第 38 卷 第 3 期
Vol.38 No.3

计 算 机 工 程
Computer Engineering

2012 年 2 月
February 2012

• 网络与通信 •

文章编号：1000—3428(2012)03—0067—04 文献标识码：A 中图分类号：N945

无标度网络中的链路预测问题研究

王 林，商 超

(西安理工大学自动化与信息工程学院，西安 710048)

摘 要：研究无标度网络中的链路预测问题。针对人造网络和实际社会网络，分别介绍静态和动态 2 种链路预测的实现过程，探究利用相似性进行链路预测的可行性，并验证多种相似度计算方法的准确性。对预测结果进行有效性分析，同时根据不同网络特性给出相应的预测算法。

关键词：复杂网络；信息检索；无标度；链路预测；拓扑结构；相似性

Research on Link Prediction Problem in Scale-free Network

WANG Lin, SHANG Chao

(School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

【Abstract】The link prediction problem in scale free networks is studied. Based on the man-made and real social network, the general processes of static and dynamic prediction are given respectively. The accuracy of several similarity methods is verified, and therefore, the method of using similarity to predict links is proved to be feasible. The pridiction algorithms are recommendied according to the effectiveness of pridiction results.

【Key words】complex network; information retrieval; scale-free; link prediction; topological structure; similarity

DOI: 10.3969/j.issn.1000-3428.2012.03.023

1 概述

复杂网络是对复杂系统的高度抽象。目前，对于网络的进化，结构和功能许多物理学家已经给予充分研究。然而有关网络分析的另一项重要问题——信息检索和恢复却始终没有引起足够重视，这在现代信息科学中是一项长期的挑战[1]，链路预测就属于此方向的一个新兴问题。网络中链路预测的目的在于通过网络的已知拓扑结构或者节点属性等信息来估计在 2 个尚未连接的节点之间产生一条连边的可能性。对链路预测问题的研究有着重要的理论和现实意义[2-3]。从理论层面讲，链路预测本质上是对网络演化规律的猜测，高精度的预测算法能够很好揭示网络的演化行为，有助于理解网络演化的内在机制。链路预测的重要意义还体现在应用方面。对于生物网络中隐而未知链接的揭示是需要耗费高额实验成本的，如果可以预测，而并非盲目地检测所有链接，并以此指导实验，就可节约实验开销。同时，对于演化的在线社会网络而言，可以去预测哪些现在尚未结交的用户之间应该是朋友关系，即预测那些未来可能出现但目前并不存在的链接。预测后可以将结果作为朋友推荐信息发送给目标用户，显然这可以帮助用户结交到新的朋友，避免对海量信息进行筛选，也有助于提高相关网站在用户心目中的地位，从而提高用户对该网站的忠诚度。

链路预测在计算机领域主要是提出一些基于马尔科夫链和机器学习过程的算法，但这些方法在物理上不简洁。从复杂网络的角度来研究链路预测是一种全新的方式，这种方法简单可靠，且具有普适性。基于拓扑结构相似性的方法是现在链路预测主流的方法，这种方法有一个重要的假设前提，就是如果 2 个节点之间的相似性越大，那么在它们之间存在链接的可能性也就越大，因此该方法的一个核心问题就归结为该如何来定义节点之间的相似性[4]。本文对无标度网络中的链路预测问题进行研究。

2 相似性链路预测算法

本节将介绍一些典型的相似性预测算法。

(1)CN 指标

CN(Common Neighbors)[1-2]共同邻居指标的基本思想是：如果 2 个节点拥有更多的共同邻居，那么在它们之间更倾向于存在一条连边。这里让 Γ(*)表示节点*的邻居集合，那么这种基于邻居重叠的算法就可以表示为：

$$S_{xy}=\left|\Gamma(x)\cap\Gamma(y)\right|$$

其中，符号|表示取维数，即计算集合中元素的个数；S_{xy} 是节点 x 和节点 y 之间的相似性分数。

(2)Salton 指标

Salton 指标[2]考虑了两端节点度的影响性，定义为：

$$S_{xy}=\frac{\left|\Gamma(x)\cap\Gamma(y)\right|}{\sqrt{k(x)\times k(y)}}$$

其中，$k(x)=|\Gamma(x)|$是节点 x 的度值，即与节点 x 相连的边或者节点的总个数。

(3)accard 指标[2]

$$S_{xy}=\frac{\left|\Gamma(x)\cap\Gamma(y)\right|}{\left|\Gamma(x)\cup\Gamma(y)\right|}$$

