基于核心标签的可重叠微博网络社区划分方法

马慧芳[1,2],谢 蒙[1],何廷年[1,3],蔺想红[1]
(1. 西北师范大学计算机科学与工程学院,甘肃兰州 730070;2. 中国科学院计算技术研究所智能信息处理重点实验室,北京 100190
3. 北京师范大学信息科学与技术学院,北京 100875)

摘 要: 针对传统微博社区发现算法内聚低重叠度不可控制等问题,以自顶向下的策略,提出一种基于核心标签的可重叠微博社区发现策略 Tag Cut. 先利用用户标签的共现关系及逆用户频率对标签进行加权,并基于标签之间的内联及外联关系并将用户的标签进行扩充,然后在整体社区中提取包含某一标签的用户作为临时分组并利用评价函数评估划分的优劣,最后选出最合适的核心标签根据其对应分组与其他分组距离的远近来决定将其划分为新的分组还是并入其他分组. 用此策略反复迭代直到满足要求. 该算法划分的组由若干个拥有核心标签的分组组成且综合利用微博用户已声明的及隐含的兴趣、用户之间的关注规律、结果的实用性对划分结果进行修正. 经真实数据实验表明该方法内聚高社区重叠度可控且拥有实际意义.

关键词: 微博网络; 可重叠社区划分; 核心标签; 用户关注关系; 标签划分
中图分类号: TP393. 092 **文献标识码:** A **文章编号:** 0372-2112 (2017)04-0769-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.04.001

An Overlapping Microblog Community Detection Algorithm via Core Tags

MA Hui-fang[1,2], XIE Meng[1], HE Ting-nian[1,3], LIN Xiang-hong[1]
(1. *College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu* 730070, *China*;
2. *Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*;
3. *College of Information Science and Technology, Beijing Normal University, Beijing* 100875, *China*)

Abstract: The traditional microblog community detection algorithm has the characteristic of low coupled clustering and the overlapping degree can not be controlled. In this paper, we present a divisive approach for overlapping microblog community detection algorithm via core tags. Firstly, the key idea is to develop a tag weighing strategy by taking advantage of the co-occurrence of tags and inverse user frequency. Then tag correlation can be exploited, which investigates both inter and intra correlation of tags, and the tags for users can therefore be expanded. Users containing certain tag in the whole community are extracted as a temporary group and the quality value is calculated under the current partition. The most appropriate core tag is selected and the corresponding group is then updated until certain requirements are satisfied. The community detected by this algorithm share common core tags and the partition results can be revised based on the explicit and implicit interest of users, together with the users' attention and practical application. Experimental results show that the method is effective and has practical significance.

Key words: microblog network; overlapping community detection; core tag; user attention relationship; tag cut

1 引言

复杂网络在现实生活中起着非常重要的作用,大量研究人员已针对复杂网络开展了相关研究. 这些网络普遍拥有一种类社区型的结构,这种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. 针对这种社区结构进行分析使得对复杂网络的功能理解和行为预测变的可能,复杂网络的社区发现研究成果已经被成功地运用到诸如恐怖组织鉴别、蛋白质功能预测,Web 社区发掘等众多领域中[1].

收稿日期:2016-01-08;修回日期:2016-08-01;责任编辑:马兰英
基金项目:国家自然科学基金(No. 61363058, No. 61163039);甘肃省青年科技基金(No. 145RJYA259, No. 1606RJYA269);甘肃省自然科学研究基金(No. 145RJZA232);中国科学院计算技术研究所智能信息处理重点实验室开放基金(No. IIP2014-4)

根据是否允许一个节点同时属于多个社区,社区发现主要分成两类:传统的社区发现节点只能属于一个社区,代表算法有 GN 算法和 FN 算法[2,3];重叠的社区发现节点可以同时属于多个社区[4],通常分成两大类:基于节点相似性的算法和基于边相似性的算法[5,6]. 也有一些人从优化目标函数的角度出发进行社区发现[7~9],由于这些办法有着不需要先验信息等优点,所以逐渐成为了社区发现的主流.

微博作为一种新兴复杂网络有着数据规模大、节点用户之间链接不对称、节点分布的度不统一等特点[10],传统办法在微博网络上的应用并不理想,所以越来越多的人开始探索微博社区发现. 其中在微博网络中用标签表示用户的兴趣爱好[11]是进行社区发现时划分的重要标准. 但要想更准确地研究微博社区发现必须结合用户节点自身标签和用户之间的关注关系,较为典型的为 LCA(Lowest Common Ancestor)算法[12],以此为基础周小平[13]发现 LCA 算法的不足提出了微博网络 R-C 模型并对其进行了改进,但该算法的处理对象是用户所发的微博内容,其干扰信息太多,导致该方法很难准确地提取用户的兴趣所在. 孙怡帆[14]在用户关注关系基础上加入了用户标签概念,但此方法并没有考虑到用户标签过少的情况且该算法为不可重叠的社区发现算法. 本文提出一种融合微博用户标签和用户链接关系的可重叠社区发现方法. 通过计算标签之间关联关系对用户标签进行扩充,然后根据是否包含核心标签对微博网络社区进行可重叠的划分,并根据划分更新结果,作为再次选择核心标签的依据. 此外,还提出了一种综合标签与边的评价指标. 以此为基础逐步修正划分结果.

