IPI: 一种基于影响力和兴趣的链接预测算法①

杨林瑞, 廖 倡

(复旦大学 计算机科学技术学院, 上海 201203)
(上海市数据科学重点实验室(复旦大学), 上海 201203)

摘 要: 链接预测的一个关键问题在于如何合理高效地结合链接属性、节点属性等相关信息以用于预测的目的, 针对该问题提出了一种基于节点影响力和兴趣的链接预测算法 IPI(Influence Plus Interest), 即通过拓扑结构信息来量化用户的影响力, 通过文本信息来模拟用户兴趣. 结合两类信息对节点间的联系进行打分, 得分高的节点对即代表具有较强的联系. 在真实数据集上的实验表明, 我们提出的方法具有一定的可行性.

关键词: 社会网络; 链接预测; 影响力度量; 兴趣模型

IPI: A Link Prediction Algorithm Based on Users' Influences and Interests

YANG Lin-Rui, LIAO Chang

(School of Computer Science and Technology, Fudan University, Shanghai 201203, China)
(Shanghai Key Laboratory of Data Science, Fudan University, Shanghai 201203, China)

Abstract: Currently one of core issues of link prediction is how to rationally and efficiently combine link attributes, node attributes and other relevant information for forecasting purposes. Aim at this problem, we propose a link prediction algorithm based on influence and interest, which mainly consists of quantizing the influence of nodes by topology structure information and simulating users' interests by text information. These two types of information are aggregated to give the relation score to the node pairs. High-scoring pair of nodes which represents a strong link. Experiments on real datasets show the method proposed in this paper is feasible.

Key words: social network; link prediction; influence measure; interest model

随着互联网的发展和社交网络的普及, 网络已经成为人们生活中不可缺少的一部分, 它是当今时代信息传播的最主要的媒介和载体. 在大数据时代的今天, 网络上的信息呈现出爆炸式的增长, 每分每秒都会有大量的信息在网络上传播与扩散, 互联网上庞大的用户、信息等可以构成很多不同的社会网络. 这些社会网络值得去研究和挖掘, 尤其是用户与用户之间的联系——链接关系, 就是目前社会网络分析的研究热点之一.

给定一个社会网络, 链接预测的任务, 即根据已知的、可观察到的节点联系, 来预测节点间的其他可能存在的联系. 链接预测对于理解复杂化网络的演化机制, 发现网络潜在的功能特性有着重要的意义, 所以链接预测是很多研究问题的基础, 得到了各领域广泛的关注.

目前大部分的链接预测算法都是基于所在网络的结构特征的. 根据网络的拓扑结构, 链接预测的方法可以分为以下三类: (1)基于全局信息, 如 Katz 指数[1], 随机游走相似性[2]等; (2)基于共同邻居, 如共同邻居指数[3]等; (3)基于路径, 如 SimRank[4]等.

传统方法很少考虑节点的影响力, 即网络中的节点更容易跟随高影响力节点的行为, 此即意见领袖[5]. 影响力度量主要可以分为四个方向: 基于网络拓扑结构的度量, 基于用户行为的度量, 基于用户交互信息的度量, 基于话题的度量; 影响力传播主要包括意见领袖发现问题和影响传播最大化问题. 在影响力传播中, PageRank 算法是一种基本模型, 为 Larry Page 和 Sergey Brin 于早期的搜索系统原型时提出的链接分析

① 收稿时间:2015-04-28;收到修改稿时间:2015-06-08

算法[6]. 其中网页质量的评估基于以下两个假设: (1)数量假设, 即如果一个页面节点接收到的其他网页指向的入链数量越多, 那么这个页面越重要; (2)质量假设, 即越是质量高的页面指向某个页面, 则该页面越重要.

另一方面, 在社会网络中很少综合地考虑用户的兴趣以用于链接预测的任务. 提取用户的兴趣一个很直接的思路体现在用户所属文本的相似性上. 关于文本特征方面的工作, 主要有 TF-IDF, LSA, PLSA, LDA 等[7]. 其中 LDA 模型可以把词项空间中的文本变换到主题空间. 由于主题的个数远远小于词项的个数, 因此 LDA 模型可以看作是一种降维方式. 目前 LDA 模型已经成为了主题建模中的一个标准, 通过对其进行改进, 挖掘当前用户文本中所关注的主题概率, 以此建立用户的兴趣向量, 可以很好地应用于社交网络的链接预测中. 通过建立用户的兴趣相似性得分, 作为用户链接预测分数的一个重要组成部分.

