微博社交网络社区发现方法研究

范超然[1],黄曙光[2],李永成[1]

(1.合肥电子工程学院 研究生管理大队,安徽 合肥 230037;
2. 合肥电子工程学院 网络工程系,安徽 合肥 230037)

摘 要:基于分析微博社交网络用户之间关系,提出了一种适用于微博的社区发现方法。实验表明这种方法能够有效地发掘微博社交网络中的社区结构。

关键词:社区发现;微博;社交网络

中图分类号:TP393 **文献标识码**:A **文章编号**:1674-7720(2012)23-0067-04

Study on microblog social network community detection

Fan Chaoran[1], Huang Shuguang[2], Li Yongcheng[1]

(1.Department of Graduate, Electronic Engineering Institute, Hefei 230037, China;
2.Department of Network, Electronic Engineering Institute, Hefei 230037, China)

Abstract: In this paper, we propose a new method of finding community structure in microblog social network through analyzing the relations among users, experimental results confirm the validity of this method.

Key words: community detection; micorblog; social network

微博作为一种新兴的社交媒体,其用户以及影响力越来越广泛,微博从一开始的社交娱乐工具到现在的重要营销手段,得到了前所未有的关注。微博不同于传统的社交媒体一对多的信息传播模式,它的传播具有迅捷性和裂变性[1],这种信息传播的模式使得微博在突发事件的传播以及舆论的扩散方面具有更强的作用力。随着复杂网络研究的不断深入,以此为基础理论的社交媒体研究正成为社会网络研究的一大分支。复杂网络中的一个主要特征是社区性[2],社区的一般定义是同一社区内的节点与节点之间的连接很紧密,而社区与社区之间的连接比较稀疏[3]。社区发现对于挖掘网络中的功能模块以及研究网络的演化是非常重要的。本文提出了一种基于关系分析的社区发现方法。

1 社区发现相关研究

社区发现从算法的角度可以分为两种[4]:(1)基于优化的算法,其中包括著名的谱方法,基本思想是采用二次型优化技术最小化预定义的“截”函数,具有最小“截”的划分被认为是最优的网络划分。(2)Kernighan和Lin在1970年提出KL算法[5],该算法是一种试探优化算法,它将网络分割成两个大小已知的子网络即社区,并且应用了贪婪算法的原理。由于以上两种算法的开销较大,Newman提出了一种快速聚类算法[6],该算法优化的目标是模块度函数Q,该函数定义为簇内实际连接数目与随机连接情况下簇内期望连接数目之差,用来衡量社区划分的质量,该算法通过合并使ΔQ最大的点的方法形成一个自底向上的聚类过程,该算法在效率上有了很大的提高。AaronClauset等人提出的CNM算法[7]在效率上有了更进一步的提高,算法复杂度为$O(n\times\log^2 n)$,接近线性复杂度,这也是本文采用此算法的重要原因。除了优化方法以外还有一种基于启发式的方法,该类算法能够快速找到网络中社区的近似最优解,其中包括最经典的GN算法[8],它通过计算迭代分割有最大边介数边的方法来划分网络。除了以上两类方法以外有学者还提出了一类基于模型的社区发现方法,其中包括标签传播算法LPA[9],基于随机游走的Infomap算法[10]等。传统意义上的社区发现方法仅仅从网络拓扑结构出发挖掘连接紧密的簇结构,随着复杂网络研究的不断扩展特别是在线社交网络的深入研究,相关学者试图利用节点和边的内容来发现在线社交网络社区。燕飞[11]等人提出了一种综合兴趣和网络拓扑结构的社区发现方法,Yang

Tianbao 等人[12]提出了一种将内容与链接结合的概率模型。针对 Twitter 的社区发现，Mohit Naresh Kewalramani[13]在他的硕士论文中利用Twitter 多个属性的相似性并通过传统聚类算法的方法发现社区。

然而，类似于微博的在线社交网络是典型的有向网络，用户之间的指向关系反映了用户与用户之间的紧密联系。单纯地利用用户之间兴趣以及联系内容的相似度来发现社区，会伴随用户兴趣和用户的活跃程度的波动产生划分的歧义，此类划分还会造成大量的重叠社区。微博用户之间转发等关于内容的联系是基于用户关注关系之上的，用户之间关注关系往往是稳定的，针对此项特点本文首先对微博用户之间的关系进行分析构建网络，然后利用用户之间基于内容联系的频繁程度定义用户之间的紧密程度，再利用加权社区发现算法来完成社区发现。

