基于用户关联与主题关注的朋友圈兴趣组发现方法①

石小丹, 王海侠, 吴爱华

(上海海事大学 信息工程学院, 上海 201306)

摘 要: 传统社区发现算法大多考虑因素单一, 联系密切的友人间关注点可能差异较大, 而关注点相同的用户却又可能不在一个朋友圈内. 为此, 提出了一种混合社区发现算法 HCDA, 它既考虑社区网络中的节点关注点, 又考虑了社区网络拓扑结构, 以社区用户间的公共邻居比、关注度及发布微博相似度为依据, 度量相邻节点间的社区关联紧密度. 并以此为基础, 依据相邻节点间的社区增益值, 迭代地扩展社区, 发现朋友圈中真正的兴趣小组. 实验表明, 相较于其他方法, 本算法能够更准确的发现社区.

关键词: 社区发现; 网络分裂; 社区网络拓扑结构; 微博主题模型; 朋友圈

Discovering Interest Groups among Friend Circle Based on Users' Association and Common Topics

SHI Xiao-Dan, WANG Hai-Xia, WU Ai-Hua

(Information Engineering College, Shanghai Maritime University, Shanghai 201306, China)

Abstract: Most traditional community detection algorithms always consider single factor. Friends who have close relationship may have different concerns and users who have common concerns may not be in a circle of friends. To solve the problems, this thesis presents a hybrid community detection algorithm HCDA, which takes into account the concerns of the community network nodes, but also consider the topological structure of community network. On this basis, it expands iteratively the community by the community gain value between adjacent nodes to find the real interest groups among friend circles. The experimental results illustrate that compared with other methods the proposed algorithm can find the community more accurately.

Key words: community detection; network splitting; community network topological structure; micro-blog topic model; circle of friends

微博, 作为一种流行的在线沟通工具, 拥有庞大的用户群体. 随着发布与制造消息的用户数量增大, 一种新的需求被提出, 即精确地为每位用户寻找适合的归属社区. 作为微博社区节点, 每位用户会关注并发表多种主题的博文, 也会与其他节点建立朋友关系. 如此, 将每个节点归属到社区时, 可考虑节点本身内容, 如: 用户浏览行为及发表微博等. 关注的话题与历史博文能反映用户兴趣点, 是区别于他人的节点标签. 此外, 朋友(相互关注的节点)关系也在一定程度上体现了用户的兴趣相似度. 以上两种因素都对节点的社区归属造成影响.

目前已存在许多优秀的社区发现算法, 如 Newman、Girvan[1,2]社区挖掘及 Weiss 与 Jacobson[3]政府部门社团研究等. 然而, 大多社区发现算法存在考虑因素单一的缺点. 如基于主题模型的社区发现方法仅根据用户自身内容, 如关注话题或历史消息划分社区. 虽然主题相关性较强, 但节点间可能毫无联系. 而根据节点间关联情况如基于网络拓扑结构发现社区的算法, 只关注用户的社交关系, 忽略了用户的兴趣点. 同时, 随着用户关注话题及好友联系的变动, 单一方法划分社区会产生不一致现象. 在实际需求中, 内部主题一致与结构连接紧密是划分社区不可缺失的标

① 基金项目:国家自然科学基金(61202022)

收稿时间:2016-09-20;收到修改稿时间:2016-11-07 [doi:10.15888/j.cnki.csa.005800]

准. 因此, 本文紧密结合用户关注的兴趣主题与用户间的相互关联, 同时考虑节点文本信息与网络拓扑结构, 提出了一种综合社区发现算法 HCDA(Hybird community detection algorithm), 使微博用户能够在庞大的社区网络中找到兴趣相同且关系紧密的朋友.

1 相关工作

目前的社区发现方法[4-6], 大多使用如聚类分析[7-9]、基于网络拓扑结构[1,2,10]或基于主题模型的方法[11,12].

聚类分析如模糊聚类、K-Medoids 或 K-Means 算法[7]能够将社区网络节点联系问题转化为数值聚类, 并构造相似性矩阵代表用户的连接情况. Kernighan-Lin 算法[8], 定义两个社区内部总边数减去社区间连接的边数为网络划分的增益函数 Q, 采用贪婪算法原理, 使 Q 值最大. 谱平分法[9]则根据 Laplace 矩阵特征, 即特征值不为 0 对应的特征向量元素中, 同一社区节点对应元素相似的特点进行社区划分, 该方法能够良好的发现社区, 但不适用大规模网络.