在本文中, 我们采用拓扑结构信息来模拟节点的影响力大小, 采用基于文本信息使用 LDA[8]模型来模拟节点的兴趣向量, 提出了一种综合网络节点影响力和兴趣信息的链接预测的新方法 IPI(Influence Plus Interest).

1 基于影响力和兴趣的链接预测算法

1.1 用户影响力建模

在社会网络中, 网络中的每个节点不是孤立的, 每个节点每时每刻都可能会受到其邻居节点的影响. 而且网络中的每个节点都具有各自的影响力, 影响力的传播在实质上就是消息的传播, 具有不同影响力的节点的消息传播深度和广度都会不同.

显然, 影响力高的节点更能影响与其有联系的其他节点, 并且影响会一层层地传递下去. 因此, 量化社会网络中每个节点的影响力, 找出节点对中具有高影响力的共同邻居, 对判断两个节点之间是否存在联系具有较高的借鉴意义.

1.1.1 影响力计算

例如, 给定一个微博影响力网络 $G=\{V,E,P\}$ 如下, 其中:

1) 每个微博用户被看作网络上的一个节点, V 表示所有微博用户的集合, 其中节点的权重 $vw\in P$ 可以表示一个用户的发微博数量;

2) E 是所有用户之间边的集合, 每条边表示两个用户所发的微博之间发生了消息传播, 我们可以理解为当用户 B 对用户 A 的微博进行了转发或者评论, 则 A 到 B 之间产生链接, 且每条边有权重 $ew\in \mathrm{E}$, 其中边的权重可以表示两个用户间微博的传播次数;

如图 1 所示, 节点 A 对节点 B 存在影响, 节点 B 对节点 E 存在影响, 我们可以认为节点 A 对节点 E 存在影响, $(\mathrm{A}\to B\to E)$ 即可表示为一条从 A 到 E 的影响传播路径, 且路径长度为 2.

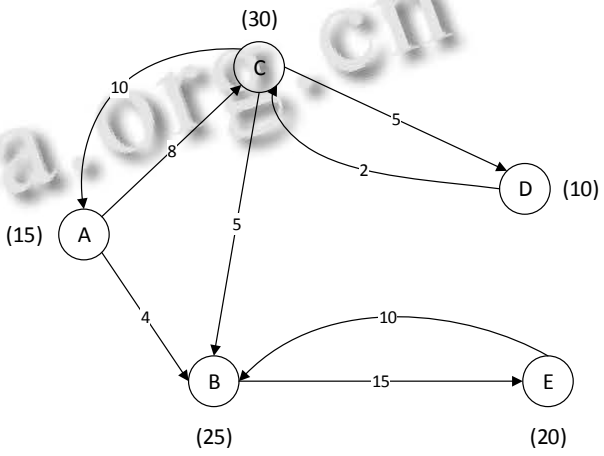

图 1 影响力网络示例

基于以上几点, 本文认为节点间的影响力计算方式如算法 1 所示:

算法 1. 影响力量化算法 IQ(Influence Quantification)

输入: 影响力网络 $G=\{V,E,P\}$

输出: 影响力大小集合 I

1)为所有节点赋予初始权重;

2)计算每个节点的影响力:

3)依次计算传播路径长度为 1,…, T 的节点间的传播强度;

4)基于所有的传播路径长度, 综合该节点到所有节点间的影响力;

5)给定用户的影响力大小得分, 即如式(1)所示:

$$\mathrm{influence}(z)=\sum_{t=1}^{T}\sum_{j=1}^{|V|}E^{t}(z,j)\cdot P(z) \quad (1)$$

其中, $P(z)$ 表示节点 z 的权重, E 为整个网络经过归一化后的邻接矩阵, 若在网络中, 节点 z 可以直接到达节点 j, 则在一阶矩阵 E 中, $E(z,j)$ 可以表示为节点 z 到节点 j 的距离, 这个值经过归一化后为一个介于 0 到 1 之间的值.

需要注意的是, 节点属性 P 表示用户在社会网络

中的活跃程度, 而链接属性 E 表示用户在社会网络中的受欢迎程度, 本文认为, 影响力应该满足以上两个因素, 需要融合二者的信息.