2 关系分析

Granovetter[14]提出社会网络中普遍存在的两种关系：强关系与弱关系，社会学家普遍认为强关系是一种基于信任的关系，而弱关系是一种信息流通的渠道。微博社交网络中从类型上讲有四种关系：关注关系、提及关系、转发关系以及互粉关系，关注关系是指用户以粉丝的形式关注另外一个用户，这种关注形式是单向的，关系展现的是一种拓扑结构。而提及关系以及转发关系是一种以关注关系为基础的关系，这种关系是用户因关注者的内容吸引而产生的关系链接。互粉关系是用户双向关注的关系模式，由此可见在微博社交网络中是一种单向关系与双向关系并存的网络，为了能够在这样的网络中发现关系紧密的社区，首先必须对关注关系与互粉关系对关系的紧密程度的影响进行分析。

本文所采用的数据集是通过 Twitter API 的方式爬取 2012 年一月份部分用户关注关系网络以及用户之间的转发和提及关系，所爬取的网络包括 12 563 个用户和 716 129 条关系数据，此网络记为 G(V,E)，V 代表网络中的节点，E 代表网络中的边。首先分析互粉关系在用户关系中的比重，图 1 是粉丝数与互粉数在所有用户中所占比重的分布情况。

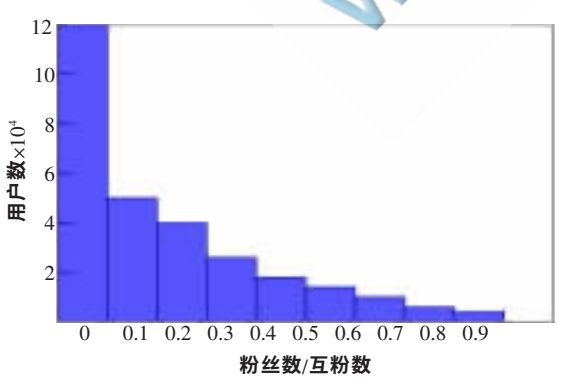

图 1 粉丝数与互粉数在用户中的比例

通过图 1 可以看出大部分的微博用户的粉丝数即双向关系在两种关系中所占的比例较小大多分布在 0.1 之内。其次分析粉丝数与互粉率之间的关系，图 2 是统计曲线图。

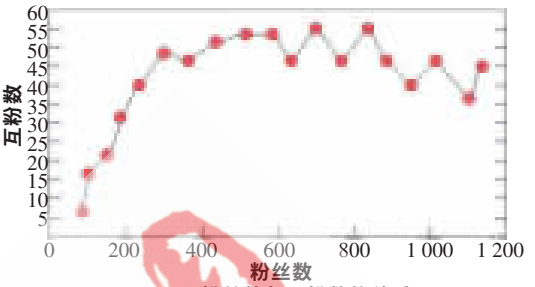

图 2 粉丝数与互粉数的关系

由图 2 可以看出粉丝数与互粉数之间没有必然的线性关系。最后分析互粉数与粉丝数之间的比率和粉丝数之间的关系，图 3 是统计结果。

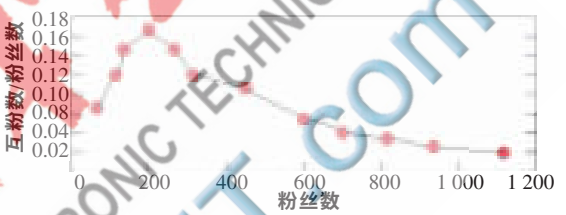

图 3 互粉数与粉丝数比例和粉丝数的关系

由图 3 可以看出随着粉丝数的增加，互粉数所占的比率越来越小。综合以上的统计分析可以得出单纯的关注关系其实是一种很松散的结构，用户关注一个用户完全是“免费”的，所以这种关系的建立带有一定的随意性，或者说这种单向关注关系是一种弱关系，而用户之间的互粉关系往往需要基于两者之间的信任关系或者两者之间有共同的兴趣点，此类是一种强关系，而一个社区内的用户往往联系紧密或者具有一定的共同属性点。

3 构建网络

3.1 网络简化

通过以上分析，可以得出单向的关注关系是一种很弱的单向关系，这种随意的关系在一个强关系社区中影响很小，所以在构建网络的第一步首先过滤掉网络中用户之间的单向链接得到纯粹的具有互粉关系的无向网络 G(V,E)。

3.2 边权值计算

边权[15]是网络中用来衡量节点 i 和节点 j 共享的边的关联度大小的量，记为 r_{ij}。r_{ij} 的值越大，说明节点 i 和 j 之间传输信息的可能性越大，即两点联系的较紧密；反之，则说明节点 i 和 j 之间信息传输比较困难，即两点之间的联系较稀疏。

