软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2824−2836 [doi: 10.13328/j.cnki.jos.004721]

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

微博网络上的重叠社群发现与全局表示*

胡 云[1,2], 王崇骏[1], 吴 骏[1], 谢俊元[1], 李 慧[2]

[1](南京大学 计算机科学与技术系,江苏 南京 210093)

[2](淮海工学院 计算机工程学院,江苏 连云港 222005)

通讯作者: 王崇骏, E-mail: chjwang@nju.edu.cn

摘 要: 微博网络是新兴的覆盖海量用户、涉及广泛话题并具有复杂重叠社群结构的多模网络.在深入研究微博网络各类实体和属性内在联系的基础上,提出了以用户-话题关系为主要划分原则的重叠社群表达模型及相应的社群结构发现算法.该方法不仅考虑网络中的用户-话题关系,还融合了这一网络特有的用户关注关系、博文评论与转发关系等所形成的复合网络关系.同时,改进了传统的社群隶属矩阵表述模型,通过引入虚拟社群,使隶属矩阵不仅合理反映个体对社群的隶属度,同时标识了个体在社群中的核心度.通过基于新浪微博数据集的实验验证,结果表明:该模型与方法能够高效合理地刻画该数据集包含的重叠社群结构,实验结果具有良好的可解释性,所提出的模型和算法可以有效地应用于类似多模网络社群划分和演化分析研究中.

关键词: 微博网络;实体关系模型;重叠社群;隶属矩阵;虚拟社群

中图法分类号: TP311

中文引用格式: 胡云,王崇骏,吴骏,谢俊元,李慧.微博网络上的重叠社群发现与全局表示.软件学报,2014,25(12):2824−2836. http://www.jos.org.cn/1000-9825/4721.htm

英文引用格式: Hu Y, Wang CJ, Wu J, Xie JY, Li H. Overlapping community discovery and global representation on microblog network. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2824−2836 (in Chinese). http://www.jos.org.cn/1000-9825/4721.htm

Overlapping Community Discovery and Global Representation on MicroBlog Network

HU Yun[1,2], WANG Chong-Jun[1], WU Jun[1], XIE Jun-Yuan[1], LI Hui[2]

[1](Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

[2](School of Computer Engineering, Huaihai Institute of Technology, Lianyungang 222005, China)

Corresponding author: WANG Chong-Jun, E-mail: chjwang@nju.edu.cn

Abstract: Micro-Blog cyberspace is a booming multiple mode network of numerous overlapping communities covering huge amount of users and topics relating to the nature, the society and the everyday life. Based on in depth analysis on the entities and inherent relationships among the network, this paper purposes a user-topic relation dominated structural module for overlapping community representation and detection, and also infuses the follow relationship along with the blog-forward and blog-comment relationship into the module. By introducing a virtual community into the actual communities of the network, the paper also puts forward an improved global belongingness matrix as user's role representation which has the ability to properly describe a user's degree of participation and importance in the network. Experimental results on Sina's micro-blog dataset show that the new method is favorable and efficient for finding meaningful communities from the micro-blog. Furthermore, the proposed module and algorithms can be adapted in various ways for similar social network analysis and helpful for community evolution research.

Key words: microblog network; entity relationship module; overlapping community; belongingness matrix; virtual community

* 基金项目: 国家自然科学基金(61403156, 61375069, 61105069); 国家博士后基金(2011M500846); 江苏省自然科学基金(11KJB520001, 13KJB520002); 江苏省科技支撑计划(BE2012181)

收稿时间: 2014-04-10; 定稿时间: 2014-08-21

微博作为目前最为流行社会媒体,为公众传递信息、参与话题讨论、表达自己的见解提供了前所未有的服务功能.微博网络是一种包含多种复杂实体关系的新型网络,它之所以有别于以往博客、论坛等传统网络平台,在于其构建了一种新的交互关系,即关注(follow)关系.所谓关注关系是指一个用户通过对其感兴趣的用户标注“关注”而形成的关系.通过关注关系,用户可即时获得他所关注用户的博文.关注关系为信息流动构建了强大的管道,使微博链式信息推送变得格外强劲.由于关注关系的引入,微博网络不同于现实中的其他网络结构,并显著区别于传统的用户-话题二部网络,因此被视为多模网络.目前,从多模网络的视角研究微博社群关系的研究尚不多见.微博上的社群结构体现为典型的重叠社群模式——用户依据自身的话题兴趣和人物关注关系同时参与多个不同的社群,通过关注关系接受特定用户的博文,参与特定话题的转发与评论.由于用户关注的用户与话题又是多方面的,由此形成了多个社群共同参与的微博网络关系.

微博现象在学术界引起了关注,吸引研究者从不同角度开展相关的研究工作,包括微博网络结构特征[1,2]、社群特征[3−5]、意见领袖发现、消息传播与舆情演化等.Zhang 等人[3]利用反映用户兴趣的文本内容、社会结构等特征来计算不同微博用户间的相似性,并利用 K-means 聚类识别微博群体.Tang 等人[4]利用基于正则化时间的多模型聚类算法发现动态的网络群体.而 Yu 等人[5]使用归一化切割方法(normalized cut)求解模块化聚类并解决群体发现问题.Huberman 等人[6]从微观角度结合马尔可夫链和聚类算法将微博用户划分为不同社群.Gao 等人[7]以用户作为节点、关系作为边,利用 MSCC(maximal strongly connected components)方法将微博用户划分为具有不同拓扑结构的社群,并且在不同用户社群上对用户的关系形成进行了分析.

在重叠社群发现研究方面,Palla[8]于 2005 年首先讨论了重叠社群发现问题,提出通过滚动 *K*-完全图发现重叠社群的 CPM(clique procolation method)方法.文献[9]提出了一种能够探测层次化社群结构的凝聚算法 BGLL.文献[10]同样提出了一种用于探测社群结构重叠性的算法 CONGA.文献[11]没有单纯从网络节点及其连接关系的角度出发,转而将个体-话题关系弧转化为图节点,把用户-话题二模网络转化为以个体-话题关系为节点,个体-话题相似性表示节点链接度的加权单模网络社群分析模型,能够较自然地解释和发现重叠社群,为重叠社群发现提供了新的思路.OSLOM[12]则基于求解节点度统计量局部最优的图聚类算法,尽管能够实现重叠社群的发现,但因其计算复杂性而难以应用于大规模网络.

在社群形态的定性研究方面,Gruzd 等人[13]认为,微博用户虽然处于一种虚拟网络社群之中,但这些虚拟社群能够体现用户的真实社会关系,提出微博社群具有成员性、影响性、整合需求性及分享性等特点.Java 等人[1]通过研究 Twitter 网络的拓扑属性和地理属性,发现具有高度相关性或强互动性的用户会逐渐聚合形成群体.Ryan 等人[14]的研究指出,社群核心用户会对整个社群的群体行为产生主导作用.

用户与话题热点的动态变化,是影响微博社群结构的主要因素.Kivran 等人[2]通过对群体结构的动态分析,总结并提出了判断微博用户间关系持续性的因素.Meeder 等人[15]利用时间戳信息分析微博社群的动态变化规律.Li 等人提出了一种事件演化脉络识别框架(event storyline from microblogs,简称 ESM)[16],该框架能够有效识别目标事件随时间演化所形成的话题脉络,从而更有利于分析网络社群的结构演化.Chen 等人[17]针对微博文本的海量性和话题发散性特征,研究了基于动态伪相关反馈思想的微博话题提取方法.袁毅等人[18]通过跟踪微博用户在时间周期内就某一话题的交流数据,发现用户在信息交流过程中形成关注、评论、转发和引用等 4 种社会关系网络,研究了 4 种关系网络不同的结构形态.Teutle 等人[19]则从网络动态性角度对 Twitter 进行分析,包括出入度增长、网络密度和介数等参数在微博网络中的变化等.

