DOI:10.13624/j.cnki.issn.1001-7445.2013.06.021

文章编号: 1001-7445(2013)06-1413-05

基于 GN 算法的微博社区识别方法

徐 杨,蒙祖强

(广西大学 计算机与电子信息学院,广西 南宁 530004)

摘要: 近年来,社交网络用户数量剧增,关于社交网络上的社区发现成为一种新的需要解决的问题。这里获取微博上的用户以及用户之间的关系作为研究样本,基于微博用户以及用户之间的关系,构建网络社区模型,在此基础上,利用 GN 算法对微博用户进行社区划分; 为了提高算法的运行速度,采用模块度增量,在得出近似结果时就停止,减少运行时间。并在获取的数据上加以验证,GN 算法适合用于社交网络中的社区发现,引入模块度增量有助于提高算法的速度。

关键词: 社交网络; GN 算法; 社区发现
中图分类号: TP391 文献标识码: A

Identify communities in the microblogging based on the GN algorithm

XU Yang, MENG Zu-qiang
(School of Computer, Electronics and Information, Nanning 530004, China)

Abstract: Recently, the increasing number of social network users makes the community finding on the social network become a new problem. By obtaining the users´ ID lists on the microblogging and taking the relationship between the users as the study sample, a network model using GN algorithm to identify the communities based on the relationship between the users was built. In order to improve the running speed, the algorithm uses the increment of the modularity, and stop running when the approximate results are obtained. Experiment results show that GN algorithm is suitable for the community detection on the social network and the increment of the modularity can improve the speed.

Key words: social network; GN algorithm; community detection

0 引 言

社会网络中的部分个体因为一些共同的兴趣爱好或背景形成社会团体,被称为“社区”。社区内的用户有某个共同话题,联系紧密,社区间的用户没有明显的共同话题,联系较少[1]。社区结构作为大规模复杂网络研究的一个基本问题,可以为网络社会关系挖掘、网络稳定性分析等问题奠定基础,也可以从整体上把握网络的结构,了解个体的作用[2]。

社区发现就是要识别出这些社区。社区发现算法主要有两类,一类是基于图论的算法[3-5],算法迭代将网络分割,却无法预知将要把网络分割到哪一步。另一类是层次聚类方法,主要根据节点之间连接

收稿日期: 2013-02-28; 修订日期: 2013-09-13
基金项目: 国家自然科学基金资助项目(61063032,61363027); 广西自然科学基金资助项目(2012GXNSFAA053225)
通讯联系人: 蒙祖强(1974-),男(壮族),广西罗城人,广西大学教授,博士; E-mail: mengzuqiang@163.com。

强度将节点自然分割。该算法主要分成凝聚算法和分裂算法两种。凝聚算法[6-7]通过耦合性判断节点间的联系,它能快速准确的找到社区中耦合性大的节点,但无法较好的划分外围节点。分裂方法[8]逐步删除链接关系最小的边,但如何衡量边的关联度是一个关键问题。这两类算法共同的缺点在于无法划分重叠社区,而派系过滤方法[9]可以解决这个问题,它利用网络中的完全子图,因此若完全子图数目稀少,算法会受到很大影响。

目前,微博作为新生代的网络应用,人气高涨,最近几年发展迅猛。据统计,截止到 2012 年 12 月,新浪微博的用户数量已经超过 5 亿,日活跃用户数量超过 4 620 万。微博消息具有及时性[10],更能够展现出每个用户对当前事件的看法,微博的重点在于建立"人与人"对事件的互动和交流,在微博网络上,更能够形成类似于现实人际交往中形成的"圈子",也就是社区。本文以微博作为研究对象,建立用户之间的关系网络,然后利用 Girvan-Newman 算法[8](以下简称 GN 算法),对这个关系网络进行社区划分,并采用模块度增量提高算法的运算速度。

1 社区发现算法

1.1 社交网络

社交网络是一群拥有相同兴趣爱好或者活动的人创建的在线社区。它为用户之间的信息交流提供了新的方式,拥有和线下活动类似的功能。由于社区内部的用户具有相同的兴趣爱好或者相似的背景[11],网络上会形成一个个基于某种共同爱好或话题的社区。社区是大规模复杂网络中的一个密集子网,相当于网络系统中的一个功能单元。

挖掘出社交网络上的社区很有实用价值。社区内用户的共同话题,会形成一个定向的区别用户的方式,为智能搜索、个性化服务、商业应用推广等应用提供理论依据。从广告投放的角度来看,容易找到定向的广告受众; 从朋友推荐的角度看,能够方便的将同一个社区内的个体向其他个体进行推荐,为找寻多年没有联系的朋友提供一个方便快捷的途径。

