

微博社区交流网络结构的实证分析

王晓光^{1,2} 袁毅¹ 滕思琦¹

(1. 华东师范大学信息学系 上海 200241;
2. 连云港师范高等专科学校计算机系 连云港 222006)

摘要 微博客是继博客之后迅速发展起来的一种新的网络社区平台。以“Myspace9911”网站为研究样本,通过核心-边缘分析和聚类分析,探讨微博社区用户交流网络结构,界定核心与边缘区域,描述聚类群组,发现核心区域和聚类群组的联系,为完善微博社区信息交流系统提供借鉴。

关键词 微博客 社会网络分析 核心-边缘分析 聚类分析

中图分类号 G350 **文献标识码** A **文章编号** 1002-1965(2011)02-0199-04

Empirical Analysis on Communicating Structure of Micro-blog Community

WANG Xiaoguang^{1,2} YUAN Yi¹ TENG Siqi¹

(1. Department of Information Science, East China Normal University, Shanghai 200241;
2. Department of Computer, Lianyungang Teacher's College, Lianyungang 222006)

Abstract Micro-blog is a new social network platform which has developed rapidly after blog. Based on the study of Myspace9911, this paper tries to research into the communicating structure displayed by its users by means of both core-periphery analysis and cluster analysis, define the core area and periphery area, describe cluster groups, and find the relationship between the core area and cluster groups, and offers ideas to improve the information exchange system of the micro-blog community.

Key words micro-blog social network analysis core-periphery analysis cluster analysis

0 引言

微博客是一种非正式的迷你型博客。在维基百科中,微博客被描述为“一种允许用户及时更新简短文本(通常少于200字)并公开发布的博客的形式,允许任何人阅读或者只能由用户选择的群组阅读”^[1]。微博社区用户之间的交流基于一种“关注与被关注”的跟随机制,即用户可随时“关注”他人,成为他人的“粉丝”,其他用户也可“关注”自己,成为自己的“粉丝”,此过程为双向可逆过程。微博用户以最简单的“关注”方式进行信息交流,形成一个个大小不一的交流网络^[2]。

目前,国内关于微博客的研究更多集中在与各种行业结合应用的探析,如文献[3]探讨了图书馆利用微博客开展互动的读者服务、信息传播、舆情监控、学

术交流等活动的可行性;文献[4]基于对微博客的内涵与特征的理解,探讨微博客在教育中的应用前景;文献[5]认为极具竞争力的传播速度使得微博客在传播突发新闻方面具有其他媒体无法比拟的先天优势。

虽然有个别文章对微博客的用户行为特征和关系特征进行了实证分析^[2],但总体来说,有关微博客的实证研究还非常少,尤其对于由微博用户形成的交流网络还没有深入研究与实证分析。本文针对以上研究背景和不足,从实证的角度,分别根据核心-边缘理论和聚类分析方法,界定微博社区中核心区域与外围区域,描述聚类群组结构,分析群组间成员彼此关系,并探求核心结点与聚类群组结点之间的异同关系,以期对交流网络的结构和特点进行把握的同时促进社区用户之间信息沟通和交流。

收稿日期:2010-07-26 修回日期:2010-09-23
基金项目:本研究受2008国家社会科学基金项目“网络社区信息运动模式研究”资助(编号:08BTQ029)。
作者简介:王晓光(1980-),男,硕士研究生,讲师,研究方向为信息分析与计量;袁毅(1963-),女,教授,研究方向为信息咨询、信息分析、网络计量;滕思琦(1982-),男,硕士研究生,研究方向为信息分析与计量。

1 理论与方法

1.1 核心-边缘理论 核心-边缘理论是由 J. R. Fridemna 在其学术著作《区域发展政策》一书中正式提出的。Stephen P. Borgatti 和 Martin G. Everett 从 3 个角度归纳了核心-边缘结构的特点^[6]。

a. 从子群或派系的角度, 核心-边缘结构是一种不能分割为多个互斥子群的网络结构(尽管有些行动者的联系比另外一些更为紧密)。换句话说, 核心-边缘结构可以比喻成一个簇, 每个行动者都或多或少地属于这个簇内。

b. 从分块的角度, 核心-边缘网络可分成两块, 核心可视为 1-块, 边缘可视为 0-块。而核心与边缘的关系可以是 1-块, 也可以是 0-块。

c. 从可视化的角度, 点群在欧式距离空间中呈现物理上的核心-边缘结构。处于核心的点之间的联系紧密, 而处于边缘的点之间彼此的联系比较稀疏, 且均有与核心点建立关系的倾向。

