社交网络中的链路预测方法改进

李旗旗,徐 敏
(南京航空航天大学 计算机科学与技术学院,江苏 南京 211106)

摘 要:社交网络在近些年得到了迅速发展,如今各个行业都在努力加入社交元素,如何提高链路预测方法在社交网络中的预测准确度成为一个热门研究方向。链路预测方法由于网络结构的不同会表现出不同的预测效果,因此可以根据社交网络的结构特性对链路预测方法进行改进,从而提高在社交网络中的预测准确度。社交网络是对人与人之间某种社会关系的描述,因此和其他复杂网络相比,会表现出独特的网络性质和结构,其中最主要的是“小世界”特性和无标度特性。针对社交网络的这种特性,对原有的链路预测方法进行改进,在共同邻居方法的基础上加入了优先连接对节点相似性的贡献。真实社交网络数据集的对比实验结果表明,改进后的方法在没有增加时间复杂度的情况下提高了预测准确度。

关键词:社交网络;链路预测;网络结构;无标度网络
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2017)11-0037-04
doi:10.3969/j.issn.1673-629X.2017.11.008

Improvement of Link Prediction Method in Social Networks

LI Qi-qi, XU Min
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: The social network has been developing rapidly in recent years. Various industries are now trying to integrate social elements, so how to improve the accuracy of link prediction methods in social networks has become a popular research. Due to the different network structures, the link prediction methods will be different in prediction performance so that it can be improved according to the characteristics of social network structure, improving of the accuracy of prediction. The social network is a description of certain social relations between people, so compared with other complex networks, it will exhibit its unique properties and network structure, of which the most important is the “small world” and scale-free characteristics. According to the characteristics of social network, the previous link prediction methods can be improved, adding the contribution of priority connection based on common neighbors. The experiments on real social network data sets show that the improved method can improve the accuracy of prediction without increasing time complexity.

Key words: social networks; link prediction; network structure; scale-free network

0 引 言

近年来蓬勃发展的互联网技术,为社交网络提供了一个优质的平台,在线社交网络应运而生。在线社交网络不仅继承了人们在线下的社会关系,而且突破了地区、领域的限制,使得不同的人之间交流更加方便。在线社交网络出现以来,一直被人们所喜爱,社交网络的规模也因此越来越大。如今人们对于在线社交网络的依赖越来越强,同时对社交网络的要求也越来越高[1]。因此,如果可以根据社交网络已经存在的信息预测未来可能产生的连边,从而用来向用户推荐他们可能感兴趣的人,将会是一项非常有价值的工作。

复杂网络中的链路预测通过计算网络中尚未连边的两个节点之间在未来产生连边的概率达到预测的目的,其研究思路和方法主要基于马尔可夫链和机器学习。Sarukkai 使用马尔可夫链对网络进行了链路预测和路径分析[2]。之后人们在此基础上进行了扩展,并提出了一系列模型,但是这些模型大多结合了网络中的节点属性,然而很多网络中节点属性的可靠性并不能得到保证。因此,国内外学者越来越关注基于网络结构的链路预测方法。相比于节点属性信息,网络的

收稿日期:2016-12-30 **修回日期**:2017-05-04 **网络出版时间**:2017-08-01
基金项目:国家“973”重点基础研究发展计划项目(2014CB744900)
作者简介:李旗旗(1991-),男,硕士研究生,CCF 会员(E200041166G),研究方向为机器学习、链路预测;徐 敏,博士,副教授,研究方向为模式识别、机器学习。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20170801.1600.084.html

结构更加可靠,并且基于网络结构的链路预测方法对于结构相似的网络具有一定的普适性。

Liben-Nowell 和 Kleinberg[3] 提出了基于网络结构的链路预测方法,并在社会合作网络中分析了一些链路预测方法的预测效果。之后,周涛等[4] 在 6 种网络中对 9 种链路预测方法的效果进行对比,发现同一种方法会随着网络结构的变化表现出不同的预测效果,同时也可以根据网络的某些结构特点对链路预测方法进行改进[5]。

