计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(3):43−51 [doi: 10.15888/j.cnki.csa.009430]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于节点交互度的社会网络链路预测①

徐瑞阳[1,2], 徐振宇[1,2], 李家印[1,2], 许　力[1,2]

[1](福建师范大学 计算机与网络空间安全学院, 福州 350117)
[2](福建省网络安全与密码技术重点实验室, 福州 350007)
通信作者: 李家印, E-mail: lijiayin@fjnu.edu.cn

摘　要: 链路预测是通过已知的网络拓扑和节点属性挖掘未来时刻节点潜在关系的重要手段, 是预测缺失链路和识别虚假链路的有效方法, 在研究社会网络结构演化中具有现实意义. 传统的链路预测方法基于节点信息或路径信息相似性进行预测, 然而, 前者考虑指标单一导致预测精度受限, 后者由于计算复杂度过高不适合在规模较大网络中应用. 通过对网络拓扑结构的分析, 本文提出一种基于节点交互度 (interacting degree of nodes, IDN) 的社会网络链路预测方法. 该方法首先根据网络中节点间的路径特征, 引入了节点效率的概念, 从而提高对于没有公共邻居节点之间链路预测的准确性; 为了进一步挖掘节点间共同邻居的相关属性, 借助分析节点间共同邻居的拓扑结构, 该方法还创新性地整合了路径特征和局部信息, 提出了社会网络节点交互度的定义, 准确刻画出节点间的相似度, 从而增强网络链路的预测能力; 最后, 本文借助 6 个真实网络数据集对 IDN 方法进行验证, 实验结果表明, 相比于目前的主流算法, 本文提出的方法在 *AUC* 和 *Precision* 两个评价指标上均表现出更优的预测性能, 预测结果平均分别提升 22% 和 54%. 因此节点交互度的提出在链路预测方面具有很高的可行性和有效性.

关键词: 链路预测; 节点交互度; 网络拓扑; 相似性; 社会网络

引用格式: 徐瑞阳,徐振宇,李家印,许力.基于节点交互度的社会网络链路预测.计算机系统应用,2024,33(3):43–51. http://www.c-s-a.org.cn/1003-3254/9430.html

Link Prediction for Social Networks Based on Interacting Degree of Nodes

XU Rui-Yang[1,2], XU Zhen-Yu[1,2], LI Jia-Yin[1,2], XU Li[1,2]

[1](College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China)
[2](Fujian Provincial Key Laboratory of Network Security and Cryptology, Fuzhou 350007, China)

Abstract: Link prediction is an important means of mining potential relationships between nodes in the future through known network topology and node attributes, which is an effective method for predicting missing links and identifying false links and has practical significance in the study of social network structure evolution. Traditional link prediction methods are based on the similarity of node information or path information. However, the former considers a single index, resulting in limited prediction accuracy, and the latter is not suitable for application in large-scale networks due to excessive computational complexity. Through the analysis of network topology, this study proposes a social network link prediction method based on the interacting degree of nodes (IDN). The method first introduces the concept of node efficiency based on the path characteristics between nodes in the network, which improves the accuracy of link prediction between nodes without common neighbors. In order to further explore the relevant attributes of common neighbors between nodes, by analyzing the topology of common neighbors between nodes, the method also innovatively integrates the path characteristics and local information to propose the definition of the IDN in a social network, which accurately portrays the degree of similarity between nodes and thus enhances the prediction ability of network links. Finally, this

① 基金项目: 国家自然科学基金 NSFC 海峡联合基金 (U1905211); 福建省科技项目 (2022G02003, 2021L3032); 福建省教育厅中青年项目 (JAT220814)
收稿时间: 2023-09-08; 修改时间: 2023-10-08; 采用时间: 2023-10-25; csa 在线出版时间: 2024-01-02
CNKI 网络首发时间: 2024-01-03

study validates the IDN method with the help of six real network datasets, and the experimental results show that, compared with the current mainstream algorithms, the method proposed in this study shows better prediction performance in both *AUC* and *Precision* evaluation indexes, and the prediction results have been improved by an average of 22% and 54%, respectively. Therefore, the proposal of node interaction degree has high feasibility and effectiveness in link prediction.

Key words: link predication; interacting degree of nodes (IDN); network topology; similarity; social network

社会网络是指个人或组织所构成的社会结构, 构成主体是一些人或社会实体, 在这些人或社会实体之间, 存在一些关系或互动模式, 如个人之间的友情关系、公司之间的合作关系、国家之间的贸易关系等. 社会网络可以建模为一个图数据类型, 其中节点映射到个人或社会实体, 节点之间的连边对应于相应的人或社会实体之间的关联. 个人或社会实体之间的关系在不断变化, 因此会发生多个连边和顶点的添加或删除, 这种变化导致社会网络呈现高度的动态性和复杂性. 社会关系是对个人之间人际关系的描述, 它可以看作是一对节点之间的定义. 具有社会关系的两个节点不仅存在稳定长期的联系, 而且这一对节点要具有更多相同的兴趣爱好. 人们经常根据自己的社会属性来互相接触, 并且与兴趣相似、行为相似和经常见面的人交朋友. 因此, 社会网络局部会出现一些内部连接比外部连接更紧密的子图, 这也是社会网络社会性和高聚集性的一种体现.