基于网络拓扑的社区发现算法主要根据节点的连接关系划分社区, 如分裂算法 GN[1], 通过使用迭代的方式去除网络结构中满足条件的边即经过所有边且最短路径数目最大, 进行社区分裂. Newman 在 GN 的基础上提出了一种凝聚算法 Fast_Newman[2]将单一节点作为独立社区, 依次合并与节点有边相连的其他节点形成社区对. 同时, 标签传播算法 LPA[10]初始化节点为不同的标签, 迭代地将自身标签替换成邻居节点出现频数较多类型, 进行标签传播从而划分社区.

基于主题模型的社区发现算法主要根据用户发布文档语义关联划分社区, 如使用非监督学习的 LDA 主题模型. 算法分割文档 w 为词频向量集, 并获取文档中各单词 d 出现概率. 文档集中所有不重复单词集合为 D. D 中各文档对应不同主题 t 的概率表示为多项分布 θ, 而任意主题产生不同单词的概率表示为多项式分布 ϕ. 对于任一文档 w_l, 计算其 θ 分布获取主题 t_l, 同时根据 ϕ 分布, 从主题 t_l 中获取其对应单词 d_l, 重复遍历所有单词 d, 直至生成文档. 公式表示为:

$$p(d_l)=\sum_{T}^{T} p(d_l \mid t_l=k)p(t_l=k) \tag{1}$$

其中 T 为主题数, $p(t_l=k)$为第 l 个单词抽取第 k 个主题概率, $p(d_l|t_l=k)$表示主题为 k 时抽取单词 d_l 的概率. 它的变形包括 AT (Author-Topic) model[11], HMM-LDA 模型[12]等, 这些算法将主题相关性纳入划分过程中, 改进了原有 LDA 模型主题假设独立缺陷.

2 基于用户关联与主题关注的朋友圈兴趣组发现方法

2.1 节点预处理

微博社区中任意两个用户的共同朋友数越多, 则这两位用户的交友圈越相似. 若两用户的连接强度越大(即互相关注), 则他们的关系越紧密. 而任意两个用户发布微博的内容及话题越雷同, 则两位用户关注的事件及兴趣越相近. 将符合以上条件的用户划分到同一社区既能体现社区兴趣爱好及关注焦点的一致性又能保证用户的关系紧密度. 为此文章综合微博社区用户公共邻居比、节点互相关注情况以及历史微博内容作为考虑节点关联性的标准, 以下给出公式定义:

定义 1. 公共邻居比: 指两相邻用户 U_i 与 U_j 的公共节点重叠度, 是两节点共同邻居与各自相邻节点个数比值, 公式如下:

$$\varphi(U_i,U_j)=\frac{|U_i\cap U_j|}{|U_i\cup U_j|-2} \tag{2}$$

其中 U_j 为与节点 j 相连的节点个数.

定义 2. 关注度: 指相邻用户 U_i 与 U_j 好友关系紧密程度, 主要由节点相互关注与否决定, 用户 U_i 与 U_j 互相关注, 则关注度为 1, 若为单向关注则为 0.5, 公式如下:

$$\eta(U_i,U_j)=\begin{cases}0.5 & one-way\ follow\\ 1.0 & bidirectional\ follow\end{cases} \tag{3}$$

定义 3. 发布微博相似度: 指两相邻用户 U_i 与 U_j 发布微博话题及文章相似程度. 根据用户在不同主题上的概率分布计算而得, 公式如下:

$$\delta(U_i,U_j)=\frac{\sum_t P_{it}P_{jt}}{\sqrt{\sum_t P_{it}^2\sum_t P_{jt}^2}} \tag{4}$$

其中 t 为主题词, P_{it} 为用户 i 在主题 t 上出现的概率, 该概率通过 LDA 模型计算 UDT_{ij} 矩阵得出.