它的基本思想是：一个随机选择的节点 x 或 y 的邻居是节点 x 和 y 的共同邻居的可能性。

(4)Sørensen 指标[2]

$$S_{xy}=\frac{2\times\left|\Gamma(x)\cap\Gamma(y)\right|}{k(x)+k(y)}$$

作者简介：王 林(1963－)，男，教授、博士，主研方向：复杂网络，网络通信；商 超，硕士研究生

收稿日期：2011-08-15 **E-mail：**wanglin@xaut.edu.cn

(5)Hub Promoted 指标[2]

$$S_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{\min\{k(x),k(y)\}}$$

在这一指标中，分母是由低度值来决定的。

(6)Hub Depressed 指标

Hub Promoted[2]分母由高度值来决定，意在削弱高度值节点对相似性的影响。

$$S_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{\max\{k(x),k(y)\}}$$

(7)LHN-Ⅰ指标

LHN-Ⅰ[5]使得拥有更多共同邻居的节点对具有更高的相似性，且是相对于期望的邻居数目而言。

$$S_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{k(x)\times k(y)}$$

(8)PA 指标

PA(Preferential Attachment)[2]优先连接指标是一个只考虑节点度的相似性指标。在无标度网络中，一个新生成的连接与节点 x 相连的概率是与 x 的度 $k(x)$呈正比的，受到这种机制的启发，一个相对应的相似性指标就可以表示为:

$$S_{xy}=k(x)\times k(y)$$

(9)Katz 指标

Katz 指标[1]考虑的是所有路径的集合。

$$S_{xy}=\sum_{l=1}^{\infty}\beta^{l}.\left|paths_{xy}^{<l>}\right|$$

其中，$paths_{xy}^{<l>}$ 表示的是连接节点 x 和 y 的长度为 l 的所有路径的集合；β 是用来控制路径权重的参数。

(10)LHN-Ⅱ指标

LHN-Ⅱ指标[5]的理论基础是：如果节点 x 和节点 y 中任意一个的邻居节点 v 与另一个节点相似，那么节点 x 与节点 y 相似。其矩阵定义式为:

$$S=2m\lambda D^{-1}\left(I-\frac{\alpha}{\lambda}\boldsymbol{A}\right)^{-1}\boldsymbol{D}^{-1}$$

其中，λ 是邻接矩阵 $\boldsymbol{A}$ 的最大特征值；m 是网络中边的总数目；$\boldsymbol{D}$ 是一个对角矩阵，对角元素为各个节点的度值；α 是自由参数，取值范围必须满足 $0<\alpha<1$，可用来控制长短路径对于相似性贡献的程度。

(11)LP 指标

LP(Local Path)[1]局部路径指标是在 CN 的基础上不仅考虑了直接邻居的影响，同时也注意到了次级邻居(邻居的邻居)的贡献。

$$S=A^2+\varepsilon A^3$$

其中，ε 为可调节参数，用来控制三阶路径的影响。

3 静态无标度网络中的链路预测

将度分布符合幂律分布的网络称为无标度(scale free)网络，它的节点的连接度没有明显的特征长度。

3.1 无标度网络的建立与实现

考虑到实际网络具有的增长和优先连接两大特性，学者 Barabási 和 Albert 提出了著名的 BA 无标度网络演化模型。

基于 BA 无标度模型构造算法，编程实现无标度网络。本文选择初始节点 m_0=20，每次引入一个新节点时，让它分别与 m=2,4,6,8,10 个已经存在的节点相连，且最终生成一个具有 300 个节点的网络。设置好这些初始参数后，经过运行程序，最终生成 5 个具有无标度性质的网络，如表 1 所示。

表 1 网络相关情况

网络	总边数	训练集边数	测试集边数
网络 1(m=2)	750	675	75
网络 2(m=4)	1 310	1 179	131
网络 3(m=6)	1 870	1 683	187
网络 4(m=8)	2 430	2 187	243
网络 5(m=10)	2 990	2 691	299