近年来, 许多学者开始关注社会网络的信息挖掘, 主要包括链路预测、社团检测、聚类分析等. 其中, 链路预测可用于提取缺失信息、识别虚假交互、评估网络演化机制等, 具有广泛的应用场景和重要的现实意义, 成为近期的研究热点. 在社交网络平台中, 链路预测可以帮助用户在社交平台中寻找感兴趣的朋友, 并进行相应的推荐, 节约用户的社交成本; 对社交平台而言, 链路预测增加个人用户的朋友关系有利于增加网络用户的在线率, 提高用户黏性, 增加社交平台的影响力. 在电子商务领域, 链路预测可以基于个人及其朋友以往的购买信息和相关的属性信息, 预测用户的下次购买行为并进行相应推荐. 在引文网络和科研合作方面, 链路预测可以评估一篇论文是否可以成为一篇高影响力论文, 以及帮助科研工作者找到高水平的科研合作者, 从而促进学术产出. 在生物分子网络中, 链路预测可以利用给定的蛋白质相互作用网络, 从网络结构中挖掘不同的信息, 设计不同的拓扑相似度度量方法, 根据已知的相互作用预测未知的相互作用, 高效挖掘潜在的蛋白质相互作用关系. 在社会网络隐私控制领域, 由于许多用户将个人帖子、音频、视频和其他敏感信息分享到社交网站, 社交平台对用户的隐私保护非常重要, 链路预测可以根据连接权重形式的信任来确定两个用户之间关系的强度, 计算用户之间的信任级别来识别种子用户的所有可能的可信用户. 到目前为止, 在社会网络链路预测的研究中, 基于节点相似性的链路预测算法通过测量两个节点的相似性来进行预测, 具有较强的可扩展性, 应用领域也较为广泛, 长期以来一直受到学者的青睐.

Chai 等[1]考虑如何选择合适的基矩阵以及重建网络的结构特征对链接预测的影响, 提出采用全连接网络的邻接矩阵作为低秩表示的基矩阵, 将重构网络邻接矩阵的核规范作为惩罚项. Zhao 等[2]利用元路径投影和语义图聚合, 提出可以从不同的元路径中学习节点对的嵌入, 用于异构网络的端到端链路预测方法. 李巧丽等[3]提出将节点间最优路径应用于网络中的节点传输能力问题, 将 6 阶范围内最优路径数和最优路径长度进行融合, 定义节点紧密中心性函数来构建相似度传输矩阵. Mishra 等[4]发现多种类型的链路可以被编码到不同的层中, 节点本身及其本质关系得以保留, 由此在考虑各个层相对密度的同时, 将高阶路径和不同的层进行关系融合. Orzechowski 等[5]考虑到社交网络中的交互表现出不对称性, 提出一种新的局部连边聚类指标, 使用不对称邻域重叠对有关社会关系不对称的信息进行提取. Liu 等[6]通过局部信息对单纯分解权重和闭合比率权重进行相似性度量, 提出基于局部特征的高阶链路预测算法, 并捕获局部高阶路径信息以预测网络中未来可能出现的交互作用. 上述方法作为基于路径信息相似性的链路预测方法, 依据近似完整的网络拓扑结构信息, 链路预测准确率较高, 但是算法的计算过程复杂度过大, 不适合应用于大规模的数据集, 并且在实际应用场景中完整的网络拓扑结构信息较难获取.

Mishra 等[7]基于合并节点和连边相关性的概念, 利用边的邻近信息计算边相关性, 而根据节点对图形整体结构的重要性并综合局部和全局的属性计算节点相关性. Liu 等[8]考虑源节点和目的节点之间的中间节点的平均度, 量化节点的初始信息贡献, 应用节点之间通过 3 种信息传输方式接收的总信息量来衡量它们之间的相似性. 郁湧等[9]将节点的共同邻居聚类系数和度融合作为局部信息的指标, 并且使用节点中心性对在网络中节点的重要性进行表示. Sarhangnia 等[10]提出一种新的二分网络中链路预测的相似性度量方法, 基于节点的邻域结构对现有经典方法进行修改并重新定义. Ghasemi 等[11]考虑将网络的拓扑特征与节点的动态特征进行结合, 并提出将子空间聚类算法应用于社会对象的聚类, 有效区分聚类的强度. Zhu 等[12]考虑无向未加权网络邻接矩阵的对称性, 进一步分析不同中心性对链路的影响, 提出一种结合节点重要性和网络拓扑属性的链路预测算法, 说明了节点度中心性和邻近中心性对链路预测精度的重要性. 以上基于节点信息相似性的链路预测方法主要考虑网络拓扑中的节点信息, 算法复杂度较低, 所适用的网络数据集范围更广, 但在考虑单一指标的情况下预测精度受限, 易受到网络拓扑结构变化的影响.

许多现实世界网络中节点之间的平均路径长度较短, 同时网络的拓扑结构又具有高聚集性, 因此网络内部会出现局部的高度连接. 上述方法并未同时对节点间的路径特征和局部信息进行分析, 容易丢失重要的网络拓扑结构信息, 影响预测精度. 本文通过对网络拓扑信息进行挖掘, 充分考虑社会网络结构的社会性和高聚集性, 根据节点的局部信息和路径特征并将其作为节点相似性的衡量依据, 提出一种基于节点交互度的社会网络链路预测方法. 为了验证该方法的预测性能, 本文针对 6 个真实的网络数据集进行了链路预测实验. 结果表明, 与其他方法相比, 本文方法在 *AUC* 和 *Precision* 两个评价指标上均表现出更优的预测性能, 预测结果平均分别提升 22% 和 54%, 结果证明节点交互度在社会网络链路预测方面的可行性和有效性, 可以在各种真实世界的网络中获得更好的准确性.