定义 4. 节点关联度: 指两相邻用户 U_i 与 U_j 的综合节点紧密程度, 由公共邻居比、关注度与微博相似度共同决定, 公式如下:

$$R(U_i,U_j)=\sqrt{\varphi^2(U_i,U_j)+\eta^2(U_i,U_j)+\delta^2(U_i,U_j)}\Big/3 \tag{5}$$

将两用户间的关联度看成节点边的权值. 为此,

构造一个微博用户节点关联矩阵 A_{ij}, 矩阵的行与列是微博社区中的用户, 值为用户间的节点关联度. 无边相连的节点, 该值为 0.

2.2 基本思想

通过计算节点间的关联度, 划分同类节点形成关系紧密的社区问题即转换成判断新节点进入社区是否导致社区更加紧密及社区中各节点是否具有更强的关联性问题. 为引入算法, 给出以下定义:

定义5. 节点相似度: 节点 i, j, k 与节点 m 相邻, 则 m 与其相邻节点的节点相似度为 i 与 m, j 与 m 及 k 与 m 的节点关联度之和, 用 $\Gamma_{m-i,j,k}$ 表示, 公式为:

$$\Gamma_{m-i,j,k} = R(U_m,U_i)+R(U_m,U_j)+R(U_m,U_k) \quad (6)$$

定义 6. 候选节点: 与社区内节点相邻(除已形成社区)的节点集合. m 是社区 C_m 内任一节点, 与节点 m 相邻且未被划分到其他社区的节点集合为节点 m 的候选集合, 用 Λ_m 表示.

定义7. 社区节点相似度: 社区 C 内节点相似度为 C 内相邻节点的相似度, 表示为 λ_{in}. 社区 C 内节点与社区外节点相似度为与社区 C 内节点相邻且未被划分到其他社区的节点相似度, 表示为 λ_{out}.

定义8. 社区关联紧密度: 是评价社区内节点关联密度指标, 用社区内所有节点相似度与社区内及社区外节点相似度比值表示. 当社区内节点数为 1 时, 该值为 0, 公式表示为:

$$\Phi_C = \begin{cases} \dfrac{\lambda_{in}}{\lambda_{in}+\lambda_{out}} & |U_i|>0 \\ 0 & |U_i|=1 \end{cases} \quad (7)$$

公式中 $|U_i|$ 表示为社区内节点个数.

定义9. 节点社区增益值: 将节点加入社区的判定标准. 节点 i 加入社区 C 的社区增益值为加入该节点后与加入前的社区关联紧密度之差. 社区 C 中加入节点 i 的社区增益值表示为:

$$\Delta\Phi_{i,C} = \frac{\lambda_{in}^{C+i}}{(\lambda_{in}^{C+i}+\lambda_{out}^{C+i})^{\varepsilon}} - \frac{\lambda_{in}^{C-i}}{(\lambda_{in}^{C-i}+\lambda_{out}^{C-i})^{\varepsilon}} \quad (8)$$

ε 为人工参数, 用以控制社区大小.

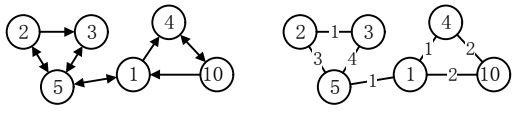
将微博社区中每个节点都视为独立社区, 融合节点相似度越高的两个独立社区对社区的紧密性贡献越大. 因此, 算法从微博网络中挑选任一未被划分的节点进行社区扩张. 构造与社区内节点相邻的用户作为候选集合, 迭代的计算候选节点加入社区与未加入社区时的社区增益值大小. 选择价值最大的候选节点加入社区, 更新社区节点, 重复进行社区扩张直至候选社区中不存在使得社区增益值大于零的节点, 则该社区完成归并. 以同样方式对其余未被划分节点进行扩张, 直至所有社区划分完毕.

