软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2014,25(12):2808−2823 [doi: 10.13328/j.cnki.jos.004720]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

# 基于 R-C 模型的微博用户社区发现*

周小平[1,2,3], 梁 循[1], 张海燕[1]

[1](中国人民大学 信息学院,北京 100872)

[2](北京建筑大学 电气与信息工程学院,北京 100044)

[3](北京市建筑安全监测工程技术研究中心,北京 100044)

通讯作者: 梁循, E-mail: xliang@ruc.edu.cn

**摘 要**: 在微博市场营销、个性化推荐等应用中,发现兴趣和网络结构双内聚的用户社区起着至关重要的作用.现阶段,绝大多数的用户社区发现算法往往将用户联系与用户内容相隔离,从而导致其社区发现结果不够合理,而少数综合用户联系和内容的用户社区发现算法较为复杂;LCA 算法是重叠社区发现算法中算法效率较高且社区质量较好的算法,然而,其在聚类时未考虑边的真实兴趣体现.针对这些问题,构建了以关注关系为网络节点、以关注关系之间是否有共同用户为关注关系潜在的边、以关注关系所关联用户的兴趣集的交集为关注关系的兴趣特征,构建微博网络 R-C 模型,并探讨了其进行微博用户社区发现的方法,分析了该方法的复杂度.最后,以新浪微博数据集为实验,对照节点 CNM 算法和 LCA 算法,从兴趣内聚和网络结构内聚两方面进行分析,发现该方法能够发现更好的微博用户社区.

**关键词**: 微博;社区发现;关注关系;重叠社区

**中图法分类号**: TP311

中文引用格式: 周小平,梁循,张海燕.基于 R-C 模型的微博用户社区发现.软件学报,2014,25(12):2808−2823. http://www.jos.org.cn/1000-9825/4720.htm

英文引用格式: Zhou XP, Liang X, Zhang HY. User community detection on micro-blog using R-C model. Ruan Jian Xue Bao/ Journal of Software, 2014,25(12):2808−2823 (in Chinese). http://www.jos.org.cn/1000-9825/4720.htm

## User Community Detection on Micro-Blog Using R-C Model

ZHOU Xiao-Ping[1,2,3], LIANG Xun[1], ZHANG Hai-Yan[1]

[1](School of Information, Renmin University of China, Beijing 100872, China)

[2](School of Electrical & Information Engineering, Beijing University of Civil Engineering & Architecture, Beijing 100044, China)

[3](Beijing Engineering Research Center of Monitoring for Construction Safety, Beijing 100044, China)

Corresponding author: LIANG Xun, E-mail: xliang@ruc.edu.cn

**Abstract**: Detecting user communities with denser common interests and network structure plays an important role in target marketing and self-oriented services. User-Generated content and the relationship between the users are often separated in the current methods on community detection, which results in the unreasonable community structures. Though some methods tried to combine the two factors, they are complex. Link community algorithm (LCA) is an efficient state-of-art method on overlapping community discovery. However, LCA does not take into account the real interest characteristics when calculating the similarity between the links. To solve the issues on user community detection on Micro-blog, this paper proposes a R-C model which takes the user relationships as the network nodes, treats the intersection of the interest characteristics of the two users in a link as the link's interest characteristics, and makes the shared user between two links as the underlying link between the links. Also, the community detection method based on the R-C model is discussed,

* 基金项目: 国家自然科学基金(71271211); 北京市自然科学基金(4132067); 中国人民大学自然科学基金(10XNI029); 北京高等学校青年英才计划(21147514040)

收稿时间: 2013-11-13; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

and the complexity in clustering is analyzed. Finally, compared with node CNM and LCA, the method using R-C model is proved to be better in finding closer relationship and denser common interest user communities.

**Key words**: micro-blog; community detection; following relationship; overlap community

社区发现是指在社会网络中发现内聚的子群.社区发现是社会网络分析的重要问题,它有助于人们进一步认识、理解和掌握所研究的复杂网络对象,进而实现更深入的应用研究,例如个性化推荐[1]、朋友推荐[2]、大规模网络压缩求解[3]、异质网络分析[4]、社会网络演变[5]等.兴趣和网络结构双内聚的用户社区发现,是精准的市场营销和准确的个性化推荐服务等的重要研究内容[6−8].现实生活中,人们往往传播其所能接触到的感兴趣的信息.因此,好的用户社区发现应同时满足网络结构和兴趣双方面的内聚.网络结构是社区内部节点间信息传播的桥梁,兴趣是信息传播的原因.

得益于移动互联网的发展,微博用户规模及其社会影响力迅速增长.Twitter 有不少于 5 亿的注册用户,每月活跃用户为 2.3 亿,而日活跃用户为 1 亿,每天推文 5 亿次(http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city).新浪微博也拥有超过 5 亿的注册用户,每天有高达 4.62 千万的活跃用户和不少于 1 亿的微博(http://news.xinhuanet.com/ info/2013-02/21/c_132181760.htm).微博是现实社会的缩影,它为人们提供了巨量的有价值的研究数据.人们使用微博进行政治[9]、市场营销[10]等活动,微博已成为一个公认的发表意见与看法的平台[11].

目前,针对微博用户社区发现的方法大致可分为 3 种:

(1) 基于用户内容[12−16].将用户微博内容进行兴趣特征提取,然后,基于兴趣特征进行用户聚类.该类方法忽略了微博网络结构(关注关系)在信息传播中的桥梁作用;

(2) 基于用户联系[17−25].提取微博网络的关注或好友关系,将问题转化为图论等问题进行社区发现.该类方法没有考虑用户的兴趣特征,因此,无法证明其兴趣的内聚性;

(3) 综合方法[26,27].将微博内容和用户联系相结合,基于内容提取基于兴趣的用户社区,基于用户联系提取基于联系的用户社区,再采用某种方法将两个社区进行融合,形成兴趣和网络结构双内聚的用户社区.该类方法由于需要进行两次社区发现,且需要进行社区融合;因此,算法效率较低.

真实情况中,用户往往对多种兴趣感兴趣,每个用户都应根据其兴趣归属于多个用户社区.因此,用户社区实际上是一个重叠社区.LCA 算法[28]是目前较好的重叠社区发现算法,它以边为单元进行边聚类,从而根据边分属不同的社区,将节点划分到多个不同的社区.LCA 算法较好地平衡了社区发现中兴趣和网络结构双方面的因素,并且其聚类复杂度较低;然而,LCA 算法在边之间的相似度计算上忽略了边的真实兴趣特征.

关注关系是微博网络形成的基础,也是微博信息传播的纽带.关注双方往往因为某个或某几个共同兴趣而建立关注关系.因此,关注关系还体现了关注关系双方的共同兴趣特征.针对现有微博用户社区发现算法的不足,本文以关注关系作为聚类节点,根据用户微博内容提取关注关系的兴趣特征,构建微博网络 R-C 模型,进而根据关注关系的兴趣特征计算关注关系之间的相似性,将问题转化为加权无向网络社区发现问题,解决了现有算法考虑不周全或效率较低等问题.本文的学术贡献主要有:

(1) 提出了微博网络 R-C 模型,并探讨了其进行用户社区发现的方法;

(2) 分析了基于 R-C 模型进行社区发现聚类的时间复杂度;

(3) 以新浪微博为实例,对照节点 CNM 算法和 LCA 算法,分析了基于 R-C 模型的用户社区发现算法所发现的用户社区在兴趣和网络结构上都有更优的内聚性.

本文第 1 节介绍现阶段微博用户社区发现的相关文献.第 2 节详细描述微博网络 R-C 模型及其社区发现方法,并分析使用该方法进行聚类的时间复杂度.第 3 节以新浪微博数据为实验对象,对照节点 CNM 和 LCA 算法,从兴趣内聚和网络内聚两方面验证本文方法能够发现更好的微博用户社区.第 4 节对本文的工作进行总结.

## 1 相关工作

近几年,随着在线网络社区的发展,社区发现算法得到了广泛的研究.针对用户社区的发现,人们已经提出

了许多方法,这些方法主要可以分为 3 类:文本聚类法、网络结构法和综合法.

文本聚类法主要通过计算社区内节点的文本内容的相似性,根据相似性将文本内容相似的节点划分为社区.早在 1999 年,Kleinberg 等人提出了基于内容的网页聚类方法,即著名的 HITS 算法[12].主题模型是文本聚类法最典型的算法.2003 年,Blei 等人提出了 LDA 模型[13],认为文档是多个主题的概率分布.2004 年, Syeyvers 等人认为主题是多个关键词的概率分布,用户也以某种概率分布对多个主题感兴趣,并提出了 AT (author-topic)模型[14],用于发现用户、文档、主题和关键词之间的关系.2007 年,McCallum 等人基于发送-接受关系提出了 ART(author-recipient-topic)模型[15],用于聚类具有相似兴趣的用户.在 ART 模型的基础上,2008 年, Pathak 等人提出 CART(community-author-recipient-topic)模型[16].这些模型都忽略了用户之间显著的关注关系,从而可能导致社区发现结果的不合理.

