

一种改进的加权网络链接预测方法

陈 旭 陈可佳

(南京邮电大学计算机学院 南京 210003)

摘 要 目前,复杂网络的链接挖掘问题已得到了广泛研究,而加权网络的相关研究还较少且结果不甚理想。鉴于此,提出一种新的针对加权网络的链接预测方法,对以往方法中的加权相似性度量进行改造。新方法主要基于这一假定:链接 xz 为强关系而链接 zy 为弱关系时,链路 $\langle x, z, y \rangle$ 对节点 x 和 y 之间形成链接的贡献最低。因此,新方法中链接 xz 为强关系而链接 zy 为弱关系时,链路 $\langle x, z, y \rangle$ 对节点 x 和节点 y 之间的相似性得分 $S(x, y)$ 的贡献度的削弱程度最大。在带权网络数据集 USAir 和 NetScience 上的比较实验表明,新方法在 AUC 指标上具有一定的优势。

关键词 加权复杂网络,链接预测,相似度指标,弱关系理论

中图分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.018

Improved Link Prediction Method for Weighted Networks

CHEN Xu CHEN Ke-jia

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract Currently, link mining in complex networks has been extensively studied. However, there are only a few related works on weighted networks and the results are not satisfactory. A new link prediction method for weighted networks was proposed by improving the weighted similarity measure of network structure. The new method is based on the assumption that when the link xz is strong and the link zy is weak, the link $\langle x, z, y \rangle$ has the least contribution to the link between node x and y . Therefore, in the new method, as the link xz is strong and the link zy is weak, the degree of weakening of the link $\langle x, z, y \rangle$ to the contribution degree of the similarity score $S(x, y)$ between the node x and y is maximal. Comparative experiments on weighted dataset USAir and NetScience show that the proposed method has better performance in AUC indicators.

Keywords Weighted complex networks, Link prediction, Similarity index, Weak ties theory

1 引言

许多复杂系统可由网络来描述,如航空网络、社交网络等。网络中的节点代表系统中的个体,节点之间的边代表个体之间的关系。近年来,复杂网络分析已成为数据挖掘领域的热点研究问题,而链接预测则是其中的重点任务之一。链接预测的目标为:基于给定网络在某个时刻 t 已存在的节点和链接,预测网络在另一个时刻 $t'(t' > t)$ 中可能出现的链接。此外,链接预测还可用于预测网络的缺失链接和错误链接(spurious links)^[1]。

目前,链接预测算法主要分为3类^[2]:基于相似度的方法,基于最大似然估计的方法和基于概率模型的方法。最常用的基于相似度的方法主要有3种:基于局部信息的相似度,基于路径的相似度和基于随机游走的相似度。大多数已有链接预测方法将链接状态视为存在或不存在,而较少考虑链接的权重。但在现实网络中,每条边的重要性和意义可能有所不同。以航空网络为例,假设航站A、航站B和航站C每日

均有航班往来,若A和B之间每日往来的航班数为119,A和C之间每日往来的航班数为44,则航线AB较航线AC更重要,对航站A形成新航线的影响也更大。因此,在度量网络结构相似度时还应考虑链接权值。如何充分利用链接权值信息来提高复杂加权网络链接预测方法的准确度,是本文的主要研究目标。

最近,加权网络链接预测问题得到了部分研究者的关注。Murata等人^[3]将共同邻居数(Common Neighbors, CN)、Adamic-Adar系数(AA)和优先连接(Preferential Attachment, PA)等指标进行加权改造,并在问答公告板(Question Answering Bulletin Boards)中进行链接预测。实验发现,通过对上述指标加权,链接预测的准确度有了显著提高。然而,Zhou等人^[4]在将加权指标应用于合著者网络NetScience,CGScience以及航空网络USAir中时却发现,采用加权指标的方法比不采用加权指标的方法效果更差。Liben-Nowell和Kleinberg^[5]在应用加权Katz指标时也发现了与期望不一致的类似结果。

