基于二部配置模型的网络用户分组及重要回复者识别

张亚茹[1,2]，唐锡晋[1,2]

(1. 中国科学院数学与系统科学研究院, 北京 100190;
2. 中国科学院大学, 北京 100049)

摘要: 在以社交媒体为载体的舆论场中, 用户的个人诉求和回复行为潜在地反映其兴趣偏好, 根据这一关联为用户分组, 从而推动群体新闻推荐系统的发展. 识别用户的重要回复者以由消息传播机制干预正负面新闻的传播. 以天涯论坛为例, 定义顶级用户与普通用户, 结合二部配置模型建模顶级用户、帖子和普通用户间基于发帖关系以及回复关系的二部图, 获取了具有不同话题偏好的用户群, 每组内的用户交互密切. 得到的重要回复者回复偏重性明显, 且易将相应顶级用户的负面情绪放大, 因此识别重要回复者并适时施以制约为网络舆情管理提供新视角.

关键词: 用户分组; 重要回复者; 二部配置模型; 二部图
中图分类号: N945.12; TP391; N32 **文献标识码**: A **文章编号**: 1000−5781(2021)05−0577−13

doi: 10.13383/j.cnki.jse.2021.05.001

Internet user grouping and important responder identification based on bipartite configuration model

Zhang Yaru[1,2], **Tang Xijin**[1,2]

(1. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: In the public opinion field with social media as the carrier, users' personal appeals and response behavior potentially reflect their interest preferences. Grouping users according to this association helps to develop group news recommendation system. This paper also identifies the users' important responders to intervene in the dissemination of positive and negative news by the message dissemination mechanism. Selecting the case of Tianya forum, the study defines the top users and ordinary users. Bipartite configuration models are used to model bipartite graphs among top users, posts and ordinary users based on the posting relationship and reply relationship. The results show that users in the same group have similar topic preferences, and their interactions are close. Important responders thus obtained are inclined to corresponding top users in terms of the number of replies. The negative emotions of the top users tend to be amplified by their important responders. Thus, the identification of important responders and timely restriction provide a new perspective for online public opinion management.

Key words: user grouping; important responders; bipartite configuration model; bipartite graph

收稿日期: 2020−08−02; 修订日期: 2021−07−26.
基金项目: 国家重点研发计划资助项目(2016YFB1000902); 国家自然科学基金资助项目(71731002; 71971190).

1 引 言

当今社交媒体迅速发展, 越来越多的人习惯通过微博、论坛和媒体网站浏览感兴趣的内容, 获取网络舆情, 参与话题讨论, 发表个人所感. Mohbey 等[1]以意大利选举期间的帖子为语料, 定义了政治界主要讨论的农业、基础设施建设、教育以及就业等 9 大类选举问题, 并使用深度学习方法将用户发表的言论划分为上述9类, 以预测用户关注的焦点. 目前也有许多关于舆情事件与社会网络用户影响力的研究. 学者们研究了确定高影响力事件, 抽取事件要素的方法[2,3], 开展了文本聚类、话题分析等工作[4−6]. 对于社会网络中用户影响力的研究, 主要是从用户网络属性, 用户行为方式和互动规律, 用户距离等角度出发进行探索的[7−11]. 随着粉丝经济, 直播带货, 广告推荐, 新闻推送的兴起, 挖掘网民兴趣偏好, 精准定位用户群, 为不同群体推荐其感兴趣的事物, 对于提高网络营销效益, 及时获取民意具有重要意义. Liao 等[12]研究了一种基于关联规则的推荐方法. 首先通过问卷调查考察了多种社会网络用户进行在线社会网络营销的经历. 接着根据这些用户的网络行为和偏好对他们聚类, 并探索用户画像、社会网络管理、社会网络行为、在线购买行为、社会网络营销以及个性化推荐间的关联. 最后基于关联规则得到的知识图为每个群体推荐其可能会购买的物件. 用户分组的效果是群体推荐的关键.

社会网络用户的分组依赖于图上的社区划分. 目前已有许多对单部图进行社区划分的方法, 如基于模块度优化的算法[13], 基于节点表示学习的方法[14,15]. 由于现实中存在很多包含两种节点类型的二部图, 学者们也研究了二部图上的社区划分方法. 二部图的社区划分, 分为两种情形, 一种是在整个二部图上同时对两种类型的节点进行社区划分, 另一种是通过某种方式将二部图映射为仅包含同一节点类型的单部图, 从而对每一类节点执行单部图上的社区划分. 对于前者, 主要是在单部图社区划分算法基础上改进距离度量方式[16], 或者是模块度衡量方式[17,18]. 这类方法没有区分节点类型, 未充分利用二部关联关系. Tackx 等[19]提出 COMSIM 算法用于二部图社区划分. 首先以同类节点的共同邻居数作为单部图中相应边的权重(节点间的相似度), 接着将映射得到的单部图中每个环具有最大连接权重的两节点作为各社区的核心, 最后对于每个非核心节点选择其与社区所有节点间相似度总和最大的社区作为其所属社区. 该方法通过直接映射的方式获取单部图, 会造成节点连接稠密, 且包含相似性不高的连边. Cui 等[20]使用二部网络中的单部社区结构实现节点聚类. 首先通过二部网络的拓扑性质, 构造二部聚类三角形, 接着通过这种二部聚类三角形将二部网络映射为两个加权的单部网络, 然后从加权的单部网络中抽取全部最大子图, 通过聚类阈值合并最大子图实现节点聚类. 该方法基于对子图的合并得到社区, 可能会造成同一社区多种不一致子类节点群的情况. 已有的社会网络用户分组研究主要根据用户画像相似性、直接的交互关系建立用户间的连边, 进而借助单部图的社区划分算法得到用户群[21−23]. 但是面临着很多用户未填写个人信息或者所提供信息与分组类别关联性不大、大型社会网络中用户连接稀疏等瓶颈. 社交媒体中的用户按照热度, 可以分为两个层级, 一类是发布热点话题的热门用户或者顶级用户, 另一类是普通用户, 通常普通用户倾向与热门用户建立社会关系, 如回复、转发等, 此关联可能蕴含了用户的话题偏好. 如果能够利用这种关联进一步衡量顶级用户间的相似性, 那么将为用户分组提供新思路.

社会影响力研究用于获取社会网络层面富有影响力的关键节点, 而对于个体层面的自我网络, 则更关注与其关联紧密的群体, 以及他们间的交互、相互作用. 已有的相关研究大多集中于后者, 如预测个体所发帖子的回复者, 预测回复内容等. 回复者预测是网络用户回复行为研究方向的热点, 按照任务设定有分类、排序两大类方法. Schantl 等[24]基于回复者关注话题与帖子话题相似性这一话题特征, 描述社会关系的社会特征, 以及帖子流行特征来对用户是否会回复某条帖子进行二分类预测, 并发现相比于话题偏好, 社会关系是更为重要的回复行为影响因子. Yuan 等[25]基于互惠、时序和上下文特征考虑友谊关系动态, 并结合排序模型预测用户的哪些朋友将更有可能回复其发布的某条帖子. 但是很多时候用户的好友仅仅点赞或者浏览,

特别是对于顶级用户, 好友关系并不是一个用于预测回复者的较为有效的社会特征. 如果事先识别出用户的历史重要回复者, 并添加这一社会关系特征可增加预测的准确性. 本文聚焦于确定个体的历史重要回复者, 即易与该个体建立回复关系的群体, 这将有助于回复者预测任务的开展, 且在消息传播机制中, 其它用户的回复也加大了原帖的可见性, 因此对于用户发布的负面新闻, 舆情管理者通过限制其重要回复者的发言或制约消息传播; 对于用户发布的正面新闻, 即时推送给重要回复者, 引起追随, 加速消息传播.