2 用户标签扩充

2.1 用户标签加权方案

尽管用户在添加各个标签的时候彼此是独立的,但标签与标签之间却客观存在着一种潜在的共现关系. 这种关系使标签对用户显示出不同的重要性.

2.1.1 标签共现关系

从整个社区网络上来看,若两个标签经常被同一个用户标注,则认为这两个标签存在共现关系,这种条件概率下的共现关系定义如下:

$$p(t_i|t_j)=\frac{p(t_it_j)}{p(t_j)} \tag{1}$$

通常情况下 $p(t_i|t_j)\neq p(t_j|t_i)$,对其进行对称化处理如公式(2)所示:

$$co(t_i,t_j)=p(t_i|t_j)\cdot p(t_j|t_i) \tag{2}$$

用户的标签互相之间拥有着共现关系,若用户某标签与该用户的其他标签共现关系都很强则认为这个标签对该用户是重要的:

$$cow(v_k,t_i)=\frac{\sum_{t_j\in v_k} co(t_i,t_j)}{|v_k|} \tag{3}$$

在公式(3)中 t_i 和 t_j 都被同一用户 v_k 所标注,$|v_k|$ 表示用户 v_k 中标签的数量.

2.1.2 标签加权

为了更准确地表示标签权重需要考虑标签对用户的代表性,类似逆文档频率 IDF(inverse document frequency)定义逆用户频率 IUF(inverse user frequency),即标注某标签用户数量占总用户数的比值:

$$IUF(t_i)=\log_2\left(\frac{n}{uf(t_i)}+1\right) \tag{4}$$

其中 n 表示社区中用户的个数,$uf(t_i)$ 表示所有标注有标签 t_i 的用户的个数. 综合标签的关联权重和 IUF 值对其进行加权,在用户 v_k 中标签 t_i 的权重定义如下:

$$w_{ki}=cow(v_k,t_i)\cdot IUF(t_i) \tag{5}$$

2.2 标签扩充

2.2.1 标签内联关系

对标签的内联关系可作如下定义[15]:

定义 1 若两个标签被同一个用户所标注,则这两个标签间存在内联关系.

t_i 和 t_j 是被用户 v_k 标注的标签,故标签 t_i 和 t_j 在用户 v_k 中具有内联关系,由 Jaccard 相似度公式,t_i 和 t_j 的内联关系由公式(6)定义:

$$LIR(t_i,t_j)=\frac{1}{|H|}\times\sum_{v_k\in H}\frac{w_{ki}w_{kj}}{w_{ki}+w_{kj}-w_{ki}w_{kj}} \tag{6}$$

其中:w_{ki}和 w_{kj}分别表示被用户 v_k 标注的第 i 个标签 t_i 与第 j 个标签 t_j 的权重. $H=\{v_k|(w_{ki}\neq 0)\&(w_{kj}\neq 0)\}$. 若 $H=\phi$,则 $LIR(t_i,t_j)=0$. 由公式(7)可得标签间内联关系并归一化如下:

$$N-LIR(t_i,t_j)=\begin{cases}1, & i=j\\ \dfrac{LIR(t_i,t_j)}{\sum_{i=1,i\neq j}^{n}LIR(t_i,t_j)}, & i\neq j\end{cases} \tag{7}$$

其中,n 表示标签的数量,$N-LIR(t_i,t_j)$表示被同一用户标注的标签 t_i 和 t_j 的内联关系.

2.2.2 标签外联关系

对标签的外联关系可作如下定义:

定义 2 若有标签同时被用户 v_1 和 v_2 共同标注,那么分别在 v_1,v_2 中与该标签有内联关系的两个标签间会存在外联关系.

给定了标签 t_i 和 t_j,至少存在一个标签 t_q,使得 $N-LIR(t_i,t_q)>0$ 且 $N-LIR(t_j,t_q)>0$,那么称标签 t_i 和 t_j 是具有外联关系的,其中标签 t_q 为标签 t_i 和 t_j 的链接标签,标签 t_i 和 t_j 通过标签 t_q 的外联关系如下:

$$LOR(t_i,t_j \mid t_q)=\min(N-LIR(t_i,t_q),N-LIR(t_j,t_q)) \tag{8}$$

对所有与标签 t_i 和 t_j 具有外联关系的链接标签计算外联关系并规范化至[0,1]之间如下:

$$N-LOR(t_i,t_j)=\begin{cases}0, & i=j\\ \dfrac{\sum\limits_{\forall t_q\in T_q} LOR(t_i,t_j \mid t_q)}{\mid T_q \mid}, & i\neq j\end{cases} \tag{9}$$

其中,$T_q=\{t_q \mid (N\text{-}LIR(t_i,t_q)>0)\&(N\text{-}LIR(t_j,t_q)>0)\}$. 若 $T_q=\phi$,则标签间没有外联关系. 公式(9)表明若两个标签共有的链接标签越多,这两个标签的关系就越密切,其标签间的外联关系就越大.

至此,充分挖掘出标签间的全部的关联关系. 通过如下公式计算出标签间关联关系:

$$LR(t_i,t_j)=\begin{cases}1, & i=j\\ \alpha\times N-LIR(t_i,t_j)+(1-\alpha)\cdot N-LOR(t_i,t_j), & \text{otherwise}\end{cases} \tag{10}$$

其中,$\alpha\in[0,1]$决定了标签间的内联关系与外联关系所占的比例.

