Journal of Computer Applications
计算机应用,2019,39(8):2366 - 2373
ISSN 1001-9081
CODEN JYIIDU
2019-08-10
http://www.joca.cn

文章编号:1001-9081(2019)08-2366-08
doi:10.11772/j.issn.1001-9081.2019010213

基于高阶近似的链路预测算法

杨燕琳[1,2,3],冶忠林[1,2,3,4],赵海兴[1,2,3,4*],孟 磊[1,2,3]

(1.青海师范大学 计算机学院,西宁 810016; 2.青海省藏文信息处理与机器翻译重点实验室(青海师范大学),西宁 810008;
3.藏文信息处理教育部重点实验室(青海师范大学),西宁 810008; 4.陕西师范大学 计算机科学学院,西安 710062)

(*通信作者电子邮箱 h.x.zhao@163.com)

摘 要:目前大部分链路预测算法只研究了节点与邻居节点之间的一阶相似性,没有考虑节点与邻居的邻居节点之间的高阶相似性关系。针对此问题,提出一种基于高阶近似的链路预测算法(LP-HOPA)。首先,求出网络的归一化邻接矩阵和相似度矩阵;其次,利用矩阵分解的方法将相似度矩阵进行分解,得到网络节点的表示向量以及其上下文的表示向量;然后,通过高阶网络表示学习的网络嵌入更新(NEU)算法对原始相似度矩阵进行高阶优化,并利用归一化的邻接矩阵计算出更高阶的相似度矩阵表示;最后,在四个真实的数据集上进行大量的实验。实验结果表明,与原始链路预测算法相比,大部分利用 LP-HOPA 优化后的链路预测算法准确率提升了 4% 到 50%。此外,LP-HOPA 算法能够将基于低阶网络局部结构信息的链路预测算法转换为基于节点高阶特征的链路预测算法,在一定程度上肯定了基于高阶近似链路预测算法的有效性和可行性。

关键词:链路预测;高阶近似;相似度矩阵;矩阵分解;网络嵌入更新算法

中图分类号:TP393 **文献标志码**:A

Link prediction algorithm based on high-order proximity approximation

YANG Yanlin[1,2,3], YE Zhonglin[1,2,3,4], ZHAO Haixing[1,2,3,4*], MENG Lei[1,2,3]

(1. *College of Computer, Qinghai Normal University, Xining Qinghai* 810016, *China*;
2. *Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province (Qinghai Normal University), Xining Qinghai* 810008, *China*;
3. *Key Laboratory of Tibetan Information Processing of Ministry of Education (Qinghai Normal University), Xining Qinghai* 810008, *China*;
4. *School of Computer Science, Shaanxi Normal University, Xi'an Shaanxi* 710062, *China*)

Abstract: Most of the existing link prediction algorithms only study the first-order similarity between nodes and their neighbor nodes, without considering the high-order similarity between nodes and the neighbor nodes of their neighbor nodes. In order to solve this problem, a Link Prediction algorithm based on High-Order Proximity Approximation (LP-HOPA) was proposed. Firstly, the normalized adjacency matrix and similarity matrix of a network were solved. Secondly, the similarity matrix was decomposed by the method of matrix decomposition, and the representation vectors of the network nodes and their contexts were obtained. Thirdly, the original similarity matrix was high-order optimized by using Network Embedding Update (NEU) algorithm of high-order network representation learning, and the higher-order similarity matrix representation was calculated by using the normalized adjacency matrix. Finally, a large number of experiments were carried out on four real datasets. Experiments results show that, compared with the original link prediction algorithm, the accuracy of most of the link prediction algorithms optimized by LP-HOPA is improved by 4% to 50%. In addition, LP-HOPA can transform the link prediction algorithm based on local structure information of low-order network into the link prediction algorithm based on high-order characteristics of nodes, which confirms the validity and feasibility of the link prediction algorithm based on high order proximity approximation to a certain extent.

Key words: link prediction; high-order proximity approximation; similarity matrix; matrix decomposition; Network Embedding Update (NEU) algorithm

0 引言

随着网络科学的不断进步,网络的演化机制[1]受到了学者们的广泛关注,而链路预测为网络的演化提供了一个高效简单的比较机制,因此,对链路预测的研究也受到了学者们的广泛关注。网络中的链路预测是指如何通过已知网络的特征、结构和节点信息等预测不相连的两个节点之间产生链接的可能性[2]。链路预测的应用对实际生活产生了重要的意义。例如,蛋白质网络[3]可预测没有产生相互作用的蛋白质节点未来产生相互作用的可能性,将最可能产生相互作用的

收稿日期:2019-01-30;**修回日期**:2019-03-26;**录用日期**:2019-03-26。

基金项目:国家自然科学基金资助项目(11661069, 61663041, 61763041);藏文信息处理与机器翻译重点实验室项目(2013-Z-Y17)。

作者简介:杨燕琳(1995—),女,四川南充人,硕士研究生,CCF 会员,主要研究方向:复杂网络、链路预测; 冶忠林(1989—),男,青海民和人,博士研究生,CCF 会员,主要研究方向:数据挖掘、自然语言表示学习; 赵海兴(1969—),男,青海湟中人,教授,博士生导师,CCF 会员,主要研究方向:复杂网络、超图理论; 孟磊(1994—),男,河南项城人,硕士研究生,CCF 会员,主要研究方向:复杂网络、超网络。

蛋白质做实验,可提高实验的成功率;社交分析网络[4]可预测陌生人成为朋友的可能性;标签分类[5]可通过节点的特征和性质去预测节点的类别;异常检测[6]可以通过链路预测预测网络中的错误链接,对错误链接进行纠正;信息推荐系统[7]则通过链路预测向用户自动推荐可能需要的物品。除此之外,链路预测还应用到了网络建模[8]、知识获取[9]等领域。为了将链路预测应用到实际生活中,研究者们已经提出了很多链路预测算法,大多是基于相似性和最大似然估计的链路预测算法,可是现实生活中产生的数据越来越多,网络越来越复杂,规模越来越大,而这些链路预测算法大多存在着高计算复杂性和低精确性的问题,主要适用于小规模网络,这就造成了链路预测应用的局限性。

邻接矩阵可以将网络简单直接地表示出来,但是邻接矩阵占用了大量的存储空间,数据十分稀疏,因此,研究者们转而思考如何将网络数据高效地表示出来。网络表示学习(Network Representation Learning, NRL)[10]将网络的节点信息转化为低维稠密的向量来表示。因此,将链路预测与网络表示学习相结合,可以更全面地读取网络节点信息,使链路预测结果更精确。DeepWalk[11]和 LINE (Large-scale Information Network Embedding)[12]是最具代表性的基于神经网络的网络表示学习算法。DeepWalk 算法利用了网络结构的随机游走序列信息,并通过节点及其上下文节点之间的关系训练神经网络;LINE 算法考虑了网络的两种相似度,用两节点是否直接相连来刻画一阶相似度,用不相连的两个节点的共同邻居来刻画二阶相似性,该算法可被应用于大规模网络表示学习任务,但其精度却不如 DeepWalk 算法。网络嵌入更新(Network Embedding Update, NEU)算法[13]是一种通过简单的矩阵转换构建高阶网络表示的方法,但并不需要重新训练网络表示学习模型,该算法可以应用到任意的 NRL 方法,来提高它们的性能。例如,将该算法应用在 DeepWalk 算法时,只用了 DeepWalk 算法运行时间的 1%,即有显著的提升。

