基于社会网络分析的微博用户网络结构研究

宋 洋 田爱奎 张 义
(山东理工大学计算机科学与技术学院 淄博 255049)

摘 要 运用社会网络分析法,以新浪微博特定标签圈下的用户为研究样本,结合用户之间的"关注"与被"关注"关系,构建了用户"相互关注"网络,然后分别从点度中心度、中间中心度和凝聚子群分析等几个方面对该特定网络进行了分析。提出了特定网络的核心用户分析方法,将得到的实验结果与现实数据进行对比,揭示了该特定标签下的核心人物以及网络成员间的关系,实验结果表明该方法对特定网络结构分析具有可行性。最后提出了相应的启示。

关键词 社会网络分析,微博,中心性,核心人物
中图法分类号 G350 **文献标识码** A

Research on Network Structure of Microbloggers Based on Social Network Analysis

SONG Yang TIAN Ai-kui ZHANG Yi
(College of Computer Science and Technology, Shandong University of Technology, Zibo 255049, China)

Abstract This paper took the users who are under the tag of entertainment in Sina microblog as sample. Combined with "concerning" and "concerned" relationship between users, it built "mutual-concern" network by using social network analysis, then analyzd the network respectively from degree centrality, betweenness centrality and cohesive subgroups analysis. Comparing with real data by using the given core users analysis method, it revealed the core microblogger of the tag and the relationship between network members. Experimental results indicate that the method is feasible for specific network. In the end, it put forward the corresponding enlightenments.

Keywords Social network analysis, Microblog, Centrality, Core microblogger

微博,即微型博客(MicroBlog)的简称,是一个基于用户关系信息共享、传播以及获取的平台[1]。它具有明显的"短、灵、快"特性,用户可以随时随地利用计算机、手机等各种连网终端,以短信息的形式(通常为140字)发布最新的动态和想法,实现即时分享[2]。2006年Twitter的出现使得微博在全球得到了广泛的推广和使用,自2009年中国各大门户网站相继推出自己的微博产品后,微博在中国也得到了迅速发展,用户数目与日俱增。

近两年来,学术界对微博的研究也越来越多。国外由于Twitter发展比较早,定量研究比较多,研究也比较成熟。国内学者对微博的研究才刚刚起步,相对来说还不够成熟,主要以定性分析为主。

1 社会网络分析

社会网络是由点(行动者)和各点之间的连线(关系)组成的集合。任何一个社会单位或者实体都可以看成"点"或行动者(actor),"连线"可以是朋友、上下级、国家之间的贸易等各种关系[3]。分析社会网络,主要是研究社会实体的关系连结以及这些连结关系的模式、结构和功能。国外的John Scott在其著作Social Network Analysis: A Handbook中对社会网络分析做了详细介绍[4];国内的刘军在《社会网络分析导论》中也比较全面地介绍了"社会网络分析"的基本概念和方法[3]。微博用户及其间的相互关系(关注、被关注),使得用户之间因为相互认识或者由于共同兴趣爱好而形成了庞大的社会网络结构。因此,通过社会网络视角来对微博进行结构分析具有很大优势。

目前仍然缺乏特定群体之间的微博社会网络的研究[5],本文选取新浪微博"风云影响力榜"娱乐标签下这一特定群体形成的社会网络作为研究对象,利用网络分析软件UCINET,从密度、中心性和凝聚子群等几个方面来对这一特殊标签圈下的网络结构进行分析,以期揭示新浪微博特定标签圈的网络结构特征,而不是以往整个泛化的微博网络,旨在为将来分析不同特定群体之间的关系奠定基础工作。通过探究该特定标签圈的网络特征,分析该特定标签圈下的核心人物及其网络成员间的关系,并结合现实对分析出来的核心用户与"名人风云榜"娱乐标签下的热门用户进行比较,看所谓的热门用户是否与我们分析出来的核心用户相一致,若不一致对我们又有什么启示。同时将分析得出的数据结果与现有的研究成果相比较,揭示该特定群体下网络结构与以往"泛化"的微博网络结构有何异同,并对现有微博发展提出相应的建议。

2 数据获取及处理

鉴于娱乐圈这一特定群体受关注度较高,成员之间交流

本文受教育部人文社科青年基金项目(13YJC790017)资助。

宋 洋(1988—),男,硕士生,主要研究方向为无线通信网络,E-mail:zbsongyang@163.com;田爱奎(1964—),男,博士,教授,硕士生导师,主要研究方向为计算机教育游戏、虚拟现实等;张 义(1983—),女,硕士生,主要研究方向为图形与图像处理。