隶属矩阵(belongingness matrix)是表达个体相对于聚类隶属度的常用方法[8,20−21],其中,隶属矩阵的每行代表一个个体,行的第 i 个(非负)分量代表个体对第 i 类的隶属度,总和为 1.隶属矩阵为定量研究重叠聚类结构及变化提供了定量表达方法,但是在面对真实社会网络时,该表示方法存在明显缺陷.主要体现在隶属矩阵要求“所有个体的社群隶属度和全为 1”的限制,即个体对所有社群贡献的总和相等.该限制使其无法真实反映现实网络中个体的行为差异.例如:微博“意见领袖”博文的质与量及与其他用户的交互程度都与普通用户不同;而学术合作网中,著名学者对合作网络结构及变化的影响力也与一般学者显著不同.如果简单地将网络划分为若干社群,并给每个用户赋以等同社群隶属和值,则无法体现个体在网络中的实际地位和作用.如何恰当地描述个体

在网络社群中的地位和关系以真实反映个体在网络中的角色,对网络社群研究发挥重要作用.

总之,微博网络重叠社群研究对于揭示这一新型社交网络内在结构及其演化规律,提高对显著影响微博网络演化关键节点和要素的识别能力具有重要意义.因此,微博网络重叠社群发现及其表示框架是本文研究工作的着眼点,主要创新点在于:

(1) 提出将用户-话题连接关系作为维系微博网络的关键节点集,用户-话题对间相似关系为连边的加权多模网络结构模型.通过综合考虑个体对话题兴趣和对特定用户的关注关系,实现网络重叠社群结构的发现,能够真实反映网络用户的社群多重性;

(2) 研究了微博多模网络实体间相似关系度量方法,全面覆盖该网络各类实体间的内在联系,并基于优势集聚类给出社群发现方法,实现以话题群-用户群为社群划分标准的重叠网络社群.所提出算法具有类团核心性度量功能,能够实现社群核心成员的发现;

(3) 在分析现有隶属矩阵表示方法不足的基础上,提出新的能够反映个体对网络全局参与度与社群核心度的隶属矩阵表示方法.新的隶属矩阵表示模型是定量研究网络社群及其演化规律的基础.

1 微博网络上的实体关系

1.1 微博网络与用户的形式化描述

微博平台可以表示为如下的五元组:

$$W=(U,BLog,E_T,E_U,F,C) \tag{1}$$

其中,U 为全体用户的集合,$U=\{u|u\in$注册微博用户$\}$,$BLog$ 为全体用户博文的有限集合.类似地,一条博文可以视为一个四元组:

$$BLog_i=\{B_{id}\ BLog_i\_message,BLog_i\_topic,t\} \tag{2}$$

其中 B_{id} 为唯一标识符;$BLog_i\_message$ 为博文主体;t 为发布时间;$BLog_i\_topic$ 为从微博内容中抽取的话题信息,需要通过专门的分析技术提取获得.由于用户发表的博文可能很多,不妨对一个用户的全部博文加以分析,从而抽取出其博文关注的主要话题信息.实验中,将一个用户博文的全部话题限制在 10 个以内.

定义 $E_T\subseteq U\times T$ 为全体用户到其发表博文所属主题的连接弧的集合:

$$E_T=\{e=(u_i,t_j)|u_i\in U,t_j\in T\} \tag{3}$$

而 $E_U\subseteq U\times U$ 为用户通过加关注所形成的连接关系集.这里,关注关系是从关注者 i 到被关注者 j 的有向弧.

$$E_U=\{u_i\leftarrow u_j|u_j\ follows\ u_i\} \tag{4}$$

集合 E_U 存在一个虚拟伴随集,记为 $\tilde{E_U}=\{u_i\mapsto u_j\,|\,u_i\ befollowed\ by\ u_j\}$,它是由用户 j 关注用户 i 所形成的被关注关系.在观察一个具体用户(例如 i)时,这一虚拟边集合可以方便地计算用户 i 的粉丝数.

定义 $F\subseteq U\times BLog$ 为由全体用户与其所转发博文之间关系的集合:

$$F=\{(u_i,BLog_j)|u_i\in U,u_i\ forwarded\ BLog_j\} \tag{5}$$

定义 $C\subseteq U\times BLog$ 为由全体用户与其所评论的博文之间关系的集合:

$$C=\{(u_i,BLog_j)|u_i\in U,u_i\ commented\ on\ BLog_j\} \tag{6}$$

与微博平台的表述类似,可以进一步定义微博用户为七元组:

$$u=(BLog,Topic,Follow,Followed,forward,Comment,t) \tag{7}$$

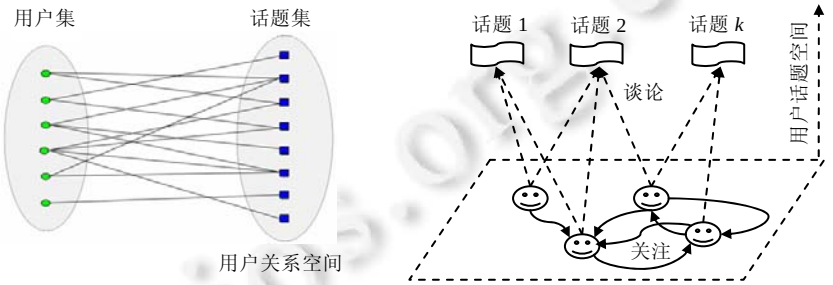
其中,$u.BLog$ 为用户发表的博文集,$u.Topic$ 为从该用户 $u.BLog$ 中学习获得的话题集,$u.Follow$ 为用户关注的其他用户集,$n_{Follow}=|u.Follow|$为关注用户数,$u.Followed$ 为关注该用户的其他用户集,$n_{Followed}=|u.Followed|$为被关注用户数(粉丝数),$u.forward$ 为用户转发其他用户的博文合集,$u.Comment$ 为其评论博文集,t 为用户注册时间.

1.2 微博网络中的实体关系模型

现实世界中,许多网络都可以表征为二部图(bipartite graph network)结构.如演员合作网中的演员-电影关系,科学研究中的共同发表论文关系等.二部图由两类性质不同、互不相交的节点构成:一类称为用户节点集 X,

另一类称为事件节点集 Y(如图 1(a)所示),且同一个集合内的节点间无连边.即,X 中元素可与 Y 中元素间存在连边,但 X 或 Y 内部节点间均无连边.由于只在不同类型节点间存在连边,二部图网络中不存在三角关系.

传统意义上,用户-话题二部图网络能够描述十分丰富的网络模型,如科研合作网络等.但是微博网络关系表现更为复杂,这是因为用户间可以用他们的博文作为联系的纽带,通过对博文的推送、转发和评论形成按话题的社会关系.同时,用户之间还存在关注与被关注关系,即使是关注某个用户,也并不意味着关注他所发表的全部话题,而是有选择地加以阅读、评论和转发.因此,本文研究将其拓展为融合多种实体关系的多模网络.微博网络上的社会关系可如图 1(b)所示.