Saracco 等[26]提出了一种基于熵的空模型——二部配置模型(bipartite configuration model, BICM), 实现了将二部图映射为单部图, 这为二部图上单模节点的社区划分提供了可能. 已有基于该模型的实际应用, 如在世界贸易网络中,确定国家群以及产品群; 在用户影评网络中, 确定电影的分组; 以关于选举的帖子为数据源, 根据未验证用户对验证用户的转帖行为, 确定用户的政治联盟, 考虑用户与帖子间的发帖与转发贴有向关系识别社会网络的重要传播者[27]. 受以上研究工作的启发, 本文尝试将 BICM 应用到社交媒体场景下顶级用户、帖子和普通用户间基于发帖关系以及回复关系的二部图中, 期望获取具有不同话题偏好的用户组, 并识别顶级用户的重要回复者. 相较于文献[28]在移动情境感知环境下挖掘用户行为模式, 以开展精准营销的个性化推荐服务, 本文所开展的用户分组研究工作则强调从个体行为获取群体特征, 以推动下游基于群体的新闻推荐的任务. 具体的, 以天涯论坛为例, 视天涯杂谈版块“年度拾英”用户为顶级用户, 顶级用户一年内在天涯杂谈发帖的回复者为普通用户, 由普通用户的回复行为使用 BICM 建立顶级用户间的连边, 进而对图划分, 实现顶级用户分组, 并根据用户发帖类型探索每组用户的话题偏好. 接着根据普通用户对各顶级用户组的极化回复, 确定普通用户组别. 再建立顶级用户与发帖间的二部图、全部回复者与回帖间的二部图, 联合 BICM 与二部部分配置模型(bipartite partial configuration model, BIPCM)确定顶级用户的重要回复者. 通过该方式得到的每组用户, 内部关联紧密, 外部稀疏连接, 且具有较为一致的话题偏好. 获取的重要回复者对相应的顶级用户依附度很高, 是相应顶级用户的高概率消息受体及反馈者. 此外, 发现存在重要回复者的顶级用户多发表负面情绪的新闻, 此时重要回复者带有负面情绪的回复也居多. 因此对于影响网络环境的负面新闻, 除了对相应发帖人进行管控外, 限制重要回复者的回复也很重要.

2　顶级用户发帖回帖标题聚类分析

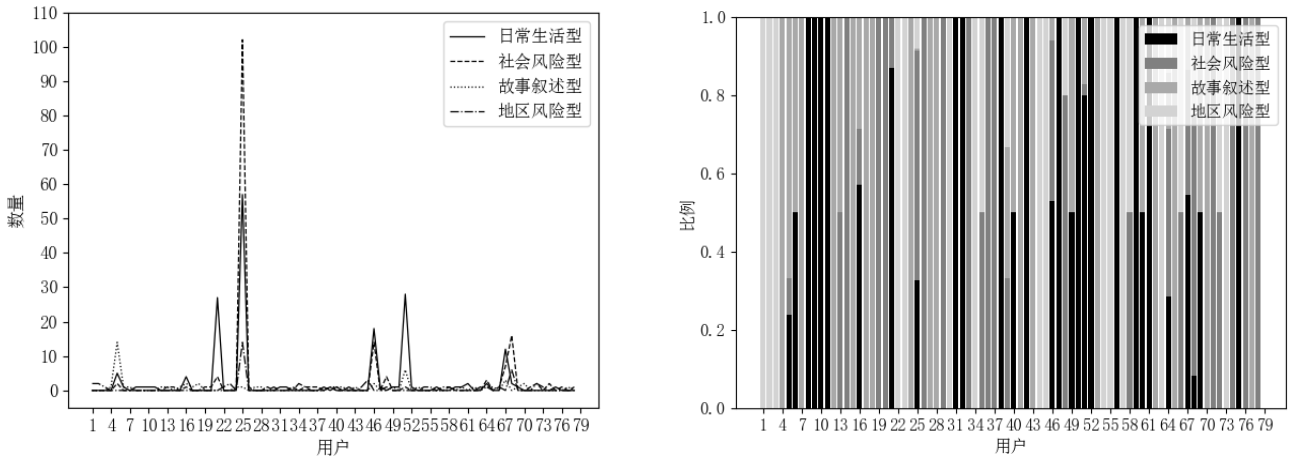
天涯论坛是目前中国最活跃的论坛之一, 包括天涯杂谈、时尚资讯和球迷一家等多个版块, 其中天涯杂谈是关于民生的版块, 它以一年内用户在该版块所发帖子的总点击量为排序指标, 给出了前 80 位天涯拾英用户, 本文以之为顶级用户, 于 2019–11–07 爬取这些顶级用户过去一年中的发帖, 发帖下的回复(并解析得到各顶级用户的回复者), 用户画像, 以及顶级用户所做的回复. 其中 2 位用户因为被封杀或者删除账户无法爬取.

顶级用户一年内在天涯杂谈的发帖共 566 条, 一年内在天涯杂谈回复的帖子数共 742 条, 使用自然语言处理中的预训练模型 Bert[1]表示每个帖子标题向量, 形成 1 308×768 的二维数组, 已有研究表明低维度的向量聚类效果更好[4], 因此本文训练自编码器, 将 768 维数据压缩到 2 维.

具体的, 从均隶属于天涯杂谈版块的顶级用户发帖以及所回复的原贴的标题向量中选取 1 108 条作为训练集, 另外的 200 条作为测试集, 设置一个编码器与一个解码器, 当解码得到的向量与原向量的均方差损失很小时, 以编码器结果作为标题低维向量表示. 设置 $\mathrm{batch\_size} = 64$, 当迭代轮数为 200 时, 训练损失达到 0.068 7, 测试集损失达到 0.065 0, 迭代终止, 保存模型, 得到高维向量编码结果.

使用 K-Means 对表达成二维向量的顶级用户一年内在天涯杂谈的发帖进行聚类, 根据帖子标题向量在二维坐标系中的分布, 本文将簇数设置为 4, 最终得到的 4 个帖子簇分别是日常生活型, 社会风险型, 故事叙述型, 地区风险型. 图 1 分别显示每位用户各种类型发帖数目以及比例. 发帖数目多的用户, 其发帖类型呈现多样性, 但仍有偏重.

