计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191000199

# 共引增强有向网络嵌入研究

吴 勇 王斌君 翟一鸣 仝 鑫
中国人民公安大学警务信息工程与网络安全学院 北京 100240
(wyppsuc@hotmail.com)

**摘 要** 网络嵌入旨在将网络节点嵌入到一个低维向量空间且最大程度地保存原有网络的拓扑结构及其属性。相比无向网络,有向网络具有特殊的非对称传递性,可体现在节点之间的高阶相似度量中,如何较好地保存这一特性是当前有向网络嵌入研究的热点和难点。针对此问题,通过引入有向网络的共引网络,设计了共引信息的度量函数,给出了一种有向网络高阶相似度量指标融合共引信息的统一框架,提出了可以保存非对称传递性的共引增强的高阶相似保存网络嵌入模型(Co-Citation Enhancing High-Order Proximity preserved Embedding,CCE-HOPE)。在4个真实数据集上进行链路预测实验的结果表明,不同高阶相似度量指标下,不同比重共引信息对效果影响具有一般规律,因此可以给出比重的最佳取值范围;在此范围内,与现有方法相比,CCE-HOPE方法可有效提高链接预测的准确度。

**关键词**:有向网络嵌入;非对称传递;共引网络;链路预测
**中图法分类号** TP391

## Study on Co-citation Enhancing Directed Network Embedding

WU Yong,WANG Bin-jun,ZHAI Yi-ming and TONG Xin
College of Police Information Engineering and Cyber Security,People's Public Security University of China,Beijing 100240,China

**Abstract** Network embedding algorithms embed a network into a low-dimensional vector space where the structure and the inherent properties of the graph can be preserved to the greatest extent. Compared with undirected networks,directed networks have special asymmetric transitivity which can be reflected in the high-order similarity measurement between nodes. A hot spot and difficulty of current directed network embedding research is how to preserve this feature well. Aiming at this problem,this paper introduces the co-citation network of directed networks and designs a metric function of the co-introduction information. At the same time,a unified framework is created for fusing the co-citation information and the high-order similarity metrics of directed networks. Then,this paper proposes a co-citation enhancing high-order proximity preserved embedding method,called CCE-HOPE,which can preserve the asymmetric transitivity well. In experiments,the proposed model is evaluated on link prediction using four real data sets. The results show that under different high-order similarity metrics,the performance of different proportions of co-introduction information follows a general regularity,so the optimal range of the proportion can be determined. Compared with other state-of-the-art methods,the method can effectively improve the accuracy of link prediction when the proportion of co-introduction information is within the optimal range.

**Keywords** Directed network embedding,Asymmetric transitivity,Co-citation network,Link prediction

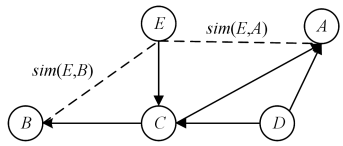
对网络进行分析的核心问题之一就是网络表示方式。采用原始的邻接矩阵直接表示网络,不仅是高维稀疏的,而且仅能够获取节点的邻居关系。而网络嵌入采用低维稠密向量表示网络中的节点,且保存节点的结构信息及其关系信息,能处理很多网络分析任务,如链接预测、节点分类、社区发现等。

目前,Deepwalk[1],LINE[2],Node2vec[3]等经典的网络嵌入方法适用于无向网络。相比无向网络,有向网络具有特殊的非对称传递性,直接影响着节点的相似度计算。对于有向网络的高阶相似度量方法主要有共同邻居相似性指标CN(Common Neighbors)、Katz指标、Rooted PageRank(RPR)指标等。文献[4]提出APP(Asymmetric Proximity Preserving)方法,对随机游走进行采样,嵌入向量能有效保存高阶相似Rooted PageRank(RPR)指标;文献[5]提出HOPE(High-Order Proximity preserved Embedding)方法,通过高阶相似度量的普适性研究,利用生成式SVD(Singular Value Decomposition)进行相似度矩阵分解得到嵌入向量,可有效保存多种高阶相似度量指标。

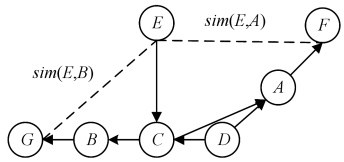
在上文提到的非对称相似保存算法中,高阶相似度量指

到稿日期:2019-10-30 返修日期:2020-06-28 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:公安部科技强警基础专项(2018GABJC03);中国人民公安大学拔尖人才培养专项资助研究生科研创新项目(2019bsky002)
This work was supported by the Science and Technology Strengthening Police Basic Program of Ministry of Public Security(2018GABJC03) and Top Talent Training Special Funding Graduate Research and Innovation Project of People's Public Security University of China(2019bsky002).
通信作者:王斌君(wangbinjun@ppsuc.edu.cn)

