融合标签平均划分距离和结构关系的微博用户可重叠社区发现

马慧芳[1,2],陈海波[1],赵卫中[3],邴 睿[1],黄乐乐[1]
(1. 西北师范大学计算机科学与工程学院,甘肃兰州 730070;2. 桂林电子科技大学,广西可信软件重点实验室,广西桂林 541004;
3. 湘潭大学信息工程学院,湖南湘潭 411105)

摘 要: 提出了一种融合标签平均划分距离和结构关系的微博用户可重叠社区发现算法. 首先从信息论与距离的概念出发,定义基于核心标签平均划分距离的准划分算法;再根据用户关注关系定义结构属性向量,并计算用户结构相异度,进而对核心标签平均划分距离和用户结构相异度进行权重调节,得到综合划分相异度;最后将综合划分相异度最低的标签所划分出的分组作为本次循环的新社区;实验表明,该方法能够识别可重叠社区且具有实际应用意义.

关键词: 可重叠划分; 核心标签; 平均划分距离; 结构相异度; 综合划分相异度
中图分类号: TP393.09 **文献标识码:** A **文章编号:** 0372-2112 (2018)11-2612-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.11.007

Leveraging Tag Mean Partition Distance and Social Structure for Overlapping Microblog User Community Detection

MA Hui-fang[1,2], CHEN Hai-bo[1], ZHAO Wei-zhong[3], BING Rui[1], HUANG Le-le[1]
(1. *Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu* 730070, *China*;
2. *Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi* 541004, *China*
3. *College of Information Engineering, Xiangtan University, Xiangtan, Hunan* 411105, *China*)

Abstract: In this paper, a microblog user community detection algorithm via tag mean partition distance and social structure is proposed. Firstly, through the concept of information theory and distance, a community pre-partition algorithm based on the mean partition distance of core tags is established. Furthermore, a structure attribute vector is defined according to the user's following and follower relationships, based on which the user structure dissimilarity is calculated. Then, the comprehensive division dissimilarity is derived by adjusting the weight of mean distance of core tag and user structure dissimilarity. Finally, the subgroup corresponding to the tag with the lowest comprehensive division dissimilarity degree is considered as a new community for one iteration. Experiments show that the proposed method is effective and has practical significance.

Key words: overlapping community detection; core tag; mean partition distances (MPD) structure dissimilarity; comprehensive division dissimilarity (CDS)

1 引言

微博作为一种复杂的社交网络存在着一种类社区型结构,该结构的特点是网络内部存在节点连接紧密的社区,而社区之间连接比较松散,因此对微博社区的划分有着重要的现实意义.

国内外对于社区发现的研究中,在考虑一个节点是否可以包含在两个以上社区中,可将社区发现分为可重叠社区发现和不可重叠社区发现. 对于不可重叠社区发现,基于模块度最优思想的凝聚类算法成为目前网络社区挖掘的主流,经典算法包括 Fast-Newman 算法[1]等. 另外基于标签传播的社区识别算法是未考虑重叠社区检测的硬聚类算法 RAK[2]. 基于相似度的社区发现方法如相似度模块化函数方法[3]在用户关注关

收稿日期:2017-10-24;修回日期:2018-01-07;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 61762078, No. 61762080);广西可信软件重点实验室研究课题(No. kx201705)

系基础上加入了用户标签概念,但该方法并未考虑到用户标签过少的情况. 随着网络结构的日益复杂化,社区结构中开始出现彼此包含的关系,可重叠社区发现是近年来的研究热点. 有代表性的方法有基于核心标签的可重叠划分方法[4],该方法通过定义划分质量函数可以很好地控制社区重叠度,但对标签作划分依据时所产生的信息关注不够. 基于非负矩阵的半监督社区划分方法[5]利用了标签的先验知识进行训练,在重叠社区发现任务上具有良好的解释性,但无法脱离半监督的先决条件.

本文受信息论和结构相似度的启发,提出一种融合用户自身属性和结构关系的可重叠社区划分方法(Mean Partition Distance and Social Structure, MPDSS). 该方法首先进行用户标签扩充,通过衡量标签区分度大小获取核心标签集,将核心标签作为划分依据,将包含某核心标签的用户群看作一个准社区并计算用该核心标签进行社区划分的标签平均划分距离,再结合该准社区的结构相异度进行权衡,选出综合划分相异度最小的分组作为当前循环产生的新社区,用此策略反复迭代直到满足要求.

