Journal of Computer Applications
计算机应用,2016,36(12):3251-3255,3268
ISSN 1001-9081
CODEN JYIIDU
2016-12-10
http://www.joca.cn

文章编号:1001-9081(2016)12-3251-05
DOI:10.11772/j.issn.1001-9081.2016.12.3251

基于复杂网络节点重要性的链路预测算法

陈嘉颖[1],于　炯[1*],杨兴耀[1],卞　琛[2]
(1.新疆大学 软件学院,乌鲁木齐 830008;　2.新疆大学 信息科学与工程学院,乌鲁木齐 830046)
(*通信作者电子邮箱 yujiong@xju.edu.cn)

摘　要:提升链路预测精度是复杂网络研究的基础问题之一,现有的基于节点相似的链路预测指标没有充分利用网络节点的重要性,即节点在网络中的影响力。针对以上问题提出基于节点重要性的链路预测算法。该算法在基于局部相似性链路预测算法的共同邻居(CN)、Adamic-Adar(AA)、Resource Allocation(RA)相似性指标的基础上,充分利用了节点度中心性、接近中心性及介数中心性的信息,提出考虑节点重要性的CN、AA、RA链路预测相似性指标。在4个真实数据集上进行仿真实验,以AUC值作为链路预测精度评价指标,实验结果表明,改进的算法在4个数据集上的链路预测精度均高于共同邻居等对比算法,能够对复杂网络结构产生更精确的分析预测。

关键词:复杂网络;中心性;相似性;链路预测;共同邻居
中图分类号:TP393　　**文献标志码**:A

Link prediction algorithm based on node importance in complex networks

CHEN Jiaying[1], YU Jiong[1*], YANG Xingyao[1], BIAN Chen[2]
(1. *School of Software, Xinjiang University, Urumqi Xinjiang* 830008, *China*;
2. *School of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang* 830046, *China*)

Abstract: Enhancing the accuracy of link prediction is one of the fundamental problems in the research of complex networks. The existing node similarity-based prediction indexes do not make full use of the importance influences of the nodes in the network. In order to solve the above problem, a link prediction algorithm based on the node importance was proposed. The node degree centrality, closeness centrality and betweenness centrality were used on the basis of similarity indexes such as Common Neighbor (CN), Adamic-Adar (AA) and Resource Allocation (RA) of local similarity-based link prediction algorithm. The link prediction indexes of CN, AA and RA with considering the importance of nodes were proposed to calculate the node similarity. The simulation experiments were taken on four real-world networks and Area Under the receiver operation characteristic Curve (AUC) was adopted as the standard index of link prediction accuracy. The experimental results show that the link prediction accuracies of the proposed algorithm on four data sets are higher than those of the other comparison algorithms, like Common Neighbor (CN) and so on. The proposed algorithm outperforms traditional link prediction algorithm and produces more accurate prediction on the complex network.

Key words: complex network; centrality; similarity; link prediction; Common Neighbor (CN)

0　引言

复杂网络链路预测是根据已知、可观察到的节点的拓扑结构、节点属性等特征,预测网络中其他节点之间缺失的链接和未来可能产生的链接[1]。在社会网络分析、蛋白质交互作用、神经网络、电力网络等领域中,链路预测方法可广泛应用于分析网络数据的缺失、分析复杂网络演化机制等问题,在理论和实际应用中都发挥着巨大的作用,受到各领域的科学家的广泛关注[2]。

网络结构链路预测方法主要有基于相似性的链路预测、基于最大似然估计的链路预测和概率模型等方法[3]。基于相似性的方法是目前运用最多的链路预测算法之一,其前提是刻画节点相似性指标,由于基于相似性的方法计算简单、速度快、准确率高,吸引了很多研究学者的关注,但该方法在节点信息使用方面不够充分[4]。

在复杂网络中,一些具有重要作用的成员节点可能具有更大的影响力或者更强的信息传播能力,网络中节点的重要性可以用节点中心性来表示[5]。由于社交网络中大量活动都是围绕一些重要成员节点开展或与其具有密切的关系,因此,节点中心性在复杂网络研究中具有重要的理论价值和现实意义。杨建祥等[6]针对无权网络的介数中心性提出了快速更新算法;李静茹等[5]将度量节点中心性的方法应用于有权社交网络中,证明了加权网络中节点中心性的有效性及作用。

现有的基于局部相似性的链路预测方法,仅仅考虑了节点度这类局部信息,忽略了节点在网络中的重要程度,导致预测算法正确率降低。本文提出了基于复杂网络节点重要性的链路预测算法,在基于局部信息链路预测相似性指标研究的