标建立在网络的原始拓扑之上，考虑了原始拓扑的邻域和高阶邻域，但没有充分考虑有向网络中的三角形结构，即有向网络中节点的共引信息。图 1 给出了两个社交信任网络（其中有向边表示信任关系），图 1(a)中 $E$ 经过 $C$ 与 $B$ 和 $A$ 相连，那么在嵌入空间中 $E$ 和 $B$ 的相似度 $sim(E,B)$ 与 $E$ 和 $A$ 的相似度 $sim(E,A)$ 是否相等呢？这里我们做出一个假设：$C$ 信任 $A$，同时 $D$ 既信任 $C$ 又信任 $A$，那么我们认为 $D$ 的信任关系可以对 $C$ 和 $A$ 的信任关系进行增强，也就是说，$sim(E,A)>sim(E,B)$，从而认为 $E$ 和 $A$ 连接的概率大于 $E$ 和 $B$ 连接的概率。如果改变方向，$C$ 和 $A$ 同时指向 $D$，同样 $D$ 可以增强 $C$ 对 $A$ 的信任关系。图 1(b)展示了三阶的情况，即 $E$ 经过 3 条边到达 $F$ 和 $G$。同样，我们认为 $C$ 和 $A$ 的信任关系得到了增强，从而可以得出 $sim(E,F)>sim(E,G)$。更高阶的情况同理。分析图 1 所示的网络结构可以发现，我们实质上是对网络中三角形 $(A,C,D)$ 的共引边进行增强，为了验证我们的假设，将在后文进行链接预测的对比实验。




(a)二阶




(b)三阶

图 1 两个社交信任网络

Fig. 1 Two social trust network

如何在嵌入空间中保存共引信息是一个关键问题，本文尝试将这个共引信息转换为网络权重并提出了对于有向网络的共引增强的高阶相似保存网络嵌入模型（CCE-HOPE）。首先构建原始网络的共引网络，并将此信息与原网络结构进行融合，进而改变高阶相似矩阵，然后通过高阶相似矩阵设计损失函数，并利用多核 GPU 版本的 NMF-mGPU 方法进行优化。

本文的主要贡献有：

1）根据有向网络的共引属性，设计了一种融合有向网络高阶相似度量指标和共引信息的统一框架。

2）提出了共引增强高阶相似保存网络嵌入模型，通过多核 GPU 版本的 NMF-mGPU 方法对损失函数进行优化，以有效保存有向网络的非对称传递特性、共引信息及高阶相似信息。

3）通过实验，对共引网络影响比重进行了探究，发现了其变化的一般规律，找出了比较合适的取值范围。在此基础上将所提算法与相关算法进行对比实验，验证了所提方法在链接预测任务上的优势。

## 1 相关研究

### 1.1 网络嵌入

网络嵌入是一种降维方法。对于不同形式的网络和应用场景，需要嵌入空间保存网络的结构[6]、标签[7]、符号[8]等信息。无向网络的网络嵌入方法大致分为 3 种：基于矩阵分解的方法[9-11]、基于随机游走的方法[1-3]及基于深度学习的方法[12-14]。其中，有代表性的 Deepwalk[1]通过在网络上进行截断随机游走进行节点序列的采样，并利用语言模型 Skip-Gram 进行训练，通过最大化一定窗口内和当前节点共同出现的其他节点出现的概率来学习节点表示向量。LINE[2]分别为网络一阶相似和二阶相似设计目标函数，通过优化目标函数分别学习节点的一阶和二阶向量表示，最后将两者合并作为最终的表示向量。这些方法若直接应用到有向网络则均无法保存有向网络的非对称性。

针对有向网络的非对称性保存的网络嵌入，文献[4]提出了 APP 方法，其基于 Monte-Carlo 采样方法，针对 Rooted PageRank (RPR)相似指标对有向网络的随机游走进行采样，然后利用基于负采样的 Skip-Gram[15]模型进行训练，并采用有向的梯度更新来有效保存高阶相似 Rooted PageRank (RPR)指标。文献[5]提出了 Katz 指标、Rooted PageRank (RPR)指标、共同邻居相似性指标 CN (Common Neighbors)和 Adamic-Adar 相似性指标(AA)等多种相似度量作为指标的统一形式，利用生成式矩阵分解学习节点的原节点表示向量和目标节点表示向量，然后将两者合并作为节点的最终表示向量。

### 1.2 高阶相似

基于网络拓扑信息的高阶相似算法大致可以分为如下 3 类：基于邻居的相似性指标、基于路径的相似性指标和基于随机游走过程的相似性指标。

基于邻居信息的相似性指标主要包括共同邻居相似性指标、Adamic-Adar 相似性指标、资源分配相似性指标、Jaccard 指标、大度节点有利指标、大度节点不利指标、优先链接指标等。邻居信息直接反映了节点的关系情况，其复杂度相对较低且计算简单；基于路径的相似性指标主要包括 Katz 指标、局部路径指标、连接增强相似指标、LHN-II 指标等。其将邻居信息扩大，考虑更大范围的拓扑信息，复杂度高于基于邻居信息的方法；基于随机游走过程的相似性指标主要包括平均通勤时间、基于随机游走的余弦相似性指标、SimRank 指标、局部随机游走的相似性指标、有叠加效应的随机游走相似性指标和 Rooted PageRank 等。基于随机游走的方法大多是基于网络的全局信息，是一种相对灵活的探索网络结构的方式。另外，还存在一些其他的高阶相似算法，如矩阵森林指数和自洽转移相似性指标等。