目前, “核心-边缘”理论的研究逐渐向横向和纵向两方面深化, 涉及到多学科门类, 主要应用于地方产业群规划、旅游区及其规划等方面^[7-8], 随着 web 2.0 技术的发展, 网络社区的兴起, “核心-边缘”理论的研究也从地理空间范围拓展到了虚拟网络空间, 如张玥和朱庆华用核心-边缘分析方法探析了图书情报领域博客的网络结构^[9]。

1.2 聚类分析法 聚类分析又称群分析, 是研究分类问题的一种多元统计方法, 即将待处理的对象分配到相应的聚类中, 使得同一聚类中的对象差别较小, 而不同聚类之间的对象差别较大。目前聚类分析广泛应用于数据挖掘、web 服务等领域^[10-11]。

2 数据与预处理

本文研究样本来自“Myspace9911”网站, 该网站是由 Myspace 公司开发的国内知名的微博社区。因为选取名人用户进行研究更具有代表性, 而名人中尤以传媒记者更善于对信息进行加工与传播, 媒体人间的沟通与交流更为密切, 更易于从中探究交流结构的特性, 所以抓取“传媒记者”板块的 54 位名人用户资料, 工具为通用爬虫“火车头采集器”^[12], 抓取时间为 2010 年 5 月 22 日 15 点至 18 点, 具体抓取和处理过程如下:

a. 提取表征用户基本属性的数据资料, 包括他关注的人数、关注他的人数、微博数、收藏数四项数据, 再根据注册时间和微博数计算出使用天数和平均每日发博数两项数据, 上述六项数据组成一条记录, 共得到 54 条记录。

b. 提取每位用户的关注对象名单, 共得到 8 257 条数据, 以用户为单位进行保存。

c. 为了在网络分析时更加高效, 对用户名称进行顺序编码, 表 1 显示部分编号及对应的用户名称。

d. 构建一个关注网络矩阵, 其中微博用户是网络的点, 关注与被关注关系是边。由于关注是一种单向关系, 因此矩阵的行代表关注者, 列代表被关注者, 边的方向是从关注者指向被关注者。将关注对象名单导入到 VFP 环境中, 通过编程实现以下赋值: 54 位用户中, 如果某位用户的关注对象中出现另一用户, 则对应行列的元素值取 1, 否则元素值取 0, 最终得到一个关注网络二值矩阵, 部分数据如表 2 所示。

表 1 微博用户编码(部分)

编号	用户名称	编号	用户名称
17	关军	25	刘颖 GTO
18	郭小寒	26	刘铮
19	郭志凯	27	流氓柚子
20	黑皓普	28	卢世伟
21	贾维	29	咪咩
22	蓝蝴蝶	30	平客
23	梁纯-1626	31	朴九月
24	刘博	32	乔小刀

表 2 关注网络矩阵(部分)

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
18	0	0	0	1	1	1	0	0	1	0	1	1	0	0	1	1
19	0	1	0	1	1	1	1	0	1	0	0	1	0	1	0	1
20	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
21	1	1	0	1	0	0	1	0	1	1	1	1	1	0	1	0
22	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1
23	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0
24	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0
25	0	1	0	1	1	0	1	0	1	0	0	1	1	1	0	1
26	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
27	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0
28	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
29	0	1	0	0	1	0	0	0	1	1	0	0	0	0	0	0
30	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0
32	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0

3 分析工具

本文使用 Ucinet 进行核心-边缘分析, Ucinet 是一个用来处理社会网络数据的软件包, 具有核心-网络分析、中心性分析、子群分析、角色分析等功能; 使用 SPSS 进行聚类分析。

4 分析与结果

4.1 核心-边缘分析 在 Ucinet 中进行核心-边缘分析结果如下:

Starting fitness:0.378

Final fitness:0.378

Core/Periphery Class Memberships:

- 1:1 2 3 7 9 10 11 12 14 18 19 20 21 22 23 24 25 26 27 28 29 30
31 32 35 37 43 44 45
- 2:4 5 6 8 13 15 16 17 33 34 36 38 39 40 41 42 46 47 48 49 50 51
52 53 54

结果中给出了初始矩阵与理想矩阵的相关系数(Starting fitness)和经过重排后的矩阵与理想矩阵的相关系数(Final fitness)。Final fitness 的数值越大,表明实际数据与理想模型越相似,实际数据的核心-边缘结构模型越显著,本文结果的相关系数为 0.378。核心区域包含了 29 个结点,边缘部分的结点数为 25 个,将此结果通过 Ucinet 中可视化模块 Netdraw 绘制出核心-边缘结构图,如图 1 所示。