收稿日期:2016-05-26;**修回日期**:2016-06-27。　**基金项目**:国家自然科学基金资助项目(61462079,61363083,61262088)。
作者简介:陈嘉颖(1988—),女,新疆沙湾人,硕士研究生,CCF会员,主要研究方向:推荐系统、社交网络、数据挖掘;　于炯(1964—),男,新疆乌鲁木齐人,教授,博士生导师,博士,主要研究方向:网络安全、网格与分布式计算;　杨兴耀(1984—),男,湖北襄阳人,博士,主要研究方向:推荐系统、网格计算、云计算、可信计算;　卞琛(1981—),男,江苏南京人,博士研究生,CCF会员,主要研究方向:内存计算、高性能计算、分布式系统。

基础上,考虑了节点在网络中的重要性。本文使用 Matlab 作为仿真工具,在真实数据集上进行仿真实验,实验结果表明改进的相似性指标在链路预测精度上有了一定的提升。

1 相关工作

1.1 链路预测

1.1.1 问题描述

定义 $G(V,E)$ 为无向网络,其中 V 为节点集合,E 为边集合。网络中总的节点数为 N,边数为 M,则该网络最多有 $N(N-1)/2$ 条链接,记为全局 U。链路预测问题可描述为:对于给定的一个网络拓扑,设计一种链路预测方法,对每个未连接的节点对 $(x,y) \in U-E$ 赋予分数值 S_{xy},然后按照该分数值从大到小排序,则排序越靠前的节点对之间就越有可能产生链接关系。

1.1.2 基于相似性的链路预测

应用相似性进行链路预测的重要前提是:假设两个节点之间相似性越大,它们之间存在链接的可能性就越大。如何定义节点之间的相似性成为基于节点相似性的链路预测方法的核心问题,根据不同的相似性度量方法,可将该类相似性指标分为基于局部信息的相似性指标、基于路径的相似性指标和基于随机游走的相似性指标。

基于局部信息相似性链路预测算法出现较早,应用较多,该方法认为两个端点的共同邻居越多则两个节点越相似,它们之间存在链接的可能性越大。共同邻居(Common Neighbor, CN)指标是基于局部信息相似性链路预测算法中衡量节点间相似性最简单的指标,此外,如果考虑共同邻居节点的度,还有优先连接指标[3,7]、AA(Adamic-Adar)指标及 RA(Resource Allocation)指标[8-9]。

基于路径相似性的链路预测算法从整体网络出发,考虑了所有长度路径的影响。如果两个节点之间最短路径的长度越短,只需要较少个数的节点就能相互访问,说明节点间关系相对密切。该方法需要考虑网络中所有长度的路径,复杂度高,计算开销大,不适合于实际网络应用。基于路径的相似性指标有局部路径指标、Katz 指标和 LHN-II(Leicht, Holme, Newman)[9]等。

基于随机游走的相似性指标根据随机游走模型定义,随机游走指的是任何无规律行走者所带的守恒量都各自对应着一个扩散运输定律[10],通常包括平均通勤时间(average commute time)、Cos + 指标、有重启的随机游走(random walk with restart)、SimRank 指标[3]等。

1.2 典型的局部信息相似性指标

基于局部信息相似性的链路预测算法中,最简单的相似性指标是共同邻居(CN)指标,AA 指标与 RA 指标也因考虑了节点度的信息而被广泛应用[11]。

1)CN 指标:共同邻居算法将共同邻居的数量作为衡量两节点间建立直接链路可能性的指标,即两节点间共同邻居越多,产生连接的可能性越大。对于节点 x、y,它们的邻居节点集合分别为 $\Gamma(x)$、$\Gamma(y)$,则节点 x、y 共同邻居集合为 $\Gamma(x) \cap \Gamma(y)$,CN 指标被定义为:

$$S_{xy} = |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

2)AA 指标:该算法的思想是度小的共同邻居节点的贡献大于度大的共同邻居节点。例如:在社交网络中,共同关注一个比较冷门话题的两个人之间相连的概率往往会比关注同一热门话题的两个人之间连接的概率高。因此,根据共同邻居节点的度为每个节点赋予一个权重值,AA 指标被定义为:

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\mathrm{lb}\, k_z} \tag{2}$$

其中:z 为节点 x,y 的共同邻居;k_z 表示节点 z 的度。

3)RA 指标:该算法从资源分配的角度出发,认为网络中每个节点都有一定的资源,并将资源平均分配给它的邻居。网络中没有直接相连的两个节点 x、y 之间,可从节点 x 传递一些资源到节点 y,在此过程中,节点 x 和 y 的共同邻居成为传递媒介。则节点 y 接收到的资源数就定义为节点 x、y 的相似度。RA 指标被定义为:

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \tag{3}$$

2 基于节点重要性的链路预测算法

2.1 节点中心性计算方法

在链路预测中,节点的重要性在一定程度上影响了预测的正确性。所谓节点的重要性指的是节点在网络中的地位,也可以看作节点在网络中的影响力[12]。目前,人们对节点影响力已有很多研究。评价节点影响力的方法有度中心性(Degree Centrality, DC)、接近中心性(Closeness Centrality, CC)和介数中心性(Betweenness Centrality, BC)[12-13]。