目前大部分链路预测算法只研究了节点与邻居节点之间的一阶相似性,忽略了节点与邻居的邻居节点的高阶相似性关系,比如二阶相似性、三阶相似性等。本文基于 NEU 表示学习算法,提出了一种基于高阶近似的链路预测算法(Link Prediction algorithm Based on High Order Proximity Approximation, LP-HOPA)。该方法能够将基于低阶网络局部结构信息的链路预测算法转化为节点高阶特征相似链路预测算法,提升其链路预测性能。该方法在相似矩阵分解的结果上进行高阶转换,以获得节点之间高阶的关系,从而可以得到高阶近似的相似度矩阵,该相似度矩阵给了节点之间的一阶相似性、二阶相似性,并可以推广到节点之间的 n 阶相似性,因此可以更精准地预测节点间的相似性。

本文的主要工作有:1) 将 NEU 表示学习算法引入到网络的链路预测中,提出了一种基于高阶近似的链路预测算法 LP-HOPA。2)基于四个真实的数据集在 17 个常见的链路预测指标进行了链路预测实验,结果表明,LP-HOPA 可有效学习网络的结构特征,具有一定程度的可行性和有效性,而且它的链路预测性能优于本文中用来对比的链路预测指标。LP-HOPA 能够将基于低阶网络局部结构信息的链路预测算法转换为基于节点高阶特征的链路预测算法,从而提升链路预测性能。

1　相关工作

近 10 年,得益于 Clauset 等 2008 年在 Nature 上发表的论文[14]以及 Redner 对这篇论文的评论文章[15],链路预测的研究方法被相继提出。目前,主要包括以下三大类方法:

第一类是基于相似性的链路预测方法,即节点相似性越大,说明连边可能性越大。主要有如下三小类:

1)基于网络局部结构信息相似性方法,主要包括基于共同邻居(Common Neighbors, CN)[7]的相似性指标、基于 Adamic-Adar(Adamic-Adar, AA)[7]算法的相似性指标、基于资源分配(Resource Allocation, RA)[7]的相似性指标和基于优先链接(Preferential Attachment, PA)[7]的相似性指标。在共同邻居的基础上,可以详细分成 6 种相似性指标,包括基于余弦相似性指标 Salton[7]、Jaccard 相似性指标[7]、Sorenson 相似性指标[7]、大度节点有利相似性指标 HPI(Hub Promoted Index)[7]、大度节点不利相似性指标 HDI(Hub Depressed Index)[7]和节点对分配相似性指标 LHN-1(Leicht-Holme-Newman)[16]。在 CN、AA 和 RA 的基础上,文献[17]中提出了基于局部朴素贝叶斯算法的相似性指标 LNBCN(Local Naive Bayes model-CN)、LNBAA(Local Naive Bayes model-AA)和 LNBRA(Local Naive Bayes model-RA)。

2)基于路径的相似性方法,包括:基于局部路径(Local Path, LP)[18]的相似性指标、基于节点声望的相似性指标 Katz[19]和 LHN-II(Leicht-Holme-Newman II)指标[16]。

3)基于随机游走的相似性方法,包括:基于平均通勤时间(Average Commute time, ACT)[20]的相似性指标、基于随机游走的余弦相似性指标 Cos +[21]、局部随机游走(Local Random Walk, LRW)[18]的相似性指标、有叠加效应的随机游走(Superposed Random Walk, SRW)[18]相似性指标和有重启的随机游走(Random Walk with Restart, RWR)[18]相似性指标。

第二类是基于概率和最大似然估计的链路预测方法。基于概率的链路预测方法通过构建贝叶斯、马尔可夫等数学模型预测未知的链接,基于最大似然估计的链路预测方法利用网络的结构信息得到最大似然数。Clauset 等[14]最初提出基于最大似然估计的链路预测方法,将其应用到有明显层次的网络结构中,发现有比较高的精确度;田甜等[22]提出了一种基于最大似然估计的链路预测模型,将脑网络的数据建立了层次随机图,再结合马尔可夫算法计算脑网络边的连接概率,结果显示出了良好的预测性能。

第三类是基于机器学习的链路预测方法。该类方法在相似性链路预测方法的基础之上进一步获取网络特征。如廖亮等[23]针对机会网络研究了基于支持向量机(Support Vector Machine, SVM)的链路预测,构建了基于节点对空间相似性和时间特征加权融合的支持向量机模型,从空间相似性和时间特征两个角度分析单节点对的连接概率,并证明了其具有很好的预测效果;吴祖峰等[24]将 AdaBoost 集成学习算法应用到了链路预测中,在论文合作网络和电子邮件网络等进行了

实验验证;吕伟民等[25]将基于机器学习的链路预测方法应用到了科研合作网络中,提高了推荐合作的精确度。

基于相似性的链路预测方法不能充分挖掘网络的结构特征,尤其是基于网络局部结构信息的相似性方法,预测准确度较低;基于概率和基于最大似然估计的链路预测方法预测准确度高,但算法复杂度较高,导致了此类方法应用的局限性,主要适用于预测低阶小规模网络;基于机器学习的链路预测方法的预测效果较好,可以预测大规模网络,但是节点特征矩阵占用了大量的存储空间,数据稀疏,因此大大增加了特征读取时间;而网络表示学习可以将特征信息用低维稠密的向量表示出来,这就降低了算法的时间复杂度,应用到链路预测中,具有低算法复杂度和高精确度的特点。因此,有越来越多的学者提出了基于网络表示学习的链路预测算法。如杨晓翠等[26]提出了基于网络表示学习的链路预测算法,冶忠林等[27]提出了基于矩阵分解的 DeepWalk 链路预测算法,刘思等[28]提出了基于网络表示学习与随机游走的链路预测算法,均得到了较好的预测结果,但他们都只考虑了节点与邻居节点之间的一阶相似性,忽略了节点与邻居的邻居节点的高阶相似性关系;而本文提出的 LP-HOPA 能够将基于低阶网络局部结构信息的链路预测算法转化为节点高阶特征相似链路预测算法,提升链路预测性能。

2 基于高阶近似的链路预测

2.1 定义描述

信息网络[29]:定义一个信息网络为 $G=(V,E)$,其中 V 表示顶点集,E 表示边集。定义 G 的特征矩阵为 $\boldsymbol{X}$,$\boldsymbol{X}$ 是 $|V|\times m$ 维的,m 表示节点的特征属性个数。如果网络 G 对应的特征矩阵 $\boldsymbol{X}$ 是非空的,则 G 是一个信息网络。

网络表示学习[29]:给定一个信息网络 $G=(V,E)$,$\boldsymbol{X}$ 为 G 的特征矩阵,满足任意顶点 $v\in V$,学习将网络用低维向量 $\boldsymbol{r}_v\in\mathbf{R}^k$ 表示,其中 $\boldsymbol{r}_v$ 是一个低维稠密的实数向量,且满足 $k\ll|V|$。

