计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600031

一种基于加权网络拓扑权重的链路预测方法

袁 榕[1] 宋玉蓉[1] 孟繁荣[2]
1 南京邮电大学自动化学院人工智能学院 南京 210003
2 南京邮电大学计算机学院软件学院网络空间安全学院 南京 210003
(15738774278@163.com)

摘 要 近年来,复杂网络中的链路预测问题受到越来越多的关注,链路预测的应用场景也越来越广泛,因此如何提高链路预测精度是一个重要问题。目前已提出了很多方法,其中加权相似性指标的预测方法取得了很好的效果。然而传统的加权网络链路预测方法仅考虑了链接的自然权重,忽略了链接的拓扑权重对预测精度的影响。因此,针对加权网络的链路预测,综合考虑网络中边的聚类和扩散特性并将其作为边的拓扑权重,提出了基于链接拓扑权重的 WCD 含权预测指标,包括 WCD-CN,WCD-AA,WCD-RA 和 WCD-LP4 个相似性指标。文中以 Matlab 为实验平台,在两个带权数据集(USAir,Bibble)和两个无权数据集(Pblogs,Dolphins)上进行实验,并以 AUC 作为评价指标。仿真结果表明,与基于自然权重的含权指标、基于簇系数的结构含权指标相比,所提算法具有更好的预测精度。

关键词:复杂网络;拓扑结构;链路预测;相似性指标;结构权重
中图法分类号 TP393.02

Link Prediction Method Based on Weighted Network Topology Weight

YUAN Rong[1],SONG Yu-rong[1] and MENG Fan-rong[2]
1 College of Automation & College of Artificial Intelligence,Nanjing University of Posts and Telecommunications,Nanjing 210003,China
2 School of Computer,Network Space Security,Nanjing University of Posts and Telecommunications,Nanjing 210003,China

Abstract In recent years,with more and more attention drawning to link prediction in complex networks,and with the application of link prediction becoming increasingly extensive,a crucial question is raised on how to improve the accuracy of link prediction. Many proposals are made,among which the weighted similarity indices have already achieved a promising result. However,the traditional weighted network link prediction only considers the natural weight of the link neglects the influence of the topological weights on prediction accuracy. Therefore,aiming at the weighted networks,this paper takes the clustering and diffusion characteristics of edges into consideration and regard them as the topological weights of edges,and consequently recommended four similarity indices based on the topology weight of links,namely WCD-CN,WCD-AA,WCD-RA,and WCD-LP. This paper takes Matlab as the experimental platform and carries out experiments on two weighted datasets(USAir,Bibble) and two weightless datasets(Pblogs and Dolphins),in which AUC is used as the evaluation index. The results of the simulation indicate that compared with two weighted indices,which are based on natural weight and cluster coefficient respectively,the proposed algorithm has higher accuracy in prediction.

Keywords Complex network,Topological structure,Link prediction,Similarity index,Structural weight

1 引言

复杂网络中的链路预测指利用网络中节点的属性信息和已知的网络拓扑结构,预测网络中没有连边的两个节点之间产生连边的可能性,预测包含了对未知链接和未来链接的预测[1]。链路预测作为复杂网络的一个重要研究方向,在互联网[2]、生物网络[3]、移动通信网络[4]和社交网络[5]等领域得到了广泛应用。近年来,随着信息技术的进步和发展,对链路预测的精度要求也愈来愈高,如何提高预测精度成为目前链路预测面临的一个严峻挑战。

到稿日期:2019-06-06 返修日期:2019-09-22 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:国家自然科学基金(61672298,61873326,61373136,61802155);江苏高校哲学社会科学研究重点项目(2018SJZDI142);教育部人文社会科学研究规划基金(17YJAZH071)
This work was supported by the National Natural Science Foundation of China(61672298,61873326,61373136,61802155),Philosophy Social Science Research Key Project Fund of Jiangsu University(G2018SJZDI142) and Social Sciences of Ministry of Education of China(17YJAZH071).
通信作者:宋玉蓉(songyr@njupt.edu.cn)