较频繁，相对来说比较有代表性，所以选取娱乐圈这一群体作为研究对象，但同时由于无法获取一份相对完整的名单，因此本文只选取新浪微博“风云影响力榜”娱乐标签下的前 100 名成员作为研究对象。选取 2013 年 8 月份的排名结果，并于 9 月 20 日至 25 日对前 100 名成员进行观察记录，得到如表 1 所列的样本。然后将用户之间的“关注”与被“关注”关系用一个 100×100 的邻接矩阵形式表示出来，为了方便表示，我们用排名序号来代替用户昵称，并将所得数据保存于 Excel 表中待后续分析使用，结果如表 2 所列。

表 1　用户排名(部分)

排名	用户昵称	排名	用户昵称
1	王力宏	8	炎亞綸
2	杨幂	9	阿信
3	夢想家林志穎	10	舒淇
4	陈坤	11	古川雄輝
5	陈晓	12	汪东城 TD
6	八卦_我实在是太 CJ 了	13	吴奇隆 TD
7	柯震東 Kai	14	姚晨

表 2　用户“关注矩阵”(部分)

	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	1
2	1	0	1	1	0	0	1	0	0
3	1	0	0	0	0	0	1	0	0
…	…	…	…	…	…	…	…	…	…
7	1	1	1	0	0	0	0	0	1
8	1	0	1	0	0	0	0	0	1
9	1	0	0	0	0	0	1	0	0

方阵中，行表示“关注”者，列表示被“关注”者，“1”表示“关注”关系存在，“0”表示关系不存在，如 $a_{ij}=1$ 表示用户 i “关注”用户 j，反之，表示不“关注”。矩阵主对角线上的值无意义，定义为 0。由于“关注”与被“关注”是双向的，用户 a 可以单方面关注用户 b，而 b 可以不用关注 a，即 a_{ij} 与 a_{ji} 可以不是完全相等的，因此得到的“关注矩阵”是一个非对称矩阵，在后续操作中还需要对其进行对称化处理[6]。

将该 100×100 邻接矩阵导入 UCINET 软件，通过绘图功能，得到我们所研究的 100 名微博用户的网络结构关系图，如图 1 所示。

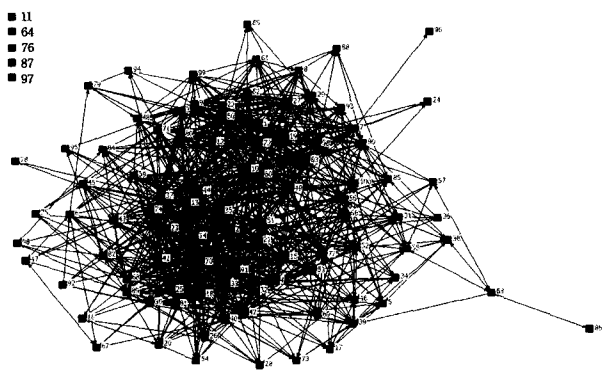

图 1　微博用户社会网络关联图

3　结构分析

利用 UCINET 软件，从密度、中心性、凝聚子群分析与小世界效应等方面对所得数据进行分析。

3.1　中心性分析

中心性是一个重要的个人结构位置指针，评价一个人重要与否，衡量他的职务的地位优越性或特权性，以及社会声望等常用这一指标[7]。中心度指标主要有：点度中心度(Degree Centrality)、中间中心度(Betweenness Centrality)和接近中心度(Closeness Centrality)。

点度中心度。这个概念最初来源于社会计量学的“明星”(star)这个概念，经常用来衡量群体的中心人物，即处于网络的中心位置，影响力较大的人物。一个点的点度中心度为网络中与该点有联系的其他点的总数目。简单来说，如果一个点与其他许多点直接相连，我们就说该点具有较高的点度中心度。通过 UCINET 计算得到网络的点度中心度分析结果如表 3 所列。

表 3　点度中心度分析结果(部分)

用户	OutDegree	InDegree
37	46.000	1.000
2	39.000	44.000
13	36.000	20.000
74	32.000	3.000
41	30.000	26.000
91	30.000	12.000
…	…	…
7	20.000	16.000
89	19.000	18.000