链路预测[2]:给定一个无向网络 $G=(V,E)$,其中 V 表示顶点集,E 表示边集。定义 M 为该网络中的最大边数,满足 $M=|V|(|V|-1)/2$,$M-E$ 表示该网络中不存在的边集,而链路预测则是在集合 $M-E$ 中找出未来可能连边的顶点对。通过某种链路预测方法可计算出每对未连边的顶点对的相似性分数,分数越高则连边可能性越大。

2.2 高阶网络表示学习 NEU 算法

NEU 算法是由 Yang 等[13]提出的一种基于矩阵转换的高阶近似网络表示方法,可以应用到任意的网络表示学习算法中,以提高基类网络表示学习算法的性能。

给定超参数 $\lambda\in(0,1/2]$,归一化的邻接矩阵 $\boldsymbol{A}$,$\boldsymbol{R}$ 和 $\boldsymbol{C}$ 分别表示信息网络的网络表示和上下文表示,通过 NEU 算法更新后,$\boldsymbol{R}'$ 和 $\boldsymbol{C}'$ 分别为更新后的网络表示和上下文表示的方法如下:

$$\begin{cases}\boldsymbol{R}'=\boldsymbol{R}+\lambda\boldsymbol{A}\cdot\boldsymbol{R}\\ \boldsymbol{C}'=\boldsymbol{C}+\lambda\boldsymbol{A}^{\mathrm{T}}\cdot\boldsymbol{C}\end{cases}\tag{1}$$

计算出 $\boldsymbol{A}\cdot\boldsymbol{R}$ 和 $\boldsymbol{A}^{\mathrm{T}}\cdot\boldsymbol{C}$ 的时间复杂度是 $O(|V|d)$,因为矩阵 $\boldsymbol{A}$ 是稀疏的并且有 $O(|V|)$ 个非零项,因此,式(1)一次迭代的整体时间复杂度为 $O(|V|d)$。

式(1)可以在进一步推广到二阶形式,以获得二阶近似的网络表示。首先更新 $\boldsymbol{R}$ 和 $\boldsymbol{C}$:

$$\begin{cases}\boldsymbol{R}'=\boldsymbol{R}+\lambda_1\boldsymbol{A}\cdot\boldsymbol{R}+\lambda_2\boldsymbol{A}\cdot(\boldsymbol{A}\cdot\boldsymbol{R})\\ \boldsymbol{C}'=\boldsymbol{C}+\lambda_1\boldsymbol{A}^{\mathrm{T}}\cdot\boldsymbol{C}+\lambda_2\boldsymbol{A}^{\mathrm{T}}\cdot(\boldsymbol{A}^{\mathrm{T}}\cdot\boldsymbol{C})\end{cases}\tag{2}$$

式(2)的时间复杂度仍然是 $O(|V|d)$,但是它在一次迭代中可以得到比式(1)更高的相似矩阵近似。同时也可以使用比式(2)更复杂的更新公式来探索更高的近似,例如三阶、四阶…… n 阶近似的网络表示。

NEU 算法避免了高阶相似矩阵的精确计算,也避免了通过模型训练高阶网络表示模型,但可以产生高阶近似的网络表示,因此该算法可以有效提高网络表示的质量。直观上,式(1)和(2)允许学习到的节点表示进一步传播到它们的邻居,因此,顶点之间距离较长的相似将被嵌入。

2.3 基于高阶近似的链路预测算法

通过 NEU 算法中节点向量的表示过程发现,该算法在矩阵分解的结果上更新以获得更高阶的网络表示,使最后的向量效果蕴含更多的网络结构特征,因此将 NEU 算法运用到链路预测后可以得到高阶相似矩阵,可以更精准地预测节点间的相似性。在此基础上,本文将 NEU 算法融入到链路预测中,提出了一种基于高阶近似的链路预测算法 LP-HOPA。

LP-HOPA 流程如图 1 所示,具体如下:

1)输入网络 $G=(V,E)$,其中 V 为顶点集,E 为边集。

2)将网络分割成训练集和测试集,将训练集转化为邻接矩阵 $\tilde{\boldsymbol{A}}\in\mathbf{R}^{|V|\times|V|}$,如果 v_i 和 v_j 之间有连边,则 $\tilde{\boldsymbol{A}}_{ij}=1$,否则 $\tilde{\boldsymbol{A}}_{ij}=0$。对角矩阵 $\boldsymbol{D}\in\mathbf{R}^{|V|\times|V|}$,$\boldsymbol{D}_{ii}$ 的数值即为 v_i 的度。

3)将 $\tilde{\boldsymbol{A}}$ 转化为归一化邻接矩阵 $\boldsymbol{A}$,$\boldsymbol{A}=\boldsymbol{D}^{-1}\tilde{\boldsymbol{A}}$,即每行之和等于 1。

4)通过 AA、CN、RA 和 MFI 等链路预测基准指标计算相似矩阵 $\boldsymbol{S}$,并使用 SVDS 算法将 $\boldsymbol{S}$ 分解为 $\boldsymbol{U}_{|V|\times k}\boldsymbol{\Sigma}_{k\times k}$ 和 $\boldsymbol{V}^{\mathrm{T}}_{k\times|V|}$ 三个矩阵。SVDS 是奇异值分解(Singular Value Decomposition,SVD)的一种 Matlab 方法,表示取最大的 6 个特征值。接着,将分解得到的这三个矩阵转化为两个矩阵的乘积:

$$\boldsymbol{ne}=\boldsymbol{U}\cdot\sqrt{\boldsymbol{\Sigma}}\tag{3}$$

$$\boldsymbol{nc}=\sqrt{\boldsymbol{\Sigma}}\cdot\boldsymbol{V}^{\mathrm{T}}\tag{4}$$

使得 $\boldsymbol{S}\approx\boldsymbol{ne}\cdot\boldsymbol{nc}$。根据 NEU 算法得知,$\boldsymbol{ne}$ 为节点的表示向量,$\boldsymbol{nc}$ 为上下文的表示向量。

5)利用 NEU 算法在矩阵分解结果的基础之上进行更新,以获得更高阶的网络特征相似度矩阵结果:

$$\boldsymbol{ne}'=\boldsymbol{ne}+\lambda_1\boldsymbol{A}\cdot\boldsymbol{ne}+\lambda_2\boldsymbol{A}\cdot(\boldsymbol{A}\cdot\boldsymbol{ne})\tag{5}$$

$$\boldsymbol{nc}'=\boldsymbol{nc}+\lambda_1\boldsymbol{A}^{\mathrm{T}}\cdot\boldsymbol{nc}+\lambda_2\boldsymbol{A}^{\mathrm{T}}\cdot(\boldsymbol{A}^{\mathrm{T}}\cdot\boldsymbol{nc})\tag{6}$$

通过以上更新结果,可得到接近高阶近似的相似矩阵:$\boldsymbol{S}'\approx\boldsymbol{ne}'\cdot\boldsymbol{nc}'$。将该相似矩阵应用到链路预测算法中,利用链路预测准确度度量指标 AUC(Area Under the receiver operating characteristic Curve)指标计算出 AUC 值,从而评估本文所提出的链路预测性能。