当前，链路预测方法主要分为4类：基于节点属性相似性的方法、基于网络结构相似性的方法、基于似然分析的方法和机器学习方法。其中，基于网络结构相似性的方法因计算复杂度低、效率高且适用于大规模网络而受到了普遍关注。该方法主要通过网络的局部结构信息来预测链路。例如，Liben等[6]提出了利用网络拓扑结构来定义节点间的相似性这一概念。Yang等[7]考虑被预测节点间共同邻居较少的情况，提出了一种基于共同邻居和距离的链路预测算法，在预测无共同邻居节点之间的缺失链路方面取得了显著的效果。Wu等[8]利用被预测节点对共同邻居的聚类系数来定义相似性，提出了CCLP算法，该算法考虑了更多的网络局部结构信息，其理论依据是被预测节点对的共同邻居节点的邻居节点之间的链路和被预测节点对之间的链路具有相同的结构位置。Liu等[9]提出了度相关聚类系数来评价节点的聚类能力，因为当链路出现观测偏差时，度相关聚类系数在估计节点聚类能力方面表现出较高的鲁棒性，并据此提出了DCP算法来获得更高的链接预测精度。

随着复杂网络研究的不断深入，无向无权网络已经不能涵盖大部分的网络特性。大部分真实网络链接是带有权重的，比如共同作者网络，权重代表两位作者合作过的文章数量；航空网络，权重代表两个机场之间的航班频次。因此，针对加权网络的链路预测逐渐受到重视。研究表明，考虑了链接权重的预测方法可以有效提高预测精度[10]。Zhu等[11]基于局部网络结构的互信息提出了加权相似性指标，实验结果表明链路权重在链路预测中起到了积极作用。Sett等[12]提出了最小流和乘积加权模型，并考查了加性、最小流和乘积加权3种加权模型对预测方法和数据集的影响。结果显示，最小流和乘积加权两种加权模型在大多数数据集和预测方法中，预测效果均优于传统的加性加权模型。

以上针对加权网络的链路研究仅依据其链接的自然权重(以连边的自然属性定义链路权重，如边权表示论文合作数、航班频次等)，忽略了链接的拓扑结构权重对链路预测精度的影响[13]。近期的研究表明，链接的结构权重(通过网络的结构属性，如聚类系数、度分布、簇系数等定义链路权重)对预测也起到积极作用。例如，Zhu等[14]使用归一化聚类系数进行链接权重构造，从而提高了链路预测的精度。Wang等[15]将边的簇系数作为结构权重，实验结果表明，该结构权重在提高网络链接预测精度的过程中起到了重要作用。Yao[16]将边的链接度作为边的拓扑权重，提出了加权网络中基于链接耦合权重(Coupling Weight-Based)的CW-Based含权预测指标，实验验证该指标在链路预测中表现较好。可见，链接拓扑结构权重对链路预测精度也起到积极的作用。

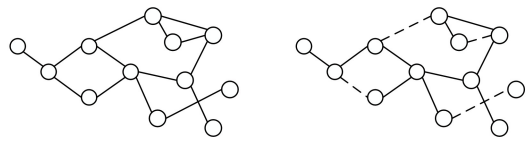
因此，本文以网络的链接拓扑结构权重为重点，基于WCN，WAA，WRA，WLP4个局部结构相似性指标，运用边的聚类和扩散特性[17]提炼链接拓扑结构权重，提出一种考虑边的聚类和扩散特性的含权预测算法——WCD(Weight of Clustering and Diffusion)含权预测算法。该算法中权重的设置充分利用了网络的拓扑结构特性，权威数据集上的实验结果表明，本文方法可显著提升预测精度。

2 链路预测

2.1 链路预测问题描述

给定一个加权网络 $G(V,E,W)$，其中 V，E 和 W 分别表示网络中节点、链接和权重的集合，网络中没有自连边。$w(e_{xy})=w(x,y)=w(y,x)$ 代表两个节点 x 和 y 之间边的权重值。为了找到该网络缺失的链接以及将来可能出现的链接，为每一对没有连边的节点对 $(x,y)\in V$ 分配一个相似度分数 S_{xy}，用来量化链接 $e_{xy}\in E$ 存在的可能性。相似度分数越高，意味着两个节点间的相似性越大，存在连边的可能性也就越大。所有未连接的节点对按相似度分数降序排列，排在前面的链接可被视为存在概率较高的链接。

为了检验本文算法的准确性，将链路集合 E 随机划分为训练集 E_T 和测试集 E_P，使得 $E_T\cup E_P=E$，$E_T\cap E_P=\varnothing$，如图1所示，测试网络中的虚线表示测试集，实线表示训练集。其中 E_T 被认为是已知信息，用来计算相似度分数，E_P 用来测试算法的精确度[18]。