3.2 静态链路预测过程的实现

依托生成的静态无标度网络，对链路预测问题进行实证研究，以下是整个链路预测过程的具体实施步骤。

Step1 将每个网络中所有生成的边按照 90%和 10%的比例随机地分成训练集和测试集 2 个部分。

Step2 设网络在训练集中节点的集合为 V，连边集合为 E_{tra}，这里 V 中都包含 300 个节点。则当前尚未连边节点对的集合为$(V\times V)$-E_{tra}，并找出此集合中所有未连边的节点对。

Step3 根据指定的相似性算法，会定量的得到所有节点对<x,y>的相似性分数，这会以一个 300×300 的相似性矩阵的形式表现出来。这里要关心的是所有未连边节点对的相似性分数大小，因此需要找出所有未连边节点对所对应的相似性分数值。

Step4 按照此分数值从大到小(降序)的顺序将对应的节点对及其相似性分数排列在表 L 中，表中从上到下出现连边的机率会依次减少。

Step5 设网络在测试集中连边的集合为 E_{pro}，$n=|E_{pro}|$ 为网络中实际新增连边的数目。

Step6 选取表 L 中的前 n 对节点建立连边，设其为预测的连边，表示为 E_{pre}，也就是最有可能存在连边的集合。

Step7 根据正确率 precision 的定义，为预测网络新增连边 E_{pre} 与实际网络新增连边 E_{pro} 中重合的边的数目占总的新增连边的比例。将表 L 中的前 n 条边与测试集中的边进行比较，找出两者中相同的边，即预测正确的连边。并计算出正确率 p。

$$p=\frac{|E_{pre}\cap E_{pro}|}{n}\times 100\%$$

3.3 预测结果及评价分析

基于以上过程，运用多种相似性算法来进行连边预测，表 2 给出了它们分别对应的正确率 p 以及随机预测的效果。表中黑体字标出了预测效果最好的一些值。

表 2 预测正确率比较

预测算法	网络 1	网络 2	网络 3	网络 4	网络 5
随机预测	0.17	0.3	0.43	0.57	0.71
CN	**26.67**	14.5	11.23	11.11	9.7
Salton	0	2.29	5.35	8.23	8.03
Jaccard	0	2.29	5.35	8.23	7.36
sorensen	0	2.29	5.35	8.23	7.36
Hub Promoted	5.33	0.76	0.16	2.06	1.67
Hub Depressed	9.33	3.05	5.35	8.23	1.67
LHN-I	0	0	0	0.82	0.33
PA	**26.67**	**16.79**	**12.3**	**11.93**	**11.71**
LHN-II	0	0	0	0	0.33
Katz	25.33	15.27	11.23	**11.93**	11.37
LP	25.33	15.27	11.23	11.52	11.37

为了对预测效果进行直观比较，图 1 用直方图的形式表述预测结果的平均正确率。可以看出，基本上所有采用了预测算法的正确率都大于随机预测的，说明利用相似性来进行链路预测的方法是可行的。

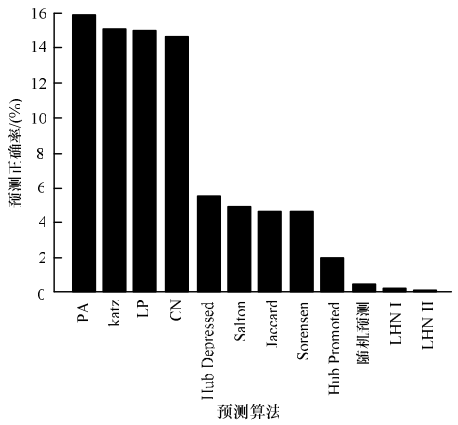


图 1 多种算法的预测正确率

PA 算法的预测效果最好，平均准确率高达 15.88%，是随机预测的 36.4 倍，其实本文的无标度网络正是基于优先连接原理而构造的，而 PA 算法也正是因为符合了网络的这一特征，所以能够取得如此好的效果。