## 2 CCE-HOPE 模型

### 2.1 基本概念

在有向图中，共引信息的表示采用网络结构的形式。根据共引边方向的不同，共引又分为正向共引和反向共引两种形式。

设 $G=\langle V,E\rangle$ 是一个有向图，$V=\{v_i\mid i=1,\cdots,|V|\}$ 是节点的集合，$E=\{e_{ij}=(v_i,v_j)\mid v_i,v_j\in V\}$ 是有向边的集合。

**定义 1**（正向共引网络） $\forall a,b,c\in V$，如果 $\exists(b,a)\wedge(c,a)\in E$，则称无向边 $(b,c)$ 为正向共引的边，由这些无向边所构成的网络被称为正向共引网络。

**定义 2**（反向共引网络）　$\forall a,b,c\in V$，如果 $\exists(a,b)\wedge(a,c)\in E$，则称无向边 $(b,c)$ 为反向共引的边，由这些无向边所构成的网络被称为反向共引网络。

### 2.2　问题描述

本文的目的是学习一个将网络节点映射到一个低维向量空间的映射函数 $f:v_i\rightarrow h_i\in R^d$，其中 $d(d\ll|V|)$ 为低维向量的维度。在这个向量空间中，每个节点的向量表示可以保存原有的高阶相似信息和共引网络信息。本文使用的主要符号及定义如表 1 所列。

表 1　文中的主要符号及定义

Table 1　Main symbols and definitions

| 符号 | 描述 |
|---|---|
| $\|V\|$ | 节点数目 |
| $\|E\|$ | 边数目 |
| $\boldsymbol{A}$ | 网络邻接矩阵 |
| $\boldsymbol{A}_{\mathrm{r}}$ | 网络融合邻接矩阵 |
| $\boldsymbol{A}_{\mathrm{o}}$ | 正向共引网络矩阵 |
| $\boldsymbol{A}_{\mathrm{n}}$ | 反向共引网络矩阵 |
| $\boldsymbol{S}_{\mathrm{o}}$ | 正向共引相似矩阵 |
| $\boldsymbol{S}_{\mathrm{n}}$ | 反向共引相似矩阵 |
| $\boldsymbol{S}_{\mathrm{c}}$ | 共引网络相似矩阵 |
| $\boldsymbol{S}_{\mathrm{com}}$ | 共引相似过滤矩阵 |
| $d$ | 低维向量空间维度 |
| $\boldsymbol{H}$ | 嵌入向量 |
| $diag(\cdot)$ | 取对角矩阵 |

### 2.3　模型整体框架

CCE-HOPE 模型采用如下步骤获得具有非对称传递的有向网络嵌入表示。首先，计算正向共引网络矩阵和反向共引网络矩阵；然后根据节点度来调整共引网络权值，在共引网络中提取出原网络中具有的边，并与原网络进行融合构成赋有新权重的网络；最后利用多核 GPU 版本的 NMF-mGPU 方法优化损失函数。

#### 2.3.1　共引网络的计算

根据定义 1、定义 2 和网络邻接矩阵，正向共引网络矩阵和反向共引网络矩阵的计算式分别为：

$$\boldsymbol{A}_o=\boldsymbol{A}\cdot\boldsymbol{A}^{\mathrm{T}}-diag(\boldsymbol{A}\cdot\boldsymbol{A}^{\mathrm{T}}) \tag{1}$$

$$\boldsymbol{A}_n=\boldsymbol{A}^{\mathrm{T}}\cdot\boldsymbol{A}-diag(\boldsymbol{A}^{\mathrm{T}}\cdot\boldsymbol{A}) \tag{2}$$

#### 2.3.2　权值的调整

在图 2 所示的无权引文网络中，节点 $b$ 和 $c$ 的相似度与节点 $b'$ 和 $c'$ 的相似度是否相等呢？在实际情况中，如果一个节点的出度越高，则认为它对相连节点的影响力越低。在图 2 中，$a$ 对 $b$ 和 $c$ 的影响大于 $a'$ 对 $b'$ 和 $c'$ 的影响，因此 $b$ 和 $c$ 的相似度大于 $b'$ 和 $c'$ 的相似度。为反映这一特性，本文根据节点度对共引矩阵进行标准化。