社交网络的规模非常大,社区发现方法在时间复杂度和准确率上受了极大的影响,这两点目前是国内外的重点研究对象。

1.2 经典社区识别算法

经过近几年的发展,社区发现算法取得了重要的进展,算法主要分成两类: 基于图论的算法和层次聚类算法。

① 基于图论的算法

基于图论算法的基本思想是将网络分解成子网络,子网络内的节点数基本相同,子网络之间的连接比较少。其中,谱平分法是根据矩阵的特征向量的分量判断节点的相似性。算法预处理时会删除矩阵中接近 0 的边,而且网络的 Laplace 矩阵是稀疏矩阵,因此算法时间复杂度较低。Kernighan-Lin 算法[3]是一种试探优化算法。它是基于贪婪算法原理将网络划分成为两个大小已知的社区的二分法。该算法引入增益函数,按一定条件交换最初两个社区中的节点,直到所有的节点都被交换一次为止。

② 层次聚类算法

层次聚类算法的重点在于分析网络中边之间的连接关系。主要分成凝聚方法和分裂方法两种。凝聚算法的代表有 Newman 快速算法和 CNM 算法,这两种算法均是基于贪婪算法的思想,首先将网络视为无边网络,迭代的对相似性最高的节点对加边,这个过程可以在任意时刻停止。与之相反的分裂方法是逐步移除边的相似性最低的边,将原始网络逐步分解成为较小的子网络。

1.3 近期社区识别算法

① 适应性聚类算法

适应性聚类算法的思想[12]是根据节点的结构,重复的移动节点,使其到吸引力最大的模块,直到与最好的模块度取值相符合为止。这种算法在节点移动过程中的时间复杂度是 $O(n)$,在稀疏图识别过程中的时间复杂度是 $O(n^2)$。

② 数据场算法

将网络视为包含所有节点及其相互作用的物理系统,每个节点周围存在一个作用场。场中节点均将受到其他节点的作用[13]。根据社区的定义,网络场势高的区域可以视为一个社区,社区的边界就具有较小的拓扑势。数据场的方法可以识别出重叠社区,比较有实用性,且时间复杂度比较低。

社区发现虽然方法很多,但是在图论的方法中,分裂出的社区大小比较接近,社交网络中的社区规模大小可能会相差比较大,凝聚算法不能很好的处理非中心节点[14]。因此本文采用 GN 算法来实现社区发现,并采用了模块度增量的概念来提高算法的运算速度。

1.4　GN 算法

GN 算法是一种自顶向下的过程,算法的关键问题有两个: 如何选择分裂的标准和如何确定社区划分的好坏,前者的选择关系到如何度量两个节点或社区的拓扑距离,以及如何度量边对网络连通性,后者是如何从所有可能的网络划分中选择最好的一个来表示网络社区结构。

算法利用边介数来确定分裂的标准,反复移除边介数最大的边,直到所有的边被移除。在这一过程中,对于有 m 条边,n 个点的网络来讲,计算边介数的算法时间复杂度是 $O(mn)$。由于每移除一条边都要进行重计算,因此时间复杂度为 $O(m^2n)$,稀疏网络的时间复杂度为 $O(n^3)$。

为了衡量划分的结果,算法采用模块度作为度量方式,它是指社区结构内部的边数减去随机生成网络中的期望的边数。模块度是基于随机网络没有社区结构的假设为前提,它通过度量网络划分的好坏度量网络社区结构。模块度的定义如下:

$$Q = \sum_i (e_{ii} - a_i^2) = Tre - \| e^2 \| ,$$

其中 e_{ij} 表示网络中连接社区 i 和社区 j 的节点的边在所有边中的比例,a_i 是每行(或者列)中各个元素之和,定义 $a_i = \sum_j e_{ij}$,$Tre = \sum_i e_{ii}$,表示对角线上元素之和。

算法会产生出一个树状图,图中每一层对应的是原网络的一种划分方式,模块度最大的值就是找出这些划分方式中的最佳划分。

2　面向微博的社区识别方法

由于 GN 算法的时间复杂度比较大,在分裂过程中,可以利用模块度增量作为衡量分裂停止的标志,减少分裂的次数,提高算法的运行速度。

2.1　模块度增量

算法的最终结果取值是找出模块度最大的值的情况,因此引入模块度增量作为循环停止的标志。设网络每次划分的社区为 $P_i(0 < i \leqslant n)$,其中 i 为社区编号。当新产生社区时,设将社区 P_m 分裂成两个社区,此时一个社区的编号维持 P_m,另一个社区编号设为 P_n。将模块度增量 ΔQ 定义为社区 P_m 和社区 P_n 内节点有关的边的权值的比例与社区间边的权值的比例的差。因为,模块度增量仅仅与新产生的社区有关,与其他社区无关。