算法: HCDA (U,A_{ij}, ε)

Input: Node information U, NodeArray A_{ij}, Parameter ε;
Output: Community-num i, Community C_i

1.Initialize community $u_{iC} \leftarrow null$, $C_i = null$
2.$k = 1$
3.For each u_{iC} is null do //任意未划分节点开始划分
4. $C_k \leftarrow u_i$, $u_{iC} \leftarrow k$
5. While(the nodes of C_k have other connect nodes u_j and u_{jC} is null) //社区存在未划分关联节点
6. Update the candidate node set $\Lambda_{C_k} = \{u_1, u_2 \cdots u_j\}$
7. Calculate the community increment $\Delta\Phi_{\Lambda_{C_k},C_k}$ of candidate nodes //计算候选节点社区增益
8. If (any of $\Delta\Phi_{\Lambda_{C_k},C_k} > 0$)
9. Find out the node u_j max the increment value
10. $u_{jC} \leftarrow k$, $C_k += u_j$ //将该节点加入社区
11. Else
//候选节点社区增益值都为负, 社区扩张结束
12. $k += 1$
13. Break //寻找未被划分其他节点
14. End if
15. End while
16.End for
17.Return(i, C_i)

图 1(a)是一个微博社区的局部节点, 边的指向表示节点关注方向. 根据节点预处理过程可形成如图 1(b). 以节点 1 为例, 进行综合社区发现 HCDA 算法社区划分, 以下给出详细过程. 为方便计算, ε 设置为 1.



(a)局部节点 (b)经节点预处理后社区

图 1 HCDA 算法举例

① 图 1 节点个数为 6, 以各自为中心划分社区.

② 观察以 1 为中心的社区 C_1 ={1}, 与节点 1 相关联的节点分别为 4, 5 与 10. 因此候选邻居节点集 $\Lambda_{\{1\}} = \{4,5,10\}$, 计算候选节点加入社区 1 的增益值:

$$\Delta\Phi_{4,C_1}=\frac{1}{1+2+1+2}-0=\frac{1}{6};\quad \Delta\Phi_{5,C_1}=\frac{1}{1+3+4+1+2}-0=\frac{1}{11};$$

$$\Delta\Phi_{10,C_1}=\frac{2}{1+2+1+2}-0=\frac{1}{3}$$

③ 节点 10 对 C_1 社区增益最大, 因此将节点 10 加入社区 C_1 中, 更新社区 C_1 ={1, 10}, 候选节点集为 $\Lambda_{\{1,10\}}$={5, 4}, 计算候选节点加入社区增益值.

$$\Delta\Phi_{5,C_1}=\frac{1+2}{1+2+3+4+1+2}-\frac{1}{3}=\frac{3}{13}-\frac{1}{3}<0;$$

$$\Delta\Phi_{4,C_1}=\frac{1+2+2}{1+2+2+1}-\frac{1}{3}=\frac{5}{6}-\frac{1}{3}=\frac{1}{2}$$

④ 将节点 4 加入社区的增益值最大, 因此加入此节点. 此时社区 C_1 ={1, 10, 4}, 候选节点为 $\Lambda_{1,10,4}$ ={5}, 计算将节点 5 加入社区 C_1 的社区增益值.

$$\Delta\Phi_{5,C_1}=\frac{1+1+2+2}{1+1+2+2+3+4}-\frac{1}{2}=\frac{6}{13}-\frac{1}{2}<0$$

⑤ 由于加入节点 5, 社区 C_1 增益值小于零, 因此社区 C_1 不再增添新节点. K=2, C_2 ={5}, 更新候选节点集 Λ_5 ={2,3}, 重复以上操作, 直至社区发现完毕.

HCDA 算法主要步骤如图 2 所示.

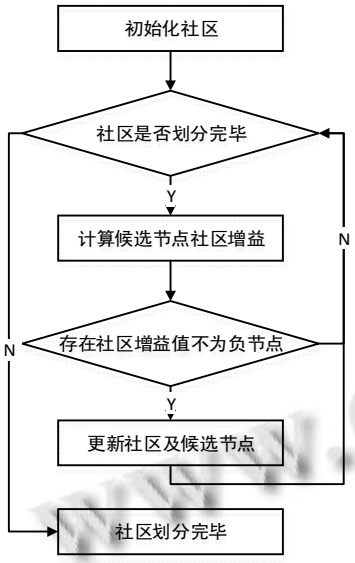


图 2　HCDA 算法流程图

算法 HCDA 可通过调节参数 ε 控制社区大小, 因此, 可以达到局部最优解. 算法计算增益值时间复杂度为 $O(ck)$, k 为迭代的次数. 返回 $\max \Delta\Phi_{\Lambda_{C_k},C_k}$、更新社区 C_k 及邻居节点的时间复杂度为 $O(cn)$, 构造节点相似度矩阵时间复杂度为 $O(n^2)$, 因此算法总时间复杂度为 $O(n^2+ck+cn)$.