[1] https://zhuanlan.zhihu.com/p/48612853

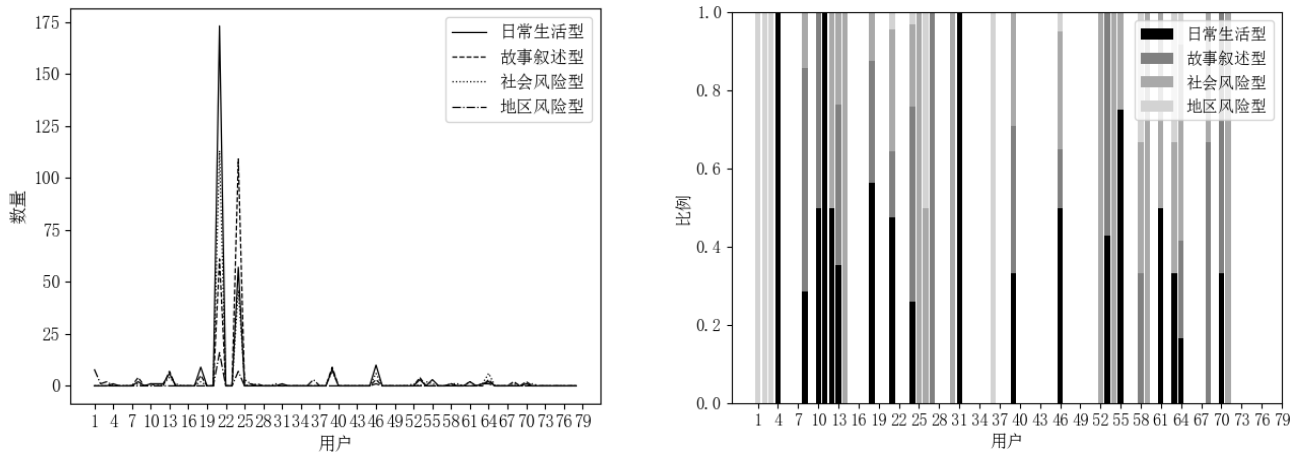


(a) (b)

图 1 顶级用户发帖类型分布

Fig. 1 Distribution of post types of top users

使用 K-Means 对二维向量表示的顶级用户一年内在天涯杂谈所回复的原帖对应的标题聚类, 仍然聚集成了上述 4 种类型的簇. 图 2 分别显示每位用户回复的帖子中各种类型帖子的数目以及比例.



(a) (b)

图 2 顶级用户回复的帖子类型分布

Fig. 2 Distribution of types of posts replied by top users

各用户回复的帖子类别分布与发贴类别分布相似, 这种相似性源于用户的话题偏好.

3 顶级用户分组

若要实现对顶级用户的分组, 一种以发帖类型为指导的方法是将用户归到发帖类型最多的那一组. 但是这种方式忽略了用户兴趣的多样性与不同类别帖子间的关联性. 下面介绍二部配置模型, 并应用该模型实现顶级用户分组.

3.1　二部配置模型

R, S 分别表示顶级用户集与普通用户集, 各自有 N_R, N_S 位用户. 若某用户 s 回复过顶级用户 r 的发帖, 则建立两者间的无向连边. 该部分旨在根据普通用户的回复行为, 确定相似的顶级用户, 以实现分组. 单射指的是若两个顶级用户有共同回复者则建立两者间的连边, 但仅仅依据单射, 会形成一个较为密集的顶级用户网络, 并且这种相似性很不可靠. 一般来说, 只有两个顶级用户拥有统计意义上足够多的共同邻居, 才能够认为它们是相似的, 如图 3 中 r_1, r_2 用户有 3 个共同回复者, 若“3”为统计意义上的“大量”, 那么可以认为这两个顶级用户间存在相似性连边. BICM 提供了一种假设检验的方法来确定两顶级用户间连边的存在性, 这使得顶级用户间的相似性连接更可信.

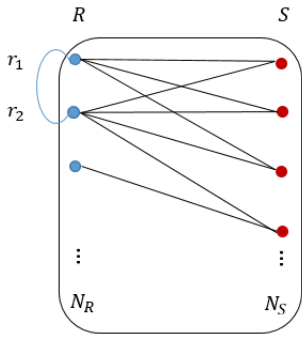


图 3 顶级用户与普通用户关于回复关系的二部图

Fig. 3　Bipartite graph between top users and ordinary users about reply relation

在同节点的所有可能图中某种图结构 M 出现的概率可表示为

$$\Pr(\mathrm{M}) = \mathrm{e}^{-H(\mathrm{M})}/Z. \tag{1}$$

二部图情况下, 哈密顿量 $H(\mathrm{M}) = \sum_{r\in R}\alpha_r k_r + \sum_{s\in S}\beta_s k_s$, k_r, k_s 为两种类型节点的度, α_r, β_s 为约束相应节点度的拉格朗日乘子, 大于 0, Z 为归一化因子.

若设定两种类型节点的连接概率为 p_{rs}, m_{rs} 为图 M 的 0/1 邻接矩阵中相应的值, 则图 M 出现的概率也可使用下列概率公式表示, 即

$$\Pr(\mathrm{M}) = \prod_{r\in R}\prod_{s\in S}(p_{rs})^{m_{rs}}(1-p_{rs})^{(1-m_{rs})}. \tag{2}$$

综合式(1)和式(2), 可得

$$p_{rs}(\alpha_r, \beta_s) = \frac{\mathrm{e}^{-(\alpha_r+\beta_s)}}{1+\mathrm{e}^{-(\alpha_r+\beta_s)}} = \frac{x_r y_s}{1+x_r y_s} \approx x_r y_s.$$

用 $\langle k_r\rangle, \langle k_s\rangle$ 表示两类节点的期望度, k_r^*, k_s^* 为两类节点的实际度, 最大化实际图出现的概率, 两者的关系为$\langle k_r\rangle = \sum_{s\in S} p_{rs} \approx k_r^*, \langle k_s\rangle = \sum_{r\in R} p_{rs} \approx k_s^*$. 进而有

$$k_r^* k_s^* \approx \left(x_r \sum_{s\in S} y_s\right)\left(y_s \sum_{r\in R} x_r\right) \approx p_{rs}\left(\sum_{s\in S}\sum_{r\in R} y_s x_r\right) \approx p_{rs}\left(\sum_{s\in S}\sum_{r\in R} p_{rs}\right) \approx p_{rs} L_{\mathrm{M}}, \tag{3}$$

其中 L_{M} 为二部图 M 中实际的边数, 则得到两类节点的连接概率 $p_{rs} = k_r^* k_s^*/L_{\mathrm{M}}$.

对于任意的两个顶级用户 r, r' 有共同回复者 s 的概率 $\Pr(V_{rr'}^s) = p_{rs}p_{r's} = (k_r^* k_s^*/L_{\mathrm{M}})(k_{r'}^* k_s^*/L_{\mathrm{M}})$, 两者的期望回复者数目以及实际回复者数目分别为

$$\langle V_{rr'}\rangle = \sum_{s\in S} p_{rs}p_{r's} = \frac{k_r^* k_{r'}^*}{L_{\mathrm{M}}^2}\sum_{s\in S}(k_s^*)^2, \quad V_{rr'}^* = \sum_{s\in S} m_{rs}m_{r's}.$$

假设 r, r' 之间不存在连边(即 r, r' 的共同回复者不是足够的多), 以 $V_{rr'}$ 作为代表 r, r' 间共同回复者数目的随机变量, 取值范围 $0, 1, \ldots, N_S$, 其服从泊松二项分布 f_{PB}, S_n 为全部可能的 S 中的 n 节点集构成的

集合, 则 $f_{\mathrm{PB}}(V_{rr'}=n)=\sum\limits_{E\in S_n}\prod_{s\in E}\Pr(V^s_{rr'})\prod_{s\notin E}(1-\Pr(V^s_{rr'}))$.