(a)完整的网络 (b)测试网络

图1 网路链接预测示例

Fig. 1 Example of network link prediction

2.2 链路预测的相似性指标

本文研究工作主要基于局部结构的相似性指标。这里介绍几种考虑自然权重的相似性指标和考虑网络拓扑结构的链路权重，表1列出了文中提到的符号的含义。

表1 符号解释

Table 1 Symbolic interpretation

符号	含义
$\Gamma(x)$	节点 x 的邻居集合
w_{xz}	连接节点 x 和 z 的边的权重值
k_z	节点 z 的度
S_z	节点 z 的强度
N_{xy}	节点 x 和 y 的共同邻居的个数
S_{xy}	节点 x 和 y 的相似度分数

常见的考虑自然权重信息的加权相似性指标有WCN，WAA，WRA，WLP。

WCN[10]指标是含权的CN指标，其中 $\Gamma(x)\cap\Gamma(y)$ 表示节点 x 和 y 的共同邻居。其定义如下：

$$S_{xy}^{\mathrm{WCN}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}(w_{xz}+w_{zy}) \tag{1}$$

WAA[10]指标是含权的 AA 指标，其中 $S_z=\sum_{x\in\Gamma(z)}w_{xz}$ 表示节点 z 的强度。其定义如下：

$$S_{xy}^{\mathrm{WAA}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{w_{xz}+w_{zy}}{\log(1+s_z)} \tag{2}$$

WRA[19]指标是含权的 RA 指标，其定义如下：

$$S_{xy}^{\mathrm{WRA}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{w_{xz}+w_{zy}}{s_z} \tag{3}$$

WLP[20]指标是含权的 LP 指标，其定义如下：

$$S_{xy}^{\mathrm{WLP}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}(w_{xz}+w_{zy})+\varepsilon\sum_{(i,j)\in l_{x\to y}}(w_{xi}+w_{ij})(w_{ij}+w_{jy}) \tag{4}$$

其中，ε 为可调参数，本文 ε 为常用值 0.001。

此外，考虑网络拓扑结构的链路权重有基于簇系数的拓扑权重[15]、链接度和自然权重的耦合权重[16]。

簇系数拓扑权重的定义为：

$$TW_{xy}=\frac{N_{xy}}{\max[k_x,k_y]} \tag{5}$$

链接度和自然权重的耦合权重定义为：

$$TW_{xy}=\frac{w_{xy}}{k_xk_y} \tag{6}$$

3　基于边的聚类和扩散特性的含权预测算法

3.1　基于边的聚类和扩散特性的链接权重

从主观角度来说，网络中的新链接往往由事件、合作、利益、共同兴趣等信息流的驱动而产生，并朝着利于网络繁殖的方向发展。文献[21]表明，信息的传播与边的重要性有着很大关系，边的重要性又对于网络拓扑结构的形成起到重要作用。因此，在链路预测中，对网络连边拓扑特性的考量，有利于提高预测的精准度。

下面以信息传播为例，简述聚类特性和扩散特性对网络链接形成的影响，如图 2 所示。

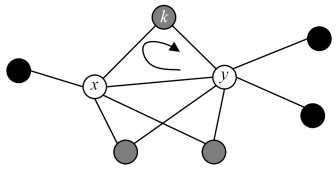


图 2　边 e_{xy} 的局部拓扑结构

Fig. 2　Local Topological Structure of Edges e_{xy}

如果只考虑网络中边的聚类特性，那么图 2 中灰色节点的数目越多，以 e_{xy} 为边的三角形个数就越多，相应的边 e_{xy} 的聚类特性也越高，其重要性也越高。但若仅考虑边的聚类特性，信息流就更易在三角形结构中流转，使其在整个网络中的流通性不好，因此还要考虑网络中边的扩散特性。图 2 中黑色节点的个数代表了边 e_{xy} 的扩散能力，黑色节点越多，说明连边的扩散特性越好。但是对于网络边缘的一些边，虽然其扩散能力较强，但聚类特性较差，不利于网络的进一步发展。

可见，单独考虑边的聚类特性和扩散特性，都会导致信息流通性都不佳。因此，本文综合考虑边的聚类特性和扩散特性[17]，并引入参数 α 作为衡量这两个拓扑结构特性的相对重要程度的因子。