收稿日期:2016-09-13 返修日期:2017-01-18 本文受国家自然科学基金面上项目(61571238),国家自然科学基金青年科学基金(61302158,61302157)资助。

陈旭(1992—),男,硕士,主要研究方向为社会网络分析,E-mail:ntchenxu@126.com(通信作者);陈可佳(1980—),女,博士,副教授,主要研究方向为机器学习、数据挖掘、社会网络分析等。

基于这一背景,本文将进一步研究加权网络链接预测问题,着重思考如何对相似度指标进行更合理且有效的加权。

2 研究背景

2.1 加权网络链接预测问题

令 $G(V, E, W)$ 表示一个无向带权网络图,其中 V 是图中节点的集合, E 是图中链接的集合, W 是图中链接权值的集合。假设 G 中不存在自环,两个节点之间也不存在多个链接。在 G 中的链接预测问题可形式化为:首先将 E 分成两部分,即训练集 E^T 和测试集 E^P ($E^T \cup E^P = E, E^T \cap E^P = \emptyset$)。在 E^T 中,根据相似性指标计算所有节点对 $\langle x, y \rangle$ 的相似度得分 $S(x, y)$,其中它们的一阶相邻节点集分别为 $\Gamma(x)$ 和 $\Gamma(y)$ 。最后,根据相似度得分和 E^P ,计算该方法的 AUC (Area Under the Receiver Operating Characteristic Curve) 评价指标。

本文沿用了基于相似度的链接预测方法。Zhou 等人^[6]比较了几种常见的局部相似度指标,发现 CN 指标和 AA 指标的表现最好。随后, Zhou 等人^[7]又提出了一种优于 CN 和 AA 的 RA (Resource Allocation) 指标。因此,本文也采用 CN, AA 和 RA 这 3 个指标。

与无权网络链接预测问题不同,加权网络链接预测在计算相似度指标时,需考虑节点间链接的权值。已有的部分方法是将权值直接累加^[8-9](如表 1 所列),而未考虑到两个节点与其共同邻居节点间链接权值的差异对这两个节点间形成新链接的贡献不同。

表 1 常用的加权度量

$s_{WCN}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} w(x, z) + w(z, y)$
$s_{WAA}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(z, y)}{\log(1 + s(z))}$
$s_{WRA}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z) + w(z, y)}{s(z)}$

表 1 中, $s(z) = \sum_{m \in \Gamma(z)} w(z, m)$ 。

2.2 链接权值差异问题的理论分析

已有研究表明,链接权值的差异对链接预测的结果存在重要影响。Granovetter^[10]于 1974 年提出了弱关系理论,发现强关系圈中存在大量的冗余信息,而真正有用的信息往往来自弱关系。

Lv 等人^[11]率先研究了弱关系理论在加权网络链接预测中的作用,并通过实验证明链接预测问题也存在弱关系效应,即权值较小的边在链接预测中起着重要的作用。他们将数据集中的链接按权值从大到小进行排序,将排名前 50% 的链接定义为强关系,将排名后 50% 的链接定义为弱关系。

通过对 3 个经典的数据集 (USAir, C. elegans 和 CG-Science) 的分析^[12-13]发现:若任意两个不直接相连的节点 x, y 与它们的一阶共同邻居 z 的链接 $\langle x, z \rangle, \langle y, z \rangle$ 均为强关系,则 x, y 之间能形成直接链接 L 的概率为:

$$P_1(L) = \frac{3a+b}{3a+b+c}$$

若 $\langle x, z \rangle, \langle y, z \rangle$ 均为弱关系,则 x, y 之间能形成直接链接 L 的概率为:

$$P_2(L) = \frac{3d+e}{3d+e+f}$$

由此,可进一步计算得到,若 $\langle x, z \rangle$ 为强关系,而 $\langle y, z \rangle$ 为弱关系,则 x, y 之间能形成直接链接 L 的概率为:

$$P_3(L) = \frac{2b+2e}{2b+2e+g}$$

本文对上述 3 个网络中任意两节点间形成链接的 3 种不同情况分别进行概率统计,如表 2 所列。

表 2 网络中任意两节点间形成链接的不同情况的概率统计/%

	$P_1(L)$	$P_2(L)$	$P_3(L)$
USAir	73.9	75.7	62.5
C. elegans	43.2	58.2	37.7
CGScience	43.5	34.4	39.3