令 $\phi=\sum\limits_{V_{rr'}\geqslant V^*_{rr'}}^{N_S} f_{\mathrm{PB}}(V_{rr'})$, 如果 ϕ 小于阈值, 说明 $V_{rr'}$ 比实际值 $V^*_{rr'}$ 大的概率小, 那么实际的共同回复者数目已经够多了, 此时拒绝原假设, 认为两者间存在连边, 即有相似性. 而根据

$$\sum_{V_{rr'}=0}^{N_S}|f_{\mathrm{PB}}(V_{rr'})-f_{\mathrm{poi}}(V_{rr'})|<2\sum_{s\in S}\Pr(V^s_{rr'})^2,$$

不等式右边较小, 为了简化计算, 用泊松分布代替泊松二项分布, 则 $\phi=\sum\limits_{V_{rr'}\geqslant V^*_{rr'}}^{N_S} f_{\mathrm{poi}}(V_{rr'})$, 泊松分布以期望值 $\langle V_{rr'}\rangle$ 为参数.

得到系列 ϕ 后, 使用多重检验方法 FDR 对原假设进行联合检验. 将计算得到的 ϕ 从小到大排列

$$\phi^{(1)}\leqslant\phi^{(2)}\leqslant\cdots\leqslant\phi^{(K)},\quad K=\mathrm{C}^2_{N_R},$$

设 $t=0.05$, 求满足 $\phi^{(i)}\leqslant it/\mathrm{C}^2_{N_R}$ 的 i 最大值 i^*, 并以 $\phi^{(i^*)}$为阈值, 拒绝小于等于阈值的原假设, 确定顶级用户间连边. 当两顶级用户没有共同回复者时, $V^*_{rr'}=0$, 因为 N_S 很大, ϕ 接近 1, 肯定会接受原假设, 即两者间不存在连边.

3.2 顶级用户分组结果

顶级用户与普通用户的二部图中共有 43 336 个节点, 66 963 条边. 如果采用文献[19]的方法, 对顶级用户单射后, 得到 78 个节点, 1 635 条加权边, 其中的 71 个节点形成一个最大的闭环, 其余 7 个节点与该闭环相连, 则最终会形成一个社区, 无法区分不同用户组.

使用上述 BICM 确定顶级用户间的相似性连边, 得到顶级用户单模网络. 使用基于模块度优化的 Louvain 算法[13] 对网络进行社区划分. 顶级用户网络包含 78 个节点, 520 条边, 平均聚类系数 0.635, 图密度 0.173. 经图划分后得到 4 个大的用户组, 另有 6 个孤立节点.

为说明每个社区用户的话题偏好, 统计各社区顶级用户发布各种类型帖子的数目, 结果如图 4 所示.

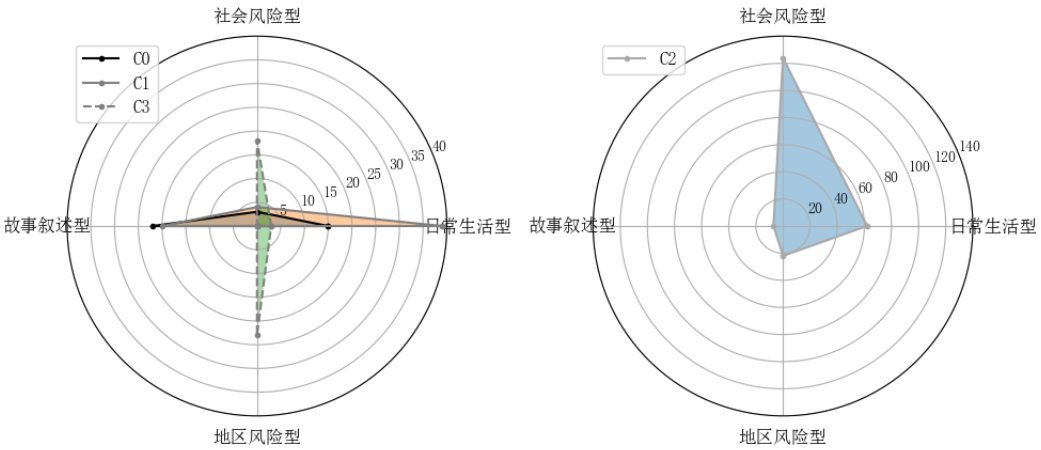


图 4 各社区顶级用户发帖类型统计

Fig. 4 Statistics of post types of top users in each community

图 4 中, C0 顶级用户组的发帖以故事叙述型为主, 另有部分日常生活型, 少部分社会风险型; C1 顶级用户组的发帖以日常生活型为主, 兼具故事叙述型与社会风险型; C2 顶级用户组发帖以社会风险型为主, 另有部分日常生活型及少量地区风险型与故事叙述型; C3 顶级用户组发帖以地区风险型为主, 兼具社会风险型, 日常生活型.

图 5 为顶级用户社区分布图, 各个社区基本呈现内部连接紧密, 外部稀疏连接的状态, 但 C0 与 C1 社区外部连接也相对紧密, 这是由于故事叙述型发帖与日常生活型发帖两者间存在共性. 介数中心性最大的

前 4 个节点: (49)“cotton 2000”(日常生活型与故事叙述型对半)、(58)“厚黑教主李宗吾”(社会风险、地区风险对半)、(2)“少华集团高董”(地区风险)和(24)“百财 2019”(故事叙述型、日常生活型、社会风险型和地区风险型), 从这些用户所在社区与发帖类型来看, 确实起着连接各社区的中介作用.