基于网络结构的社区发现算法是目前较为流行且研究较多的方法,这类方法根据用户之间的相互关系将社区网络划分为社区内联系紧密、社区之间联系稀疏的多个子社区.1970 年,Kernighan 和 Lin 针对图分割问题提出了 KL 算法[17],该算法应用于复杂网络社区发现,就是社区发现图分割法的典型算法.图分割法通过迭代的方式将图分解为最优的两个子图,反复处理,直至得到足够数目的子图.2002 年,Girvan 和 Newman 提出了 GN 算法[18],它通过反复识别和删除网络中边介数最大的连接,实现复杂网络聚类.GN 算法的复杂度较高,但它启发了人们对复杂网络社区发现的思路.2004 年,Newman 和 Girvan 提出的网络模块性评价函数——模块度 $Q$[19].$Q$ 函数为社区内的实际连接数目与随机连接下社区内的期望连接数目之差,它描述了所发现社区的优劣.$Q$ 值越大,社区结构越好.在此基础上,Newman 提出了基于局部搜索的快速复杂网络聚类算法,即快速 Newman 算法[20].快速 Newman 算法通过局部搜索,找到极大化的 $Q$ 值,从而实现社区划分.同年,Newman 等人从算法复杂度的角度出发,通过引入模块度增量矩阵和堆结构,将快速 Newman 算法演进为了 CNM 算法[21].2005 年,Guimera 和 Amaral 以优化目标函数 $Q$ 为目标,提出基于模拟退火(simulated annealing,简称 SA)算法的复杂网络聚类算法——GA 算法[22].SA 的引入,使得 GA 算法具有找到全局最优解的能力,因而,GA 算法具有很好的聚类精度.基于模块度优化的聚合方法是目前比较流行的社区发现算法,并被扩充到了加权网络社区发现[23]、有向网络社区发现[24]和重叠社区发现[25]等.虽然基于网络结构(用户关系)的社区发现算法能够对用户进行聚类,但由于其忽略了用户之间的共同兴趣特征,因此不能保证社区发现的兴趣内聚性.

针对上述两种社区发现在兴趣社区发现上的不足,2012 年,Zhang 等人[26]提出了将用户关系同用户内容进行结合,发现用户社区.他们采用 NMF 方法进行基于用户关系的社区发现,采用 AT 模型用于兴趣社区的发现,并在此基础上将两种社区发现结果进行融合,并在 Tweets 和 Delicious 上进行了验证.燕飞等人[27]首先对个人兴趣进行聚类,得到基于兴趣的行动者社区,然后使用社会网络拓扑结构信息对兴趣社区进行扩展,并在 Flickr 上进行了实验分析.这些方法虽然得到了较好的兴趣社区发现,并能将用户根据其兴趣划分到多个不同的社区,符合实际情况,但其算法逻辑较为复杂,而且复杂度较高.

真实世界中,社区结构大多数都是重叠且具有层次结构[28],微博用户往往具有多样化的兴趣特征,因此微博用户社区发现是重叠社区发现问题.CPM 算法[29]是目前流行的重叠社区算法,其在自然和社会学等领域[30,31]都有所应用,且被推广到了加权网络的重叠社区发现[32].然而 CPM 算法认为社区是强连通的簇,其对社区苛刻的定义使得在稀疏网络(如新浪微博用户联系网络[33,34])中社区发现效果较差.此外,CPM 算法需要指定 $k$ 值,且复杂度较高,制约了 CPM 算法在大数据网络中的运用.2010 年,Ahn 等人提出了边社区概念及其算法——LCA 算法[28],并在生物网络、社会网络和其他代表性网络(哲学家关系网、单词关系网和 Amazon.com 产品联系网)上对照 CPM 算法、Infomap 算法[35]和快速 Newman 算法[20]验证了 LCA 算法能发现质量更好的重叠社区.

LCA 算法以边作为聚类节点,对边进行聚类,并根据边所属的社区,将节点划分到多个不同的社区.在一个具有 $N$ 个节点的加权网络中,LCA 算法假定对于任意节点 $i$ 都有属性向量 $\boldsymbol{a}_i=(\tilde{A}_{i1},...,\tilde{A}_{iN})$,且:

$$\tilde{A}_{ij}=\frac{1}{k_i}\sum_{i'\in n(i)}w_{ii'}\delta_{ij}+w_{ij},$$

其中,$w_{ij}$ 为边 $e_{ij}$ 的权重;$n(i)$为与节点 $i$ 有连接关系的所有邻居节点集合;$k_i$ 为集合 $n(i)$的元素数量;当 $i=j$ 时,$\delta_{ij}=1$,

其他情况为 0.在 LCA 算法中,边 $e_{ij}$ 的权重 $w_{ij}$ 表征具有联系的两个节点 $i$ 和 $j$ 在某种性质上相关度.通常,权重值越高,相关度越大.根据不同的应用,$w_{ij}$ 的具体含义也略有不同,在具体应用中,$w_{ij}$ 可根据社区发现的不同目的和网络的不同特征采用不同的方法进行计算.如:为了发现电影演员之间的协作关系,可以以演员为节点、演员之间是否有合作电影为边、演员间合作的电影数为边的权重构建演员关系网络,此时,$w_{ij}$ 将表示演员间协作程度[28].又如:为了发现内容和结构双内聚的微博用户社区,可以以用户为节点、相互关注关系为边、用户发布内容之间的相似性为边的权重构建微博网络模型,此时,$w_{ij}$ 表示微博用户之间兴趣的相似程度.再如:为了挖掘 Amazon 上不同产品间的关系,可以构建以产品为节点、用户是否同时购买某两种产品为边、产品所包含的用户标签的相似度值为边的权重构建产品网络模型,此时,$w_{ij}$ 表示产品间用户标签的相似程度[28].

在此基础上,LCA 算法采用 Tanimoto 系数计算公式计算具有公共节点 $k$ 的两条边 $e_{ik}$ 和 $e_{jk}$ 之间的相似度.由于边 $e_{ik}$ 和 $e_{jk}$ 具有公共节点 $k$,LCA 算法认为:从网络结构上看,节点 $k$ 的邻居节点对该两条边相似度的贡献不大,边 $e_{ik}$ 和 $e_{jk}$ 的计算只考虑节点 $i$ 和节点 $j$ 的属性向量 $\boldsymbol{a}_i$ 和 $\boldsymbol{a}_j$,忽略节点 $k$ 的属性向量 $\boldsymbol{a}_k$.因此,边 $e_{ik}$ 和 $e_{jk}$ 的相似度计算公式为

$$S(e_{ik},e_{jk})=\frac{\boldsymbol{a}_i\cdot\boldsymbol{a}_j}{|\boldsymbol{a}_i|^2+|\boldsymbol{a}_j|^2-\boldsymbol{a}_i\cdot\boldsymbol{a}_j} \tag{1}$$

在计算边边之间相似度的基础上,LCA 算法采用单边聚类方法[36]对边聚类,直至形成一个社区.最后,采用最优社区密度对层次进行切分,形成多个社区.在微博网络中,边(关注关系)是微博信息传输的纽带,也是其所相互关联的两个用户共同兴趣特征的体现,即,边的真实兴趣特征为其所关联的两个节点共同的兴趣特征.显然,公式(1)在边边的相似度计算上仅从网络结构出发,忽略了边的真实兴趣特征.

图 1 给出了 LCA 算法未考虑边的真实兴趣特性而导致社区发现不准确的一个案例,该案例由 3 个节点 $n_1$,$n_2$,$n_3$ 和两条边 $e_{12}$ 和 $e_{13}$ 组成.节点 $n_1$,$n_2$ 和 $n_3$ 的兴趣特征及其权重分别为($I_1$:0.5,$I_2$:0.5),($I_1$:0.5)和($I_2$:1).采用 Tanimoto 系数计算公式分别求得边 $e_{12}$ 和 $e_{13}$ 的权重 $w_{12}$ 和 $w_{13}$ 为 0.5 和 0.5.进而根据公式(1)可知,边 $e_{12}$ 和 $e_{13}$ 之间的相似度为 0.5.因此若采用 LCA 算法,由于边 $e_{12}$ 和 $e_{13}$ 间较高的相似度,使得 $e_{12}$ 和 $e_{13}$ 将被划分到一个社区,即,节点 $n_1$,$n_2$ 和 $n_3$ 都归属于同一个社区(图 1 左下图所示).而事实上,$n_1$ 和 $n_2$ 的共同兴趣为 $I_1$,$n_1$ 和 $n_3$ 的共同兴趣为 $I_2$,而 $n_2$ 和 $n_3$ 之间无共同兴趣.因此,好的社区发现应能将其划分为 $n_1$,$n_2$ 和 $n_1$,$n_3$ 两个不同的社区结构(如图 1 右下图所示).显然,LCA 算法因未考虑 $e_{12}$ 和 $e_{13}$ 的真实兴趣特征,使得其社区发现不够合理.

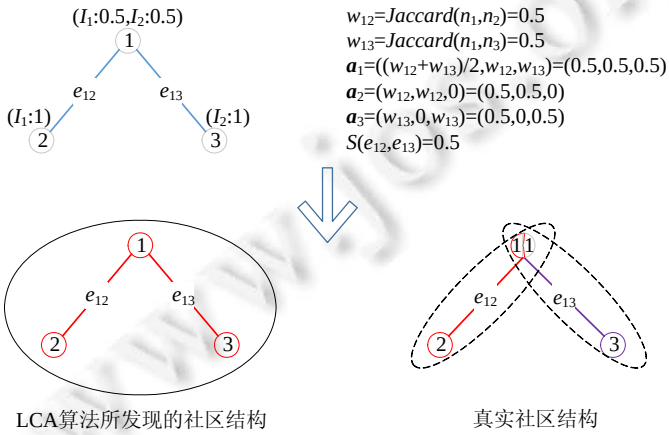



Fig.1 An example of unconsidered case in LCA

图 1 LCA 算法不足案例示意图

本文对微博网络进行 R-C 模型构建,建立以关注关系为节点、以关注关系之间是否有共同用户为边、从用户的微博内容提取用户的兴趣特征、进而转化为关注关系的兴趣特征.在此基础上进行微博用户社区发现,解

决了现有算法存在的考虑不全面、效率较低和 LCA 算法未考虑边的真实兴趣特征等问题.