2　核心标签划分

本文将微博网络视为有向图,图内元素包括用户节点、有向链接和用户的标签集合;将所有包含某一标签的用户提取出来划分为一个分组,利用标签的划分距离和结构的策略进行修正.

微博网络定义为三元组 $G=(V,E,T)$,用户集合 $V=\{v_1,v_2\cdots v_n\}$,$v_i\in V$ 表示用户节点. $E=\{<v_i,v_j>|v_i$ 关注 $v_j,v_i,v_j\in V\}$. $T=\{t_1,t_2\cdots t_m\}$ 为用户标签集合,m 为标签总数,$T_i=\{t_{i1},t_{i2}\cdots t_{ir_i}\}$ 表示扩充之后的用户的标签集合,且 $\cup_{i=1}^{n}T_i'=\cup_{i=1}^{n}T_i=T=\{t_1,t_2\cdots t_m\}$. T_{core} 为核心标签集合,$T_{\text{core}}\subset T$ 且 $|T_{\text{core}}|=l$. 标签 t_i 对用户集合 V 的划分表示为 $V/t_i=\{C_i,\overline{C}_i\}$,其中子用户集合 C_i 表示包含标签 t_i 的用户集合,$C_i\cup\overline{C}_i=V$.

2.1　标签扩充与核心标签

标签作为用户自身标注信息,代表了用户的总体偏好,是经过长时间积累形成的长期兴趣. 标签所包含的信息有助于建立精确的用户描述[6],由于用户标注的标签数量较少,用户-标签矩阵高维稀疏,被选取的核心标签指示性不强,标签扩充就更具有实际应用意义. 依文献[4]对标签进行扩充,构造 $n\times m$ 的用户-标签矩阵 $\boldsymbol{M}$,其元素 $M(k,i)$ 为 v_k 的标签 t_i 的权重 w_{ki}. 记 $\boldsymbol{M}$ 中第 i 列元素的权重均值为 E_i;记 $\boldsymbol{M}$ 中第 i 列元素的权重方差为 D_i,则 t_i 的标签区分度 TD_i 可表示为 $\text{TD}_i=E_i/D_i$. 标签区分度作为区别核心标签与非核心标签的指示性强度,旨在获取 TD_i 较大的 top-l 个标签作为核心标签并用作后续步骤的社区划分.

2.2　划分熵与归一化距离

Mantaras[7]基于条件熵和联合熵提出了一种基于距离的属性划分方法,将由某属性测试后分出的一组子集称为一个划分(Partition). 本文利用划分的概念,以核心标签作为属性得到相应地划分,并做如下定义:

定义 1(划分熵)　给定核心标签 $t_i\in T_{\text{core}}$,设 t_i 构成的划分为 $V/t_i=\{C_i,\overline{C}_i\}$,$P(C_i)$ 表示取到包含标签 t_i 的用户集合的发生概率,则核心标签 t_i 的划分熵 $H(t_i)$(Partition Entropy)定义为:

$$H(t_i)=-P(C_i)\log_2P(C_i)-P(\overline{C}_i)\log_2P(\overline{C}_i)\tag{5}$$

定义 2(划分条件熵)　给定核心标签 $t_i,t_j\in T_{\text{core}}$,设 t_i,t_j 构成的划分分别为 $V/t_i=\{C_i,\overline{C}_i\}$ 和 $V/t_j=\{C_j,\overline{C}_j\}$,则核心标签 t_j 关于核心标签 t_i 的划分条件熵 $CH_{t_i}(t_j)$(Partition Conditional Entropy)定义为:

$$\begin{aligned}CH_{t_i}(t_j)=&-P(C_j)P(C_i|C_j)\log_2P(C_i|C_j)\\&-P(C_j)P(\overline{C}_i|C_j)\log_2P(\overline{C}_i|C_j)\end{aligned}\tag{6}$$