矩阵 E^t 表示矩阵 E 经过矩阵运算后得到的 t 阶矩阵, $E^t(z,j)$ 表示节点 z 经过 t 次传播到达节点 j 后的影响强度, 由于矩阵经过归一化, 所以随着传播层次的加深, $E^t(z,j)$ 的值逐减小趋向于 0.

为避免因弱化的间接影响使得计算复杂, 最大传播路径长度需要设置一个最大值, 若 T 的值取值过大, 可能会使后面的链接预测结果变差.

1.1.2 基于影响力的相似性

传统的链接预测模型的思路多为基于共同邻居数目, 如公式(2)所示:

$$S_{xy}=\frac{|\Gamma(x)\cap\Gamma(y)|}{\max\{|\Gamma(x)|,|\Gamma(y)|\}} \tag{2}$$

其中, $\Gamma(t)$ 表示节点 t 的邻居节点集合. 该模型基于如下考虑: 当两个节点之间的共同邻居越多, 则它们存在链接关系的可能性越大.

同样我们也基于以上的考虑, 引入影响力网络这个概念, 并在上文中给网络中的每个节点计算出了一个影响力大小. 基于这个影响力, 我们将传统的只考虑共同邻居数目的链接预测模型, 扩展为基于影响力的链接预测.

设节点z是节点x和节点y的共同邻居, *influence*(z)为计算出的节点 z 的影响力大小, 则从网络拓扑结构角度来考虑, 节点x和节点y之间的相似性 S_{xy}^R 可以定义为如公式(3)所示:

$$S_{xy}^R=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\text{influence}(z) \tag{3}$$

其中, 将节点x和节点y的所有共同邻居的影响力大小相加, 将得到的值作为二者的相似性大小的基础之一. 当相加之和越大, 即二者具有很多高影响力的共同邻居时, 二者越可能存在链接关系.

1.2 用户兴趣建模

兴趣可以反映用户的行为, 而行为分析在链接预测中同样起着重要的作用. 通常, 我们认为用户联系关系的产生来源于相同的行为轨迹. 例如, 在微博文本数据中, 通过对文本信息进行处理, 可以得到用户的兴趣向量, 用以表示用户的行为特征.

不同的微博可以抽象成多个相同或不同的主题, 而不同的用户可能拥有共同的兴趣, 通过不同用户之间发布的微博文本内容, 来得到他们各自的兴趣, 从而计算他们的相似性.

1.2.1 兴趣向量计算

LDA 是一个完备的主题模型, 我们可以依据 LDA 的方法来模拟节点所感兴趣的主题. 根据文本的生成规则和已知数据, LDA 通过概率推导可以得到文本的主题结构. 通过 LDA 得到每个用户的兴趣分布的算法 2 所示:

算法 2. 兴趣抽取算法 IE(Interest Extraction)

输入: 微博文本数据 D

输出: 兴趣向量集合 θ

1)对用户 U_i 所有的文本数据进行分词;

2)为每一个词的主题分布进行初始赋值;

3)通过 MCMC 进行反复抽样, 得到每个词在各个主题之间出现的次数;

4)根据式 4 求出用户的兴趣(主题)分布.

$$\theta_{ij}=\frac{C_{ij}^{MK}+\beta}{\sum_{k=1}^{K}C_{ik}^{MK}+K\alpha} \tag{4}$$

其中, C_{ij}^{MK} 表示在重复实验中, 主题 T_j 在用户 U_i 所有主题中出现的次数[9].

1.2.2 基于兴趣的相似性

给定 R 维欧式空间中的节点 x 和节点 y 的主题兴趣向量 θ_1 和 θ_2, 则从二者的兴趣相似性出发, S_{xy}^I 定义如公式 5 所示. S_{xy}^I 表示节点 x 和节点 y 之间基于文本内容的兴趣相似性大小.

$$S_{xy}^I=\frac{\vec{\theta_1}\cdot\vec{\theta_2}}{|\vec{\theta_1}||\vec{\theta_2}|} \tag{5}$$

其中, 将节点x和节点y的兴趣向量的余弦相似度作为二者的相似性大小的基础之一, 当二者的兴趣分布类似, 即使得概率相近时, 二者越可能存在链接关系.