3　相关工作实验结果与分析

3.1 数据预处理

本节主要介绍算法数据预处理过程, 包括微博数据爬取、存储、清洗以及获取微博用户主题概率分布.

实验数据: 论文实验数据来源于新浪微博. 使用 Java 实现基于页面解析方式的爬虫设计. 程序通过设定的初始种子 URL, 迭代的访问下载内容. 去掉页面上 HTML 标识, 解析页面内容并保存至数据库中. 同时抽取当前页面新的 URL, 保存至下载队列, 重复爬取直至满足设定条件. 爬取数据过程主要包括预登陆(获取服务器参数), 登录(利用获取参数用户加密及请求参数填充)及解析(提取<script>代码块, 将标准数据存储在数据库中).

数据存储在 oracle 中, 创建用户信息表、微博表与关系表分别存储用户个人信息, 发布微博信息以及关注信息. 数据表设计如图 3, user 为用户表, blog 为用户微博表, relationship 为用户关系表.

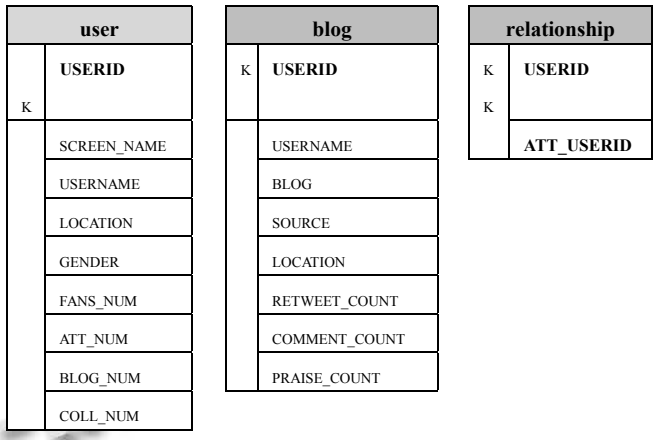


图 3　微博数据表格

实验前, 对获取的数据进行清洗, 包括淘汰无用数据、分词及去停用词. 关注或被关注数小于 3, 历史发布微博数小于 20 的用户微博使用率较少, 本文不予考虑. 经淘汰, 数据表中用户信息为 3296 条, 微博信息为 145365 条, 用户关系信息为 85597 条. 使用 ICTCLAS201 方法对筛选的微博信息进行中文分词, 并去除部分停用词. 根据微博数据实际情况, 构造实验去停用词规则, 包括不能表达语义的助词、叹词、标点符号、特殊数字、英文字母以及一些频繁出现却无实际意义的连词. 下图为实现读取用户发布历史微博数据、分词去停用词及数据存储的用时记录. 图 4

中, 最长用时主要由微博内容较长所致, 当微博内容较短时, 运行时间可降为 1 秒以下.

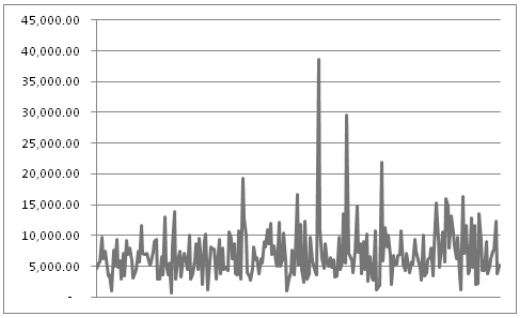


图 4 处理微博内容用时

节点预处理阶段, 发布微博相似度 $\delta(U_i,U_j)$中的 P_{it}, 需要使用 LDA 主题模型计算用户在各主题上的概率分布. 采用 Gibbs sampling 方法实现 LDA 参数估计, 使用 GibbsLDA++-0.2.tar.gz Linux C++进行 LDA 建模. 实验环境为 Redhat6.4, 配置 GibbsLDA++主题数 K 为 50 个, Dirichlet 先验参数 α 为 50/K, Dirichlet 先验参数 β 为 0.01, 保存本地迭代间隔次数为 1000 次, Gibbs 迭代次数为 2000 次, 实验结果将存储为 UDT_{ij} 矩阵. 表 1 展示了部分微博用户及其主题提取内容.