结合式(1)和式(2)，正向共引相似矩阵和反向共引相似矩阵的计算式分别如式(3)和式(4)所示：

$$\boldsymbol{S}_o=((\boldsymbol{D}^{\mathrm{out}})^{-1}\cdot\boldsymbol{A})\cdot((\boldsymbol{D}^{\mathrm{out}})^{-1}\cdot\boldsymbol{A})^{\mathrm{T}}-\mathrm{diag}(((\boldsymbol{D}^{\mathrm{out}})^{-1}\cdot\boldsymbol{A})\cdot((\boldsymbol{D}^{out})^{-1}\cdot\boldsymbol{A})^{\mathrm{T}}) \tag{3}$$

$$\boldsymbol{S}_n=((\boldsymbol{D}^{\mathrm{in}})^{-1}\cdot\boldsymbol{A})\cdot((\boldsymbol{D}^{\mathrm{in}})^{-1}\cdot\boldsymbol{A})^{\mathrm{T}}-diag(((\boldsymbol{D}^{\mathrm{in}})^{-1}\cdot\boldsymbol{A})\cdot((\boldsymbol{D}^{\mathrm{in}})^{-1}\cdot\boldsymbol{A})^{\mathrm{T}}) \tag{4}$$

其中，$\boldsymbol{D}^{\mathrm{out}}$ 为对角矩阵且满足 $\boldsymbol{D}_{ii}^{\mathrm{out}}=\sum_{j=1}^{n}\boldsymbol{A}_{ij}$，$\boldsymbol{D}^{\mathrm{in}}$ 为对角矩阵且满足 $\boldsymbol{D}_{ii}^{\mathrm{in}}=\sum_{j=1}^{n}\boldsymbol{A}_{ji}$。

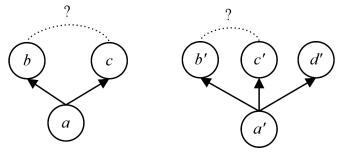



图 2　两个有不同出度的引文网络

Fig. 2　Two citation networks with different outdegrees

至此，我们已经得到正向和反向共引网络相似矩阵，根据上文的分析，正向共引网络的边和反向共引网络的边共同影响原网络节点的相似度，因此将正向和反向共引网络相似矩阵进行结合为下一步的网络融合做准备。

共引网络相似矩阵如式(5)所示：

$$\boldsymbol{S}_{\mathrm{c}}=\beta\cdot\boldsymbol{S}_{\mathrm{o}}+\gamma\cdot\boldsymbol{S}_{\mathrm{n}} \tag{5}$$

其中，$\beta$ 和 $\gamma$ 是控制比重的调节参数，且 $\beta+\gamma=1$。

#### 2.3.3　网络融合

为提取有向共引网络中原网络所包含的边，采用共引相似过滤矩，如式(6)所示：

$$\boldsymbol{S}_{\mathrm{com}}=\boldsymbol{S}_{\mathrm{c}}\Theta\boldsymbol{A}=\beta\cdot\boldsymbol{S}_{\mathrm{o}}\Theta\boldsymbol{A}+\gamma\cdot\boldsymbol{S}_{\mathrm{n}}\Theta\boldsymbol{A} \tag{6}$$

其中，$\Theta$ 为 Hadamard 积，表示将两个矩阵中的对应元素相乘。

原有向网络与共引网络的融合可表示为：

$$\boldsymbol{A}_{\mathrm{r}}=\alpha\cdot\boldsymbol{S}_{\mathrm{com}}+\boldsymbol{A} \tag{7}$$

其中，$\alpha$ 为调节参数，用于控制原有向网络中的共引网络边和原有向网络其他边的比值。$\alpha$ 为算法中的重要参数，如果 $\alpha$ 过小，则无法体现共引网络边的作用，如果 $\alpha$ 过大，则会过于突出共引网络边而破坏原网络的结构。

#### 2.3.4　高阶相似矩阵的构建与求解

为了获得不同高阶相似指标下的相似矩阵，本文采用文献[5]提出的高阶相似度量的一般形式。为了保存有向网络的非传递性，我们基于求得的高阶相似矩阵 $\boldsymbol{S}$ 设定目标函数为：

$$\min_{\boldsymbol{U}^s,\boldsymbol{U}^t}(\boldsymbol{S}-\boldsymbol{U}^s\cdot(\boldsymbol{U}^t)^{\mathrm{T}})^2+\frac{\lambda}{2}(\|\boldsymbol{U}^s\|^2+\|\boldsymbol{U}^t\|^2) \tag{8}$$

其中，$\boldsymbol{U}^s,\boldsymbol{U}^t\in\boldsymbol{R}^{|V|\times d}$，选取 $\boldsymbol{U}^s$ 中的每一行为节点的原向量，$\boldsymbol{U}^t$ 中的每一行为节点的目标向量。$\|\cdot\|$ 为 L2 范数，$\lambda$ 为控制参数。

将式(8)看作一个非负矩阵分解的过程，即：

$$\boldsymbol{S}\approx\boldsymbol{U}^s\cdot(\boldsymbol{U}^t)^{\mathrm{T}} \tag{9}$$