1.3 结合影响力和兴趣的链接预测方法

以上的部分为分别从两个方面来对社会网络进行建模, 即基于影响力来建模和基于兴趣信息来建模, 并分别从这个两个模型中得出了计算节点相似性的基础部分. 这两个方面从不同角度来量化潜在的链接关系可能性大小. 链接关系的产生需要同时受用户所在网络和自身兴趣两方面的影响. 具体而言:

1)如果两个节点的共同的邻居节点的影响力越大, 则二者存在链接关系的可能性就越大;

2)如果两个节点的兴趣范围越相似, 则二者存在链接关系的可能性就越大.

综合影响力和兴趣相似性, 链接预测可以扩展为如公式(6)所示, 其中 α 和 $1-\alpha$ 分别代表两者在这个综合模型中所占的权重. 当 α 取值不同时, 会影响最终的预测结果的准确率.

$$score(x,y)=\alpha * S_{xy}^{R}+(1-\alpha) * S_{xy}^{I} \quad (6)$$

2 实验与分析

2.1 实验描述

本文实验使用真实数据集, 数据集来源为在 2011 年 1 月到 2011 年 5 月的 Plurk micro-blog 数据, 其数据从[10]下载得到.

Plurk, 中文名称为噗浪, 是一个提供基于时间轴的可视化微博客服务的的社交网站. 该数据集包含用户个人信息(包括用户 id、用户昵称)、用户之间的好友关系以及这些用户在该段时间内的发布 Plurk 内容及回复情况.

本文算法全部采用 Java 实现, 实验平台是 Intel(R) Celeron(R) CPU E3200 @ 2.40 GHz, 2.00GB RAM, 32 位 Windows 操作系统的计算机.

2.2 评价指标

衡量链接预测算法的精确度的指标有 AUC 和 precision[11]. AUC 从整体上来衡量算法的精确度; 而 precision 只考虑排在靠前的 N 个链接的预测结果.

AUC 的值大于 0.5 的程度可以衡量出算法比随机选择的方法更加精确的程度[12]. AUC 可以定义为如式(7)所示:

$$AUC=\frac{n^{'}+0.5n^{''}}{n} \quad (7)$$

而 Precision 表示排在靠前的 L 个待预测边中被预测准确的比例. 如果排在前 L 的边中有 L_r 个出现在测试集中, 则 Precision 可以定义为:

$$precision=\frac{L_r}{L} \quad (8)$$

2.3 实验结果分析

我们将基于影响力的链接预测方法、基于兴趣的链接预测方法和融合二者的链接预测方法应用于 Plurk 数据集进行实验验证, 同时将基于共同邻居数的方法作为基准方法进行对比, 其结果如下图 2 和图 3 所示.

从中我们可以看出: 通过结合节点影响力网络和兴趣相似性网络, 我们相对于比较方法获得了一定的准确度上的提升. 同时可以看出, 综合了影响力和兴趣相似度的预测模型在 AUC 和 Precision 上都获得了最好的效果.

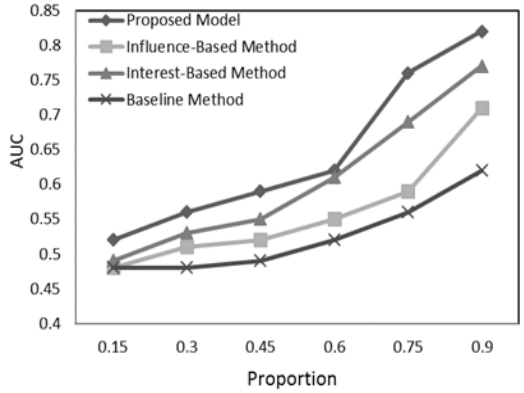

图 2 不同比例的训练集在算法中的 AUC 效果

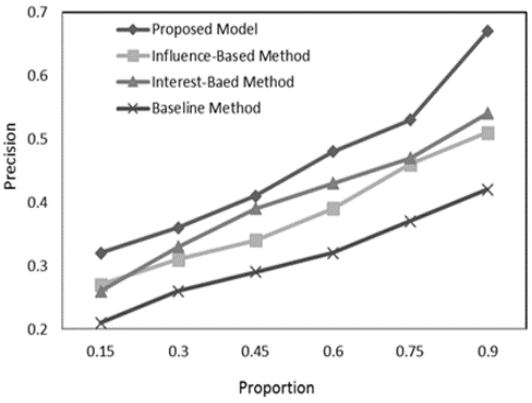

图 3 不同比例训练集在算法中的 Precision 效果

在 1.1 节提出的影响力量化算法 IQ 中, 计算节点影响力需要设置影响力的传播路径长度 *T*, 这个参数的大小会对影响力大小的计算产生影响, 从而也会影响模型最终的预测效果.

