一种微博用户情感影响者发现模型

朱 江,王 柏,吴 斌,李小明
(北京邮电大学计算机学院,北京 100876)

摘 要: 情感在微博网络中传播并感染用户,对微博网络甚至现实世界都有重要影响.发现具有情感影响力的用户(情感影响者)对社会管理或制定市场策略等具有重要意义.本文建立了包含两种节点(用户,微博)和三种关系(转发,关注,发帖)的异质微博网络,利用微博情感相似性和用户情感行为相似性将其转化为只包含用户节点的同质网络,进而在该网络中使用随机游走模型发现情感影响者.贡献包含以下方面:利用微博情感相似性和用户的情感行为相似性验证了本文所构建微博网络的情感同配性,确认了情感影响在该网络中存在;提出 EmotionRank 模型用以寻找情感影响者;基于微博数据的实验结果有效验证了该模型的有效性和优越性.

关键词: 微博网络; 情感影响; 情感相似性; 情感同配性
中图分类号: TN911.23 **文献标识码:** A **文章编号:** 0372-2112 (2015)12-2497-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.12.023

A Model for Finding Emotional Influencers in Microblog

ZHU Jiang, WANG Bai, WU Bin, LI Xiao-ming
(*School of Computer Science, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

Abstract: In microblog network, emotion propagates among users and infects user behavior. These emotional behaviors have an important impact on the microblog network or even the real world, so finding emotional influencers in microblog network is very significant for society management or marketing strategies. In this paper, a heterogeneous microblog network that contains two types of nodes (user, microblog) and three types of relations (forwarding, following, posting) is constructed. Utilizing the similarity of microblogs' emotion and users' emotional behavior, the heterogeneous network is transformed to be a homogeneous network that contains only users. Random walk model is employed to find emotional influencers in this network. Our contribution is summarized as follows: we verify emotional homophily in the microblog network constructed by our dataset, which confirms the existence of emotional influence in this network; we propose a novel model (EmotionRank) to find emotional influencers; experimental results effectively illustrate the utility and superiority of EmotionRank.

Key words: microblog network; emotional influence; emotional similarity; emotional homophily

1 引言

社交网络在全世界范围内展现出惊人的发展速度,微博作为重要组成部分,吸引了众多用户并传播各种信息[1],是人们交流和传播各类情感的重要平台.

单条微博最多有 140 字,但其传递的情感却对社会网络有重要影响.当今社会热点事件层出不穷,用户群体的情感互动甚至能影响事件进展.所以发现具有情感影响力的用户-情感影响者,对社会管理或市场策略制定等都具有重要意义.

微博网络中用户可参与多个主题,每个主题也可吸引成千上万的用户,该过程可唤起用户多样的情感.如此复杂的场景意味着不同用户在不同主题下有不同的情感影响力,也有一些用户在很多主题中都有较大影响力.所以寻找情感影响者是一项极具挑战的任务.

微博网络包含多种节点(用户、微博等)和关系(关注、转发等),这些节点和关系构成一个异质社会网络,用户情感在该网络中传播并产生影响.图 1 是微博网络示例,有 3 个用户,2 条微博.用户 1、2 相互关注,用户 3 关注 2,用户 1、3 之间无‘关注’关系.大多时候微博用户主要与自己‘关注’的朋友互动,但用户 3 却转发了用户 1 的微博,说明用户 3 被用户 1 所发内容影响.转发行为可看做微博用户间情感传播影响过程.一些经典算法仅考虑‘关注’关系用户间的互动,较少考虑‘非

收稿日期:2015-01-30;修回日期:2015-07-26;责任编辑:蓝红杰
基金项目:国家 973 重点基础研究发展计划(No.2013CB329603);国家自然科学基金(No.7123100,No.61375058)

关注'关系用户间的'转发'行为,这可能导致发现影响者结果偏差.本文将利用微博网络'转发'、'关注'等更丰富的用户关系,有效发现微博网络某主题情感影响者和那些超越了主题的整体情感影响者.

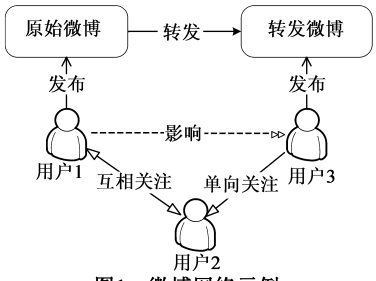


图1 微博网络示例

2 相关工作

国内外学者近年提出多种方法研究社会网络用户影响力.文献[2]提出 TwitterRank 模型,利用 LDA 模型构建权重网络并随机游走来度量 Twitter 用户影响力.文献[3]提出双峰公平对赌模型,利用 LinkIn 数据发现真实世界影响力用户.文献[4]利用图模型挖掘某主题帖子影响力,最终发现影响者.文献[5]提出并利用属性列表来确定微博网络主题权威用户.文献[6]利用异质网络挖掘了论坛意见领袖社区.文献[7]研究 160 0000 个 Twitter 用户关注关系网络的 7400 0000 个事件扩散过程,分析用户属性和影响力.

而上述工作未考虑情感因素,较少考虑微博内容,更多是关注网络拓扑结构和用户行为.可拓扑结构并不能有效发现用户真正影响力[8].文献[9]利用 Twitter 数据探讨情感与影响力的关系,认为用户影响力不能简单通过网络图属性得到,而应依赖用户行为背后隐含的情感.

在微博二元情感挖掘方面,RostEA 软件[10]通过构建情感词典确定语料所表达情绪,可计算中文微博正负情感分值,其针对微博情感挖掘测试结果经交叉判定正确率达 80.6%.