最后对用户标签进行扩充,若用户 v_k 标注了标签 t_i 且没有标注 t_j 同时 $LR(t_i,t_j)\geqslant\theta$ 则将标签 t_j 扩充到用户 v_k 中,对微博用户标签进行扩充后的新标签能够更充分的表示用户的潜在兴趣,使下一步的社区划分更为准确.

3　划分算法

微博社区网络中存在着可重叠的用户节点、有方向的关注关系及用户节点自身的标签,本文采用一种图的策略对其进行更合理的修正. 已经有研究人员对图聚集进行了详细的研究及应用[16],受此启发本文将微博网络视为一种有向图. 图内元素包括用户节点,有向链接和用户节点的标签集合.

设整个社区网络为 $G=(V,E,T)$,其中 $V=\{v_1,v_2,v_3\cdots v_n\}$,$v_i\in V$ 表示用户节点,$E=(<v_i,v_j>\mid v_i\in V,v_j\in V)$表示用 v_i 对用户 v_j 的关注关系,$T=\{t_1,t_2,t_3\cdots t_m\}$ 表示用户的标签. $C=\{C_1,C_2,C_3\cdots C_k\}$ 是 G 的一个划分结果,C_i 内部所有标签的集合为 T_{Ci}.

3.1　划分评价

直观上,一个好的社区划分结果应该满足三个条件:

(1)单个社区 C_i 中用户的标签应该尽量相似,关注关系应趋向于社区内部.

(2)不同社区 C_i,C_j 的标签应该尽量的不相似,关注关系应趋于一致化.

(3)划分结果应具有实用意义.

针对以上三个条件提出一种改善划分结果的评价函数,该函数分别引进三个评价指标对划分结果进行综合修正.

3.1.1　社区内聚度

定义 3　设用户 v_i 的标签集合为 T_i,用户 v_j 的标签集合为 T_j,则两个用户的标签相似度为:

$$TagSim(v_i,v_j)=\frac{\mid T_i\cap T_j\mid}{\mid T_i\cup T_j\mid} \tag{11}$$

定义 4　设用户 v_i,v_j 都是社区 C_l 的用户,T_i 和 T_j 分别是其对应的标签集合,则社区 C_l 的标签内聚度为:

$$TagCohesion(C_l)=\frac{\sum\limits_{v_i\in C_l}\sum\limits_{v_j\in C_l} TagSim(v_i,v_j)}{\mid C_l\mid^2} \tag{12}$$

定义 5　设 $C=(C_1,C_2,C_3\cdots C_k)$ 是网络社区 $G=(V,E,T)$的一种划分结果,则 C 的标签内聚度为:

$$TagCohesion(C)=\frac{\sum\limits_{C_l\in C} TagCohesion(C_l)}{k} \tag{13}$$

社区的标签内聚度反应了社区内部用户标签的重叠程度,若标签的重叠度高则认为社区的标签内聚度高,重叠度低则认为社区的标签内聚度低.

定义 6　设用户 v_i 是社区 C_l 中的用户,e_{ij}是用户 v_i 向用户 v_j 关注关系所对应的边,则社区 C_l 的边分散度为:

$$EdgeDispersion(C_l)=\frac{\left\mid\sum\limits_{v_i,v_j\in C_l} e_{ij}\right\mid}{\left\mid\sum\limits_{v_i\in C_l} e_{ij}\right\mid} \tag{14}$$

社区的边内聚度反应了社区内用户向关注社区内部用户所对应的边占总边数的比例.

定义 7　设 $C=\{C_1,C_2,C_3\cdots C_k\}$ 是网络社区 $G=(V,E,T)$的一种划分结果,则 C 的边内聚度为:

$$EdgeCohesion(C)=\frac{\sum\limits_{C_l\in C} EdgeDispersion(C_l)}{k} \tag{15}$$

定义 8　设 $C=\{C_1,C_2,C_3\cdots C_k\}$ 是网络社区 $G=(V,E,T)$的一种划分结果,则 C 的内聚度为:

$$Cohesion(C)=TagCohesion(C)\cdot EdgeCohesion(C) \tag{16}$$

3.1.2　社区耦合度

定义 9　设 T_{Ci}是社区 C_i 的用户所有用户标签的并集,T_{Cj}是社区 C_j 内所有用户标签的并集,则两个社区的相似度为:

$$Sim(C_i,C_j)=\frac{\mid T_{C_i}\cap T_{C_j}\mid}{\mid T_{C_i}\cup T_{C_j}\mid} \tag{17}$$