(a) 二模网络 (b) 微博多模网络

Fig.1 Illustration of a Bi-mode network and the micro-blog multi-mode network

图 1 简单的二部网络(a)与微博多模网络(b)示意图

图 1(b)中,4 个微博用户的话题集包含话题 1~话题 k,可以构成标准的话题-用户二部图.但同时,4 个用户间还存在关注与被关注关系.此外,网络中还存在用户对特定博文的转发、评论关系(未在图中标识).显然,微博网络的这种复合关系无法直接转化为单模网络.如何综合考虑用户节点的话题属性和用户间的关注属性,既在统一的框架下解决微博社群发现问题,又能恰当刻画节点在社群中的地位,是本文研究的关键问题.它显著区别于传统的二部网络领域的研究工作.

为此,本文拓展文献[11]将二模网络转化为单模加权网络的研究方法,通过综合权衡微博网络中存在的多种实体关系,以最基本的用户-话题关系为主线,综合考虑关注、评论与转发关系,构建该复杂多模网络向单模加权网络的映射方法,并据此实现微博网络上的重叠社群发现.

首先给出描述微博网络的图表示为:微博网络图 $G=(V,E)$是一个复合图,其中,节点集 $V=U\cup T$ 满足 $U\cap T=\varnothing$,U 为全体用户节点的集合,T 为全体用户博文所涉话题的集合,$T=\{t_1,\ldots,t_m\}$.连边集 $E=E_T\cup E_U$,$E_T\cap E_U=\varnothing$.其中,连边集 E_T 为用户节点到其微博所涉话题节点的连边,它可以视为一个映射函数 $E_T:U\times T\to\{0,1\}$,即:

$$e_{ip}(u_i,t_p)=\begin{cases}1, & u_i \text{ post messages on } t_p\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

连边集 E_U 是全体用户间由关注(follow)而引发的关系集,即 $E_U:U\times U\to\{0,1\}$,满足:

$$e_{ij}(u_i,u_j)=\begin{cases}1, & u_i \text{ follows } u_j\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

定义 1. 给定图 G 中用户-话题关系集合 E_T 中的任意两条连边 $e_{ip}(u_i,t_p)$和 $e_{jq}(u_j,t_q)$,其相似度定义为

$$sim(e_{ip},e_{jq})=\alpha\cdot sim(u_i,u_j)+(1-\alpha)sim(t_p,t_q) \tag{10}$$

上述定义将两条用户-话题连边之间的相似关系定义为对应用户之间的相似度与话题之间相似度的线性组合,$\alpha(0<\alpha<1)$为权重调节系数.这既考虑了话题间的接近程度,同时又兼顾了用户间的同质性,具有现实意义与理论上的合理性.

值得注意的是,由于 $e(u,topic)$是用户与话题间的连边,因此既存在多个用户联向同一话题的连边,又存在一个用户向多个不同话题的连边.前者对应于多个用户博文涉及同一话题,而后者对应于一个用户博文涉及多个

不同话题域.正是由于多对多的关系,为发现重叠社群提供了可能.为了实现诸如公式(10)所定义的用户-话题相似度计算,以下分别讨论用户相似度与话题相似度的定义与计算方法.

1.3 博文话题相似性度量

为了建立用户博文间的相似性,首先必须从博文中提取相应的话题.有关文本话题提取的方法主要有TF-IDF 算法[22]和 LDA 算法[23]等,在此不再赘述,仅假设通过对用户博文集的分析处理获得了用户的话题集.然而对于海量的微博用户,简单汇聚所有用户的话题形成用户话题集显然将非常庞杂并存在大量冗余.为此,可以利用话题间的语义相似性和概念聚类方法对用户话题集加以压缩,形成规模适度的话题集.

事实上,虽然存在海量的用户共享着微博网络,其话题也涉及现实社会生活各个领域,但从统计学角度看,网络用户所涉话题同样符合幂率分布,即少量热门话题吸引了大量用户,而大量冷门话题则相对稳定地只被少数人关注.因此,可以利用话题语义间相似性实现话题压缩,将博文所涉主要话题控制在一定范围内.

设 T 为从全体用户博文中提取的主题词集,$T=\{t_1,\ldots,t_n\}$, $T^*=\{t_1^*,\ldots,t_m^*\}$ 为由 T 中话题组成的隐语义网络空间的基,则对网络中任一个体 u,其所涉话题子集 $u_T=\{t_{i1},\ldots,t_{ik}\}$可通过如下映射转化为 T^*中的话题子集:

$$u_{T^*}=\{t_{i1}^*,\ldots,t_{ik'}^*\}=\Theta(u_T) \tag{11}$$

奇异值分解(singular value decomposition,简称 SVD)方法是常用于实现上述话题集压缩映射的有效工具.设 m 阶对角矩阵Σ的全体对角元素为由话题词 $t_1,\ldots,t_n$ 构成的语义相似矩阵的前 m 个特征值,L 与 R 为相应的左右奇异特征向量组成的矩阵,则有 $M=L\Sigma R^T$.此时, $u_T=u_T(t_{i1},\ldots,t_{ik})=\{L\Sigma\}u_TR^T=u_{T^*}(t_{i1}^*,\ldots,t_{ik'}^*)R^T$.于是,

$$u_{T^*}(t_{i1}^*,\ldots,t_{ik'}^*)=u_T(t_{i1},\ldots,t_{ik})R\,.$$

这里,$u_T(t_{i1},\ldots,t_{ik})$与 $u_{T^*}(t_{i1}^*,\ldots,t_{ik'}^*)$ 分别为用户话题在原话题空间及其在压缩空间中的映像.考虑到微博话题的发散性和高动态性,微博话题选择与时间跨度之间应建立一定的联系.根据 Dumais 等人的研究[24]:随时间跨度的不同,微博热门话题可选择在 50~1000 范围内.因此,本文实验中将话题总数控制在 100 个范围内.

根据上述讨论,两个用户讨论的话题尽管可能略有不同,但当对话题进行高度浓缩后,这些细微的差异已无法区分.因此,公式(10)中两话题的相似度定义为

$$sim(t_p,t_q)=\begin{cases}1, & t_p=t_q\\ 0, & t_p\neq t_q\end{cases} \tag{12}$$

以下讨论用户博文所涉话题时,均假设其为对全体用户博文集话题经奇异值分解压缩后的话题集合.

1.4 用户关注相似性度量

本节从用户行为的维度探讨其与微博社群结构形成该关系.袁毅等人[18]通过跟踪微博用户在特定时间周期内的话题交流数据,发现用户在信息交流过程中形成关注、评论、转发和引用这 4 种关系网络,指出 4 种网络具有不同的结构形态. Teutle 等人[19]则从网络动态性的角度对 Twitter 加以分析,包括出入度的增长、网络密度和介数等参数来描述微博网络特征.本文依据文献[18]的相关结论,综合构造用户行为相似度的描述方法.

关注网络是由用户间因关注关系而构成的级联式网络.由于博文随关注关系自动即时到达一级关注用户,因此信息传播快,关注网络通过关注对象的转发、引用和评论而使信息进一步转播给各自的关注用户,形成更大的多级关注网络.转发网络是在时间周期内用户间因博文转发形成的网络,其规模远比关注网络小得多.这是因为关注用户中仅有一部分对当前博文内容具有转发动机,其他关注者对当前博文并不感兴趣,转发大多发生在一次范围内,二次转发衰减显著.评论网络是通过对博文评论形成的关系网络,其规模更小于转发网络,说明用户更愿意做较为省力的简单转发而不加评论,只有当用户对话题具有更大兴趣时,才会加以评论[18].