边 e_{xy} 的聚类特性表示如下：

$$JL(e_{xy})=|\{\Delta_{xyk}:\Delta_{xyk}\in\Delta_G\}| \tag{7}$$

其中，Δ_{xyk} 表示以 x,y,k 为顶点组成的三角形，Δ_G 表示网络中所有的三角形构成的集合，$JL(e_{xy})$ 表示以节点 x,y 为顶点组成三角形的数目。$JL(e_{xy})$ 值越大，边 e_{xy} 的聚类特性越好。

边 e_{xy} 的扩散特性表示如下：

$$KS(e_{xy})=\sum_{k\in\Gamma(x,y)\backslash x,y}\theta_k \tag{8}$$

其中，$\Gamma(x,y)$ 表示节点 x,y 的邻居节点集合，$\Gamma(x,y)\backslash x,y$ 表示节点 x，y 的邻居节点集合中去除节点 x 和 y 后构成的集合，$\sum_{k\in\Gamma(x,y)\backslash x,y}\theta_k$ 表示在集合 $\Gamma(x,y)\backslash x,y$ 中不能与节点 x,y 构成三角形的节点的集合。

在此基础上，本文定义了基于边的聚类与扩散特性的链接拓扑权重。在加权网络 $G(V,E,W)$ 中，边 $e_{xy}=(x,y)\in E$，$x,y\in V$ 的拓扑权重表示如下：

$$CD(e_{xy})=\alpha\times JL(e_{xy})+(1-\alpha)\times KS(e_{xy}),\alpha\in(0,1) \tag{9}$$

其中，$CD_{xy}=CD_{yx}$，CD 代表连边的拓扑权重；α 为可调参数，用于衡量聚类特性和扩散特性之间的相对重要程度。

3.2　基于边的聚类和扩散特性的含权预测方法

本文 2.2 节介绍的 WCN，WAA，WRA 和 WLP 4 个加权相似性指标都可以看作基于共同邻居信息并结合链接自然权重的函数，即：

$$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}f(w_{xz},w_{yz}) \tag{10}$$

WCN，WAA，WRA 和 WLP 指标都是这个函数的展开。其中，w_{xy} 表示连接节点 x 和 y 的边的自然权重值。$\Gamma(x)\cap\Gamma(y)$ 表示节点 x 和 y 的共同邻居，$f(w_{xz},w_{yz})$ 表示共同邻居节点 z 和被预测节点对两端点 x,y 之间链接的自然权重函数。

本文将链路的自然权重扩展到拓扑权重，提出考虑边拓扑权重的加权预测，基于链接拓扑权重的函数表示如下：

$$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}f(TW_{xz},TW_{yz}) \tag{11}$$

其中，TW_{xz} 表示连接节点 x 和 z 的边的拓扑权重值；$z\in\Gamma(x)\cap\Gamma(y)$ 表示节点 x 和 y 的共同邻居，$f(TW_{xz},TW_{yz})$ 表示共同邻居节点 z 和被预测节点对两端点 x,y 之间链接的拓扑权重函数。

这里，进一步将式(11)推广到 WCN，WAA，WRA，WLP 指标，并使 $TW_{xz}=CD_{xz}$，提出扩展 WCN，WAA，WRA 和 WLP 的 WCD 含权预测方法，其对应的预测指标定义如下：

$$S_{xy}^{\text{WCD-CN}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}(CD_{xz}+CD_{yz}) \tag{12}$$

$$S_{xy}^{\text{WCD-AA}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{CD_{xz}+CD_{yz}}{\log(1+s_z)} \tag{13}$$

$$S_{xy}^{\text{WCD-RA}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{CD_{xz}+CD_{yz}}{s_z} \tag{14}$$

$$S_{xy}^{\text{WCD-LP}}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}(CD_{xz}+CD_{yz})+\varepsilon\sum_{(i,j)\in l_{x\to y}}(CD_{xi}+CD_{ij})(CD_{ij}+CD_{jy}) \tag{15}$$

其中，$s_z=\sum_{x\in\Gamma(z)}CD_{zx}$ 表示共同邻居节点 z 和式(12)－式(15)邻居节点之间链接的拓扑权重之和。上述方法被称为基于拓扑结构权重的 WCD 含权预测指标。

WCD 含权预测函数的流程如算法 1 所示。

算法 1　WCD 含权预测算法

输入：4 个真实网络(network)的边列表

输出：评价指标结果(AUC)