由于“关注矩阵”的非对称性，我们得到的是一个具有方向性的图，因此点度中心度测度有内中心度(in-centrality)和外中心度(out-centrality)，分别对应“点入度”和“点出度”[4]。在这里，点入度表示一个微博用户“关注”其他用户的程度，点出度表示一个用户被其他人“关注”的程度。从分析结果来看，点出度最大的是 37 号用户(点出度为 46)；点入度最大的前 6 名用户分别是 2 号用户(点入度为 44)、14 号和 60 号用户(点入度都为 36)、35 号(点入度为 34)、4 号和 21 号用户(点入度为 33)；11 号、64 号、76 号、87 号和 97 号用户的点入度和点出度都为 0，在网络中成为孤立的节点，如图 1 所示。2 号(杨幂)、14 号(姚晨)、60 号(大 S)、35 号(李冰冰)、4 号(陈坤)和 21 号(赵薇)具有高于其他人的点入度，即更受人“关注”，拥有较大的“权力”，他们发布的信息也更容易被他人所注意。在此分析中我们更加关注点入度，所以不再讨论点出度。同时，可以得到整个网络的标准化点出度中心势为 32.068%，点入度中心势为 30.028%，两者相差不大，且值较低，说明整个网络的中心势不是很大，不具有明显的集中性、向心性。

中间中心度。如果一个点处于许多其他点的测地线上，即两点之间的长度最短的途径上，那么该点就具有较高的中间中心度[8]。它用来衡量一个人作为媒介者的能力，起到沟通其他行动者的作用。

通过分析表 4 可以清晰地看出，2 号用户具有明显高于其他用户的中间中心度，在该网络中控制信息流动的能力最强；其次是 55 号(姜潮)、72 号(潘玮柏)、10 号(舒淇)和 14 号(姚晨)，这些节点处在资源流向其他节点的关键性位置，是信息流动的枢纽。整个网络的中间中心势为 0.0639，因此可以认为整个网络中的大部分节点不需要别的节点作为桥节点就可以获得信息。

表4 中间中心度分析结果(部分)

用户	Betweenness	nBetweenness
2	711.128	7.330
55	335.386	3.457
72	329.127	3.392
10	314.623	3.243
14	285.398	2.942
18	277.771	2.863
…	…	…
51	92.139	0.950
40	86.095	0.887

接近中心度。如果一个点与网络中所有其他点距离都很短,称该点是整体中心点。接近中心度的计算要求很高,必须是完全相连的图形才能计算接近中心度,且度中心性与它高度相关,即度中心性高的人接近中心性往往也较高,所以在此不计算这一指标。

通过对比表3、表4可以发现,2号(杨幂)、14号(姚晨)、18号(宁财神)和50号用户(蔡卓妍)在点度中心度和中间中心度的排名都比较靠前,相对其他用户来说,他们处于该网络比较关键的位置,拥有较大的"权力",因此可以认为他们是整个网络的核心人物。

3.2 凝聚子群——派系

凝聚子群(cohesive subgroups)分析是社会结构研究的重要方法,在社会网络研究中没有明确的定义,大体上说,它能揭示社会行动者之间实际存在的或者潜在的关系,利用一些算法找出行动者集合中具有相对较强的、直接的、紧密的、经常的或者积极的关系[8]。由于可以从多角度来分析"较强、紧密、经常以及积极"等关系的属性,因此凝聚子的概念也就有多种,主要分析方法有成分、派系、n-派系、n-宗派等。

"派系"指的是至少包含3个点的最大完备子图,即派系中任何两点之间都是直接相关的。派系分析根据群体互惠性关系进行凝聚子群分析[9],即成员之间的关系都是互惠的。如图2所示,分别是3成员派系、4成员派系、5成员派系,它们分别包含3条线、6条线以及10条线。

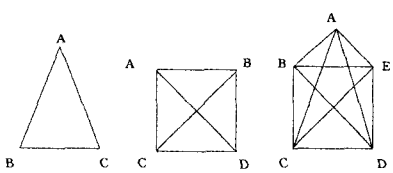

图2 不同规模派系图示

有向图中的派系较严格,称为强派系,经分析得到41个"强派系",如表5所列。

表5 派系分析结果

派系	派系成员
1	2、4、14、15、18、21、35
2	2、14、15、18、21、35、41
3	2、14、18、21、23、35
4	2、14、15、18、35、41、59
5	2、14、18、23、35、59
…	…
40	18、21、25、35、41、89
41	4、18、21、25、35、89

3.3 密度分析及小世界效应验证

密度是社会网络分析最常用的一种测度,用来描绘网络中各个节点关联的紧密程度,固定规模的节点间连线越多该网络的密度就越大。在有向图中,密度 $\rho = l/n(n-1)$,其中 l 为网络中实际拥有的连线数,n 为节点总数[10]。完备图(Complete Graph)每个节点都与其他节点直接相连,密度为1,因此,通常我们分析密度也是为了探究网络图在多大程度上具有这种完备性。经过分析,得到该群体网络的密度为0.1472。一般来说,关系紧密的群体,信息流动比较容易;关系疏远的群体,则常有信息不通等问题,该网络密度值偏低,说明该特定群体形成的网络节点之间信息交流或人际交流不太紧密,完备性较低。

