基于个人意愿的社会网络团体结构与信息检测方案

汪林玉[1,2],谷 科[1,3],余 飞[1,3],尹 波[1,3],廖年冬[1,3]
(1. 长沙理工大学综合交通运输大数据智能处理湖南省重点实验室,湖南长沙 410114;
2. 湖南信息学院电子信息学院,湖南长沙 410151;3. 长沙理工大学计算机与通信工程学院,湖南长沙 410114)

摘 要: 个人意愿对于形成网络社团和传播信息有着重要的影响力,因此本文提出一种基于个人意愿的社团结构与信息检测方案.该方案中的社团检测算法初次检测以融入节点属性的模块度,再次检测以兴趣度并能发现重叠社团,最后精细检测以个人意愿,本文社团检测算法(ε_CSDA)较之前的算法更有效的是可以发现重叠社团;同时,该方案建立的信息传播模型在指数模型基础上构建边特征向量(边属性)、节点特征向量(节点属性)和意愿向量(用户意愿、社团意愿和节点意愿),并以传播概率和传播延迟构建模型基本关系,从而使得该模型实现了基于个人意愿的信息传播.实验结果表明,加入个人意愿的社团检测和信息传播方案,能够保证社团检测的有效性和实用性,能够实现用户间信息传播的主动性和可靠性.

关键词: 社会网络;个人意愿;社团检测;重叠社团;信息传播
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2019)04-0886-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.04.017

Social Community Structure and Information Detection Scheme Based on Personal Willingness

WANG Lin-yu[1,2], GU Ke[1,3], YU Fei[1,3], YIN Bo[1,3], LIAO Nian-dong[1,3]
1. *Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science & Technology, Changsha, Hunan* 410114, *China*;
2. *School of Electronic Information, Hunan Institute of Information Technology, Changsha, Hunan* 410151, *China*;
3. *School of Computer & Communication Engineering, Changsha University of Science & Technology, Changsha, Hunan* 410114, *China*)

Abstract: Personal willingness is one of the most important factors influencing the construction of social community and the information dissemination in social network. In this paper, we propose a social community structure and information detection scheme based on personal willingness in social network. In our proposed scheme, the social community detection algorithm uses the node attributes to detect social community structure and further find overlapping communities; the information dissemination method is based on the exponential model, which constructs the feature vector by the edge feature and the node feature, the willingness vector by the personal willingness and the community willingness, and the basic relationship by the dissemination probability and dissemination delay. Experimental results show that our proposed scheme can ensure the effectiveness of social community detection and the initiative and reliability of information dissemination.

Key words: social network; personal willingness; community detection; overlapping community; information dissemination

1 引言

随着以互联网为主的社会网络的广泛应用,越来越多的人们参与社会网络的信息传播和信息获取.社会网络以其庞大的规模、复杂的结构和海量的信息,成为学者研究的热点.随着社交网络应用的不断进步,用户之间的交互方式呈现出多样化,社会网络能够尽快地将不同的内容呈现在用户的面前.同时随着互联网

收稿日期:2017-03-31;修回日期:2018-03-27;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61402055, No. 61462048, No. 61504013);湖南省自然科学基金(No. 2018JJ2445, No. 2016JJ3012);"综合交通运输大数据智能处理"湖南省重点开放基金(No. JTXY16B03, No. JTXY16B01, No. JTXY16B05)

的迅猛发展,使得信息传播的方式发生了巨大变革. 信息传播可表现为人们的交互行为,其呈现出多对多、实时性、速度快等多重特性. 正是因为有着信息内容的关联性,才促使用户之间交互行为的产生,进一步动态地改变和影响社会网络的结构和信息传播的方式. 因此,研究社会网络中的相关特性对于锁定关键目标群体、保护用户隐私、互联网舆情监测等具有重要意义[1].

在社会网络中,节点属性、节点的意愿已经成为社会网络社团构成和信息传播的重要影响因素. 其中,个人意愿是节点对外界的主动性或用户对于外界信息获取的能动性. 个人意愿充分考虑了节点自身意愿,因此在社团检测和信息传播中扮演着重要角色,个人意愿值越高,表明节点对外开放程度越高,更容易接收信息和传递信息,反之亦然.

所以,本文在基于个人意愿的基础上提出了社团检测算法和信息传播模型. 一方面,在社团检测算法中,不仅使用模块度检测社团,还使用兴趣度和个人意愿进一步检测社团和发现重叠社团;另一方面,在信息传播中,融合了节点属性、边属性,构建意愿向量,充分考虑用户的主动性. 本文主要利用节点意愿、社团意愿、用户意愿参与社团检测和信息传播模型构建,充分考虑节点的主观能动性,降低出现超大社团的机率,有利于信息传播的稳定性和持续性.

2　相关工作

2.1　基于节点属性的社团检测

目前社会网络不断复杂化,与之最相关的结构就是社团结构. 社团检测算法大多是基于网络结构划分的,主要包括:①基于图划分算法[2];②模块度[3];③边聚类算法[4];④层次聚类[5];⑤种子扩散方法[6];⑥随机游走[7];⑦标签传播方法[8]. 随着对社会网络研究的加深,许多学者开始考虑把节点的属性加入到社团检测中. Viennet 等人[9]在基于鲁汶算法的基础上结合了节点的属性,以模块度和节点属性相似度进行加权求和来构建社团检测算法. Kewalramani 等人[10]利用元数据的相似性(基于相关属性)通过聚类的方式在 Twitter 中检测社团. Deitrick 等人[11]基于用户在一段时间内发送的 Twitter 内容来进行情感分析,进而提高社团检测效率. 孙怡帆等人[12]提出相似度的模块化函数,依据贪心算法思想设计出基于模块化函数最大化的社团发现方法.

2.2　信息传播机制

刻画出社会网络中的信息传播方式,最常用的方式就是建模. 早期的有独立级联模型(Independent Cascades Model,IC)[13]和线性阈值模型(Linear Threshold,LT)[14]. Kempe 等人[15]在独立模型的基础上提出了一种递减级联模型(Decreasing Cascades Model). Lagnier 等人[16]提出了一个基于线性阈值模型的 DRUS(Decaying Reinforced User-Centric),该模型可预测信息如何在网络中传播,影响信息在网络中传播因素包括用户的兴趣度,邻接用户的影响力,用户的传播意愿等. Saito 等人[17]在信息传播 AsIC (Asynchronous IC Model)模型为基础,以传播概率和相邻节点的节点属性向量构建函数,用最大似然方法建立模型,求解传播率和时间延迟参数. Spiro 等人[18]提出了一个时间模型,在 Twitter 平台中给出了信息传播的统计分析,得出影响时间延迟的因素:①用户的相关属性,例如:权威度、活跃度等;②消息的相关属性,例如:标签、url、是否包含热搜事件等.