由以上结果可以看出,交流网络中处于核心区域的结点数量和处于边缘区域的结点数量比较接近,核心结点之间的联系紧密,处于边缘处的结点之间彼此联系比较稀疏,有与核心点建立关系的倾向。

在社会网络分析中,点的中间中心度是一个用来测量该点在多大程度上控制他人之间交往的指标,测

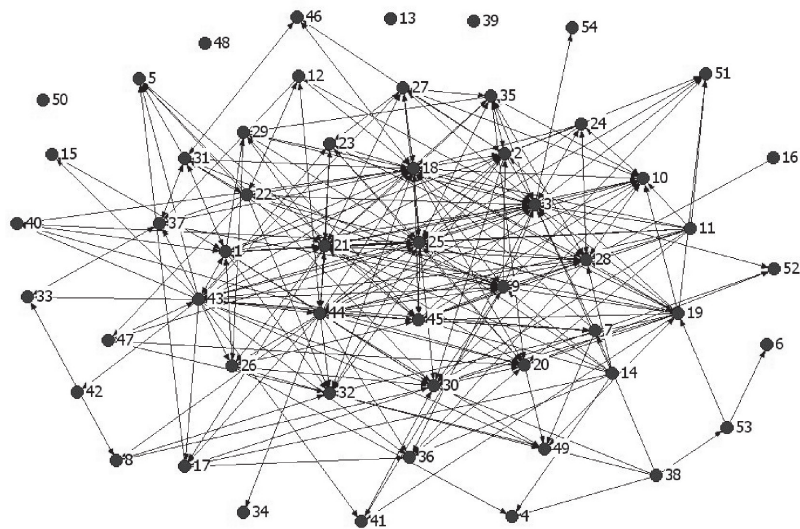


图1 交流网络核心-边缘结构图

量的是该点对资源控制的程度。现利用 Ucinet 测量关系网络中结点的中间中心度,并与核心-边缘分析结果进行比较,如表 3 所示。

由表 3 可以看出,中间中心度排名靠前的结点基本都在核心区域,排名前 20 个结点均为核心结点,说明这 20 个结点本身的联系非常紧密,既属于一个密不可分的团体,同时也处在其他结点的最短路径上,控制着大部分结点的交流。

将排名范围扩大到前 27 个结点(结点集的一半),也只有 8 号、36 号和 53 号为外围结点。外围结点的存在方式主要为三种:a. 属于刚刚参与此网络结构的新成员;b. 作为此结构和其他结构中间的“桥”存

表3 中间中心度与核心区比较(部分)

结点	1		2	
	绝对中间中心度	相对中间中心度	是否核心区	
18 18	250.900	9.104	是	
25 25	222.626	8.078	是	
44 44	175.627	6.373	是	
3 3	162.125	5.883	是	
43 43	153.187	5.558	是	
21 21	150.100	5.446	是	
9 9	139.713	5.069	是	
10 10	127.508	4.627	是	
1 1	108.928	3.952	是	
20 20	102.402	3.716	是	
19 19	97.507	3.538	是	
30 30	94.104	3.415	是	
45 45	70.685	2.565	是	
28 28	63.654	2.310	是	
22 22	54.464	1.976	是	
26 26	42.779	1.552	是	
2 2	41.220	1.496	是	
37 37	38.755	1.406	是	
32 32	28.931	1.050	是	
31 31	28.811	1.045	是	
23 23	27.497	0.998	是	
8 8	23.909	0.868		
7 7	19.060	0.692	是	
36 36	18.329	0.665		
27 27	17.044	0.618	是	
49 49	14.045	0.510		
35 35	13.316	0.483	是	
12 12	7.114	0.258	是	
... ..				