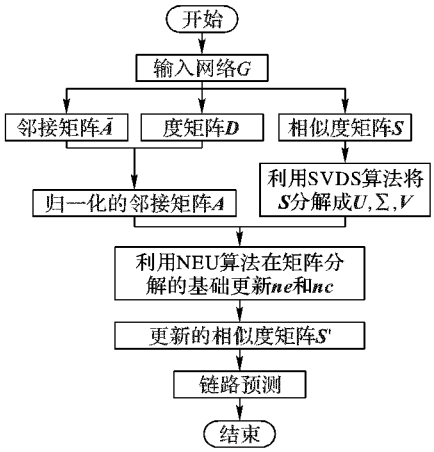

图 1　LP-HOPA 框架

Fig. 1　Framework of LP-HOPA

3　实验结果与分析

3.1　实验数据

本文选择四个真实网络数据集进行测试,分别为:

1) Citeseer 网络: http://citeseerx. ist. psu. edu/index。它是由 3 312 篇世界顶级会议论文构成的引文网络,包含 4 732 篇文章之间引用或被引用的关系。

2) DBLP(DataBase systems and Logic Programming) 网络: https://dblp. uni-trier. de/。它是由 3 119 个作者构成的合作网络,顶点表示作者,连边表示作者之间的合作关系,包含 39 516个作者间的合作关系。

3) Cora 网络: http://www. cs. umd. edu/ ~ sen/lbc-proj/data/cora. tgz。它是由 2 708 份科学出版物组成的引文网络,包含 5 429 条连边。

4) Wiki 网络: https://www. wikipedia. org/。该网络是维基百科网页链接网络,本文只取了 2 405 个网页之间的 17 981 个链接关系。

表 1 进一步列出了这四个数据集的网络拓扑结构特征,其中:$|V|$ 表示节点数,$|E|$ 表示连边数,$|Y|$ 表示网络标签数,K 表示平均度,D 表示网络直径,L 表示平均路径长度,P 表示密度,C 表示平均聚类系数。

表 1　数据集的拓扑结构特征

Tab. 1　Topology features of datasets

网络	$\|V\|$	$\|E\|$	$\|Y\|$	K	D	L	P	C
Citeseer	3 312	4 732	6	2.857	28	9.036	0.001	0.257
DBLP	3 119	39 516	4	25.339	14	4.199	0.008	0.259
Cora	2 708	5 429	7	4.010	19	6.310	0.001	0.293
Wiki	2 405	17 981	19	14.953	9	3.650	0.006	0.480

3.2　评价指标

本文链路预测准确度的度量指标采用 AUC 指标[30]。AUC 可以从整体上衡量算法的精确度,可以描述为在测试集中随机选择一条存在连边的分数值高于随机选择一条不存在连边的分数值的概率,如果独立重复比较 n 次,有 n_1 次在测试集中存在连边的分数大于不存在连边的分数,有 n_2 次在测试集中存在连边的分数等于不存在连边的分数,则 AUC 值可以定义为:

$$AUC = (n_1 + 0.5n_2)/n \tag{7}$$

一般意义上,计算出的 AUC 值至少应大于 0.5,至多不超过 1。AUC 值越高,算法的准确度越高。

3.3　基准方法

本文将常用的 17 种基于相似性的链路预测算法作为基准进行性能比较。其中包括基于网络局部结构信息的基准方法:CN、Salton、HPI、HDI、LHN-1、AA、RA、PA、LNBAA、LNBCN、LNBRA;基于路径的基准方法:LP、Katz;基于随机游走的基准方法:ACT、Cos^+;基于矩阵森林理论的相似性指标(Matrix-Forest theory Index, MFI)[18];基于传递的相似性指标(Transferring Similarty Common Neighbor, TSCN)[31]。下面分别对各基准方法作简要介绍。

1) 基于共同邻居的相似性指标 CN。

两个未连接的节点 x、y,如果有更多的共同邻居,则 x、y 更倾向于连边,且点 x 和 y 的相似性就等于它们共同的邻居数,其中,$\Gamma(x)$、$\Gamma(y)$ 分别表示节点 x 和 y 的邻居节点集。

$$S_{xy}^{\text{CN}} = |\Gamma(x) \cap \Gamma(y)|/|\Gamma(x) \cup \Gamma(y)| \tag{8}$$

2) 基于共同邻居的余弦相似性指标 Salton。

Salton 指标在 CN 指标的基础之上引入了两个节点度的信息,当这个比值越大,则节点 x 和 y 越相似,更倾向于连边。其中,k_x 和 k_y 分别表示节点 x 和 y 的度。

$$S_{xy}^{\text{Salton}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_y}} \tag{9}$$

3) 基于共同邻居的大度节点有利相似性指标 HPI。

HPI 指标的分母是由较小的节点度决定的,因此,度越大的节点越容易与其他节点具有更高的相似性,与枢纽节点链接的连边会被分配更高的相似度分数。

$$S_{xy}^{\text{HPI}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min\{k_x, k_y\}} \tag{10}$$

4) 基于共同邻居的大度节点不利相似性指标 HDI。

与 HPI 指标相反,度越小的节点容易与其他节点具有更高的相似性,与枢纽节点链接的连边会被分配更低的相似度分数。

$$S_{xy}^{\text{HDI}} = |\Gamma(x) \cap \Gamma(y)|/\max\{k_x, k_y\} \tag{11}$$

5) 基于共同邻居的节点对分配相似性指标 LHN-1。

相比于 CN 指标,LHN-1 指标是给拥有很多共同邻居的节点对分配更高的相似性,而且该指标不会无限制地变大。

$$S_{xy}^{\text{LHN-1}} = |\Gamma(x) \cap \Gamma(y)|/(k_x \times k_y) \tag{12}$$

6) 基于 Adamic-Adar 算法的相似性指标 AA。

AA 指标通过共同邻居节点的度为每个节点赋予权重,为度小的节点赋予更大的权重,从而提高度小的共同邻居节点的贡献值,因为度小的共同邻居节点比度大的共同邻居节点贡献大,其中,z 表示节点 x 和 y 的共同邻居节点,$|\Gamma(z)|$ 表示节点 z 的度值。

$$S_{xy}^{\text{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\lg k(z)} \tag{13}$$

7) 基于资源分配算法的相似性指标 RA。

RA 指标的基本思想是资源的重新分配。虽然节点 x 和 y 未相连但是 x 可以传递资源到节点 y,假设它们每经过一个共同邻居就有一个单位的资源平均分配传给它的邻居,那么节点 x 和 y 的相似度就定义为 y 可以接收到的资源数。

$$S_{xy}^{\text{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)} \tag{14}$$

8)基于优先链接的相似性指标 PA。

PA 指标只考虑了节点度对相似度的影响,因此该指标的算法复杂度相比于其他算法低。节点 x 的度越大,则它与其他节点的连边多,那么它与未连接的节点产生连边的可能性越大。

$$S_{xy}^{\mathrm{PA}} = k_x \times k_y \tag{15}$$

9)基于局部路径的相似性指标 LP。

LP 指标在 CN 指标的基础上不仅考虑了直接邻居的贡献,同时也考虑了邻居的邻居的贡献。其中,∂ 为可调参数,用来控制邻居的邻居的贡献度,当 ∂ 等于 0 时,LP 指标即 CN 指标。

$$S_{xy}^{\mathrm{LP}} = A^2 + \alpha A^3 \tag{16}$$

10)基于 CN 指标的局部朴素贝叶斯算法的相似性指标 LNBCN。

基于局部朴素贝叶斯算法提出了一个角色函数,旨在量化每一个节点作为共同邻居节点时的作用大小,并将该函数与 CN 指标相结合,得到 LNBCN 指标。其中,$|O_{xy}|$ 为节点 x 和 y 的共同邻居数,s 为连边相连的概率与连边不相连的概率之比,R_w 为提出的角色函数。

$$S_{xy}^{\mathrm{LNBCN}} = |O_{xy}| \operatorname{lb} s + \sum_{v_w \in O_{xy}} \operatorname{lb} R_w \tag{17}$$