Katz 算法的预测平均准确率为 15.026%，是随机预测的 34.5 倍，效果也很好。Katz 是一种基于全局信息的算法，它考虑到了网络中所有长度的路径，但是这种算法存在的最大缺陷是计算的复杂度比较高，需要较大的存储空间和较长的耗时。

LP 算法的预测平均准确率为 14.944%，是随机预测的 34.3 倍，效果也比较好。LP 指标不仅考虑到了二阶路径的，同时也考虑到了三阶路径的影响力，因此上它应该算是一种半局部指标。LP 不仅有着高的预测准确率，同时因为它不是全局算法，所以复杂度也不高，是一种值得推荐的链路预测算法。

最简单的基于共同邻居数目的 CN 算法的预测准确率为 14.642%，是随机预测的 33.6 倍，效果也很不错。这种朴素的共同邻居思想应用在现实生活中就是如果 2 个人拥有的共同朋友越多，那么这 2 个人之间也是朋友的几率就越大。CN 算法虽然比较简单，但是经证实这种预测的思想在几乎所有的网络中表现都非常不错，特别是在聚类性比较高的网络中，是值得推广的。

相比较而言，LHN-Ⅰ和 LHN-Ⅱ算法的效果就很差，甚至不如随机预测的好。对于 LHN-Ⅰ算法，因为它计算所得的相似性分数值一般会很小，这势必会导致相似性之间的差异不明显，从而使得算法的预测效果不好。LHN-Ⅱ算法是基于一种简单的传递思想，经证实这种相似性的方法在无标度网络中预测效果很差，因为没有能够反映出此类网络的特性，也许在其他类型的网络中会有不错的效果。

4 动态 BBS 兴趣网络上的链路预测

4.1 BBS 简介

BBS 即论坛，这是一种交互性强，内容非常丰富，同时具有及时性的因特网电子服务系统。随着互联网的大力普及，BBS 更是以惊人的速度发展起来，并且一直占据着互联网的主导地位，对它的研究有着重要的意义。同时由于其信息复杂性与海量性的特点，也非常适合采用复杂网络的方法进来行描述和研究。

本文分析的 BBS 数据来自于人民网强国论坛，它是目前国内最具有影响力的论坛之一。

4.2 BBS 兴趣网络定义

通常情况下认为，如果 2 个用户都对同一个帖子进行了回复，那么这 2 个用户有着共同的兴趣，因此可以在这 2 个用户之间连上一条边，同时也分别在这 2 个回帖用户和发帖用户之间各连上一条边，这样就会得到能够反映出用户共同兴趣的网络，将其称为 BBS 用户兴趣网络[6]。

4.3 数据前期预处理及说明

基于 BBS 兴趣网络的定义，在 SQL Server 2000 数据库中结合 Matlab 程序，提取数据库中帖子的主键信息进行以下处理：建立帖子之间的相互关系；找出帖子和发帖人之间的关系；建立会员和会员之间的关系。经过这样的数据前期预处理后就可以提取出一个 BBS 用户兴趣网络模型。

已经证实 BBS 网络具有自相似性[7]，且由于网络的帖子数量庞大，全部拿来分析不现实，在此选取论坛 2003 年 3 月 1 日−10 日的数据进行分析。同时在此动态网络中选取 1 日−5 日的数据作为训练集来使用，6 日−10 日的数据作为测试集来使用，将这 2 组数据分别进行前期预处理，表 3 列出了这 2 组数据的一些基本信息。