在;c. 是一种稀缺资源并且从属于另一个网络结构之中。通过核心-边缘结构图可以看出,上述三个结点虽然与其他结点的联系不够紧密,属边缘结点,但自身也处在其他几个结点的交往网络路径上,在一定程度上控制着他人的交往,因此中间中心度相对比较靠前。再通过这三个结点用户的基本资料可以看出,8 号程宫是使用微博时间最短的 5 个用户之一,属于外围结点存在方式的第一种情况;36 号王小鱼和 49 号谭飞关注别人的人数分别为 9 和 4,与本文研究的关注网络联系不够紧密,而关注这两人的人数却达到了 5 621 人和 1

132 人,从某种程度上讲,两人属于一种稀缺资源,吸引了如此多人去关注他们,属于外围结点存在方式的第三种情况见表 4。

表4 用户基本资料

序号	名称	他关注的	关注他的	微博数	使用天数	平均日发数	他收藏数
8	程宫	78	80	176	178	0.99	0
36	王小鱼	9	5621	100	242	0.41	0
49	谭飞	4	1132	204	274	0.74	0

4.2 核心结点二次分析 由于核心结点较多,相互间的联系复杂,现将处于核心部分的结点抽出,对其进行二次核心-边缘结构分析,进一步观察其中结构特征,核心结点二次分析结果如下:

Starting fitness:0.544

Final fitness:0.586

Core/Periphery Class Memberships:

- 1: 1 2 3 9 10 18 21 25 27 28 43 44 45
- 2: 7 11 12 14 19 20 22 23 24 26 29 30 31 32 35 37

重排后的矩阵与理想矩阵的相关系数达到 0.586,表明实际数据的核心-边缘结构模型更加显著。同时也可得出交流网络中联系最紧密,处于绝对核心的 1 号、2 号、3 号等 13 个结点。

4.3 聚类分析 将用户基本数据资料组成的 54 条记录导入 SPSS,进行系统聚类法分析。系统聚类法的基本思想是:首先视 n 个观测值(或者变量)各自成为一类,然后找性质最接近的两个类合并成一个新类,计算在新的类别分划下各类之间的距离,再将性质最接近的两类合并,直到所有模式聚成一类为止。以用户为结点,以他关注的人数、关注他的人数、微博数、收藏数、使用天数和平均每日发博数作为聚类变量,做标准化处理后进行聚类分析。生成如下树状图(见图 2),横向距离表示差异大小。

根据结果,可将用户结点划分为若干群组:

第一组为最大规模群组,包括 36 个结点,分别是 51、54、29、48、52、23、15、42、12、16、13、39、37、35、5、49、26、50、17、34、1、6、4、31、45、40、7、11、41、24、27、14、33、8、30、32,其中核心结点和外围结点数量相当,由于结点数过多,无太多研究意义;第二组包括:28、47、25、44、20、21、10、9 八个结点;第三组包括:3、18、43、22 四个结点;余下的 19、38、36、2、46 五个结点相对独立。

聚类分析是根据观测值和变量对样本进行分组,而核心区域和外围区域的划分则是根据核心-边缘理论,表面上看两种分析没有共同点,也不具备可比较性。但仔细对比聚类群组情况和核心-边缘划分情况后发现:第二组聚类除了 47 号其余结点均为核心节点,第三组结点全部为核心结点,聚类分离出的小群组结点和核心-边缘划分的核心结点存在一定程度的重叠。

进一步将聚类分组情况与核心结点二次分析结果进行比较后发现:第二组和第三组聚类包括 3、9、10、18、20、21、22、25、28、43、44、47 共 12 个结点,核心-边缘二次分析后核心结点包括 1、2、3、9、10、18、21、25、27、28、43、44、45 共 13 个结点,其中共有结点数为 9 个,共有比例分别占到 75% 和 69%。也就是说,以观测值为基础的聚类分析结果中小群组结点和以交流网络为基础的核心-边缘分析结果中核心结点有很高的重叠性。

根据上述发现,作者认为,之所以产生这样的结果是因为在微博社区中,一部分人发布微博频繁,发博数量越来越多,渐渐引起彼此的注意,比较多地关注别人或被别人关注,联系日趋紧密,形成了社区交流网络中

的核心成员,其他更多的人也通过这部分人聚在一起,彼此之间的交流也由这部分人所控制,形成了更大的交流网络。核心成员的形成是因为彼此有一些共性,如博文更新快,数量多,喜欢关注别人等,这些共性是聚类分析的基础,也是渐渐现成交流网络核心的基础,反映在结果上就是聚类分析和核心-边缘分析能够发现一些特点相同或相似的结点。