3　相关定义

加权图 $G(V,E)$ 表示社会网络中的用户关系,V 是网络中的节点集合,E 是两节点间边集合. k 表示图 G 所有的节点数目,$k=|V(G)|$,n 表示图 G 中所有的边数目,$n=|E(G)|$. 即,一个具有 k 个节点的社会网络用邻接矩阵 $\boldsymbol{B}_{k\times k}$ 表示,则

$$B_{i,j}=\begin{cases}1, & \text{如果节点 } i \text{ 与节点 } j \text{ 相连接}\\ 0, & \text{否则}\end{cases} \tag{1}$$

网络的总边数 n 为

$$n=\sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k}B_{i,j} \tag{2}$$

定义 1　节点的度 m_i

与点 i 关联的所有边的数目就是节点的度 m_i,可定义为

$$m_i^{\text{in}}=\sum_{j=1}^{k_{\text{in}}}B_{i,j},m_i^{\text{out}}=\sum_{j=1}^{k_{\text{out}}}B_{i,j} \tag{3}$$

其中,k_{in}是指向节点 i 相关联的所有边的数目,k_{out}是指从节点 i 发出的所有相关联的边的数目. 因此,m_i^{in} 表示节点 i 的入度,m_i^{out} 表示节点 i 的出度.

定义 2　节点意愿 $\varepsilon_{i,j}$

节点意愿 $\varepsilon_{i,j}$是以节点 i 与节点 j 之间的亲密度来衡量. $\varepsilon_{i,j}\in[0,1]$. 取特殊值时,当 $\varepsilon_{i,j}=0$ 时,表示该节点无对外传播的意愿,不具社会行为;当 $\varepsilon_{i,j}=1$ 时,表示该节点对外开放程度很高,具有社会行为. 算法中给出 $\varepsilon_{i,j}$的计算的公式.

定义 3　社团意愿 ε_c

社团意愿 ε_c 以社团与社团之间的联系程度来衡量,$\varepsilon_c\in[0,1]$. 其值是根据社团内部要求来设定,每个社团的总体意愿值是不同的,其信息交互也不同.

定义 4　用户意愿 ε_u

用户意愿 ε_u 是每个用户加入一个社团时,用户根据意愿梯度选择用户意愿值,例如:0,0.25,0.5,0.75,

$1,\varepsilon_u\in[0,1]$. 当 $\varepsilon_u=0$ 时,不考虑用户的个人意愿,对外界不具备传播意愿. 当 $\varepsilon_u=1$ 时,用户接受外界传来的信息和对外转发信息.

定义 5 节点亲密度 $r^{\varepsilon}_{i,j}$

节点亲密度 $r^{\varepsilon}_{i,j}$ 表示节点 i 和节点 j 相互间信息传递的紧密程度,节点意愿 $\varepsilon_{i,j}$ 影响着节点亲密度 $r^{\varepsilon}_{i,j}$. 其定义为

$$r^{\varepsilon}_{i,j}=\frac{p^{\varepsilon}_{i,j}}{\sqrt{p^{\varepsilon}_i\cdot p^{\varepsilon}_j}} \tag{4}$$

$$p^{\varepsilon}_i=\sum_{x=1}^{m^{\text{out}}}p^{\varepsilon}_{i,x},p^{\varepsilon}_j=\sum_{x=1}^{m^{\text{in}}}p^{\varepsilon}_{x,j} \tag{5}$$

$$p^{\varepsilon}_{i,j}=p_{i,j}\cdot\varepsilon_{i,j} \tag{6}$$

其中,$p_{i,j}$ 表示节点 i 与节点 j 之间的传输的原始信息量,与节点意愿 $\varepsilon_{i,j}$ 无关;$p^{\varepsilon}_{i,j}$ 表示由节点意愿 $\varepsilon_{i,j}$ 控制的节点 i 与节点 j 之间传输的信息量.

定义 6 边权重 $w_{i,j}$

边权重 $w_{i,j}$(指有向边的边权重)由加入了节点意愿 $\varepsilon_{i,j}$ 的节点亲密度 $r^{\varepsilon}_{i,j}$ 构建[19]. 其定义为

$$w_{i,j}=\eta\cdot r^{\varepsilon}_{i,j}+(1-\eta)\sqrt{\frac{B_{i,j}}{m_i^{\text{out}}\cdot m_j^{\text{in}}}} \tag{7}$$

其中 η 为影响因子,$\eta\in[0,1]$. 则 η 的值越大,节点亲密度对边权重的影响也越大,反之,则越小. 取特殊值时,当 $\eta=0$,社团结构划分只考虑了拓扑结构;当 $\eta=1$ 时,连边权重就是两节点的亲密度. 因而,引入 η 合理平衡了加入节点属性给边权重带来的影响.

定义 7 模块度增量 ΔD^*

文献[19]提出的模块度增量公式,任意节点 i 加入与其邻接邻居节点 j 所在的社团 c_j,其模块度增量为

$$\Delta D^*=\left[\frac{w_{c_j}+w_{i,c_j}}{w}-\frac{(w^{\text{in}}_{c_j}+w^{\text{in}}_i)(w^{\text{out}}_{c_j}+w^{\text{out}}_i)}{w^2}\right]-\left[\frac{w_{c_j}}{w}-\frac{w^{\text{out}}_{c_j}\cdot w^{\text{in}}_{c_j}}{w^2}-\frac{w^{\text{out}}_i\cdot w^{\text{in}}_i}{w^2}\right] \tag{8}$$

其中,w_{c_j} 表示社团 c_j 内部边权重值之和,w_{i,c_j} 表示节点 i 与社团 c_j 内部节点所有边权重值之和,$w^{\text{in}}_{c_j}$ 是从社团 c_j 外部节点指向社团 c_j 内部节点的所有边的边权重之和,$w^{\text{out}}_{c_j}$ 是从社团 c_j 内部节点指向社团 c_j 外部节点的所有边的边权重之和,w^{in}_i 指向节点 i 的所有边的权重之和,w^{out}_i 是从节点 i 出发的所有边的权重值之和,w 是网络中的所有边的边权重之和.

定义 8 节点兴趣度 $Int(i,j)$

一个消息 cnt_i,可以表示为

$$cnt_i=\{(n_1,w_1);(n_2,w_2);\cdots;(n_N,w_N)$$

其中,消息 cnt_i 中的第 i 个关键字用 n_i 表示,第 i 个关键字的权重用 w_i 表示,按权重降序排列.

节点 i 与节点 j 的兴趣相似度,即节点兴趣度[20]可定义为

$$Int(i,j)=sim(i,j)=\frac{\sum_{n=1}^{m}w_{i,n}\cdot w_{j,n}}{\sqrt{\sum_{n=1}^{m}w^2_{i,n}}\cdot\sqrt{\sum_{n=1}^{m}w^2_{j,n}}}\cdot\varepsilon_{i,j} \tag{9}$$

其中,m 表示公共的关键字数目,$w_{i,n}$ 和 $w_{j,n}$ 表示节点 i 和节点 j 第 n 个公共关键词的权重,节点意愿 $\varepsilon_{i,j}$ 体现了节点间具有相互交流的程度. 如果 $Int(i,j)$ 的值越大,则节点 i 和 j 被检测到一个社团的可能性越大,反之亦然.

