

基于多元属性的社团挖掘在意见领袖发现中的应用

申艳光, 李金伟

(河北工程大学 信息与电气工程学院, 河北 邯郸 056038)

摘 要: 意见领袖是网络舆论传播中的关键, 在舆情分析中有重要意义. 采用无向图建模, 将链接两端节点的 PageRank 值与相似度, 作为链接的双重属性, 通过二次判别式对链接进行主次划分, 将整个网络划分为核心与非核心两类用户, 并将网络中的大社团进行细化, 结合细化后的社团跟核心用户的划分结果找出其中相应的子社团中的意见领袖, 提出了一种基于二元分类的意见领袖发现算法.

关键词: 数据挖掘; 社团挖掘; 意见领袖识别; 机器学习; 二次判别式

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2014)07-0129-04

Community Identification Based on Multivariate Properties and Utilizing It to Identify Network Opinion Leader

SHEN Yan-guang, LI Jin-wei

(College of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China)

Abstract: Opinion leader is the key point in the process of network public opinion spreading, has important significance in the analysis of public opinion. This article uses undirected graph for modeling, Take the value of PageRank and similarity into account, as the attribute of the link, and uses Quadratic Discriminant to classify the model of links. We can find the opinion leader of the inner community.

Key words: data mining; community identification; opinion leader identification; machine learning; quadratic discriminant

1 引言

意见领袖 (Opinion Leader) 最早在文献[1]中提出, 是指在人际传播网络中经常为他人提供信息、意见、品论, 对他人施加影响的活跃分子^[1]. 网络意见领袖是指在网络社区中比较活跃, 对该社区中的网民具有影响力, 在特定时空条件下能够制造和引领社会舆论的一类人^[2]. 意见领袖在传播和引导舆论等方面具有十分重大的意义, 通过对意见领袖的有效识别, 可以使政府更加有效地提高对突发性公共事件舆情管理水平, 可以更加方便地引导网络舆情朝积极、健康的方向发展.

目前国内外学者对网络意见领袖的识别方法的研究主要采用统计测量方法、PageRank 权重识别

方法和文本倾向性识别方法等. 有的方法具有很强的理论意义, 但是在实际应用中, 效率往往不尽如人意. 另外, 大多方法只是将某一个大类作为一个整体进行分析, 并没有将其内部的社团细化, 进而寻找细化后内部子社团的意见领袖.

在大型在线社区中帖子数量是极其庞大的, 进行文本分析帖子的倾向性将会极大的影响意见领袖发现算法的效率. 因为采用无向图建模, 赋予图中链接双重属性, 通过二次判别式对链接进行主次划分, 将整个网络划分为核心与非核心两类用户, 同时将网络中的大社团进行细化, 并给出相应的细化后子社团中的意见领袖.

收稿日期: 2013-09-05; 修回日期: 2013-10-31

基金项目: 国家自然科学基金资助项目(61075053)

2 核心链接

Ahn 等^[3]提出了以“链接”间“相似度”作为每个“链接”的特征对整个网络进行划分,使用层次聚类建立聚类树状图.由于每个节点拥有一条到多条“链接”,每条“链接”存在于一个社团中,因此对该树状图不同层次的切割得到的社团结构对应着不同层次的重叠社团.所谓的“相似度”指的就是两个节点间所具有的共同节点的一个量化值.利用“相似度”这个量化值可以很好的刻画出被某个“链接”所连接的两个节点是否同时归属于某个网络子社团.

但是,“相似度”高,只能说明这两个节点同时归属于某一个社团,并不能说明这个“链接”所连接的两个节点在这个网络子社团中的地位,如图 1 所示.在图 1 中连接 1、2 两个节点的链接同连接 9、10 两个节点的链接相似度都比较高,但是节点 1、2 的 PageRank 值高于 9、10 两个节点,所以认为 1、2 两个节点之间的链接更具有研究价值.