随着社交网络的快速发展,对社交网络的研究也逐渐成为一个热门方向,而如何预测社交网络中的关系成为社交网络研究的重要任务。各种社交网络中的链路预测方法因此被提出,如针对微博网络的链路预测研究[6]。社交网络作为复杂网络的一种典型表现形式,反映了社会成员之间复杂的社会关系。社交网络具有复杂网络的一般特征,又因其连边存在的社会性,表现出独特的结构特征[7-8]。根据社交网络的结构特性对链路预测方法进行改进,可以在一定程度上提高链路预测方法在社交网络中的预测效果。

1 链路预测方法

用 $G(V,E)$ 表示一个无向网络,其中 V 表示节点集,E 表示边集。$G[t,t_1]$ 表示 G 在 $[t,t_1]$ 时间段的情况,那么在 $(t_1,t_2]$ 时间段 G 的情况就是 $G(t_1,t_2]$。链路预测关注的就是如何预测网络 G 从 $[t,t_1]$ 到 $(t_1,t_2]$ 的变化。

链路预测经过多年的研究,提出了各种各样的方法。为了能更好地介绍这些方法,首先介绍一些在文章中使用的符号。其中,x,y 表示节点,N 表示网络中节点的数量。k_x 和 k_y 表示节点 x 和 y 的度数,$\Gamma(x)$ 和 $\Gamma(y)$ 分别表示节点 x 和 y 的邻居节点集合。

基于局部信息的相似性方法是指仅通过节点的局部信息(如节点的度和最近邻等)就可以计算出相似度的方法。这种方法的优势在于时间复杂度低,适用于大型的网络应用。

1.1 共同邻居(CN)

共同邻居是一种比较简单的链路预测方法,其基本假设为:如果两个尚未连边的节点有更多的共同邻居,那么它们更倾向于连边。比如在社交网络中,如果两个陌生人有很多共同的朋友,那么他们将来成为朋友的概率也很大[5]。又比如,Newman 等发现在科学家合作网络中,如果两个科学家的共同合作伙伴很多,那么他们将来也很有可能会合作[6]。其定义为:

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

在共同邻居的基础上,考虑两个端点的度对节点相似性的影响,又产生了 Salton[9]、Jaccard[10] 和 LNH-I[11] 等方法。后来的研究表明,这些更复杂的变体在很多情况下都不如 CN 方法的预测效果好,所以选取 CN 作为基于共同邻居的方法的代表。

1.2 Adamic-Adar(AA)

其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[11-13]。例如,在微博网络中,受关注较多的人往往是某个领域的专家或者名人,因此共同关注他们的人之间可能并不拥有特别相似的兴趣。相反,如果两个人共同关注了一个粉丝很少的人(非专家),那么说明这两个人确实具有相同的兴趣爱好或者重叠的社交圈,因此有更高概率相连。

$$s_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(k_z)} \tag{2}$$

周涛等根据网络资源分配的启发,提出了 RA 方法[4]。该方法和 AA 唯一不同的是将公共邻居的权重赋值为度的倒数。

1.3 局部路径(Local Path,LP)

LP 方法在 CN 方法的基础上考虑了三阶路径的影响,其定义为:

$$s_{xy} = A^2_{xy} + \alpha A^3_{xy} \tag{3}$$

其中,α 为可调参数;$\boldsymbol{A}$ 为网络的邻接矩阵,A^3_{xy} 表示节点 x 和 y 之间路径长度为 3 的路径长度。当 α 为 0 时,LP 指标就退化为 CN 指标。实验中选取的 α 值为 0.001。

1.4 Katz

Katz 方法考虑了两个节点在网络中的所有路径对相似性的贡献,其定义为[14]:

$$s_{xy} = \sum_{l=1}^{\infty} \beta^l \mid \mathrm{path}^l_{x,y} \mid = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \cdots \tag{4}$$