以下着手构建用户行为维度上的相似关系度量.首先,两用户间的关注对象集的交叠性反映了其在社会网络中社群角色的异同.因此,基于关注关系的用户相似性度量可以如下定义:

定义 2. 设 $u_i,u_j\in U$ 为微博用户,$u_i.Follow,u_j.Follow$ 分别为 u_i,u_j 的关注对象集,而 $u_i.Followed,u_j.Followed$ 分别为 u_i,u_j 的粉丝集,则 u_i,u_j 的关注相似度为

$$Sim_F(u_i,u_j)=\beta\cdot\frac{|u_i.Follow\cap u_j.Follow|}{|u_i.Follow\cup u_j.Follow|}+(1-\beta)\cdot\frac{|u_i.Followed\cap u_j.Followed|}{|u_i.Followed\cup u_j.Followed|} \tag{13}$$

式(13)右端第 1 项表示两用户共同关注用户在他们关注的所有用户集的比例,而后者表示两者的粉丝集内共同粉丝所占的比值.显然,前一指标值越大,说明他们在网络上共同关注的对象越相似;而后一项指标越大,则说明其被同一群人追捧的程度越高.公式(13)通过调节$\beta\in(0,1)$的值,调整关注行为和被关注状态在用户相似度中的比值.由于"关注"属于用户的自主行为,因此其在自我认同中的比重较大.本文实验中,选取β=0.6.

1.5 转发和评论关系引入的用户-话题相似性增量

由用户间转发和评论行为所引起的相似性既不能简单归结到用户间的相似性,也不能视为用户间的话题认同.这是因为一个用户未必会转发或评论另一个用户涉及所有话题的博文,而只会就自己感兴趣的话题博文进行转发或评论.因此,转发或评论反映的是两用户间感兴趣话题上的相似性.为此,可以采用以下策略在考虑用户-话题相似性时加入用户转发、评论的因素:

定义 3. 设用户 u_i 转发了用户 u_j 话题为 t_p 的博文,即:存在 u_i 的某篇评论 $u_i.comment_l\in u_i.cmt$,使得 $u_j.t_p\in(u_i.cmt).t$,则按以下策略定义一对用户-话题弧间的评论相似性增量 $sim(e_{ip},e_{jq})$:

$$sim(e_{ip},e_{jq})=\begin{cases}\alpha\cdot sim(u_i,u_j)+(1-\alpha), & t_p=t_q\\ \alpha\cdot sim(u_i,u_j)+\delta_{cmt}(1-\alpha), & t_p\neq t_q,0<\delta_{cmt}<1,t_p\in(u_j.cmt).topic\ or\ t_q\in(u_i.cmt).topic\\ \alpha\cdot sim(u_i,u_j), & t_p\neq t_q,t_p\notin(u_j.cmt).topic\ and\ t_q\notin(u_i.cmt).topic\end{cases} \tag{14}$$

其中,$(u_j.cmt).topic$ 表示 u_j 所有评论所涉及话题的集合.

上述相似性定义策略的含义是:如果 u_i 评论用户 u_j 话题为 t_p 的博文,而其本身也是话题 t_p 的博文原创者(即 t_q=t_p),则可不需考虑这一转发关系,直接设置该对用户-话题弧的话题相似度为 1;反之,尽管 $t_q\neq t_p$,但因 u_i 评论了话题为 t_q 的博文或 u_j 评论了话题为 t_p 的博文,这种评论关系反映该对用户话题弧上两用户间在话题上具有相似的兴趣,因此提高了这对用户-话题弧的相似度.最后,如果两用户间不存在相同话题的评论关系,则该对用户-话题弧的话题相似度为 0.公式(14)中,本文用参数 δ_{cmt} 调节因评论关系形成的相似度增量.

同样地,对用户参与其他用户博文的转发行为也可以采用类似的策略,在一对用户-话题弧的话题相似关系上通过设置参数 δ_{fwd} 加以体现.

综合公式(12)~公式(14),由公式(10)形式定义的用户-话题关系对 e_{ip} 和 e_{jq} 之间的相似度最终定义为

$$sim(e_{ip},e_{jq})=\alpha\cdot sim(u_i,u_j)+\beta(1-\alpha) \tag{15}$$

其中,β的取值可分为如下几种情况:

(1) 当 t_q=t_p 时,β=1;

(2) 当 $t_q\neq t_p$,但 u_i 或 u_j 评论且转发对方弧上对应的话题时,$\beta=\delta_{fwd}+\delta_{cmt}$;

(3) 当 $t_q\neq t_p$,但 u_i 或 u_j 评论或转发对方弧上对应的话题时,$\beta=\delta_{fwd}$ 或$\beta=\delta_{cmt}$;

(4) 当 $t_q\neq t_p,u_i,u_j$ 不存在对方弧上对应的话题的评论或转发,β=0.

显然,用户对博文发表评论较之简单转发更能反映其对相关话题的兴趣度.因此,实验中将由转发和评论行为导出的用户-话题相似度调节系数 $\delta_{fwd},\delta_{cmt}$ 分别设为 δ_{fwd}=0.25,δ_{cmt}=0.5.

2 微博网络重叠社群发现

2.1 基于优势集划分的加权图聚类方法

优势集(dominant set)聚类是由 Pavan 和 Pelillo 首次给出确切定义的图聚类方法[25],该方法源于聚类直觉观念和主导集合之间的类比关系,优势集将最大类团的概念推广到边赋权图中.文献[26]则证明了优势集检测与标准单纯形的二次规划极大值问题之间的对应关系,使得算法可应用连续最优技术,并可在局部交互计算单元的并行网格上简单实现.本文采用优势集聚类的方法主要基于两点考虑:(1) 该方法具有扎实的理论基础,并能够清晰地反映网络图上节点的类团隶属关系;(2) 在实现类团划分的同时还标识了节点在类团中的核心度,具

有描述网络个体角色属性的便利.以下我们给出基于优势集聚类算法实现上述用户-话题连接关系为节点的加权图聚类相关的定义与算法过程.

给定无向加权图$\Omega=(E_T,L)$,其中,节点集 E_T 为由公式(10)定义的用户-话题关系集,L 为由公式(15)定义的用户-话题对间相似度构成的加权边.而全体边上的权值构成用户-话题关系集上的相似矩阵 $A=[a_{ij}]$:

$$a_{ij}=sim(e_i,e_j),e_i,e_j\in E_T \tag{16}$$

定义 4. 设 S 为 E_T 的非空子集,E_T 中任意元素 $e_i\in E_T$ 相对于子集 S 的平均权值相似度定义为

$$dg_S(e_i)=\frac{1}{|S|}\sum_{e_j\in S}a_{ij} \tag{17}$$

注意到,由于图 E_T 中不存在自连边,因此有 $dg_{e_i}(e_i)=0$.进一步地,对 $e_j\notin S$,定义:

$$\varphi_S(e_i,e_j)=a_{ij}-dg_S(e_i) \tag{18}$$

定义 5. 设 S 为 E_T 的非空子集,则 S 中任意元素 e_i 相对于 S 的权定义为

$$w_S(e_i)=\begin{cases}1, & |S|=1\\ \sum\limits_{e_j\in S\setminus\{e_i\}}\varphi_{S\setminus\{e_i\}}(e_j,e_i).w_{S\setminus\{e_i\}}(e_j), & |S|>1\end{cases} \tag{19}$$

S 的总权值定义为

$$w(S)=\sum\nolimits_{e_i\in S}w_S(e_i) \tag{20}$$

根据上述定义,$w_S(e_i)$的计算过程仅需在由 S 导出的子图上进行.值 $w_S(e_i)$直观地表示在 S 中添加节点 e_i 前后该集合平均相似度的变化情况.