1)度中心性(DC)。

节点的度定义为目标节点与其他节点存在连接或关系的数量,即邻居个数。度中心性认为一个节点的邻居数目越多,影响力就越大。n 阶网络中的任一节点 v_i 的度中心性被定义为:

$$DC(i) = \frac{k_i}{n-1} = \sum_j a_{ij}/(n-1) \tag{4}$$

其中:a_{ij} 是网络 G 的邻接矩阵 A 中的 i 行和第 j 列的元素;k_i 代表节点 i 的度;n 代表网络中节点个数。节点度中心性指标是对节点最直接影响力的描述,拥有简单、直观、计算复杂度低的特点。度中心性越接近 1 表示节点越重要。

2)接近中心性(CC)。

接近中心性通过计算网络中的任意节点与其他节点的距离的平均值来取得。一个节点与网络中其他节点的平均距离越小,该节点的接近中心性就越大。对有 n 个节点的连通网络 G,G 中任意一个节点 v_i 到网络中其他节点的平均最短距离为 d_i:

$$d_i = \frac{1}{n-1} \sum_{j \neq i} d_{ij} \tag{5}$$

其中:d_{ij} 表示节点 i 到节点 j 的距离,d_i 的取值越小,意味着节点 v_i 更能接近网络中的其他节点。节点 v_i 接近中心性定义如下:

$$CC(i) = \frac{1}{d_i} = (n-1) \Big/ \sum_{j \neq i} d_{ij} \tag{6}$$

接近中心性指标采用了平均值来计算节点的重要性,故能在一定程度上消除对特殊值的干扰。

3)介数中心性(BC)。

节点的介数用来衡量一个节点位于其他节点最短路径上的次数,即网络中所有其他节点之间的最短路径中经过该节点的次数,代表该节点控制其他节点的能力[10]。也就是说,

如果其他节点的最短路径都有经过该节点,那么该节点具有较高的重要性,节点的介数中心性指标值就越高。对于 n 阶网络 G 中的任一节点 v_i,其介数中心性指标定义如下:

$$BC(i) = \sum_{i\neq s,i\neq t,s\neq t} \frac{g_{st}^{i}}{g_{st}} \tag{7}$$

其中:g_{st} 是从节点 v_s 到 v_t 的所有最短路径数目,g_{st}^{i} 是从节点 v_s 到 v_t 最短路径中经过节点 v_i 的最短路径数。由此可见,当一个节点不在任何一条最短路径上时,节点的介数中心性为 0。归一化的介数中心性定义为:

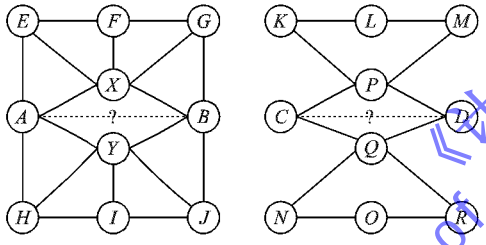
$$BC'(i) = \frac{2}{(n-1)(n-2)} \sum_{i\neq s,i\neq t,s\neq t} \frac{g_{st}^{i}}{g_{st}} \tag{8}$$

节点的介数中心性越高,说明该节点越重要,其控制信息流动的能力就越强,其他节点的依赖性就越大[14-15]。

2.2　考虑节点重要性的 CN 算法

CN、AA、RA 指标均为基于局部结构信息的算法,其中,CN 指标均一地为所有邻居节点分配权重,简单地将邻居节点的个数作为链路关联度的衡量标准[16],此算法虽然能够简洁地度量两点间存在链路的可能性,但没有充分考虑不同邻居节点在网络中重要性的差距。

图 1 为两个节点个数为 10 的社交网络:图 1(a)中,用户 A、B 的共同邻居为 X 和 Y;图 1(b)中,用户 C、D 的共同邻居为 P 和 Q。如果仅考虑节点共同邻居节点的个数,则用户 A、B 和用户 C、D 间存在链路的可能性相同。



(a) 10节点较复杂网络　　(b) 10节点简单网络

图 1　两种不同的节点链接方式

由图 1 可以看出:用户 A、B 在社交网络中较用户 C、D 具有更多的共同邻居节点和更强的连通性,即用户 A、B 较用户 C、D 重要程更高,因此可以根据共同邻居节点重要性的总和优化用户间存在链路可能性的值。考虑节点重要性的 CN 算法在共同邻居链路预测指标的基础上分别考虑了网络中节点的度中心性、接近中心性及介数中心性[17],改进的算法如下。

1)DCCN 算法:如果节点 X、Y 的共同邻居节点拥有较大的度中心性,那么节点 X、Y 之间建立链路的可能性更大。DCCN 算法相似性指标定义如下:

$$S_{xy}^{\mathrm{DCCN}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} DC(z) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{k_z}{n-1} \tag{9}$$