表 1 部分微博用户及其主题提取内容

微博用户	主要提取内容
人民网	时政、中国、人民、报道、传播、改革、法治、反腐、财经、奥运
新浪旅游	旅行、探秘、中国、游记、出境、老照片、北京、摄影、秒拍、自驾游
精彩电影	电影、主演、导演、电影票、作品、大片、好莱坞、高清、烂片、经典
潮音乐	歌曲、颁奖、专辑、花絮、现场、秒拍、华人、献唱、超燃、音乐节
微博汽车	发动机、汽车、宝马、名车、刹车、试驾、离合、电量、车型、舒适

3.2 算法实现与比较

实验环境: Intel 酷睿 i5 CPU, 内存 4G, win7 64 位, 编程环境为 MATLAB R2010b

实验数据: 经过预处理的微博用户信息、发布微博信息以及用户关系信息

比较算法: LPA、GN、HCDA

3.2.1 主题相似度比较

比较社区中各节点的主题相似度是考察算法划分社区能力的重要标准. 社区中各节点主题越相似, 则社区内节点的关联性越强. 因此, 文章对各算法进行了实现, 截取了节点个数前五的社区, 使用 LDA 模型进行社区内节点微博相似度比较. 在实现 HCDA 算法时, 设置社区大小参数 $\varepsilon=1.3$, 以下实验均使用此参数. 社区按照节点个数从大到小排列, HCDA 算法获取节点数前五社区如表 2 所示, 实验结果如图 5.

表 2 节点数前五社区

社区 C	C_7	C_{23}	C_{35}	C_{39}	C_{62}
节点数	247	165	129	93	87

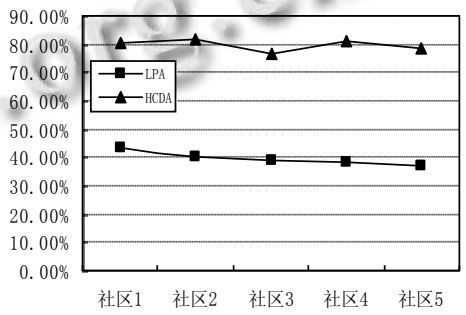


图 5 算法社区主题相似度比较

观察图 5, 在节点数前五的社区中, HCDA 算法各节点微博相似度均大于 LPA 算法, 且相差较多. 同时, 随着节点个数的减少, LPA 算法节点主题相似度逐渐减弱. 而 HCDA 算法节点相似度较稳定, 且能够保持在一个较高水平. HCDA 算法除了考虑 LPA 强调的节点连接属性信息外, 还考虑到了节点发布的微博信息, 因此算法在主题相似度比较上表现良好.

3.2.2 模块度比较

使用 Newman 模块度量值 Q 度量算法划分网络社区的结构强度. 通过改变实验节点个数, 观察比较各算法分解社区的质量. 在实验节点数相同情况下, 各算法均进行 5 次实验, 获取 5 次实验的平均度量值作为最终比较结果, 实验结果如图 6 所示.

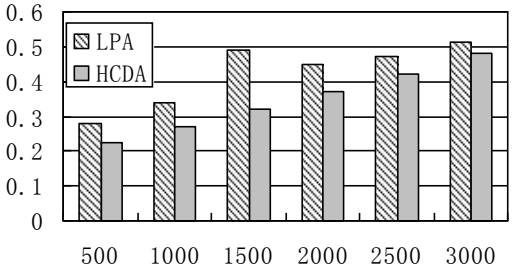


图 6 算法模块度比较

观察图 6, 通过增加实验节点个数, 各算法的模块度 Q 值显著增大. 本文提出的 HCDA 算法模块度 Q 值稍低于 LPA 算法, 但随着数据节点个数增加, 差距逐渐减少, 与 LPA 算法较为接近. 由于 HCDA 算法着重平衡节点连接关系与节点文本信息, 与重点考虑节点连接结构的 LPA 算法相比, 在网络拓扑节点连接紧密度方面能力稍弱. 然而当社区节点数较多时, 平均关联度较大的用户在结构上也会更加完整. 考虑整个实验比较结果, HCDA 算法在划分社区方面能力较强.