由表 2 可知,在不同网络中,强关系或弱关系对形成链接的贡献是不同的。在稠密图(如 USAir 和 C. elegans)中,当 $\langle x, z \rangle$ 和 $\langle y, z \rangle$ 为强关系或弱关系时 x 和 y 之间形成链接的概率,要大于当 $\langle x, z \rangle$ 为强关系而 $\langle y, z \rangle$ 为弱关系时 x 和 y 之间形成链接的概率。在稀疏图(如 CGScience)中,当 $\langle x, z \rangle$ 和 $\langle y, z \rangle$ 均为弱关系或 $\langle x, z \rangle$ 为强关系而 $\langle y, z \rangle$ 为弱关系时 x 和 y 之间形成链接的概率。

图 1 表示某稠密的社交网络的局部结构,节点 A, B, C, D 是网络中的 4 个用户,节点间的链接表示一个月内用户之间建立过联系,链接上的权值表示这个月内用户间联系的频次。

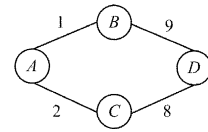


图 1 某社交网络局部图

图 1 中,链接 $\langle A, B \rangle$ 和 $\langle A, C \rangle$ 为弱关系,链接 $\langle D, B \rangle$ 和 $\langle D, C \rangle$ 为强关系。由之前的分析可知,节点 B, C 之间形成链接的概率要大于节点 A, D 之间形成链接的概率。然而,采用表 1 中加权相似度指标得到的结果却是上述两个概率值相等。

3 本文提出的方法

本文方法对 CN, AA 和 RA 的加权相似度指标进行了改进,得到 New_WCN, New_WAA 和 New_WRA。根据上述对链接权值差异问题的理论分析可知,在不同网络中,当 $\langle x, z \rangle$ 为强关系而 $\langle y, z \rangle$ 为弱关系时 x 和 y 之间形成链接的概率最小,因此链路 $\langle x, z, y \rangle$ 对节点 x 和 y 之间形成链接的贡献最低。新的相似度指标不再简单累加链接 $\langle x, z \rangle$ 和 $\langle y, z \rangle$ 的权值,而考虑了权值大小对新链接的不同影响,即对任意两个不直接相连的节点 x 和 y 与它们的一阶共同邻居 z 的链接 $\langle x, z \rangle$ 和 $\langle y, z \rangle$ 均有 $s(x, y) = \min\{w(x, z), w(z, y)\}$ 。可以明显地看出,该改进加权指标中 $\langle x, z \rangle$ 和 $\langle y, z \rangle$ 均为强关系或弱关系时,链路 $\langle x, z, y \rangle$ 对节点 x 和节点 y 之间的相似性得分 $S(x, y)$ 的贡献度的削弱程度较小;而当 $\langle x, z \rangle$ 为强关系而 $\langle y, z \rangle$ 为弱关系时链路 $\langle x, z, y \rangle$ 对节点 x 和节点 y 之间的相似性得分 $S(x, y)$ 的贡献度削弱程度较大,直接有效地降低了 $\langle x, z \rangle$ 为强关系而 $\langle y, z \rangle$ 为弱关系时链路 $\langle x, z, y \rangle$ 对节点 x 和 y 之间形成链接的贡献度。上述具体做法可由式(1)~式(3)表示。