1 链路预测概述

1.1 问题描述

一个无向网络可以被具体表示为$G=(V,E)$, 其中$V=\{v_1,v_2,v_3,\cdots,v_N\}$表示网络中节点的集合, 共包含$N$个节点, 即$|V|=N$; $E=\{e_1,e_2,e_3,\cdots,e_M\}$表示网络中节点连边的集合, 共有$M$条连边, 即$|E|=M$. 令网络中所有的节点对可能组成边的全集为$U$, 则全集$U$为所有$N\times(N-1)/2$个节点对可能组成所有边的集合. 链路预测问题即利用网络的已知结构和节点属性等信息, 计算尚未形成连边的一对节点之间形成连边的概率. 基于节点相似性的链路预测算法对于每个未连接的节点对(x,y)计算节点相似度, 计算的相似性分数越高, 那么节点对在未来形成链路的概率就越大. 按照该链路预测分数值从大到小进行排序, 排序越靠前的节点对之间就越可能产生链接, 即当计算出的相似性度量S_{xy}越大, 那么节点之间出现连边的可能性就越大.

1.2 基准算法

本文所使用到的符号及其说明如表 1 所示. 本文所对比的链路预测方法如表 2 所示.

表 1 符号说明

符号	含义
S_{xy}	节点x和y的相似性分数
$\Gamma(x)$	节点x的邻居节点集合
k_x	节点x的度
CC_x	节点x的聚类系数
d_{xy}	节点x和y的最短路径

表 2 对比方法的指标

指标	计算公式
CN[13]	$S_{xy}=\lvert\Gamma(x)\cap\Gamma(y)\rvert$
AA[14]	$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log k_z}$
RA[15]	$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{k_z}$
PA[16]	$S_{xy}=k_x\cdot k_y$
JAC[17]	$S_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
CCLP[18]	$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}CC_z$
ERA[19]	$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\left(\frac{1}{\log k_z}\right)^2$
NDCC[20]	$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{CC_z}{k_z}+\frac{1}{k_z}$
TNDCC[21]	$S_{xy}=\sum_{\substack{z\in\Gamma(x)\cap\Gamma(y)\\ t\in\{\Gamma[\Gamma(x)]\cap\Gamma(y)\}\cup\{\Gamma[\Gamma(y)]\cap\Gamma(x)\}}}\frac{CC_z}{k_z}+\frac{k_z+k_t}{k_zk_t}$
CN2D[22]	$S_{xy}=\lvert\Gamma(x)\cap\Gamma(y)\rvert+\beta\left(\frac{\sum_{z\in\Gamma(x)\cap\Gamma(y)}k_z}{\max(k_x,k_y)}\right)$

在以上基于相似性的链路预测算法的指标中, CN 指标和 JAC 指标考虑节点之间共同邻居的数量; AA 指标、RA 指标和 ERA 指标考虑共同邻居节点度的影响, 认为度小的共同邻居节点对连边影响要更大; PA 指标将待预测节点对的度作为相似性的衡量指标; CCLP 指标利用共同邻居节点的聚类系数定义节点间的相似性; NDCC 指标和 TNDCC 指标在共同邻居算法的基础上, 考虑节点聚类系数和度这两种重要的拓扑属性; CN2D 指标利用节点的度分布以及共同邻居数量, 对节点间的相似性进行估计.

2 基于节点交互度的社会网络链路预测方法

在本节我们提出一种新的链路预测方法. 根据小世界网络的特性, 本方法从路径特征的角度对节点间信息交换的效率进行评估, 充分考虑社会网络的社会性和高聚集性, 并在没有公共邻居的节点之间的链路预测方面得到成功应用. 除此之外, 本方法从节点间拓扑结构的角度对公共邻居的局部信息进行分析, 将共同邻居节点的聚类系数和度作为计算节点交互度的重要信息, 提高对局部信息的有效利用, 达到提高预测精度的目的. 由于不同规模数据集中路径特征和局部信息的不同, 本方法基于网络数据集规模大小使用定义的参数α控制路径特征和局部信息的权重构成零和条件, 将计算得到的节点交互度作为链路预测相似性的衡量指标, 从而有效地提高了预测精度. 基于节点交互度的链路预测方法的基本原理是如果两个节点之间的交互度越大, 那么这两个节点之间的关系更为亲密, 则存在连边的可能性也就越大. 社会网络的网络拓扑结构复杂, 节点本身蕴含着丰富的属性特征, 依据节点间的路径特征和局部信息可以有效地挖掘社会网络中节点间的交互度. 基于以上分析, 本文融合路径特征和局部信息对社会网络节点交互度进行定义, 并分别从上述两个角度对节点间的相似性进行度量.

2.1 小世界网络下的节点效率

小世界网络模型在图像上介于规则网络和随机网络之间, 其特点是节点之间特征路径长度小, 接近随机网络, 而聚合系数依旧相当高, 接近规则网络. 在社会网络这种图结构中, 绝大多数节点之间并不相邻, 但任意给定节点的邻居们却很可能彼此相邻, 并且大多数任意节点, 都可以用较少的步数访问到其他节点. 其次社会网络内部存在聚集的子网络, 这种子网络的特点是网络内部几乎任意两个节点之间都存在连接. 具体体现在一些彼此并不相识的人, 却可以通过一条很短的熟人链条被联系在一起; 具有相同社会背景的人形成小团体, 团体内部彼此关系紧密. 这说明社会网络中同样存在小世界现象.

Latora 等[23]用小世界网络的效率概念衡量了网络信息交换的效率, 认为图中一对节点的效率是节点间最短路径距离的乘积倒数, 图的全局平均效率是所有节点对的平均效率. 显然, 若两节点间最短路径长度越短, 只需经过较少节点就能相互访问, 说明节点间的效率越高, 交互关系也更为密切.

2.2 节点交互度的分析与量化

在社会网络的结构中, 两个节点之间的共同邻居节点会充当信息载体的角色, 源节点的信息在共同邻居节点的传递下发送到目的节点. 因此节点之间的公共邻居是衡量两个节点相似性的重要依据, 并且公共邻居的聚类系数和度也可以表达节点之间链路结构的信息.