个人情绪状态模型(GPOMS)[11]可度量文本 6 维情感.它分析了 Twitter 用户 6 维情感时间序列,描述了用户群体情感变化.

文献[12]利用 7 元(乐、好、怒、哀、惧、恶、惊)情感时间序列描述社会网络用户情感行为.该方法描绘的用户情感曲线很好匹配了真实世界人们的节日情感表现.本文将继续使用此法,并度量用户情感行为相似性.

社会网络情感影响研究方面,文献[13]利用 Filickr 数据,基于图片挖掘用户情感,不仅验证情感影响在网络中存在,还构建网络用户情感影响概率图模型预测用户情感.文献[14]定量分析用户情感如何被影响,并预测用户情感.文献[15]利用社会网络用户间影响,从 Filickr 图片推断用户情绪.上述工作仅从二元角度描述情感,不能从多元角度分析社会网络用户情感行为.本文将结合二元情感和多元情感度量微博网络用户情感行为以及相似性.

情感影响是否存决定了本文工作的必要性.文献[16]探讨了用户影响与相似性间的关系,认为用户间相互影响提高了用户行为相似性.文献[17]认为社会网络同配性与用户间影响自相关.文献[18]使用网络用户情感相似性来度量用户间情感影响.以上工作认为社会网络用户同配性与用户间影响直接相关,本文也认为情感同配性是微博网络用户情感影响存在前提.

社会网络同配性研究方面,文献[19]认为社会网络幸福度具有同配性.文献[20]研究了 Twitter 对话情感模式,发现对话者间往往有相似情感,相似比例高达 96%.文献[21]通过 Facebook 大规模实验,发现情绪感染现象.实验证明情绪传染不需要通过人与人之间直接互动就能发生.上述工作表明社会网络情感同配性存在.由于本文使用微博数据,验证该数据情感同配性至关重要.

本文将用微博情感相似性和用户多元情感行为相似性验证所建立微博网络情感同配性,确认该网络用户间情感影响存在,进而挖掘情感影响者.

3 微博数据来源

新浪微博注册用户已超 5 亿,每天发帖量超 1 亿,为反应中国网民情感行为提供了坚实而丰富的数据基础.

新浪应用程序接口(API) 被用以抓取微博数据.该 API 抓取单个用户微博最大量为 2000,所以抓取的一部分不活跃用户微博时间跨度较长,一部分较活跃用户微博时间跨度很短.时间是描述用户情感行为重要因素,实验有必要保证在某一时间段内用户微博数据完整性.2013/10/22 至 2013/11/9 被选为实验时间段,该时间段内实验用户微博全被抓取.

用户可用多种表达方式发布微博,如各种语言、符号等,甚至使用无法识别的垃圾信息.因此有必要过滤实验数据.过滤过程使用正则表达式发现微博内容是否包含中文文本,如包含则保留,反之则放弃该微博.本文目标是发现情感影响者,那些从未发言的用户也被过滤掉.

最终实验数据包含 2470 个用户和 106137 条关注关系.这些用户在 19 天内发布了 174984 条微博,包含 81455 条转发关系.

4　情感相似性和同配性

4.1　同配性概念

微博网络拥有两种节点(用户,微博),情感同配性意味用户更愿关注与之有相似情感行为的用户,更愿转发与之具有相似情感的微博.

微博网络情感同配性是该网络用户间情感影响存在的前提.以下两个问题有助于确认该微博网络是否存在情感同配现象.

(1)具有'转发'关系的微博间是否比不具有转发关系的微博间有更加情感相似的情感内涵?

(2)具有'关注'关系的微博网络用户间是否比不具有关注关系的用户间有更加相似的情感行为?

分析情感同配性之前,先定义微博情感相似性和微博用户情感行为相似性.

4.2　情感相似性

4.2.1　微博情感相似性

单条微博内容短小,包含较少情感词.虽然微博可表达多样情感,但从多元情感角度计算微博间情感相似性并不合理.本文采用二元情感计算微博间情感相似性.

RostEA 是经典的二元情感计算软件.其计算结果认为中性情感值分布在[-5, +5],积极和消极情感值分别分布在(+5, + ∞)和(- ∞, -5),[-25, +25]之外是最激烈情感.RostEA 按照文本包含情感词进行情感评分,单条微博即使表达最激烈情感,也不会包含大量情感词.所以认为 [-25, +25]之外表示最激烈情感较合理.微博间情感相似性定义为:

$$S_{\mathrm{m}}=1-\frac{|e_1-e_2|}{50},\ e_i=\begin{cases}\max(e_i,-25), & e_i\leqslant 0\\ \min(e_i,+25), & e_i\geqslant 0\end{cases}\tag{1}$$

e_i 为微博情感值,式(1)意味拥有相近情感值微博间具有更相似情感.

4.2.2　用户情感行为相似性

文献[12]提出多元情感时间序列描述网络用户情感行为,并使用 PCA 相似性(S_{PCA})和距离相似性(S_{dist})分别度量用户间情感波动和情感强度相似性.本小节对这两种相似性做简要介绍.

用户每天聚合微博都可抽取 7 元情感向量 $\boldsymbol{\beta}_i=(e_{\mathrm{happy}}\ e_{\mathrm{good}}\ e_{\mathrm{anger}}\ e_{\mathrm{sorrow}}\ e_{\mathrm{fear}}\ e_{\mathrm{hate}}\ e_{\mathrm{shock}})^{\mathrm{T}}$.每个用户所有微博以天为单位可构建一个多元情感时间序列($\boldsymbol{\beta}_1\ \boldsymbol{\beta}_2\ \boldsymbol{\beta}_3\ \cdots\ \boldsymbol{\beta}_n$).该序列可描绘用户情感波动和强度.