定义 3(划分联合熵)　给定核心标签 $t_i,t_j\in T_{\text{core}}$,则划分联合熵 $H(t_i,t_j)$(Partition Joint Entropy)定义为:

$$H(t_i,t_j)=H(t_j)+CH_{t_j}(t_i)\tag{7}$$

定义 4(标签划分距离)　划分 V/t_i 和划分 V/t_j 的标签划分距离 $d(V/t_i,V/t_j)$(Tag Partition Distance)定义为:

$$d(V/t_i,V/t_j)=CH_{t_i}(t_j)+CH_{t_j}(t_i)\tag{8}$$

定义 5(归一化标签划分距离)　划分 V/t_i 和划分 V/t_j 的归一化标签划分距离 $d_n(V/t_i,V/t_j)$(Normalization Tag Partition Distance)即 V/t_i 和 V/t_j 的标签划分距离与划分联合熵的比值,定义为:

$$d_n(V/t_i,V/t_j)=\frac{d(V/t_i,V/t_j)}{H(t_i,t_j)}\tag{9}$$

由于上式分母不会为零,所以不会有未定义的情况. 其次,因为分母大于分子永远成立,不会因为分母过小而放大比值,促使选用划分效果不理想的标签.

2.3　平均划分距离

定义 6(标签平均划分距离)　给定核心标签 $t_i,t_j\in T_{\text{core}}$,则用标签 t_i 进行划分的标签平均划分距离 MPD(t_i)(Mean Partition Distance)定义为:

$$\text{MPD}(t_i)=\frac{\sum_{j=1,j\neq i}^{l}[\,d_n(V/t_i,V/t_j)\,]}{l-1}\tag{10}$$

MPD(t_i)作为衡量划分效果差异的度量,可表征用某一标签进行划分与用其他标签进行划分的相异程度. 某标签 t_i 的该值越小,则表明 t_i 划分所得社区与用其他标签划分得社区的相似程度越高,选择 t_i 作为最终社区划分的核心标签理论依据越充足.

3 用户的结构关系

微博是一个典型的有向网络,关注和被关注所产生的连边关系不能被忽略,应作为社区划分的基本要素加以权衡. 本文从用户的社交属性入手引入了基于关注关系的相似度度量方法. 在用户关注关系建模方面,Sun[3]等人通过考虑用户的共同粉丝和共同关注的计数方法来建模用户的结构相似度;Hsu[8]等人提出的AttriRank模型不仅考虑图的结构,还考虑节点的结构属性,在大型社交网络上有良好的扩展性. 考虑到该方法的可移植性和效率均适用于本文微博社区结构关系度量场景,为更好体现准社区的结构相似度偏差程度,本节提出一种结构相异度划分计算方法.

3.1 基于关注关系和节点度的结构相似度

由某核心标签所形成的准社区中用户节点的关注关系邻接矩阵记作 A,记 $A_{ij}=1$ 表示节点 i 关注了节点 j. 同一用户对自身也应有关注关系,同一用户节点自身的直接连边关系也应被考虑进来. 因此初始化矩阵 A 对角线为1,记作 A^*,定义自身是自身的关注者,即 v_1 关注了 v_2,则 v_2 是 v_1 和 v_2 的共同关注,v_1 是 v_1 和 v_2 的共同粉丝.

为了衡量微博用户节点之间的结构相似程度,考虑节点 v_i 结构的6个特征信息构造属性向量:$\boldsymbol{R}_i(\boldsymbol{a}1,\boldsymbol{a}2,\boldsymbol{a}3,\boldsymbol{a}4,\boldsymbol{a}5,\boldsymbol{a}6)$的形式,每个维度的含义如表1所示:

表1 节点的结构属性表

属性	定义	属性	定义
$\boldsymbol{a}1$	粉丝数(入度)	$\boldsymbol{a}2$	关注数(出度)
$\boldsymbol{a}3$	粉丝同配性($\boldsymbol{a}1$/粉丝节点度的平均值)	$\boldsymbol{a}4$	关注同配性($\boldsymbol{a}2$/关注节点度的平均值)
$\boldsymbol{a}5$	共同关注度	$\boldsymbol{a}6$	共同粉丝度