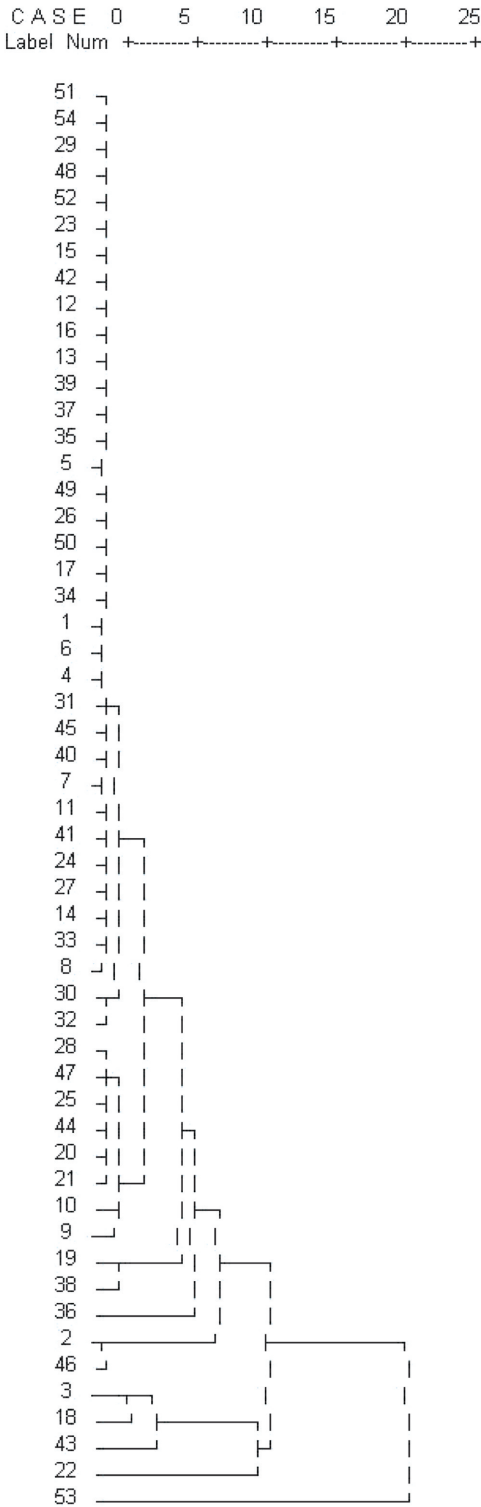


图 2 聚类分析树状图

(上接第 202 页)

5 结束语

本研究较为系统地考察了“Myspace 9911 微博”的网络交流结构和用户一般特征,利用核心-边缘方法和中心性分析界定了核心结点和边缘结点,并通过聚类方法对用户进行了分组,在比较两种分类结果后发现:微博用户的关注数、发博数等属性是聚类分析的直接基础,是核心-边缘分析的间接基础。

本文对于促进微博社区用户的交流及对信息资源的挖掘与利用具有一定的现实意义。在研究方法上,从数理统计的角度尚无法对本文的发现进行解释,后续论文将尝试从其他角度进一步研究。另外,由于“传媒记者”板块的微博用户仅能从一个侧面反映微博社区交流网络的结构,无法反映整体社区的情况,因此在后续研究中还要选取更多领域的样本进行综合研究,得出更加全面的结论。

参 考 文 献

[1] 维基百科. 微博客的定义[EB]. [2010-05-22]. <http://zh.wikipedia.org/zh-cn/微博客>

[2] 王晓光. 微博客用户行为特征与关系特征实证分析[J]. 图书情报工作,2010,54(14):66-70

[3] 李 华,赵文伟. 微博客:图书馆的下一个网络新贵工具[J]. 图书与情报,2009(4):78-82

[4] 郑燕林,李卢一. 微博客教育应用初探[J]. 中国教育信息化,2010(2):29-32

[5] 赵战花,来向武. 微博客对新闻信息传播的影响探析[J]. 理论导刊,2010(4):93-95

[6] Borgatti P, Everett G. Models of core/Periphery Structures[J]. Social Networks,2000,21(4):375-395

[7] 包 卿,陈 雄,朱华友等. 基于核心-边缘理论的地方产业集群升级发展探讨[J]. 国土与自然资源研究,2005(3):3-5

[8] 史春云,张 捷,尤海梅等. 四川省旅游区域核心-边缘空间格局演变[J]. 地理学报,2007,62(6):631-639

[9] 张 玥,朱庆华. 学术博客交流网络的核心-边缘结构分析实证研究[J]. 图书情报工作,2009(6):25-29

[10] 孙 雷,孙庆苏. 数据挖掘在高校图书馆智能分析中的应用[J]. 现代情报,2009,29(8):185-190

[11] 孙 萍,蒋昌俊. 聚类分析及关联挖掘在 Web 服务组合中的应用研究[J]. 高技术通讯,2008,18(11):1187-1194

[12] 火车头采集器. <http://www.locoy.com/>[CP],2010-05-22

(责编:贺晓利)