基于节点相似性的链接预测

东昱晓 柯 庆 吴 斌
(北京邮电大学计算机学院 北京 100876)

摘 要 链接预测是图数据挖掘中的一个重要问题。它是通过已知的网络结构等信息预测和估计尚未链接的两个节点存在链接的可能性。目前大部分基于节点相似性的链接预测算法只考虑共同邻居节点的个体特征,针对目前预测算法对共同邻居节点间相互关系的考虑不足,提出了一种新算法:节点引力指数算法。该算法在保持低时间复杂度的同时,提高了预测的准确率。通过多个现实网络实验证实了算法的预测效果。

关键词 复杂网络,数据挖掘,链接预测,节点相似度,节点引力指数

中图法分类号 TP391 **文献标识码** A

Link Prediction Based on Node Similarity

DONG Yu-xiao KE Qing WU Bin
(School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract Link prediction is an important issue in graph mining. It aimed at estimating the likelihood of the existence of links between nodes by the known network structure information. Currently, most link prediction algorithms based on node similarity consider only the individual characteristics of common neighbor nodes. We designed a new algorithm exploiting the interactions between common neighbors, namely Individual Attraction Index. While maintaining low time complexity, this algorithm remarkably improved the accuracy of prediction. This paper proved well the best overall performance of this new algorithm by comparing three well-known node similarity algorithms on eight real networks with Individual Attraction Index.

Keywords Complex network, Data mining, Link prediction, Node similarity, Individual attraction index

1 引言

链接预测(Link Prediction)问题作为数据挖掘的方向之一,在计算机领域已有长时间的研究,它是指如何通过已知的网络结构等信息评估预测网络中尚未链接的两个节点之间存在或产生链接的可能性[1]。链接预测包括:(1)预测已存在但尚未被发现的链接,即预测未知链接;(2)预测现在未存在但未来可能新产生的链接,即预测未来链接[2]。本文主要研究对未知链接的预测问题,研究方法为将完整网络随机切分成训练集和测试集,根据训练集进行链接预测,根据测试集得到预测结果。

早期的研究思路和方法为基于马尔可夫链进行网络的链接预测和路径分析[3],之后 Zhu[4] 等人在自适应网站(adaptive web sites)的预测中使用了基于马尔可夫链的预测方法。后来,Popescul 和 Ungar[5] 提出一个回归模型,并在文献引用网络中预测科学文献的引用关系。2008 年,Clauset, Moore 和 Newman[6] 发表在《自然》上的论文提出了一种利用网络层次结构进行链接预测的方法,该方法在具有明显层次结构的网络中表现最好。另外,Lichtenwalter[7] 等人提出了一种监督学习的链接预测框架,该框架使用分类算法进行预测。最近几年,基于节点相似性的预测方法已经成为研究热点,在该方法中,两个节点之间相似性指数越大,就认为它们之间存在或产生链接的可能性越大。Liben-Nowell 和 Kleinberg[8] 将相似性指标分为基于节点和基于路径的,并分析了若干指标对社会合作网络中链接预测的效果。

传统的基于节点相似性的算法只考虑被预测的两个节点共同邻居的个数和共同邻居的度数,并未深入研究共同邻居之间的相互关系对预测结果的影响。本文结合已有的基于节点相似性算法,提出了一种新的基于节点相似性的预测算法——节点引力指数算法(IA),节点引力指数是在考虑被预测两个节点共同邻居的度数的基础上,同时考虑被预测节点与共同邻居组成的小型网络对预测结果的影响。新算法与其他主流的基于节点相似性算法进行比较,取得了较好的实验结果。

本文第 2 节对目前已有的基于节点相似性算法进行了简要介绍;第 3 节对链接预测问题与算法评价方法进行了简要

到稿日期:2010-08-14 返修日期:2010-11-25 本文受国家自然科学基金(60905025,61074128),国家高技术研究发展计划(2009AA04Z136)资助。