其中 $\boldsymbol{a}1$、$\boldsymbol{a}2$ 直观地给出了用户节点粉丝关系和关注关系. $\boldsymbol{a}3$、$\boldsymbol{a}4$ 体现了一个节点与周围邻居节点的相似匹配程度,表征用户社交参与度与热情度. $\boldsymbol{a}5$、$\boldsymbol{a}6$ 是微博社区用户共同关注与共同粉丝数量,揭示用户节点的社交相似程度,其定义如下[3]:

$$\boldsymbol{a}5=\sum_{k=1}^{n} I(\boldsymbol{A}_{ik}^{*}=1)\times I(\boldsymbol{A}_{jk}^{*}=1)\times\frac{1}{\log_2 I_k} \tag{11}$$

$$\boldsymbol{a}6=\sum_{k=1}^{n} I(\boldsymbol{A}_{ki}^{*}=1)\times I(\boldsymbol{A}_{kj}^{*}=1)\times\frac{1}{\log_2 O_k} \tag{12}$$

社交网络中某节点的入度越大,则该节点对于关注该节点相似度贡献越小;出度越大,此节点对于他所关注的节点之间的相似度贡献越小. 因此,为每个被关注的节点 k 赋权值 $\frac{1}{\log_2 I_k}$ 和关注的节点 k 赋权值 $\frac{1}{\log_2 O_k}$,其中 I_k 是被关注节点 k 的入度,O_k 是关注的节点 k 的出度.

3.2 结构相异度

定义7(节点结构相异度) 给定用户节点 $v_i,v_j\in C_k$,综合用户节点的结构属性信息,v_i 与 v_j 之间的节点结构相异度 $\mathrm{DS}(v_i,v_j)$(Node Structure Dissimilarity Degree)定义为:

$$\mathrm{DS}(v_i,v_j)=1-e^{-\|\boldsymbol{R}_i-\boldsymbol{R}_j\|_2^2}=1-e^{-(\|\boldsymbol{R}_i\|^2+\|\boldsymbol{R}_j\|^2-2\boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{R}_j)} \tag{13}$$

定义8(准社区结构相异度) 给定核心标签 t_i 划分所得的准社区 C_k,则标签 t_i 的准社区结构相异度 $\mathrm{PDS}(t_i)$(Pre-community Structure Dissimilarity Degree)定义为:

$$\mathrm{PDS}(t_i)=\frac{\sum_{v_i,v_j\in C_k}\mathrm{DS}(v_i,v_j)}{|C_k|^2-|C_k|} \tag{14}$$

$\mathrm{PDS}(t_i)$ 是准社区 C_k 结构相似度的偏差程度的直观反映,某准社区的 $\mathrm{PDS}(t_i)$ 越低,该准社区中用户节点间的结构相似性越强,将该准社区作为最终划分结果输出的可能性越大,反之亦然.

3.3 社区综合划分

标签作为用户自身属性信息,是自身兴趣偏好的直接体现,用户间的结构关系作为用户的社交属性信息,隐含了用户社交兴趣的间接信息. 如何选取适当的标准来权衡两者的权重关系,对于最终社区划分尤为重要,因此提出了综合划分相异度的概念对该标准予以阐释.

定义9(综合划分相异度) 给定调节因子 α,综合划分相异度 $\mathrm{CDS}(t_i)$(Comprehensive Division Dissimilarity Degree)定义如下:

$$\mathrm{CDS}(t_i)=\alpha\times\mathrm{MPD}(t_i)+(1-\alpha)\times\mathrm{PDS}(t_i) \tag{15}$$

用核心标签 t 得到的最优划分就是使得 $t=\arg\min_{t_i}\mathrm{CDS}(t_i)$,$t_i\in T_{\mathrm{core}}$成立.

定义10(社区重叠度) 社区重叠度 $\mathrm{overlap}(C_i,C_j)$(Community Overlapping Degree)定义如下[9]:

$$\mathrm{overlap}(C_i,C_j)=\frac{|V_i\cap V_j|}{|\min(V_i,V_j)|} \tag{16}$$

其中 V_i 表示社区 C_i 中的用户集合,$\min(V_i,V_j)$ 表示社区 V_i 或 V_j 中节点最少的某个社区的节点数. 算法1给出了社区综合划分算法流程.