定义 10　设 $C=\{C_1,C_2,C_3\cdots C_k\}$ 是网络社区 $G=(V,E,T)$的一种划分结果,则 C 的标签相似度为:

$$TagSim(C)=\frac{\sum_{C_i\in C}\sum_{C_j\in C}Sim(C_i,C_j)}{k^2-k} \tag{18}$$

定义 11 设 $C=\{C_1,C_2,C_3\cdots C_k\}$ 是网络社区 $G=(V,E,T)$ 的一种划分结果,社区 C_i 有 p 个用户,其中有 p_m 个用户关注社区 C_j 中的用户,则分组 C_i 对分组 C_j 的关注度为:

$$Follow_{C_i}(C_j)=\frac{p_m}{p} \tag{19}$$

定义 12 设 $C=\{C_1,C_2,C_3\cdots C_k\}$ 是网络社区 $G=(V,E,T)$ 的一种划分结果,对于任意 $C_i,C_j\in C$ 且 $C_i\neq C_j$,从 C_i 到 C_j 的关注熵为:

$$H_{C_j}(C_i)=\begin{cases}-Follow_{C_i}(C_j)\log_2 Follow_{C_i}(C_j), & \text{if } Follow_{C_i}(C_j)>0\\ 0, & \text{else}\end{cases} \tag{20}$$

熵值越小说明分组 C_i 中用户关注 C_j 的一致性越高.

定义 13 设 $C=\{C_1,C_2,C_3\cdots C_k\}$ 是网络社区 $G=(V,E,T)$ 的一种划分结果,若 $C_i\neq C_j$ 则 C 的边熵为:

$$EdgeEntropy(C)=\frac{\sum_{C_i\in C}\sum_{C_j\in C}H_{C_j}(C_i)}{k^2-k} \tag{21}$$

定义 14 设 $C=\{C_1,C_2,C_3\cdots C_k\}$ 是网络社区 $G=(V,E,T)$ 的一种划分结果,则 C 的耦合度为:

$$Coupling(C)=TagSim(C)\cdot EdgeEntropy(C) \tag{22}$$

3.1.3 关于意见领袖社区的影响

在网络社区中存在着影响力不同的用户群体,有些人组成的"意见领袖社区"具有比较大的威望,往往能够一呼百应,所以需要与普通用户划分开[17]. 若某用户的"粉丝"比较多则认为该用户的影响力较大,于此同时若用户的"粉丝"影响力较大则会使被关注用户的影响力变大. 本文引入网页链接分析中的 PageRank 技术,并依此对用户依照影响力大小进行排序,希望社区划分中同一分组的用户的影响力相似.

定义 15 设 $C=\{C_1,C_2,C_3\cdots C_k\}$ 是网络社区 $G=(V,E,T)$ 的一种划分且 $v_i\in C_i$,$VarPageRank(v_i)$ 表示分组 C_i 中用户 *PageRank* 值的方差,那么 C 的实用性表示为:

$$Utility(C)=\sum_{C_i\in C}\frac{[VarPageRank(v_i)\mid v_i\in C_i]}{\max[PageRank(v_i)-\overline{PageRank(v_i)}]^2} \tag{23}$$

为了与之前的内聚度及耦合度对应,式(23)对分组的方差进行了归一化处理.

3.1.4 社区划分质量评价函数

定义 16 对于微博社区网络 $G=(V,E,T)$ 的划分结果 $C=\{C_1,C_2,C_3\cdots C_k\}$ 的归一化质量函数为:

$$Quality(C)=\frac{Coupling(C)\cdot Utility(C)}{Cohesion(C)+1} \tag{24}$$

Cohesion 从社区内部的紧密程度反映了划分结果的内聚度;*Coupling* 从社区之间联系的松散程度衡量划分结果的耦合度;*Utility* 从实际应用方面考虑,把影响力大的用户与普通用户分开. 三者是相互制约的,因此社区划分的质量评价函数是三者的折中.

3.2 社区划分算法

3.2.1 寻找核心标签

通过最小化质量评价函数,可为社区网络找到最优标签. 将整个社区网络视作一个簇,对所有标签 t,将所有包含 t 的用户提取出来划分为一个分组. 对于一个新的划分,能够最小化 *quality* 值的标签将被标注为核心标签用作划分依据,关于发现核心标签的伪代码如下:

算法 1 发现核心标签算法(微博网络社区 G,当前用户分组 C)

输入:微博网络社区 G 及当前用户分组 $C=\{C_1,C_2,\cdots,C_k\}$
输出:核心标签 t 及 t 所对应的分组 C_b

```
初始化核心标签为空,MinQ 赋极大值,分组 C_b 为空;
For G 中的每个非核心标签 t_i   //之前的核心标签不再遍历
  提取所有含有标签 t_i 的用户划分为 C_b1 组;
    If  quality(C_1,C_2,…,C_k,C_b1) < MinQ;
      MinQ = quality(C_1,C_2,…,C_k,C_b1);
      t = t_i;
      C_b = C_b1;
    End if
  End For
Return < CoreTag, C_b >
```

3.2.2 基于核心标签的微博社区划分策略

本文所提出的微博社区划分策略基本思想是不断迭代的依照核心标签提取最优分组,而为了防止发生社区过分重叠的现象需要在算法里对分组进行分类处理,引入以下概念:

定义 17 设 C_i 是划分结果 C 中的一个分组,则社区 C_i 中的用户的集合为 V_{C_i},则对于任意用户分组 C_i 和 C_j 的相似度为

$$Jaccard(C_i,C_j)=\frac{|V_{C_i}\cap V_{C_j}|}{|V_{C_i}\cup V_{C_j}|} \tag{25}$$

分组的相似度代表着两个组中相同用户的个数,引入社区相似度可以有效地控制社区的重叠度,其算法的伪码如下:

算法 2 核心标签划分算法(终止条件 p,社区重叠度 γ,微博网络社区 G)

输入:终止条件 p,社区重叠度 γ,微博网络社区 G
输出:微博网络社区 G 中用户的分组结果 $C=\{C_1,C_2,\cdots,C_k\}$

1. Repeat
2. 初始化核心标签为空,分组 C_b 为空;
3. 调用发现核心标签算法,得到核心标签 t 及所对应的分组 C_b;
4. for 所有其他分组 C_i
5. If Jaccard(C_b,C_i)$\leqslant\gamma$ 将新分组加入到已有分组 C;
6. else 找到最大 Jaccard(C_b,C_i)将 C_b 并入到 C_i 中;
7. end for
8. Until 满足条件 p
9. Return C

根据 TagCut 终止条件的不同本文提出了两种社区发现算法:NSTC(Number Specified Tag Cut)算法和 QBTC(Quality Based Tag Cut)算法. 前者需预先设定社区划分中社区的数目并将其作为终止条件,后者则是以评价质量函数 quality 变差作为终止条件.

NSTC 算法指定社区划分算法结果中社区的数目,在算法中每次迭代划分出新社区后都会检测社区数目是否超过了所设定的阈值 η,如果超过阈值则停止算法,NSTC 的伪代码如下:

算法 3　NSTC 算法(社区数目 η,社区重叠度 γ,微博网络社区 G)

输入:社区数目 η,社区重叠度 γ,微博网络社区 G
输出:微博网络社区 G 中用户的分组结果 $C=\{C_1,C_2,\cdots,C_k\}$ 及社区数目 *ClusterNum*

1. $U=G$ 中的用户;
2. Repeat
3. 初始化核心标签为空,分组 C_b 为空,*ClusterNum* = 0;
4. 调用发现核心标签算法,得到核心标签 t 及所对应的分组 C_b;
5. $U=U-C_b$;
6. For 所有其他分组 C_i;
7. If Jaccard(C_b,C_i)$\leqslant\gamma$;
8. 将新分组 C_b 加入到当前分组 C;
9. *ClusterNum* + +;
10. Else
11. $C_j=\underset{C_i\in C}{\operatorname{argmax}}[\text{Jaccard}(C_b,C_i)]$;
12. $C_j=C_j+C_b$
13. Until $\eta\leqslant$*ClusterNum or* U 等于空;
14. If $U\neq\varnothing$;
15. 将 U 作为最后一个社区加入到 C 中;
16. Return C,*ClusterNum*;

随着划分算法的迭代运行,宏观上的评价函数会逐渐变好再变差,QBTC 算法的终止条件为一旦 quality 值变大或者用户被分完立即终止. QBTC 算法的伪代码如下:

算法 4　QBTC 算法(社区重叠度 γ,微博网络社区 G)

输入:社区重叠度 γ,微博网络社区 G
输出:微博网络社区 G 中用户的分组结果 $C=\{C_1,C_2,\cdots,C_k\}$,*ClusterNum*

1. $U=G$ 中的用户;
2. Repeat
3. 初始化核心标签为空,分组 C_b 为空;*ClusterNum* = 0;
4. *originC* = *C*;*originQuality* = *quality*(*C*);
5. 调用发现核心标签算法,得到核心标签 t 及所对应的分组 C_b;
6. For 所有其他分组 C_i
7. If Jaccard(C_b,C_i)$\leqslant\gamma$
8. 将新分组加入到已有分组 C;*ClusterNum* + +;计算 *currentQuality*;
9. If *originQuality* > *currentQuality*
10. Return *originC*,*ClusterNum*
11. Else
12. $C_j=\underset{C_i\in C}{\arg\max}[Jaccard(C_b,C_i)]$;$C_j=C_j\cup C_b$;计算 *currentQuality*;
13. If *originQuality* > *currentQuality*
14. Return *originC*,*ClusterNum*
15. $U_b=C_b$ 中的用户;
16. $U=U-U_b$;
17. End For
18. Until $U=\varnothing$;
19. Rturn C,*ClusterNum*;

4　实验

4.1　实验数据分析

本文实验数据通过新浪微博公开 API 进行抓取,具体的做法是从核心用户开始,以广度优先遍历原则,沿关注关系进行逐层抓取,抓取的数据包括新浪微博用户数据和用户标签数据. 实验使用某在校学生作为核心用户,分别抓取微博用户,用户标签,用户所发微博内容,用户关注关系,形成 3 组数据集,数据集 1 选取核心用户及其关注列表里的所有人;数据集 2 在数据集 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数据集 3 的抓取所以最后需要剔除出度大于 100 的用户;数据集 3 以数据集 2 为基础并沿用其获取方法,但是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). 其数据信息如表 1 所示:

表 1　实验数据

数据集	用户数	标签数	关注关系数
*S*1	85	237	950
*S*2	598	2034	6463
*S*3	1265	4493	11695

4.2　实验结果与分析

为了验证本文方法的有效性设计了四个实验. (1)分析标签扩充对划分结果的影响找出合适的 θ 并调节参数 α 观察内联关系和外联关系对划分结果的影响;

(2)调节参数 γ,通过观察划分结果中社区个数选出最贴合实际重叠度的 γ;(3)研究 QBTC 和 NSTC 两个算法的性能指标;四是选取了典型的两个微博社区划分方法与本文方法进行比较.