## 2 微博网络 R-C 模型

本节将针对现有微博用户社区发现算法的不足构建微博网络 R-C 模型,为了更好地进行模型说明,本文有如下定义:

**定义 1(*X* 社区)**. 如果一个社区内的对象为 *X*,则称该社区为 *X* 社区.例如:在微博社区中,其社区是用户的集合,称为用户社区;在一个节点网络中,节点是社区的元素,称为节点社区;在 LCA 算法中,其以边为单位进行聚类,所形成的社区成为边社区.

### 2.1 微博网络R-C模型

一个微博社区的真实内容通常包含 3 部分内容:用户集合 *U*、用户关系集合 *L* 和由 *U* 所产生的各类内容 *T*(主要为微博及其评论内容).因此,一个微博通常可以表示为 *S*=(*U*,*L*,*T*),其中,*S* 表示微博社区.针对不同的研究和应用,该模型略有不同.图 2 下半部分是一个微博网络真实内容及其关系示意图:$U=\{U_1,U_2,U_3\}$为微博用户集合;$L=\{L_1,L_2\}$为用户联系的集合,也是微博内容 *T* 传播的纽带;$T=\{T_1,T_2,T_3\}$为微博内容集合,$T_i$ 为 $U_i$ 的所有微博内容集合.

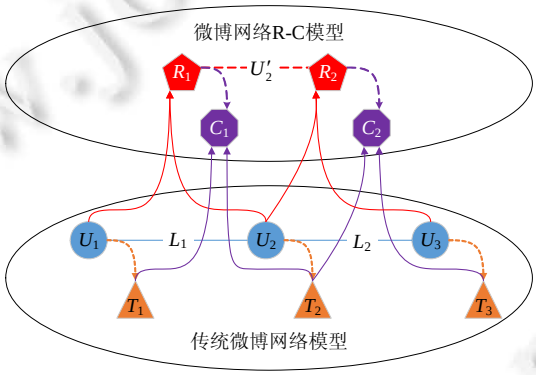



Fig.2 Micro-Blog model, top is R-C model, while bottom is traditional model

图 2 微博模型示意图,上半部分为微博网络 R-C 模型,下半部分为现有微博网络模型

微博用户社区发现即是在微博网络 *S* 中发现 *L* 和 *T* 同时内聚的 *U* 社区.若以 *T* 作为研究对象,采用文本聚类的方法进行社区发现,该方法能够形成兴趣内聚的 *U* 社区;但由于忽略了关系 *L* 的重要作用,不能保证信息在所发现的社区内部能够畅通传播.若以 *L* 作为聚类条件进行 *U* 社区发现,无法保证所形成的社区的兴趣内聚.因此,合理的 *U* 社区发现应综合考虑 *L* 和 *T*.现有的综合方法采用某种方法融合由 *L* 和 *T* 所发现的两类 *U* 社区,形成网络结构和兴趣双内聚的 *U* 社区.先后两次社区发现及社区融合,导致了该类社区发现算法效率较低.而导致该算法需要进行两次社区发现,其最根本的原因是没有充分利用 *L* 的信息和价值.*L* 作为用户之间的相互关系,已经体现了 *U* 的存在,因此在兴趣社区发现中,如果以 *L* 为社区发现对象、以 *T* 作为 *L* 的属性进行 *L* 社区发现,通过一次社区发现找出 *L* 社区,进而转化为 *U* 社区,将能简化社区发现复杂度.

由于双向关注关系(好友关系)更能体现真实社会情况[10],本文所讨论的关注关系是指双向关注关系.图 2 上半部分显示了微博网络 R-C 模型示意图,它将原有模型中的关注关系 $L=\{L_1,L_2\}$映射成网络节点 $R=\{R_1,R_2\}$. $U_2'$是关注关系 $R_1$ 和 $R_2$ 潜在的连接关系,它体现了 $R_1$ 和 $R_2$ 之间存在着共同用户.同时,关注关系 *L* 还潜在着所关联的两个用户之间的共同兴趣特征.微博内容 *T* 是用户兴趣集的具体表现,因此,通过对关注关系所关联的两个用户的微博内容 *T* 进行兴趣特征提取,可进一步获得关注关系的所关联的用户的共同兴趣特征 *C*,实现对 R-C 模型中关注关系兴趣特征的描述,从而将原有微博网络模型转化为 R-C 模型,即 $S=\{R,C\}$.

由于用户往往具有多个不同的兴趣,现有的方法通常根据用户内容计算出用户对各不同兴趣感兴趣的程度.因此,用户兴趣集是一个带权值的兴趣集合.关注关系 $R_x$ 的兴趣特征 $C_x$ 为其所关联的两个用户 $U_i$ 和 $U_j$ 兴趣特征的公共部分,因此本文认为,关注关系 $R_x$ 的兴趣特征 $C_x$ 为 $U_i$ 和 $U_j$ 兴趣的共同兴趣,其兴趣的权值为该兴趣在 $U_i$ 和 $U_j$ 兴趣中权重的较小值.为了计算关注关系的兴趣特征,本文做如下定义.

**定义 2(权值集合)**. 若给定一个集合 $A=\{a_1,a_2,\ldots,a_m\}$,其每个元素都含权值,即第 $i$ 个元素 $a_i$ 的权值为 $w_{ai}$,则称 $A$ 为权值集合.$A$ 又表示为 $A=\{(a_1,w_{a1}),(a_2,w_{a2}),\ldots,(a_m,w_{am})\}$.

**定义 3(权值集合交集)**. 假定权值集合 $A=\{(a_1,w_{a1}),(a_2,w_{a2}),\ldots,(a_m,w_{am})\}$和 $B=\{(b_1,w_{b1}),(b_2,w_{b2}),\ldots,(b_n,w_{bn})\}$,则集合 $A$ 和 $B$ 的交集为:$A\cap B=\{(c,w_c)|c$ 为 $A$ 和 $B$ 的共同元素,若 $c=a_i=b_j$,有 $w_c=\min(w_{ai},w_{bj})\}$,其中,$\min(\cdot)$函数为取最小值.例如,若权值集合 $A=\{(a,1),(b,2),(c,3)\}$,权值集合 $B=\{(b,1),(c,2),(d,3)\}$,则 $A\cap B=\{(b,1),(c,2)\}$.对于无权值的集合,可以假定其各元素的权重值为固定常数,如 1,也可用该定义进行交集运算.在微博网络中,若用权值集合 $I_i$ 表示用户 $U_i$ 的兴趣集,则关联用户 $U_i$ 和 $U_j$ 的关注关系 $R_x$ 的兴趣特征 $C_x$ 可以使用定义 3 进行计算,即:

$$C_x=I_i\cap I_j.$$

在微博网络 R-C 模型的基础上进行 $R$ 社区发现,最后将 $R$ 直接映射为其关联的用户,转化为 $U$ 社区.它在综合考虑用户联系和用户内容的基础上提高了用户社区发现效率,并解决了 LCA 算法在社区发现上没有充分考虑边的兴趣特征的问题.以图 1 为例,在 R-C 模型中,假定边 $e_{12}$ 和 $e_{13}$ 所对应的关注关系为 $R_1$ 和 $R_2$,则不难得出 $R_1$ 和 $R_2$ 所对应的兴趣特征分别为 $C_1=\{(I_1,0.5)\}$和 $C_2=\{(I_2,0.5)\}$.由于 $C_1$ 和 $C_2$ 完全不同,因此不论采用哪种聚类方法,$R_1$ 和 $R_2$ 都分属于不同的兴趣社区,最终发现真实的兴趣社区.

虽然 R-C 模型和 LCA 算法都采用边进行聚类,但两者具有本质的不同,具体表现在:

(1) LCA 算法只是将边作为一个聚类的对象,其边并不具有兴趣特征描述.而 R-C 模型在社区发现上将关注关系作为实体进行聚类.在 R-C 模型中,关注关系不仅仅只是聚类的对象,其还具有其所关联的两个用户的兴趣特征描述.因此,R-C 模型更有利于挖掘内容和结构双内聚的社区结构;

(2) LCA 算法仅仅只是从网络结构的角度出发进行社区发现,且认为两条具有公共节点的边,其公共节点的属性对该两条边的相似度的贡献不大,即,LCA 算法忽略了公共节点的属性特征.因此,LCA 算法忽略了边的真实特征.而 R-C 模型通过对边所关联的两个节点的特征取交集,保留了边的真实特征;

(3) 针对各类型的网络,LCA 算法根据不同的社区发现目标构建加权或无权网络,进而从边的角度出发进行社区发现,各节点的属性特征在构建网络时就已转化为数值.而 R-C 模型首先将关注关系构建为网络节点,并从关注关系所关联的两个用户的兴趣获取该关注关系的特征;接着,根据关注关系的特征计算关注关系间的权重,最后进行社区发现.由于 R-C 模型在进行社区发现前才将属性特征转化为数值,因而能挖掘更为真实的社区结构.