扩展 Frobenius 范数(Eros)被用以定义 PCA 相似性:

$$S_{\mathrm{PCA}}=\mathrm{Eros}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{w})=\sum_{i=1}^{m}w_i\cdot|\langle\boldsymbol{a}_i,\boldsymbol{b}_i\rangle|=\sum_{i=1}^{m}w_i\cdot|\cos\theta_i|\tag{2}$$

$\boldsymbol{A},\boldsymbol{B}$ 为 $m\times n$ 多元时间序列,m 为情感维数,n 为天数.$\boldsymbol{V}_A(\boldsymbol{V}_B)$是对其协方差矩阵 $\boldsymbol{M}_A(\boldsymbol{M}_B)$做 SVD 分解后的右特征向量矩阵.令 $\boldsymbol{V}_A=[\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots\boldsymbol{a}_m]$,$\boldsymbol{V}_B=[\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots\boldsymbol{b}_m]$,$\boldsymbol{a}_i\boldsymbol{b}_i$ 为 m 维列特征向量.令 $\boldsymbol{\lambda}_A=(\lambda_1^A,\lambda_2^A,\cdots\boldsymbol{\lambda}_m^A)$,$\boldsymbol{\lambda}_B=(\lambda_1^B,\lambda_2^B,\cdots\lambda_m^B)$,$\lambda_i^A(\lambda_i^B)$ 为 $\boldsymbol{M}_A(\boldsymbol{M}_B)$特征值,$\lambda_1^A\geq\lambda_2^A\cdots\geq\lambda_m^A\geq 0$,$\lambda_1^B\geq\lambda_2^B\cdots\geq\lambda_m^B\geq 0$.

$\langle\boldsymbol{a}_i,\boldsymbol{b}_i\rangle$为 $\boldsymbol{a}_i,\boldsymbol{b}_i$ 内积,$\boldsymbol{w}=(w_1,w_2,\cdots w_m)$为基于多元时间序列特征值所得出的权值向量,$\sum_{i=1}^{m}w_i=1$ 且 $w_i\geq 0$.$\cos\theta_i$为 $\boldsymbol{a}_i,\boldsymbol{b}_i$ 夹角余弦.显然 S_{PCA}介于 0,1 间,0 代表完全不相似,1 代表完全相似.

权重 w_i 基于特征值的启发式算法计算.$\lambda_i^A(\lambda_i^B)$反映了 $\boldsymbol{A},\boldsymbol{B}$ 主成分所包含信息,定义 $w_i=w_i^{\boldsymbol{AB}}/\sum(w_i^{\boldsymbol{AB}})$,$w_i^{\boldsymbol{AB}}=0.5(\lambda_i^A+\lambda_i^B)$.

S_{dist}可区分空间中具有相似方向却距离很远的两点.若 $\boldsymbol{A},\boldsymbol{B}$ 具有相同主成分,却有不同元素值时,S_{dist}作用则体现出来.

S_{dist}基于空间距离范数定义.$\boldsymbol{A},\boldsymbol{B}$ 间范数 $\phi_{\boldsymbol{AB}}=\|\boldsymbol{A}-\boldsymbol{B}\|$,距离相似性定义为:

$$S_{\mathrm{dist}}=\sqrt{\frac{2}{\pi}}\int_{\phi_{\boldsymbol{AB}}}^{\infty}\mathrm{e}^{-z^2/2}\mathrm{d}z\tag{3}$$

$0\leqslant S_{\mathrm{dist}}\leqslant 1$,马氏距离 $\phi_{\boldsymbol{AB}}=\sqrt{(\boldsymbol{C}_A-\boldsymbol{C}_B)^{\mathrm{T}}\boldsymbol{\Sigma}_A^{*-1}(\boldsymbol{C}_A-\boldsymbol{C}_B)}$用以描述 $\boldsymbol{A},\boldsymbol{B}$ 间范数.$\boldsymbol{C}_A,\boldsymbol{C}_B$ 为 $\boldsymbol{A},\boldsymbol{B}$ 中心点.$\boldsymbol{\Sigma}_A^{*-1}$ 为 $\boldsymbol{A}$ 协方差矩阵的伪逆矩阵.

PCA 相似性和距离相似性的结合可度量用户情感波动和情感强度相似性,如下:

$$\mathrm{SF}=\alpha S_{\mathrm{PCA}}+(1-\alpha)S_{\mathrm{dist}},\quad 0<\alpha<1\tag{4}$$

4.3　情感同配性

微博网络有无数'转发'、'关注'关系,本节使用假设检验验证微博网络情感同配性.

4.3.1　微博情感同配性

令 $\mu_{\mathrm{forwarding}}$为转发微博对之间的情感相似性均值,$\mu_{\mathrm{unforwarding}}$为非转发微博对之间的情感相似性均值,使用双样本 t-检验(假设方差不等)测试二者间关系.

假设 $H_0:\mu_{\mathrm{forwarding}}\leqslant\mu_{\mathrm{unforwarding}}$,$H_1:\mu_{\mathrm{forwarding}}>\mu_{\mathrm{unforwarding}}$.