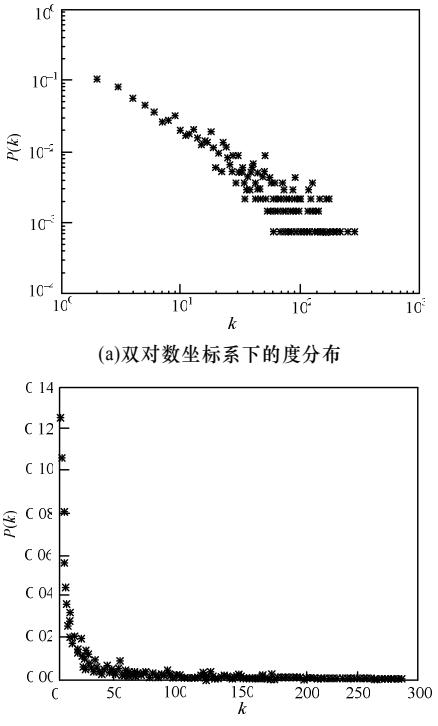
表 3 BBS 网络基本信息

采样时间	网络总节点数	网络总边数	连通子图规模	连通子图内的连边数目
2003.03.01(00:00:00)−2003.03.05(23:59:59)	1 374	16 476	1 363	16 469
2003.03.06(00:00:00)−2003.03.10(23:59:59)	1 311	16 672		

由于部分链路预测算法只有在网络连通时才有意义，因此仅考虑网络中的最大连通子图。

4.4 BBS 网络特性分析

对收集到的 BBS 网络的度分布进行分析，如图 2 所示。



(a)双对数坐标系下的度分布

(b)正常坐标系下的度分布

图 2 双对数坐标系及正常坐标系下的度分布

通过正常坐标和双对数坐标可以发现此网络的度分布近似为幂律分布，在正常坐标系中拖着长长地重尾，在双对数坐标系中可以近似为一条直线，所以 BBS 兴趣网络是一个不太严格的无标度网络，它具有无标度网络的一些性质，同时其自身又带有一些特别的性质。

BBS 兴趣网络的聚类系数 C 为 0.485 194，网络的平均最

短路径长度 L 为 2.762 957，网络直径 D 为 6，平均度<k> 24.165 8，度的异质性指标 H 是 3.509 4。从这些参数中可以反映出，BBS 兴趣网络是一个高度聚集的小世界网络，它的聚类系数非常高，且最短路径很小。

4.5 动态链路预测的实现

由于此处的 BBS 兴趣网络是一个动态的演化网络，因此针对该网络的链路预测的具体实现过程与上一部分静态无标度网络上的过程存在差别。

Step1 将预处理后的 1 日到 5 日数据作为训练集来使用，这是一个包含有 1 363 个节点和 16 469 条边的网络。同时将 6 日−10 日的数据作为测试集来使用，它是一个含有 1 311 个节点和 16 672 条边的网络。但是此处必须注意，因为网络是动态的，随着时间的推移，会有新的节点和连边添加进来，即在测试集中会出现一些并不存在于训练集中的节点，同时在这些新的节点之间或者在这些新节点和旧节点(既存在于训练集也存在于测试集的节点)之间也可能会形成一些连边，而此处所关心的仅是通过预测后旧节点之间可能会出现哪些新的连边，因此对于上面说的这种情况在后面必须进行一些相应的处理。

Step2 设网络从开始到 t 时刻形成的快照中节点的集合为 G，连边集合为 E_{old}，可以找出网络中未连边的节点对<x,y>的集合为$(G\times G)-E_{\text{old}}$。经计算此集合中有 911 734 对节点。

Step3 根据指定的链路预测算法，定量的得出上面集合中所有未连边节点对<x ,y>之间的相似性分数 s_{xy}。

Step4 按照这一分数值由大到小的顺序将其排列在表 L 中。

Step5 选取 t 时刻至其后的某一时刻 t' 形成新的网络，设新的连边集合为 E_{new}。在此为了使 Step1 中所描述的新加入节点所成的连边的麻烦情况得以解决，需要定义新的变量 E^*_{new}，并使 $E^*_{\text{new}}=E_{\text{new}}\cap(G\times G)$ 为对应初始网络节点集合 G 实际新加入的连边，且令 $n=\left|E^*_{\text{new}}\right|$ 为新增连边的数目，此处其值为 10 850，而并非 16 672。

Step6 选取表 L 中的前 n 对节点建立连边，设其为预测的网络新增连边集合 E_{pre}。

Step7 将 E_{pre} 中的边与 E^*_{new} 中的边进行比较，找出两者中相同的边，并计算出正确率 p：

$$p=\frac{\left|E^*_{\text{new}}\cap E_{\text{pre}}\right|}{n}\times 100\%$$

4.6 预测结果及其分析

依照以上动态预测过程，给出 BBS 兴趣网络中多种预测方式的预测结果及其优于随机预测效果的比例，如表 4 所示，分析结果如下：

CN 算法预测效果很不错，正确率达到 18.76%，其实在无标度网络中它的表现就很好，可以说这是一个虽然简单，但是很优秀的算法，这种基于共同邻居的相似性思想在链路预测领域是值得大力推广的。