为了加速训练过程，根据文献[16]，本文采用多核 GPU 版本的 NMF-mGPU 方法进行训练。

综上所述，CCE-HOPE 算法的主要流程如算法 1 所示。

**算法 1**　CCE-HOPE

输入：邻接矩阵 **A**，嵌入维度 d，参数 α，β，γ

输出：嵌入向量 **H**

1. 根据式(3)计算正向共引相似矩阵 $\mathbf{S}_{\mathrm{o}}$；
2. 根据式(4)计算反向共引相似矩阵 $\mathbf{S}_{\mathrm{n}}$；
3. 根据式(6)计算处理后的共引相似矩阵 $\mathbf{S}_{\mathrm{com}}$；
4. 根据式(7)计算融合后的网络矩阵 $\mathbf{A}_{\mathrm{r}}$；
5. 利用高阶相似一般形式计算高阶相似矩阵 **S**；
6. 根据式(9)计算节点的源向量 $\mathbf{U}^{\mathrm{s}}$ 和目标向量 $\mathbf{U}^{\mathrm{t}}$；
7. 将节点源向量和目标向量合并，并将其作为最终的向量表示 $\mathbf{H}=[\mathbf{U}^{\mathrm{s}};\mathbf{U}^{\mathrm{t}}]$。

## 3 实验

### 3.1 数据集和参数设置

#### 3.1.1 数据集

实验数据集包括:2004 年美国大选背景下博客之间的超链接数据库 Blog[1]、免费的软件开发人员在线社区平台上信任关系数据库 Advogato[2]、维基百科管理员的投票网络数据库 Wikivote[3]以及科学出版物的引文网络数据库 DBLP[4]。各数据集的结构特征如表 2 所列。表 2 中,$\langle C\rangle$为平均聚类系数,$\langle T\rangle$为网络中三角形的个数。

表 2 数据集的网络结构特征

Table 2 Structure features of each dataset

| Dataset | Blog | Advogato | Wikivote | DBLP |
|---|---|---|---|---|
| $\lvert V\rvert$ | 1 224 | 6 541 | 7 115 | 12 600 |
| $\lvert E\rvert$ | 19 025 | 51 127 | 103 689 | 49 700 |
| $\langle C\rangle$ | 0.022 6 | 0.092 2 | 0.140 9 | 0.119 2 |
| $\langle T\rangle$ | 101 043 | 98 300 | 608 389 | 133 900 |

#### 3.1.2 参数设置和对比算法

对比算法有 Deepwalk[1]、LINE[2]、APP[4]、Common Neighbors、Adamic-Adar、Jaccard、HOPE[5]以及共引处理的 HOPE(C-HOPE)。其中,C-HOPE 为 HOPE 的简单变体,其将 HOPE 原始的方法与本文提出的共引信息相融合来学习网络嵌入。Common Neighbors 和 Adamic-Adar 为分别使用共同邻居和 Adamic-Adar 值作为相似度的度量值。Jaccard 系数采用节点邻居交集个数和节点邻居并集个数的比值作为相似度度量值。

本文实验基于 python3.6,tensorflow 1.13 和 cuda9.0,采用基于多核 GPU 版本的 NMF-mGPU[5]进行优化。对于 Deepwalk[6],设定随机游走窗口的大小为 10,每个节点的游走次数为 10,游走长度为 80。对于 LINE[7],分别学习保存节点间一阶相似和二阶相似的节点表示向量,然后采用两者的合并作为最终的节点表示。对于 APP[8],利用原文提供的 JAVA 环境下的代码并采用默认参数进行训练。实验参数设置如下:嵌入维度为 128;计算相似度时 Deepwalk,LINE 和 APP 算法采用向量内积形式,其余算法采用有向边起始节点的源向量和到达节点的目标向量内积的形式;正向和反向具有相同的重要性(即参数 $\beta=\gamma=0.5$);Katz 系数参数为 0.01;RPR 系数参数为 0.5。我们随机选取 80%的边作为训练集,其余 20%作为测试集。

### 3.2 评价指标

链路预测准确度的度量指标采用 *AUC* 指标[17],其公式如式(10)所示:

$$AUC=(n_1+0.5n_2)/n \tag{10}$$

其中,$n$ 表示实验次数,每次实验从测试集随机抽取一条边,从不存在的边中随机抽取一条边;$n_1$ 是前者分数大于后者分数的次数,$n_2$ 是前者分数等于后者分数的次数。实验重复进行 30 次,选取 30 次结果的平均值作为最终的 *AUC* 值。

### 3.3 权重参数 a 值的影响

$\alpha$ 值控制着共引网络信息和原始网络信息的比例。前面提到,如果 $\alpha$ 值过大或过小都会影响链路预测的准确度,因此首先将 $\alpha$ 值从 0.001 变化到 1,以相应的不融合共引信息的 CCE-HOPE 结果作为比较值,观察在各种高阶相似指标下链接预测 *AUC* 的变化趋势。

图 3 给出了 Katz 指标下,CCE-HOPE 在 4 个数据集上不同 $\alpha$ 值下的 *AUC* 值变化趋势。可以看出,随着 $\alpha$ 值的增加,所有数据集的 *AUC* 值都总体呈现出先上升到最大值然后下降的趋势。同时,Blog,Advogato,Wikivote,DBLP 4 个数据集分别在 $\alpha=0.2$,$\alpha=0.2$,$\alpha=0.07$ 和 $\alpha=0.7$ 时出现断点。另外,这 4 个数据集分别在 $\alpha=0.04$,$\alpha=0.05$,$\alpha=0.02$,$\alpha=0.2$ 时出现分界点。当 $\alpha$ 的取值小于分界点时,链接预测 *AUC* 的值大于比较值;当 $\alpha$ 的取值大于分界点时,链接预测 *AUC* 的值次于比较值。