11)基于 AA 指标的局部朴素贝叶斯算法的相似性指标 LNBAA。

与 LNBCN 指标相似,LNBAA 指标将角色函数与 AA 指标相结合。其中,k_w 表示 w 节点的度。

$$S_{xy}^{\mathrm{LNBAA}} = \sum_{v_w \in O_{xy}} \frac{1}{\operatorname{lb} k_w}(\operatorname{lb} s + \operatorname{lb} R_w) \tag{18}$$

12)基于 RA 指标的局部朴素贝叶斯算法的相似性指标 LNBRA。

与 LNBCN 指标相似,LNBRA 指标将角色函数与 RA 指标相结合。

$$S_{xy}^{\mathrm{LNBRA}} = \sum_{v_w \in O_{xy}} \frac{1}{k_w}(\operatorname{lb} s + \operatorname{lb} R_w) \tag{19}$$

13)基于矩阵森林理论的相似性指标 MFI。

MFI 指标利用了矩阵森林理论,考虑了节点对之间存在多条加权连边的情况,认为网络中两个节点的相似性可定义为两个节点所在的相同网络生成树数量与网络中所有的生成森林数量之比。其中,$\boldsymbol{I}$ 为单位矩阵,$\boldsymbol{L}$ 为拉普拉斯矩阵。

$$S_{xy}^{\mathrm{MFI}} = (\boldsymbol{I} + \partial \boldsymbol{L})^{-1} \tag{20}$$

14)基于自洽(节点传递)的相似性指标 TSCN。

TSCN 指标是通过中间节点传递相似性,更加完整地利用了网络的结构信息,因此,在某种程度上可以提高预测的准确性。其中,ε 是常量参数,$\boldsymbol{T}$ 是已知的相似性矩阵,$\boldsymbol{S}$ 是传递相似性矩阵,即最终得到的相似性矩阵。

$$\boldsymbol{S}_{xy}^{\mathrm{TSCN}} = \varepsilon \boldsymbol{T}\boldsymbol{S} + \boldsymbol{T} \tag{21}$$

15)基于节点声望的相似性指标 Katz。

Katz 指标考虑了 x 和 y 之间的所有路径数,对于短路径赋予大权重,对于长路径赋予小权重。其中,β 为可调参数。

$$S_{xy}^{\mathrm{Katz}} = \beta \boldsymbol{A} + \beta^2 \boldsymbol{A}^2 + \beta^3 \boldsymbol{A}^3 \cdots = (\boldsymbol{I} - \beta \boldsymbol{A})^{-1} - \boldsymbol{I} \tag{22}$$

16)基于平均通勤时间的相似性指标 ACT。

一个随机粒子从节点 x 到达节点 y 平均要走的步数 $m(x,y)$,那么,节点 x 和 y 的平均通勤时间定义为:

$$n(x,y) = m(x,y) + m(y,x)$$

则其数值求解可以通过求该网络拉普拉斯矩阵的伪逆 $\boldsymbol{L}^+$ 得到。如果节点 x 和 y 的平均通勤时间越短,则这两个节点的相似度越高。

$$S_{xy}^{\mathrm{ACT}} = \frac{1}{l_{xx}^+ + l_{yy}^+ - 2l_{xy}^+} \tag{23}$$

17)基于随机游走的余弦相似性指标 $\cos^+$。

在由向量 $\boldsymbol{v}_x = \boldsymbol{\Lambda}^2 \boldsymbol{U}^{\mathrm{T}} \boldsymbol{e}_x$ 展开的欧氏空间中,$\boldsymbol{U}$ 是一个标准正交矩阵,$\boldsymbol{\Lambda}$ 为对角矩阵,对角线元素为特征根,$\boldsymbol{e}_x$ 表示一个只有第 x 个为 1,其他元素为 0 的一维向量;$\boldsymbol{L}^+$ 中的元素 l_{xy}^+ 为 v_x 和 v_y 的内积。

$$S_{xy}^{\cos^+} = \cos(x,y)^+ = (l_{xy}^+) \Big/ \sqrt{l_{xx}^+ l_{yy}^+} \tag{24}$$

3.4 实验设置

在 LP-HOPA 中,本文设置了四个网络训练集的训练比例分别为 0.7、0.8、0.9,测试集的训练比例分别为 0.3、0.2、0.1;特征维度 d 设为 100;超参数 $\lambda_1 = 0.5$,$\lambda_2 = 0.25$;迭代次数 $maxIter$ 设为 3;最终实验结果为各网络独立运行 10 次的平均值。

3.5 实验结果

首先求出网络的归一化邻接矩阵 $\boldsymbol{A}$ 和相似度矩阵 $\boldsymbol{S}$:其次,通过 SVDS 将 $\boldsymbol{S}$ 进行分解,得到网络节点的表示向量 $\boldsymbol{ne}$ 以及其上下文的表示向量 $\boldsymbol{nc}$:接着,通过高阶网络表示学习 NEU 算法在原始相似度矩阵的结果上进行高阶优化;然后,利用归一化的邻接矩阵计算出更高阶的相似度矩阵表示;最后,在 Citeseer、DBLP、Cora 和 Wiki 四个数据集上进行实验验证。为验证上述方法的可行性及有效性,本文使用 3.3 节的所有相似性链路预测基准指标进行对比。

表 2 列出了在上述四个数据集上,其训练集的训练比例分别为 0.7、0.8、0.9 时,3.3 节中原始方法和本文方法在各链路预测基准算法的 AUC 值对比。

观察表 2 中的数据结果,对于原始链路预测方法,在四个数据集上链路预测准确率都高于 80% 的仅有 4 个算法,即 LP、Katz、Cos^+ 和 MFI。其中 Katz 和 MFI 算法链路预测性能较优,尤其是在 Citeseer 数据集上准确率都达到了 97% 以上;而 CN、Salton、HPI 等算法在 Citeseer 数据集上显示链路预测准确率较差,最低低至 65%。对比相关工作中所列出的基于相似性的链路预测算法发现,基于路径的相似性方法链路预测性能较优,基于网络局部结构信息相似性方法性能较差。对于本文方法,在四个数据集上链路预测准确率都高于 80% 的有 14 个算法,即为 CN、Salton、HPI、HDI、LHN-1、AA、RA、LP、Katz、LNBAA、LNBCN、LNBRA、Cos^+ 和 MFI。其中 HDI、LNBRA 和 Salton 算法链路预测性能较优,尤其是在 DBLP 数据集上准确率都达到了 93.5% 以上;而 ACT 在 Citeseer 数据集上显示链路预测准确率较差,最低低至 35.9%。对比相关工作中所列出的基于相似性的链路预测算法发现,基于网络局部结构信息相似性方法性能较优,基于随机游走的相似性方法链路预测性能较差。通过对比原始链路预测方法和本文方法可以发现,利用 LP-HOPA 后,链路预测准确率都高于 80% 的算法个数比原始链路预测算法多 10 个,这些算法为 CN、Salton、HPI、HDI、LHN-1、AA、RA、LNBAA、LNBCN 和 LNBRA。大部分相比原始链路预测算法准确率提升了 4% 到 50% 不等,仅有极个别算法有较小幅度的下降。原始算法基于路径的相似性方法链路预测性能较优,而利用 LP-HOPA 后