每次分裂的过程中,只需要判断模块度增量,若 $\Delta Q \geqslant 0$,则继续分裂; 否则,则停止分裂,前一次分裂为最终社区划分的结果。

2.2　*Community Id Vector*

在实现过程中,为了确定每一步分解的过程中,哪些社区内的节点发生了变化,需要另外引入一个向量 $\boldsymbol{E}_i$,称为 Community Id Vector,记录网络中每个节点所在的社区编号,以便每次分裂过程中记录节点所属社区编号的改变,也能很容易得出模块度增量。

在分裂的过程中,开始时所有的点聚集成为一个社区 $\{P_1\}$。第一次分裂,对 $\{P_1\}$ 删除边,分裂成两个社区 $\{P_1, P_2\}$。以此类推,分裂过程中第 $n-1$ 次分裂后状态为 $\{P_1, P_2, \cdots, P_n\}$。进一步分裂时,设此时需要分裂社区 P_k,分裂后的两个社区,其中一个社区保持社区编号不变为 P_k,另一个社区编号,为此时所有社区编号中最大的的数值加 1,即为 $n+1$。则在这次分裂后结果为 $\{P_1, P_2, \cdots, P_n, P_{n+1}\}$。

设原始网络中顶点的编号依次为 0,1,2,$\cdots$,n,在每一次分裂过程中 Community Id Vector 定义为 $\boldsymbol{E}_i = (e_1, e_2, \cdots, e_n)$,表示第 i 次分裂后,网络中每个节点归属的社区编号。当网络处于第 $i+1$ 次分裂时,

所得向量 $\boldsymbol{E}_{i+1}$ 与第 i 次分裂时的向量 $\boldsymbol{E}_i$ 的差记为 $\Delta\boldsymbol{E}_i$ 。$\Delta\boldsymbol{E}_i$ 中不为 0 的项所对应的点为此次分裂过程中,所属社区编号发生变化的点; 相反,为 0 的项所对应的点所属的社区没有发生变化。这样就可以方便的计算出模块度增量。

2.3 算法描述

基于以上分析,下面给出微博社区发现算法。

输入: 经过预处理的原始网络

输出: 社区列表,以及社区包含的节点

Begin:

(1) 将原始网络视为一个社区 $\{P_1\}$,记录此时的 Community Id Vector: $\boldsymbol{E}_1$

(2) for 对每一个社区 do

(2.1) 计算每条边的介数

(2.2) 去掉最大介数的边,将这个社区分裂为两个社区

(2.3) 记录分裂后的 Community Id Vector: $\boldsymbol{E}_i$

end for

(3) 计算模块度增量 ΔQ

(4) 若 $\Delta Q \geqslant 0$,转(2); 反之,算法结束。

End

传统的 GN 算法在分裂过程中,将原始网络从全部节点归属于一个社区的情况,分裂成每个节点单独形成社区,然后选择模块度最大的分裂结果,为最终结果。一般情况下,随着分裂的进行,模块度会逐渐增大,直到最大值,然后慢慢减小。但是,在模块度逐渐减小过程中的得出的社区划分并没有对最终划分结果有帮助。因此,当模块度取最值的状态下停止分裂,输出最终结果,能够避免对网络进行无用的分裂,到达减小运行时间的目的。

基于以上分析,本算法在模块度增量为负数的情况下,即在模块度为极值的状态下,停止分裂,得出最后划分结果,并输出,这样就会比传统的 GN 算法快。

3 实验分析

3.1 数据获取和预处理

本文调用相对应的 API 依次获取微博用户,以及他们的关注列表、粉丝列表。在获取数据时,以某一个微博用户的 ID 号为起始点,用这个 ID 用户的第一个关注的用户 ID 作为下一个遍历的微博用户,以此顺序来遍历获取微博用户的关注列表和粉丝列表。每次获取结果以文件名为 Friendship-id 的文档存储。同时将每次运行时,遍历的用户的 id 写入 Nodes-id。其中 id 均为用户的账号。

在社区划分前需要先对这些数据进行预处理。首先遍历 Nodes 文件,去掉重复的 id,每个 id 对应的微博用户详细信息对应一个 friendship-id 文件。再根据遍历 Nodes 文件得到的微博用户 id 列表,读取对应的 Friendship-id 文件。

由于微博上存在一些恶意注册的僵尸粉,和很少使用的用户,这两类用户不存在实验分析价值,需要删除。对于有效用户,根据关注列表和粉丝列表,查找这个 id 的用户与其他用户,如果两个用户有关系,则记为这两个 id 的节点之间有边相连。遍历完后,得到用户关系原始网络。