3.2.3 综合比较

为体现算法的整体设计, 文章从算法受初始点的影响、算法的时间复杂度等方面进行了综合比较.

不难分析, LPA 算法的时间复杂度为 O(m), GN 算法的时间复杂度为 O(m2n), HCDA 算法时间复杂度为 O(n2+ck +cn), 其中 n ,m 分别为网络中的节点数与边数, k 为 HCDA 算法计算增益值迭代的次数. 初始点的选取对三个算法均无影响, 且三个算法均能考虑节点的拓扑结构, 但只有 HCDA 算法能考虑节点的属性信息. 相比考虑单一因素的 GN 及 LPA 算法, HCDA 能综合考虑网络拓扑结构及节点属性信息, 优势显然. 算法无需选取初始节点, 时间复杂度也适中. GN 算法在社区数目不确定时, 无法确定迭代数目, 且时间复杂度高. LPA 算法时间复杂度低, 但由于其随机选取标签的策略, 识别准确率不稳定. 同时, 在实验中, 使用 HCDA 算法能发现社区数 69 个, 而 LPA 算法仅能发现 47 个. 通过综合评估, HCDA 算法时间性能稳定, 在控制社区节点连接紧密的基础上保持了节点的关注一致度, 能够更准确地划分微博社区.

4 结语

本文提出了一种新型的综合社区发现算法 HCDA, 该方法结合了社区用户间的关联关系、用户关注主题及发布微博的相似度, 综合考虑了节点的文本信息与网络拓扑结构, 能更好地发现朋友圈中的有效兴趣小组. 此外, 算法还可以着重考虑如评论、@、点赞等更具实际意义的用户互动行为, 扩展节点属性信息, 从而提高算法准确性. 同时, 文章还做了充分的实验, 比较了提出的算法与其他几类算法在发现社区拓扑结构、主题相似度及划分结果上的差异. 结果表明, 本文的算法在各项比较指标中均表现较好, 能将社区成员话题兴趣爱好与社交关系紧密结合, 划分出较高质量的社区. 解决了目前已有算法的一些弊端, 且能以新视角构建社会网络. 同时, 随着用户兴趣与时间的改变, 社区划分也会受影响, 因此社区的动态演化支持也可是未来的研究内容之一.

参考文献

1 Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821–7826.

2 Newman MEJ. Fast algorithm for detecting community structure in networks. Physical review E, 2004, 69(6): 066133.

3 Weiss RS, Jacobson E. A method for the analysis of the structure of complex organizations. American Sociological Review, 1955, 20(6): 661–668.

4 淦文燕,赫南,李德毅,王建民.一种基于拓扑势的网络社区发现方法.软件学报,2009,20(8):2241–2254.

5 李磊,倪林.基于模块度优化的标签传播社区发现算法.计算机系统应用,2016,25(9):212–215.

6 孙怡帆,李赛.基于相似度的微博社交网络的社区发现方法.计算机研究与发展,2014,51(12):2797–2807.

7 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the emalgorithm. Journal of the Royal Statistical Society, Series B (Methodological), 1977, 39(1): 1–38.

8 Kernighan BW, Lin S. An efficient heuristic procedure for partitioning graphs. Bell System Technical Journal, 1970, 49(2): 291–307.

9 Pothen A, Simon HD, Liou KP. Partitioning sparse matrices with eigenvectors of graphs. SIAM Journal on Matrix Analysis and Applications, 1990, 11(3):430–452.

10 张俊丽,常艳丽,师文.标签传播算法理论及其应用研究综述.计算机应用研究,2013,30(1):21–25.

11 Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. Meek C, Halpern J, eds. Proc. of the Twentieth Conference Annual Conference on Uncertainty in Artificial Intelligence. Arlington, Virginia. AUAI Press. 2004. 487–494.

12 Griffiths TL, Steyvers M, Blei DM, Tenenbaum JB. Integrating topics and syntax. Saul LK, Weiss Y and Bottou L, eds. Advances in Neural Information Processing Systems. Cambridge, Massachusetts. MIT Press. 2005. 537–544.