4.2.1 标签扩充及内联外联权重

用户标签扩充时 θ 对划分结果中社区的个数影响很大,所以使用划分的社区数目对 θ 进行试验,由于另一参数 γ 对社区数目的影响也很大所以先将 γ 定为 0.8,此数值会在后面的实验说明.

标签扩充对 NSTC 算法结果的 *ClusterNum* 影响较小但是对 QBTC 算法结果的 *ClusterNum* 影响较大,所以首先要用 NSTC 算法对 3 个数据集进行社区划分,划分时不进行标签扩充即 $\theta=1$,将参数 η 从 1 开始逐步增大到 20,并记录每个参数所对应的 *quality* 值,*quality* 值最小的划分结果所对应的 η 即是该数据集最贴合实际社区数量如图 1. 再用 QBTC 算法对三个数据集进行社区划分,同样对 θ 从 0 到 1 开始赋值并依此进行标签的扩充,数据集划分结果中最接近最优 η 的划分结果所对应的 θ 即是最合适的 θ,结果如表 2.

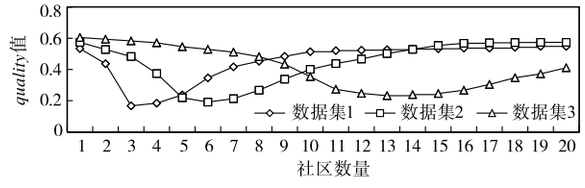


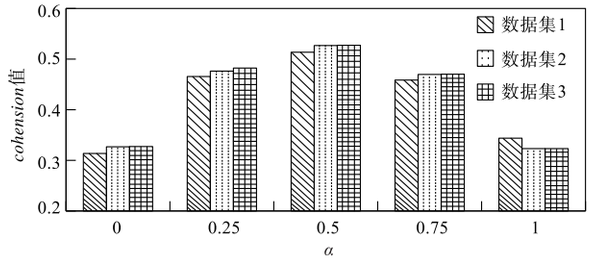
图1 NSTC算法在三个数据集上的结果

表 2 θ 取值对划分结果的影响

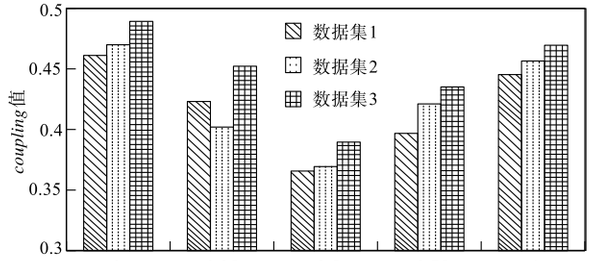
	数据集 1	数据集 2	数据集 3
$\theta=0$	1	1	1
$\theta=0.1$	1	1	2
$\theta=0.2$	1	2	2
$\theta=0.3$	2	3	3
$\theta=0.4$	2	4	5
$\theta=0.5$	3	4	7
$\theta=0.6$	3	5	13
$\theta=0.7$	4	7	15
$\theta=0.8$	5	10	17
$\theta=0.9$	6	11	20
$\theta=1$	6	14	24

从表 2 中可以看出当 $\theta=0$ 时,算法在归并阶段这些社区都被归入到同一个社区导致结果严重失真;而当 $\theta=0.6$ 的时候各社区数目最贴合实际的划分标准.

α 值是否合适对划分结果的内聚度及耦合度影响非常大,本实验用划分结果的内聚度和耦合度的值来对其进行评估. 在图 2 中从整体上来看当 α 取值为 0.5 时内聚度和耦合度都能达到最佳状态.



(a) α取值对内聚度的影响



(b) α取值对耦合度的影响

图2

4.2.2 标签距离参数 γ 的取值

在核心标签划分策略中标签距离 Jaccard(C_i,C_j)决定了新划分出的用户是组成新的分组还是归到之前已划分的分组,因此 γ 对于 QBTC 算法结果的影响非常大,所以本节实验采用与 4.2.1 中关于 θ 实验相似的策略用社区划分结果中社区数量对 γ 进行评估并对 θ 的选取提供了数据支持.

由图 3 可知经过比较得出当 $\gamma=0.8$ 时得到最佳结果,此时社区划分中允许重叠部分的比例最为接近真实结果.

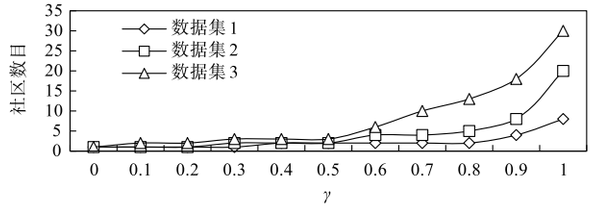


图3 γ取值对划分结果社区数目的影响

4.2.3 NSTC 算法和 QBTC 算法性能分析

在 NSTC 算法中需要指定划分结果中社区的数量,随着社区被划分出来质量评价函数在整体上会趋于某一定值,由于数据集 1 和数据集 2 经过 NSTC 算法后社区数量过少,不能很好地体现出其变化的过程,所以本节实验对数据集 3 进行处理,实验结果如图 4 所示:

在 QBTC 算法中由评价指标的优劣来作为社区划分的终止条件,总体上来看社区划分结果的评价指标是从差趋向好,数值从高趋向低,在达到最合适的划分

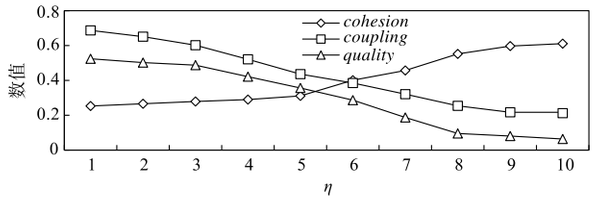


图4 NSTC算法的性能

结果时即评价指标数值最小时算法结束. 算法的具体表现如图 5.