3.2 实验结果

对于获取的数据,这里选取 26 个节点 33 条边的网络、369 个点 1 437 条边的网络和 1 892 个点 19 965条边的网络进行测试。实验在 intel i-5 @2.30 GHz 2.30 GHz 的处理器,4G 内存的联想笔记本电脑上进行实验。

在 369 个节点网络中,节点的度数相差非常大,其中度数最高的为 259 号节点,度数为 124,有 61 个节点的度数节点为 1。259 号节点可能正好是这个网络中的中心节点用户,这类用户的活跃度和知名度均比较大,表现在微博上是粉丝数目很多,微博数量也比较大,发表的微博内容比较有内涵和深度,被转

发和被评论的数目多。也有可能是这个节点的用户本身是名人,这类用户即使是仅发表少量微博也有很多粉丝,表现在微博中是粉丝数目比较大,微博数量比较少。

经过算法的划分,选用的三个网络分别被划分成了 4 个、19 个、43 个社区。其中 26 节点网络被划分成为四个社区,每个社区的社区内成员都是微博上由某个共同话题聚集在一起的用户,他们之间的关系或为粉丝,或为关注,或为互粉,联系紧密。而任意两个社区间的用户没有明显的关系,联系较弱。

除了获取的数据,本文还利用了著名的空手道俱乐部网络进行实验,划分的时间如表 1 所示。

表 1　算法时间对比

Tab. 1　The contrast of the working time between the two algorithms

算法	26 个节点网络/ms	空手道俱乐部/ms
GN 算法	286	342
本文的算法	194	239

如表 1 所示,改进后的算法在时间性能上有所提高。这是因为采用模块度增量的概念划分社区的过程中会在到达近似最优解的过程中就退出,减少了划分的时间。

综上所述,改进后的 GN 算法是比较适用于类似微博这样的社交网络上的社区发现的,利用模块度增量可以提高算法的运算速度。

4　结　语

本文利用微博用户联系的数据,构建用户关系网络,用 GN 算法对社交网络进行社区划分,同时采用了模块度增量来提高算法的运算效率。实验证明,相较于文献 [7] 中的 GN 算法,引入了模块度增量在一定程度上可以加快算法的运算速度,由此也可以推广到一般的社交网络进行社区划分问题。为大规模复杂信息网络研究、网络节点区域性特征的研究等提供了一个基础性的作用。

算法的缺点在于,极少部分网络若产生的第一个模块度极值不是模块度最值得情况下,得到的不是最优解。因此,如何解决这个问题,是今后工作的重点。

参考文献:

[1] NEWMAN M E J. Communities, modules and large-scale structure in networks [J]. Nature Physics, 2012,8(1):25-31.

[2] 朱小虎,宋文军,王崇骏,等. 用于社团发现 Girvan-Newman 改进算法 [J]. 计算机科学与探索,2010,4(12):1101-1108.

[3] KERNIGHAN B W, LIN S. An efficient heuristic procedure for partitioning graphs [J]. Bell System Technical Journal, 1970, 49: 291-307.

[4] FIEDLER M. Algerbraic connectivity of graphs [J]. Czech Math, 1973,23(98):298.

[5] POTHEN A, SIMON H, LIOU K P. Partitioning sparse matrices with eigenvectors of graphs [J]. SIAM Journal on Matrix Analysis and Applications, 1990, 11(3): 430-452.

[6] NEWMAN M E J. Fast algorithm for detecting community structure in networks [J]. Physical Review E, 2004, 69(6): 41-53.

[7] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks [J]. Physical Review E, 2004, 70(6):135-142.

[8] 汪小帆,李翔,陈关荣. 复杂网络理论及其应用 [M]. 北京: 清华大学出版社,2006: 171-175.

[9] GREGORY S. An algorithm to find overlapping community structure in networks [C] //Knowledge Discovery in Databases: PDKK 2007. Berlin, Germany: Springer Berlin Heidelberg, 2007: 91-102.

[10] 丁苣,涂浩. 微博感知突发重大新闻事件的研究与分析 [J]. 广西大学学报: 自然科学版,2011,36(S1):335-338.

[11] 姜秀芳. 面向复杂网络的社区发现算法研究 [D]. 合肥: 中国科学技术大学,2011.

[12] NEWMAN M E J. Communities, modules and large-scale structure in networks [J]. Nature physics, 2012,8(1):25-31.

[13] YE Z S, HU S N, YU J. Adaptive clustering algorithm for community detection in complex networks [J]. Physical Review E, 2008,78(4):046115.

[14] 淦文燕,赫南,李德毅,等. 一种基于拓扑势的网络社区发现方法 [J]. 软件学报,2009, 20(8):2241-2254.

(责任编辑　梁碧芬)