4 社团检测算法 ε_CSDA

本文提出的 ε_CSDA 算法,在基于以模块度测度的同时,以兴趣度精细划分社团,兴趣度相同的,更容易划分到同一个社团,一段时间后,以意愿值进行社团内部调整,将不满足意愿条件的节点移出该社团,保证了社团成员的结构稳定以及信息的流通,并综合考虑了各方面的因素,本文的 ε_CSDA 算法不仅能够优于其他算法检测出重叠社团,还能够实现较高的检测品质. 社团结构检测过程如图 1 所示.

(1)算法 1 以模块度检测社团. 在具有 k 个节点的网络 $G(V,E)$ 中,初始化 $\varepsilon_{i,j}$ 为节点 i 和节点 j 的用户意愿均值,由式(7)得到每条边的新权重值. 将网络 G 中每个节点初始化形成一个社团,社团总个数为 k. 对于任意社团 $i(i\in G)$,将社团 i 分别加入其邻居社团集中的每个社团成员 j,并由式(8)计算相应 ΔD^*,比较 ΔD^* 的值,以 arg max 函数选取 ΔD^*($\Delta D^*>0$)值最大时所对应的邻居社团 j,将社团 i 加入到此时的邻居社团 j 中. 只要 ΔD^* 的值在发生变化,合并社团的过程将循环迭代,直到不能在划分更高社团层次为止. 划分结束后返回一组节点组成的社团集列表 C_1,算法 1 检测完成.

算法 1 以模块度检测社团

输入:邻接网络 $G(V,E)$,节点总数 k.

输出:以模块度检测后的节点集组成的社团集列表 C_1.

```
C1←{{},…,{}};  //List 列表集
for 节点 i∈V do
    for 对节点 i 的邻居节点集 N(i) 中所有的成员节点 j do
        初始化节点意愿, ε_{i,j} = (ε_{u_i} + ε_{u_j})/2;  //节点意愿初始化时值为节点 i 与节点 j 的用户意愿值的均值
    end for
end for
将邻接网络 G 中各节点初始化为独立社团;  //初始化 G 为社团集网络
while 还能划分出更高社团层次 do
    Q←{};  //清空
```

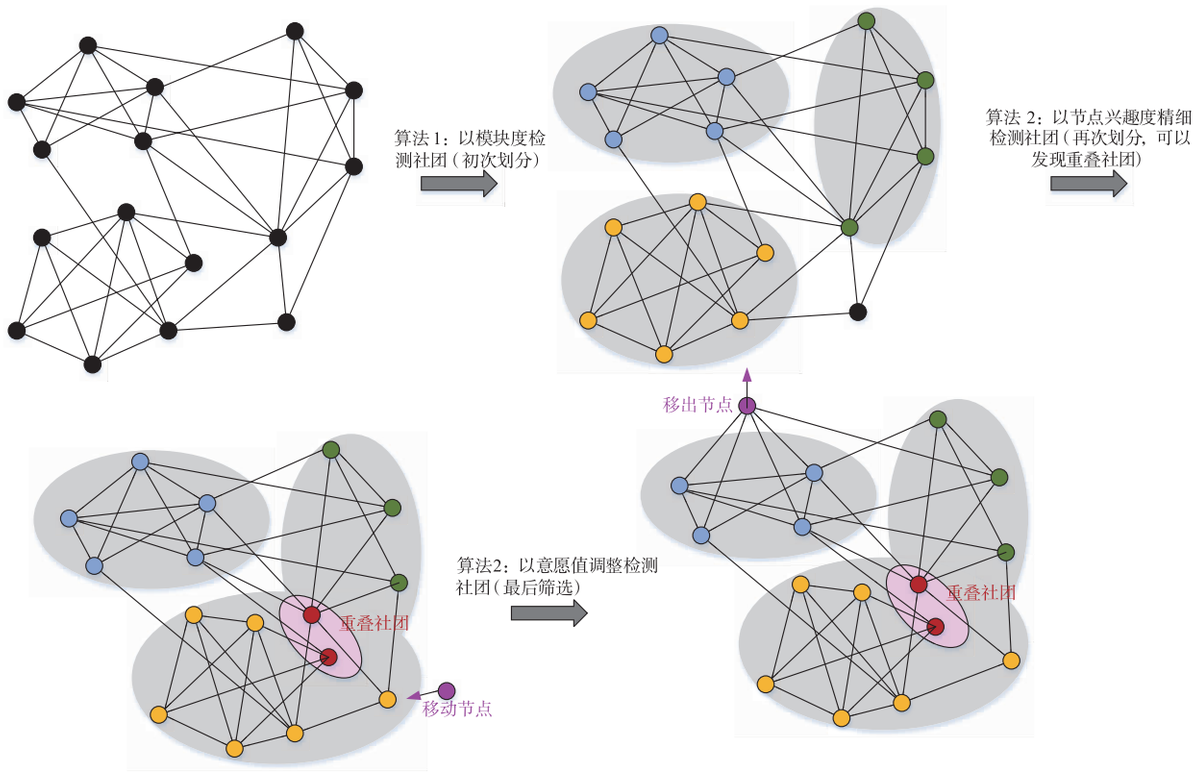


图1　社团检测示意图

```
10.    while ΔD* 值在发生变化 do
11.      for 社团 i∈G do
12.        for 对社团 i 的邻居社团集 N(i) 中所有的成员社团 j do
13.        将社团 i 加入到相邻邻居社团 j,计算加入后的模块度增量 ΔD*,并记录 ΔD* 的值;  //计算模块度 ΔD* 是以整体的社团 i 中的所有节点加入社团 j 来计算的.
14.        if ΔD* >0 then
15.          Q←push(ΔD*,index);  //保存 ΔD* 的值,以及所对应的邻居社团 j 的索引值 index
16.          end if
17.        end for
18.        Max_member←arg max(Q);  //模块度最大时所对应的邻居社团 j 所在的索引值
19.        将社团 i 加入到此时 Max_member 值对应的邻居社团 j;  //此时的社团 i 是整个加入社团 j,包含了社团 i 中的所有节点.
20.      end for
21.    end while
22.  end while
23.  C1←最后不能再划分的邻居节点集组成的社团集列表;
24.  return C1;
```

(2)算法 2 以节点兴趣度精细检测社团. 以算法 1 返回的社团集列表 C_1 为邻接的社团集网络,对于任意社团 $i(i\in C_1)$ 由式(9)计算社团 i 中每个节点所在社团的兴趣度均值,以及与其邻居社团 j 中所有节点的 $Int(i_k,j_k)$ 的值,累加求和,计算邻居社团 j 的兴趣度均值,保存在集合 C_{ner} 中. 将当前节点 $i_k(i_k\in i)$ 所在社团的兴趣度均值与集合 C_{ner} 中的邻居社团兴趣度均值进行比较,满足邻居兴趣度均值大于所在社团均值,且有唯一最大差值时,将节点 i_k 从当前社团 i 移除,加入到唯一最大差值所在的社团中;如果邻居兴趣度均值大于所在社团均值且存在几个社团差值相等且最大,将节点 i_k 从当前社团 i 移除,随机加入到差值最大的其中一个所在的社团,并输出 i_k、i_k 移动后所在社团的编号以及可移入的其他社团编号(如图 1 中移动节点). 直到循环完所有节点,循环结束. 然后计算社团集列表 C_2 中每个社团结构的社团意愿,社团意愿值为社团成员的用户意愿值的均值,将社团意愿的值保存在社团意愿值集 Ω 中. 返回当前划分完节点组成的社团集列表 C_2 和社团意愿值集 Φ,算法 2 检测完成.