图 1 中, 节点 1 和节点 2 是一组待预测的节点对, 节点 3 是它们的共同邻居节点. 从图 1 中可以看出, 与节点 3 相邻的节点彼此之间没有任何连边, 是相互独立的, 此时节点 3 的邻居节点彼此相互联系的程度处在较低水平.

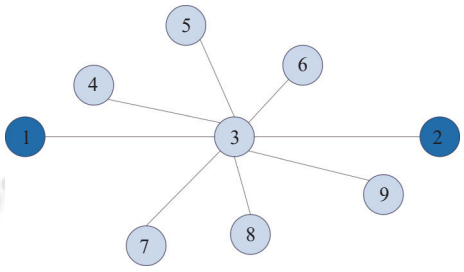

图 1 示例网络 1

图 2 中, 节点 7 和节点 8 建立连接, 以及节点 8 和节点 9 也形成连边, 此时节点 3 作为节点对的共同邻居节点发挥联系作用, 让节点对彼此建立连接. 那么作为共同邻居节点, 节点 3 使得待预测节点之间交互度更大, 从而相邻节点之间产生连接的可能性更高.

图 3 中, 节点 3 的度相较于图 2 中节点 3 的度有所减少, 在以节点 3 为共同邻居的节点对之间, 彼此产生连接的数量比例显著增加. 另一方面, 节点 3 作为共同邻居节点承担传递信息的作用, 在节点 3 的度减少的同时, 节点 3 的邻居节点得到的信息会相应增加. 这说明节点 3 作为待预测节点对的共同邻居节点, 对待

预测节点之间的交互度的贡献上升, 连接可能性也会更大.

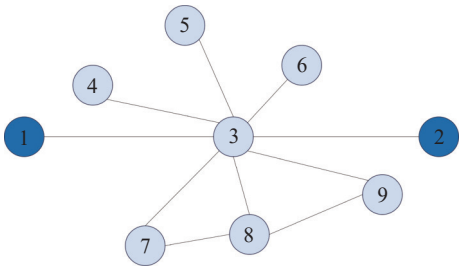

图 2 示例网络 2

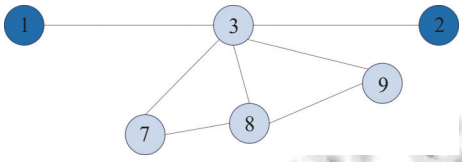

图 3 示例网络 3

节点的聚类系数使用了共同邻居的邻居节点之间的连接, 描述了该节点的相邻节点之间也相互连接的可能性, 并且这些连接和链路预测的连边概率具有相同的拓扑结构[18]. 节点的度描述了与该节点相关联的节点个数, 由于节点通过邻居节点作为信息载体来传递信息, 那么经由邻居节点传递的这个信息将会在接下来的信息传递再均匀分布给它的所有邻居[15]. 基于以上分析, 待预测节点之间共同邻居节点的聚类系数越大或节点的度越小, 那么待预测节点之间的交互关系会更强, 彼此之间更为相似, 存在连接的概率也就越高.

显然, 仅考虑节点局部信息而忽视路径特征难以应用于没有公共邻居的节点之间的链路预测. 基于上述分析, 对于节点相似性的度量可以从共同邻居的拓扑属性和节点效率两个方面进行考察. 节点间共同邻居的信息可作为待预测节点对的局部信息, 而节点效率是通过分析节点间路径特征得到的. 因此, 本文用路径特征和局部信息描述节点间相似性定义节点交互度.

定义 1. 节点交互度. 对社会网络中的任意两节点 x和节点 y, 节点交互度可用于对节点之间的相似性进行度量, 考虑路径特征和局部信息后并通过归一化处理使得前后量纲统一, 计算表达式如下:

$$S_{xy}=\alpha\frac{1}{d_{xy}}+(1-\alpha)\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{1+\mathrm{e}^{-\frac{CC_z}{k_z}}} \tag{1}$$

其中, d_{xy}为节点x和节点y之间的最短路径长度; $\Gamma(x)$为节点x的邻居节点; CC_z为节点x和节点y的共同邻居节点z的聚类系数; k_z为节点x和节点y的共同邻居节点z的度; α为可调节参数, 用于根据网络具体拓扑结构的不同进行调节.

本文将上述所提方法命名为节点交互度 (interacting degree of nodes, IDN), 节点交互度可作为节点x和y形成连边的可能性, 实现社会网络的链路预测. 此方法相较于传统基于相似性的链路预测指标, 适用于较大规模网络, 通过路径特征和局部信息更为全面地考量节点相似性. 对于小规模网络数据集, 本算法在考虑节点间局部信息的同时, 会优先考虑完整的网络拓扑结构信息和路径特征, 进而克服在小规模数据集上预测精度受限的问题. 对于较大规模数据集而言, 本算法则是主要对节点间局部信息进行充分利用, 减少路径特征信息在节点交互度所占的权重, 在一定程度上降低了计算复杂度. 本算法创新性地整合了路径特征和局部信息, 基于数据集规模的不同提出社会网络节点交互度的定义, 对节点间相似度进行准确刻画, 在预测精度和计算复杂度两个方面实现了有效的平衡, 从而增强对不同网络链路的预测能力, 降低了计算复杂度.

2.3 算法流程

在算法 1 中, α是可调节参数, 并且取值范围为 (0, 1). 由于α的最佳取值受到不同网络拓扑结构的影响, 这里将取值范围内的取不同值的α值代入算法 1, 不同的α取值会生成不同的链路预测结果, 本文从不同预测结果中选择结果最优时的参数值, 将其作为该网络的最优参数. 此时把参数α的最优取值代入到 IDN 算法中, 得到节点的相似性分数矩阵. IDN 算法的具体步骤如算法 1 所示.