小世界现象(Small-world Phenomenon)也称为六度分隔理论(Six Degrees of Separation),该理论出自约翰·格雷的一部电影,其中的一句台词简单地概括了这个理论:"在这个世界,任意两个人之间,只隔着6个人"。经过分析,该网络图节点之间的平均距离为2.134,说明该群体中任意两个微博用户之间平均只要通过2个人就可以相互连通,一般情况该值介于3至7之间,因此该微博群体网络小世界效应明显,这也说明该网络中的微博用户之间具有很好的信息交流渠道。

结束语 通过运用社会网络分析法对该娱乐标签下的微博用户进行分析发现:

(1)该特定网络的核心用户与新浪"风云影响力榜"用户排名并不完全一致。这可能是由于新浪微博影响力是根据活跃度、传播力以及覆盖度3方面计算得到的,而分析出的核心用户是单纯地针对这一特定群体进行某一指标的验证。因此可以认为,这一特定群体的交流并不局限于他们内部,而是更倾向于与外部网络的节点进行交流。

(2)该特定群体存在多个"强派系",而核心用户又存在于半数以上的派系中,因此它能够增进该娱乐圈成员之间的深入交流,对该群体的发展具有积极作用。同时该群体网络中存在多个孤立节点,结合实际发现他们大部分是国外用户,说明该群体缺乏与国外娱乐人员的交流,不利于我国娱乐圈的现实发展。

(3)具有典型的小世界现象效应,此微博用户之间的平均距离在2个人左右,说明特定标签下的用户信息流通效率、群体成员之间的关联程度比较高。

通过分析,也可以对我国现有微博好友推荐机制提出相应的启示:即人气指标与学术分析指标相结合的推荐方法。特定样本的研究有助于对不同群体、兴趣圈进行深层次了解,对于促进群体发展有着积极意义,但是实验样本偏少,所选取样本也具有特殊性,对于更大的特定网络群体还有待验证,下一步的重点是将对规模更大的、不同特定群体之间的关系进行验证,而验证方法也有待进一步优化处理。

参考文献

[1] 平亮,宗利永. 基于社会网络中心性分析的微博信息传播研究——以Sina微博为例[J]. 图书情报知识,2010(6):92-97

[2] Java A, Song X D, Finin T, et al. Why We Twitter: Understand Microblogging Usage and Communities [C]// Proceedings of the 9th WEBKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. University of Maryland. Baltimore County, ACM Press, 2007: 56-65

[3] 刘军. 社会网络分析导论[M]. 北京:社会科学文献出版社,2004:4-9

[4] Scott J. Social Network Analysis: a Handbook [M]. SAGE Pub-

lications Ltd,2000

[5] 罗文伯. 社会网络视角下的微博研究[J]. 今传媒,2013(2):108-109

[6] 刘军. 整体网分析讲义——UCINET 软件应用[C]//第二届社会网络与关系管理研讨会. 哈尔滨:哈尔滨工程大学社会学系,2007:119

[7] 罗家德. 社会网络分析讲义[M]. 北京:社会科学文献出版社,2005:150-151

[8] 丁兆云,贾焰,周斌,等. 社交网络影响力研究综述[J]. 计算机科学,2014,41(1):48-53

[9] 党洪莉,孙红霞. 图书情报学博客的社会网络分析[J]. 情报杂志,2009(1):180-181

[10] 袁圆,孙霄凌,朱庆华. 微博用户关注兴趣的社会网络分析[J]. 现代图书情报技术,2012(2):68-75

[11] 魏顺平. 社会网络分析及其应用案例[J]. 现代教育技术,2010,20(3):29-34

(上接第 194 页)

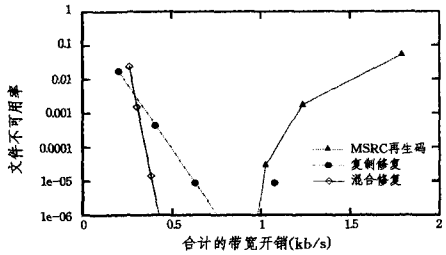

图 4 $n=20, k=10$ 时不同冗余策略的带宽/可用性曲线

从图中可以得出,要获得同样的数据可用性,MSRRC 编码比其它策略需要更多的网络带宽,比如在 0.0001 位置(即若需要 99.9999%的数据可用性),但存储开销则大大减少。但值得注意的是:MSRRC 冗余再生码只需维护一种冗余类型,复合策略需同时维护完整副本和纠删码分块,故冗余再生码比复合策略的系统设计简单,又由于冗余再生码具有纠删码的一切属性,因此其整体可用性是复制策略所无法比拟的。