4.2.4 与其他算法的比较

为了验证算法的有效性,选取了针对微博网络的两个较为典型的方法:基于 R-C 模型的方法与基于相似度的方法[13,14]与本文的 QBTC 算法在数据集 3 进行比较并采用内聚度、耦合度、实用性综合对算法进行评估. 由于 *cohesion* 和 *utility* 从局部即可进行评价,且在划分时遵循的原则是局部最优所以这两个指标的结果会从好变差,而 *coupling* 与 *quality* 必须从整体来进行计算,所以呈现出从差变好的过程且在实验中为了方便与其他方法比较所以在计算时只考虑所选取的最大 10 个社区. 如图 6 所示.

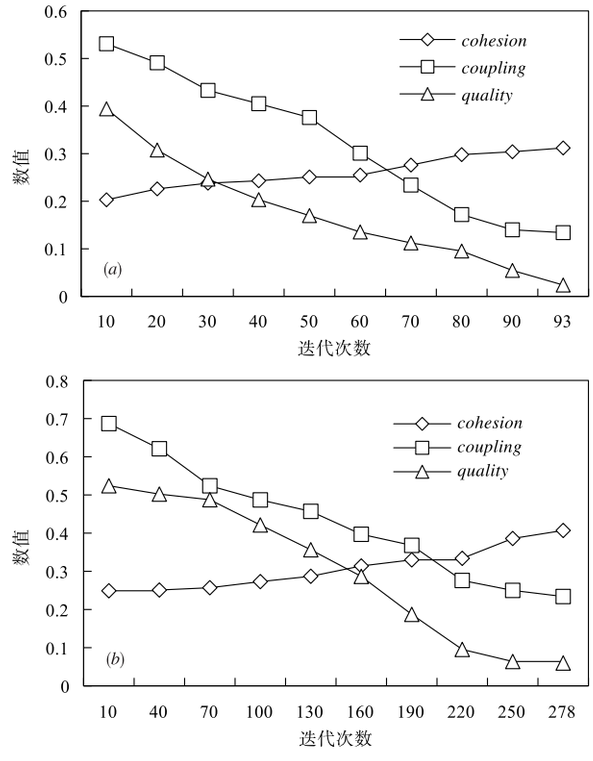
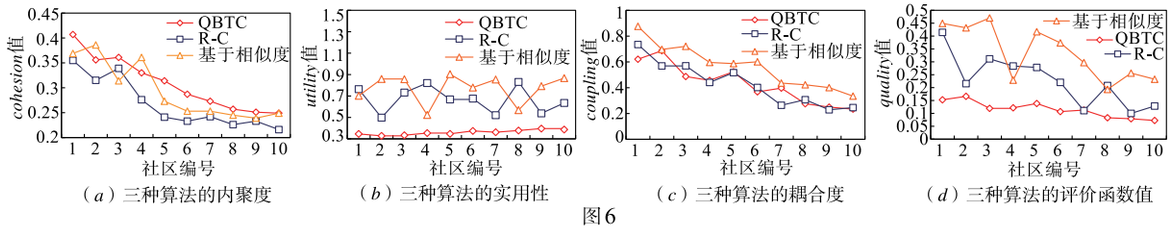


图5 QBTC在数据集2,3上的迭代结果



(a) 三种算法的内聚度　(b) 三种算法的实用性　(c) 三种算法的耦合度　(d) 三种算法的评价函数值

图6

从图 6 可以看到基于相似度的方法与 QBTC 算法的内聚程度要好于基于 R-C 模型的方法,这是由于微博文本中有太多的杂乱信息对用户的兴趣表示起到了极大的干扰作用;在耦合度上基于 R-C 模型的方法却要好于基于相似度的方法,因为基于 R-C 的方法将边映射为点,并将其加上属性,所以其方法划分结果的边熵要低于基于相似度的方法,而 QBTC 算法在耦合度上的表现不仅要略高于基于 R-C 模型的方法而且在社区间关注关系的一致性上考虑的更全面,所以在耦合度的比较上结果更倾向于 QBTC;在验证 *quality* 指标的时候 QBTC 算法表现出了突出的优势,这是由于其他两个方法的划分结果的实用性都有所欠缺,在加上了 *utility* 后 QBTC 算法在评价指标 *quality* 上有了非常突出的表现.