### 2.2 基于R-C模型的用户社区发现方法

微博网络 R-C 模型是一个以关注关系为节点的网络模型,其 $R$ 社区发现可借鉴现有成熟的节点社区发现算法.一般来说,基于微博网络 R-C 模型进行 $R$ 社区发现有两种基本方法:

(1) 将 $R$ 视为孤立的节点,基于兴趣特征 $C$ 之间的相似度进行聚类,实现 $R$ 社区发现.由于 $R$ 自身为 $U$ 社区的边结构,因此在一定程度上能够解决文本聚类方法在 $U$ 社区发现中网络结构内聚不足的问题;但由于忽略了边边之间潜在的关联关系,因此其所发现的社区往往网络结构内聚不足;

(2) 根据 $R$ 之间是否有共同节点建立 $R$ 之间的连接关系,接着,根据 $R$ 的兴趣集计算相互连接的 $R$ 之间的相似度,并将该相似度设为 $R$ 之间连接关系的权重值,建立加权无向 $R$ 网络,将问题转化为加权无向网络的社区发现问题进行求解.该方法在兴趣聚类的同时,更好地考虑了网络结构问题,因此,本文使用该方法进行问题求解.

图 3 为使用 R-C 模型进行微博网络 $U$ 社区发现的基本框架,大致可以分为两个阶段:微博网络 R-C 模型构建和社区发现.本文使用 LDA 模型[13]从微博内容 $T$ 提取用户兴趣集 $I=\{I_1,I_2,\ldots\}$,进而通过交集运算,计算关注关系的兴趣特征集 $C$.关注关系的兴趣特征集 $C$ 和关注关系集合 $R$ 构成微博网络 R-C 模型.接着,本文通过计算有

潜在联系的关注关系之间的兴趣相似度,将微博网络 R-C 模型转换为加权无向网络,并使用较为成熟的加权无向网络社区发现算法进行 *R* 社区发现.由于 CNM 算法的聚类复杂度较低,本文使用加权 CNM 算法进行 *R* 社区发现.最后,将 *R* 直接映射为相应的 *U*,形成 *U*.

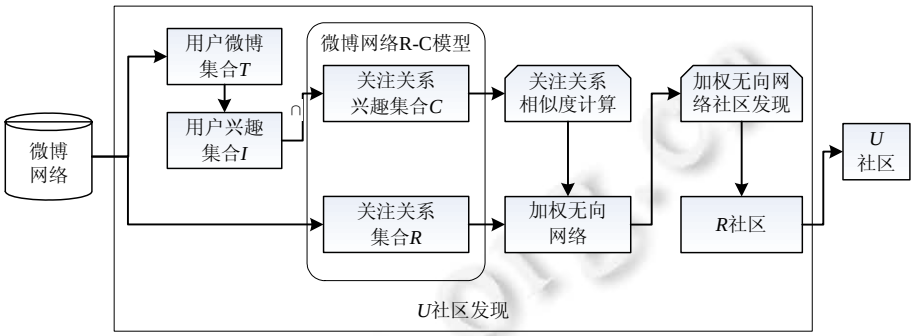



Fig.3 Framework of user community detection using R-C model

图 3 基于 R-C 模型的用户社区发现方法框架

具体地,使用微博网络 R-C 模型进行社区发现的方法步骤如下:

(1) 微博内容 *T* 集合构建.将微博内容依据其所属的用户进行归类,形成 *T* 集合;

(2) 用户兴趣集 *I* 计算.对 *T* 集合中的微博内容进行分词,并采用相关模型(如 LDA 模型等)构建用户兴趣集合 *I*;

(3) 关注关系特征集 *C* 计算.依据关注关系所对应的两个用户兴趣特征集,使用定义 3 所描述的方法,取交集形成关注关系兴趣特征集 *C*;

(4) 关注关系相似度计算.对于无共同用户的关注关系,将不进行相似度计算.对于有公共用户的两个关注关系,采用 Tanimoto 系数计算公式计算其的相似度.计算公式如下:

$$Sim(R_1,R_2)=\frac{C_1\cdot C_2}{|C_1|^2+|C_2|^2-C_1\cdot C_2};$$

(5) *R* 社区发现.采用加权无向网络社区发现算法(如 CNM 算法等)对上述网络进行 *R* 社区发现;

(6) *U* 社区形成.R-C 模型中,任意 *R* 都包含两个具有相互关注关系的用户.对于某个 *R* 社区,其所包含的所有 *R* 所对应的用户集形成该 *R* 社区所对应的 *U* 社区.依次遍历所发现的所有 *R* 社区,形成 *U* 社区.

### 2.3 算法复杂度分析

**定理 1**. 在一个有 *m* 个节点和 *n* 条边的图中,将边转化为节点,节点转化为边,可形成一个包含 *n* 个节点和 $\frac{1}{2}\sum_{i=1}^{m}L_i^2-n$ 条边的图,其中,$L_i$ 为第 *i* 个节点的度数,且有 $\sum_{i=1}^{m}L_i=2n$.

证明:

(1) 将原始图中的边转化为节点后,原始图的 *n* 条边自然形成转化后图的 *n* 个节点;

(2) 以图中某个节点作为考察对象,若图中节点 *i* 的度数为 $L_i$,则其所关联的 $L_i$ 条边可形成 $C_{L_i}^2$ 种不同的两两组合,在转换后的图中可形成 $C_{L_i}^2$ 条边.遍历原始图中所有的节点,易知转化后图的边数为 $\sum_{i=1}^{m}C_{L_i}^2$ .又由于一条边关联两个节点,因此在上述遍历过程中,每条边遍历 2 次,故有 $\sum_{i=1}^{m}L_i=2n$ .展开 $C_{L_i}^2$ ,有:

$$\sum_{i=1}^{m}C_{L_i}^2=\sum_{i=1}^{m}\frac{L_i(L_i-1)}{2}=\sum_{i=1}^{m}\frac{L_i^2-L_i}{2}=\frac{1}{2}\sum_{i=1}^{m}L_i^2-n.$$

考虑某节点的度数为 1 的特殊情况,此时,围绕该节点的边之间可形成 $C_1^2=\frac{1\times(1-1)}{2}=0$ 种边边关系,符合实际情况. □

假定在一个有 $m$ 个用户和 $n$ 条关注关系的微博社区中,将其转化为微博网络 R-C 模型后,可形成 $n$ 个节点、$\frac{1}{2}\sum_{i=1}^{m}L_i^2-n$ 条边的无向网络.若所使用的社区发现算法的复杂度为 $O(f(m,n))$,则转化后,其复杂度为

$$O\left(f\left(n,\frac{1}{2}\sum_{i=1}^{m}L_i^2-n\right)\right).$$

案例 1:若使用 CNM 算法进行 $R$ 社区发现,CNM 算法的时间复杂度为 $O(f(m,n))=O(nd\log m)$[21],$d$ 为图的深度.转换后的算法复杂度为 $O\left(f\left(n,\frac{1}{2}\sum_{i=1}^{m}L_i^2-n\right)\right)=O\left(\left(\frac{1}{2}\sum_{i=1}^{m}L_i^2-n\right)d\log n\right)$. 特殊地,在稀疏网络(如新浪微博[33,34])中,有 $m\sim n$,$d\sim\log m$,此时有 $O(f(m,n))=O(m\log^2 m)$,故而,其复杂度为

$$O\left(f\left(n,\frac{1}{2}\sum_{i=1}^{m}L_i^2-n\right)\right)=O(n\log^2 n)=O(m\log^2 m).$$

该时间复杂度基本等同于基于节点的时间复杂度.

案例 2:若使用 LCA 算法中所用的单边聚类算法,单边聚类算法的时间复杂度为 $O(f(m,n))=O(m^2)$[37].转换后,算法复杂度为 $O\left(f\left(n,\frac{1}{2}\sum_{i=1}^{m}L_i^2-n\right)\right)=O(n^2)$. 相应地,在新浪微博等稀疏网络中,由于 $m\sim n$,因此其时间复杂度为 $O\left(f\left(n,\frac{1}{2}\sum_{i=1}^{m}L_i^2-n\right)\right)=O(n^2)=O(m^2)$.

## 3 实验及分析

### 3.1 实验数据

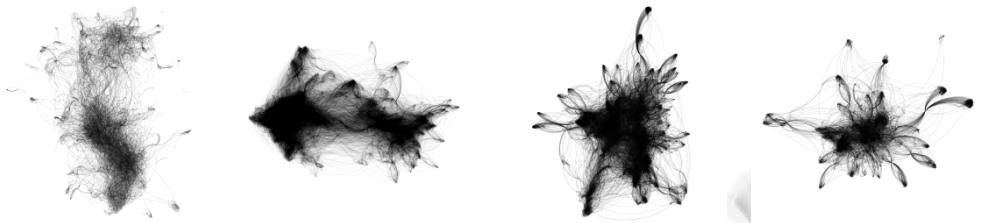
本文实验通过新浪微博开放平台提供的 API 抓取新浪微博数据进行.通过从种子用户出发,采用广度优先遍历的原则,沿关注关系方向逐层抓取新浪微博用户数据和微博内容数据.由于用户的兴趣可能随时间不断变化,因此本文实验只抓取每个用户最新的 200 条微博.由于好友数太少可能是僵尸粉,好友数太多可能是组织机构账户,因此本文随机从微博网络中选取好友数在 50~100 之间的 4 个不同用户作为种子用户,并分别抓取微博用户、用户双向关注关系及用户的最新 200 条微博,形成 4 组数据集.该 4 组数据集的数据信息见表 1.