Step1　将输入的数据集转换为相应的邻接矩阵。

Step2　计算每条边的聚类特性和扩散特性，并计算出 CD(e_{xy})。

Step3　生成带权重的邻接矩阵并将其划分为训练集和测试集。

Step4　计算被预测节点对 x，y 的共同邻居节点个数。

Step5　根据式(12)－式(15)分别计算被预测节点对 x，y 之间的相似度分数 S_{xy}，并生成相应的相似度矩阵。

Step6　根据相似度矩阵和测试数据集计算评价指标 AUC。

4 实验结果与分析

4.1 实验数据

本文使用 Matlab 作为实验平台，在以下来自不同领域的 4 个真实网络上测试所提算法的准确度。

美国航空网络(USAir)[22]，该网络中的节点代表机场，连边代表机场之间的航班数目。

Bibble 网络[23]，该网络中的节点代表《圣经》中的人物，连边代表两个人物在相同章节同时出现的次数。

美国政治博客网络(Pblogs)[24]，即美国政治博客之间有指向的超链接网络。

海豚社交网络(Dolphins)[25]，该网络中的节点代表海豚，连边代表海豚成对出现的次数高于预期次数。

4 个真实网络的网络拓扑特征如表 2 所列，其中 N 代表网络节点数，M 代表网络连边数，$\langle k\rangle$ 表示网络的平均度，$\langle d\rangle$ 表示网络的平均距离，ρ 为网络密度，C 为网络的聚类系数。

表 2 4 个真实网络的基本统计特征

Table 2 Basic statistical characteristics of four real networks

Networks	N	M	$\langle k\rangle$	ρ	$\langle d\rangle$	C
USAir	332	2126	12.8	0.039	2.74	0.74
Bibble	1773	9131	10.3	0.006	3.38	0.72
Pblogs	1224	19090	27.3	0.026	2.73	0.36
Dolphins	61	159	5.13	0.083	3.36	0.26

在 USAir 数据集中，α 取 0.4；在 Bibble 数据集中，α 取 0.6；在 Pblogs 数据集中，α 取 0.6；在 Dolphins 数据集中，α 取 0.7。

4.2 实验设置

为评估预测结果的准确性，本文使用 AUC(Area Under the receiver operating characteristic Curve)作为精度测量。AUC 指 ROC 曲线下的面积，在实际计算过程中，由于数据过大，一般采取抽样的方法。AUC 可以理解为在测试集中随机选择一条边的分数值比随机选择的一条不存在的边的分数值高的概率[26]。在 n 次独立比较的过程中，随机选择一条缺失链接和一条不存在的链接来比较它们的相似度分数，如果有 n' 次测试集中边的相似度分数大于不存在的边的相似度分数，有 n'' 次二者的相似度分数相同，那么 AUC 的定义如下：

$$AUC=\frac{n'+0.5n''}{n}$$

对于数据集中训练集和测试集的比例，本文进行了多次实验。最终，将训练集和测试集的比例划分为 9∶1。

本文所提预测指标为 WCD-CN，WCD-AA，WCD-RA 和 WCD-LP，令式(11)中的 $TW_{xz}=CD_{xz}$。本文比较了式(11)中的 TW_{xz} 为不同值时，相应预测指标的 AUC 值。以下为 TW_{xz} 的不同取值：

$TW_{xz}=1$(即为无权状态下)

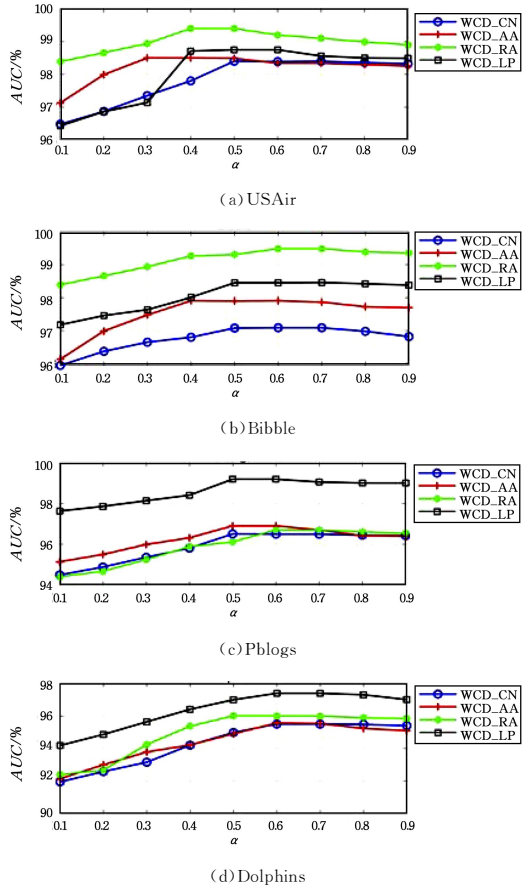
$TW_{xz}=w_{xz}$(即为自然权重)