定义 6. 非空子集 S 称为优势集,如果该集合满足以下两个条件:

(1) 对 S 中的每个元素 e_i,均有 $w_S(e_i)>0$;

(2) 对任意的 $e_i\notin S$, $w_{S\cup\{e_i\}}(e_i)<0$.

定义 6 描述了类团的两个基本特征,即,类团内部同质性及类团成员与外部个体的异质性.它说明:如果在 S 加入某个节点 e_i 后,集合 S 仍然保持为同质的,则可以认为该点是类团成员;否则,应剔除该节点.

根据文献[26]的证明,在加权图上寻找优势集的问题可以转化为求解标准单纯形二次型极值问题:

$$\text{Maxsize } f(x)=\frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x}\ \text{ s.t. } \boldsymbol{x}\in\Delta \tag{21}$$

其中,$\Delta=\{\boldsymbol{x}\in\boldsymbol{R}^n,\boldsymbol{x}\geqslant 0,\|\boldsymbol{x}\|=1\}$.上述单纯形上求解二次型极大值问题可转化为繁殖方程法求解:

$$u_i(t+1)=u_i(t)\frac{(A\boldsymbol{x}(t))_i}{\boldsymbol{x}(t)^T A\boldsymbol{x}(t)} \tag{22}$$

其中,$u_i(t)$是向量 $\boldsymbol{x}(t)$的分量,t 为迭代步数.迭代结果值的大小,表示分量对应的节点属于当前类团的可能性.向量 $u_i(t)$的支持集即为优势集对应的顶点,每一分量 $u_i(t)$对应原图中节点集 E_T 中一个节点的特征向量,数值越大的分量之间相似性越高.因此可以利用该信息,按照 $u_i(t)$值从大到小对原始节点排序,直到完成当前类成员的搜索.基于优势集的图聚类算法描述见算法 1.

算法 1. 优势集聚类算法(dominant sets clustering algorithm).

输入:图 G 的节点集和节点间的相似度矩阵 A;

输出:类团及成员列表.

1. 初始化 $A^k=A,k=1$
2. 利用迭代方程(22)求解方程(21)获得解 u^k
3. 计算 $f(u^k)$的值
4. 求解 u^k 对应的特征向量获得矩阵对应行向量,输出为 $cluster_k=\rho(u^k)$
5. 从矩阵 A^k 中删除 $cluster_k$ 元素对应的行和列,形成新的相似矩阵 A^{k+1}
6. 若 A^{k+1} 不为空矩阵,$A^k\leftarrow A^{k+1}$,$k=k+1$;转步骤 2

7. 否则,程序终止.

2.2 用户-话题重叠社群发现

尽管算法 1 将节点划分为互不重叠的类团,但这些节点事实上包含的是用户与其博文话题之间的连接信息.因此,仍需将类团的节点转化为对应的用户-话题关系,从而可得到该类团相应的社群信息.

给定类团 $S=\{e_1,\ldots,e_k\}$,将节点集转化为用户-话题关系后的集合为 $S=\{(u_i,t_i),i=1,\ldots,k\}$,于是可得社群用户集为 $U_S=\bigcup u_i$,S 的话题集为 $T_S=\bigcup t_i$.

显然,由此得到的用户话题关系集是一个多用户、多话题复合集,且 T_S 中存在重复话题,通过剔除重复项,即可得到一个社群涉及的话题集.当然,判别算法的有效性在于检验所发现的社群是否包含相对集中的话题.

注意到:在对用户的博文进行话题聚合时,每个用户涉及同一话题的连边进行了归并.因此,通过预处理后得到的用户-话题数据中,每个用户相对于一个话题要么没有连接关系,要么仅有一条连接.但当类团中存在多个话题时,同样存在用户的重复,也需要加以剔除.

以下总结并给出完整的用于面向用户-话题复合网络的重叠社群检测模型(overlapping community detection on multi-mode network,简称 OCDM)的完整步骤如下:

输入:有限用户集 U,微博网络 $W=(U,Blog,E_T,E_U,F,C)$;

输出:用户-话题(重叠)社群列表、每个社群核心成员集.

1. 对每个用户 u_i,从 $u_i.BLog$ 博文集合中提取用户话题集 T_i(本文采用 TF-IDF 算法)
2. 对全体话题集 $T=\bigcup T_i$,采用 SVD 算法压缩形成新的全局话题集 T
3. 依据话题集 T,剔除用户-博文关系中重复连接,压缩形成用户-话题关系集 E_t
4. 依据公式(12)~公式(14)计算 E_t 中元素的两两相似度,构建加权网络 $\Omega=(E_T,L)$.
5. 采用 DS-Cluster 算法划分 $\Omega=(E_T,L)$ 上的类团 $\{S_i\}$.
6. 剔除 S_i 中的重复话题和重复用户,依次输出用户-话题社群 S_i.
7. 采用颜色标注算法提取每个重叠社群 S_i 的核心成员集.

3 网络社群隶属关系的全局表示

以上我们研究了微博网络上的重叠社群发现问题.为了从全局层面上研究所有社群及其个体的演化行为,必须给出网络成员相对于社群的隶属关系.

隶属矩阵是表达个体相对于聚类隶属度的常用方法[8,20,21].在隶属矩阵中,每行表示一个数据对象,而每列代表一个聚类.于是,可将 N 个成员相对于 k 个类团的隶属关系表示为如下的形式:

$$\boldsymbol{P}=[a_{ij}]_{N\times k},0\leqslant a_{ij}\leqslant 1,\sum_{j=1}^{k}a_{ij}=1 \tag{23}$$

上述隶属矩阵表达方法在描述空间数据点对重叠聚类的隶属关系时具有合理性,这是因为在聚类时每个点在数据集中的地位是等同的,不存在数据对象在聚类中地位的差异.但是,直接将这种描述方法应用于现实网络中的社群关系描述则具有明显的缺陷.以下我们以微博网络为例,说明公式(23)在描述个体社群隶属关系时的不合理性.

假设用户 A 是微博网络的积极参与者,并在 3 个社群 S_1,S_2,S_3 中因博文数、被关注数、被评论和转发数领先而具有核心地位.显然,该用户应成为网络社群中的重点研究对象.但从社群隶属关系来说,由于该用户同时是多个社群的成员,因此在总隶属度值为 1 的限制下,该用户在 3 个社群中的平均隶属度只有 1/3.再假设用户 B 仅是社群 S_1 的普通用户,在社群划分时仅被归入 S_1.根据前述,则 B 对社群 S_1 的隶属度为 1,远远高于用户 A 对该社群的隶属度.在研究社群演化时,如果 B 在下一时刻离开社群 S_1 而加入到社群 S_2,则在两个相邻的隶属矩阵视图间,用户 B 对应的分量从(1,0,0,…)转化为(0,1,0,…).显然,个体 B 的社群隶属度变化导致的演化异动量较之个体 A 要显著得多,这一情景体现了传统隶属矩阵描述方式的被忽略的不合理性.