其中:x、y 表示任意两用户。

2)CCCN 算法:如果节点 X、Y 的共同邻居节点拥有较大的接近中心性,那么节点 X、Y 之间建立链路的可能性更大。CCCN 算法相似性指标定义如下:

$$S_{xy}^{\mathrm{CCCN}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} CC(z) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{n-1}{\sum_{j\neq z} d_{zj}} \tag{10}$$

3)BCCN 算法:如果节点 X、Y 的共同邻居节点拥有较大的介数中心性,那么节点 X、Y 之间建立链路的可能性更大。BCCN 算法相似性指标定义如下:

$$S_{xy}^{\mathrm{BCCN}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} BC'(z) = \frac{2}{(n-1)(n-2)} \sum_{z\in\Gamma(x)\cap\Gamma(y)} \sum_{z\neq s,z\neq t,s\neq t} \frac{g_{st}^{z}}{g_{st}} \tag{11}$$

2.3　考虑节点重要性的 AA 算法

AA 指标根据共同邻居节点的度值进行刻画,根据共同邻居节点的度,将节点度对数的倒数,即 $1/(\lg k)$ 作为权重值赋予每个节点。因此,AA 指标对于度值相同的节点仍看作是一样的。然而,重要性不同的节点具有不同的中介能力和信息转移能力,AA 指标没有考虑节点的重要性。考虑节点重要性的 AA 算法分别考虑了网络中节点的度中心性、接近中心性及介数中心性,改进的算法如下。

1)DCAA 算法:如果节点 X、Y 共同关注了度中心性较低的节点,则节点 X、Y 之间建立链路的可能性更大。DCAA 算法相似性指标定义如下:

$$S_{xy}^{\mathrm{DCAA}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\mathrm{lb}\, DC(z)} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\mathrm{lb}\,(k_z/(n-1))} \tag{12}$$

2)CCAA 算法:如果节点 X、Y 共同关注了接近中心性较低的节点,则节点 X、Y 之间建立链路的可能性更大。DCAA 算法相似性指标定义如下:

$$S_{xy}^{\mathrm{CCAA}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\mathrm{lb}\, CC(z)} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\mathrm{lb}\left((n-1)/\sum_{j\neq z} d_{zj}\right)} \tag{13}$$

3)BCAA 算法:如果节点 X、Y 共同关注了介数中心性较低的节点,那么节点 X、Y 之间建立链路的可能性更大。DCAA 算法相似性指标定义如下:

$$S_{xy}^{\mathrm{BCAA}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\mathrm{lb}\left(\frac{2}{(n-1)(n-2)} \sum_{z\neq s,z\neq t,s\neq t} \frac{g_{st}^{z}}{g_{st}}\right)} \tag{14}$$

2.4　考虑节点重要性的 RA 算法

RA 指标根据共同邻居节点的度值,从资源的角度出发,将共同邻居节点作为传递的媒介,使用共同邻居节点度的倒数为节点赋值。与 AA 指标相同,RA 指标对于度相同的节点也看作是一样的,没有充分利用网络节点的重要性。考虑节点重要性的 RA 算法分别考虑了网络中节点的度中心性、接近中心性及介数中心性,改进的算法如下。

1)DCRA 算法:如果节点共同关注了度中心性较低的节点,那么共同邻居节点将会因为分到节点 X、Y 更多的资源而具有更强的信息传递能力,则节点 X、Y 之间建立链路的可能性就越大。DCRA 算法相似性指标定义如下:

$$S_{xy}^{\mathrm{DCRA}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{DC(z)} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{n-1}{k_z} \tag{15}$$

2)CCRA 算法:如果节点共同关注了接近中心性较低的节点,那么共同邻居节点将会因为分到节点 X、Y 更多的资源而具有更强的信息传递能力,则节点 X、Y 之间建立链路的可能性就越大。CCRA 算法相似性指标定义如下:

$$S_{xy}^{\mathrm{CCRA}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{CC(z)} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \left(\sum_{j\neq z} d_{zj}\right)/(n-1) \tag{16}$$

3)BCRA 算法:如果节点共同关注了介数中心性较低的节点,那么共同邻居节点将会因为分到节点 X、Y 更多的资源而具有更强的信息传递能力,则节点 X、Y 之间建立链路的可能性就越大。CCRA 算法相似性指标定义如下:

$$S_{xy}^{\mathrm{BCRA}} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\frac{2}{(n-1)(n-2)}\sum_{z\neq s,z\neq t,s\neq t}\frac{g_{st}^{z}}{g_{st}}} \tag{17}$$