5 结束语

鉴于微博网络社区所具有的复杂网络特性,本文提出一种基于核心标签的社区划分算法,首先通过标签之间的共现关系及出现频率算出标签之间的关联关系并对其进行扩充,然后算法采用最优核心标签分的策略对其所对应的用户进行分组. 使用质量评价函数对划分结果进行评价并修正,且在算法中引入参数让社区的重叠度变得可控. 本算法在处理零散社区时由于评价函数的值变得极端,所以效果并不理想,今后的工作将考虑用户数较少的社区的处理方式.

参考文献

[1] Fortunato S. Community detection in graphs[J]. Physics Reports,2009,486(3-5):75-174.

[2] Girvan M,Newman M E J. Community structure in social and biological networks[J]. Proceedings of National Academy of Science of the United States of America,2001,99(12):7821-7826.

[3] Newman M E. Fast algorithm for detecting community

structure in networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2004,69(6):066133-1 - 066133-5.

[4] Palla G,Derényi I,Farkas I,et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature,2005,435(7043):814 - 818.

[5] Latouche P,Birmelé E,Ambroise C. Overlapping stochastic block models with application to the french political blogosphere[J]. The Annals of Applied Statistics,2011,5(1):309 - 336.

[6] Shi C,Cai Y,Fu D,et al. A link clustering based overlapping community detection algorithm[J]. Data & Knowledge Engineering,2013,87(9):394 - 404.

[7] 冷作福. 基于贪婪优化技术的网络社区发现算法研究[J]. 电子学报,2014,42(4):723 - 729.
Leng Zuofu. Discovery algorithm optimization technique based on greed online community[J]. Acta Electronica Sinica,2014,42(4):723 - 729. (in Chinese)

[8] 张桂杰,张健沛,杨静,等. 基于链接相似性聚类的重叠社区识别[J]. 电子学报,2015,43(7):1329 - 1335.
Zhang Guuijie,Zhang Jianpei,Yang Jing,et al. Overlapping community detection based on link similarity clustering[J]. Acta Electronica Sinica,2015,43(7):1329 - 1335. (in Chinese)

[9] 于海,赵玉丽,崔坤,等. 一种基于交叉熵的社区发现算法[J]. 计算机学报,2015,38(8):1574 - 1581.
Yu Hai,Zhao Yuli,Cui kun,et al. Community detection algorithm basedon cross-entropy method[J]. Chinese Journal of Computers,2015,38(8):1574 - 1581. (in Chinese)

[10] 王诗懿,董一鸿,李志超. 大规模复杂网络下重叠社区的识别[J]. 电子学报,2015,43(8):1575 - 1582.
Wang Shiyi,Dong Yihong,Li Zhichao. The identification of over lapping communities in large-scale complex networks[J]. Acta Electronica Sinica,2015,43(8):1575 - 1582. (in Chinese)

[11] 张引,张斌,高克宁,等. 面向自主意识的标签个性化推荐方法研究[J]. 电子学报,2012,40(12):2353 - 2359.
Zhang Yin,Zhang Bin,Gao Kening,et al. Autonomy oriented personalized tag recommendation[J]. Acta Electronica Sinica,2012,40(12):2353 - 2359. (in Chinese)

[12] Ahn Y Y,Bagrow J P,Lehmann S. Link communities reveal multiscale complexity in networks[J]. Nature,2010,466(7307):761 - 764.

[13] 周小平,梁循,张海燕. 基于 R-C 模型的微博用户社区发现[J]. 软件学报,2014,25(12):2808 - 2823.
Zhou Xiaoping,Liang Xun,Zhang Haiyan. User community detection on micro-blog using R-C model[J]. Journal of Software,2014,25(12):2808 - 2823. (in Chinese)

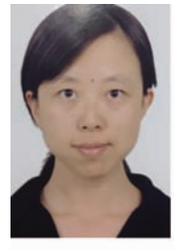
[14] 孙怡帆,李赛. 基于相似度的微博社交网络的社区发现方法[J]. 计算机研究与发展,2014,51(12):2797 - 2807.
Sun Yifan,Li Sai. Similarity-based community detectionin social network of microblog[J]. Journal of Computer Research and Development,2014,51(12):2797 - 2807. (in Chinese)

[15] Ma Huifang,Jia Meihuizi,Xie meng,et al. A microblog recommendation algorithm based on multi-tag correlation [A]. International conference on Knowledge Science Engineering and Management[C]. Chongqing,2015. 483 - 488.

[16] 尹丹,高宏,邹兆年. 一种新的高效图聚集算法[J]. 计算机研究与发展,2011,48(10):1831 - 1841.
Yin Dan,Gao Hong,Zou Zhaonian. A novel efficient graph aggregation algorithm[J]. Journal of Computer Research and Development,2011,48(10):1831 - 1841. (in Chinese)

[17] 张伟哲,王佰玲,何慧,等. 基于异质网络的意见领袖社区发现[J]. 电子学报,2012,40(10):1927 - 1932.
Zhang Weizhe,Wang Bailing,He Hui,et al. Public opinion leader community mining based on the heterogeneous network[J]. Acta Electronica Sinica,2012,40(10):1927 - 1932. (in Chinese)

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. 研究领域为人工智能、数据挖掘与机器学习.
E-mail:mahuifang@ yeah. net

谢 蒙 男,1990 年 6 月出生,河北邢台人. 西北师范大学计算机科学与工程学院硕士. 研究方向为:互联网数据挖掘与机器学习.
E-mail:xiemengh@ hotmail. com