算法 1. 基于节点交互度的链路预测算法

输入: 无向网络$G=(V,E)$, 参数α;
输出: 节点相似性分数矩阵 $S=[s_{ij}]$.

1) 初始化节点相似性分数矩阵 $S=[s_{ij}]$, 节点间最短距离矩阵 $D=[d_{ij}]$;
2) 计算每两个节点之间最短路径长度, 并更新节点间最短距离矩阵 $D=[d_{ij}]$;
3) 通过数据预处理, 得到每个节点的聚类系数和度;
4) 根据式 (1) 计算节点之间的相似性分数;
5) 更新并生成节点相似性分数矩阵 $S=[s_{ij}]$.

3 实验设置

本实验将来自不同领域的 6 个网络将分为两部分, 分别是训练集 E_Training 和测试集 E_Test, 两部分的比例为 8:2, E=E_Training∪E_Test, E_Training∩E_Test =

∅. 本文介绍的算法和基准方法将应用于这些训练集, 获得链路预测的相似性分数矩阵.

3.1 实验数据

在本实验中, 我们将使用从公共学术网站下载的 6 个真实世界网络数据来验证我们的算法, 并将结果与上面介绍的基准算法进行比较. 这些网络通常用于链路预测研究, 分别包括: Dolphins[24], Polbook[25], FWFW[26], Jazz[27], USAir[28], Polblogs[29], 每个数据集的简要描述如下.

(1) Dolphins 网络: 这是海豚种群中成员关系构建的网络, 节点表示海豚, 节点之间的连边表示种群成员之间的联系.

(2) Polbook 网络: 这是亚马逊商城的买者购买图书情况的网络, 节点表示出售的有关美国政治的图书, 节点之间的连边表示同一买家频繁共同购买图书.

(3) FWFW 网络: 这是佛罗里达海湾雨季的食物链网络, 节点表示生物, 节点之间的连边则表示生物之间的捕食关系.

(4) Jazz 网络: 这是爵士音乐家合作网络, 节点表示爵士音乐家, 节点之间的连边则表示音乐家之间是朋友关系.

(5) USAir 网络: 这是美国航空网络, 节点表示机场, 节点之间的连边表示机场之间有直飞航线.

(6) Polblogs 网络: 这是博客网页之间的超链接关系构成的网络, 节点表示博客网页, 节点之间的连边则表示博客网页之间的包含有超链接.

本文使用到的上述网络数据集主要参数如表 3 所示, 其中 N 表示网络中节点数, M 表示网络中连边数, <d>表示网络平均路径, <k>表示网络中节点的平均度, <c>表示网络平均聚类系数, D 表示网络密度.

表 3 6 个真实数据集的网络属性

Network	N	M	<d>	<k>	<c>	D
Dolphins	62	159	3.36	5.13	0.26	0.08
Polbook	105	441	3.08	8.40	0.49	0.08
FWFW	128	2 106	1.78	32.42	0.34	0.26
Jazz	198	2 742	2.24	27.70	0.62	0.14
USAir	332	2 126	2.74	12.81	0.75	0.04
Polblogs	1 490	19 090	2.74	27.36	0.36	0.02

3.2 评价指标

AUC 是一种在链路预测中最常用的评价指标, 可以解释为在测试集中随机选择的缺失链路被分配的相似性高于在未知链路集中随机选择的链路的概率[30]. 当在随机情况下, $AUC\approx0.5$, 一个链路算法的性能越好, 那么它的 *AUC* 会越接近于 1. 在实验中, 我们独立地比较n次, 其中有n_1次测试集中的边的分数值大于不存在的边的分数值, 有n_2次两者分数相等, 则 *AUC* 的计算公式为:

$$AUC=\frac{n_1+0.5n_2}{n} \tag{2}$$

Precision 同样是链路预测中常见的评价指标, 相比于对预测整体准确性进行评价的 *AUC*, 它更侧重于关心前L个预测节点对中预测正确的比例[31]. 在本文中, 我们设置L=100, 即在实验中若排在前 100 个预测节点对中有m个节点对在测试集中, 则 *Precision* 的计算公式为:

$$Precision=\frac{m}{L} \tag{3}$$

4 实验与结果分析

4.1 *AUC* 评价方法

首先我们对所提出的 IDN 算法在 *AUC* 评价指标下的性能与基准算法进行比较. 图 4 给出了不同网络中 IDN 算法与其他算法的 *AUC* 值. 从图 4 中可以看出, 与其他基准算法相比, IDN 算法在 Dolphins、Polbook、Jazz、USAir、Polblogs 这 5 个不同的网络数据集上都取得了 *AUC* 的最大值, 预测结果平均提升比例为 22%, 这也验证了考虑节点交互度的 IDN 算法在链路预测中能更有效地挖掘网络拓扑结构. 需要注意的是在 FWFW 数据集上, 各算法的 *AUC* 表现都不理想, 而除 PA 算法外其他算法 *AUC* 值均小于 0.7, 本文算法并未发挥出优势. 造成此现象的原因在于 PA 算法不需要节点的邻域信息, 仅考虑待预测节点对的度, 而该数据集网络密度较大, PA 算法得以发挥优势, 这说明网络中节点的度对预测结果有很大的影响.