具有互粉关系的微博用户之间的联系有转发数和提及数，设两个微博用户 A 和 B，A 和 B 之间具有互粉关系，A 转发 B 的次数为 r_sum1，B 转发 A 的次数为 r_sum2，则转发权重为：

$$R=r\_sum1+r\_sum2 \tag{1}$$

A 提及 B 的次数为 m_sum1，B 提及 A 的次数为 m_sum2，则提及权重为：

$$M=m\_sum1+m\_sum2 \tag{2}$$

则 A 和 B 链接的权重为：

$$\omega_{AB}=R+M \tag{3}$$

根据以上步骤可以构建出微博社交网络中具有互粉关系的无向权重图 $G'(V,E)$。

4 算法与实验

4.1 算法改进

CNM 算法采用快速贪婪规则合并划分得出社区结构，是凝聚型算法的典型代表。为了能够快速地找到模块度增长最快的节点，CNM 算法定义了以下数据结构：

(1)一个用来存储每对有连接的点的 ΔQ_{ij}，矩阵的每一行又同时用平衡二叉树(因此插入和查询每个点的时间为 $O(\log n)$)和一个大顶堆来(最大的元素可以最快找到)存放。

(2) 大顶堆 H 包含 ΔQ_{ij} 矩阵中每一行的最大元素，以及标签 i,j 标志社区对。

(3)一个存储 a_i 的向量组。

CNM 算法具有一个很好的特性：在整个算法过程中，模块度 Q 仅有一个峰值(最大值)。当模块度增量矩阵中最大元素都小于 0 以后，Q 的值就只可能一直下降。因此，只要模块度增量矩阵中最大由正变负以后，就可以停止合并，并认为此时的社团结构就是网络的社团结构。

为了适用于无向加权网络，对模块度计算方法做相应改动。ΔQ_{ij} 表示节点 i 加入到邻居节点 j 所在社团时模块度的变化，ΔQ_{ij} 定义如下：

$$\Delta Q_{ij}=\frac{1}{2m}\left[\sum_{ij}\left(\omega'-\frac{k'_i k'_j}{2m}\right)\right]-\frac{1}{2m}\left[\sum_{ij}\left(\omega-\frac{k_i k_j}{2m}\right)\right]$$

其中 ω' 是 i 加入 j 所在社区时 i 与其临边的权值和加上原有的社区边权值的总和。$\frac{k'_i k'_j}{2m}$ 代表 i 加入到 j 所在社区后，i 与原有社区组成新的社区的边数目的期望值。算法的基本流程如表 1 所示。

表 1 加权 CNM 算法

算法：加权社区算法 CNM
输入：无向加权图 $G'(V,E)$，ΔQ_{ij} 矩阵，大顶堆 H，向量组 a_i
输出：社区划分结果以及 Q 值
初始化：赋予每一个节点一个单独的社区号，初始化 ΔQ_{ij} 矩阵构建最大堆 H
repeat
从 H 中找出最大的 ΔQ_{ij} 值，合并对应的社区
更新 ΔQ_{ij} 矩阵、H、a_i，以及 Q 增量
until
所有的 ΔQ_{ij} 由正值变为负值，即 Q 值达到峰值得到网络的最好划分

4.2 实验结果与分析

通过构建网络得到微博无向加权图连通网络 $G'(V,E)$，$G'(V,E)$ 包含 98 327 条互粉关系，通过式(3)赋予每条边相应的权值 ω。算法通过迭代划分网络试图找到最优的社区划分数量。图 4 是模块度变化趋势。

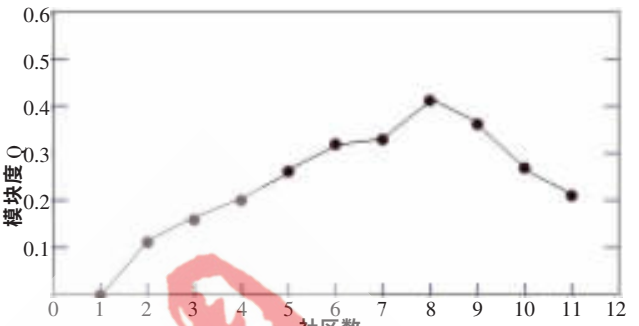

图 4 模块度变化曲线图

由图 4 可知在社区划分数为 8 的时候，Q 值达到峰值 0.401，而通常社区结构较明显的网络模块度介于 0.3~0.7 之间[16]，这时网络的社区划分达到一个最优的效果，实验结果说明该算法实现的网络划分在模块度衡量上有较强的社区结构。社区划分的可视化效果如图 5 所示。

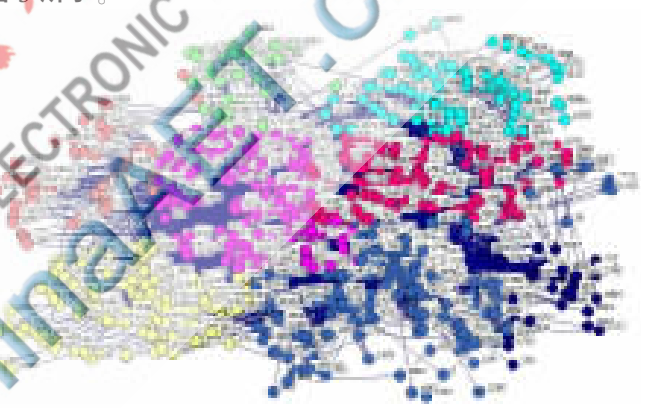

图 5 社区划分可视化效果

为了能够进一步评估社区划分的质量，依据以上的社区划分结果，构建每个社 C_i 区内用户所发 Tweet 中以及转发和提及当中词频较高的词汇集，列出频率较高(>%10)的词汇作为社区的主题标注，统计结果如表 2 所示。