算法1 社区综合划分算法

输入:调节因子 α、社区重叠度阈值 β、核心标签数目 l

输出:所有社区 $C=\{C_1,C_2\cdots C_{\mathrm{sum}}\}$

1. 初始化社区总数 sum=0,社区集合 $C=\varnothing$;
2. 将包含某核心标签的用户群看作一个准社区,依次得到 l 个准社区;
3. 对于每个准社区,依次计算该社区的 $\mathrm{PDS}(t_i)$;
4. 执行准社区划分得到每个准社区对应的 $\mathrm{MPD}(t_i)$;
5. 利用(15)分别计算每个准社区的 $\mathrm{CDS}(t_i)$;

6. 取 $t = \arg\min_{t_i} \mathrm{CDS}(t_i)$ 所对应的核心标签 t 划分出来的准社区 C_k 为当前社区;
7. 如果社区集合 $C = \varnothing$,则将当前社区 C_k 加入社区集合 C 中,更新 sum = sum + 1;
8. 当 $C \neq \varnothing$,则将当前社区 C_k 与 C 中已存在的社区两两计算社区重叠度;如果 $\arg\max_{C_t} \mathrm{overlap}(C_t, C_k) > \beta, C_t \in C$,合并社区 C_t 与 C_k:$C_t = C_t \cup C_k$;否则将当前社区 C_k 加入社区集合 C 中,更新社区总数 sum = sum + 1;
9. 更新 $l = l - 1$,删除当前标签 t_i;
10. 如果 l 不为 0,返回执行步骤 4,否则结束循环并输出所有社区 C;

4　实验与性能分析

为了验证本文提出的方法的有效性,本节设计实验对社区发现算法进行验证,并设计评价指标对实验结果进行评价与分析.

4.1　实验数据分析

本文实验数据通过微博 API① 进行抓取,首先从某学生微博用户开始采用广度优先搜索,沿关注关系获取用户节点数据、用户关注关系和用户标签数据. 形成 3 个数据集:*S*1 选取某用户及其关注列表里的所有人;*S*2 在 *S*1 的基础上进行了扩充,由于用户出度太大的话可能并不是有效用户并且会影响 *S*3 的抓取所以最后需要剔除出度大于 100 的用户;*S*3 以 *S*2 为基础并沿用其获取方法,由于数据量较大导致边界点也较多,所以还需要除去边界点用户(入度为 1 出度为 0),实验数据集如表 2 所示.

表 2　实验数据

数据集	用户数	标签数	关注关系数
*S*1	95	277	941
*S*2	632	2146	6245
*S*3	2339	8567	22675

4.2　评价指标

4.2.1　有向网络的社区模块度

传统社会网络的重叠社区的评价标准是通过文献[10]所建立的重叠社区模块度(Community Modularity)进行评价的,其定义为:

$$\mathrm{EQ} = \frac{1}{R}\sum_{t=1}^{|Cl|}\sum_{v_i \in C_t, v_j \in C_t}\frac{1}{B_i \times B_j}\left(\boldsymbol{A}_{i,j} - \frac{D_i \times D_j}{R}\right) \quad (17)$$

其中,R 为网络节点的总度数,B_i 为节点 v_i 所隶属的社区个数,$\boldsymbol{A}$ 为网络邻居矩阵,D_i 为节点 v_i 的入度出度总和.

4.2.2　社区划分质量函数

定义社区划分质量函数 QFCD(Quality Function of Comprehensive Partition)作为评价指标来综合衡量社区划分结果的质量如下:

$$\mathrm{QFCD} = \frac{d \times \sum_{t=1}^{|C|} NJ(C_t)}{c} \quad (18)$$

其中,将社区链接密度 d(Density)与传导率 c(Conductance)[3] 作为社区内用户节点的社交紧密程度度量,用标签平均相似度 NJ(C_t)(Tag Average Similarity)衡量某社区内不同用户间标签的平均相似程度. QFCD 越高则社区集合内用户标签一致性越强,且兼顾"高社区链接密度"与"低传导率"特征. 重叠社区的模块度 EQ 作为衡量社区一致性的标准,反映的是社区间的模块化程度. 因此 EQ 与 QFCD 均作为后续实验的评价指标.

4.3　实验结果与分析

为了验证本文方法的有效性设计了三个实验. 一是选择适宜的参数 α 并观察不同的 α 对 CDS(t_i)值的影响;二是给定不同的参数 β,通过模块度、社区划分质量函数、社区划分中的社区数量选出最贴合实际的控制重叠度阈值 β;三是选取两个典型的微博社区划分算法与本文方法进行比较.

4.3.1　调节因子 α 的取值

为选择合适的调节因子 α,将 α 作为自变量,给定不同的 α 观察 CDS(t_i)均值的变化. 如图 1 所示,x 轴为 α 在 0 到 1 之间的不同取值,y 轴为 CDS(t_i)均值的变化情况.