东昱晓(1987—),男,主要研究方向为复杂网络、数据挖掘、分布式计算,E-mail:ericdongyx@gmail.com;柯 庆(1987—),男,主要研究方向为分布式计算、复杂网络、数据挖掘;吴 斌(1969—),男,副教授,博士生导师,主要研究方向为智能信息处理、复杂网络、基于图的数据挖掘、商务智能。

描述；第 4 节提出了新的链接预测算法；第 5 节采用几个实际数据集比较了各种算法的实验结果；最后是对全文的简单总结。

2 相关工作

通过对众多基于节点相似性算法的研究，Liben-Nowell 和 Kleinberg 发现 Common Neighbors(CN)和 Adamic-Adar (AA)[9]算法比另外 7 种已知算法有更好的实验结果[8]。另外周涛等人[10]提出一种新的基于节点相似性的 Resource Allocation(RA)算法，该算法比 CN 和 AA 算法有更高的预测准确率。因此，本文简要介绍 CN、AA 和 RA 3 种著名的算法。

Common Neighbors(CN)：对于网络中的节点 m，定义 m 的邻居集合为 $\varphi(m)$，那么两个节点 x 和 y 的节点相似性 S_{xy} 就定义为它们的共同邻居数量，即 $S_{xy}=|\varphi(x)\bigcap\varphi(y)|$，CN 为最简单的相似性指标，它的主要思想是如果两个节点有较多的共同邻居，它们更倾向于存在或产生链接。

Adamic-Adar Index(AA)：对于网络中节点 m，定义 m 的度为 $k(z)=|\varphi(x)|$，该算法是在共同邻居算法的基础上赋予其权重，节点 x 和 y 的节点相似性定义为 $S_{xy}=\sum_{z\in\varphi(x)\cap\varphi(y)}\frac{1}{\log k(z)}$，其思想是共同邻居中度小的节点对于相似性分数 S_{xy} 的贡献权重大于度大的节点。该权重为 $1/\log k$[9]。

Resource Allocation Index(RA)：该算法是考虑网络中没有直接相连的两个节点 x 和 y，可以从 x 中传递一些资源到 y，在传递过程中，它们的共同邻居就成为传递资源的媒介。节点相似性定义为 $S_{xy}=\sum_{z\in\varphi(x)\cap\varphi(y)}\frac{1}{k(z)}$[10]。在算法中，每个媒介都有一个单位的资源并且将其平均分配传给它的邻居，y 可以接收到的资源数就是节点 x 和 y 的相似度[2]。

我们看到，CN，AA 和 RA 3 种算法都只考虑共同邻居个体对节点相似性的影响，并未考虑共同邻居节点之间的相互关系的影响。

3 问题描述与评价方法

定义无向网络 $G(V,E)$，V 为节点集合，E 为边集合，即链接集合。网络 G 中节点数为 N，边数为 M，网络中不存在自环结构。对于每一对节点 $x,y\in V$，$E_{xy}\notin E$，链接预测算法计算其存在链接的概率，概率值表示节点 x 和 y 之间存在链接的可能性大小。基于节点相似性的预测算法使用节点相似性表示其存在链接的概率，每对节点存在链接的概率从大到小排序，排在前面的节点对出现链接的可能性大。

为了测试链接预测算法的准备性，将已知链接 E 分为训练集 E_t 和测试集 E_p 两部分，明显地，$E=E_t+E_p$。本文中，公用数据集的训练集 E_t 包含 90%的边，剩下 10%的边存在于测试集中。同时，我们使用 AUC(area under the receiver operating characteristic curve)指标[11]和 Precision 指标[12]来衡量预测算法的准确度。

AUC 可以理解为在测试集中边的分数值比随机选择的一个不存在的边的分数值高的概率，也就是说，每次随机从测试集中选取一条边与随机选择的不存在的边进行比较，如果测试集中的边的分数值大于不存在的边的分数值，就加一分；如果两个分数值相等，就加 0.5 分。独立地比较 n 次，如果 n' 次测试集中的边的分数值大于不存在的边的分数，n'' 次两个分数值相等，则 AUC 定义为：