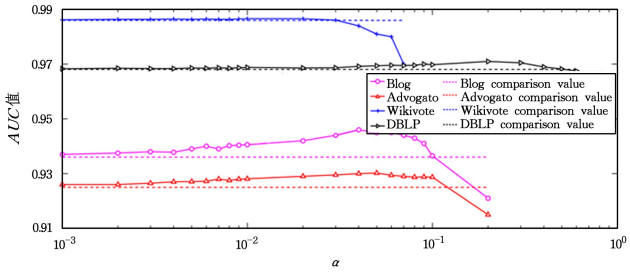



图 3 Katz 指标下 4 个数据集在不同 $\alpha$ 值下的 AUC 值

Fig. 3 AUC values of four data sets on Katz index with different weight factor-$\alpha$

图 4 和图 5 描述了 CN 和 AA 指标下,CCE-HOPE 在 4 个数据集上不同 $\alpha$ 下的 *AUC* 值变化趋势。可以发现,随着 $\alpha$ 值的增加,所有数据集的 *AUC* 值虽然变化曲线相对波折,但总体上仍然呈现出先上升到最大值然后下降的趋势。在 4 个数据集上同样存在分界点,在 CN 指标下,Blog,Advogato,Wikivote 和 DBLP 数据集 $\alpha$ 值的分界点分别为 0.07,0.03,0.09和 0.2。AA 指标下,Advogato,Wikivote 和 DBLP 相应的分界点分别为 0.06,0.5 和 0.1,而 Blog 数据集上 *AUC* 值全部优于比较值,$\alpha$ 值在变化范围内尚未达到分界点。

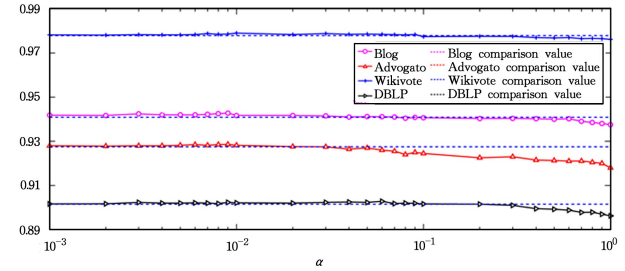



图 4 CN 指标下 4 个数据集在不同 $\alpha$ 值下的 AUC 值

Fig. 4 AUC values of four data sets on CN index with different weight factor-$\alpha$

1) http://konect.uni-koblenz.de/networks/moreno_blogs

2) http://konect.uni-koblenz.de/networks/advogato

3) http://snap.stanford.edu/data/wiki-Vote.html

4) http://networkrepository.com/cit-DBLP.php

5) https://github.com/bioinfo-cnb/bionmf-gpu

6) https://github.com/phanein/deepwalk

7) https://github.com/tangjianpku/LINE

8) https://github.com/AnryYang/APP

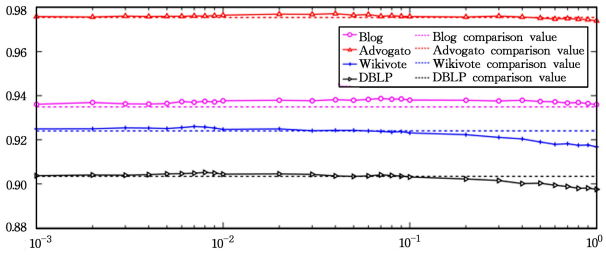



图 5　AA 指标下 4 个数据集在不同 $\alpha$ 值下的 AUC 值

Fig. 5　AUC values of four data sets on AA index with different weight factor-$\alpha$

图 6 给出了 RPR 指标下，CCE-HOPE 在 4 个数据集上不同 $\alpha$ 值下的 *AUC* 值的变化趋势。可以看出，随着 $\alpha$ 值的增加，取值范围内 *AUC* 值都优于比较值。其中，在 Blog 和 DBLP 两个数据集上，AUC 值呈现出总体逐渐上升的趋势；在 Advogato 和 Wikivote 两个数据集上呈现出总体先上升后下降的趋势，其中最高点分别在 $\alpha=0.5$ 和 $\alpha=0.6$ 时出现。Blog 和 DBLP 两个数据集在 $\alpha=1$ 时尚未达到最高点。共引信息的引入实际上改变了 RPR 指标的随机游走，从无差别的随机游走改为有偏向的随机游走，具体而言是偏向于网络中的三角形。这也说明了有策略的偏向随机游走可以提高链接预测的准确度。整体来看，$\alpha$ 值在局部相似和全局相似相结合的 Katz，CN 和 AA 3 种方法上存在分界点，当 $\alpha$ 取值小于分界点时，链接预测 *AUC* 的值大于比较值；当 $\alpha$ 取值大于分界点时，链接预测 *AUC* 的值次于比较值。而对于基于随机游走的全局相似 RPR 方法，实验结果表明，在变化范围内不存在分界点，即 *AUC* 值均大于比较值。如果 $\alpha$ 值过大会过度强调共引网络信息，导致原始网络信息的破坏而影响链路预测效果；如果 $\alpha$ 值过小则会由于共引网络信息所占比例太低而使得链路预测效果不明显。综合在 4 个数据集中 4 种高阶相似方法的比较结果可知，$\alpha$ 值的适宜取值范围为[0.005，0.03]。