算法 2　以节点兴趣度精细检测社团

输入:C_1.

输出:重叠情况下节点的移动信息,以兴趣度检测后的节点集组成的社团集列表 C_2,社团意愿值集 Φ.

```
1.  C2←{{},…,{}};  //List 列表集
2.  C2←C1;  // 将 C1 保存在 C2 中
3.  Cner←{};  //存放邻居社团的兴趣度均值中
4.  for 社团 i∈C1 do
5.    sum1←0;  //初始化保存累加值.
```

6. $n_1 \leftarrow 0$; //初始化变量.
7. $average_{own} \leftarrow 0$; //保存节点 i_k 自身兴趣度均值.
8. for 节点 $i_k \in$ 社团 i do
9. for 节点 $i_m \in$ 社团 i 且 $i_m \neq i_k$ do
10. 计算兴趣度 $Int(i_m, i_k)$;
11. $n_1 + +$;
12. $sum_1 + = Int(i_m, i_k)$;
13. end for
14. $average_{own} = \frac{sum_1}{n_1}$;//计算节点 i_k 自身社团的兴趣度均值
15. $sum_2 \leftarrow 0$; //初始化保存累加值
16. $n_2 \leftarrow 0$; //初始化变量.
17. $average_{ner} \leftarrow 0$;//初始化变量.
18. for 对社团 i 的邻居社团集 $N(i)$ 中所有的成员社团 j 且成员社团 j 随机选取,然后按顺序执行 do
//随机选择一个邻居社团开始,然后按照顺序进行下一个邻居社团.
19. for 节点 $j_k \in$ 社团 j do
20. 计算兴趣度 $Int(i_k, j_k)$;
21. $n_2 + +$;
22. $sum_2 + = Int(i_k, j_k)$;
23. end for
24. $average_{ner} = \frac{sum_2}{n_2}$; //计算当前邻居社团的兴趣度均值
25. $C_{ner}.push(average_{ner})$;
26. $sum_2 \leftarrow 0$; //清零
27. $n_2 \leftarrow 0$; //清零
28. end for
29. if 如果 C_{ner} 集合中的成员邻居社团兴趣度均值 $average_{ner} > average_{own}$ 且有唯一的最大差值 then
30. 将节点 i_k 从当前社团 i 移除,加入到唯一最大差值所在的社团;
31. end if
32. if 如果 C_{ner} 集合中的成员邻居社团兴趣度均值 $average_{ner} > average_{own}$ 且存在几个社团差值相等且最大 then
33. 将节点 i_k 从当前社团 i 移除,随机加入到差值最大的其中一个所在的社团中;
34. record(i_k,i_k 移动后所在社团的编号,可以移入的其他社团编号);//发现重叠情况,记录输出可重叠节点的节点号、移动后所在社团编号,以及可以移动到的社团编号
35. end if
36. end for
37. end for
38. $C_2 \leftarrow$循环完所有节点得到的邻居节点集组成的社团集列表;
39. $\Omega \leftarrow \{\}$; //初始化为空集合,用来存放各社团的社团意愿.
40. for 社团 $i \in C_2$ do
41. $sum_{\varepsilon_c} \leftarrow 0$;
42. for 节点 $i_k \in$ 社团 i do
43. $sum_{\varepsilon_c} \leftarrow sum_{\varepsilon_c} + \varepsilon_{u_{i_k}}$
44. end for
45. $\varepsilon_{c_i} = \frac{sum_{\varepsilon_c}}{|i|}$; //$|i|$表示为社团 i 中节点的个数,ε_{c_i} 为社团各节点的用户意愿值的均值
46. $\Omega \leftarrow push(\varepsilon_{c_i}, i)$; //保存社团 i 的社团意愿值
47. end for
48. return C_2,Ω;

(3)算法 3 以意愿值调整检测社团. 以算法 2 返回的社团集列表 C_2 为邻接社团集网络,社团意愿值集 Ω,对于社团集列表中的每个社团查找其相应的社团意愿 ε_{c_i} 的值. 观察和记录一段时间后,循环社团列表集中的各个社团以及社团中的成员节点,求其成员节点的所有邻居节点的节点意愿之和,根据和求其意愿均值. 如果意愿均值满足条件小于社团意愿值的 $1/\varpi$ (ϖ 为控制社团内部调整的意愿值参数,根据需求给定),说明该节点对外开放意愿太低,不满足社团要求,将被移出当前社团(如图 1 中移出节点),直到循环结束,返回当前检测完成的社团集列表 C,记录社团结构数目,算法 3 检测完成,算法结束.

算法 3 以意愿值调整检测社团

输入:C_2,社团意愿值集 Ω.
输出:以意愿值调整后的节点集组成的社团集列表 C,社团结构数目
1. $C \leftarrow \{\{\},\cdots,\{\}\}$; //List 列表集
2. $C \leftarrow C_2$; // 将 C_2 保存在 C 中
3. for 社团 $i \in C_2$ do
4. for 节点 $i_k \in$ 社团 i do
5. $\varepsilon_{c_i} \leftarrow find(\Omega, i)$; //查找 Ω 集合中社团 i 对应的社团意愿值.
6. $sum_{\varepsilon_{node}} \leftarrow 0$;
7. for 节点 $j_k \in$ 社团 i 且 $j_k \in$ 节点 i_k 的邻居节点集 $N(i_k)$ do
8. $sum_{\varepsilon_{node}} \leftarrow sum_{\varepsilon_{node}} + \varepsilon_{i_k j_k}$;
9. end for
10. $average_{i_k} \leftarrow \frac{sum_{\varepsilon_{node}}}{|N(i_k)|}$; // $|N(i_k)|$为邻居节点集 $N(i_k)$ 中的节点个数,$average_{i_k}$ 表示节点 i_k 与其所有邻居节点的节点意愿值均值
11. if $average_{i_k} < \frac{\varepsilon_{c_i}}{\varpi}$ then //ϖ 表示控制社团内部意愿调整的参数,根据需求给定,若节点意愿值的均值小于 $\frac{\varepsilon_{c_i}}{\varpi}$,表示节点 i_k 对外开放意愿很低,自我封闭意识太过强烈,影响了社团中的信息传播和内部结构的稳定,因而将节点 i_k 从当前社团移除.
12. 将 i_k 从社团 i 移除;
13. end if
//我们给定环境,以一段时间为例,那么在一段时间之后,可能需要重新进行动态的筛选检测,重复算法一,算法二和算法三,在这里需要重新计算社团的相关参数.
14. $compute\_w_{ij}(\varepsilon_{i,j})$; //以调整后的 ε_{ij} 值和相关属性计算社团内部所有节点的边权重 w_{ij}.
15. $sum\_c_i(w_{i,j})$; //社团 i 的边权重为调整后的各节点边权重值之和.
16. end for
17. end for
18. $C \leftarrow$意愿值检测后的邻居节点集组成的社团集列表;
19. return C,记录社团结构数目.