图 5 给出了不同的链路预测方法在各数据集中的 *AUC* 值的堆积柱形图. 从图 5 中可以看出, IDN 算法每种颜色的面积几乎均匀, 说明该算法对各个网络数据集能够较为稳定地预测. 而 CN 算法、AA 算法、JAC 算法每种颜色面积差别很大, 表明这些算法的稳定性相对较差. 虽然在 FWFW 数据集上 IDN 算法的 *AUC* 值要低于 PA 算法, 但 IDN 算法的 *AUC* 累计值最大, 相反 PA 算法和 JAC 算法的累计值要处于劣势, 这也验证了本文基于节点交互度的链路预测算法有效性.

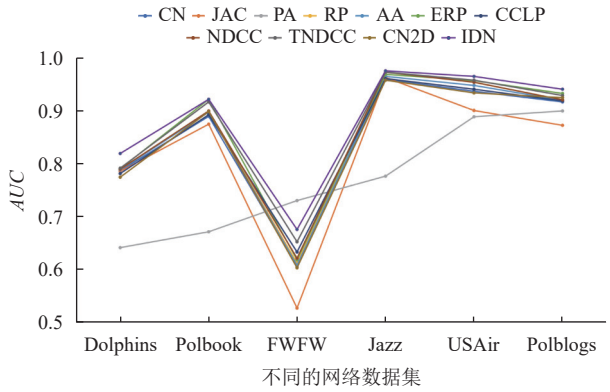

图 4 各算法对不同网络进行链路预测的 *AUC* 值实验结果对比

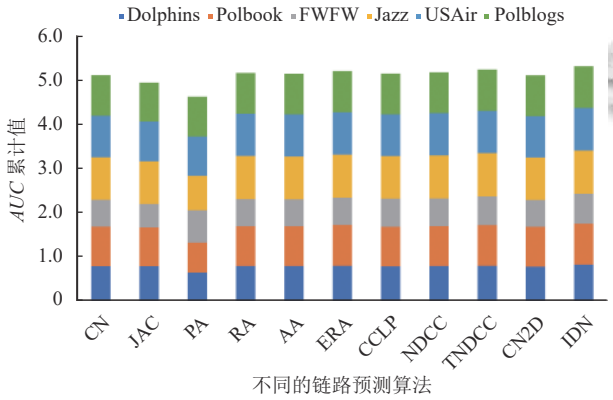

图 5 不同的链路预测算法在各数据集中 *AUC* 值的堆积柱形图

4.2 *Precision* 评价方法

为了进一步验证我们提出的 IDN 算法的性能, 本文在 *Precision* 评价指标下将本文算法与其他基准算法进行比较. 图 6 给出了不同链路预测算法在各网络中的 *Precision* 值. 可以看出, 与其他基准算法相比, IDN 算法在 Dolphins、Polbook、FWFW、Jazz、USAir 这 5 个不同的网络数据集的 *Precision* 值都是最高的, 在 *Precision* 评价指标下平均提升比例为 54%. 值得注意的是, JAC 算法和 TNDCC 算法在个别数据集中的 *Precision* 为 0, 这是因为 JAC 算法的原理是计算节点间公共邻居占彼此邻居数量的比例, 在遇到大规模数据集时, 比例急剧减小从而导致性能不佳; 而 TNDCC 算法结合两层节点度进行预测, 在遇到小规模数据集时, 也会导致预测效果不理想的情况. 而 PA 算法的思想是基于网络模型中新加入节点会倾向于和度大的节点相连的优先连接机制而提出的, 在这种机制的链路预测中, 节点链路预测的概率就正比于节点之间度的乘积, 在遇到网络密度较大数据集时有一定优势, 相反, 对一般数据集的预测性能较差.

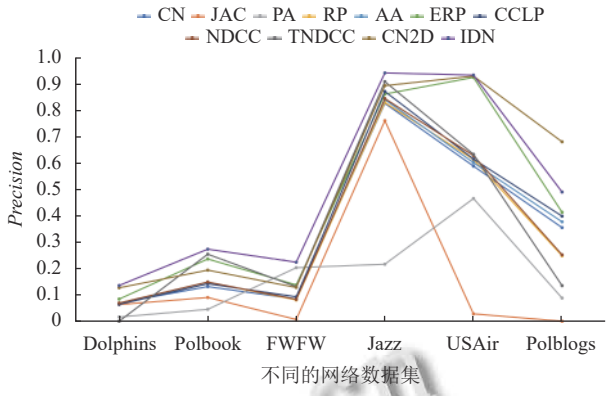

图 6 各算法对不同网络进行链路预测的 *Precision* 值实验结果对比

图 7 给出了不同的链路预测方法在各数据集中的 *Precision* 值的堆积柱形图. 从图 7 中可以明显看出, JAC 算法、PA 算法、TNDCC 算法在部分颜色的面积上有缺失且每种颜色之间的面积差别较大, 说明算法在不同数据集上的稳定性较差. 从整体看来, 各个算法在 Jazz 数据集上整体预测效果较好, 而在 Dolphins 数据集上的预测效果普遍较差. 相比于其他算法, 本文提出的 IDN 算法 *Precision* 累计值最大, 算法的预测结果也较为稳定. 这也进一步证明本文引入节点交互度作链路预测的性能优越性.