其中,β 用来调节高阶路径的贡献。当 β 值很小时,高阶路径的贡献也就很小了,此时 Katz 的预测结果接近于局部路径方法。

2 社交网络的结构特性

2.1 “小世界”特性

“小世界”现象最早出现在 20 世纪 60 年代 Milgram 的信件投递实验中。Milgram 通过实验发现每封信平均经过 6 个人就可以到达目标收件人手中,并据此提出了“六度分离”理论[15]。此后,又接连涌现了一些社交网络上的“小世界”实验,比较著名的是“Kevin Bacon”游戏,它的主要目的是计算电影演员 Bacon 与其他电影演员之间的距离。该实验基于美国 Virginia 大学的电影演员数据库,计算得到全世界 60 万演员与 Bacon 的平均距离为 2.944。2011 年,Facebook 网站与米兰大学等共同进行的六度分离实验[16] 显示,Face-

book 上两个用户之间的平均距离仅为 4.74,实验采用了包含 Facebook 中 7 亿多用户和 687 亿条朋友关系的数据,是自 2011 为止最大规模的小世界实验。有研究表明,不同种类的社交网络都具有“小世界”现象[17]。

“小世界”现象引发人们对如何构造一个小世界网络模型的思考,比较著名的小世界网络模型是 WS 小世界模型[18]和 NW 小世界模型[19]。两个网络都从规则网络开始。WS 小世界模型以概率 p 进行随机重连,而 NW 小世界模型以概率 p 随机选择一定数量的节点对,然后在这些节点对之间添加边。研究表明,两个模型都满足“小世界”网络的结构特性,但是在可搜索性角度没有进行考虑。Kleinberg 从可搜索性角度对 NW 小世界模型进行了修改。Kleinberg 认为,在两个节点之间增加连边不应该是随机的,而是要随着两个节点之间的网格距离的增加来降低产生连边的概率[20]。Kleinberg 模型在满足“小世界”特性的基础上增加了网络的可搜索性,更加符合实际的网络。

2.2　无标度特性

无标度特性指网络的度分布服从幂律分布,由于这类网络中节点的度没有比较明显的特征长度,因此称这类网络为无标度网络。无标度网路中的度分布为:

$$P(k) \sim k^{-\lambda} \tag{5}$$

其中,λ 为幂指数,取值通常在 2 ~ 3 之间。

最著名的无标度网络模型是 BA 无标度网络模型[21]。该模型的核心思想在于考虑了网络的增加和优先连接。网络的增长指的是节点和连边的增加,优先连接指的是新加入的节点更容易与度数大的节点相连,也称为“马太效应”。

根据对 BA 无标度网络模型构造的网络研究,发现网络中度为 k 的节点出现的概率 $P(k)$ 近似与 k^{-3} 成正比,满足幂律分布。正是这个原因,文中认为同样满足度分布为幂律分布的社交网络也存在优先连接的现象。随后有研究表明,社交网络确实具有无标度性,即网络中大度数节点更倾向于与其他大度数节点建立连接,而小度数节点更倾向于与小度数节点相连。

3　链路预测方法的改进方案

针对社交网络中链路预测方法的改进主要是基于社交网络的小世界特性和无标度特性,因此改进方案的核心思想是在两个节点共同邻居的基础上加入对优先连接特性的考虑。在具有无标度特性的网络中,一个节点 x 与另一个节点 y 产生连边的概率与 y 的度 k_y 成正比[22],同理,节点 y 与节点 x 产生连边的概率与 x 的度 k_x 成正比,因此节点 x 和节点 y 产生连边的概率与 $k_x k_y$ 的乘积成正比。再根据“小世界”现象中两个节点之间的连边概率随着距离的增加而衰减,需要对 $k_x k_y$ 设置一个衰减系数 α,用于惩罚距离比较大的节点对。由于是在 CN 的基础上进行改进,所以将改进的方法命名为 ImprovedCN,其定义为:

$$s_{xy} = A^2_{xy} + \alpha k_x k_y \tag{6}$$

其中,$\boldsymbol{A}$ 为网络的邻接矩阵;A^2_{xy} 表示节点 x 和节点 y 的共同邻居个数;k_x 和 k_y 分别表示节点 x 和节点 y 的度;α 为一个可调参数,用于调节两个节点度的乘积对两个节点相似性的贡献。当 α 很小时,两个节点的度对相似性的影响就微乎其微,而当 $\alpha = 0$ 时,公式退化为 CN 方法。如果 α 较大,那么即使两个节点的度数都不是很大并且没有共同邻居,也会获得比较高的分数,相比之下共同邻居节点的个数对相似性的贡献则微乎其微。然而在社交网络中,显然具有共同邻居的两个节点之间产生连边会更有说服力,所以这里的 α 需要取一个很小的值。

4　实　验

4.1　实验平台

实验平台的硬件信息为:酷睿 i5 处理器,8 G 内存,系统版本为 Windows 7。

编程语言为 Matlab,编程环境为 Matlab2014b。

4.2　数据集

实验选取了 5 个具有不同功能的社交网络,包括朋友网络、信息网络和通信网络,介绍如下:

Wikivote:维基百科中的活跃用户可以被提名为管理员,当一个用户被提名时,维基百科会组织选举,获得支持最多的用户晋升为管理员。用户表示为节点,选举行为对应于网络中的边,如果用户 A 给用户 B 投票,那么就有一条边从 A 指向 B。数据集中包含 7 115 个节点和 103 689 条边。

Youtube:一个以视频分享为主的网站,用户可以在该网站上分享和下载视频,也可以通过朋友关系组成网络社区进行互动。网络节点和连边分别表示用户和朋友关系,其中包含 8 500 个节点和 77 007 条边。

Facebook:该网络包含了 Facebook 中的朋友列表,是通过其官方应用 Facebook. app 从调查参与者中收集到的,数据集通过将 Facebook 内部用户的 id 用一个新的值代替对真实网络进行了匿名化。其中包含 4 039个节点和 88 234 条边。

Email:数据集来自国外某公司员工的 email 通信记录。网络中每个节点对应一个电子邮箱地址,如果一个地址 i 发送给地址 j 至少一封电子邮件,网络中就会包含一条从 i 到 j 的无向边。不考虑公司内员工与公司外的电子邮箱地址之间的通信,其中包含 1 133 个节点和 10 903 条边。

Gplus:数据集包含 Google+中的“圈子”,Google+

是 Google 的一个扩展功能,其中心要点是朋友和熟人组成的“圈子”(Circles)。用户可以把联系人按不同的圈子分组,如家庭成员、同事、大学同学等,并在“圈子”里分享照片、视频及其他资讯。用户之间的每一次互动都会在两个用户对应的节点之间出现一条连边。其中包含 6 600 个节点和 189 167 条边。

4.3 实验结果与分析

实验中用 4 种常用链路预测方法和改进的方法 ImprovedCN 对五个网络数据集进行了预测,预测准确度使用 AUC 来衡量。AUC 从整体上衡量算法的准确度,它可以理解为测试集中的边的分数值比随机选取的一个不存在的边的分数值高的概率。实验结果见表 1。

表 1 链路预测方法的预测准确度(AUC)

预测方法	Wikivote	Facebook	Email	Youtube	Gplus
CN	0.939 9	0.993 2	0.852 8	0.924 2	0.962 5
ImprovedCN	0.963 7	0.985 5	0.903 5	0.961 6	0.979 1
AA	0.941	0.993 9	0.852 4	0.927 3	0.963 5
LP	0.970 2	0.993 9	0.919 4	0.959 7	0.983 2
Katz	0.969	0.993 3	0.923 1	0.960 9	0.984 5

从表 1 可以看出,改进后的方法在预测准确度上相比 CN 有了一定的提高,并且在 Wikivote、Facebook、Email 和 Gplus 四个网络上都优于 AA 方法,达到了与 LP 相当的预测水平,在 Facebook 网络中虽然预测准确度不如其他四种方法,但仍然取得了比较理想的效果。