5　信息传播

本文提出的信息传播模型,结合了节点属性特征和信息内容特征,基于指数模型的建模方案,以传播概率 $p(i,j,cnt,\varepsilon)$ 和传播延时 $\tau(i,j,cnt,\varepsilon)$ 构建基础函数,并且引入个人意愿的相关特性. 该模型量化了信息传播的各个因素,保证信息传播的可靠性和安全性.

5.1　提取特征

社会网络中的社团网络进行特征提取,根据其实际情况采取相应的特征提取方法,特征提取可分为节点特征和边特征. 特征提取刻画了相关属性,属性与提取的相关特征是一一对应的[21].

5.1.1　节点特征

(1)传播主体的特征 ψ_s

节点影响力(influence):该节点一段时间内所发布的消息所带来的影响力之和. 则

$$Inf(i) = m_i^{\text{out}} = \sum_{i=1}^{k_{\text{out}}} B_{i,j} \tag{10}$$

节点权威度(authority):该节点的入度 r^{in} 与出度 r^{out} 的差值. 则

$$Auth(i) = \begin{cases} rand()\text{产生 0 到 1 的随机数}, & r^{\text{in}} - r^{\text{out}} < 0 \\ r^{\text{in}} - r^{\text{out}}, & \text{否则} \end{cases}$$

节点活跃度(activity):该节点发布消息的数量,以天计算. 则

$$Act(i) = r_i^{\varepsilon} = \sum_{x=1}^{m^{\text{out}}} r_{i,x}^{\varepsilon} \tag{11}$$

(2)传播客体的特征 ψ_r

节点传播意愿(willing):用户是否愿意传播接收到的信息. 则

$$Will(j) = \lg\left(\frac{transmit}{original} + 1\right) \times \varepsilon_u \tag{12}$$

节点传播特性 ζ:定义 $\zeta_{i\leftarrow j}$ 为消息从节点 j 到节点 i 的传播特性度量.

$$\zeta_{i\leftarrow j} = \frac{\log(1 + \text{Inf}(i))}{\log((1 + \text{Inf}(i)) \cdot (1 + \text{Inf}(j)))} \cdot \varepsilon_u \tag{13}$$

节点 i 和节点 j 的影响力决定节点传播特性的值,特此说明,一般情况下 $\zeta_{i\leftarrow j} \neq \zeta_{j\leftarrow i}$,当 $\text{Inf}(i) \ll \text{Inf}(j)$ 时,$\zeta_{i\leftarrow j} \approx \varepsilon_u$,节点 i 极易接受节点 j 传播的消息,j 的影响力较大,反之亦然.

(3)信息的特征 ψ_{cnt}

针对消息是否包含 url 链接:url 链接到的是一个网址页面,能够更详细的解读消息的内容. 则

$$url(cnt) = \begin{cases} 1, & \text{信息 } cnt_i \text{ 中包含 } url \text{ 链接} \\ 0, & \text{否则} \end{cases}$$

针对消息是否包含标签(label #标签内容#):内容中带有标签信息,能够产生共同的兴趣. 则

$$lab(cnt) = \begin{cases} 1, & \text{信息 } cnt_i \text{ 中包含\#标签内容\#} \\ 0, & \text{否则} \end{cases}$$

5.1.2　边特征

(1)传播主体与传播客体的特征关系 $\psi_{s,r}$

兴趣相似度(interesting):兴趣相同或相似的用户更容易传播信息,完整公式见式(9).

$$Int(i,j) = sim(i,j)$$

是否相互提及:发布过的消息中是否有提到对方的用户名,即为“@ 对方用户名”. 则

$$T(i,j) = \begin{cases} 1, & i,j \text{ 相互提及} \\ 0, & \text{否则} \end{cases}$$

(2)传播客体和传播内容的特征关系 $\psi_{r,cnt}$

传播兴趣(spreading):传播内容的兴趣度,则

$$Spre(U,C) = dist(U,C) = \sum_{k=1}^{n} |u_k - c_k| \tag{14}$$

其中,$\boldsymbol{U} = (u_1, u_2, \cdots, u_n)$ 代表用户 u 的文档向量,$\boldsymbol{C} = (c_1, c_2, \cdots, c_n)$ 代表内容 c 的文档向量. 采用词项 TF-IDF 值来构建文档向量.

以上提取的特征,以其代表的物理意义不同,对其中的一些特征采用了 Min-Max 标准化方法,将取值映射到[0,1]区间.

5.2　建立模型

建立模型[21]的重点是传播概率函数和传播延迟函数. 本文的处理方式是将 5.1 特征提取的结果集表示成节点特征向量、边特征向量,向量的维数分别为节点特征数量 k、边特征的数量 n 和资源数量 r,以节点的意愿值 ε_u、节点间的意愿值 ε_{ij} 和节点所在的整体意愿值 ε_c 构建一个个人意愿向量 ψ_ε. 充分考虑用户、用户之间和用户所在整体的意愿值,体现出意愿的执行力,向量构建过程如下

$$\boldsymbol{\Psi}_k = \begin{Bmatrix} \psi_s \\ \psi_r \\ \psi_{cnt} \end{Bmatrix} \quad \boldsymbol{\Psi}_n = \begin{pmatrix} \psi_{s,r} \\ \psi_{r,cnt} \end{pmatrix} \quad \boldsymbol{\Psi}_\varepsilon = \begin{Bmatrix} \varepsilon_u \\ \varepsilon_{i,j} \\ \varepsilon_c \end{Bmatrix}$$

本文用一个基础函数线性表示,关联相关特性 $\boldsymbol{\Psi}_k$,$\boldsymbol{\Psi}_n$,$\boldsymbol{\Psi}_\varepsilon$,参数 i、j 表示节点 i、节点 j,cnt 表示节点信息,ε 表示加入意愿量,则

$$f'(i,j,cnt,\varepsilon) = \alpha_0 + \alpha_1^{\text{T}}\boldsymbol{\Psi}_k + \alpha_2^{\text{T}}\boldsymbol{\Psi}_n + \alpha_3^{\text{T}}\boldsymbol{\Psi}_\varepsilon \tag{15}$$

用贝叶斯逻辑斯谛函数(Bayesian logistic function)表示传播概率 $p(i,j,cnt,\varepsilon)$,则

$$p(i,j,cnt,\varepsilon) = \frac{1}{1 + \exp\{-f'(i,j,cnt,\varepsilon)\}} \tag{16}$$

其中,α_0 表示一个常量,α_1 表示节点的特征权重,α_2 表示边权重,α_3 表示节点的传播意愿权重. 权重越大,其对所对应的传播概率 $p(i,j,cnt,\varepsilon)$ 的影响越大. (下文出现的 $\boldsymbol{\alpha}$,其表示为 $\boldsymbol{\alpha} = (\alpha_0, \alpha_1^{\text{T}}, \alpha_2^{\text{T}}, \alpha_3^{\text{T}})^{\text{T}}$).