New_WCN 度量:

$$s(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \bar{w}(x, z, y) - \sigma \quad (1)$$

New_WAA 度量:

$$s(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\bar{w}(x, z, y) - \sigma}{\log(1 + s(z))} \quad (2)$$

New_WRA 度量:

$$s(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\bar{w}(x, z, y) - \sigma}{s(z)} \quad (3)$$

其中, $\bar{w}(x, z, y) = \frac{w(x, z) + w(z, y)}{2}$, $\sigma = |w(x, z) - \bar{w}(x, z, y)|$,

$z, y |, s(z) = \sum_{m \in \Gamma(z)} w(z, m)$.

本文方法的具体步骤如下:

(1) 计算所有当前不相连节点对的相似性得分。

(2) 按照相似性得分从大到小排列, 对于相似性得分相同的节点对, 计算它们的 σ 值。 σ 值越大, 排名越靠后。

(3) 得到相似性得分的最终排名, 越靠前的节点对形成链接的可能性越大。

4 实验

4.1 实验设置

本文实验在两个真实的带权网络数据集 USAir 和 NetScience^[14] 上进行。USAir 中的链接权值代表两个航站间的航班频次, NetScience 中的链接权值表示两个作者间的合作频次。值得注意的是, 这两个网络均不存在自环, 且节点间无多条链接。

实验比较了基于 WCN, WAA 和 WRA 指标的方法与本文提出的基于 New_WCN, New_WAA 和 New_WRA 指标的方法在上述网络中的 AUC 值。AUC 可以理解为在测试集中边的分数值 $S(x, y)$ 比随机选择的一个不存在的边的分数值 $S(x, y)$ 高的概率^[15]。具体为: 独立比较 n 次, 每次从测试集中随机选取一条边并从未链接集合中随机选取一条不存在的边, 如果前者的分数值大于后者, 则加 1 分; 如果两个分数值相等, 则加 0.5 分。如果总共有 n' 次大于的情况和 n'' 次相等的情况, 则 AUC 定义为:

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

实验步骤如下:

(1) 构建网络的无向带权图, 并将链接权值存入稀疏矩阵 W 中。

(2) 根据 W , 在所有已连接的节点对中随机选取 10% 的节点对作为测试集 probe。根据表 1 和式(1)~式(3)计算 probe 集中已选节点对的各个相似性指标得分。

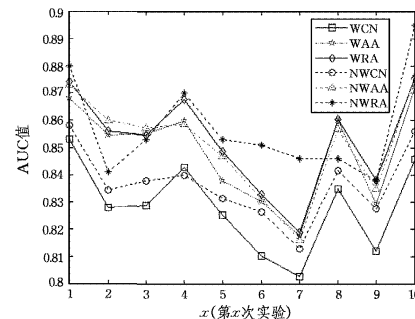
(3) 根据 W 选取所有未连接的节点对作为 no_link 集。根据表 1 和式(1)~式(3)计算 no_link 集中所选取的节点对的各个相似性指标得分。

(4) 分别从 probe 集与 no_link 集中随机选取一对节点对, 并比较它们的相似性指标得分, 重复这一操作 10000 次, 计算该指标的 AUC 值。

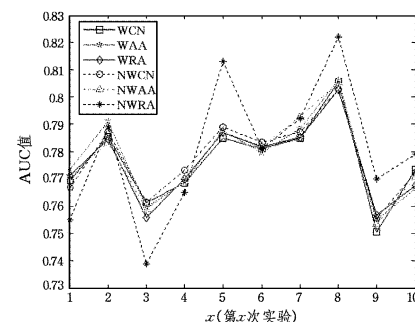
(5) 重复步骤 2~步骤 5 10 次, 最后计算出各个指标的平均 AUC 值。

4.2 实验结果与分析

图 2 比较了在数据集 USAir 和 NetScience 上各方法基于各加权相似度指标的链接预测方法的 AUC 值。图 3 比较了在数据集 USAir 和 NetScience 上各方法基于各加权相似度指标的平均 AUC 值。