首先计算具有'转发'关系微博对间的情感相似性,可得 $\mu_{\mathrm{forwarding}}$样本数据.然后对每个转发微博,随机选择配对一个原始微博.选择原始微博时,注意被选原始微博与转发微博间应不具有'转发'关系.这样可获得 $\mu_{\mathrm{unforwarding}}$样本数据.两样本具有相同数量,可使用 Matlab 的 ttest2 函数检测.

结果表明应拒绝 $H_0:\mu_{\mathrm{forwarding}}\leqslant\mu_{\mathrm{unforwarding}}$,显著性水平为 alpha = 10^{-6}.这说明具有转发关系的微博对间拥有更相近情感.

4.3.2 用户情感行为同配性

令 $\mu_{\text{following\_pca}}(\mu_{\text{following\_dist}})$ 为‘关注’关系用户间的PCA(距离)情感相似性均值.且 $\mu_{\text{unfollowing\_pca}}(\mu_{\text{unfollowing\_dist}})$ 为不具有‘关注’关系用户间情感相似性均值.依然用双样本 t-检验.

令 H_0: $\mu_{\text{following\_pca}}(\mu_{\text{following\_dist}}) \leqslant \mu_{\text{unfollowing\_pca}}(\mu_{\text{unfollowing\_dist}})$, H_1: $\mu_{\text{following\_pca}}(\mu_{\text{following\_dist}}) > \mu_{\text{unfollowing\_pca}}(\mu_{\text{unfollowing\_dist}})$.

首先计算具有‘关注’关系用户间 PCA(距离) 情感相似性,可得 $\mu_{\text{following_pca}}(\mu_{\text{following_dist}})$样本数据.然后对每个被‘关注’用户,随机选取另一用户与之配对,计算二者间的PCA(距离)情感相似性.随机选取用户时,应保证配对用户间不具有‘关注’关系.这样可获得 $\mu_{\text{unfollowing_pca}}(\mu_{\text{unfollowing_dist}})$样本数据.

检验结果拒绝 H_0,说明拥有‘关注’关系微博网络用户间具有更相似的情感波动和情感强度.

上述实验回答了 4.1 的问题,验证了本文所构建微博网络的情感同配性,确认该网络情感影响存在,用户更易受到自己所关注人情感影响,表达相似情感.与文献[18]相似,本文将用户情感相似性作为网络情感影响权重.下文使用随机游走模型寻找微博网络情感影响者.

5 EMOTIONRANK 模型

5.1 EMOTIONRANK 思想框架

为寻找微博网络情感影响者,建立有向异质微博网络 $G(V,E)$. V 是节点集合,包含两种节点(用户、微博), E 是边集合,包含三种关系(关注、转发、发布) .

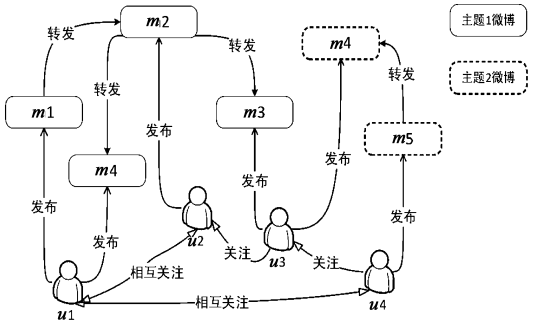


图2 拥有两个主题的微博网络

如图2,四个用户参与两个主题,若想寻找某一主题情感影响者,应关注参与该主题用户和与之有关的微博,其他节点则无用.显然 $u1$、$u2$、$u3$ 是主题 1 情感影响者候选人, $u3$、$u4$ 是主题 2 情感影响者候选人.对于网络整体情感影响者,所有用户都是候选人,应考虑全体用户情感行为.

用户有时会对非关注人某个微博主题感兴趣,然而大多时候用户只对自己所关注用户微博感兴趣.所以如果想找到主题情感影响者,‘转发’关系应该被考虑;如果想找到整体情感影响者,‘关注’关系则应作为主要被考虑对象来构建网络.

为寻找情感影响者,有必要对微博网络用户情感影响力排名.微博网络拥有多种节点和关系,所以需将异质微博网络转化为只有用户节点的网络,进而使用随机游走模型排名.基于上述思想,本文提出了一个全新模型(EmotionRank),图 3 为该模型框架流程.

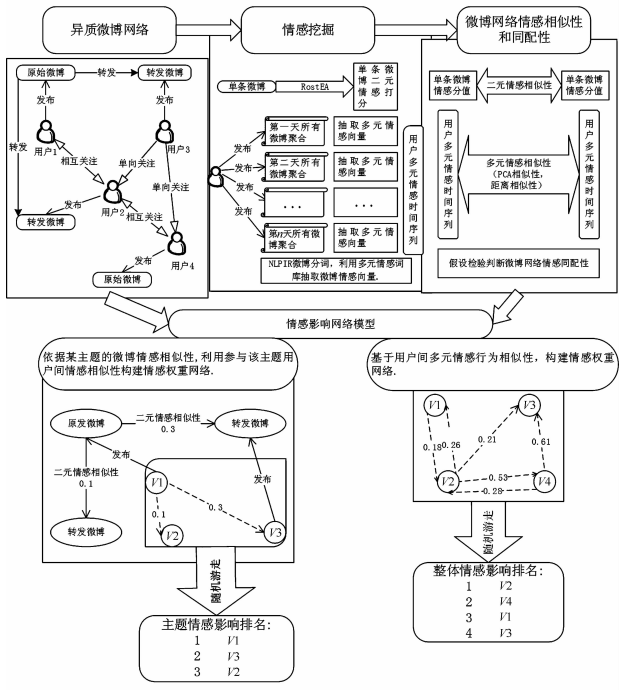


图3 EmotionRank 模型框架

5.2 主题情感影响者发现

为发现主题微博情感影响者,算法如下:

算法 1 EmotionRank – 发现主题情感影响者

输入: $G_{\text{topic}}(V_{\text{topic}}, E_{\text{topic}})$

输出: $\boldsymbol{R}_{\text{topic}}$

(1)Foreach edge_ij in E_{topic}

/* 将两用户间该主题所有转发微博相似性值加和作为情感影响权重. */

(2) $W_{ij-\text{topic}} = \sum S_{\text{m-}ij\text{-topic}}$

/* 利用权重值构造情感影响矩阵. */

(3) $\boldsymbol{H}_{\text{topic}} = \text{Matrix}(W_{ij-\text{topic}})$

/* 利用随机游走模型对用户排序 */

(4) $\boldsymbol{R}_{\text{topic}} = \text{RandomWalk}(\boldsymbol{H}_{\text{topic}})$

(5)End

执行算法 1 前,先使用关键字过滤技术寻找该主题的相关微博,再利用所构建异质微博网络的‘发布’关系,找到参与该主题的用户.将这些用户构建成有向网络 $G_{\text{topic}}(V_{\text{topic}}, E_{\text{topic}})$. V_{topic}是参与该主题的用户集合.

E_{topic}是以'转发'关系为基础构建的网络边集合,即该主题中具有'转发'行为的用户间才有一条有向边. $W_{ij-\text{topic}} = \sum S_{\text{m-}ij\text{-topic}}$是用户 i 指向 j 的有向边权重. $S_{\text{m-}ij\text{-topic}}$是用户 j 转发用户 i 微博之间的情感相似性. $S_{\text{m-}ij\text{-topic}}$的加和视为该主题下用户 i 对用户 j 的情感影响.

'转发'关系可保证用户参与同一主题,'关注'关系却不能.所以使用'转发'关系构建主题情感影响矩阵 $\boldsymbol{H}_{\text{topic}} = (W_{ij-\text{topic}})$,矩阵元素表示用户间主题情感影响.继而使用随机游走模型对用户影响力排序,方法如下:

$$\boldsymbol{R}_{\text{topic}}(k) = \widetilde{\boldsymbol{H}}^{\text{T}}_{\text{topic}} \cdot \boldsymbol{R}_{\text{topic}}(k-1) = (\widetilde{\boldsymbol{H}}^{\text{T}}_{\text{topic}})^k \cdot \boldsymbol{R}_{\text{topic}}(0)$$

$$\widetilde{\boldsymbol{H}}_{\text{topic}} = s\bar{\boldsymbol{H}}_{\text{topic}} + (1-s)\frac{1}{N_{\text{topic}}}\boldsymbol{e}\boldsymbol{e}^{\text{T}} \tag{5}$$

$\bar{\boldsymbol{H}}_{\text{topic}}$为 $\boldsymbol{H}_{\text{topic}}$标准化矩阵,$N_{\text{topic}}$为参与该主题用户数量.$\bar{\boldsymbol{H}}_{\text{topic}}$是非负矩阵,从而 $\widetilde{\boldsymbol{H}}_{\text{topic}}$是正定矩阵.根据矩阵论中 Perron-Frobenius 定理,当 $k \to \infty$ 时,$\boldsymbol{R}_{\text{topic}}(k)$收敛.

5.3　整体情感影响者发现

为发现微博网络整体情感影响者,算法如下:

算法 2　EmotionRank-发现整体情感影响者

输入:$G_{\text{overall}}(V_{\text{overall}}, E_{\text{overall}})$
输出:$\boldsymbol{R}_{\text{overall}}$
(1)Foreach edge_ij in E_{overall}
/* 计算关注用户间情感行为相似性,作为情感影响权重. */
(2) $W_{ij-\text{overall}} = \text{SF}_{ij} = \alpha S_{ij\text{-PCA}} + (1-\alpha) S_{ij\text{-dist}} (0 < \alpha < 1)$
/* 利用权重值构造情感影响矩阵 */
(3) $\boldsymbol{H}_{\text{overall}} = \text{Matrix}(W_{ij-\text{overall}})$
/* 利用随机游走模型对用户排序 */
(4) $\boldsymbol{R}_{\text{overall}} = \text{RandomWalk}(\boldsymbol{H}_{\text{overall}})$
(5)End

算法 2 先构建有向图 $G_{\text{overall}}(V_{\text{overall}}, E_{\text{overall}})$. V_{overall}为所有用户节点集合. E_{overall}是以'关注'关系为基础构成的有向边集合.用户 i 指向 j 的有向边权重为用户间情感行为相似性,即 $W_{ij-\text{overall}} = \text{SF}_{ij} = \alpha S_{ij\text{-PCA}} + (1-\alpha) S_{ij\text{-dist}} (0 < \alpha < 1)$,可由式(4)计算.它表示用户 i 对用户 j 整体情感影响.

'关注'关系相对稳定,不像'转发'关系那样频繁变化.SF_{ij}考虑了一段时间内用户情感波动和情感表达强度.所以'关注'关系被用来寻找整体情感影响者.