在 $\alpha = 0.6$ 时,3 个数据集上 CDS(t_i)均值都达到最低,即将调节因子 α 设定为 0.6 可使得综合划分相异度趋于最小化. 社区划分中标签的比重略高于用户社交结构的比重的情况下,可使得 CDS(t_i)的值在一定范围内降低,使得划分效果达到最佳状态.

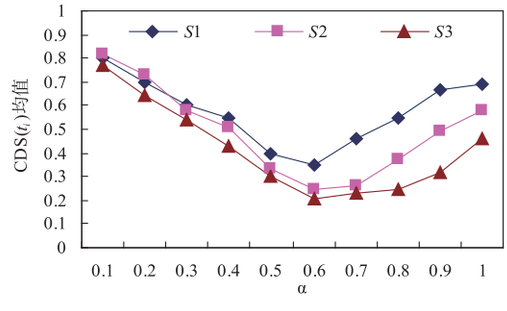

图1　CDS(*ti*)均值随调节因子α的不同取值变化图

4.3.2　控制重叠度的阈值 β 的取值

社区重叠度决定了新划分出来的社区是否与已有社区进行合并,因此控制重叠度的阈值 β 对最终社区数目至关重要,本节实验用模块度、社区划分质量函数、社

① http://open.weibo.com/wiki/API

区数量对β进行判定. 首先将不同社区数下的模块度进行比对,如图 2 所示,x 轴为社区数目的不同取值,y 轴为社区模块度的变化情况.

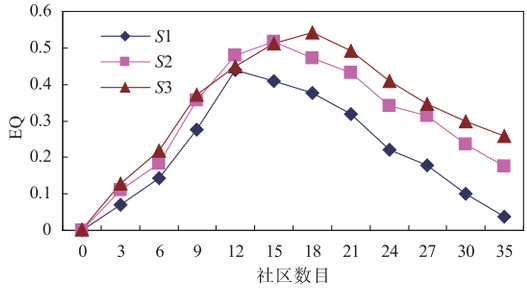

图2 重叠社区的模块度随社区数目的变化图

一个好的社区发现,其模块度应在 0.3 ~ 0.7 之间[1]. *S*1 的模块度在社区为 12 时达到了峰值,*S*2 和 *S*3 则分别在 15 和 18 达到峰值. 其次,进一步观察不同社区的 QFCD 随社区数目的变化情况,如图 3 所示,x 轴为社区数目的不同取值,y 轴为 QFCD 的变化情况.

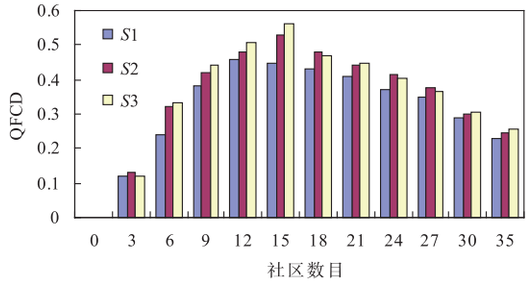

图3 社区划分质量函数QFCD随社区数目的变化图

通过比较得出,*S*1 的 QFCD 在社区数量为 12 时达到峰值,并在社区数目增加时呈平稳下降趋势;*S*2、*S*3 则同样在社区数目为 15 时达到峰值. 最后将不同β下的社区个数进行比较,如图 4 所示,x 轴为阈值β的变化情况,y 轴为社区数目的不同取值.

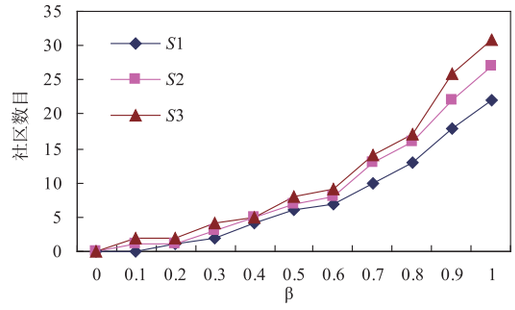

图4 社区数目随阈值β的变化图

通过观察不同β所对应的社区个数,结合模块度以及 QFCD 的变化得出,在$\beta = 0.8$时划分结果中的重叠比例最接近真实结果,即只有重叠度较大的情况下算法才考虑合并社区,避免了合并低相似的社区使得算法效果变差.