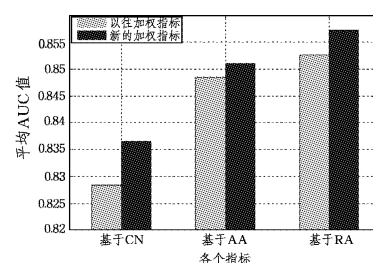


(a) USAir

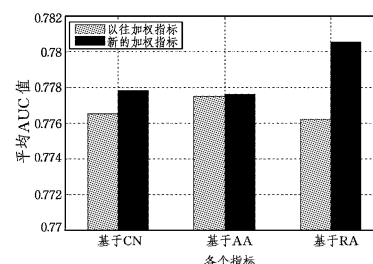


(b) NetScience

图 2 各方法的 AUC 值比较



(a) USAir



(b) NetScience

图 3 各方法的平均 AUC 值比较

由图 2 可知, 本文采用新加权指标方法的 AUC 值绝大部分都大于基于以往加权指标的方法的 AUC 值。在少数点处, 例如在数据集 USAir 上的第 4 次实验中, New_WCN 和 New_WAA 指标的 AUC 值略小于 WCN 和 WAA 指标的 AUC 值; 在数据集 NetScience 上的第 1, 3 次实验中, New_WRA 指标的 AUC 值小于 WRA 指标的 AUC 值。其可能的

(下转第 108 页)

IEEE Global Communications Conference, Xi'an, Globecom, 2014;986-992.

- [11] RINA K, NATH S, MARCHANG N, et al. Can Clustering be Used to Detect Intrusion During Spectrum Sensing in Cognitive Radio Networks[J]. IEEE Systems Journal, 2016(99): 1-10.
- [12] MIN A W, ZHANG X, KANG G S. Detection of small-scale primary users in cognitive radio networks[J]. IEEE Journal of Selected Areas in Communications, 2011, 29(2): 349-361.
- [13] KOH C W K, YAU K L A. Trust and reputation scheme for clustering in Cognitive Radio Networks[C]//International Con-

ference on Frontiers of Communications, Networks and Applications (ICFCNA 2014). Malaysia, 2014; 1-9.

- [14] QIN T, YU H, LEUNG C, SHEN Z, et al. Towards a trust aware cognitive radio architecture[J]. ACM SIGMOBILE Mobile Computer Communication Review, 2009, 13(2): 86-95.
- [15] DU Z Y, CHEN H N, SONG F. SNR based Weighted-Consensus Algorithm for Cooperative Spectrum-sensing[J]. Journal of Data Acquisition and Processing, 2013, 28(2): 184-189. (in Chinese) 杜智勇, 陈浩楠, 宋菲. 一种基于信噪比加权共识的协作频谱感知算法[J]. 数据采集与处理, 2013, 28(2): 184-189.

(上接第 98 页)

原因为, New_WCN 和 New_WAA 指标在稀疏图(如 Net-Science)中表现更佳, 而 New_WRA 指标在稠密图(如 USAir)中表现更佳。由图 3 可知, 本文方法的平均 AUC 值较以往的加权指标均有一定程度的提高, 特别在数据集 USAir 中, New_WCN 的平均 AUC 值比以往的 WCN 指标提高了约 0.01, 这说明本文方法在数据集 USAir 和 NetScience 上的总体性能优于以往的加权方法。