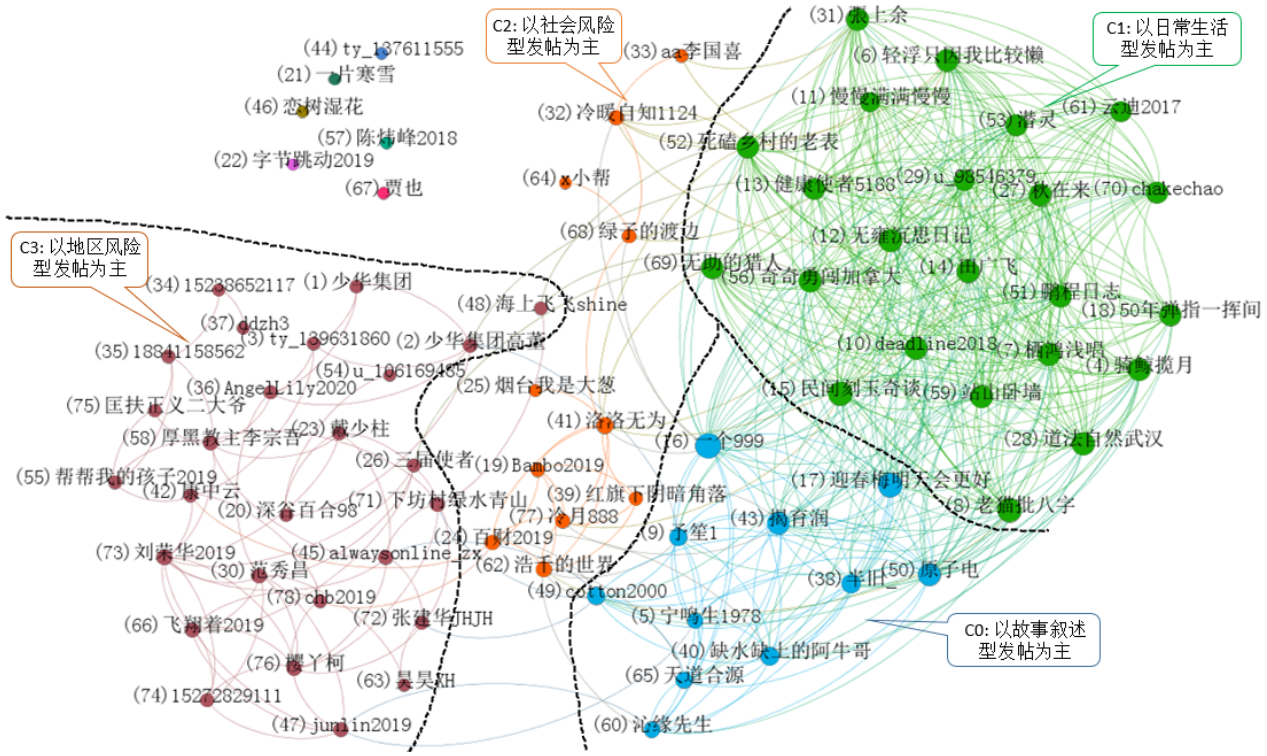


图 5　顶级用户社区分布

Fig. 5　Top user community distribution

为进一步说明每个社区用户话题的偏好性, 提取每个主贴的前 20 个 TextRank 关键词, 帖子标题的前 3 个 TextRank 关键词. 将每个顶级用户组的发帖关键词合并, 取词频最大的前 10 个关键词(对于第 10 位次的词语, 则将其相同词频的词语一并陈列):

C0(故事叙述型为主): [(‘没有’, 7), (‘人类’, 6), (‘人生’, 6), (‘朋友’, 5), (‘理论’, 5), (‘孩子’, 5), (‘七绝’, 5), (‘世界’, 4), (‘原理’, 4), (‘中国’, 4), (‘引力’, 4), (‘速度’, 4), (‘光速’, 4), (‘宇宙’, 4), (‘时间’, 4)];

C1(日常生活型为主): [(‘没有’, 39), (‘时候’, 34), (‘开始’, 25), (‘感觉’, 24), (‘孩子’, 15), (‘知道’, 13), (‘觉得’, 13), (‘看到’, 12), (‘事情’, 12), (‘生活’, 12), (‘早起’, 12), (‘锻炼’, 12), (‘可能’, 12)];

C2(社会风险型为主): [(‘图片’, 66), (‘没有’, 52), (‘发布’, 49), (‘问题’, 41), (‘社会’, 30), (‘造成’, 27), (‘不能’, 27), (‘进行’, 26), (‘知道’, 26), (‘需要’, 25)];

C3(区域风险型为主): [(‘公司’, 18), (‘没有’, 18), (‘有限公司’, 13), (‘法院’, 12), (‘相关’, 11), (‘投资’, 10), (‘诈骗’, 6), (‘法律’, 6), (‘事实’, 6), (‘项目’, 6), (‘证据’, 6), (‘师范类’, 6)].

C0 社区中的“人生”、“七绝”体现出用户在论坛中谈论自己的经历以及连载个人小说; C1社区中的“孩子”、“生活”表现出用户发布日常生活中的问题等; C2 社区的特征词“问题”、“社会”、“造成”体现了用户关注社会风险型话题; C3 社区中的“公司”、“法院”和“诈骗”等体现了用户谈论的是某个机构或者某个管辖区域的风险.

上文使用第三方普通用户来建立顶级用户间的连边, 并对顶级用户网络做了社区划分. 各社区顶级用户的实际交互情况如图 6 所示. 图 6 的横坐标表示每个顶级用户组, 纵坐标表示所回复各社区的总人数. C2 中的顶级用户回复的其它用户最多, 其中大部分为 C2 中的人. C3 组的用户所回复的用户也大多处在 C3 组. 这是由于同一社区中的用户偏好相似, 往往会产生交互, 也说明了本文社区划分的合理性.

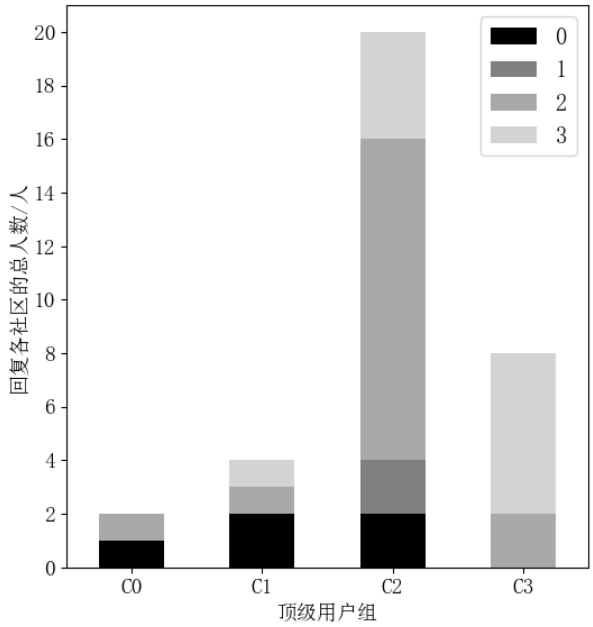


图 6 各社区顶级用户实际交互情况

Fig. 6 Actual interaction of top users in each community

4 普通用户极化分析

考虑二部图中度大于等于 2 的普通用户(回复的顶级用户数目多于 1), 计算每位普通用户回复每个组的顶级用户数占总数目的比例, 若最值仅一个且大于 0.25, 认为出现了极化, 得到 7 253 位极化用户.

根据比例, 极化于 C0 的普通用户组, 将其归到 I0 普通用户组, 计算 I0 组中普通用户对各顶级用户组回复比例的平均值, 以相同的方式计算普通用户的其它组别, 得到极化热度图(图 7). 分布结构显示普通用户极化现象明显, 特别是 I3 对于 C3 社区的极化较为突出, 即地区风险型话题更容易引起极化.

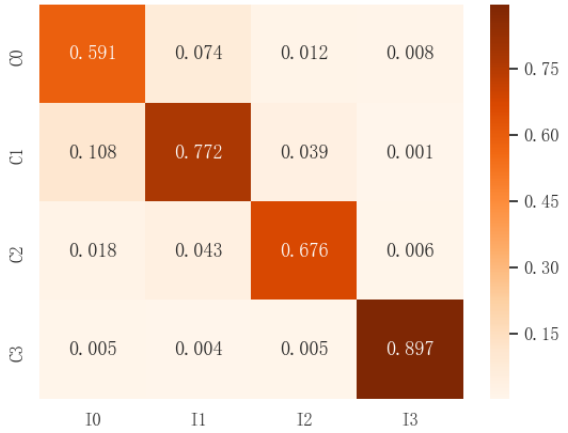


图 7 普通用户极化热度图

Fig. 7 Polarization heat map of ordinary users

极化分析根据回复体现的话题偏好实现了普通用户的分组. 将 C3 与 I3 用户合并, 该组别的用户偏向于关注地区风险型话题. 图 8 对这些用户所在地进行了统计, 网民参与了关于太原师范大学校园暴力, 张家口化工厂爆炸, 内蒙古赤峰市 2 000 名入学师范类定向大专生就业派遣诉求的地区风险型事件的讨论, 而该组别中山西、河北和内蒙古的用户居多, 即该组用户主要集中在风险发生地. 其它 3 个相应的合并用户组中用户主要所在地基本都为北京、广东和江苏等.