整体用户情感影响矩阵 $\boldsymbol{H}_{\text{overall}} = (W_{ij-\text{overall}})$的元素表示用户间整体情感影响.使用随机游走模型对所有用户情感影响力排序如下:

$$\boldsymbol{R}_{\text{overall}}(k) = \widetilde{\boldsymbol{H}}^{\text{T}}_{\text{overall}} \cdot \boldsymbol{R}_{\text{overall}}(k-1) = (\widetilde{\boldsymbol{H}}^{\text{T}}{}_{\text{overall}})^k \cdot \boldsymbol{R}_{\text{overall}}(0)$$

$$\widetilde{\boldsymbol{H}}_{\text{overall}} = s\bar{\boldsymbol{H}}_{\text{overall}} + (1-s)\frac{1}{N_{\text{overall}}}\boldsymbol{e}\boldsymbol{e}^{\text{T}} \tag{6}$$

$\bar{\boldsymbol{H}}_{\text{overall}}$为 $\boldsymbol{H}_{\text{overall}}$标准化矩阵,$N_{\text{overall}}$为全体用户数量.$k \to \infty$ 时,$\boldsymbol{R}_{\text{overall}}(k)$显然收敛.

5.4　EmotionRank 的优点

EmotionRank 模型有以下优点:(1) 异质微博网络可更快速、直接找到参与某主题用户,提升了查询效率.(2)传统发现整体影响者方法往往依赖主题影响者,可能产生偏差.模型摆脱了对主题的依赖.(3)模型发现的情感影响者与时间相关,结果更合理.

6　实验评估

6.1　主题情感影响者实验

本节选用两个中国社会热点话题实验,成千上万微博用户关注并参与到这两个话题,并表达情感.主题 1 是 2013 年湖南卫视热门节目"爸爸去哪儿了".节目中 5 个明星父亲带着 5 个孩子完成各种任务.主题 2 是社会热点"陈永洲事件".新快报记者陈永洲 2013 年 10 月被捕,许多媒体和法律界人士通过新浪微博表达关注.

利用 EmotionRank(ER)模型,PageRank(PR)和 InDegree (InD)排序,寻找到两个主题情感影响者.两主题前 5 名情感影响者如表 1 所示.

主题 1 结果中,EmotionRank 找到的前 5 名用户均与该节目有关.张亮、郭涛、林志颖是节目的直接参与者.湖南卫视芒果捞吸引了 70 万湖南卫视忠实观众.影视明星叶一茜虽没参与该节目,但她是节目参与者田亮的妻子.所以 EmotionRank 针对该主题寻找到的情感影响者结果是合理可靠的.

主题 2 是法律界和传媒界的热点,EmotionRank 找到的前 5 名情感影响者均为媒体或法律界人士.丁来峰是国际公关杂志常务副主编,袁裕来和徐昕是知名律师.头条新闻和新浪财经是知名媒体.虽然他们专注不同领域,但却较早关注该主题并表现了强大的情感影响力.

表 1　前 5 主题情感影响者

排名	爸爸去哪儿了			陈永洲事件		
	ER	PR	InD	ER	PR	InD
1	张亮	思想聚焦	何炅	丁来峰	南都周刊	头条新闻
2	郭涛	申音	林志颖	头条新闻	华尔街日报中文网	陈里
3	林志颖	新浪娱乐	林俊杰	袁裕来	五岳散人	人民日报
4	湖南卫视芒果捞	投资界微博	潘石屹	徐昕	新浪财经	央视新闻
5	叶一茜	传媒老王	任志强	新浪财经	21 世纪经济报道	新浪财经

其他方法(PageRank, InDegree)从网络结构角度对网络用户影响力排名,找到的影响者并没直接参与目标主题.这些用户往往是目标主题的转发者和评论者,并不是最有影响力的用户.

6.2 发现整体情感影响者实验

用户情感行为相似性(SF)作为网络权重来构建网络.SF 的参数 α 变化可导致结果不同.以 0.1 为步长,改变 α,得出前 20,50,100 情感影响者列表.当 α 从 0.5 变到 0.6 时,两相邻 α 列表有最多共同用户,几乎有相同排名,如图 4.这表明此时结果最稳定.$\alpha=0.6$ 也较好结合用户的情感波动和强度,所以在此场景下 $\alpha=0.6$ 是合理参数.

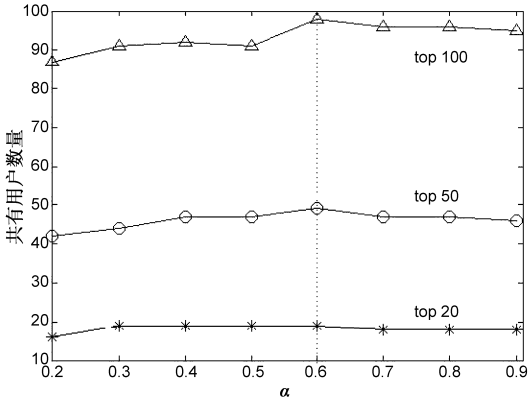


图4 共有用户数量变化

表 2 展示了前 10 位整体情感影响者.应注意这是实验用户群体中从 2013/10/22 到 2013/11/9 时间段内情感影响力排名.其中用户可分为两种,一种是个体用户(柳岩、李开复、张泉灵、黄健翔、文章、徐小平、石述思),另一种是媒体用户(当时我就震惊了、思想聚焦、Vista 看天下).个体用户发布微博较少,却有着极大吸引力并产生巨大情感影响.媒体用户主要依赖发布更多信息吸引眼球.