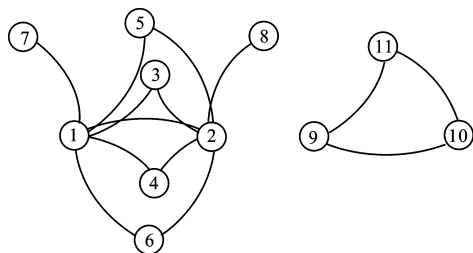


图 1 链接重要性示意图

在本文中只考虑无向图中 PageRank 权重和相似度,每条链接的 PageRank 权重等于这条边所连接的两个顶点的 PageRank 值的均值.

PageRank 算法是有向图中的一种迭代算法,网络图中的节点不断地积累 PageRank 值,并把部分传递给他“邻居”.虽然 PageRank 算法最初是被设计用来计算有向网络的,但是通过引入 Newman 权重,可以将该算法应用在无向网络中^[4].如式(1)所示,其中 ω 就是 Newman 权重.

$$PRW_i = \beta \sum_{j \neq i} \frac{a_{ij} \omega_{ij}}{\sum_{u \neq j} a_{uj} \omega_{uj}} PR_j + (1 - \beta) \frac{1}{|N|} \quad (1)$$

3 二次判别式

根据贝叶斯规则,可以把类 C_i (包括“核心链接”和“非核心链接”两类所以 $i = 1, 2$) 的后验概率写为式(2):

$$P(C_i | x) = \frac{P(x | C_i) P(C_i)}{P(x)} \quad (2)$$

高斯分布密度函数是许多自然现象的模型,因为大多数的样本都可以看作是简单高斯分布的数学期望轻微变动的版本,因此,假设核心链接和非核心链接的先验概率分布为高斯分布.

通过假设先验概率分布为高斯分布,将式(2)进行推导得到式(3).

$$g_i(x) = \frac{1}{2} \log(2\pi) - \frac{1}{2} \log |\sum i| - \frac{1}{2} (x - \mu_i)^T \sum i^{-1} (x - \mu_i) + \log P(C_i) \quad (3)$$

$$\sum i = \frac{\sum_{t=1}^N (x^t - \mu_i)(x^t - \mu_i)^T}{N} \quad (4)$$

式中, $\sum i$ 是第 i 类两个属性的协方差矩阵, μ_i 是第 i 类两个属性的均值向量,在本文中因为分为“核心链接”和“非核心链接”两类,所以 i 只有两种值可取, $i = 1, 2$. 式(4)是 $\sum i$ 的计算公式.

4 意见领袖发现算法

基于二元分类的意见领袖发现算法步骤如下:

(1)以“链接”为对象,进行数据建模.从数据集中随机抽取部分样本,获得相关参数,包括样本协方差矩阵、均值等.

(2)使用贪婪策略进行聚类,将得到的核心链接进行划分(也就是各个子社团中的核心成员的一个划分).

(3)继续使用贪心策略将非核心链接加入到各个核心成员所组成的子社团中.按照以下规则进行聚类:

①如果一条边的节点不与已经划分好的集合中的任何节点重合,则新建一个集合并将该边加入.

②如果一条边的节点与已经划分好的其中一个集合的节点相重合,则将该条边加入该集合.

③如果一条边的节点同时与多个已经划分好的集合中的节点相重合,则将该边加入到与该边联系最多的集合中去.

5 算法的仿真与分析

5.1 算法仿真

这里采用的数据集是 Zachary 空手道俱乐部,该数据集共含有 78 条链接.首先,根据式(2)计算出各个“链接”的 PageRank 值,还有“重叠度”,在这里并没有对重叠度进行量化.得到的初始化数据包括:

“核心链接”的均值向量为:[10,2.073 2],出现概率为:0.525 6。“非核心链接”的均值向量为:[5.297 3,1.351 4],出现概率为:0.474 4。如表 1 所示,左边是“核心链接”的协方差矩阵,右边是“非核心链接”的协方差矩阵。

表 1 协方差矩阵

核心链接协方差矩阵		非核心链接协方差矩阵	
1.762 5	1.537 5	2.450 8	0.489 9
1.537 5	3.869 5	0.489 9	1.067 6