4.3.3 与其他算法的比较

为了验证算法的有效性,选取了针对微博社区划分的两个较为典型的算法:基于相似度的发现方法[3]与 MUIC 算法[11]进行参照比较,并采用 EQ、QFCD 对算法进行评估.

首先,在 3 数据集上分别对 3 种方法进行实验,计算 EQ 和 QFCD 值并对实验结果进行综合评估,详细实验结果如表 3 所示.

表 3 3 种方法在 3 个数据集上的结果比较

	*S*1			*S*2			*S*3		
	本文方法	基于相似度的方法	MUIC 算法	本文方法	基于相似度的方法	MUIC 算法	本文方法	基于相似度的方法	MUIC 算法
EQ	**0.474**	0.438	0.334	0.412	**0.455**	0.371	0.514	**0.522**	0.417
QFCD	**0.527**	0.413	0.426	**0.522**	0.397	0.411	**0.537**	0.425	0.431

基于相似度的方法 EQ 值在 *S*2、*S*3 上具有一定优势,该方法所划分形成的用户集合反映出明显的社区结构,而重叠社区划分算法允许同一节点分布于不同社区,因而在一定程度上导致本文方法的 EQ 值较低,但随着用户节点数的增加,该差距在进一步缩小. 社区划分质量方面,*S*3 由于去除了边界点,使得数据集产生的社区内部两两用户节点连边概率提升,进而使得 QFCD 的值在该数据集达到最高值. 应用本文方法所生成的社区因内部用户的标签一致性最强,进而在 QFCD 指标上表现出最优的实验结果. 为了进一步阐释不同数据集中各社区的统一性,本文以 3 种算法划分所得的最大 10 个社区的模块度贡献分布图和 3 个数据集上最大的 10 个社区的模块度贡献值进行微观揭示. 本文方法的模块度贡献值在不同程度上高于其他方法,详细结果如图 5 所示.

基于相似度的方法和本文方法在每个数据集上用户数最多的 10 个社区内部模块度高于 MUIC 算法,原因在于 MUIC 算法删除了有特殊符号或包含英文的个性化标签,生成的标签集未准确反映用户兴趣,划分中产生偏差;从模块度的贡献值方面来说,基于相似度的方法和本文方法与实际社区较为贴合,前者总体趋于稳定但依然略低于本文方法,这是因为前者作为硬聚

类方法,由于标签引入使得社区内部标签相似度提高的同时节点紧密性降低,实际生成的社区中多出现社交结构差异明显但标签相似性较高的节点对,该算法未能在标签与社交关系间做到权衡. 总体来说,本文方法更适合于重叠微博网络的用户社区发现.