表 2 10 大整体情感影响者

用户	发帖量	粉丝量	平均转发量
柳岩	19	10401309	1148.95
李开复	2	51779415	12485.50
张泉灵	11	8466331	3374.45
当时我就震惊了	739	4791121	1900.18
思想聚焦	728	4768110	2050.71
黄健翔	25	16516720	1119.72
文章	3	42819347	12854.67
Vista 看天下	333	3610868	338.88
徐小平	13	10761400	658.23
石述思	71	1348651	613.338

下面用文献[2]设计的网络推荐任务验证 EmotionRank 模型优越性.任务描述如下:

(1)随机选出 $|L|$ 个具有'关注'关系的网络边;

(2)每一个 $l\in L$ 执行

(3)令 s_0 和 s_f 为 l 的关注者和被关注者;

(4)随机选择 10 个 s_0 不关注的用户,该集合为 S_t;

(5)移除关系 l 产生一个新网络,新网络中 s_0 不关注 s_f;

(6)应用不同算法度量 s_f 和 S_t 中用户网络影响力,基于此结果推荐 s_0 是否关注 s_f;

(7)对比不同算法推荐质量;

(8)结束.

若某种算法结果 s_f 排名高于 S_t 中用户,则认为该算法推荐效果较好.算法推荐质量由 S_t 中排名高于 s_f 用户数量来评估.

L 四个选择标准与文献[2]类似.

(a)基于 s_f 粉丝数量产生两组 L.一为 L_{fh},其中 s_f 有较多粉丝量,另一为 L_{fl},s_f 有较少粉丝量.

(b) 基于 s_f 的发帖量产生两组 L.一为 L_{ph},s_f 有较高发帖量,另一为 L_{pl},s_f 有较低发帖量.

(c)基于 s_0 与 s_f 间的情感行为相似性产生两组 L.一为 L_{sh},s_0 与 s_f 有较高情感行为相似性,另一为 L_{sl},s_0 与 s_f 有较低情感行为相似性.

(d)基于 s_0 与 s_f 间关注关系产生两组 L.一为 L_{rr},s_0 与 s_f 相互关注,从拥有相互关注关系的边随机选择.另一为 L_{ur},从非相互关注关系用户中随机选择.

4 种方法(EmotionRank, PageRank, InDegree, SocialInfluence)对 8 组数据完成推荐任务.PageRank 和 InDegree 从结构角度发现影响力用户,SocialInfluence 依据文献[18]的社会情感影响作为情感权重,构建情感权重网络,发现情感影响者.其情感权重定义为两用户对待第三方微博情感态度相似性.定义如下:

$$w(v_i,v_j)=\frac{\left|\{b\mid y_i(b)\cdot y_j(b)>0,b\in B_H(i),b\in B_H(j)\}\right|}{\left|\{b\mid b\in B_H(i),b\in B_H(j)\}\right|} \tag{7}$$

其中 $B_H(i)$是用户 v_i 所评论的微博集合,而 $y_i(b)$则是用户 v_i 对微博 b 的评论情感值.

由于本文数据未采集微博评论数据集,于是采用转发数据集代替评论数据集计算用户间的社会情感相似性.转发行为表明原用户与转发用户有共同的关注点,转发过程可判断转发微博与原微博的情感态度变化情况,所以此种代替合理.

共进行 4 次试验,图 5 为 4 种方法完成推荐任务的平均效果.除 L_{fh}之外,EmotionRank 在其他 7 个场景均最好完成任务.数据集 L_{fh}完成推荐实验时,该数据已倾向选择有较多追随者的 s_f,使得 InDegree 方法表现最好.

下面用另一指标 EVI 评估 EmotionRank 优越性.该

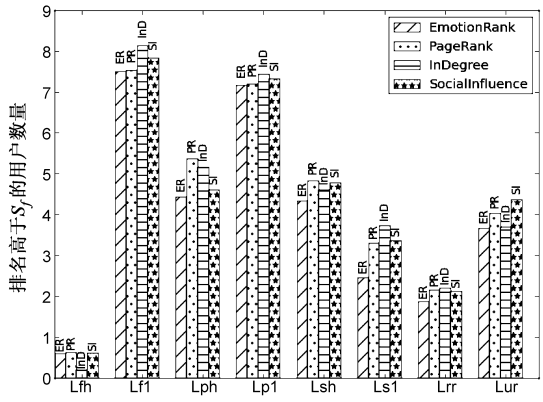


图5　推荐任务表现算法对比

指标表明某种方法的认可度.以 EmotionRank 为例,利用 4 种方法(EmotionRank, PageRank, InDegree, PostNumber)查询 top-K 影响力用户,分别为 I_{ER}, I_{PR}, I_{InD}, I_{PoN}.

$$EVI_{ER}=\frac{\frac{|I_{ER}\cap I_{PR}|}{K}+\frac{|I_{ER}\cap I_{InD}|}{K}+\frac{|I_{ER}\cap I_{PoN}|}{K}}{3} \tag{10}$$

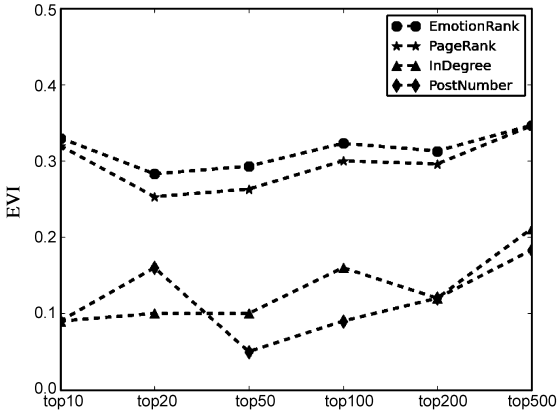


图6　影响力用户被接受程度

图 6 表明 EmotionRank 寻找的情感影响者比其他三种方法更被认可.