传播时间延迟 $\tau(i,j,cnt,\varepsilon)$,也可表示为 $\boldsymbol{\Psi}_k$,$\boldsymbol{\Psi}_n$,

$\boldsymbol{\Psi}_{\varepsilon}$ 的线性组合. 则

$$\tau(i,j,cnt,\varepsilon)=\beta_0+\beta_1\boldsymbol{\Psi}_k+\beta_2\boldsymbol{\Psi}_n+\beta_3\boldsymbol{\Psi}_{\varepsilon} \quad (17)$$

其中,β_0 表示一个常量,β_1 表示节点的特征权重,β_2 表示边权重,β_3 表示节点的传播意愿权重. 权重越大,对所对应的传播延迟 $\tau(i,j,cnt,\varepsilon)$ 的影响越大. (下文出现的 $\boldsymbol{\beta}$,其表示为 $\boldsymbol{\beta}=(\beta_0,\beta_1^{\mathrm{T}},\beta_2^{\mathrm{T}},\beta_3^{\mathrm{T}})^{\mathrm{T}}$).

模型的构建加入时间衰减因素,但并没有考虑在现实社会网络中,信息传播会随时间间隔变化的传播概率演化. 有研究指出,节点之间信息传播的能力和影响力会随着时间间隔的增大而衰减,是符合指数衰减规律[22]. 因而,本文选择用指数模型来建立信息传播模型. 模型求解过程采用随机梯度下降算法,满足似然概率 $f((j,t_j)\mid(i,t_i);\alpha,\beta,\varepsilon)$ 最大化的参数估计 $\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}$ 为模型的解. 信息传播模型求解过程与文献[21]中的模型的求解过程类似,具体可见文献[21].

6 实验

6.1 数据集

本文基于新浪微博开放的 API 接口,采用爬虫抓取数据集,并引入 Twitter 网络数据集[23],将数据集进行预处理,作为实验的数据集,如表 1 和表 2 所示.

表 1 不同规模的网络数据

网络名称	累计时间	节点数	边数
I1	2010(部分)	812	1290
I2	2012(部分)	1608	2960
I3	2014(部分)	3120	5870
I4	2015(部分)	6029	9620
Twitter	2015(部分)	6312	9267

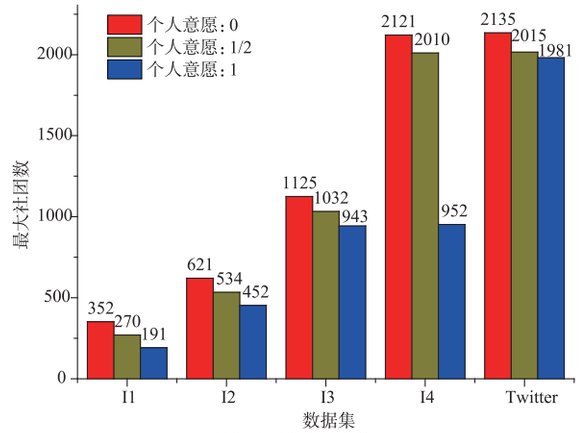
表 2 预处理后的网络数据

用户数量	1294
连接关系(边)	18155
微博消息	2 万
时间跨度	2012. 07. 01 – 2012. 11. 01

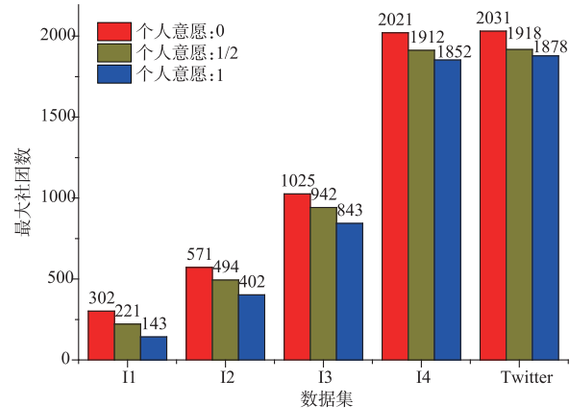
6.2 社团检测实验

本实验对表 1 中的网络数据进行了社团检测,得到社团大小的规模分布情况,给出不同取值的影响因子 η、控制社团内部参数 ϖ,得出相关参数取值不同时对社团检测的影响;给出用户个人意愿取值为 0、1/2、1(最大)时检测出的社团个数分布和最大社团成员数分布,如图 2 所示.

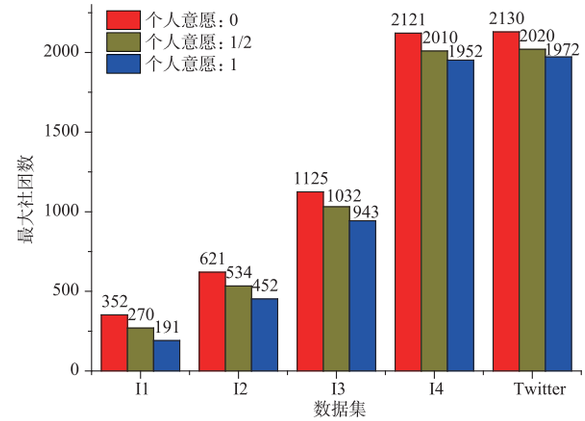
由图 2(*a*)可知,个人意愿的值越大,其最大社团中的成员数越少,说明超大规模社团出现的机会减少. 由图 2(*a*)和图 2(*c*)可知,η 的值越大,最大社团的成员个数轻微减少,说明影响因子对社团检测的影响较少. 由图 2(*a*)和图 2(*b*)可知,ϖ 值减少,最大社团的成员个数轻微减少,说明影响因子对检测出的最大社团中的成员数影响较少. 由以上可知,ε_CSDA 算法引入了个人意愿,使社团划分时可参考的信息更全面,从而降低了出现超大规模社团的机率,提高了社团划分的准确性和可靠性,增加了社团结构的稳定性和有效性.



(*a*) 最大社团中成员个数(η=0.5、ω=3)



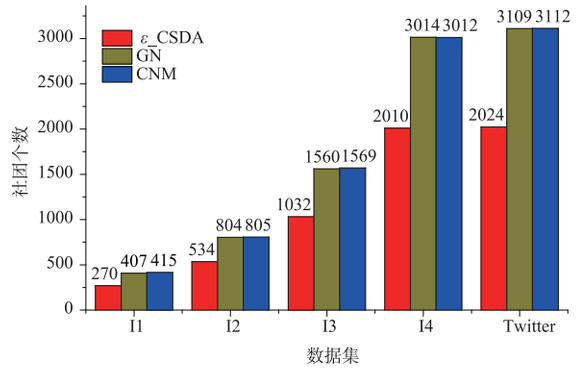
(*b*) 最大社团中成员个数(η=0.5、ω=2)



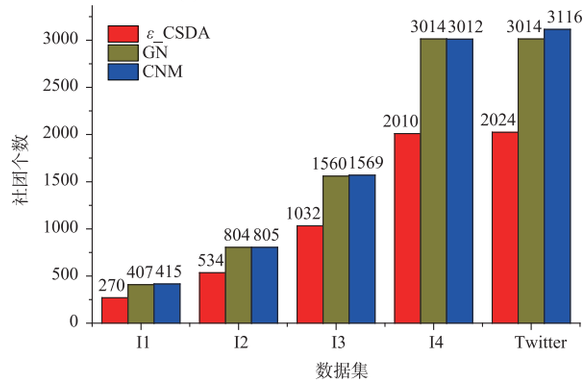
(*c*) 最大社团中成员个数(η=0.8、ω=3)

图2 检测的最大社团成员数分布

本实验采用 GN 算法[24]和 CNM 算法[25]进行对比实验,GN 算法是一种分裂型的社区结构发现算法. CNM 社团发现算法是采用堆的数据结构来计算和更新网络的模块度,它是一种贪心算法. 对表 1 中的网络数据集进行社团检测,依据检测结果得到社团大小的规模分布情况. 算法对比试验中,给出算法运行的相关参量,本文算法 ε_CSDA 给出影响因子 $\eta=0.5$,控制社团内部参数 $\varpi=3$;CNM 算法给出模块度区间为[0,0.5],根据以上参数检测出社团个数分布和最大社团成员数分布对比图,如图 3 所示.