3 实验结果与分析

在链路预测评估阶段,将网络中已存在的链接集合 E 按照随机划分的方法划分为训练集 E^{T} 和测试集 E^{P} 两个集合,$E = E^{\mathrm{T}} \cup E^{\mathrm{P}}$,$E^{\mathrm{T}} \cap E^{\mathrm{P}} = \varnothing$。

3.1 评价指标

AUC (Area Under the receiver operation characteristic Curve)是衡量链接预测算法精度最常用的一种指标,该方法从整体上衡量算法的精确度,本文选择 AUC 指标作为验证实验的评价指标。AUC 可解释为从测试集 E^{P} 中随机取一条链接的预测概率值比随机地的从不存在的链接集合 E^{O} 选择的一条链接的概率值大的可能性[10],即每次随机从测试集 E^{P} 中选择一条链接(x,y) 与随机地从不存在的链接集合 E^{O} 中选择的链接(x',y') 的测试值进行比较,若 $S_{xy} > S_{x'y'}$,则加 1 分;若 $S_{xy} = S_{x'y'}$,则加 0.5 分;否则加 0 分。独立随机比较 n 次,记加 1 分的次数为 n',加 0.5 分的次数为 n'',因此 AUC 的计算公式定义为:

$$AUC = \frac{n' + 0.5n''}{n} \tag{18}$$

由式(18)可见,如果所有分数都是随机产生的,$AUC = 0.5$,因此 AUC 值大于 0.5 的程度衡量了算法在多大程度上比随机选择的方法精确。

3.2 实验数据集

为了评价算法的有效性,本文在四个典型的真实网络数据集上开展实验,各数据集特征如下:

1) USAir 网络(http://vlado.fmf.uni-lj.si/pub/networks/data/):该网络是美国航空路线无向加权网络,网络中的节点代表机场,节点间的链接代表机场之间的航线,网络中共包含 332 个节点,2 126 条链路关系。

2) C. elegans 网络(http://www.linkprediction.org/index.php/link/resource/data):该网络为蠕线虫神经无向无权网络,网络中的节点代表神经元,节点之间的链接代表神经元之间的链接关系,网络中包含 453 个节点,2 298 条链路关系。

3) Jazz 网络(http://www-personal.umich.edu/~mejn/netdata/):该网络为爵士音乐人合作网络,网络中的节点代表音乐人,节点之间的链接代表音乐人之间的合作关系,网络中包含 192 个节点,2 742 条链路关系。

4) Email 网络(http://konect.uni-koblenz.de/networks/):该网络为西班牙 Universitat Rovira i Virgili(URV)大学邮件通信网络,网络中的节点代表个人,节点之间的链接代表通过邮件的通信关系,网络中包含 1 133 个节点,5 451 条链路关系。

3.3 实验结果

本文以 AUC 值作为精度衡量指标,分别以基于局部相似性链路预测算法的 CN 指标、AA 指标和 RA 指标作为基准方法进行对比,将改进的链接预测方法应用于 USAir 网络、C. elegans网络、Email 网络及 Jazz 网络四 4 个真实数据集,采用 Matlab 作为仿真工具,进行 100 次独立实进行验证,其结果如图 2 所示,图中竖条从左到右依次为 CN、DCCN、CCCN、BCCN、RA、DCRA、CCRA、BCRA、AA、DCAA、CCAA、BCAA。

实验过程中将训练集占数据集的比例定义为 proportion,根据 proportion 随机地划分为训练集与测试集。第一次实验 proportion 的值是 0.9,其后每次实验减少 0.1,直到 0.5。由图 2 可以看出,在四个数据集中,随着设定的训练集比例 proportion 数值的减小,AUC 值也随之降低,因此,本文将训练集的比例设为 0.9。