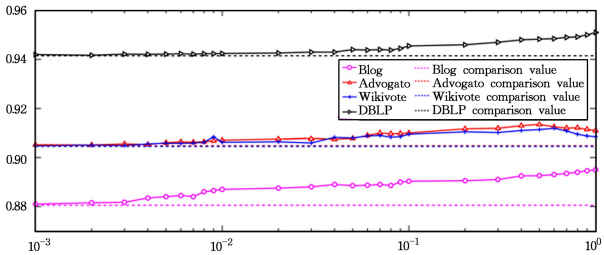



图 6　PRP 指标下 4 个数据集在不同 $\alpha$ 值下的 AUC 值

Fig. 6　AUC values of four data sets on PRP index with different weight factor-$\alpha$

### 3.4　对比实验

根据第 3.3 节的分析，本节取 $\alpha=0.01$，并观察 CCE-HOPE 与相关算法的效果对比。由表 3 可以发现：

1）在所有数据集上，可以保存有向网络非对称性的训练模型 APP，HOPE，C-HOPE 和 CCE-HOPE 优于仅适用于无向网络的训练模型 Deepwalk 和 LINE 算法，说明了有向非对称性对链接预测的积极作用。

2）相比 CN，AA，PRP 3 种高阶相似指标，以 katz 为高阶相似指标的 CCE-HOPE 在不同数据集上表现最好，此结果说明了 katz 指标在保存非对称性的高阶相似计算时的优越性。相比 CN 和 AA 仅保存二阶相似，katz 可以保存更高阶的相似；相比 PRP 在随机游走时的随机性，katz 对高阶相似的度量更准确。

表 3　链接预测的 *AUC* 值

Table 3　*AUC* scores for link prediction

| Dataset | | Blog | Advogato | Wikivote | DBLP |
|---|---|---|---|---|---|
| Deepwalk | | 0.7734 | 0.8541 | 0.8310 | 0.8867 |
| Line | | 0.5716 | 0.7867 | 0.6253 | 0.6300 |
| APP | | 0.7812 | 0.8655 | 0.8921 | 0.8893 |
| Common Neighbors | | 0.9062 | 0.8417 | 0.9034 | 0.6852 |
| Adamic-Adar | | 0.9068 | 0.8420 | 0.9035 | 0.6855 |
| Jaccard | | 0.9060 | 0.8417 | 0.9033 | 0.6849 |
| HOPE | katz | 0.8511 | 0.8621 | 0.9454 | 0.9067 |
| | CN | 0.9248 | 0.9033 | 0.9588 | 0.8258 |
| | AA | 0.9189 | 0.9043 | 0.9592 | 0.8338 |
| | PRP | 0.7599 | 0.8695 | 0.9040 | 0.8960 |
| C-HOPE | katz | 0.9059 | 0.8897 | 0.9553 | 0.9086 |
| | CN | 0.9261 | 0.9070 | 0.9609 | 0.8274 |
| | AA | 0.9210 | 0.9082 | 0.9611 | 0.8356 |
| | PRP | 0.7694 | 0.8740 | 0.9051 | 0.8971 |
| CCE-HOPE | katz | **0.9417** | **0.9293** | **0.9866** | **0.9689** |
| | CN | 0.9411 | 0.9281 | 0.9790 | 0.9021 |
| | AA | 0.9377 | 0.9247 | 0.9765 | 0.9044 |
| | PRP | 0.8860 | 0.9070 | 0.9064 | 0.9424 |

3）在 4 种高阶相似指标下，CCE-HOPE 均可以取得最优值。对比相应高阶相似指标下的 HOPE 模型的链路预测 *AUC* 值，在 4 个数据集上 katz 指标下的 CCE-HOPE 平均提高了 7.41%（4.35%～10.64%），CN 指标下的 CCE-HOPE 平均提高了 3.94%（2.11%～9.23%），AA 指标下的 CCE-HOPE 平均提高了 3.63%（1.80%～8.46%），PRP 指标下的 CCE-HOPE 平均提高了 6.58%（0.26%～16.59%）。

4）与 Common Neighbors 和 Adamic-Adar 相比，以 CN 和 AA 为高阶相似指标的训练模型 HOPE，C-HOPE 和 CCE-HOPE 在链路预测时的准确度均有提高，其中 CCE-HOPE 最好，这直接说明了模型训练的有效性。