结束语 冗余再生码(Regenerating Codes)是一种分布式存储编码,能优化失效节点修复中的带宽与存储开销。论文根据冗余再生码数据再生时的极值点:最小存储再生点(MSR),通过矩阵运算实现最小存储冗余再生码 MSRRC。文中详细给出再生码的数据重构和失效节点再生的实现原理,并利用实例描述了冗余再生码的实现过程,在理论上证明了实现原理的可靠性,文中还给出了实验运行结果。

参 考 文 献

[1] Patterson D A, Gibson G, Katz R H. A case for redundant arrays of inexpensive disks (RAID) [C] // Proc. ACM SIGMOD Chicage. USA, June 1988:109-116

[2] Kubiatowicz J, Bindel D, Chen Y, et al. OceanStore: An Architecture for Global-Scale Persistent Store[C] // Proceedings of the Ninth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2000). 2000:190-201

[3] Bhagwan R, Kati K, Cheng Y-C, et al. Total recall: System support for automated availability management [C] // Proc. of ACM/USENIX NSDI'04. 2004:337-350

[4] Wu Y, Dimakis A G, Ramchandran K. Deterministic regenerating codes for distributed storage[C] // Allerton Conference on Control, Computing and Communication. Monticello, October 2007

[5] Wang Z, Dimakis A G, Bruck J. Rebuilding for Array Codes in Distributed Storage Systems[C] // Workshop on the Application of Communication Theory to Emerging Memory Technologies (ACTEMT). Dec. 2010

[6] Akhlaghi S, Kiani A, Ghanavati M R. A Fundamental Tradeoff between the download cost and repair bandwidth in distributed Storage systems[C] // Proceeding of IEEE International Symposium on network Coding (NetCod). Toronto, Jun. 2010

[7] Shah N B, Rashmi K V, Kumar P V. A Flexible Class of generating Codes for Distributed Storage[C] // Proceeding of IEEE International symposium on information theory (ISIT). Austin, Jun. 2010:1943-1947

[8] Gaston B, Pujol J. Double Circulate Minimum Storage generating Codes. 2010[OL]. http://arXiv:1007.2401[cs. IT]

[9] Wang Z, Mateescu R, Dimakis A G, et al. Array codes for distributed storage: Results and open problems [J]. Information Tehory and Application. ITA, 2010

[10] Wu Y. Existence and construction of capacity-achieving network codes for distributed storage[J]. Journal on Selected Areas in Communications, 2010, 28(2):277-288

[11] Li J, Yang S, Wang Xin, et al. Tree structured Data Regeneration in Distributed Storage Systems with Regenerating Codes [C] // Proceedings of IEEE INFOCOM. 2010

[12] Shah N B, Rashmi K V, Kumar P V, et al. Distributed Storage Codes with Repair-by-Transfer and Non-achievability of Interior Points on the Storage-Bandwidth Tradeoff [C] // Allerton Conf. . Urbana-Champaign, Sep. 2009

[13] Dimakis A G, Godfrey P G, Wainwright M J, et al. Network coding for peer-to-peer storage[C] // Proceeding of INFOCOM. Anchorage, Alaska, May 2007

[14] Wu Y, Dimakis A G, Ramchandran K. Deterministic regenerating codes for distributed storage[C] // Allerton Conference on Control, Computing, and Communication. Monticello, October 2007

[15] Shah N B, Rashmi K V, Kumar P V, et al. Distributed storage codes with repair-by-transfer and non-achievability of interior points on the storage-bandwidth tradeoff[J/OL]. http://arxiv.org/abs/1011.236

[16] Wu Y, Dimakis A G, Ramchandran K. Deterministic Regenerating codes for distributed storage[C] // Proc. Allerton Conf. . Urbana-Champaign, Sep. 2007

[17] Ramabhadran S, Pasquale J. Analysis of durability in replicated distrusted storage systems[C] // 2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS). Atlanta, GA, April 2010:1-12

[18] 王禹,赵跃龙,侯昉. 基于副本管理的 P2P 存储系统的可靠性问题研究[J]. 华南理工大学学报,2011,39(2):148-152

[19] 王禹,赵跃龙,侯昉,分布式存储系统最小带宽再生码研究[J]. 小型微型计算机系统,2012,33(8):1710-1714