关注与自己研究领域相关的文献,因此,为了使评价结果更公正且更具实际价值,本文进一步将论文按照领域分类并进行比较。本文给出 6 种算法对于人工智能、数据库以及网络安全 3 个领域的论文预测结果对比,如表 2 所示,最优结果加粗表示。

表 2　6 种算法对于不同领域论文的 RI 值对比

算法	人工智能				数据库				网络安全			
	RI@20	*RI*@30	*RI*@50	*RI*@100	*RI*@20	*RI*@30	*RI*@50	*RI*@100	*RI*@20	*RI*@30	*RI*@50	*RI*@100
PageRank 算法	6.75	11.40	18.52	52.90	6.05	12.83	14.10	49.78	7.15	11.30	23.16	44.56
FutureRank 算法	12.99	18.63	24.56	64.17	10.95	17.63	26.14	56.05	11.50	15.77	26.70	49.58
HHGBiRank 算法	9.00	14.46	20.74	57.81	8.10	12.37	19.40	45.12	10.15	14.53	20.54	44.63
MRCoRank 算法	**13.65**	17.93	29.40	71.33	12.70	22.33	28.86	69.23	11.05	16.30	26.18	55.36
NERank-NE 算法	10.55	14.23	20.54	67.67	5.45	13.73	18.32	54.25	3.55	4.83	9.12	47.01
NERank 算法	13.45	**20.17**	**33.26**	**100.76**	**14.25**	**22.80**	**28.90**	**83.71**	**12.85**	**16.63**	**30.40**	**79.36**

由表 2 可知,各算法针对不同领域论文的影响力预测效果存在差异。以 Top-50 论文的预测结果为例,相比于 MRCoRank 算法,NERank 算法对人工智能、数据库、网络安全 3 个领域论文的 RI 值分别可提高 13.12%、0.14% 和 16.12%。从整体上看,NERank 算法、MRCoRank 算法和 FutureRank 算法结果优于 PageRank 算法和 HHGBiRank 算法,这是因为 PageRank 算法与 HHGBiRank 算法未考虑时间信息和期刊信息的影响,说明加入时间信息与期刊信息可改善论文影响力预测的准确性。NERank 算法优于其他对比算法,说明结合网络表示学习和多变量随机游走的方法可取得更精确的预测结果。上述结果表明,通过网络表示学习模型学习各节点在异构学术网络上的潜在影响力之间的相似性,并结合常用的多变量随机游走模型,能更有效地保留学术网络的局部结构以及全局结构,有助于提升论文影响力预测的准确性。

4　结束语

为了提高论文影响力预测的准确性,本文提出网络表示学习与多变量随机游走方法相结合的论文影响力预测算法 NERank。建立一个包含论文、作者、期刊/会议 3 种实体的时间感知异构学术网络,根据各节点间关系建立的时间来设置边的权重,以指示各节点的未来影响力。在此基础上,建立一种面向论文影响力预测的异构学术网络表示模型,将不同类型的节点表示在同一向量空间中并保留网络局部结构,以指示其在网络中的潜在影响力。将节点之间的向量相似度应用于多变量随机游走的转移过程中,并计算论文的未来影响力。实验结果表明,该算法可有效提高论文影响力预测的准确性。下一步考虑将 NERank 算法推广至具有相似结构的其他异构网络中,或者将保留节点属性的异构网络表示学习模型应用于多类型实体的联合排序任务,拓展该算法的应用领域。

参考文献

[1] XIA Feng, WANG Wei, BEKELE T M, et al. Big scholarly data: a survey[J]. IEEE Transactions on Big Data, 2017, 3(1): 18-35.

[2] ZENG Wei. The research of literature value prediction algorithm and author impact assessment algorithm in academic network[D]. Chongqing: Southwest University, 2014. (in Chinese)
曾玮. 文献排名预测算法及作者影响力评估算法研究[D]. 重庆:西南大学,2014.

[3] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the Web[EB/OL]. [2018-12-01]. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

[4] JIANG Xiaorui, SUN Xiaoping, HAI Zhuge. Towards an effective and unbiased ranking of scientific literature through mutual reinforcement[C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2012: 714-723.

[5] NG K P, LI Xutao, YE Yunming. MultiRank: co-ranking for objects and relations in multi-relational data[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2011: 1217-1225.

[6] LU Guangdi. Teaching quality assessment of interval number comparison algorithm[J]. Journal of Shanxi Normal University (Natural Science Edition), 2017, 31(4): 19-24. (in Chinese)
陆广地. 教学质量多属性评估的联系数算法[J]. 山西师范大学学报(自然科学版),2017,31(4):19-24.

[7] TANG Jian, QU Meng, WANG Mingzhe, et al. LINE: large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. New York, USA: ACM Press, 2015: 1067-1077.

[8] CUI Peng, WANG Xiao, PEI Jian, et al. A survey on network embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(5): 833-852.

[9] ZHOU Ding, ORSHANSKIY S A, ZHA Hongyuan, et al. Co-ranking authors and documents in a heterogeneous network[C]//Proceedings of IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2007: 739-744.

[10] LIANG Ronghua, JIANG Xiaorui. Scientific ranking over heterogeneous academic hypernetwork[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AIAA Press, 2016: 20-26.

[11] SAYYADI H, GETOOR L. FutureRank: ranking scientific articles by predicting their future PageRank[C]//Proceedings of SIAM International Conference on Data Mining. [S.l.]: SIAM Publications, 2009: 533-544.

(下转第 170 页)

(上接第 165 页)

[12] WANG Senzhang, XIE Sihong, ZHANG Xiaoming, et al. Coranking the future influence of multiobjects in biblio-graphic network through mutual reinforcement[J]. ACM Transactions on Intelligent Systems and Technology,2016,7(4):1-28.

[13] TU Cunchao, YANG Cheng, LIU Zhiyuan, et al. Network representation learning: an overview[J]. SCIENTIA SINICA Informationis, 2017, 47(8): 980-996. (in Chinese)
涂存超,杨成,刘知远,等. 网络表示学习综述[J]. 中国科学:信息科学,2017,47(8):980-996.

[14] LIU Zhengming, MA Hong, LIU Shuxin, et al. A network representation learning algorithm fusing with textual attribute information of nodes[J]. Computer Engineering, 2018,44(11):165-171. (in Chinese)
刘正铭,马宏,刘树新,等. 一种融合节点文本属性信息的网络表示学习算法[J]. 计算机工程,2018,44(11):165-171.

[15] PEROZZI B, ALRFOU R, SKIENA S. DeepWalk: online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2014: 701-710.

[16] TANG Jian, QU Meng, MEI Qiaozhu. PTE: predictive text embedding through large-scale heterogeneous text networks[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2015: 1165-1174.

[17] PAN Shirui, WU Jia, ZHU Xingquan, et al. Tri-party deep network representation [C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2016: 1895-1901.

[18] XU Linchuan, WEI Xiaokai, CAO Jiannong, et al. Embedding of Embedding (EOE): joint embedding for coupled heterogeneous networks[C]//Proceedings of the 10th ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2017: 741-749.

[19] DONG Yuxiao, CHAWLA N V, SWAMI A. Metapath2vec: scalable representation learning for heterogeneous networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2017: 135-144.

[20] TANG Jie, ZHANG Jing, YAO Limin, et al. ArnetMiner: extraction and mining of academic social networks[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2008: 990-998.

编辑 樊丽娜