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

显然，如果所有分数都是随机产生的，AUC=0.5。因此 AUC 大于 0.5 的程度衡量了算法比随机选择准确的程度[2]。

Precision 指标定义为在前 L 个预测链接中被预测准确的比例。如果有 m 个预测准确，即排在前 L 的链接中有 m 个在测试集中，则 Precision 定义为：

$$\mathrm{Precision}=m/L$$

Precision 越大说明预测结果越准确。

4 节点引力算法

本节提出一种新的基于节点相似性的链接预测算法——节点引力算法(Individual Attraction Index)。该算法在考虑节点共同邻居的个体特征时，还考虑了共同邻居间相互关系对节点相似性的影响。

在基于节点相似性的预测算法中，都是从被预测的两个节点的共同邻居出发，但是传统的算法都只是考虑单个共同邻居节点的特征，如图 1 中(a)(b)所示，求节点 X 和节点 Y 之间存在链接的可能性，CN 算法是只基于共同邻居的个数，两图中节点 X 和 Y 之间具有相同的共同邻居节点 ABCDE，所以两图中 X 和 Y 之间存在链接的可能性相同；而 AA 算法与 RA 算法是基于共同邻居节点的度数，从图中可以发现 $ABCDE$ 每个节点都具有相同的度数，所以对于节点 X 和 Y，AA 与 RA 算法也具有相同的预测结果。

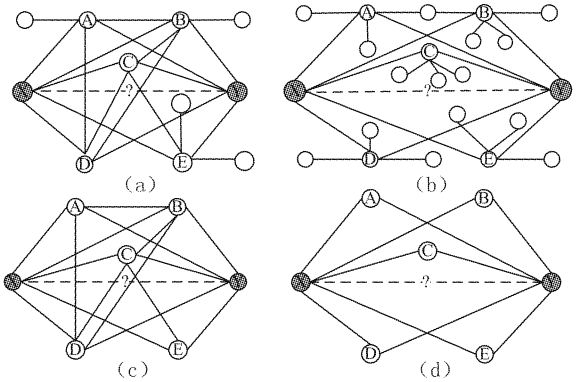

图 1 节点 X 和 Y 与它们共同邻居组成的网络

因此，传统算法对于图 1(a)、(b)中节点 X 和 Y 产生链接的可能性计算结果是相同的。考虑分别从图 1(a)、(b)中抽出的只含 $XYABCDE$ 7 个节点的子网络，如图 1(c)、(d)所示，图(c)中节点之间的链接紧密性比图(d)更强。本文认为图(a)中节点 X 和 Y 存在链接的可能性大于图(b)，因此，我们提出一种新算法，该算法在计算节点 X 和 Y 产生链接的可能性时考虑了共同邻居节点之间的相互关系对于预测可能性的影响，算法中节点 X 和 Y 的相似性定义如下：

$$S_{xy}=\sum_{z\in\varphi(x)\cap\varphi(y)}\frac{e_z}{k(z)}$$

式中，e_z 为节点 z 与其他共同邻居之间及与节点 X 和 Y 的链接数，IA 算法伪代码如表 1 所列。

表 1　节点引力算法伪代码

```
IA 算法
输入：网络 G=(V,E)，节点 x 和 y。其中 x. neighbors 表示节点 x 的邻居集合，x. degree 表示节点 x 的邻居数，xy. CommonNeighbors 表示节点 x 和 y 的共同邻居集合，x. edge 表示节点 x 与其他共同邻居节点之间的链接数
输出：节点 x 和 y 的相似性 Sxy
1 for each x in V do
2   empty xy. CommonNeighbors
3   Sxy<−0
4   for each m in x. neighbors do
5     for each n in m. neighbors do
6       if(n==y)
7         push m to xy. CommonNeighbors
8       end if
9     end for
10  end for
11  for each p in xy. CommonNeighbors do
12    for each q in p. neighbors do
13      for each r in xy. CommonNeighbors do
14        p. edge ++
15      end for
16    end for
17    add (p. edge+2)/(p. degree) to Sxy
18  end for
19 end for
```