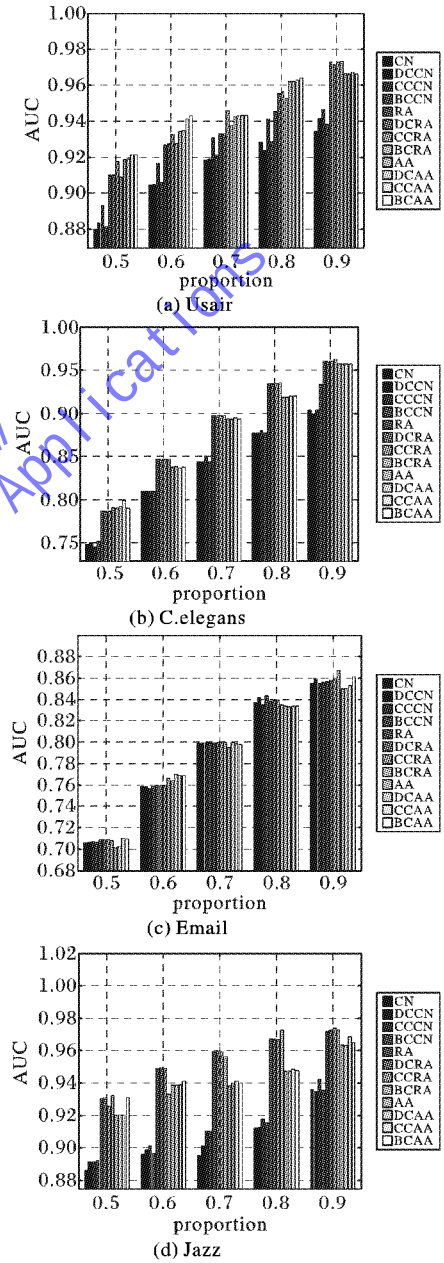


图 2 不同比例的训练集在算法中的 AUC 效果

将训练集比例为 0.9 时的几种相似性指标的预测精确度记录在表 1 中,表 2 记录了改进算法 AUC 指标相对于对比算法的改进程度。

从表 1 可以看出,在四个数据集上,考虑节点重要性的 CN 算法:DCCN、CCCN、BCCN 算法的预测精度总体优于 CN 算法的预测精度;从表 2 可以看出,DCCN、CCCN、BCCN 算法在四个数据集上相比 CN 算法,其预测精度分别提升了

0.178 1%,0.500 2%和 0.978 2%。考虑节点重要性的 AA 算法:DCAA、CCAA、BCAA 算法的预测精度总体优于 AA 算法的预测精度,以上 3 种算法在四个数据集上相比 AA 算法,其预测精度分别提升了0.014 7%,0.128 9%和0.378 9%。考虑节点重要性的 RA 算法:DCRA、CCRA、BCRA 算法的预测精度总体优于 RA 算法的预测精度,以上 3 种算法在四个数据集上相比于 RA 算法,其预测精确度分别提升了0.033 6%,0.107 6%和 0.382 1%。表 1 列出了对比算法与改进算法的预测精度的 AUC 值,可以看出,80.6% 的改进算法的预测精度高于对比算法预测精度,同时也出现了个别预测精度低于对比算法预测精度的现象,这与数据集的大小及其准确度有关。

在算法效率方面,设 N 为节点个数,CN 算法首先要查找网络中每一对被预测的节点,然后在两个节点上查找共同邻居节点,因此 CN 算法的时间复杂度为 $O(N^2)$。AA 和 RA 算法只是在共同邻居节点的基础上根据节点的度数进行一些计算[18],因此时间复杂度与 CN 算法相同。考虑节点度中心性的算法,在查找到共同邻居节点后计算每个节点的 DC 值,计算节点 DC 的时间复杂度为 $O(N)$,因此,在 CN、AA、RA 指标的基础上考虑节点度中心性后,改进算法时间复杂度不改变。同理,考虑节点接近中心性、介数中心性的算法,在共同邻居节点的基础上根据节点的 CC 值、BC 值进行计算,计算节点 CC 值、BC 值的计算复杂度分别为 $O(N^2)$、$O(N^3)$,因此,考虑节点接近中心性、介数中心性的算法的时间复杂度分别为 $O(N^2)$、$O(N^3)$,与共同邻居、AA、RA 算法相比,考虑节点介数中心性算法的时间复杂度有所提升,考虑节点度中心性、接近中心性算法的时间复杂度不变。上述实验结果表明,节点重要性在链路预测精确度方面起到了积极的作用,链路预测算法在 AUC 评价指标下较原始链路预测算法在预测精度上有了一定程度的提升。

表 1　不同算法在四个数据集上的预测精度 AUC 值比较

网络	CN	DCCN	CCCN	BCCN	AA	DCAA	CCAA	BCAA	RA	DCRA	CCRA	BCRA
USAir	0.934 3	0.941 5	0.946 5	0.938 2	0.966 2	0.966 3	0.967 1	0.966 3	0.972 7	0.971 4	0.972 7	0.973 3
C. elegans	0.903 6	0.899 6	0.903 7	0.933 8	0.956 9	0.956 9	0.957 1	0.957 0	0.960 4	0.960 6	0.960 4	0.962 6
Email	0.855 1	0.859 3	0.855 1	0.856 5	0.849 6	0.850 1	0.853 1	0.861 0	0.856 9	0.858 4	0.859 0	0.866 8
Jazz	0.935 8	0.934 8	0.942 2	0.935 7	0.963 3	0.963 2	0.963 2	0.964 8	0.971 8	0.972 5	0.973 6	0.972 6