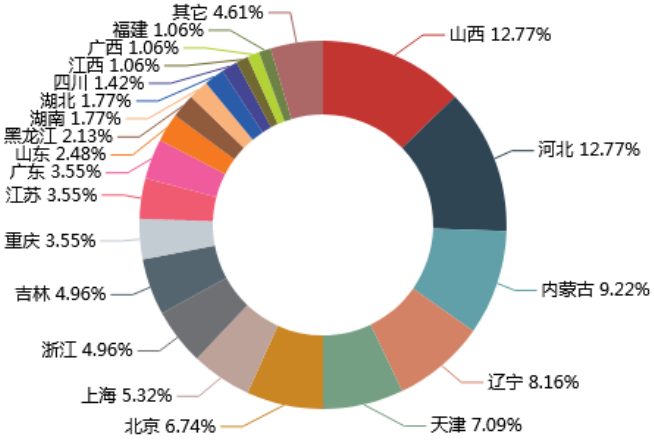


图 8　关注于地区风险型话题的用户所在地

Fig. 8　Location of users focusing on regional risk topics

5　重要回复者识别

上文基于用户发帖与回帖所体现的话题偏好实现了顶级用户与普通用户的分组, 并说明了社区划分的合理性, 这有助于为各组别用户精准推荐其感兴趣的帖子, 并进一步获取民意. 而根据这种回复关系, 寻找每位顶级用户的历史重要回复者, 对于回复者预测具有重要意义. 由于在论坛中, 通过用户界面的回复历史, 就可以溯源到相应原贴, 因此回复行为会扩大消息的传播. 重要回复者是顶级用户众多回复者中较为稳定的一部分, 当顶级用户发布负面新闻时, 通过限制这些重要回复者的发言, 有利于及时阻滞负面消息的传播, 加强网络治理. 当顶级用户发布正面新闻时, 即时推送给重要回复者, 引起追随, 起到加速消息传播的效果. 下面将结合两个二部图以及相应模型尝试寻找顶级用户的重要回复者.

5.1　模型设计

R, Q, C 分别表示顶级用户集, 帖子集, 收集到的全部用户集, 节点数目分别为 N_R, N_Q, N_C. 图 9 为基于发帖与回帖关系的联合二部图, 图中的两个部分, 一个表示发帖关系, 另一个表示回贴关系. 若某用户回复了另外一位用户统计意义上的大部分帖子, 那么, 认为这个用户是另外一位用户的重要回复者, 如 c_2 为 r_1 的重要回复者.

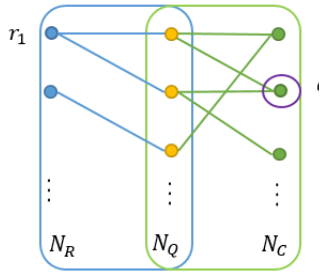


图 9　基于发帖与回帖关系的联合二部图

Fig. 9　Joint bipartite graph based on post and reply relationship

利用图 9 中的两个二部图, 寻找顶级用户的重要回复者. 将左边的二部图记为 M_1, 右边的二部图记为 M_2, 则 $\Pr(\mathrm{M}_1)=\prod\limits_{r\in R}\prod\limits_{q\in Q}(p_{rq})^{m_{rq}}(1-p_{rq})^{(1-m_{rq})}$.

M_1 中帖子的度都为 1, 不需要对其度进行限制, 因此采用二部部分配置模型来获取顶级用户 r 发布帖子 q 的概率 p_{rq}, 则

$$H(\mathrm{M}_1)=\sum_{r\in R}\alpha_r k_r,\quad p_{rq}(\alpha_r)=\frac{\mathrm{e}^{-\alpha_r}}{1+\mathrm{e}^{-\alpha_r}}=\frac{x_r}{1+x_r},\quad \langle k_r\rangle=\sum_{q\in Q}p_{rq}\approx k_r^*,\quad p_{rq}=\frac{k_r^*}{N_Q}.$$

图 M_2 仍采用二部配置模型, 按照式(1)∼式(3), 类似地得到用户 c 回复帖子 q 的概率 $p_{cq}=k_c^*k_q^*/L_{\mathrm{M}_2}$, 其中 L_{M_2} 为图 M_2 中的实际边数.

普通用户 c 回复了顶级用户 r 的发帖 q 的概率 $\Pr(V_{rc}^q)=p_{rq}p_{cq}=k_r^*k_c^*k_q^*/(N_QL_{\mathrm{M}_2})$, 普通用户 c 回复了顶级用户 r 的期望帖子数 $\langle V_{rc}\rangle$ 与实际帖子数 V_{rc}^* 分别为

$$\langle V_{rc}\rangle=\sum_{q\in Q}p_{rq}p_{cq}=\sum_{q\in Q}\frac{k_r^*}{N_Q}\frac{k_c^*k_q^*}{L_{\mathrm{M}_2}}=\frac{k_r^*k_c^*}{N_Q},\quad V_{rc}^*=\sum_{q\in Q}m_{rq}m_{cq}.$$

对于 R,C 中的每一组节点(共 N_RN_C 组), 假设 c 不是 r 的重要回复者(即 c 回复 r 的帖子不是足够的多). 令随机变量 V_{rc} 代表 c 回复 r 的帖子数, 取 $0,1,2,\ldots,N_Q$, 其服从式(4)所示的泊松二项分布 f_{PB},Q_n 为全部可能的帖子集 Q 中的 n 节点集构成的集合.

$$f_{\mathrm{PB}}(V_{rc}=n)=\sum_{E\in Q_n}\prod_{q\in E}\Pr(V_{rc}^q)\prod_{q\notin E}(1-\Pr(V_{rc}^q)).\tag{4}$$

进而

$$\psi=\sum_{V_{rc}\geqslant V_{rc}^*}^{N_Q}f_{\mathrm{PB}}(V_{rc}),\tag{5}$$

式(5)可简化为 $\psi=\sum\limits_{V_{rc}\geqslant V_{rc}^*}^{N_Q}f_{\mathrm{poi}}(V_{rc})$, 其中泊松分布 f_{poi} 以 $\langle V_{rc}\rangle$ 为参数. 得到系列 ψ 后, 仍使用多重检验方法 FDR 对原假设进行检验, 从而确定各顶级用户的重要回复者.

5.2 重要回复者识别结果

如果从超过平均值的角度确定顶级用户的重要回复者, 计算过程分为两步:

1) 对于仅回复过一个顶级用户, 且回帖数目(指回复顶级用户发布的不同帖子的数目)不为 1 的用户, 初步确定其为该顶级用户的重要回复者. 除此以外, 若用户回复某个顶级用户的帖子数超过其平均回帖水平, 则初步认为该用户为相应顶级用户的重要回复者;

2) 对于初步确定的重要回复者, 如果其对相应顶级用户的回帖数超过该顶级用户所有回复者回帖数目的平均值, 则将其确定为该顶级用户的重要回复者.

考虑 78 位顶级用户所发布的 452 条帖子以及涉及到的 43 237 位回复者, 使用上述方法, 得到 4 783 对重要回复关系, 3 742 位重要回复者, 统计重要回复者对相应顶级用户的回复比例(回复该顶级用户的帖子数目/总回贴数目), 绘制图 10 所示的频数分布直方图. 图 10 表明 4 个区间频数相差不大.