对于图 1(a)和图 1(b)两种情况，分别计算节点 X 和 Y 的节点相似性。

图 a：$S_{xy}=\frac{4}{5(A)}+\frac{5}{6(B)}+\frac{5}{5(C)}+\frac{5}{5(D)}+\frac{3}{5(E)}=\frac{127}{30}$；

图 b：$S_{xy}=\frac{2}{5(A)}+\frac{2}{6(B)}+\frac{2}{5(C)}+\frac{2}{5(D)}+\frac{2}{5(E)}=\frac{58}{30}$；

由于新算法对共同邻居节点间相互关系的考虑，因此赋予图(a)和(b)中节点 X 和 Y 不同的节点相似性。IA 算法对共同邻居之间相互关系的计算是在表 1 第 7 行获得所有的共同邻居后，在每一个共同邻居 m 的邻居中查找是否存在其他共同邻居，如果存在，则说明共同邻居间含有一条链接，当共同邻居组成的小型网络中含有的链接数越多，IA 算法就会赋予 S_{xy} 更多的分数。

5　实验

5.1　数据集

本文使用了 8 种具有代表性的公用网络：Coauthorships in network science 网络(NS)[13]、Power Grid 网络(PG)[13]、High-energy theory collaborations 网络(Hep)[13]、AlexArenas' sJazz(Jazz)[13]、AlexArenas' s Email 网络(Email)[13]、Neural network of Elegans 网络(Cele)[13]、Political blogs 网络(PB)[13]、US Air 网络(USAir)[14]。它们的网络拓扑性质如表 2 所列。其中 N，M 分别表示网络的节点数和边数，de 表示网络密度，Nc 为网络的最大连通分量，例如 379/268 表示科学家合作网络中有 268 个连通分量，最大连通分量包含 379 个节点，Cc 表示网络的极大团，例如 20/392 表示科学家合作网络中有 392 个极大团，最大极大团含有 20 个节点，C 表示该网络的聚集系数[15]，r 表示网络的同配系数[16]，K 表示网络的平均度。

表 2　6 个公用实验网络的拓扑性质

Nets	N	M	de	Nc	Cc	C	r	K
NS	1461	2742	3.888E-4	379/268	20/392	0.6937	0.462	3.7536
PG	4941	6594	3.273E-4	4941/1	6/464	0.0801	0.003	2.6691
Hep	8361	15751	6.741E-5	5835/581	24/3581	0.4420	0.2939	3.76771
Jazz	198	2742	5.269E-5	198/1	30/738	0.6175	0.020	27.6969
Email	1133	5451	7.628E-5	1133/1	12/2045	0.2202	0.078	9.6222
Cele	453	2025	2.211E-4	453/1	9/650	0.6465	−0.226	8.9404
PB	1490	16715	1.067E-5	1222/2	20/48932	0.2627	−0.221	22.4362
USAir	332	2126	1.469E-4	332/1	22/634	0.6252	−0.208	12.8072

5.2　实验结果

将 CN，AA，RA，IA 4 种算法在 8 个现实网络中进行实验，并比较预测准确率。每个实验结果是通过对原始数据集进行 200 次随机划分形成的训练集(含 90%的链接数)和测试集(含 10%的链接数)进行预测和评估得到的平均值。其中在 AUC 评估方法中，进行了 2000000 次随机抽取比较，而在 Precision 评估方法中，L 的取值为 100，即取预测结果中分数在前 100 名的链接与测试集进行比较。

5.3　实验分析

通过对不同算法的结果分析，根据 AUC 评估方法，IA 算法在 7 种不同性质的实际网络中都取得了明显的优势，在 USAir 网络中，仅次于 RA 算法，如表 3 所列。在表 4 中，根据 Precision 评估方法，IA 算法在除 Cele 网络和 USAir 网络的其他 6 个网络中都取得了最好的预测准确率，在 Cele 网络和 USAir 网络中的准确率仅次于 RA。IA 算法的结果表明，对共同邻居节点之间相互关系的关注提升了预测算法的准确率。