$TW_{xz}=\frac{N_{xz}}{\max[k_x,k_z]}$(即为簇系数拓扑权重[15])

$TW_{xz}=\frac{w_{xz}}{k_xk_z}$(即为链接度与自然权重耦合权重[16])

4.3 关参数对链路预测精度的影响分析

本文将式(9)中的 $CD(e_{xy})$ 作为边 e_{xy} 的权重，并将此权重运用到 WCD 含权预测指标中，式(9)中的参数 $\alpha(\alpha\in(0,1))$ 为可调参数，因此，本节研究 α 的变化对链路预测精度的影响，如图 3 所示。



(a) USAir

(b) Bibble

(c) Pblogs

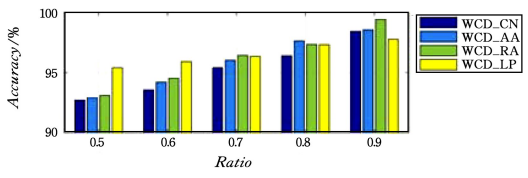
(d) Dolphins

图 3 α 值对实验结果的影响

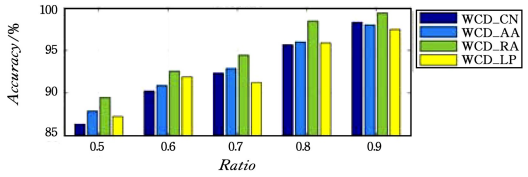
Fig. 3 Influence of different α values on the experimental results

可以看出，当 α 值逐渐增大时，聚类特性所占的比例增加，扩散特性的比例减少；当 α 增大到一定阶段时，两者维持平衡，预测准确率维持在一个稳定值；当 α 继续增加时，准确率开始降低，这表明扩散特性是不可或缺的，从而验证了本文方法的正确性。

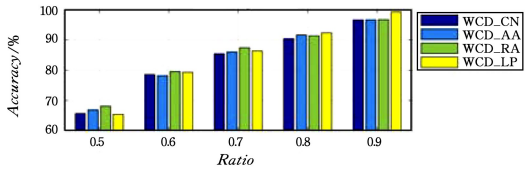
图 4 给出了在数据集 USAir，Bibble，Pblogs 和 Dolphins 上，不同的训练集比例对实验结果的影响。实验中，训练集占整个数据集的比例分别为 50%，60%，70%，80%，90%，每个训练集比例下的 AUC 值均为随机划分 1000 次后的平均值。由图 4 可以得知，随着训练集比例的增大，预测准确度也相应提高。以 USAir 网络为例，当训练集比例为 50%时，WCD-CN 指标的精确度为 0.927；当训练集比例为 60%时，精确度为 0.936，提高了将近 1%；当训练集比例增加到 90%时，精确度为 0.981，较之前提高了 6%左右，这是因为训练集比例越大，可利用的网络拓扑信息就越多，计算出的网络拓扑权重就更加精确，基于链接拓扑权重的 WCD 含权预测指标的精确度越高。上述结果在其他 3 个网络均有体现。因此，在划分网络训练集和测试集时，设置训练集占原网络数据的 90%，测试集占 10%。