基于网络局部结构信息相似性方法性能较优,在一定程度上肯定了本文方法的可行性和有效性。

利用 LP-HOPA 后,在四个数据集上,CN、Salton、HPI、HDI、LHN-1、AA、RA、LNBAA、LNBRA 和 LBNCN 算法的 AUC 值得到了大幅度提升,尤其是在 Citeseer、Cora 数据集上基本都提升了 20 个百分点,PA 算法略微下降,但 Katz、ACT、MFI 和 TSCN 算法上链路预测准确度却有一定程度的下降,尤其是在 Citeseer 数据集上 ACT 算法 AUC 值比原始方法下降了 40 个百分点。PA 算法是一种只考虑节点度对相似度影响的链路预测算法,而本文方法对 PA 算法进行优化时,由于并未考虑连边,所以性能略微下降;Katz 是基于全部路径的链路预测算法,因此 Katz 是一个 n 阶特征的链路预测算法,然而本文方法结合节点的高阶特征,仅考虑了节点间 6 阶以内的相似性,因此将 n 阶特征降至 6 阶特征导致了链路预测性能下降;MFI 是一种基于森林树的算法,考虑的是两个节点所在的相同网络生成树数量,因此考虑了网络节点的高阶特征相似性;ACT 算法实质上是考虑了两个节点之间随机游走来回的平均路径长度之和,路径越短,相似性越高,而 LP-HOPA 考虑了节点间 6 阶以内的相似性,因此性能有大幅度下降。

为了更进一步分析个别相似性指标使用本文方法后预测准确率下降的原因,表 3 比较了 1、3、5 阶对最终预测结果的影响。为使对比结果更加鲜明,2、4、6 阶对比结果未展示。

表 3 所示为 Citeseer 数据集在训练率为 0.7 时,对 1、3、5 阶使用本文方法后的 AUC 值。可以看出,除了 ACT 相似度指标外,其他指标的 AUC 值都是呈增加状态。阶数越高,ACT 的预测准确率反而越来越低,说明 ACT 路径越短,相似性越高;Katz 指标呈增加趋势,也可以由此说明若不断将阶数增加到 n, Katz 相似性指标的 AUC 值也会不断递增;CN、Salton、AA 和 LNBRA 等这些低阶的相似性指标呈现了一个很好的增加趋势。

本文所采用的方法是结合节点间的二阶相似、三阶相似、n 阶相似综合考虑节点的相似性,所以对于能够转换为高阶特征的链路预测算法使用本文方法后性能会有所下降,但对于低阶链路预测算法使用本文方法的链路预测准确率都会有一定程度的提升,由此说明本文方法主要适用于基于网络局部结构信息的低阶链路预测基准方法。

表 2　4 种数据集上两种算法的 AUC 性能对比

Tab. 2　AUC performance comparison of two algorithms on four datasets

方法	数据集	训练率	CN	Salton	HPI	HDI	LHN-I	AA	RA	PA	LP	Katz	LNBAA	LNBCN	LNBRA	ACT	Cos +	MFI	TSCN
原始链路预测算法	Citeseer	0.7	0.665	0.659	0.654	0.673	0.659	0.663	0.659	0.781	0.811	0.971	0.658	0.653	0.664	0.751	0.883	0.977	0.848
		0.8	0.704	0.718	0.721	0.721	0.728	0.716	0.725	0.800	0.864	0.976	0.712	0.718	0.708	0.751	0.902	0.980	0.856
		0.9	0.763	0.758	0.757	0.738	0.731	0.734	0.768	0.795	0.921	0.982	0.767	0.761	0.759	0.740	0.904	0.979	0.862
	DBLP	0.7	0.859	0.853	0.845	0.855	0.849	0.850	0.852	0.767	0.924	0.931	0.852	0.855	0.863	0.800	0.951	0.950	0.906
		0.8	0.885	0.872	0.881	0.880	0.871	0.886	0.885	0.774	0.939	0.934	0.880	0.887	0.886	0.798	0.954	0.956	0.905
		0.9	0.904	0.899	0.892	0.904	0.894	0.902	0.904	0.777	0.947	0.940	0.900	0.897	0.907	0.791	0.954	0.957	0.924
	Cora	0.7	0.686	0.684	0.684	0.698	0.686	0.691	0.681	0.704	0.807	0.916	0.693	0.683	0.694	0.742	0.893	0.929	0.896
		0.8	0.733	0.727	0.739	0.721	0.731	0.728	0.734	0.718	0.851	0.924	0.729	0.734	0.719	0.724	0.919	0.941	0.910
		0.9	0.757	0.768	0.746	0.761	0.781	0.752	0.771	0.729	0.873	0.932	0.772	0.757	0.771	0.729	0.933	0.961	0.921
	Wiki	0.7	0.864	0.852	0.855	0.854	0.849	0.872	0.857	0.828	0.923	0.936	0.865	0.865	0.866	0.792	0.914	0.933	0.916
		0.8	0.880	0.881	0.887	0.873	0.877	0.881	0.891	0.822	0.939	0.945	0.879	0.890	0.878	0.804	0.913	0.947	0.923
		0.9	0.906	0.886	0.885	0.893	0.879	0.914	0.907	0.826	0.941	0.946	0.905	0.901	0.893	0.791	0.924	0.946	0.787
本文方法	Citeseer	0.7	0.859	0.875	0.861	0.897	0.908	0.858	0.860	0.758	0.861	0.834	0.850	0.854	0.881	0.516	0.899	0.905	0.855
		0.8	0.909	0.888	0.891	0.906	0.916	0.872	0.892	0.770	0.890	0.876	0.888	0.894	0.893	0.359	0.915	0.905	0.880
		0.9	0.910	0.920	0.905	0.897	0.934	0.889	0.901	0.755	0.909	0.912	0.894	0.907	0.914	0.664	0.927	0.931	0.913
	DBLP	0.7	0.898	0.948	0.930	0.947	0.896	0.919	0.939	0.743	0.897	0.919	0.922	0.902	0.937	0.764	0.933	0.865	0.864
		0.8	0.901	0.939	0.942	0.949	0.913	0.922	0.949	0.744	0.906	0.916	0.932	0.906	0.944	0.770	0.941	0.863	0.852
		0.9	0.908	0.948	0.941	0.950	0.933	0.924	0.956	0.751	0.910	0.922	0.931	0.890	0.953	0.772	0.943	0.867	0.879
	Cora	0.7	0.871	0.879	0.862	0.868	0.864	0.853	0.870	0.655	0.858	0.850	0.884	0.881	0.882	0.688	0.910	0.879	0.873
		0.8	0.881	0.905	0.897	0.897	0.883	0.911	0.898	0.646	0.892	0.880	0.902	0.892	0.888	0.691	0.919	0.912	0.878
		0.9	0.914	0.916	0.901	0.910	0.908	0.922	0.925	0.660	0.911	0.887	0.912	0.910	0.930	0.689	0.936	0.917	0.899
	Wiki	0.7	0.905	0.931	0.923	0.922	0.906	0.930	0.928	0.743	0.909	0.914	0.923	0.912	0.925	0.487	0.911	0.867	0.888
		0.8	0.905	0.930	0.926	0.928	0.908	0.925	0.937	0.746	0.913	0.930	0.925	0.909	0.924	0.537	0.916	0.861	0.886
		0.9	0.925	0.928	0.921	0.923	0.901	0.925	0.93	0.750	0.914	0.934	0.923	0.917	0.925	0.514	0.931	0.868	0.737