LP 算法因为考虑的更加全面，使它的预测正确率最高，达到了 18.86%。

Katz 算法正确率 18.35%，这种基于全局信息的指标预测效果也很不错。

PA 算法正确率为 18.32%，其实通过前面对 BBS 兴趣网络拓扑特性的分析已经得出了，这是一个近似的无标度网络，那么它就具有优先连接的特性，而 PA 算法正是基于这一原理的，所以正确率高是可以理解的。同时从实际情况来看，人们平时在浏览论坛的时候，总是会情不自禁的优先选择那些点击率比较高的帖子来关住，优先针对热门话题来发表评论和看法。

小度节点有利 Hub Depressed 算法的效果也很不错，预测正确率为 10.49%，说明原来小度值的节点在链路预测中的作用变大了，它们的力量不可小视。而大度节点有利 Hub Promoted 算法因为原来大度值的节点在预测中的作用被削弱了，因此效果反而不佳。

LHN-Ⅰ和 LHN-Ⅱ算法的预测效果很差，这与在 BA 无标度网络中得到的结果相同，由此认为，这 2 种方法并不适合于具有无标度性质的网络，它们不能充分挖掘出此类网络中的隐含连边信息，因此在进行预测时应尽量避免使用这些算法。

表 4 预测正确率比较

预测算法	预测正确率/(%)	预测正确边的条数	优于随机预测效果的比例
随机预测	1.189		
CN	**18.760**	2 036	15.778 0
Salton	9.100	987	7.653 0
Jaccard	10.730	1 164	9.024 0
Sorensen	10.730	1 164	9.024 0
Hub Promoted	0.980	106	0.824 2
Hub Depressed	10.490	1 138	8.822 5
LHN-I	0.065	7	0.054 7
PA	**18.320**	1 988	15.408 0
LHN-II	0.074	8	0.062 2
Katz	**18.350**	1 991	15.433 0
LP	**18.860**	2 046	15.862 0

5 结束语

本文分别通过在构造的静态无标度网络和收集的动态 BBS 兴趣网络上实现链路预测，给出预测结果并提出利用网络拓扑结构信息的相似性预测算法，且都能获得优于随机预测的效果，这说明利用相似性进行链路预测是有效的。

针对具有无标度特性的网络，在链路预测时推荐使用 CN，PA，LP 和 Katz 4 种方式，同时应尽量避免使用 LHN-Ⅰ和 LHN-Ⅱ 2 种方式。基于结构的相似性预测方法优点是比较简单，但是它主要是对网络某一方面的结构特点进行刻画，这将会导致各个方法在不同类型网络中的预测能力各不相同，需要根据网络特征来匹配适合的预测算法，因此希望在以后的进一步研究中能够找到一些性能相对较好，且能够适应大部分网络结构的通用预测算法。下一步工作将对含权有向网络中的链路预测问题进行研究。

参考文献

[1] Lv Linyuan, Jin Cihang, Zhou Tao. Similarity Index Based on Local Paths for Link Prediction of Complex Networks[J]. Physical Review E, 2009, 80(4).

[2] Zhou Tao, Lv Linyuan, Zhang Yicheng. Predicting Missing Links via Local Information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[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.

[4] 吕琳媛. 链路预测的研究现状及展望[EB/OL]. (2010-04-30). http://blog.sciencenet.cn/?329471.

[5] Leicht E A, Holme P. Vertex Similarity in Networks[EB/OL]. (2005-10-14). http://arxiv.org/abs/physics/0510143.

[6] 王 林, 戴冠中. 基于复杂网络社区结构的论坛热点主题发现[J]. 计算机工程, 2008, 34(11): 214-216.

[7] 王 林, 戴冠中. 复杂网络的 Scale-free 性、Scale-free 现象及其控制[M]. 北京: 科学出版社, 2009.

编辑 陈 文