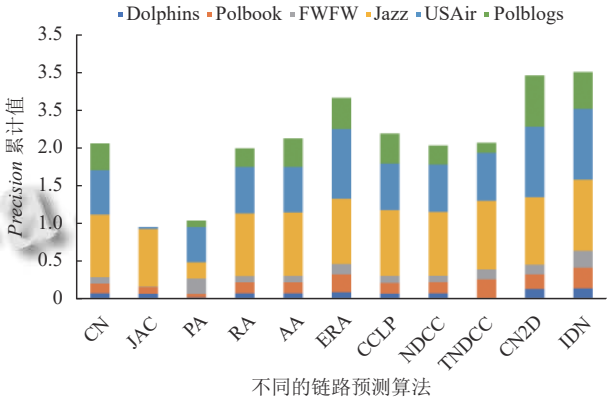

图 7 不同的链路预测算法在各数据集中 *Precision* 值的堆积柱形图

4.3 参数α对实验结果的影响

本文为了确定α的最优参数值, 分别设置了α={0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}的平均结果. 通过 6 个网络数据集对链接预测 *AUC* 和 *Precision* 评价指标进行计算, 本文方法的实验结果如图 8 和图 9 所示. 可以看出, 在不同的网络中, IDN 算法的性能对参数的敏感程度较低, 原因在于本文通过归一化处理统一量纲, 使得算法在保证性能的同时兼顾了稳定性. 实验还

发现, 本文方法对每个网络取得最佳预测性能对应的参数值并不相同. 在对实际网络应用中, 本文方法可根据网络拓扑结构适当调节α值, 提高链路预测的性能.

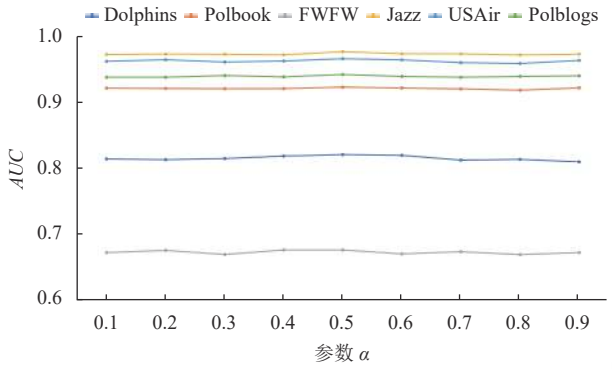

图 8 参数α对 *AUC* 实验结果的影响

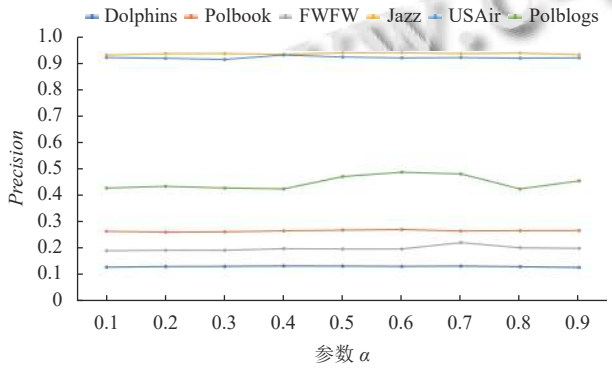

图 9 参数α对 *Precision* 实验结果的影响

4.4 实验结果分析

从上述实验结果来看, 在 6 个不同的网络数据集中, 本算法对比其他 10 个基准算法, *AUC* 和 *Precision* 指标最好的各有 5 个网络, 且在 *AUC* 和 *Precision* 两个评价指标上预测结果平均分别提升 22% 和 54%, 这些都进一步证明了 IDN 算法的优越性. 通过分析本算法的优势, 主要有以下 3 个方面. 首先, 与传统基于局部信息相似性的方法相比, 本算法考虑了节点在网络中信息交换的效率; 其次, 基于全局相似性方法通过考虑网络中的所有路径来预测链路的存在, 本算法使用较少的路径信息来降低算法的计算复杂度; 第三, 本算法在衡量节点间相似性时考虑了共同邻居节点的拓扑信息, 并使用归一化处理统一量纲, 本算法在与其他算法的稳定性比较中表现最佳.

5 结语

本文通过分析网络拓扑结构中节点间的交互关系, 提出了一种基于节点交互度的社会网络链路预测方法. 与传统的基于相似性的链路预测方法不同, 该方法不仅引入了共同邻居的相关局部信息, 而且考虑了节点效率在节点交互所起到的作用, 对节点局部信息和路径特征进行分析, 达到深度挖掘节点信息的目的, 在相似性的计算中更具有准确性和稳定性, 充分考虑社会网络结构的社会性和高聚集性. 本文借助 6 个真实网络数据集对该方法进行了验证, 结果表明本文提出的方法在 *AUC* 和 *Precision* 两个评价指标上均表现出更优的预测性能, 证明了节点交互度在社会网络链路预测方面的可行性和有效性. 本文在接下来的研究中, 可以不局限于静态网络的链路预测方法, 通过挖掘时序性特征来研究动态网络的演化趋势, 对本文的方法作进一步改进.

参考文献

1 Chai L, Tu LL, Yu XY, *et al*. Link prediction and its optimization based on low-rank representation of network structures. Expert Systems with Applications, 2023, 219: 119680. [doi: 10.1016/j.eswa.2023.119680]

2 Zhao YC, Sun YY, Huang YN, *et al*. Link prediction in heterogeneous networks based on metapath projection and aggregation. Expert Systems with Applications, 2023, 227: 120325. [doi: 10.1016/j.eswa.2023.120325]

3 李巧丽, 韩华, 李秋晖, 等. 基于最优路径相似度传输矩阵的链路预测方法. 复杂系统与复杂性科学, 2023, 20(1): 9–17. [doi: 10.13306/j.1672-3813.2023.01.002]

4 Mishra S, Singh SS, Kumar A, *et al*. HOPLP-MUL: Link prediction in multiplex networks based on higher order paths and layer fusion. Applied Intelligence, 2023, 53(5): 3415–3443. [doi: 10.1007/s10489-022-03733-8]

5 Orzechowski KP, Mrowinski MJ, Fronczak A, *et al*. Asymmetry of social interactions and its role in link predictability: The case of coauthorship networks. Journal of Informetrics, 2023, 17(2): 101405. [doi: 10.1016/j.joi.2023.101405]