表 3　不同阶数对应的 AUC 值对比

Tab. 3　Comparison of AUC values corresponding to different orders

阶数	CN	Salton	HPI	HDI	LHN-I	AA	RA	PA	LP	Katz	LNBAA	LNBCN	LNBRA	ACT	Cos +	MFI	TSCN
1	0.792	0.834	0.815	0.841	0.858	0.802	0.828	0.770	0.800	0.767	0.810	0.804	0.823	0.671	0.891	0.874	0.812
3	0.828	0.847	0.839	0.856	0.875	0.823	0.832	0.774	0.830	0.786	0.839	0.831	0.863	0.553	0.909	0.888	0.830
5	0.847	0.860	0.853	0.867	0.895	0.847	0.855	0.782	0.854	0.815	0.856	0.851	0.881	0.499	0.915	0.907	0.844

3.6　时间复杂度对比

图 2 所示为 Citeseer 数据集下训练率为 0.7 时的链路预测算法时间复杂度对比,不同的数据集在不同的训练率下结果虽然不同,但是时间变化规律一致,因此时间复杂度对比结

果具有代表性。横坐标表示相似性基准方法，纵坐标表示运行时间。通过对比可以看出，CN、Salton、HPI、HDI、LHN1、AA、RA、LP、LNBAA、LNBCN 和 LNBRA 使用本文方法优化后，时间复杂度基本与原始方法的时间复杂度持平，说明本文方法对低阶链路预测算法能够在不提升算法复杂度的情况下转化为高阶特征的链路预测算法，并提升其链路预测性能；PA、Katz、ACT、Cos + 、MFI 和 TSCN 在使用本文方法后的算法复杂度有不同程度的提高，尤其是 MFI 的运行时间有大幅度的增加，由此说明本文方法并不适用于可转化为高阶特征的链路预测算法。

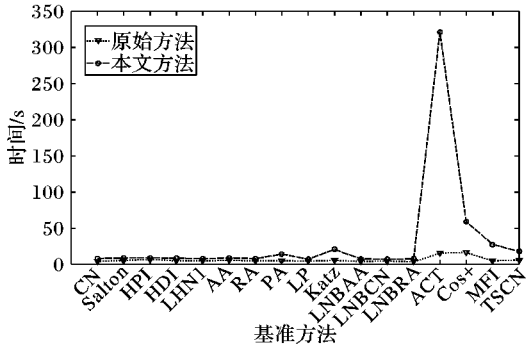

图 2 时间复杂度对比

Fig. 2 Comparison of time complexity

3.7 度分布可视化

度分布是网络的基本性质之一，指网络中节点的度的概率分布。度分布与网络的拓扑结构性质密切相关，因此，研究网络的度分布可以基本确定网络的类型。本文使用 Matlab 实现了 Citeseer、Cora、DBLP 和 Wiki 四个数据集的度分布可视化。具体结果如图 3 所示，横坐标表示该数据集节点度值，纵坐标表示该度值对应的节点个数。

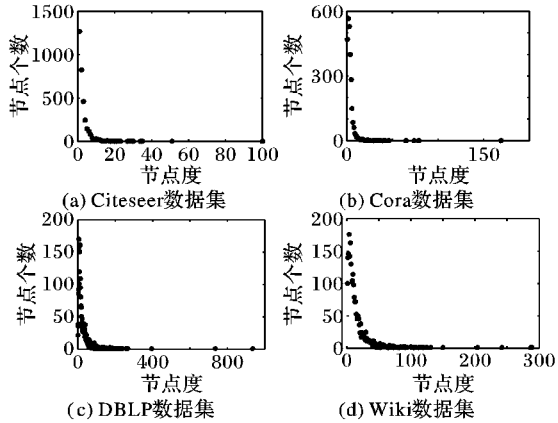

图 3 Citeseer、Cora、DBLP 和 Wiki 数据集上的度分布可视化

Fig. 3 Visualization of degree distribution on Citeseer, DBLP, Cora and Wiki datasets

通过对比四个数据集可以看出，Citeseer 数据集中节点最大度仅为 100，但度为 1 的节点高频出现，高达 1 270 次；Cora 数据集节点的最大度值为 169，度为 2 的节点较多，有 567 次；DBLP 数据集中节点的最大度高于 900，度为 6 的节点较多，有 170 次；Wiki 数据集中节点最大度大于 280，度为 4 的节点较多，有 176 次。通过对比，Citeseer 和 Cora 数据集是相对比较稀疏的网络，而 DBLP 和 Wiki 数据集是相对稠密的网络。再通过对比表 2 和表 3 可以发现，不管是原始方法还是本文方法，DBLP 和 Wiki 数据集链路预测效果比 Citeseer 和 Cora 数据集好。由此可以说明，链路预测对稠密的网络的预测结果比稀疏的网络好。

4 结语

本文针对目前大部分链路预测算法未考虑节点与邻居的邻居节点之间的高阶相似性关系，提出了一种基于高阶近似的链路预测算法 LP-HOPA。首先，利用矩阵分解将相似度矩阵进行分解；其次，通过高阶网络表示学习 NEU 算法在矩阵分解的结果上进行更新，得到高阶的相似度矩阵；最后，在 Citeseer、DBLP、Cora 和 Wiki 四个数据集上进行了实验验证。实验结果表明，在实际网络的链路预测中，利用 LP-HOPA 可以进行更加有效的高阶转换，使其链路预测性能比现有的众多链路预测算法更加优异。但是 LP-HOPA 也存在着不足，主要有以下两点：1) LP-HOPA 对于能够转换为高阶特征的链路预测算法使性能会有所下降；2) 本文只考虑了 6 阶以内的相似性，因此基于随机游走的 ACT 算法性能有大幅度下降。在下一步的研究中，将尝试考虑在高阶转换时融入外部信息，充分挖掘网络的相关特征，并且将相似性阶数提升，使其不仅能适用于网络低阶的链路预测算法，同样基于高阶特征的链路预测算法也有一定程度的提升。除此之外，还可基于本文方法结合边的权值尝试构造相似性指标。

参考文献 (References)

[1] ALBERT R, BARABÁSI A-L. Statistical mechanics of complex networks [J]. Reviews of Modern Physics, 2002, 74(1): 47 – 97.

[2] GETOOR L, DIEHL C P. Link mining: a survey [J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 3 – 12.

[3] YU H, BRAUN P, YILDIRIM M A, et al. High-quality binary protein interaction map of the yeast interactome network [J]. Science, 2008, 322(5898): 104 – 110.

[4] XIE X, LI Y, ZHANG Z, et al. A joint link prediction method for social network [C]// Proceedings of the 2015 International Conference of Young Computer Scientists, Engineers and Educators, CCIS 503. Berlin: Springer, 2015: 56 – 64.