结束语 通过对加权网络链接预测的研究, 本文认为已有的加权方法在计算相似度时存在不足, 并由此提出了一种新的加权网络链接预测方法, 对其中的加权相似度指标进行了改进。实验验证了新的加权指标比以往的加权指标更加有效。然而, 在实验中有时会出现新方法的 AUC 值反而更低的情况, 因此, 判别针对哪些指标的加权改进更为有效并追究其原因, 将是未来研究的重点。此外, 新方法性能的提高并不十分显著, 对加权指标的设计需要进一步讨论。最后, 网络科学已成为最具活力的交叉学科, 它的不断发展有望解决许多具有重要意义的问题, 如预测金融危机的产生和发展^[16], 预测人际关系和社会结构的形成^[17], 预测全球传染病的传播^[18], 揭示生物的基本功能得以实现的机制^[19]等, 给人类带来无尽的社会经济价值。

参 考 文 献

- [1] HOLLAND P W, LASKEY K B, LEINHARDT S. Stochastic blockmodels; first steps [J]. Social Networks, 1983, 5(2): 109-137.
- [2] LV L Y. Link prediction on complex networks [J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 652-660. (in Chinese) 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5): 652-660.
- [3] MURATA T, MORIYASU S. Link prediction of social networks based on weighted proximity measures[C]//IEEE/WIC/ACM International Conference on Web Intelligence. Silicon Valley, CA, USA, 2007: 85-88.
- [4] LV L Y, ZHOU T. Role of weak ties in link prediction of complex networks[C]//Proceedings of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management. New York, USA, 2009: 55-58.
- [5] LIBEN-NOWELL D, KLEINBERG J. The link prediction problem for social networks [J]. Journal of the American Society for

Information Science & Technology, 2003, 54(5): 1345-1347.

- [6] LV L Y, JIN C H, ZHOU T. Similarity index based on local paths for link prediction of complex networks [J]. Physical Review E, 2009, 80(2): 046122.
- [7] ZHOU T, LV L Y, ZHANG Y C. Predicting missing links via local information [J]. Physics of Condensed Matter, 2009, 71(4): 623-630.
- [8] GÜL S, KAYA M, KAYA B. Predicting links in weighted disease networks[C]//2016 3rd International Conference on Computer and Information Sciences (ICCOINS). Kuala Lumpur, Malaysia, 2016: 77-81.
- [9] MARTINČIĆ-IPŠIĆ S, MOČİBOB E, MEŠTROVIĆ A. Link Prediction on Tweets' Content [C]//2016 International Conference on Information and Software Technologies; Information and Technologies. 2016: 559-567.
- [10] GRANOVERTER M S. Strength of weak ties [J]. American Journal of Sociology, 1973, 78(6): 1360-1380.
- [11] LV L Y, ZHOU T. Link prediction in weighted networks; the role of weak ties [J]. Europhysics Letters, 2010, 89(1): 18001.
- [12] MILO R, SHEN-ORR S, ITZKOVITZ S, et al. Network motifs: simple building blocks of complex networks [J]. Science, 2002, 298(5594): 824-827.
- [13] MILO R, ITZKOVITZ S, KASHTAN N, et al. Super families of evolved and designed networks [J]. Science, 2004, 303(5663): 1538-1542.
- [14] 吕琳媛, 周涛. 链路预测[M]. 北京: 高等教育出版社, 2013: 290.
- [15] SHARMA S, SINGH A. An efficient method for link prediction in weighted multiplex networks[C]//International Conference on Signal-image Technology & Internet-based Systems. 2015: 453-459.
- [16] SCHWEITZER F, FAGIOLO G, SORNETTE D, et al. Economic networks; The new challenges [J]. Science, 2009, 325(5939): 422-425.
- [17] BORGATTI S P, MHEGRA A, BRASS D J, et al. Network analysis in the social sciences[J]. Science, 2009, 323(5916): 892-895.
- [18] COLIZZA V, BARRAT A, BARTHÉLEMY M, et al. The modeling of global epidemics; Stochastic dynamics and predictability [J]. B Math Biol, 2006, 68(8): 1893-1921.
- [19] BARABÁSI A L, OLTVAI Z N. Network biology; Understanding the cell's functional organization [J]. Nat Rev Genet, 2004, 5(2): 101-113.