6 Liu B, Yang RM, Lü LY. Higher-order link prediction via local information. Chaos, 2023, 33(8): 083108. [doi: 10.1063/5.0135640]

7 Mishra S, Singh SS, Kumar A, *et al*. MNERLP-MUL: Merged node and edge relevance based link prediction in multiplex networks. Journal of Computational Science, 2022, 60: 101606. [doi: 10.1016/j.jocs.2022.101606]

8 Liu YJ, Liu SH, Yu FS, *et al*. Link prediction algorithm

based on the initial information contribution of nodes. Information Sciences, 2022, 608: 1591–1616. [doi: 10.1016/j.ins.2022.07.030]

9 郁湧, 王莹港, 罗正国, 等. 基于聚类系数和节点中心性的链路预测算法. 清华大学学报 (自然科学版), 2022, 62(1): 98–104. [doi: 10.16511/j.cnki.qhdxxb.2021.21.039]

10 Sarhangnia F, Mahjoobi S, Jamshidi S. A novel similarity measure of link prediction in bipartite social networks based on neighborhood structure. Open Computer Science, 2022, 12(1): 112–122. [doi: 10.1515/comp-2022-0233]

11 Ghasemi S, Zarei A. Improving link prediction in social networks using local and global features: A clustering-based approach. Progress in Artificial Intelligence, 2022, 11(1): 79–92. [doi: 10.1007/s13748-021-00261-3]

12 Zhu JX, Dai F, Zhao FQ, *et al*. Integrating node importance and network topological properties for link prediction in complex network. Symmetry, 2023, 15(8): 1492. [doi: 10.3390/sym15081492]

13 Lorrain F, White HC. Structural equivalence of individuals in social networks. The Journal of Mathematical Sociology, 1971, 1(1): 49–80. [doi: 10.1080/0022250X.1971.9989788]

14 Adamic LA, Adar E. Friends and neighbors on the Web. Social Networks, 2003, 25(3): 211–230. [doi: 10.1016/S0378-8733(03)00009-1]

15 Zhou T, Lü LY, Zhang YC. Predicting missing links via local information. The European Physical Journal B, 2009, 71(4): 623–630. [doi: 10.1140/epjb/e2009-00335-8]

16 Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019–1031. [doi: 10.1002/asi.20591]

17 Jaccard P. Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin de la Société Vaudoise des Sciences Naturelles, 1901, 37(142): 547–579. [doi: 10.5169/seals-266450]

18 Wu ZH, Lin YF, Wang J, *et al*. Link prediction with node clustering coefficient. Physica A: Statistical Mechanics and Its Applications, 2016, 452: 1–8. [doi: 10.1016/j.physa.2016.01.038]

19 白桦, 马云龙, 毕玉, 等. 一种基于节点局部相似性的复杂网络链路预测算法. 计算机应用与软件, 2020, 37(5): 298–301, 308. [doi: 10.3969/j.issn.1000-386x.2020.05.051]

20 高杨, 张燕平, 钱付兰, 等. 结合节点度和节点聚类系数的链路预测算法. 小型微型计算机系统, 2017, 38(7): 1436–1441. [doi: 10.3969/j.issn.1000-1220.2017.07.003]

21 陈紫扬, 张月霞. 结合二层节点度和聚类系数的链路预测算法. 计算机工程与应用, 2019, 55(23): 40–44. [doi: 10.3778/j.issn.1002-8331.1811-0185]

22 Mumin D, Shi LL, Liu L. An efficient algorithm for link prediction based on local information: Considering the effect of node degree. Concurrency and Computation: Practice and Experience, 2021, 34(7): e6289. [doi: 10.1002/cpe.6289]

23 Latora V, Marchiori M. Efficient behavior of small-world networks. Physical Review Letters, 2001, 87(19): 198701. [doi: 10.1103/PhysRevLett.87.198701]

24 Lusseau D, Schneider K, Boisseau OJ, *et al*. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 2003, 54(4): 396–405. [doi: 10.1007/s00265-003-0651-y]

25 Yang JX, Zhang XD. Predicting missing links in complex networks based on common neighbors and distance. Scientific Reports, 2016, 6(1): 38208. [doi: 10.1038/srep38208]

26 Ulanowicz RE, Bondavalli C, Egnotovich MS. Network analysis of trophic dynamics in South Florida ecosystems, FY 99: The graminoid ecosystem. Technical Report, Solomons: University of Maryland System Center for Environmental Science, Chesapeake Biological Laboratory, 2000.

27 Gleiser PM, Danon L. Community structure in Jazz. Advances in Complex Systems, 2003, 6(4): 565–573. [doi: 10.1142/S0219525903001067]

28 Batagelj V, Mrvar A. Pajek —Program for large network analysis. Connections, 1998, 21(2): 47–57.

29 Adamic LA, Glance N. The political blogosphere and the 2004 U.S. election: Divided they blog. Proceedings of the 3rd International Workshop on Link Discovery. Chicago: ACM, 2005. 36–43. doi: 10.1145/1134271.1134277.

30 Kumar A, Mishra S, Singh SS, *et al*. Link prediction in complex networks based on significance of higher-order path index (SHOPI). Physica A: Statistical Mechanics and Its Applications, 2020, 545: 123790. [doi: 10.1016/j.physa.2019.123790]

31 Zhou T, Lee YL, Wang GN. Experimental analyses on 2-hop-based and 3-hop-based link prediction algorithms. Physica A: Statistical Mechanics and Its Applications, 2021, 564: 125532. [doi: 10.1016/j.physa.2020.125532]

(校对责编: 孙君艳)