为改进隶属矩阵表示方法的缺陷,有必要依据社群个体在社群中的差异性设计新的社群隶属度描述方法,以便真实反映个体对社群的隶属度及其社群地位.改进后的隶属矩阵表达模型应具备以下几项特征:

(1) 一个社群对整个社会网络的影响与其所包含个体的数量正相关,即社群规模越大,该社群的隶属矩阵总权值越大.事实上,有关网络社群的统计研究结果表明[1,2],网络社群在规模上符合幂率分布.即网络中存在少量的庞大社群,而绝大多数社群是局部的和小规模的;

(2) 个体在一个社群中的影响和地位与其在社群中的核心度存在正相关性.这一现象同样得到社会网络研究相关成果所证实,即网络社群中的少量个体对社群结构起到关键作用(如节点的连接度),而大量其他个体通过他们而形成社群;

(3) 为了保持隶属矩阵的行归一性质,可假设在网络外部存在一个虚拟社群,使得除了网络中综合权值最大的个体对真实社群的隶属度和为 1(相当于全力参与社会网络)外,其他个体因为参与这一虚拟社群(如与社会网络无关的学习工作等)而没有将全部的精力投入到实际网络中,其对实际社群的隶属度总和δ<1,而对虚拟社群的隶属度为 1−δ.在研究社群演化时,通过对所有网络个体在实际社群的隶属度评价值自动调节δ的大小,以反映每个个体对实际社群的综合影响力.

基于上述假设,我们给出改进后的网络社群隶属矩阵的定义如下:

定义 7. 给定由 N 个成员组成的社会网络 G 及其上的 k 个社群 $S_1,\ldots,S_k$,设$|S_j|$为第 j 个社群的成员数,而 $w_{S_j}(e_i)$ 为 S_j 成员 e_i 的核心度权值, $i=i_1,\ldots,i_{|S_j|}; j=1,\ldots,k$.则社群 S_j 的全局影响力权值定义为 $\log|S_j|$,e_i 基于社群 S_j 的全局影响力权值定义为 $w_{S_j}(e_i)\cdot\log|S_j|$.

根据上述定义,网络个体 e_i 相对于全体社群的影响力向量可以表示为

$$\boldsymbol{w}(e_i)=(w_{S_1}(e_i)\cdot\log|S_1|,\ldots,w_{S_k}(e_i)\cdot\log|S_k|).$$

定义 8. 设 e^*为社会网络 G 中具有最大影响力向量和值的成员,即

$$\|\boldsymbol{w}(e^*)\|_1=\max_{e\in G}|\boldsymbol{w}(e)|=\max_{e\in G}\sum_{i=1}^{k}w_{S_j}(e)\cdot\log|S_j| \tag{24}$$

则定义 G 中全体成员的社群隶属度向量为

$$\tilde{\boldsymbol{P}}_i=\frac{\boldsymbol{w}(e_i)}{\|\boldsymbol{w}(A^*)\|_1},i=1,\ldots,N \tag{25}$$

其中,$\|\cdot\|_1$ 代表向量的 1-范数(分量绝对值之和).

易见:经归一化后,除网络中全局影响力最大的网络成员 e^*的社群隶属度值之和$\|\tilde{\boldsymbol{P}}^*\|_1=1$外,其他成员的社群隶属度值之和均小于或等于 1.

定义 9. 设 $\tilde{\boldsymbol{P}}_i$ 为成员 e_i 相对于全体 k 个社群的隶属度向量, $\delta_i=\|\tilde{\boldsymbol{P}}_i\|_1$,称矩阵 $\boldsymbol{P}\in[0,1]^{N\times(k+1)}$为 G 的社群隶属矩阵:

$$\boldsymbol{P}=\begin{bmatrix}\tilde{\boldsymbol{P}}_1 & 1-\delta_1\\ \tilde{\boldsymbol{P}}_2 & 1-\delta_2\\ \ldots & \ldots\\ \tilde{\boldsymbol{P}}_N & 1-\delta_N\end{bmatrix} \tag{26}$$

其中,矩阵 $\boldsymbol{P}$ 的前 k 列由 N 个 k 维行向量 $\tilde{\boldsymbol{P}}_1,\ldots,\tilde{\boldsymbol{P}}_N$ 给出,而最后一列为全体成员对前 k 个社群隶属度总和的剩余分量组成的列向量.

由定义 9 给出的社会网络社群隶属矩阵在形式上与传统的表达方式保持一致,即矩阵元素非负性和行归一性.而改进后的隶属矩阵不仅能够反映个体相对于社群的隶属状态,同时能够更加真实地反映个体在社会网络中的地位和作用.在改进后的隶属矩阵中,每列仍代表一个社群,但其元素的取值不仅与社群的相对规模正相关,而且与个体在该社群的核心度正相关.

4 实验结果与讨论

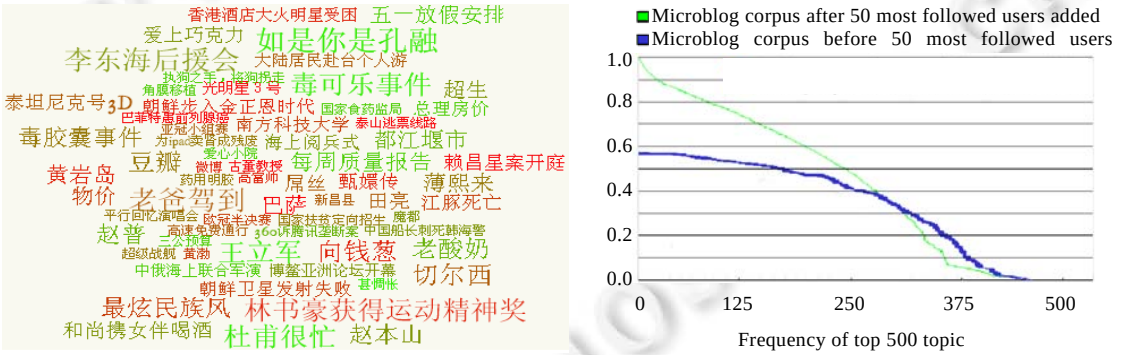
本文以数据堂公司提供的微博用户数据 5 000 条用户数据集为基础[27],通过清洗并补充必要的相关用户数据,然后采用网络爬虫获取上述用户集的微博语料信息,系统地分析和论证本文模型和算法的有效性.

该数据集包含 5 000 个微博用户的用户名、关注用户数、粉丝数和发表的微博数等信息,同时还包含每个用户所关注的用户 id 列表.经去除无效用户及关注数、粉丝数、博文数小于 10 的用户后,有效用户信息为 3 839 条.由于该数据是发布者随机抽取的,其成员间相互加关注的比例极低(<7%),而大多数关注数最多的用户恰恰不包含在该数据集内,因此,如果不将这些被高关注用户纳入到数据集合中,就无法反映该用户集的真实关注关系.为此,从全体用户的关注列表中统计出前 50 名被关注数最多的外部用户加入到用户列表中(均为微博标 V 用户,3 839 个用户对其总加关数 6 460 次.限于篇幅,所添加用户列表及关注统计信息表从略).此时,总用户数达到 3 889 名.