表 2 社区内用户 Tweet 高频词汇统计

社区	节点数	高频率词汇/出现频率/%
1	1 256	politic/40.3、public/26.7
2	1 912	sports/36.5、NBA/12.5
3	2 001	news/68.9
4	977	funny/33.5、family/22.4
5	1 823	new york/46.8
6	1 169	office/32.7、war/17.3
7	1 526	hospital/38.9、health/33.4
8	1 899	music/51.6、movie21.8

依据表 2 可以发现社区 1、2、3、5、7、8 主题相对集中，主题词之间的语义相似度较高，就社区划分解释而

言，这样的社区划分更接近一个真实的社区划分即社区内用户往往关注同一类的主题。而对于社区4、6而言，社区内用户的关注主题之间的语义相似度较低，但通过考察社区内用户之间的联系频率较高，这样的社区划分解释是用户之间的“朋友”关系而产生社区。总体而言，社区划分后的网络中具有明显的社区结构。

本文通过分析微博社交网络中关系的强弱关系对于用户紧密度的影响，通过过滤用户之间单向的关注关系以及根据用户之间的联系对边赋值的方法构造了社区发现的元数据：微博无向加权图。再通过相应的加权社区发现算法实现了在微博社交网络中的社区发现，实验效果显示这种方法能够很好地挖掘网络中的社区结构。然而以微博为代表的社交网络所包含的信息相当丰富，可以说微博社交网络中不但边是多属性的，用户也是多属性的，如何利用这些属性信息挖掘社区是值得探讨的问题。另外微博社交网络的一个重要特点是动态性，动态社区的发现如何运用在微博社交网络中也是一个重要的问题。

参考文献

[1] 李瑗瑗.微博舆论的形成机制及特点分析[J].新闻界，2010(6)：51-52.

[2] LANCICHINETTI A，FORTUNATO S，KERT J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics，2009，3(11)：15-33.

[3] NEWMAN M E J. Communities modules and large-scale structure in networks[J]. Nature Physics，2012(1)：25-31.

[4] 杨博，刘大有，Liu Jiming，等. 复杂网络聚类方法[J]. 软件学报，2009，20(1)：54-66.

[5] KERNIGHAN B W，LIN S. An efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal，1970.49(2)：291-307.

[6] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E，2004，69(6)：1-5.

[7] CLAUSET A，NEWMAN M E J. Finding community structure in very large networks [J]. Physics Review E，2004，(70)：71-76.

[8] GIRVAN M，NEWMAN M E J. Community structure in social and biological networks[J]. Proc. of the National Academy of Science，2002，12(9)：7821-7826.

[9] RAGHAVAN U N，ALBERT R，KUMARA S. Near linear time algorithm to detect community structures in large scale networks[J]. Physical Review E，2007，6(3)：47-58.

[10] ROSVALL M，CARL T. Bergstrom Maps of random walks on complex networks reveal community structure [J]. PNAS，2008，105(4)：1118-1123.

[11] 燕飞，张铭，谭裕韦，等.综合社会行动者兴趣和网络拓扑的社区发现方法[J]. 计算机研究与发展，2010，47：357-362.

[12] Yang Tianbao，Jin Rong，Chi Yun，et al. Combining Link and content for community detection[C]. Adiscriminative Approach KDD'09，Paris，France，2009.

[13] NARESH M，LRAMANI K. Community detection in twitter [D]. Dept of Comuputer Science of University of Maryland Baltimore County，2011：1-60.

[14] GRANOVETTER M S. The strength of weak ties [J]. American Journal of Sociology，1973，78(6)：1360-1380.

[15] LI M，FAN Y，CHEN J，et al. Weighted networks of scientific communication：The measurement and topological role of weight [J]. Physica A，2005，39，(11)：643-656.

[16] NEWMAN M E J，GIRVAN M. Finding and evaluating community structure in networks [J]. Physical Review E，2004，69(2)：32-46.

（收稿日期：2012-09-24）

作者简介：

范超然，男，1988年生，硕士研究生，主要研究方向：复杂网络、可视化。

黄曙光，男，1960年生，教授，博导，主要研究方向：信息安全。

李永成，男，1986年生，博士研究生，主要研究方向：信息安全，数据挖掘。