**Table 1** Four datasets for the experiments

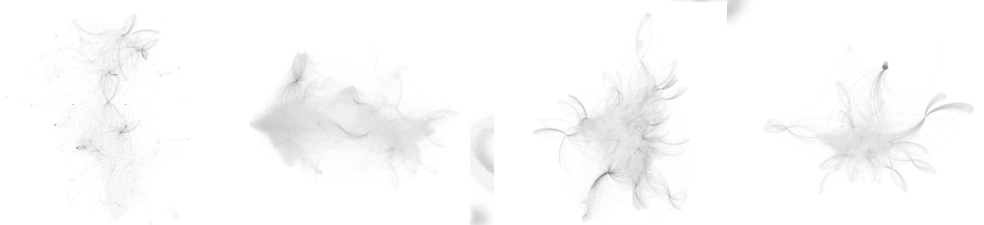
**表 1** 4 组实验数据集

| 数据集 | 用户数 | 关注关系数 | 微博数 | 抓取时间 |
|---|---|---|---|---|
| *S*1 | 6 502 | 28 909 | 872 743 | 2013 年 5 月 |
| *S*2 | 8 734 | 84 311 | 1 476 810 | 2014 年 8 月 |
| *S*3 | 5 649 | 40 284 | 939 194 | 2014 年 8 月 |
| *S*4 | 2 853 | 17 044 | 466 379 | 2014 年 8 月 |

图 4 为所抓取的 4 组数据的微博网络结构图.图 4(a)从左往右分别为数据集 *S*1,*S*2,*S*3 和 *S*4 的无权重下的网络基本结构图.由图可知:该 4 组数据所形成的网络的节点都是相互连通的,不存在孤立的节点和节点群.图 4(b)从左往右分别为数据集 *S*1,*S*2,*S*3 和 *S*4 按度进行着色后的网络结构分布(节点度数越高,其颜色越深).该图说明了在新浪微博网络中,大量的节点度数较低,少数节点度数较高.图 5 为该 4 组数据的微博网络节点的度分布情况,它进一步揭示该网络的度数分布特征.由图 5 可知,所进行实验的 4 组数据所形成的微博网络节点的度基本都呈现幂律分布,因此,所进行实验的 4 组数据集所形成的微博网络都是无标度网络[38].



(a) 无权重的网络基本结构图



(b) 按度大小进行着色后的网络结构图

Fig.4 Micro-Blog network structure of the four datasets

图 4 微博网络结构图

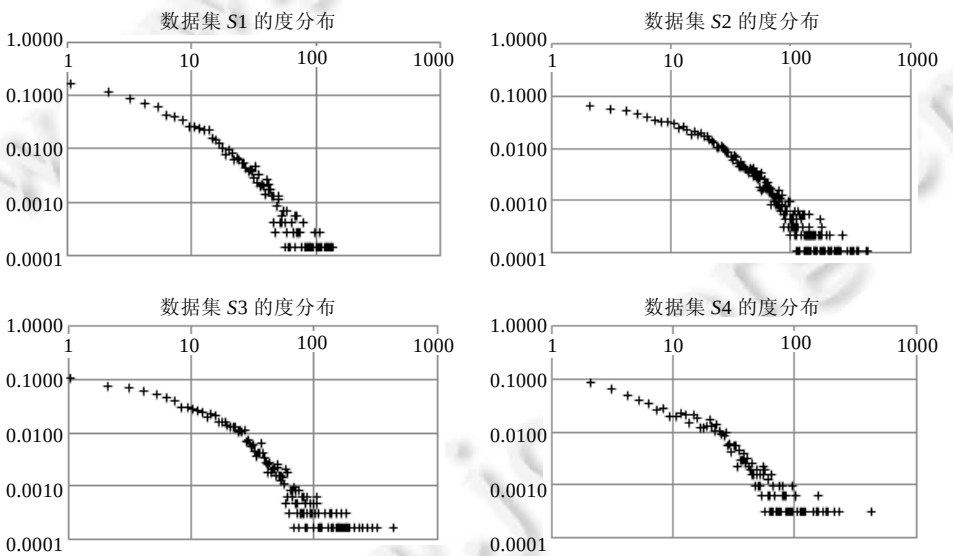



Fig.5 Degree distribution of micro-blog networks of the four datasets

图 5 微博网络的度分布

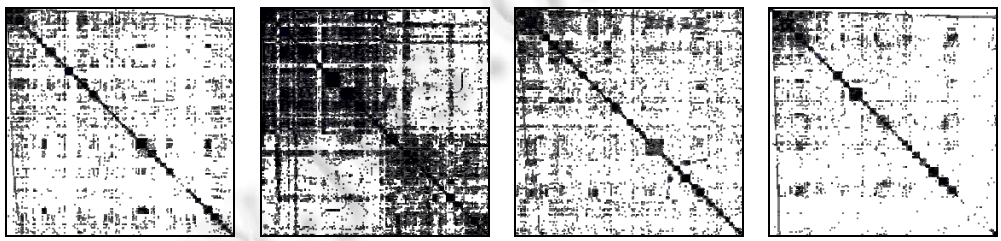
### 3.2 实验结果

由于 LCA 算法采用 Tanimoto 系数计算公式计算边边之间的相似度,为了更好地进行实验对照,本实验采用 Tanimoto 系数计算公式计算节点及关注关系的兴趣集间的兴趣相似度.由于 CNM 算法在加权无向网络社区发现中具有较好的效率,本文采用加权 CNM 算法对关注关系进行聚类.

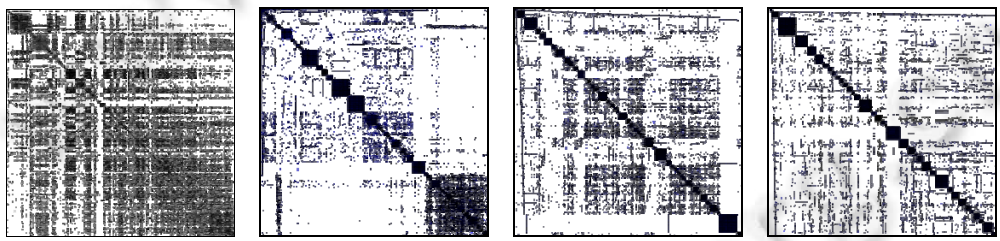
CNM 算法是目前较为流行的社区发现算法,并被广泛引用,且本文使用加权 CNM 算法[23]进行 $R$ 社区发现,因此,直接用于 $U$ 社区发现的加权 CNM 算法(后文称为节点 CNM 算法)将作为本文的对照实验之一,以更好地分析使用 R-C 模型进行社区发现的优势.LCA 算法[28]作为最新的重叠社区发现算法,且其模型与 R-C 模型最为

接近,因此也作为本文的对照算法.在本文实验中,节点 CNM 算法和 LCA 算法的边权重都采用 Tanimoto 系数计算公式进行计算.

为了更直观地显示实验网络中用户之间以及关注关系之间的兴趣相似度情况,本文以图形的形式描绘了这两个相似度矩阵.图 6(a)从左往右分别为数据集 *S*1,*S*2,*S*3 和 *S*4 用户兴趣集的相似度矩阵,图 6(b)从左往右分别为数据集 *S*1,*S*2,*S*3 和 *S*4 关注关系的兴趣相似度矩阵.在相似度矩阵图中,点的颜色越深,则与其对应的一对用户或关注关系拥有更高的兴趣相似度.从 4 组数据的用户兴趣集相似度矩阵和关注关系兴趣集相似度矩阵可以看出:用户兴趣集相似度矩阵和关注关系兴趣相似度矩阵较为显著的特点在于沿对角线方向有一条较明显的黑线,即对角线周围的用户相似度和关注关系相似度较高.由于用户兴趣集矩阵按用户的广度优先遍历所形成的层次关系有序排列,关注关系兴趣相似度矩阵按节点有序排列,因此,较为明显的对角线特征体现了节点与其邻居及同一节点的多条边之间往往具有较高的相似度.在微博网络中,它说明了用户与其邻居用户及同一用户的多个关注关系之间存在较大的相似性,同时也佐证了人以群分的观点.



(a) 4 组数据集用户兴趣集相似度矩阵图



(b) 4 组数据集关注关系兴趣集相似度矩阵图

Fig.6 Similarity matrixes of user interest sets and relationship interest sets

图 6 用户兴趣集相似度矩阵和关注关系兴趣集相似度矩阵

经社区发现,本文方法、节点 CNM 算法和 LCA 算法在 4 组数据集中的社区发现统计见表 2.