表 2　各种算法 AUC 指标改进程度比较

网络	AUC 指标改进程度/%								
	DCCN	CCCN	BCCN	DCAA	CCAA	BCAA	DCRA	CCRA	BCRA
USAir	0.770 6	1.305 8	0.417 4	0.010 3	0.093 1	0.010 3	−0.133 6	0.000 0	0.061 7
C. elegans	−0.442 7	0.011 1	3.342 2	0.000 0	0.020 9	0.010 5	0.020 8	0.000 0	0.229 1
Email	0.491 2	0.000 0	0.163 7	0.058 9	0.412 0	1.341 8	0.175 0	0.245 1	1.155 3
Jazz	−0.106 9	0.683 9	−0.010 7	−0.010 4	−0.010 4	0.155 7	0.072 0	0.185 2	0.082 3
平均值/%	↑0.178 1	↑0.500 2	↑0.978 2	↑0.014 7	↑0.128 9	↑0.378 9	↑0.033 6	↑0.107 6	↑0.382 1

4　结语

将复杂网络节点中心性信息应用到了链路预测问题中,本文提出了考虑节点重要性的 CN、AA、RA 算法,并将改进的算法在真实数据集上与经典的链路预测指标进行比较,结果表明改进的算法能够提升链路预测精度。在算法效率方面,融合节点重要性后的部分算法的时间复杂度有所增加。在接下来的研究中,将对以上算法的效率进行改进,并对节点影响力进行进一步研究。

参考文献:

[1] 陈佳璐,钱宇华,张晓琴,等.依据节点贡献的链路预测方法[J].小型微型计算机系统,2016,37(1):92 −95.(CHEN J L, QIAN Y H, ZHANG X Q, et al. Link prediction method according to node contribution [J]. Journal of Chinese Computer Systems, 2016, 37(1): 92 −95.)

[2] 高曼,陈崚,徐永成.基于投影的二分网络连接预测[J].计算机科学,2016,43(2):118 −123.(GAO M, CHEN L, XU Y C. Projection based algorithm for link prediction in bipartite network [J]. Computer Science, 2016, 43(2): 118 −123.)

[3] 吕琳媛.复杂网络链路预测[J].电子科技大学学报,2010,39(5):651 −661.(LYU L Y. Link prediction in complex networks [J]. Journal of University of Electronic Science Technology of China, 2010, 39(5): 651 −661.)

[4] LYU L Y, ZHOU T. Link prediction in complex networks: a survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150 − 1170.

[5] 李静茹,喻莉,赵佳.加权社交网络节点中心性计算模型[J].电子科技大学学报,2014,43(3):322 −328.(LI J R, YU L, ZHAO J. A node centrality evaluation model for weighted social networks [J]. Journal of University of Electronic Science and Technology of China, 2014, 43(3): 322 −328.)

[6] 杨建祥,王朝坤,白易元,等.社交网络介数中心度快速更新算法[J].计算机研究与发展,2012,49(Suppl.):243 −249.(YANG J X, WANG C K, BAI Y Y, et al. A fast algorithm for updating betweenness centrality in social networks [J]. Journal of Computer Research and Development, 2012, 49(Suppl.): 243 −249.)

[7] LÜ L Y, JIN C H, ZHOU T. Similarity index based on local paths for link prediction of complex network [J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2009, 80(4 Pt 2): 046122.

[8] LEICHT E A, HOLME P, NEWMAN M E J. Vertex similarity in network [J]. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2006, 73(2 Pt 2): 026120.

[9] ZHOU T, LÜ L Y, ZHANG Y C. Predicting missing links via local information [J]. The European Physical Journal B, 2009, 71(4): 623 −630.

(下转第 3268 页)

HUANG W. Research on incentive mechanism of water-saving cities based on system dynamics [J]. Journal of Yangtze River Scientific Research Institute, 2010, 27(6): 10 – 13, 22.)

[11] XU J P, LI X F. Using system dynamics for simulation and optimization of one coal industry system under fuzzy environment [J]. Expert Systems with Applications, 2011, 38(9): 11552 – 11559.

[12] 周李磊,官冬杰,杨华,等.重庆经济-资源-环境发展的系统动力学分析及不同情景模拟[J].重庆师范大学学报(自然科学版),2015,32(3):59 – 67.(ZHOU L L, GUAN D J, YANU H, et al. System dynamics analysis and scenarios simulation of urban economy-resource-environment development [J]. Journal of Chongqing Normal University (Natural Science Edition), 2015, 32(3): 59 – 67.)

[13] 司训练,张锐,宋泽文.累积环境影响评价方法研究综述[J].西安石油大学学报(社会科学版),2014,23(4):11 – 16.(SI X L, ZHANG R, SONG Z W. Research review on the evaluation method of accumulative environment impacts [J]. Journal of Xi'an Shiyou University (Social Science Edition), 2014, 23(4): 11 – 16.)

[14] 李琰,杨勇,钟念,等.基于知识传播的集群聚集能力系统动力学研究[J].系统管理学报,2011,11(1):94 – 97,108.(LI Y, YANG Y, ZHONG N, et al. Modeling the knowledge transferring by system dynamics to analyze the district agglomeration capacity [J]. Journal of Systems & Management, 2011, 11(1): 94 – 97, 108.)