从网络变化观察，链接预测对大部分网络都取得了不错的预测准确率，但是对 PG 网络的预测结果只是稍好于随机预测的结果。从表 2 的网络拓扑信息中可以发现 PG 网络与其他 7 种网络明显的区别是，PG 网络具有低数量级的网络聚集系数和较低的网络平均度。因此，基于节点相似性的预测算法对不同拓扑性质的网络具有不同的效果。

在算法效率方面，CN 算法对每一个被预测的节点 X，第一步得到 X 的邻居，第二步查找 X 的每一个邻居 m 的邻居中是否含有节点 Y，所以 CN 算法的时间复杂度为 $O(Nk^2)$[17]；明显地，AA、RA 算法同 CN 算法具有相同的时间复杂度；对于 IA 算法，表 1 中 2−10 行查找 X 和 Y 的共同邻居，时间复杂度为 $O(k^2)$，11−18 行是在获得节点 X 的共同邻居后，在每一个共同邻居 m 的邻居中查找是否存在其他共同邻居，其时间复杂度为 $O(kn^2)$(n 为某共同邻居节点 m 与其他共同邻居之间的链接数，k 为 m 的邻居数)，所以 IA 算法的时间复杂度为 $O(Nk^2+Nkn^2)$。算法实际运行时间如图 2 所示，由于 PB 网络具有较高的平均度，因此 IA 算法运行时间明显高于其他 3 种算法，但仍在相同的数量级范围内；而在其他 7 种网络中，IA 算法时间复杂度仅略高于 $O(Nk^2)$。所以，IA 算法在保证时间复杂度的情况下，取得了较好的预测效果。

(下转第 199 页)

2009:427-436

[14] Quinn M J. MPI与OpenMP并行程序设计[M]. 陈文光,武永卫,等译. 北京:清华大学出版社,2004,10

[15] Hanyf,Esbensenh,Song L J. Performance improvement of a genetic algorithm for floor planning with parallel computing technology[J]. IEEE,1997:642-786

[16] Calegari P,Guidec F,Kuonen P. Parallel Island2 Based Genetic Algorithm for Radio Network Design[J]. Journal of Parallel Distributed Computing,1997,47(1):68-102

[17] Arunadevi J,Johnsanjeevkumar A,Sujutha N. Intelligent transport route planning using parallel genetic algorithms and MPI in high performance computing cluster[C]//The 15th International Conference Advanced Computing and Communications. 2007

[18] 任子武,伞冶. 实数遗传算法的改进及性能研究[J]. 电子学报,2007,35(2):269-274

[19] Yussof S,Razali R A,See O H,et al. A Coarse-grained Parallel Genetic Algorithm With Migration for Shortest Path Routing Problem[C]//The 11th IEEE International Conference on High Performance Computing and Communications. 2009

[20] 汪少敏,赵猛. 基于多核处理器并发计算软件构架设计与实现[J]. 计算机科学,2008,35(7):283-285

[21] Dong Li, de Dupinski B R, Nikolopoulos D S. Hybrid MPI/OpenMP power-aware computing[C]// Parallel &Distributed Processing(IPDPS): 2010 IEEE International Symposium on Digital Object Identifier. 2010:1-12

[22] Xu Guang-hua,Liu Dan,Liang Lin. Immune Clonal selection optimization method with combining mutation strategies[J]. Journal of Xian Jiao Tong University,2007,19(2):177-181

(上接第164页)

表3 实验结果(AUC) $n=2000000$

Algo	NS	PG	Hep	Jazz	Email	Cele	PB	USAir
CN	0.9335	0.5882	0.8900	0.9252	0.8429	0.8779	0.9259	0.9130
AA	0.9321	0.5877	**0.8905**	0.9368	0.8420	0.9298	0.9282	0.9301
RA	0.9366	0.5881	0.8904	0.9425	0.8431	0.9355	**0.9301**	**0.9393**
IA	**0.9383**	**0.5884**	**0.8905**	**0.9476**	**0.8443**	**0.9373**	**0.9301**	0.9348