首先,设计程序利用新浪 API 接口从所有 3889 名用户的微博网页中抓取其 2012 年 1 月 1 日~2012 年 4 月 25 日(该数据集采集时间为 2012 年 4 月 23 日~2012 年 4 月 25 日)的全部博文数据和转发、评论信息,所得语料总规模约 47.5 万条(去除了与本数据集无关的其他用户相关博文信息).

首先,采用本文方法对粗语料进行预处理后,采用第 1.3 节所述的 TF-IDF 和 SVD 技术对文本集合进行话题提取.图 2(a)为上述时间段内该用户集的前 100 个博文话题云图.

进一步地,分析加入 50 名高关注用户微博语料对用户话题聚类的影响,分别对合并前后的语料进行话题词频统计.实验结果表明:加入名人微博语料后,热门话题出现的频率发生了显著的变化.原随机抽取 3 839 名用户语料约 31 万条博文中话题较为分散,而加入高关注用户博文语料后话题聚集度显著提升.分析原因:主要在于普通用户间重叠话题较少,而加入高关注对象的博文后,拥有共同话题的比率大幅提高.图 2(b)为加入 50 个网络名人语料前后,前 500 个高频话题词出现的频率对比(粗线为加入前,细线为加入后).



(a) 热门话题词云图 (b) 热门话题词频分布曲线

Fig.2 Word cloud (a) and the frequency curve of hot topic phrases (b) of the user dataset (4.1~4.25, 2013)

图 2 用户微博话题词云图(a)及高频话题短语词频统计曲线(b)

其次,运用本文构建的社群发现模型构建了用户-话题网络,结合用户-用户关注关系、用户-博文评论关系和转发关系形成了一个完整的多模网络.采用本文算法开展社群发现研究,在用户关注关系与话题相似性相等(α=0.5)、关注关系参数β=0.6、转发参数δ_{fwd}=0.25、评论参数δ_{cmt}=0.5 的条件下,分别就不含 50 名高关注对象和包含 50 名高关注对象的两个数据集上获得以下结果:

(1) 在未加入高关注用户数据和博文语料前,社群结构高度分散,几乎无法形成有意义的社群结构(如图 3 所示).图 3 中得到的较大规模社群,是因为该数据集本身包含一个受关注的微博用户(香港 TVB 音乐总监邓智伟),而大量孤立个体存在于网络中.在包含 50 名高关注对象的情况下,社群检测结果发生了显著变化,规模化社群发现数与社群成员之间的链接关系强度显著增长.图 4 为本文方法检测获得的一个典型的由 537 个用户组成

的社群示例图.该图中,社群当前的话题是{泰坦尼克 3D;甄嬛传;平行回忆演唱会},在 7 个具有高入度(被关注度)的节点中,除用户“1697364500”外,其中 7 个用户均来自后加入的高关注用户数据.

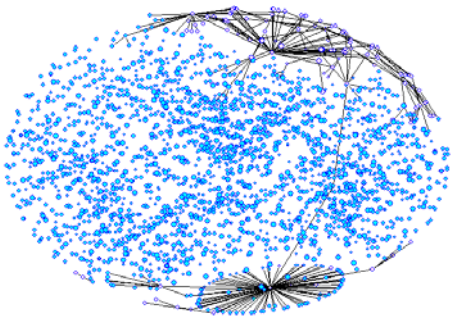


Fig.3 Graph of clustering result before 50 highly followed users' data added

图 3 不包含 50 个微博名人时的社群聚类结果

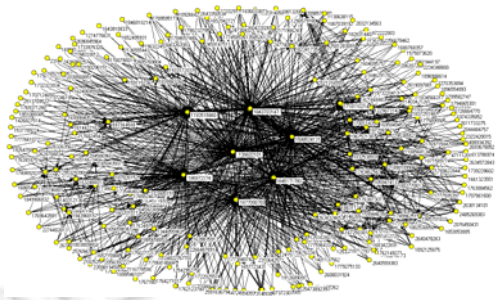
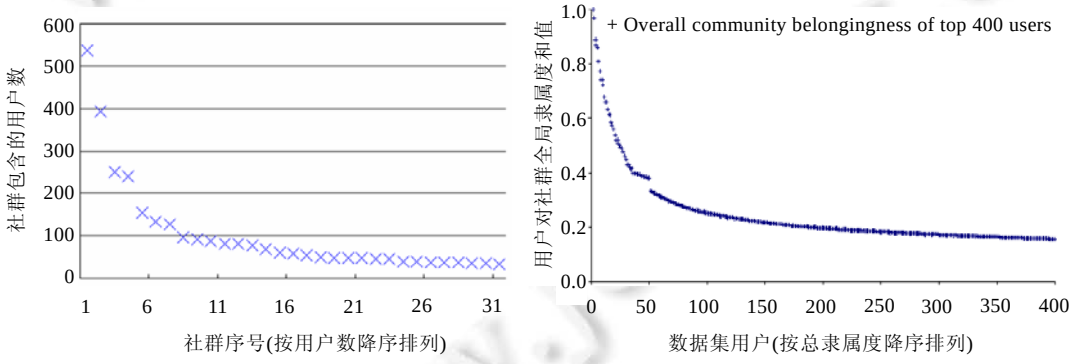


Fig.4 A typical community clustered after highly followed users' data added

图 4 包含 50 个微博名人时检测的典型社群

通过对该实现的分析表明:由于高关注对象的加入,不仅显著提升了博文话题聚焦度,同时也因为关注关系拉近了用户之间的关系,相关的转发与评论的博文显著增加.所有这些都有效提升了用户之间的相似度,从而对社群的发现发挥了显著的提升作用.

(2) 把由优势集聚类 DS-Cluster 算法发现的类团转化为用户话题重叠社群后,分析各个社群中话题和用户的重合率.所采用的数据为加入 50 名高关注用户的数据集,过程中忽略了一些小型社群.实验结果表明,重叠社群在用户规模上同样服从幂率分布.即大规模社群仅占全部社群的极少数,但具有较集中的公共话题集;而小规模社群占绝大多数,且话题相对分散.图 5(a)为所发现的前 30 个社群包含的用户数.



(a) 从数据集中检测的前 30 重叠社群包含用户数 (b) 前 400 用户相对于前 30 个社群的全局隶属度

Fig.5

图 5

(3) 依据第 3 节的社群全局隶属度表示模型,本文将全部 3 889 个用户依据其相对于前 30 个社群的隶属度形成 3889×30 的隶属矩阵.然后,依次统计每个用户对全部社群的隶属度累计值.统计结果表明:依据第 3 节的社群隶属表示模型,个体对社群隶属度(可以认为是参与度)同样符合幂率分布(如图 5(b)所示).即少量个体的社群参与度总值较大(0.5~1.0 之间),而大部分个体的隶属度较小(δ<0.2).图 5(b)中横坐标表示按降序排列的用户编号,纵坐标为其社群全局隶属度和值,横坐标上位于 51 处出现隶属值突变的原因是由用户样本本身造成的(前 50 个用户是其余用户的高关注对象,他们在社群形成中比其余用户发挥了更大的作用).

综上所述,本文通过系统采集和分析微博实验数据,并综合运用所提出的模型与分析方法,全面验证了所提模型与算法的有效性.实验结果能够真实反映微博数据所包含的重叠社群,并能够对实验结果进行合理的解释.