[15] 袁崇义.Petri 网原理与应用[M].北京:电子工业出版社,2005:10 – 30.(YUAN C Y. Principles and Applications of Petri Nets [M]. Beijing: Publishing House of Electronics Industry, 2005: 10 – 30.)

[16] 林闯.随机 Petri 网和系统性能评[M].2 版.北京:清华大学出版社,2005:126 – 135.(LIN C. Stochastic Petri-Nets and System Performance Evaluation [M]. 2nd ed. Beijing: Tsinghua University Press, 2005: 126 – 135.)

[17] 黄光球.网络攻击形式化建模理论[M].西安:陕西科学技术出版社,2010:10 – 15.(HUANG G Q. Formal Modeling Theory of Network Attacks [M]. Xi'an: Shaanxi Science and Technology Press, 2010: 10 – 15.)

[18] 胡运权,郭耀煌.运筹学教程[M].4 版.北京:清华大学出版社,2012:226 – 256.(HU Y Q, GUO Y F. Operational Research Tutorial [M]. 4th ed. Beijing: Tsinghua University Press, 2012: 226 – 256)

Background

This work is partially supported by the General Project of Humanity and Social Science Programming Foundation of Chinese Ministry of Education (15YJA910002), the Key Project-Basic Research Project of Natural Science of Shaanxi Province (2015JZ010), the Industrialization Project of Shaanxi Provincial Department of Education (16JF015), the Social Science Foundation of Shaanxi Province (2014P07), the Soft Science Research Project of Bureau of Science and Technology of Xi'an Municipality (SF1505(9)).

HUANG Guangqiu, born in 1964, Ph. D., professor. His research interests include Petri-net, system dynamics, swarm intelligent algorithm, computer simulation.

HE Tong, born in 1994, M. S. candidate. His research interests include Petri-net.

LU Qiuqin, born in 1966, Ph. D., professor. Her research interests include Petri-net, swarm intelligent algorithm, numerical simulation.

(上接第 3255 页)

[10] 刘海峰.社交网络用户交互模型及行为偏好预测研究[D].北京:北京邮电大学,2014:37 – 39.(LIU H F. Research of social network user interaction model and behavior preference prediction [D]. Beijing: Beijing University of Posts and Telecommunications, 2014: 37 – 39.)

[11] AHN M W, JUNG W S. Accuracy test for link prediction in terms of similarity index: the case of WS and BA model [J]. Physica A: Statistical Mechanics and its Applications, 2015, 429(1): 177 – 183.

[12] 张凯,马英红.基于网络结构的节点中心性排序优化算法[J].计算机应用研究,2016,33(9):2596 – 2600,2605.(ZHANG K, MA Y H. Centrality ranking algorithm based on network structure [J]. Application Research of Computers, 2016, 33(9): 2596 – 2600, 2605.)

[13] CHEN D B, LÜ L Y, SHANG M S, et al. Identifying influential nodes in complex networks [J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(4): 1777 – 1787.

[14] 任晓龙,吕林媛.网络重要节点排序方法综述[J].科学通报,2014,59(13):1175 – 1197.(REN X L, LYU L Y. Review of ranking nodes in complex networks [J]. Science China Bulletin, 2014, 59(13): 1175 – 1197.)

[15] 刘欣,李鹏,刘璟,等.社交网络节点中心性测度[J].计算机工程与应用,2014,50(5):116 – 120.(LIU X, LI P, LIU J, et al. Centrality for nodes in social networks [J]. Computer Engineering and Applications, 2014, 50(5): 116 – 120.)

[16] 郭婷婷,赵承业.基于共同邻居的链路预测新指标[J].中国计量学院学报,2016,27(1):121 – 124.(GUO T T, ZHAO C Y. A new measurement of link prediction based on common neighbors [J]. Journal of China University of Metrology, 2016, 27(1): 121 – 124.)

[17] LIU H F, HU Z, HADDADI H, et al. Hidden link prediction based on node centrality and weak ties [J]. Europhysics Letters, 2013, 101(1): 18004 – 18009.

[18] 张健沛,姜延良.一种基于节点相似性的链路预测算法[J].中国科技论文,2013,8(7):659 – 662.(ZHANG J P, JIANG Y L. A link prediction algorithm based on node similarity [J]. China Sciencepaper, 2013, 8(7): 659 – 662.)

Background

This work is partially supported by the National Natural Science Foundation of China (61462079, 61363083, 61262088).

CHEN Jiaying, born in 1988, M. S. candidate. Her research interests include recommender system, social network, data mining.

YU Jiong, born in 1964, Ph. D., professor. His research interests include network security, grid and distributed computing.

YANG Xingyao, born in 1984, Ph. D. His research interests include recommender system, grid computing, cloud computing, trusted computing.

BIAN Chen, born in 1981, Ph. D. candidate. His research interests include in-memory computing, high performance computing, distributed system.