(*a*) 不同算法检测社团个数分布对比图



(*b*) 不同算法检最大社团个数分布图

图3　不同算法检测的社团个数分布

由图 3(*a*)可知,本文社团检测算法 ε_CSDA,检测出的社团个数比 GN 算法和 CNM 算法检测出的社团个数少,说明本文社团检测算法的条件比其他算法更严格,保证了检测出社团的稳定性和有效性. 由图 3(*b*)可知,本文社团检测算法 ε_CSDA,检测出的最大社团成员个数比 GN 算法和 CNM 算法检测出的最大社团成员数少,说明本文算法能够减少出现超大社团的概率,保证了社团划分的准确性和可靠性.

在表 1 给定网络数据集取 I4 网络数据集进行实验,本文算法 ε_CSDA、GN 和 CNM 进行对比实验,计算不同算法的运行效率,网络的平均度为 16,边数从 1000 ~10000 不断递增.

如图 4 所示,不同算法的平均执行时间随着边数的增加呈现增长趋势,曲线在开始一段时间增长速度相对较慢,当边数达到一定数目(2×10^3)时,增长速度相对较快;在相同边数的条件下,本文算法 ε_CSDA 的算法平均执行时间最少,GN 和 CNM 的算法平均时间相差不大. 因而本文算法运行效率相对较高,算法性能相对较好,检测速度相对更快更准确.

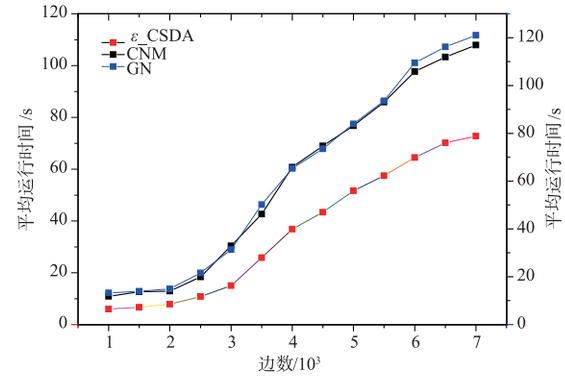


图4　算法的执行效率

6.3　信息传播实验

在信息传播模型中,预处理选取一定数量的节点进行实验. 任意选取实验中三个梯度的关联节点进行实验,为 5 个、10 个、20 个,并进行编号(1-5、1-10、1-20),提取相应的节点特性,对节点传播特性进行实验分析,如图 5 所示. 在此基础上,任意选取实验中的 6 个关联节点,并进行编号(1-6),对这 6 个节点两两节点兴趣度进行实验分析,如图 6 所示.

由图 5 可知,图中个人意愿为 0 的黑色柱状块并没有在图中显示出来,因为个人愿意值为 0 时,其节点传播特性为低特性状态. 由整体来看,5 个节点、10 个节点和 20 个节点的节点传播特性图,都可以表明,个人意愿的值越大,其所占的柱状面积越大,说明个人意愿的值越大,节点传播特性越好,节点对外状态越活跃,越容易进行信息的接收和传递. 表明了个人意愿在信息传播模型中具有一定的执行力.

由图 6 可知,图中个人意愿值为 0 的黑色柱状块并没有在图中显示出来,因为个人意愿值为 0 时,两节点都处于对外拒绝状态,其节点兴趣相似度没有可比性,无实际意义. 由整个图来看,个人意愿的值越大,其所占的柱状面积越大,节点兴趣相似度值越高. 说明个人意愿影响节点兴趣相似度. 表明了个人意愿在信息传播模型中是具有一定的有效性和影响力.

7　结论

在复杂的社会网络中挖掘社团结构,研究信息传播模型,探究与个人意愿的相关性,并深入分析影响的可

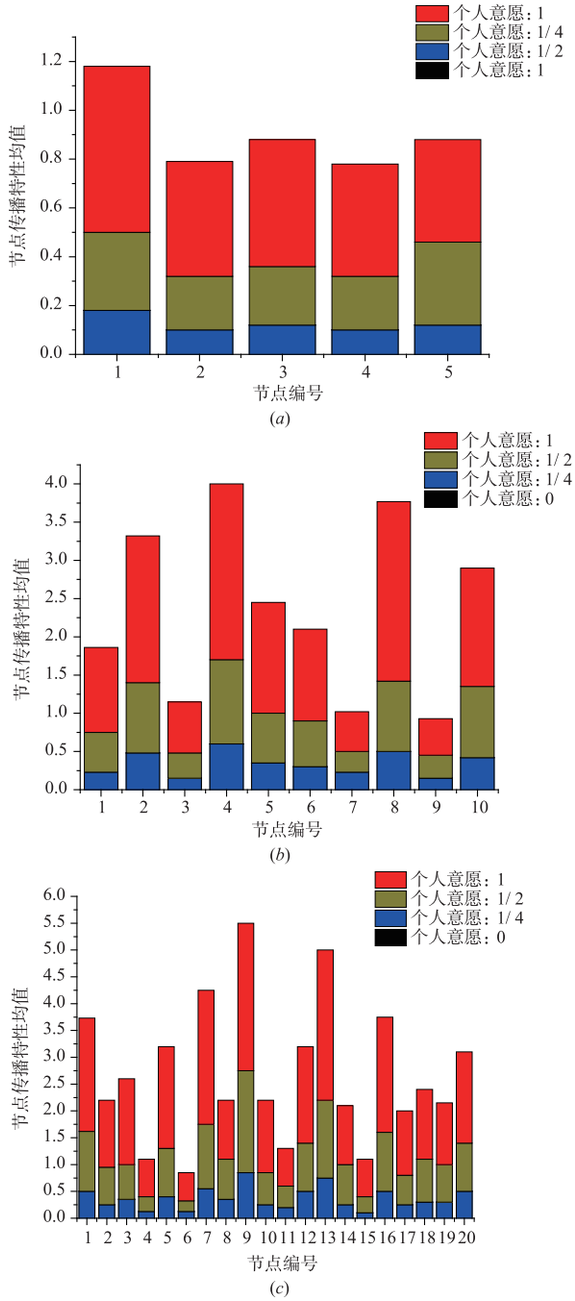


图5 不同个人意愿值的节点传播特性

能因素，对于预防网络犯罪、网络舆情监控等方面有着重要意义. 因此，本文提出了基于个人意愿的社会网络团体结构和信息传播方案：①给出了 ε_CSDA 社团检测算法，该算法结合模块度、兴趣度和个人意愿进行社团检测，在基于个人意愿的基础上，考虑节点的亲密度和兴趣度作为社团划分时可参考的节点关系，从而构建算法实现了重叠社团的发现，提高了社团划分的质量、降低了出现超大社团的规模，实验验证了基于个人意愿的 ε_CSDA 算法在合理的情况下，能够提高社团划分的稳