5 结论与展望

微博社会网络是新兴的覆盖海量用户、涉及广泛话题并具有高度动态性的复杂网络.本文以微博社会网络为研究对象,针对多模社会网络重叠社群发现问题展开相关工作,提出了以用户-话题关系为主要划分原则的多模网络重叠社群关系表达模型及相应的社群结构发现算法.该方法不仅考虑网络中的用户-话题关系,还融合了网络特有的关注关系、博文转发与评论关系所形成的复合网络关系.本文同时改进了传统的社群隶属矩阵表述模型,通过引入虚拟社群,使隶属矩阵不仅合理反映个体对社群的隶属度,还标识了个体在社群中的核心度.后续研究工作将拓展到基于本文方法和表示模型的社群演化定量研究等领域.

References:

[1] Java A, Song XD, Finin T, Tseng B. Why we Twitter: An analysis of a microblogging community. In: Zhang H, *et al*., eds. Proc. of the WebKDD/ SNA-KDD. LNCS 5439, Berlin, Heidelberg: Springer-Verlag, 2009. 118−138. [doi: 10.1007/978-3-642-00528-2_7]

[2] Kivran-Swaine F, Govindan P, Naaman M. The impact of network structure on breaking ties in online social networks: Unfollowing on Twitter. In: Desney ST, ed. Proc. of the Annual Conf. on Human Factors in Computing Systems. New York: ACM Press, 2011. 1101−1104. [doi: 10.1145/1978942.1979105]

[3] Zhang Y, Wu Y, Yang Q. Community discovery in Twitter based on user interests. Journal of Computational Information Systems, 2012,8(3):991−1000.

[4] Tang L, Liu H, Zhang JP. Identifying evolving groups in dynamic multimode networks. IEEE Trans. on Knowledge and Data Engineering, 2012,24(1):72−85. [doi: 10.1109/TKDE.2011.159]

[5] Yu LB, Ding C. Network community discovery: Solving modularity clustering via normalized cut. In: Brefeld U, ed. Proc. of the 8th Workshop on Mining and Learning with Graphs. New York: ACM Press, 2010. 34−36. [doi: 10.1145/1830252.1830257]

[6] Huberman BA, Romero DM, Wu F. Social networks that matter: Twitter under the microscope. ArXiv e-prints. http://arxiv.org/abs/0812.1045. [doi: 10.2139/ssrn.1313405]

[7] Gao Q, Qu Q, Zhang XH. Mining social relationships in micro-blogging systems. In: Ant OA, Panayiotis Z, eds. Book: Online Communities and Social Computing. Berlin, Heidelberg: Springer-Verlag, 2011. 110−119. [doi: 10.1007/978-3-642-21796-8_12]

[8] Palla G, Derienyi 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]

[9] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008,10:10008. [doi: 10.1088/1742-5468/2008/10/p10008]

[10] Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010,12(10):103018. [doi: 10.1088/1367-2630/12/10/103018]

[11] Wang XF, Tang L, Gao HJ, Liu H. Discovering overlapping groups in social media. In: Geoffrey I, ed. Proc. of the 10th IEEE Int'l Conf. on Data Mining. IEEE Computer Society, 2010. 569−578. [doi: 10.1109/ICDM.2010.48]

[12] Lancichinetti A, Radicchi F, Ramasco J. Finding statistically significant communities in networks. PLoS One, 2011,6(4):e18961. [doi: 10.1371/journal.pone.0018961]

[13] Gruzd A, Wellman B, Takhteyev YJ, Fortunate S. Imagining Twitter as an imagined community. American Behavioral Scientist, 2011,55(10): 1294−1318. [doi: 10.1177/0002764211409378]

[14] Hazlewood WR, Makice K, Ryan W. Twitterspace: A co-developed display using Twitter to enhance community awareness. In: Simonsen J, ed. Proc. of the Participatory Design Conf. The Trustees of Indiana University, 2008. 230−234. [doi: 10.1145/1795234.1795284]

[15] Meeder B, Karrer B, Sayedi A, Ravi R, Borgs C, Chayes J. We know who you followed last summer: Inferring social link creation times in Twitter. In: Sadagopan S, ed. Proc. of the 20th Int'l Conf. on World Wide Web. New York: ACM Press, 2011. 517−526. [doi: 10.1145/1963405. 1963479]

[16] Lin C, Lin C, Li JX, Wang DD, Chen Y, Li T. Generating event storylines from microblogs. In: Chen XW, ed. Proc. of the 21st ACM inter. Conf. on Information and knowledge management. Maui. ACM Press, 2012. 175−184. [doi: 10.1145/2396761.2396788]

[17] Lin C, Lin C, Lin ZY, Quan Z. Hybrid pseudo relevance feedback for microblog retrieval. Journal of Information Science, 2013, 39(6):773−788.

[18] Yuan Y, Yang CM. Empirical analysis of all kinds of social networks and their relationships formed by information communication among microblog users. Library and Information Service, 2011,55(12):11−25 (in Chinese with English abstract).

[19] Teutle ARM. Twitter: Network properties analysis. In: Palomares RA, ed. Proc. of the Int'l Conf. on 20th Electronics, Communications and Computer. Cholula: IEEE, 2010. 180−186. [doi: 10.1109/CONIELECOMP.2010.5440773]

[20] Gupta M, Gao J, Sun YZ, Han JW. Integrating community matching and outlier detection for mining evolutionary community outliers. In: Yang Q, ed. Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. New York: ACM Press, 2012. 859−867. [doi: 10.1145%2F2339530.2339667]

[21] Jakobsson M, Rosenberg NA. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics, 2007,23(14):1801−1806. [doi: 10.1093/bioinformatics/btm233]

[22] Salton G, Buckley C. Term-Weighting approaches in automatic text retrieval. Information Processing & Management, 1988,24(5): 513−523. [doi: 10.1016/0306-4573(88)90021-0]

[23] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022. [doi: 10.1162/jmlr.2003.3.4-5.993]

[24] Adar E, Teevan J, Dumais ST. Large scale analysis of Web revisitation patterns. In: Czerwinski M, ed. Proc. of the ACM Conf. on Human Factors in Computing Systems (CHI 2008). Florence: ACM Press, 2008. 1197−1206. [doi: 10.1145/1357054.1357241]

[25] Pavan M, Pelillo M. Dominant sets and hierarchical clustering. In: Proc. of the 9th IEEE Int'l Conf. on Computer Vision. Nice: IEEE, 2003. 362−369. [doi: 10.1109/ICCV.2003.1238367]

[26] Pavan M, Pelillo M. Dominant sets and pairwise clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29(1): 167−172. [doi: 10.1109/TPAMI.2007.250608]

[27] http://www.datatang.com/data/45081

附中文参考文献:

[18] 袁毅,杨成明.微博客用户信息交流过程中形成的不同社会网络及其关系实证研究.图书情报工作,2011,55(12):11−25.

胡云(1978－),女,江苏连云港人,博士生,副教授,CCF 会员,主要研究领域为数据挖掘,社会网络分析.
E-mail: 15250998131@139.com

王崇骏(1975－),男,博士,教授,CCF 高级会员,主要研究领域为智能信息处理.
E-mail: chjwang@nju.edu.cn

吴骏(1981－),男,博士,副教授,CCF 会员,主要研究领域为多 Agent 联盟.
E-mail: iip@nju.edu.cn

谢俊元(1961－),男,教授,博士生导师,主要研究领域为分布式人工智能.
E-mail: jyxie@nju.edu.cn

李慧(1979－),女,博士生,CCF 会员,主要研究领域为数据挖掘.
E-mail: shufanzs@126.com