关于参数传播层数的敏感性问题的实验结果, 如图 4 所示. 我们可以看出, 传播层数为 2 时, 准确率最高, 即证明间接影响在网络用户中的有效性.

在 1.3 节提出的影响力和兴趣的综合模型中, 需要设置它们在该模型中的权重 α 和 $1-\alpha$, 而 α 这个参数的大小会也会影响模型最终的预测效果.

关于参数权重因子 α 的敏感性问题, 如图 5 所示. 我们可以看出, 当 α =0.4 左右时, 效果达到最优, 这表明了综合考虑用户兴趣信息和影响力信息

在链接预测中的必要性.

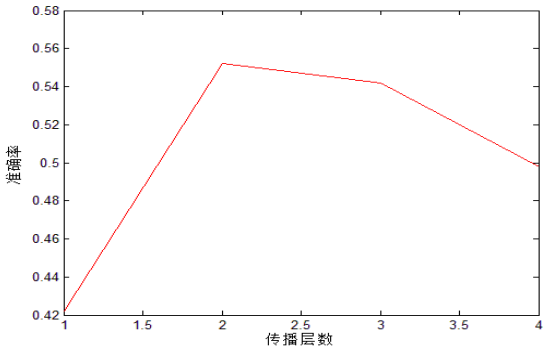

图 4 测试数据在不同的传播层数下的准确率表现

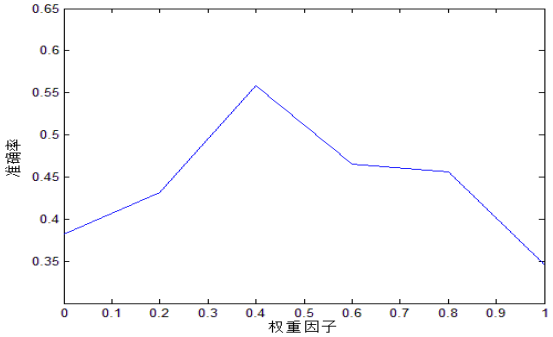

图 5 测试数据在不同的权重因子下的准确率表现

3 结语

本文研究和讨论了社会网络中的链接预测问题. 已有的大部分链接预测算法缺乏同时综合考虑文本信息和用户间的拓扑结构信息, 导致预测效率较低. 本文提出了一种综合用户的影响力建模和兴趣建模的链接预测算法 IPI, 实验表明, 本文算法的准确率较已有的算法有一定的提高.

参考文献

1 Katz L. A new status index derived from sociometric analysis. Psychometrika, 1953, 18(1): 39–43.

2 Backstrom L, Leskovec J. Supervised random walks: Predicting and recommending links in social networks. Proc. of the Fourth ACM International Conference on Web Search and Data Mining. ACM. 2011. 635–644.

3 Lv L, Zhou T. Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150–1170.

4 Jeh G, Widom J. SimRank: A measure of structural-context similarity. Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2002. 538–543.

5 Cha M, Haddadi H, Benevenuto F, Gummadi P K. Measuring user influence in twitter: the million follower fallacy. ICWSM, 2010, 10(10-17): 30.

6 Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. 1999.

7 Aggarwal CC, Zhai CX. Mining text data. Springer Science & Business Media, 2012.

8 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993–1022.

9 Liu Q, Chen E, Xiong H, Ding, CH, Chen J. Enhancing collaborative filtering by user interest expansion via personalized ranking. IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42(1): 218–233.

10 Kuo TT, Hung SC, Lin WS, Peng N, Lin SD, Lin WF. Exploiting latent information to predict diffusions of novel topics on social networks. Proc. of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers. Association for Computational Linguistics, 2012, 2: 344–348.

11 Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proc. of the 23rd International Conference on Machine Learning. ACM. 2006. 233–240.

12 吕琳媛.复杂网络链路预测.电子科技大学学报,2010, 39(5):652.