**Table 2** Community size distribution of the three methods

**表 2** 3 种算法的社区划分大小分布

| 社区大小 | *S*1 | | | *S*2 | | | *S*3 | | | *S*4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 本文方法 | 节点 CNM | LCA 算法 | 本文方法 | 节点 CNM | LCA 算法 | 本文方法 | 节点 CNM | LCA 算法 | 本文方法 | 节点 CNM | LCA 算法 |
| [100,+∞) | 18 | 13 | 4 | 14 | 11 | 15 | 13 | 10 | 5 | 9 | 7 | 2 |
| [10,100) | 40 | 28 | 119 | 34 | 26 | 474 | 35 | 33 | 345 | 23 | 22 | 124 |
| [3,9) | 44 | 26 | 1 412 | 20 | 16 | 4 073 | 14 | 118 | 2 580 | 6 | 69 | 691 |
| 2 | 630 | 45 | 7 064 | 845 | 16 | 19 545 | 1 098 | 9 | 9 776 | 457 | 8 | 2 019 |
| [2,+∞) | 732 | 112 | 8 599 | 913 | 69 | 24 107 | 1 160 | 170 | 12 706 | 495 | 106 | 2 836 |

本文算法和节点 CNM 算法所划分的社区数量相差不大,主要因为两者采用了相同的聚类算法.在 4 组实验数据集中,本文算法所发现的社区数量都要多于节点 CNM 算法,主要是因为本文算法采用边进行聚类.一方面,在微博网络中边数通常为点数的数倍(例如,数据集 *S*1 的边数为点数的 4 倍左右);另一方面,本文算法所发现的社区为重叠社区,而节点 CNM 算法挖掘非重叠社区.因此,本文算法的社区数量多于节点 CNM 算法.在 4 组数

据集中,LCA 算法所划分的社区数量都最多.此外,从 4 组实验结果可知:本文算法和 LCA 算法都产生了大量的零碎社区(2 个节点、1 条边的社区),且零碎社区的比重都比较大.以数据集 *S*1 为例,本文算法有 630 个零碎社区和 102 个非零碎社区,而 LCA 算法零碎社区和非零碎社区的数量分别为 7 064 和 1 535,两者所发现的社区有 85%左右都为零碎社区.这些零碎社区是在切割条件(最大社区密度或最大模块度)下没有被聚类产生的,也即,微博网络中存在着大量相似度较低的边边关系.而大量的相似度较低的边边关系也使得 LCA 算法在微博网络用户社区发现中的不足更加明显.

图 7 为 3 种社区发现算法在 4 组数据集中所发现的最大的 20 个社区大小分布图.相较而言,本文算法和节点 CNM 算法由于都采用了 CNM 方法进行聚类,其能发现更多用户数较大的社区结构;而 LCA 算法采用单边聚类算法进行聚类,其次大社区和最大社区的大小比值在 0.5 左右,落差较大;而这与 Ahn 等人所述的"LCA 算法中,较优社区发现情况下,次大社区和最大社区的大小比值趋近于 0.5[28]"基本吻合.此外,由于本文算法和 LCA 算法都以边为聚类对象并发现重叠社区,因此所发现的社区相对较大.

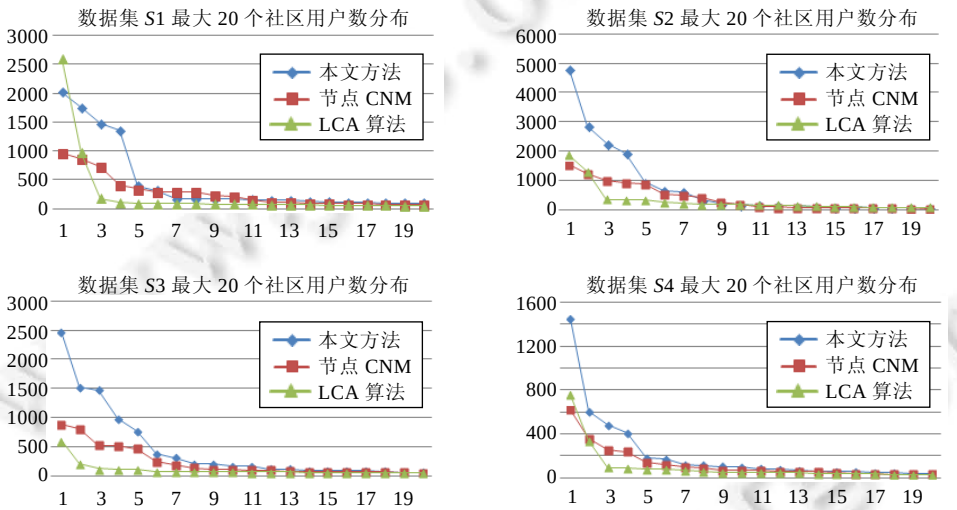



Fig.7 Distribution of the number of users on top rank 20 user communities

图 7 3 种算法所发现的最大的 20 个社区用户数分布

### 3.3 兴趣内聚分析

兴趣内聚是指兴趣社区应具有高度兴趣聚集的特性.为了更好地描述兴趣内聚特性,本文作如下定义:

**定义 4(兴趣内聚指数 *E*)**. 兴趣内聚指数用于描述社区内的用户之间的兴趣相似度与整个社区用户相似度之间的比值.若微博网络中任意两个用户 $U_i$ 和 $U_j$ 的兴趣相似度表示为 $\mu(i,j)$,则兴趣内聚指数 $E$ 可以表示为

$$E=\frac{\sum_{\text{同一子社区内的所有点对}i,j}\mu(i,j)}{\sum_{\text{网络中所有点对}i,j}\mu(i,j)}.$$

显然,$E$ 值越高,表明所发现的社区总体具有更好的兴趣内聚性,用户社区发现算法越优越.虽然 $E$ 值描述了社区发现算法的总体兴趣内聚特征,但却无法描述单个社区的兴趣内聚性.因此,本文定义兴趣平均指数用于描述单个社区的兴趣内聚情况.

**定义 5(兴趣平均指数 *e*)**. 兴趣平均指数用于描述社区内节点对社区总相似度的平均贡献值.若社区 $C$ 有 $|C|$ 个用户,则社区兴趣平均指数 $e$ 可以表示为

$$e_C=\frac{\sum_{i,j\in C}\mu(i,j)}{|C|}.$$

易知:$e$ 越高,社区内节点对社区总兴趣相似度平均贡献越大,社区兴趣内聚性越好.

本文采用 TF-IDF 模型构建用户兴趣集向量,并采用余弦公式计算用户之间的兴趣相似度.经计算,实验所用的 4 组数据集的总相似度分别为 365 911.1,470 122.9,174 433.3 和 49 293.2.4 组数据所得社区的兴趣内聚指数见表 3.由表 3 可知:在 4 组数据集实验中,本文算法所划分的社区,其社区内部的总相似度值都最高,分别为 135 393.2,329 698.1,78 055.0 和 20 674.3,相应的兴趣内聚指数分别为 0.370,0.701,0.447 和 0.419.在数据集 *S*1 和数据集 *S*4 中,LCA 算法的兴趣内聚指数要高于节点 CNM 算法.究其原因,主要在于 CNM 算法是非重叠社区发现算法,它使得一个节点只能归属于一个社区,从而漏失许多相似兴趣.而在数据集 *S*2 和 *S*3 中,LCA 算法虽然发现重叠社区,然而其兴趣内聚指数都要低于 CNM 算法,其主要由于 LCA 在进行边相似度计算时忽略了边的真实兴趣特征和公共节点.由实验可知,本文算法在兴趣内聚指数上要明显优于节点 CNM 算法和 LCA 算法.

**Table 3** *E* values of the three methods

**表 3** 兴趣内聚指数 *E*

| 算法 | *S*1 | | | *S*2 | | | *S*3 | | | *S*4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 本文方法 | 节点 CNM | LCA 算法 | 本文方法 | 节点 CNM | LCA 算法 | 本文方法 | 节点 CNM | LCA 算法 | 本文方法 | 节点 CNM | LCA 算法 |
| $\sum_{同一子社区的所有点对i,j}\mu(i,j)$ | 135393.2 | 39337.4 | 99563.6 | 329698.1 | 60681.8 | 27711.2 | 78055.0 | 17201.2 | 4738.9 | 20674.3 | 5280.6 | 6225.9 |
| $\sum_{网络中所有点对i,j}\mu(i,j)$ | 365911.1 | | | 470122.9 | | | 174433.3 | | | 49293.2 | | |
| 兴趣内聚指数 *E* | 0.370 | 0.108 | 0.272 | 0.701 | 0.129 | 0.059 | 0.447 | 0.099 | 0.027 | 0.419 | 0.107 | 0.126 |

图 8 显示了 3 种算法中在 4 组数据集中最大的 20 个社区的兴趣相似度贡献.由图可知:社区用户数越多的社区,其对社区总体相似度的贡献越大,从而说明了 3 种聚类算法在兴趣聚集上都是有效的.在 4 组数据集中,本文算法主要社区的兴趣贡献度都优于节点 CNM 算法和 LCA 算法.在数据集 *S*1 中,本文算法挖掘的最大社区兴趣贡献度略小于 LCA 算法,主要是因为 LCA 算法最大的社区用户数要明显多于本文算法的最大社区.