[5] KUMAR R, NOVAK J, TOMKINS A. Structure and evolution of online social networks [M]// Link Mining: Models, Algorithms, and Applications. New York: Springer, 2010: 337 – 357.

[6] ZHANG X, ZHAO C, WANG X, et al. Identifying missing and spurious interactions in directed networks [C]// Proceedings of the 2014 International Conference on Wireless Algorithms, Systems, and Applications, LNCS 8491. Berlin: Springer, 2014: 470 – 481.

[7] ZHOU T, LYU L, ZHANG Y. Predicting missing links via local information [J]. European Physical Journal B, 2009, 71(4): 623 – 630.

[8] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in networks [J]. Physical Review E: Statistical Nonlinear & Soft Matter Physics, 2006, 73(2): No. 026120.

[9] ZADEH P M, KOBTI Z. A knowledge based framework for link prediction in social networks [C]// Proceedings of the 2016 Interna-

tional Symposium on Foundations of Information and Knowledge Systems, LNCS 9616. Cham: Springer, 2016: 255 – 268.

[10] 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. (TU C C, YANG C, LIU Z Y, et al. Network representation learning: an overview [J]. SCIENTIA SINICA Information, 2017, 47(8): 980 – 996.)

[11] BEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations [C]// Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York: ACM, 2014: 701 – 710.

[12] TANG J, QU M, WANG M, et al. LINE: Large-scale information network embedding [C]// Proceedings of the 24th International Conference on World Wide Web. New York: ACM, 2015: 1067 – 1077.

[13] YANG C, SUN M, LIU Z, et al. Fast network embedding enhancement via high order proximity approximation [C]// Proceedings of the 2017 26th International Joint Conference on Artificial Intelligence. Pola Alto, CA: AAAI, 2017: 3894 – 3900.

[14] CLAUSET A, MOORE C, NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks [J]. Nature, 2008, 453(7191): 98 – 101.

[15] REDNER S. Networks: teasing out the missing links [J]. Nature, 2008, 453 (7191): 47 – 48.

[16] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in networks [J]. Physical Review E: Statistical Nonlinear & Soft Matter Physics, 2006, 73(2): No. 026120.

[17] LIU Z, ZHANG Q-M, LYU L, et al. Link prediction in complex networks: a local Naive Bayes model [J]. Europhysics Letters, 2011, 96(4): No. 48007.

[18] 王富田, 张鹏, 肖井华. 链路预测算法错边识别能力的评测[J/OL]. 中国科技论文在线, 2015 [2015-12-30]. http://www.paper.edu.cn/releasepaper/content/201512 – 1363. (WANG F T, ZHANG P, XIAO J H. Evaluation the ability of link prediction methods in the spurious link detection [J/OL]. Sciencepaper Online, 2015 [2015-12-30]. http://www.paper.edu.cn/releasepaper/content/201512 – 1363.)

[19] KATZ L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1): 39 – 43.

[20] KLEIN D J, RANDIC M. Resistance distance [J]. Journal of Mathematical Chemistry, 1993, 12(1): 81 – 95.

[21] FOUSS F, PIROTTE A, RENDERS J, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation [J]. IEEE Transaction on Knowledge & Data Engineering, 2007, 19(3): 355 – 369.

[22] 田甜, 杨艳丽, 郭浩, 等. 基于层次随机图模型的脑网络链路预测[J]. 计算机应用研究, 2016, 33(4): 1066 – 1069. (TIAN T, YANG Y L, GUO H, et al. Link prediction of brain networks based on hierarchical random graph model [J]. Application Research of Computers, 2016, 33(4): 1066 – 1069.)

[23] 廖亮, 张恒锋. 基于支持向量机的机会网络链路预测[J]. 信息通信, 2018(9): 28 – 30. (LIAO L, ZHANG H F. Link prediction based on support vector machine chance network [J]. Information & Communications, 2018(9): 23 – 25.)

[24] 吴祖峰, 梁棋, 刘峤, 等. 基于 AdaBoost 的链路预测优化算法[J]. 通信学报, 2014, 35(3): 116 – 123. (WU Z F, LIANG Q, LIU Q, et al. Modified link prediction algorithm based on AdaBoost [J]. Journal on Communications, 2014, 35(3): 116 – 123.)

[25] 吕伟民, 王小梅, 韩涛. 结合链路预测和 ET 机器学习的科研合作推荐方法研究[J]. 数据分析与知识发现, 2017, 1(4): 38 – 45. (LYU W M, WANG X M, HAN T. Recommending scientific research collaborators with link prediction and extremely randomized trees algorithm [J]. Data Analysis and Knowledge Discovery, 2017, 1(4): 38 – 45.)

[26] 杨晓翠, 宋甲秀, 张曦煌. 基于网络表示学习的链路预测算法[J/OL]. 计算机科学与探索, 2018[2018-06-25]. http://kns.cnki.net/kcms/detail/11.5602.TP.20180622.1301.008.html. (YANG X C, SONG J X, ZHANG X H. Link prediction algorithm based on network representation learning [J/OL]. Journal of Frontiers of Computer Science and Technology, 2018[2018-06-25]. http://kns.cnki.net/kcms/detail/11.5602.TP.20180622.1301.008.html.)

[27] 冶忠林, 曹蓉, 赵海兴, 等. 基于矩阵分解的 DeepWalk 链路预测算法[J/OL]. 计算机应用研究, 2018 [2018-12-12]. http://kns.cnki.net/KCMS/detail/51.1196.TP.20181211.1539.012.html. (YE Z L, CAO R, ZHAO H X, et al. Link prediction based on matrix factorization for DeepWalk [J/OL]. Application Research of Computers, 2018 [2018-12-12]. http://kns.cnki.net/KCMS/detail/51.1196.TP.20181211.1539.012.html.)

[28] 刘思, 刘海, 陈启买, 等. 基于网络表示学习与随机游走的链路预测算法[J]. 计算机应用, 2017, 37(8): 2234 – 2239. (LIU S, LIU H, CHEN Q M, et al. Link prediction algorithm based on network representation learning and random walk [J]. Journal of Computer Applications, 2017, 37(8): 2234 – 2239.)

[29] 陈维政, 张岩, 李晓明. 网络表示学习[J]. 大数据, 2015, 1(3): 8 – 22. (CHEN W Z, ZHANG Y, LI X M. Network representation learning [J]. Big Data Research, 2015, 1(3): 8 – 22.)

[30] 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.

[31] CHEBOTAREV P, SHAMIS E. The matrix-forest theorem and measuring relations in small social groups [J]. Automation & Remote Control, 1997, 58(9): 1505 – 1514.

This work is partially supported by National Natural Science Foundation of China (11661069, 61663041, 61763041), the Tibetan Information Processing and Machine Translation Key Laboratory (2013-Z-Y17).

YANG Yanlin, born in 1995, M. S. candidate. Her research interests include complex network, link prediction.

YE Zhonglin, born in 1989, Ph. D. candidate. His research interests include data mining, natural language representation learning.

ZHAO Haixing, born in 1969, Ph. D., professor. His research interests include complex network, hypergraph theory.

MENG Lei, born in 1994, M. S. candidate. His research interests include complex network, hyper network.