5）相比 HOPE，在所有数据集上不同的高阶相似指标下 C-HOPE 的链路预测准确度均有所提高，这直接说明了共引信息的有效性，同时也验证了本文的假设。

**结束语**　考虑到有向共引网络对节点相似度量的影响，本文设计了共引信息度量函数，将共引信息融入非对称高阶相似矩阵中，并利用多核 GPU 版本的 NMF-mGPU 方法进行优化，从而提出了一种表现良好的共引增强的高阶相似保存网络嵌入模型。针对共引影响所占的比重，根据实验中 *AUC* 的变化规律确定了 $\alpha$ 的最佳的取值范围。后续工作将考虑：1）建立异质网络的共引网络，利用共引网络对异质网络高阶相似的影响，提高异质网络节点嵌入的效果；2）针对共引信息比重问题，设计针对不同网络能自适应调整比重的算法，以进一步提高模型的效率。

## 参 考 文 献

[1]　PEROZZI B，Al-RFOU R，SKIENA S. Deepwalk：Online learning of social representations[C]// Proc of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. New York：Assoc Computing Machinery，2014：701-710.

[2] TANG J, QU M D, WANG M Z, et al. Line: Large-scale information network embedding[C]// Proc of the 24th International Conference on World Wide Web. New York: Assoc Computing Machinery, 2015: 1067-1077.

[3] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks[C]// Proc of the 22th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. New York: Assoc Computing Machinery, 2016: 855-864.

[4] ZHOU C, LIU Y Q, LIU X F, et al. Scalable graph embedding for asymmetric proximity[C]// Proc of the 31th AAAI Conference on Artificial Intelligence. PALO ALTO: ASSOC ADVANCEMENT ARTIFICIAL INTELLIGENCE, 2017: 2942-2948.

[5] OU M D, CUI P, PEI J, et al. Asymmetric transitivity preserving graph embedding[C]// Proc of the 22nd ACM SIGKDD international conference on Knowledge Discovery and Data Mining. New York: Assoc Computing Machinery, 2016: 1105-1114.

[6] RIBEIRO L F, SAVERESE P H, FIGUEIREDO D R. struc2-vec: Learning node representations from structural identity [C]// Proc. of the 23th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. New York: Assoc Computing Machinery, 2017: 385-394.

[7] HUANG X, LI J D, XIA H. Label Informed Attributed Network Embedding[C]// Proc of the 10th ACM International Conference on Web Search & Data Mining. New York: Assoc Computing Machinery, 2016: 731-739.

[8] KIM J, PARK H, LEE J E. SIDE: Representation learning in signed directed networks[C]// Proc of the 27th World Wide Web International Conference. New York: Assoc Computing Machinery, 2018: 509-518.

[9] WANG X, CUI P, WANG J, et al. Community preserving network embedding[C]// Proc of the 31th AAAI Conference on Artificial Intelligence. PALO ALTO: ASSOC ADVANCEMENT ARTIFICIAL INTELLIGENCE, 2017: 203-209.

[10] TU C C, ZHANG W C, LIU Z Y, et al. Max-Margin DeepWalk: discriminative learning of network representation[C]// Proc of the 25th AAAI Conference on Artificial Intelligence. Palo Alto: Assoc Advancement Artificial Intelligence, 2016: 3889-3895.

[11] ZHANG D K, YIN J, ZHU X Q, et al. Collective classification via discriminative matrix factorization on sparsely labeled networks[C]// Proc of the 25th International Conference on World Wide Web. New York: Assoc Computing Machinery, 2016: 1563-1572.

[12] WANG D X, CUI P, ZHU W W. Structural deep network embedding[C]// Proc of the 22th International Conferenceon on Knowledge Discovery and Data Mining. New York: Assoc Computing Machinery, 2016: 1225-1234.

[13] GU Y P, SUN Y Z, LI Y E. Rare: Social rank regulated large-scale network embedding[C]// Proc of the 27th World Wide Web International Conference. New York: Assoc Computing Machinery, 2018: 359-368.

[14] DONG Y X, CHAWLA N V, SWAMI A. Ananthrammetapath-vec: Scalable representation learning for heterogeneous networks [C]// Proc of the 22nd ACM SIGKDD international conference on Knowledge Discovery and Data Mining. New York: Assoc Computing Machinery, 2018: 135-144.

[15] WANG H W, WANG J, WANG J L, et al. GraphGAN: Graph representation learning with generative adversarial nets[C]// Proc of the 32th AAAI Conference on Artificial Intelligence. Palo Alto: Assoc Advancement Artificial Intelligence, 2018: 2508-2515.

[16] EDGARDO M-R, DANIEL T-M, JAVIER S, et al. NMF-mGPU: non-negative matrix factorization on multi-GPU systems [J]. BMC Bioinformatics, 2015, 16(1): 43-55.

[17] HANLEY J A, MCNEIL B J. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve [J]. Radiology, 1982, 143(1): 29-36.

**WU Yong**, born in 1989, Ph. D candidate, is a student member of China Computer Federation. His main research interests include knowledge graph and network representation learning.

**WANG Bin-jun**, born in 1962, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include knowledge graph, data mining and natural language processing.