根据前面所到的数据代入式(3)、(4)进行计算分别得到核心链接和非核心链接的判别式函数,得到判别函数概率分布如图 2 所示。

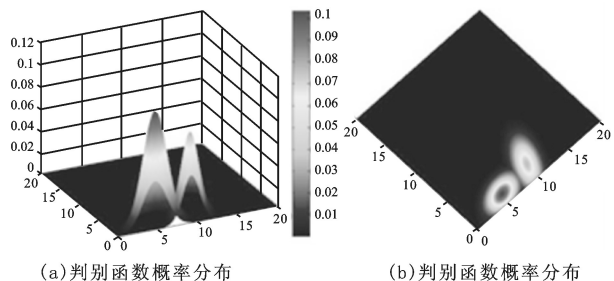


图 2 判别函数概率分布图与判别函数等概率分布图

根据得到的函数判别式,对原始数据集中的骨干链接进行基于贪婪政策的聚类得到子社团的骨干成员,如图 3 所示,本算法能够将一个大的网络社团进一步细化成多个内部社团,并且找出细化后社团相应的意见领袖。在图中深色和浅色两种颜色表示将这个空手道俱乐部成员划分为两个子社团,最大圆圈表示的点就是各个子社团中的主要成员(即子社团中的意见领袖)。节点之间的链接有粗、中、细三种连线构成,粗的连线代表了在函数判别式中距离判别式函数中心越近的连接也就是社团关系中比较重要的关系,中的连线代表的是次重要的联系,而细线代表的就是非核心链接。

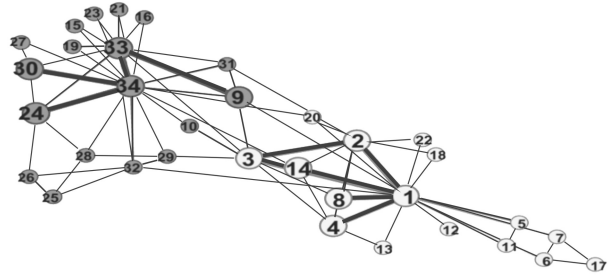


图 3 仿真实验结果示意图

5.2 算法分析

本算法主要性能指标主要包括时间复杂度和准

确度两个性能指标。

5.2.1 算法时间度分析

本算法的时间复杂度主要包括 PageRank 权值计算、相似度计算和聚类三个方面。

(1)PageRank 权值计算:依据 Atish Das Sarma 等^[5]关于快速分布式 PageRank 计算方法,得到在无向图中计算 PageRank 值的最优时间复杂度是 $O(\frac{\sqrt{\log n}}{\epsilon})$,其中 n 是节点的个数, ϵ 是重启概率, $1-\epsilon$ 也叫做阻尼系数。

(2)相似度计算:假设数据集中,节点度的平均值为 k ,共有 m 条边,因此得到时间复杂度为 $O(m * k)$ 。

(3)聚类:这里采取前面第 3 节中提到的聚类策略,因此所需要的时间复杂度为 $O(m)$, m 为数据集中边的数量。

综上,本算法的时间复杂度为: $O(\frac{\sqrt{\log n}}{\epsilon}) + O(m * k) + O(m)$ 其中 n 为节点个数, m 为边的数量, k 为节点的度的平均值。

在文献[6]中提出使用关键节点算法,每次计算单个用户平均路径的时间复杂度是 $O((n-1)^3)$, n 个用户的平均路径变化的时间复杂度就是 $O(n^4)$ 。文献[7]提出了 GlobalPR 算法,其时间复杂度也达到了 $O(n^{(k+1)}n^2)$,其中 n 是节点个数, k 是迭代次数。上述两种算法随着数据集规模的增大收敛速度将会变慢,复杂度高于本文提出的算法。

5.2.2 算法准确度分析

关于意见领袖识别准确度的评价目前尚无统一标准,文献[8]提出了采用前 N 个结果中的正确率 $P@N$:

$$P@N = \frac{\text{前 } N \text{ 个结果中人工判定为真正意见领袖的个数}}{N}$$

前 N 个结果中的平均正确率为 AVGP@N:

$$AVGP@N = \sum_{i=1}^N P@i / N$$

本文采用以上方法对正确率进行统计,根据链接所获得的判别式函数的值进行排序,以确定前 N 条链接所对应的前 $2N$ 个点的正确率。表 2 中记录了平均准确率。

表 2 前 N 个结果的平均准确率, $N=5,10,15$

AVGP@5	AVGP@10	AVGP@15
80%	80%	70%

综上所述,本算法不但能够比较全面地找出意见领袖,而且能够将网络社团进一步细化,给出细化后各子社团的划分及其意见领袖,并且错误率也比较低,且具有较强的实用性.