表4 实验结果(Precision) $L=100$

Algo	NS	PG	Hep	Jazz	Email	Cele	PB	USAir
CN	0.8214	0.1024	0.4101	0.8090	0.2821	0.1947	0.4101	0.5887
AA	0.9544	0.0689	0.3710	0.8267	0.3069	0.2483	0.3710	0.6073
RA	0.9585	0.0546	0.2411	0.8178	0.2484	**0.3123**	0.2411	**0.6209**
IA	**0.9619**	**0.1061**	**0.4628**	**0.8832**	**0.3426**	0.2887	**0.4628**	0.6100

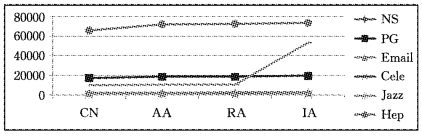

图2 算法运行时间(单位:毫秒)

结束语 随着复杂网络的快速发展,基于网络结构的链接预测方法得到了不断的发展与完善。本文从节点共同邻居之间相互关系的角度提出了一种基于节点相似性的链接预测算法——节点引力算法,该算法是在考虑单个共同邻居的同时,考虑所有共同邻居之间的相互关系。同时,本文将新算法与其他优秀的基于节点相似性的算法同时作用于8种不同的现实网络,在不丢失时间效率的情况下,新算法取得了更好的预测准确率。下一步,我们的工作将致力于研究基于节点相似性的预测算法与网络拓扑结构的关系,并积极研究大规模现实网络的链接预测问题。

参 考 文 献

[1] Getoor L,Diefl C P. Link mining:a survey[J]. ACM SIGKDD Explorations Newsletter,2005,7(2):3-12

[2] Lv Lin-yuan. Link Prediction on Complex Networks[J]. Journal of University of Electronic Science and Technology of China,2010,9:5-39

[3] Sarukkai R R. Link prediction and path analysis using markov-chains[J]. Computer Networks,2000,33:377

[4] Zhu J,Hong J,Hughes J G. Using markov chains for link prediction in adaptive web sites[J]. Computer Science,2002,2311:22

[5] Popescul A,Ungar L. Statistical relational learning for link prediction[C]// Proc. Workshop on Learning Statistical Models-from Relational Data,2003. New York:ACM Press,2003:81

[6] Clauset A,Moore C,Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature,2008,453:98-101

[7] Lichtenwalter R N,Lussier J T,Chawla N V. New Perspectives and Methods in Link Prediction[C]//ACM SIGKDD. 2010

[8] Liben-Nowell D,Kleinberg J. The LinkPrediction Problem for Social Networks[J]. Journal Am. Soc. Inform. Sci. Technol,2007,58:1019

[9] Adamic L A,Adar E. Friends and neighbors on the web[J]. Social Networks,2003,25:211

[10] Zhou Tao,Lv Lin-yuan,Zhang Yi-cheng. Predicting missinglinks via local information[J]. The European Physical Journal B,2009,71:623

[11] Hanely J A,Mcneil B J. The meaning and use of the area under a receiver operating characteristic curve[J]. Radiology,1982,143:29-36

[12] Herlocker J L,Konstann J A,Terveen K,et al. Evaluating collaborative filtering recommender systems[J]. ACM Transaction Information System,2004,22(1):5-53

[13] Newman M. Network DataSets [OL]. http://www-personal.umich.edu/~mejn/netdata/

[14] Batageli,Mrvar A. Pajek Datasets[OL]. http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm

[15] Watts D J,Strogatz S H. Collective dynamics of small-world networks[J]. Nature,1998,393:440

[16] Newman M E J. Assortative mixing in networks[J]. Phys. Rev. Lett,2002,89:208701

[17] Lv Lin-yuan,Zhou Tao. Similarity index based on local paths for link prediction of complexnetworks[J]. Phys. Rev. E,2009,80:046122