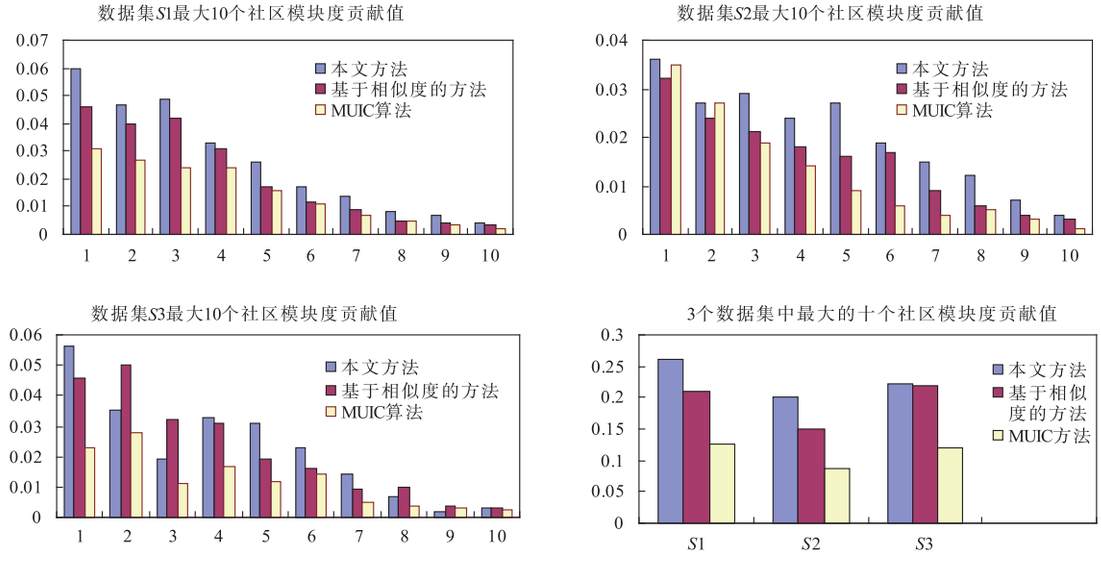

图5　3种算法用户数最多的10个社区内部模块度值分布

5　结束语

鉴于微博网络社区的复杂网络特性,本文提出一种融合标签平均划分距离和结构关系算法,首先通过标签之间的关系进行扩充并发现 top-l 个核心标签. 然后从信息论与距离的概念出发,定义了基于标签平均划分距离的用户划分方法对用户进行分组产生准划分结果,再得到结构相异度值,进而对两者进行权重的调节得到综合划分相异度,最后将综合划分相异度最低的标签所划分出来的簇作为本次循环的最终社区,且在算法中引入参数让社区的重叠度变得可控.

参考文献

[1] Newman MEJ, Girvan M. Finding and Evaluating community structure in networks[J]. Physcial Review E, 2004, 69(2): 026113.

[2] Raghavan UN, A1bert R, KumaraS. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(3): 036106.

[3] 孙怡帆,李赛. 基于相似度的微博社交网络的社区发现方法[J]. 计算机研究与发展, 2014, 51(12): 2797-2807.
Sun Yi-fan, Li Sai. Similarity-based community detection in social network of microblog[J]. Journal of Computer Research and Development, 2014, 51(12): 2797-2807. (in Chinese)

[4] 马慧芳,谢蒙,何廷年,蔺想红. 基于核心标签的可重叠微博网络社区划分方法[J]. 电子学报, 2017, 45(4): 769-776.
Ma Hui-fang, Xie Meng, He Ting-nian, Lin Xiang-hong. An overlapping microblog community detection algorithm via core tags[J]. Acta Electronica Sinica, 2017, 45(4): 769-776. (in Chinese)

[5] Wang Zhao-xian, Wang Wen-jun, Xue Gui-xiang, et al. Semi-supervised community detection framework based on non-negative factorization using individual labels[A]. The Sixth International Conference on Swarm Intelligence[C]. Beijing, China, 2015, 349-359.

[6] 邢千里,刘列,刘奕群,张敏,马少平. 微博中用户标签的研究[J]. 软件学报, 2015, 26(7): 1626-1637.
Xing Qian-li, Liu Lie, Liu Yi-qun, Zhang Min, Ma Shao-ping. Study on user tags in weibo[J]. Journal of Software, 2015, 26(7): 1626-1637. (in Chinese)

[7] RLD Múntaras. A distance-based attribute selection measure for decision tree induction[J]. Machine Learning, 1991, 6(1): 81-92.

[8] Hsu Chin-chi, Lai Yi-an, Chen Wen-hao. Unsupervised ranking using graph structures and node attributes[A]. Tenth ACM International Conference on Web Search & Data Mining[C]. ACM, 2017. 771-779.

[9] Qin Yu, Yu Zheng-tao, Wang Yan-bing, Gao Sheng-xiang, Shi Lin-bin. Approaches to detect micro-blog user interest communities through the integration of explicit user rela-

tionship and implicit topic relations[J]. Science China Information Sciences,2017,60(9):092105.
[10] Shen Hua-wei, Cheng Xue-qi, Cai Kai, Hu Mao-bin. Detect overlapping and hierarchical community structure in networks[J]. Physica A: Statistical Mechanics and its Applications,2009,388(8):1706 – 1712.

作者简介

马慧芳 女,1981 年 7 月出生,甘肃兰州人. 博士,硕士生导师,现为西北师范大学计算机科学与工程学院副教授. 研究领域为数据挖掘与机器学习.
E-mail: mahuifang@ yeah. net

陈海波 男,1993 年 1 月出生,山东淄博人. 西北师范大学计算机科学与工程学院硕士生. 研究方向为机器学习.
Email:605423127@ QQ. com