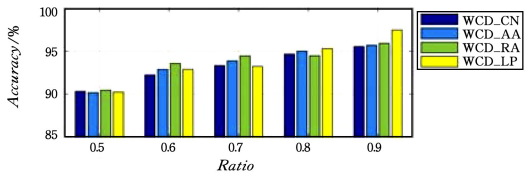
(a) USAir



(b) Bibble



(c) Pblogs



(d) Dolphins

图 4　不同训练集比例对 *AUC* 值的影响

Fig. 4　Influence of different training set proportion on *AUC* value

4.4　不同加权方式对预测精度的影响分析

针对自带权重的网络 USAir 和 Bibble，分别计算本文所提 WCD 含权预测指标，即 $TW_{xz}=CD_{xz}$ 和当 $TW_{xz}=w_{xz}$（即为自然权重）、$TW_{xz}=\frac{w_{xz}}{k_x k_z}$（即为耦合权重[16]）时，对应的扩展 WCN，WAA，WRA 和 WLP 指标的 *AUC* 值，如表 3 所列。可以看出，考虑上述 3 种加权方式时，本文提出的 WCD 含权预测指标在耦合权重预测指标的基础上，预测精度有所提升，证明了该方法对于加权网络有很好的适用性。可以注意到，由于 USAir 网络中存在“弱连接”效应，考虑链接自然权重的预测指标的精确度低于不含权的预测方法的精确度，但是当综合考虑边的聚类和扩散特性的拓扑权重时，预测精度有了明显的提升，这说明本文提出的加权方法弥补了自然权重方法的不足。

表 3　加权网络上的 *AUC* 值

Table 3　*AUC* values on weighted networks

网络		*WCN*	*WAA*	*WRA*	*WLP*
USAir	无权	0.955	0.967	0.973	0.964
	自然权重	0.948	0.961	0.968	0.963
	耦合权重	0.963	0.974	0.971	0.974
	CD	0.984	0.985	0.994	0.987
Bibble	无权	0.977	0.985	0.986	0.986
	自然权重	0.978	0.986	0.987	0.987
	耦合权重	0.988	0.987	0.988	0.987
	CD	0.983	0.979	0.994	0.974

对于无权网络 Pblogs 和 Dolphins，考虑将链接的拓扑权重作为无权网络的权重值。分别计算本文所提 WCD 含权预测指标，即 $TW_{xz}=CD_{xz}$ 和当 $TW_{xz}=1$（即为无权）、$TW_{xz}=\frac{N_{xz}}{\max[k_x,k_z]}$（即为簇系数拓扑权重[15]）时，对应的扩展 WCN，WAA，WRA 和 WLP 指标的 *AUC* 值，如表 4 所列。可以看出，对于无权网络，使用簇系数拓扑权重加权网络可以使预测效果有很好的提升，本文方法通过考虑网络中的聚类和扩散特性来加权网络，使得预测效果有了进一步的提升。

在算法的时间复杂度方面，设 n 为网络中节点的个数，WCN 算法需要计算共同邻居矩阵，首先要找到需要预测的两个节点，然后再找到两个节点的共同邻居节点以及连边的权重值，因此算法的时间复杂度为 $O(n^2)$。WAA 和 WRA 算法计算完两个节点共同邻居之后，又计算了共同邻居的强度，所以其时间复杂度和 WCN 算法相同。而计算权重矩阵的时间复杂度也为 $O(n^2)$，因此 WCD 算法的时间复杂度为 $O(n^2)$，在算法的时间复杂度不大的情况下，本文算法的预测精度有所提升。

表 4　无权网络上的 *AUC* 值

Table 4　*AUC* values on weightless networks

网络		*WCN*	*WAA*	*WRA*	*WLP*
Pblogs	无权	0.918	0.920	0.922	0.929
	簇系数	0.924	0.925	0.925	0.932
	CD	0.965	0.969	0.967	0.992
Dolphins	无权	0.828	0.825	0.822	0.828
	簇系数	0.890	0.910	0.913	0.934
	CD	0.955	0.957	0.960	0.974

结束语　将网络的拓扑结构信息作为权重进行链路预测可以有效提高链路预测的准确度。然而，获得网络的全局拓扑信息进行链路预测的算法计算成本较高，不适合大规模网络。

本文从网络的局部结构信息出发，综合考虑了边的聚类特性和扩散特性之间的制约关系，定义了 *CD* 函数，作为网络中边的拓扑权重值，并进一步将其推广至 WCN，WAA，WRA 和 WLP 指标，提出了基于链接拓扑权重的 WCD-CN，WCD-AA，WCD-RA，WCD-LP 指标。将这 4 个指标在 4 个真实网络上与其他经典的基于自然权重的含权相似性指标 WCN，WAA，WRA，WLP 和基于簇系数的结构含权指标等进行比较，实验结果表明，在大多数情况下，本文提出的 WCD 含权预测指标的预测性能要优于其他指标。在算法的时间复杂度方面，本文算法在保持了相对较低的时间复杂度的同时，取得了更优的预测精度。

本文研究的链路预测问题仅限于加权网络，后续将研究更加复杂的网络，如有向网络和多层网络等，利用层间的相关性来提高预测精度。

参 考 文 献

[1] LV L Y. Link prediction on complex networks [J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 651-661.