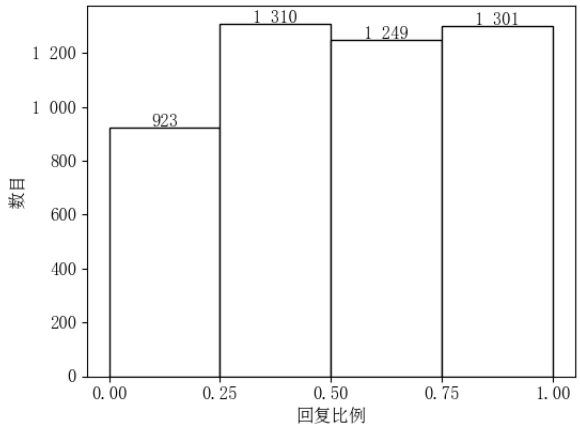


图 10 回复比例分布图

Fig. 10 Distribution of reply proportion

如果从回复比例角度确定顶级用户的重要回复者, 首先排除回复顶级用户帖子数全为 1 的回复者、发

帖数目仅为 1 的顶级用户, 若某回复者回复某顶级用户的帖子数与该顶级用户总发帖数之比大于 75 %, 将其视为相应顶级用户的重要回复者. 得到 110 对重要回复关系, 它们在上述 4 个区间的分布比例为 63 : 15 : 14 : 17, 重要回复者对顶级用户的依附不强.

根据发贴关系与回贴关系建立两个二部图, 结合 5.1 节的模型, 从统计意义上确定每位用户的重要回复者. 有 65 位顶级用户有重要回复者, 共计 8 546 对重要回复关系, 重要回复者 8 543 位, 几乎一位回复者依附于一位顶级用户. 图 11 统计了顶级用户回复者的数目.

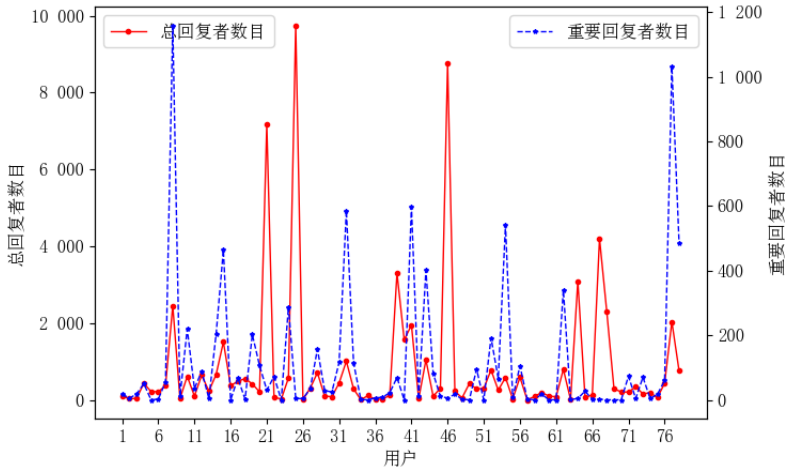


图 11　各顶级用户总回复者数目与重要回复者数目

Fig. 11　Total number of responders and number of important responders per top user

在这 8 546 对重要回复关系中, 有 8 471 对的重要回复者仅仅回复了该顶级用户的一个帖子, 再无其它发言. 统计另外 175 对中重要回复者的回复比例, 图 12 为频数分布直方图.

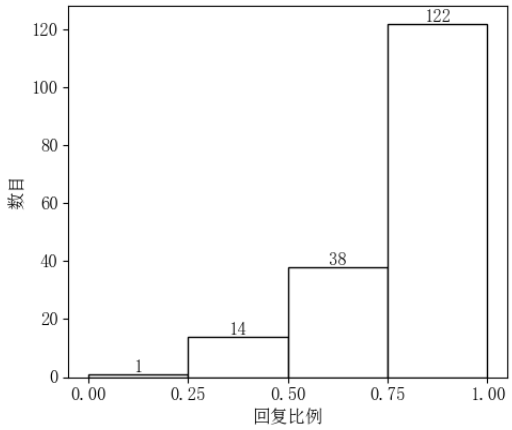


图 12　175 对重要回复关系中回复比例分布图

Fig. 12　Distribution of reply proportion in 175 important reply relationships

图 12 中有 122 对重要回复关系, 其重要回复者对所依附的顶级用户的回复比值达到 75 % 以上, 相比于图 10, 回复偏重性明显. 因此本文提出的基于统计验证确定顶级用户重要回复者的方法, 在保证重要回复者回复顶级用户足够多帖子的同时, 也保证了重要回复者对顶级用户的回复偏重, 且自动给予了“足够多”、“偏重”合理的限定. 175 对重要回复关系中包含 22 位存在重要回复者的顶级用户, 这些用户中有 9 个属于 C3 社区(以地区型风险发帖为主), 6 个属于 C2 社区(以社会型风险发帖为主), 4 个属于 C1 社区(以日常生活型发帖为主), 其余 3 个是不在社区内的孤立节点, 由此可见, 关注于风险型话题的顶级用户易存在重要回复者.

5.3 重要回复关系双方发言内容情感分析

本小节将分析顶级用户发帖内容情感极性与重要回复者相应回复内容情感极性间的关系. 由于天涯杂谈帖子正文通常很长, 且多引用事例, 而标题一般概括了作者的态度, 因此, 本文仅考虑帖子标题. 因为旨在探索首次回复关系的建立, 仅考虑重要回复者对相应帖子的第一次回复.

具体的, 对于每位存在重要回复者的顶级用户, 使用百度情感分析 API[2]分析其全部重要回复者对其发帖的回复及相应帖子标题对(总计 2 958 对)的情感极性, 获取正面情绪发帖–正面情绪回复, 正面情绪发帖–负面情绪回复, 负面情绪发帖–负面情绪回复, 负面情绪发帖–正面情绪回复的比例, 见图 13.