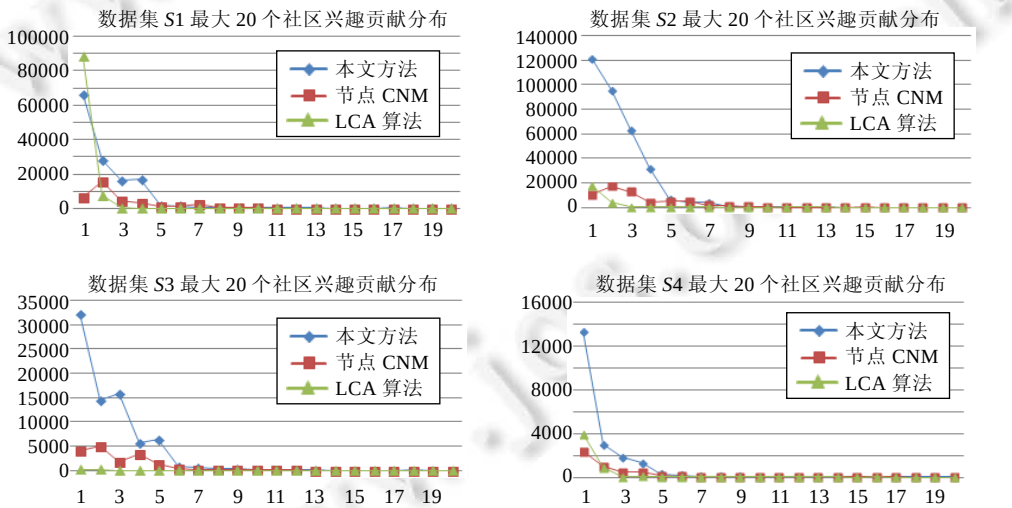



Fig.8 Interest contribution of top rank 20 user communities

图 8 3 种算法最大的 20 个社区内部相似值贡献

图 9 为 3 种算法在 4 组数据集中最大 20 个社区的社区兴趣平均指数,该图直观描述了在最大 20 个社区中,3 种算法的兴趣平均指数对照情况.显然,本文方法在单个社区的兴趣平均指数上基本都优于节点 CNM 和 LCA 算法.因此,本文方法所发现的用户社区具有更好的兴趣内聚性.

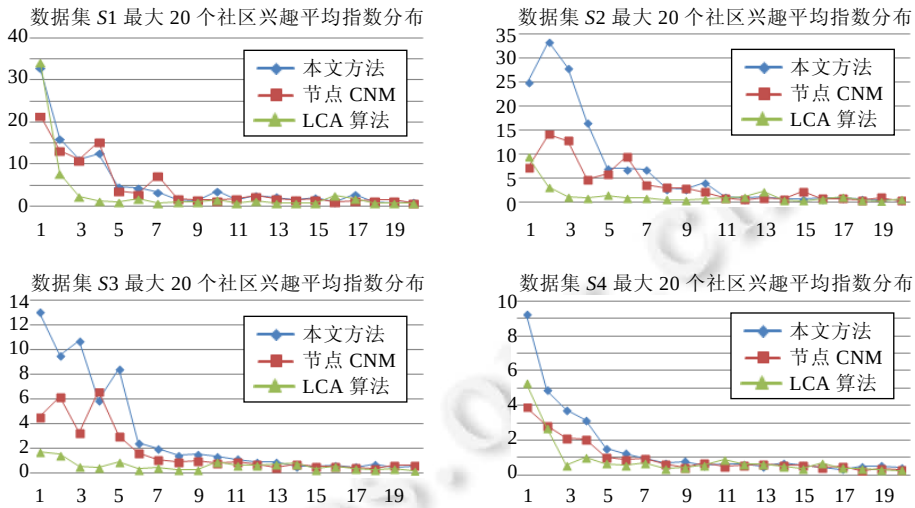



Fig.9 Average interest values *e* of top rank 20 user communities

图 9 3 种算法最大 20 个社区兴趣平均指数 *e* 值

### 3.4 网络结构内聚分析

网络结构内聚是指兴趣社区除了应具有高度兴趣凝聚外,其网络结构也应高度内聚.社区密度是 LCA 算法进行社区层次切分的标准,但其并不是较好的网络结构评价指标,且 LCA 算法本身都不以社区密度作为社区结构评价的指标,因此,本文也不将社区密度作为网络结构的评价指标.

模块度是目前评价网络节点凝聚度的常用指标.假定一个无向无权网络拥有 $m$ 条边,节点 $i$ 的度为 $k_i$,经社区发现后,总共有 $v$ 个社区,节点 $i$ 其所归属的社区数目为 $O_i$,则重叠社区模块度[25]定义为

$$Q=\frac{1}{2m}\sum_{v}\sum_{i,j}\frac{1}{O_iO_j}\left(A_{ij}-\frac{k_ik_j}{2m}\right),$$

其中,$A$ 为网络的邻接矩阵,描述节点 $i$ 和 $j$ 之间的连接关系,若节点 $i$ 和 $j$ 有连接,则 $A_{ij}$ 为 1,否则为 0.在非重叠社区划分算法(例如节点 CNM 算法)中,每个节点只能归属于 1 个社区,此时,对任何节点 $i$,有 $O_i$=1.显然,当所划分的社区内部边数越多时,$Q$ 值越大,社区结构越明显.

为了讨论单纯的网络结构凝聚性,本文在不考虑网络权重的情况下,对所发现的用户社区重新计算模块度,它避免了本文方法和节点 CNM 算法采用最优加权模块度进行层次切分所带来的评价不公平问题.同时,为了避免大量的零碎社区导致各算法模块度不均衡,本文在模块度计算上滤除了零碎社区.

表 3 为本文方法、节点 CNM 和 LCA 算法在 4 组数据集上所发现社区的模块度值.

**Table 3** *Q* values in the three methods

**表 3** 3 种算法的 *Q* 值

| 算法 | *S*1 | | | *S*2 | | | *S*3 | | | *S*4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 本文方法 | 节点 CNM | LCA 算法 | 本文方法 | 节点 CNM | LCA 算法 | 本文方法 | 节点 CNM | 本文方法 | LCA 算法 | 节点 CNM | LCA 算法 |
| 模块度 *Q* | 0.167 | 0.316 | 0.110 | 0.089 | 0.195 | 0.030 | 0.111 | 0.323 | 0.048 | 0.187 | 0.368 | 0.097 |

在 4 组数据集中,CNM 算法都具有最大的模块度值,其值分别为 0.316,0.195,0.323 和 0.368.一个好的社区发现,其模块度应在 0.3~0.7 之间[19].在 4 组数据集中,除了数据集 *S*2 外,其他 3 组数据集的节点 CNM 的模块度值较好,基本满足一个好的社区发现的模块度标准.本文方法和 LCA 算法的模块度较低,主要是因为重叠社区使得节点通常分布于多个不同社区,进而降低了模块度值.从总模块度上看,本文方法在 4 组数据集中所发现社

区的模块度值分别为 0.167,0.089,0.111 和 0.187,而 LCA 算法在 4 组数据集中所发现社区的模块度值分别为 0.110,0.030,0.048 和 0.097.因此,本文算法所发现的社区在总体网络结构上都在不同程度上优于 LCA 算法.此外,图 10 为 3 种算法最大 20 个社区的模块度贡献分布图,在 4 组数据集中,本文方法的模块度贡献值几乎都高于 LCA 算法.因此总体上说,本文方法更适合于微博网络的用户社区发现.

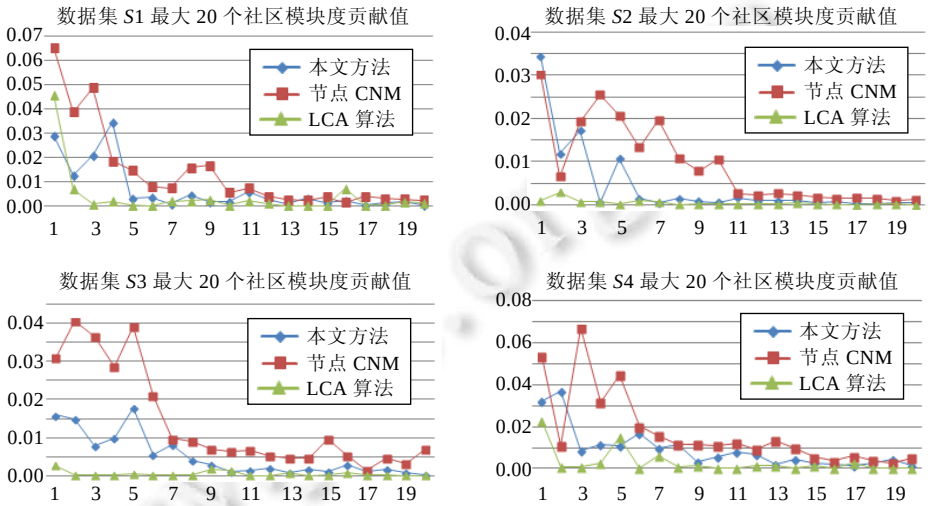



Fig.10 Distribution of modularity $Q$ on top rank 20 communities

图 10 3 种算法用户数最多的 20 个社区内部模块度值分布

为了进一步讨论所划分用户社区的质量,本文抽取本文算法和 LCA 算法所划分的最大用户社区进行的讨论.由于节点 CNM 算法社区较小,不进行对比.表 4 从图密度、网络直径、平均路径长度和平均聚类系数对两种算法最大社区进行了对比.由表 4 可知:在 4 组数据集中,本文方法和 LCA 算法所发现的最大社区都有相同的网络直径,分别为 9,7,6 和 5,在网络直径上两者持平.在数据集 *S*1 和 *S*4 中,本文方法最大社区的平均聚类系数分别为 0.128 和 0.497,略低于 LCA 算法的最大社区的 0.239 和 0.617;但在该数据集 *S*1 和 *S*4 中,本文方法最大社区的图密度分别为 0.003 47 和 0.003 4,都要优于 LCA 算法的最大社区的 0.003 42 和 0.022;且本文方法的平均路径长度也要略优于 LCA 算法.在数据集 *S*2 和 *S*3 中,虽然本文方法的平均路径长度和图密度都略差于 LCA 算法,但本文方法所发现的最大社区有着较优的平均聚类系数.因此,本文方法所发现的单个社区在网络结构上也不劣于 LCA 算法所发现的社区.