[2] KIM J, KIM S, LEE C. Anticipating technological convergence: Link prediction using Wikipedia hyperlinks[J]. Technovation, 2019, 79: 25-34.

[3] KOVÁCS I A, LUCK K, SPIROHN K, et al. Network-based prediction of protein interactions[J]. Nature Communications, 2019, 10(1): 1240.

[4] SUFIAN A, SULTANA F, DUTTA P. Data Load Balancing In Mobile Ad Hoc Network Using Fuzzy Logic(DBMF)[J]. arXiv: 1905.11627.

[5] AHUJA R, SINGHAL V, BANGA A. Using Hierarchies in Online Social Networks to Determine Link Prediction[M]// Soft Computing and Signal Processing. Singapore: Springer, 2019: 67-76.

[6] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.

[7] YANG J, ZHANG X D. Predicting missing links in complex networks based on common neighbors and distance[J]. Scientific Reports, 2016, 6: 38208.

[8] WU Z, LIN Y, WANG J, et al. Link prediction with node clustering coefficient[J]. Physica A Statistical Mechanics & Its Applications, 2016, 452: 1-8.

[9] LIU Y, ZHAO C, WANG X, et al. The degree-related clustering coefficient and its application to link prediction[J]. Physica A Statistical Mechanics & Its Applications, 2016, 454: 24-33.

[10] MURATA T, MORIYASU S. Link Prediction of Social Networks Based on Weighted Proximity Measures[C]. IEEE International Conference on Web Intelligence. IEEE, 2007.

[11] ZHU B, XIA Y. Link prediction in weighted networks: A weighted mutual information model[J]. PloS One, 2016, 11(2): e0148265.

[12] SETT N, SINGH S R, NANDI S. Influence of edge weight on node proximity based link prediction methods: An empirical analysis[J]. Neurocomputing, 2016, 172: 71-83.

[13] HUANG Z. Link prediction based on graph topology: The predictive value of generalized clustering coefficient[J/OL]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1634014.

[14] ZHU M, CAO T, JIANG X. Using clustering coefficient to construct weighted networks for supervised link prediction[J]. Social Network Analysis and Mining, 2014, 4(1): 215.

[15] WANG L, HU K, TANG Y. Robustness of Link-Prediction Algorithm Based on Similarity and Application to Biological Networks[J]. Current Bioinformatics, 2014, 9(5): 1-7.

[16] YAO Y B. Research on Link Prediction Method Based on Complex Network Topology[D]. Lanzhou: Lanzhou University, 2017.

[17] YANG L, SONG Y R, LI Y W. Network structure optimization algorithm for information propagation considering edge clustering and diffusion characteristics[J]. Journal of Physics, 2018, 67(19): 56-67.

[18] 吕琳媛, 周涛. 链路预测[M]. 北京: 高等教育出版社, 2013: 290.

[19] LV L Y, ZHOU T. Link prediction in weighted networks: The role of weak ties[J]. Epl, 2010, 89(1): 18001.

[20] MENG B, KE H, YI T. Link prediction based on a semi-local similarity index[J]. Chinese Physics B, 2011, 20(12): 128902.

[21] LIU Y, TANG M, DO Y, et al. Accurate ranking of influential spreaders in networks based on dynamically asymmetric link weights[J]. Physical Review E, 2017, 96(2): 022323.

[22] OPSAHL T. Why Anchorage is not(that) important: Binary ties and Sample selection [OL]. http://wp.me/poFcY-Vw.

[23] KUNEGIS J. Konect: the koblenz network collection[C]// Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 1343-1350.

[24] ADAMIC, LADA A, NATALIE G. The political blogosphere and the 2004 US election: divided they blog[C]// Proceedings of the 3rd International Workshop on Link Discovery. ACM, 2005.

[25] LUSSEAU D, SCHNEIDER K, BOISSEAU O J, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations[J]. Behavioral Ecology and Sociobiology, 2003, 54(4): 396-405.

[26] DU Z Y, CHEN H, SONG F. SNR Based Weighted-Consensus Algorithm for Cooperative Spectrum-Sensing [J]. Journal of Data Acquisition & Processing, 2013, 28(2): 184-189.

YUAN Rong, born in 1995, postgraduate. Her main research interests include complex network and link prediction.

SONG Yu-Rong, born in 1971, Ph. D, professor, is a member of China Computer Federation. Her main research interests include network information dissemination and its control.