从 ImprovedCN 的表现来看,它在一定程度上弥补了 CN、AA 方法在聚类系数和平均度都较低的社交网络中表现不佳的缺陷,使预测准确度不因网络结构的变化而表现出较大的差别。在 CN 的基础上增加优先连接的考虑确实取得了更好的效果。

从时间复杂度来分析,虽然 ImproveCN 在 CN 的基础上加入了节点度的乘积,但是后者可以通过邻接矩阵的列向量与自身的转置矩阵相乘计算,而 CN 的实现是邻接矩阵自身相乘求得,所以在时间复杂度上相比于 CN 没有增加。相比于 LP 和 Katz,在运行时间上有比较明显的优势。

5 结束语

研究了社交网络的结构特性,并根据研究结果对链路预测方法进行了改进,在共同邻居的基础上考虑了优先连接对节点对的相似性贡献。通过实验对比发现,改进方法的预测准确度有了一定的提高,并且在效果上达到了局部路径方法的水平,而且该方法在时间复杂度上没有增加,比较适合于如今大型的社交网络。

参考文献:

[1] 吴信东,李 毅,李 磊. 在线社交网络影响力分析[J]. 计算机学报,2014(4):735-752.

[2] Sarukkai R R. Link prediction and path analysis using Markov chains[J]. Computer Networks,2000,33(1-6):377-386.

[3] 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.

[4] Zhou T,Lü L,Zhang Y C. Predicting missing links via local information[J]. European Physical Journal B,2009,71(4):623-630.

[5] 刘大伟,吕元娜,余智华. 一种改进的复杂网络链路预测算法[J]. 小型微型计算机系统,2016,37(5):1071-1074.

[6] 傅颖斌,陈羽中. 基于链路预测的微博用户关系分析[J]. 计算机科学,2014,41(2):201-205.

[7] 许 进,杨 扬,蒋 飞,等. 社交网络结构特性分析及建模研究进展[J]. 中国科学院院刊,2015,30(2):216-228.

[8] Lorrain F,White H C. Structural equivalence of individuals in social networks[J]. Social Network,1971,1(1):67-98.

[9] Salton G,McGill M J. Inoduction to modern information retrieval[M]. Auckland:McGraw-Hill,1983.

[10] Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et du Jura[J]. Bulletin of the Torrey Botanical Club,1901,37:547.

[11] Leicht E A,Holme P,Newman M E. Vertex similarity in networks[J]. Physical Review E,2006,73(2):026120.

[12] Adamic L A,Adar E. Friends and neighbors on the web[J]. Social Networks,2003,25(3):211-230.

[13] Lü L,Jin C H,Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E,2009,80(4):046122.

[14] Katz L. A new status index derived from sociometric analysis[J]. Psychometrika,1953,18(1):39-43.

[15] Milgram S. The small world problem[J]. Psychology Today,1967,2(1):60-67.

[16] Backstrom L,Boldi P,Rosa M,et al. Four degrees of separation[C]//Proceedings of the 4th annual ACM web science conference. [s. l.]:ACM,2012.

[17] Mislove A,Marcon M,Gummadi K P,et al. Measurement and analysis of online social networks[C]//ACM SIGCOMM conference on internet measurement. San Diego,California,USA:ACM,2007.

[18] Watts D J,Strogatz S H. Collective dynamics of “small-world” networks[J]. Nature,1998,393(6684):440-442.

[19] Newman M E J,Was D J. Renormalization group analysis of the small-world network model[J]. Physics Letters A,1999,263(4-6):341-346.

[20] Kleinberg J. The small-world phenomenon: an algorithm perspective[C]//Proceedings of ACM symposium on theory of computing. [s. l.]:ACM,2010:163-170.

[21] Barabási A L,Albert R. Emergence of scaling in random networks[J]. Science,1999,286(5439):509-512.

[22] 王 林,商 超. 无标度网络中的链路预测问题研究[J]. 计算机工程,2012,38(3):67-70.