**Table 4** Comparison of network statistics over the largest user communities

**表 4** 最大社区网络参数统计对比

| 算法 | *S*1 | | *S*2 | | *S*3 | | *S*4 | |
|---|---|---|---|---|---|---|---|---|
| | 本文方法 | LCA 算法 | 本文方法 | LCA 算法 | 本文方法 | LCA 算法 | 本文方法 | LCA 算法 |
| 图密度 | 0.003 47 | 0.003 42 | 0.013 | 0.014 | 0.022 | 0.034 | 0.034 | 0.022 |
| 网络直径 | 9 | 9 | 7 | 7 | 6 | 6 | 5 | 5 |
| 平均路径长度 | 4.017 | 4.112 | 3.498 | 2.969 | 2.988 | 2.912 | 2.66 | 2.867 |
| 平均聚类系数 | 0.128 | 0.239 | 0.483 | 0.432 | 0.603 | 0.593 | 0.497 | 0.617 |

## 4 结 论

本文在分析现有社区发现算法(尤其是 LCA 算法)在微博用户社区发现上存在不足的基础上,提出了微博网络 R-C 模型,探讨了使用 R-C 模型进行微博用户社区发现的方法,分析了其社区聚类的时间复杂度,并以 CNM 算法和单边聚类算法为例,说明了该方法在微博网络等稀疏网络中,其聚类复杂度较低,近似等同于原有模型的

复杂度.在 4 组新浪微博真实数据集下,通过对比节点 CNM 算法和 LCA 算法,本文从兴趣内聚和网络结构内聚两方面分析了本文方法所发现微博用户社区具有较优的兴趣和网络结构内聚性.

本文将 R-C 模型应用于新浪微博进行社区发现,并取得了较优的社区结构.R-C 模型和本文方法将同样适用于其他具有双向关注关系并能提取用户内容的在线社区,如 Twitter、Google+、人人网、QQ 微博等.在后续的工作中,我们将会在更多的数据集上使用 R-C 模型并验证其能发现更优的用户社区.

**References**:


[1] Lim KH, Datta A. Following the follower: Detecting communities with common interests on Twitter. In: Proc. of the 23rd ACM Conf. on Hypertext and Social Media. New York: ACM Press, 2012. 317−318. [doi: 10.1145/2309996.2310052]

[2] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007,58(7):1019−1031. [doi: 10.1002/asi.20591]

[3] Dourisboure Y, Geraci F, Pellegrini M. Extraction and classification of dense communities in the Web. In: Proc. of the 16th Int'l Conf. on World Wide Web. New York: ACM Press, 2007. 461−470. [doi: 10.1145/1242572.1242635]

[4] Tang L, Wang X, Liu HF. Uncoverning groups via heterogeneous interaction analysis. In: Proc. of the Ninth IEEE Int'l Conf. on Data Mining. Miami: IEEE, 2009. 503−512. [doi: 10.1109/ICDM]

[5] Asur S, Parthasarathy S, Ucar D. An event-based framework for characterizing the evolutionary behavior of interaction graphs. ACM Trans. on Knowledge Discovery from Data (TKDD), 2009,3(4):16. [doi: 10.1145/1281192.128129]

[6] Richardson M, Domingos P. Mining knowledge-sharing sites for viral marketing. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2002. 61−70. [doi: 10.1145/775047.775057]

[7] Iyer G, Soberman D, Villas-Boas JM. The targeting of advertising. Marketing Science, 2005,24(3):461−476. [doi: 10.1287/mksc.1050.0117]

[8] Kaplan AM, Haenlein M. Two hearts in three-quarter time: How to waltz the social media/viral marketing dance. Business Horizons, 2011,54(3):253−263. [doi: 10.1016/j.bushor.2011.01.006]

[9] Larsson AO, Moe H. Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media & Society, 2012,14(5):729−747. [doi: 10.1177/1461444811422894]

[10] Lim KH, Datta A. Finding twitter communities with common interests using following links of celebrities. In: Proc. of the 3rd Int'l Workshop on Modeling Social Media. New York: ACM Press, 2012. 25−32. [doi: 10.1145/2310057.2310064]

[11] Jansen BJ, Zhang MM, Sobel K, Chowdury A. Micro-Blogging as online word of mouth branding. In: Proc. of the 27th Int'l Conf. on Extended Abstracts on Human Factors in Computing Systems. New York: ACM Press, 2009. 3859−3864. [doi: 10.1145/1520340.1520584]

[12] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 1999,46(5):604−632. [doi: 10.1145/324133.324140]

[13] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

[14] Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T. Probabilistic author-topic models for information discovery. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2004. 306−315. [doi: 10.1145/1014052.1014087]

[15] McCallum A, Wang XR, Corrada-Emmanuelm A. Topic and role discovery in social networks with experiments on Enron and academic email. The Journal of Artificial Intelligence Research, 2007,30:249−272.

[16] Pathak N, Delong C, Banerjee A, Erickson K. Social topic models for community extraction. In: Proc. of the 2nd SNA-KDD Workshop 2008. Las Vegas: ACM Press, 2008.

[17] Kernighan BW, Lin S. An efficient heuristic procedure for partitioning graphs. The Bell System Technical Journal, 1970,49(1):291−307. [doi: 10.1002/j.1538-7305.1970.tb01770.x]

[18] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences, 2002, 99(12):7821−7826. [doi: 10.1073/pnas.122653799]

[19] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004,69(2):026113. [doi: 10.1103/PhysRevE.69.026113]

[20] Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E, 2004,69(6):066133. [doi: 10.1103/PhysRevE.69.066133]

[21] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E, 2004,70(6):066111. [doi: 10.1103/PhysRevE.70.066111]

[22] Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. Nature, 2005,433(7028):895−900. [doi: 10.1038/nature03288]

[23] Newman MEJ. Analysis of weighted networks. Physical Review E, 2004,70(5):056131. [doi: 10.1103/PhysRevE.70.056131]

[24] Leicht EA, Newman MEJ. Community structure in directed networks. Physical Review Letters, 2008,100(11):118703. [doi: 10.1103/PhysRevLett.100.118703]

[25] Shen HW, Cheng XQ, Cai K, Hu MB. Detect overlapping and hierarchical community structure in networks. Physica A: Statistical Mechanics and its Applications, 2009,388(8):1706−1712. [doi: 10.1016/j.physa.2008.12.021]

[26] Zhang ZF, Li QD, Zeng D, Gao H. User community discovery from multi-relational networks. Decision Support Systems, 2013,54(2): 870−879. [doi: 10.1016/j.dss.2012.09.012]

[27] Yan F, Zhang M, Tan YW, Tang J, Deng ZH. Community discovery based on actors' interests and social network structure. Journal of Computer Research and Development, 2010,47:357−362 (in Chinese with English abstract).

[28] Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature, 2010,466(7307):761−764. [doi: 10.1038/nature09182]

[29] Derényi I, Palla G, Vicsek T. Clique percolation in random networks. Physical Review Letters, 2005,94(16):160202. [doi: 10.1103/PhysRevLett.94.160202]

[30] Palla G, Barabási AL, Vicsek T. Quantifying social group evolution. Nature, 2007,446(7136):664−667. [doi: 10.1038/nature05670]

[31] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005,435(7043):814−818. [doi: 10.1038/nature03607]

[32] Farkas I, Ábel D, Palla G, Vicsek T. Weighted network modules. New Journal of Physics, 2007,9(6):180. [doi: 10.1088/1367-2630/9/6/180]

[33] Xiong X, Niu X, Zhou G, Xu K, Huang YZ. Microgroup mining on tsina via network structure and user attribute. In: Proc. of the 7th Int'l Conf. on Advanced Data Mining and Applications. Berlin: Springer-Verlag, 2011. 138−151. [doi: 10.1007/978-3-642-25856-5_11]

[34] Xiong X, Zhou G, Niu X, Huang YZ, Xu K. Remodeling the network for microgroup detection on microblog. Knowledge and Information Systems, 2013:1−23. [doi: 10.1007/s10115-013-0626-x]

[35] Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proc. of the National Academy of Sciences, 2008,105(4):1118−1123. [doi: 10.1073/pnas.0706851105]

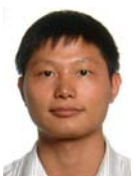
[36] Gower JC, Ross GJS. Minimum spanning trees and single linkage cluster analysis. Applied Statistics, 1969:54−64.

[37] Sibson R. SLINK: An optimally efficient algorithm for the single-link cluster method. The Computer Journal, 1973,16(1):30−34. [doi: 10.1093/comjnl/16.1.30]

[38] Barabási AL, Albert R. Emergence of scaling in random networks. Science, 1999,286(5439):509−512. [doi: 10.1126/science.286.5439.509]

## 附中文参考文献:

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



**周小平**(1985－),男,福建寿宁人,博士生,讲师,CCF 学生会员,主要研究领域为社会计算,数据挖掘.
E-mail: zhouxiaoping@bucea.edu.cn

**张海燕**(1975－),女,副教授,CCF 学生会员,主要研究领域为社会计算,推荐系统.
E-mail: zhy_rabbit@ruc.edu.cn

**梁循**(1965－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据挖掘,商务智能,社会计算.
E-mail: xliang@ruc.edu.cn