6 结束语

意见领袖的筛选对于网络舆情的分析与监控具有重要的现实意义.我们通过抽取样本建立模型将网络中的链接划分为核心链接和非核心链接,迅速将网络的骨干进行划分,并依据划分后的网络骨干将网络进一步细化成多个子社团,给出了子社团的划分及其意见领袖.但是本方法在大型网络社区中进行意见领袖挖掘的效果还有待于完善,因此如何提高大型网络社区中的意见领袖挖掘方法的时间复杂度和空间复杂度,将是以后研究的重点方向.

参考文献:

- [1] Lazarafeld P F, Berelso B, Gaudet H. The people's choice: how the voter makes up his mind in a presidential campaign[M]. New York: Columbia University Press, 1948.
- [2] 倪邦文. 中国网络青年意见领袖的构成、特征及作用

[J]. 中国青年研究, 2011(9):05-09.

- [3] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks[J]. Nature, 2010, 466(7307):761-764.
- [4] Grolmusz V. A note on the page rank of undirected graphs[J]. CoRR, 2012(2):1205-1260.
- [5] Das Sarma A, Rahaman Molla A, Pandurangan G, et al. Fast Distributed PageRank Computation[J]. Internet Mathematics, 2012, 2(3):251-273.
- [6] 高俊波, 杨静. 在线论坛中的意见领袖分析[J]. 电子科技大学学报, 2007, 36(6):1249-1252.
- [7] Zhai Z W, Xu H, Jia P F. Identifying opinion leaders in bbs[C]//Proc of WI-IAT'08, Sydney, Australia: IEEE, 2008:398-400
- [8] 樊兴华, 赵静, 方滨兴, 等. 影响力扩散概率模型及其用于意见领袖发现研究[J]. 计算机学报, 2013, 36(2): 360-367.

作者简介:

申艳光 女, (1970-), 教授. 研究方向为数据挖掘、隐私保护.
李金伟 男, (1983-), 硕士研究生. 研究方向为数据挖掘、社团挖掘.

(上接第 128 页)

- [6] Dong Bo, Qiu Jie, Zheng Qinghua, et al. A novel approach to improving the efficiency of storing and accessing small files on hadoop: A case study by power point files [C]//International Conference on Services Computing, Miami, FL: IEEE, 2010:65-72.
- [7] 郭琦娟, 陈通照. 一种动态更新索引结构的设计与实现[J]. 计算机系统应用, 2008(12):76-69.
- [8] 史长琼, 唐铭, 张大方, 等. 基于 Trie 树的哈希表[J]. 计算机应用, 2010, 30(8):2193-2196.
- [9] Liu Xuhui, Han Jizhong, Zhong Yunqin, et al. Implementing WebGIS on hadoop: A case study of improving small file I/O performance on HDFS[C]// Proc of the 2009 IEEE Conf on Cluster Computing and Workshops, New Orleans, LA: IEEE, 2009:1-8.
- [10] 泰冬雪. 基于 Hadoop 的海量小文件处理方法的研究[D]. 沈阳: 辽宁大学, 2011.
- [11] 余思, 桂小林, 黄汝维, 等. 一种提高云存储中小文件件

存储效率的方法[J]. 西安交通大学学报, 2011, 45(6): 59-63.

- [12] 刘琨, 肖琳, 赵海燕. Hadoop 中云数据负载均衡算法的研究及优化[J]. 微电子学与计算机, 2012, 29(9): 18-22.

作者简介:



樊超 男, (1989-), 硕士. 研究方向为网络与信息安全技术, Hadoop 云计算.

凌捷 男, (1964-), 博士, 副院长. 研究方向为网络与信息安全技术.