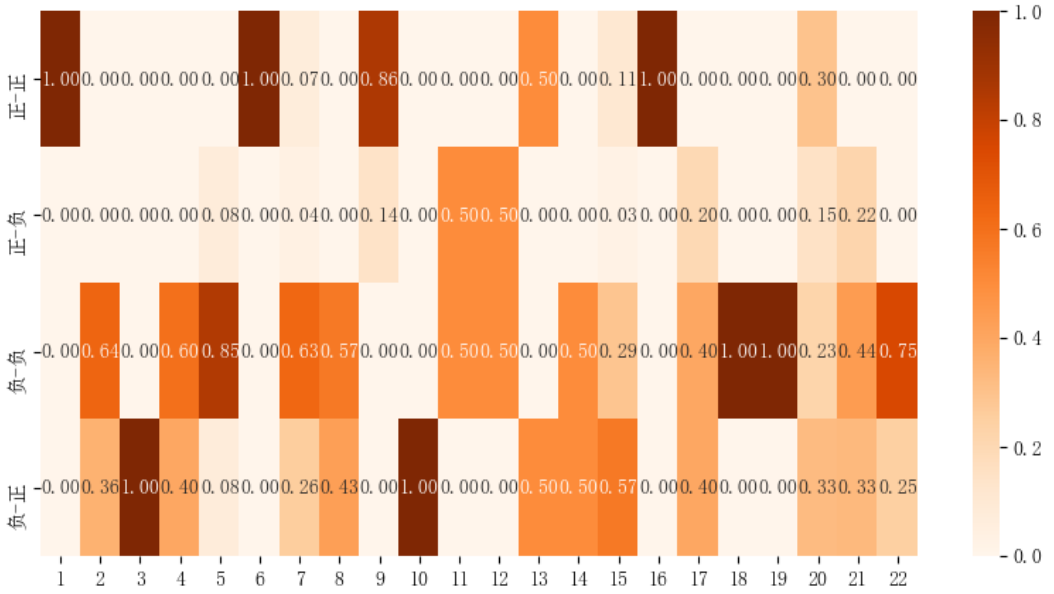


图 13 顶级用户不同情感对比例分布

Fig. 13 Proportion distribution of top users' different emotion pairs

这些顶级用户多发表负面情绪的新闻, 且负–负比值大于负–正比值的用户有 11 个, 前者小于后者的用户有 5 个, 两者相等的用户有 6 个. 这说明了在负面情绪新闻居多, 负面情绪易被重要回复者放大的网络环境中, 识别重要回复者并在相应顶级用户发表极负面新闻情境下对该重要回复者进行制约的重要性.

6 结束语

本文以天涯论坛为例, 定义顶级用户与普通用户, 开展了用户分组与重要回复者识别研究工作. 用户分组研究借助于二部配置模型, 通过第三方普通用户的统计意义上足够多的回复行为来构建顶级用户网络, 进而实现顶级用户社区划分. 不仅所得到的相似性连边是可信的, 而且避免了由直接回复关系构建顶级用户稀疏单模网而无法划分社区的后果. 帖子标题聚类结果表明网民们关注的话题包含日常生活型、社会风险型、故事叙述型、地区风险型 4 大类, 得到的 4 个用户组各自主要发帖类型对应这 4 个帖子簇类型, 同组的用户具有相似的话题偏好, 且交互密切. 对于普通用户, 则以回复行为能够体现兴趣偏好为视角, 使用极化分析的方法确定所属组别. 用户分组有助于下游任务——用户个性化推文, 这对于网络精准营销与民意及时获取具有实际意义. 本文着眼于使用统计验证的方法确定顶级用户的重要回复者, 从而推动回复者预测研究. 具体的, 结合了 BICM 与 BIPCM 两种模型建模发帖和回帖关系的二部图, 这是对于配置模型仅用于单一二部图的扩展. 筛选出的重要回复者, 不仅是经过验证的高频回复者, 且对相应顶级用户的回复偏重性明显. 此外, 发现存在重要回复者的顶级用户多发表负面情绪的新闻, 此时重要回复者带有负面情绪的回复也居多. 因此, 识别重要回复者并适时对其进行制约有助于舆情管理与净化网络环境.

[2]https://ai.baidu.com/tech/nlp/sentiment_classify

文章所建立的回复关系二部图未考虑权重, 多次回复与单次回复在强度上还是有差异的, 今后尝试将频次因素加入到研究中, 探索其对实验结果的影响, 并进一步分析这种影响是否带来了本质的改变. 未来也将参考不同流派的研究工作, 集成各自优势, 改进模型.

参考文献:

[1] Mohbey K K. Multi-class approach for user behavior prediction using deep learning framework on twitter election dataset. Journal of Data, Information and Management, 2020, 2(1): 1–14.

[2] Balalau O, Castillo C, Sozio M. Evidense: A graph-based method for finding unique high-impact events with succinct keyword-based descriptions // 12th International AAAI Conference on Web and Social Media. Palo Alto, CA: AAAI Press, 2018.

[3] 许　诺, 唐锡晋. 基于百度热搜新闻词的社会风险事件 5W 提取研究. 系统工程理论与实践, 2020, 40(2): 334–342.
Xu N, Tang X J. A study on 5W extraction for societal risk events based on hot news search words. Systems Engineering: Theory & Practice, 2020, 40(2): 334–342. (in Chinese)

[4] 刘　勘, 袁蕴英. 基于自动编码器的短文本特征提取及聚类研究. 北京大学学报(自然科学版), 2015, 51(2): 282–288.
Liu K, Yuan Y Y. Short texts feature extraction and clustering based on auto-encoder. Acta Scientiarum Naturalium Universitatis Pekinensis(Nature Science Edition), 2015, 51(2): 282–288. (in Chinese)

[5] Pinto S, Albanese F, Dorso C O, et al. Quantifying time-dependent media agenda and public opinion by topic modeling. Physica A: Statistical Mechanics and its Applications, 2019, 524: 614–624.

[6] 吴　俊, 欧阳书凡, 李晓华. 基于 STM 和格兰杰因果分析的网络新闻媒体倾向研究.系统工程学报, 2020, 35(4): 446–458.
Wu J, Ouyang S F, Li X H. Media slant of online news based on structural topic model and Granger causality analysis. Journal of Systems Engineering, 2020, 35(4): 446–458. (in Chinese)

[7] Hu Y, Wang S, Ren Y, et al. User influence analysis for Github developer social networks. Expert Systems with Applications, 2018, 108: 108–118.

[8] 魏杰明, 何　慧. 社交网络中用户行为及影响力评估算法研究. 智能计算机与应用, 2019, 9(2): 162–167.
Wei J M, He H. Research on user behavior and influence algorithm in social network. Intelligent Computer and Applications, 2019, 9(2): 162–167. (in Chinese)

[9] Jain L, Katarya R. Discover opinion leader in online social network using firefly algorithm. Expert Systems with Applications, 2019, 122: 1–15.

[10] Garibay I, Mantzaris A V, Rajabi A, et al. Polarization in social media assists influencers to become more influential: Analysis and two inoculation strategies. Scientific Reports, 2019, 9(1): 1–9.

[11] 李泉林, 李超然, 马静宇. 超图结构下的在线社交网络中隐性影响力评估. 系统工程学报, 2020, 35(1): 130–144.
Li Q L, Li C R, Ma J Y. Online social networks under hypergraph structure and their hidden influence evaluation. Journal of Systems Engineering, 2020, 35(1): 130–144. (in Chinese)

[12] Liao S H, Yang C A. Big data analytics of social network marketing and personalized recommendations. Social Network Analysis and Mining, 2021, 11(1): 1–19.

[13] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10): P10008.

[14] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 701–710.

[15] Tang J, Qu M, Wang M, et al. Line: Large-scale information network embedding // Proceedings of the 24th International Conference on World Wide Web. New York: ACM, 2015: 1067–1077.

[16] Sun H, Ch'ng E, Yong X, et al. A fast community detection method in bipartite networks by distance dynamics. Physica A: Statistical Mechanics and its Applications, 2018, 496: 108–120.

[17] Barber M J. Modularity and community detection in bipartite networks. Physical Review E, 2007, 76(6): 066102.

[18] Zhou C, Feng L, Zhao Q. A novel community detection method in bipartite networks. Physica A: Statistical Mechanics and its Applications, 2018, 492: 1679–1693.

[19] Tackx R, Tarissan F, Guillaume J L. ComSim: A bipartite community detection algorithm using cycle and node's similarity // International Conference on Complex Networks and their Applications. Berlin: Springer, 2017: 278–289.

(下转第 708 页)