7　结束语

本文使用真实微博数据,以微博网络情感同配性为基础,验证了所构建微博网络情感影响的存在.进而提出一个全新模型寻找微博网络情感影响者.模型利用微博网络'转发'和'关注'等关系,将异质微博网络转化为只包含用户节点的同质网络,通过随机游走模型成功发现情感影响者.实验结果验证了模型有效性和优越性.

本文工作存在以下不足:(1)由于数据缺乏,模型未考虑'评论'关系,而用户微博收到的'评论'可体现该用户对其他用户的影响.(2)数据规模并不大,而微博网络每天产生海量数据.

以后工作中,'评论'关系将用以改进模型,并考虑用并行图挖掘算法提升实验数据规模.

参考文献

[1] Java A, Song X, et al. Why we twitter: understanding microblogging usage and communities[A]. Proceedings of the Ninth WebKDD and First SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis[C]. San Jose: ACM, 2007.56 – 65.

[2] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitters[A]. Proceedings of the Third ACM International Conference on Web Search and Data Mining[C]. New York: ACM, 2010.261 – 270.

[3] Budalakoti S, et al. Bimodal invitation-navigation fair bets model for authority identification in a social network[A]. Proceedings of the Twenty-First International Conference on World Wide Web[C]. Lyon: ACM, 2012.709 – 718.

[4] Sun B, Ng V T. Identifying Influential Users by Their Postings in Social Networks[M]. Berlin Heidelberg: Springer, 2013.128 – 151.

[5] Pal A, Counts S. Identifying topical authorities in microblogs [A]. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining[C]. Hong Kong: ACM, 2011.45 – 54.

[6] 张伟哲,王佰玲,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 W Z, Wang B L, et al. Public opinion leader community mining based on the heterogeneous network[J]. Acta Electronica Sinica, 2012, 40(10): 1927 – 1932. (in Chinese)

[7] Bakshy E, Hofman J M, et al. Everyone's an influencer: quantifying influence on twitter[A]. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining[C]. Hong Kong: ACM, 2011.65 – 74.

[8] Cha M, Haddadi H, et al. Measuring user influence in twitter: the million follower fallacy[A]. Proceedings of the Fourth International Conference on Weblogs and Social Media[C]. Washington, DC: AAAI, 2010.10 – 17.

[9] Quercia D, Ellis J, et al. In the mood for being influential on twitter[A]. Proceedings of the Third IEEE International Conference on Information Privacy, Security, Risk and Trust[C]. Boston: IEEE, 2011.307 – 314.

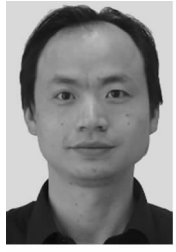
[10] Shen Y, et al. Emotion mining research on micro-blog[A]. Proceedings of the First IEEE Symposium on Web Society [C]. Lanzhou: IEEE, 2009.71 – 75.

[11] Bollen J, et al. Modeling public mood and emotion: twitter sentiment and socio-economic phenomena[A]. Proceedings of Nineteenth International World Wide Web Conference[C].

Raleigh, North Carolina: ACM, 2010.1 – 10.
[12] Zhu J, Wang B, Wu B. Social network users clustering based on multivariate time series of emotional behavior[J]. The Journal of China Universities of Posts and Telecommunications, 2014, 21(2): 21 – 31.
[13] Wang X, Jia J, Tang J, et al. Modeling emotion influence in image social networks[J]. IEEE Transactions on Affective Computing, 2015, 63): 1 – 13.
[14] Tang J, Zhang Y, Sun J, et al. Quantitative study of individual emotional states in social networks[J]. IEEE Transactions on Affective Computing, 2012, 3(2): 132 – 144.
[15] Wu B, Jia J, et al. Inferring emotions from social images leveraging influence analysis[A]. The Third National Conference of Social Media Processing[C]. Beijing: Springer, 2014. 141 – 154.
[16] Crandall D, Cosley D, et al. Feedback effects between similarity and social influence in online communities[A]. Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Lasvegas: ACM, 2008. 160 – 168.
[17] Fond T La, Neville J. Randomization tests for distinguishing social influence and homophily effects[A]. Proceedings of the Nineteenth International Conference on World Wide Web [C]. Raleigh, North Carolina: ACM, 2010. 601 – 610.
[18] Yang Y, Cui P, et al. User interest and social influence based emotion prediction for individuals[A]. Proceedings of the Twenty-First ACM International Conference on Multimedia [C]. Barcelona: ACM, 2013. 785 – 788.
[19] Bollen J, Gonçalves B, et al. Happiness is assortative in online social networks[J]. International Society for Artificial Life, 2011, 17(3): 237 – 251.
[20] Kim S, et al. Do you feel what I feel? Social aspects of emotions in twitter conversations[A]. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media [C]. Dublin: AAAI, 2012. 495 – 498.
[21] Kramer A D I, Guillory J E, et al. Experimental evidence of massive-scale emotional contagion through social networks [J]. Proceedings of the National Academy of Sciences, 2014, 111(24): 8788 – 8790.

作者简介

朱 江 男, 1982 年生于吉林. 博士研究生. 研究方向为社会网络分析, 复杂网络算法. E-mail: zhujiang@bupt.edu.cn



吴 斌 男, 1969 年出生. 教授, 博士生导师. 研究方向为图数据挖掘、智能信息处理.

王 柏 女, 1962 年出生. 教授, 博士生导师. 研究方向为数据挖掘、智能信息处理.

李小明 男, 1989 年出生. 硕士研究生. 研究方向为复杂网络算法.