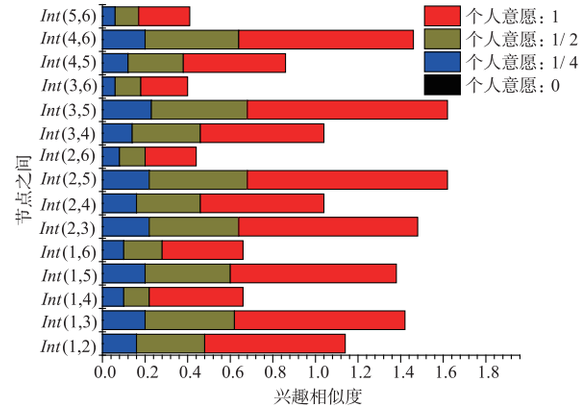


图6 不同个人意愿值的兴趣相似度

定性和可靠性；②在信息传播方面，提出了考虑个人意愿的信息传播模型，该模型提取了用户属性和信息内容等特性，并结合了节点特征向量、边特征向量以及意愿向量，以传播概率和传播延迟进行建模，实验验证了考虑个人意愿的信息传播能够充分考虑用户信息传播的主动性和有效性，保证了信息传播的质量.

参考文献

[1] 余智欣，黄天戍，等. 一种新型的分布式隐私保护计算模型及其应用[J]. 西安交通大学学报，2007，41(8)：954－958.
YU Zhi-xin，HUANG Tian-shu，et al. Novel privacy-protecting distributed computation model and its applications[J]. Journal of Xi′an Jiaotong University，2007，41(8)：954－958. (in Chinese)

[2] SHEN H，CHENG X，CAI K，et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A：Statistical Mechanics and Its Applications，2009，388(8)：1706－1712.

[3] BLONDEL V D，GUILLAUME J L，et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics：Theory and Experiment，2008，(10)：P10－008.

[4] AHN Y Y，BAGROW J P，LEHMANN S. Link communities reveal multiscale complexity in networks[J]. Nature，2010，466(7307)：761－764.

[5] VON LUXBURG U. A tutorial on spectral clustering[J]. Statistics and Computing，2007，17(4)：395－416.

[6] WHANG J J，GLEICH D F，DHILLON I S. Overlapping community detection using seed set expansion[A]. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management[C]. US：ACM，2013. 2099－2108.

[7] PONS P，LATAPY M. Computing communities in large networks using random walks[J]. Journal of Graph Algorithms and Applications，2006，10(2)：191－218.

[8] RAGHAVAN U N，ALBERT R，KUMARA S. Near linear

time algorithm to detect community structures in large-scale networks[J]. Physical Review E,2007,76(3):036106.

[9] DANG T A, VIENNET E. Community detection based on structural and attribute similarities[A]. International Conference on Digital Society[C]. US:ICDS,2012. 7 – 12.

[10] KEWALRAMANI M N. Community Detection in Twitter [M]. Baltimore County, US: University of Maryland, 2011. 231 – 300.

[11] DEITRICK W,HU W. Mutually enhancing community detection and sentiment analysis on twitter networks[J]. Journal of Data Analysis & Information Processing,2013,1(3):19 – 29.

[12] 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)

[13] GOLDENBERG J,LIBAI B,MULLER E. Talk of the network:A complex systems look at the underlying process of word-of-mouth[J]. Marketing Letters,2001,12(3):211 – 223.

[14] YOUNG H P. The diffusion of innovations in social networks[J]. General Information, 2000, 413(1): 2329 – 2334.

[15] KEMPE D, KLEINBERG J, TARDOS É. Maximizing the spread of influence through a social network[A]. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining[C]. US:ACM,2003. 137 – 146.

[16] LAGNIER C, DENOYER L, et al. Predicting information diffusion in social networks using content and user's profiles[A]. European Conference on Information Retrieval [C]. Berlin Heidelberg:Springer,2013. 74 – 85.

[17] SAITO K,OHARA K,et al. Learning diffusion probability based on node attributes in social networks[A]. International Symposium on Methodologies for Intelligent Systems[C]. Berlin Heidelberg:Springer,2011. 153 – 162.

[18] SPIRO E, IRVINE C, et al. Waiting for a retweet: modeling waiting times in information propagation[A]. NIPS Workshop of Social Networks and Social Media Conference[C/OL]. http://snap. stanford. edu/social2012/papers/spiro-dubois-butts. pdf. 2012,12.

[19] 刘瑶,康晓慧,高红,等. 基于节点亲密度和度的社会网络社团发现方法[J]. 计算机研究与发展,2015,52(10):2363 – 2372.

[20] 李致远,陈汝龙,王汝传. 基于兴趣和行为预测的移动社交网络动态资源发现机制[J]. 通信学报,2016,37(4):34 – 43.
LI Zhi-yuan, CHEN Ru-long, WANG Ru-chuan. Exploiting interests and behavior prediction for dynamic resource discovery in mobile social networking[J]. Journal on Communication,2016,37(4):34 – 43. (in Chinese)

[21] 周东浩,韩文报,王勇军. 基于节点和信息特征的社会网络信息传播模型[J]. 计算机研究与发展,2015,52(1):156 – 166.
ZHOU Dong-hao,HAN Wen-bao,WANG Yong-jun. A fine-grained information diffusion model based on node attributes and content features[J]. Journal of Computer Research and Development,2015,52(1):156 – 166. (in Chinese)

[22] GOYAL A,BONCHI F,LAKSHMANAN L V S. Learning influence probabilities in social networks[A]. International Conference on Web Search and Web Data Mining (WSDM)[C]. New York,US:ACM,2010. 241 – 250.

[23] AGARWAL A,XIE B,VOVSHA I,et al. Sentiment analysis of Twitter data[A]. The Workshop on Languages in Social Media[C]. US:Association for Computational Linguistics,2011. 30 – 38.

[24] GIRVAN M, NEWMANM E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences,2002,99(12):7821 – 7826.

[25] NEWMAN M E J. Analysis of weighted networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2004,70(5 Pt 2):056131.

作者简介

汪林玉　女,1993 年 5 月出生,湖南长沙人. 现为湖南信息学院电子信息学院教师. 主要研究方向为社会网络和信息安全.

谷　科　男,1980 年 4 月出生,湖南长沙人. 博士,现为长沙理工大学计算机与通信工程学院硕士生导师. 主要研究方向为网络和信息安全.