# 基于概率生成模型的微博话题传播群体划分方法

陈 静 刘 琰 王煦中
(数学工程与先进计算国家重点实验室 郑州 450001)

**摘 要** 事件以话题形式在微博中迅速传播,并能够产生巨大的影响力。因此,对参与话题传播过程的用户进行分析以及发现具有不同主题兴趣情感倾向性的群体受到政府和企业的广泛关注。现阶段,绝大多数应用到微博的群体发现算法都是从单个用户出发,仅考虑了用户社会联系,与用户共享内容相隔离,其群体发现的结果不具有语义信息。少数算法综合了用户社会联系与内容,却忽略了微博本身的结构特性。因此从微博话题的角度出发,综合考虑话题传播过程中的用户交互、微博文本内容以及情感极性,同时结合用户的行为信息,提出了一个基于概率生成模型的微博话题传播群体划分方法BP-STG。采用吉布斯抽样对模型进行推导,不仅能够挖掘出具有不同主题倾向性的群体,同时还能够挖掘出群体的情感倾向分布以及用户在群体中的活跃度及其行为表现。此外,模型还能够推广到许多带有社交网络性质的媒体中。在获取的新浪微博两个话题数据集上的实验表明,BP-STG模型不仅能够有效地对微博话题传播群体进行划分,而且能够发现群体内部活跃用户以及用户在群体中的行为模式。

**关键词** 微博话题,概率生成模型,群体划分,情感元素,行为模式
**中图法分类号** TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.8.045

## Group Partition in Topic-related Microblogging Spreading Based on Probability Generation Model

CHEN Jing LIU Yan WANG Xu-zhong
(State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China)

**Abstract** Event can spread rapidly in the form of topic microblog and make enormous influence. Therefore, the analysis for the users and discovering groups with different interesting and sentiments in the topic discussion obtain the concern of the government and enterprises. The generated content and relationship between the users are often separated in the current methods on community detection, which have no semantic information. Though some methods have combined the two factors, they fail to take account of the behavior information and sentiment information which exist in microblog, and they are not well to mine the groups in the microblog topic discussion. We proposed a group partition model called BP-STG which takes the text information, social contacts, text sentiment information and the users' behavior into consideration. We presented a Gibbs sampling implementation for inference of our model, mining only different interest groups, but also the sentiment distribution and participants' activeness and behavior information in a group. Besides, our model can be extended to many texts associated with a group of people such as E-mails and forum posts. Experimental results on actual dataset show that BP-STG model can offer an effective solution to group partition in topic-related microblogging spreading and provide more meaningful semantic information than the state-of-the-art model.

**Keywords** Microblogging topic, Probability generation model, Groups partition, Sentiment information, Behavior pattern

近年来,微博凭借着快速、便捷等特性成为了网民获取新闻时事、自我表达以及社会参与的重要媒介,同时也成为了社会公共舆论、企业品牌和产品推广的重要平台。微博中的话题更是成为了公众参与和获取社会焦点事件、综艺娱乐节目推广等的主要手段之一。

微博话题传播更是一把双刃剑:一方面,微博为一些社会事件中的信息公开提供了一个快速响应的平台,它在一定程度上弥补了传统媒体和其他网络工具的不足。用户可以通过#话题名#创建或者参与到特定话题的讨论中。例如,在两会期间,由人民日报创建起来的#2015两会#、由央视新闻创建的#微博看两会#和由用户为微博新鲜事创建的#两会#,成为了3个热门话题,以微博看两会为例,其中有2.9万微博账户关注了该话题,约有65万人参与了该话题的讨论。另一方面,微博不同于传统新闻媒体,其新闻的发布存在重复性,且真实性无法保证,可能会被利用成为谣言传播的载体、不满情绪的导火索,甚至给国家安全和社会稳定带来极坏的后果。

到稿日期:2015-07-23 返修日期:2016-01-04 本文受国家自然科学基金(61309007),国家863计划(2012AA012902)资助。
陈 静(1990—),女,硕士生,主要研究方向为社会网络分析、数据挖掘,E-mail:1923542221@qq.com;刘 琰(1979—),女,博士,副教授,主要研究方向为网络信息安全、网络数据分析、网络态势感知;王煦中(1991—),男,硕士,主要研究方向为数据挖掘、社会网络分析。

当新的热点事件出现后，政府部门需要及时掌握热点事件的主题、参与者以及舆论的情感倾向，以便对后期的舆论引导具备快速响应能力；同时为了防止团伙利用话题进行虚假信息扩散或者负面不良情绪煽动，政府部门需要掌握话题传播中具有不同主题情感倾向性的群体。

解决上述问题，需要用到社团发现的算法。目前社团发现的方法有很多，最早的是由 Girvan 和 Newman[1] 提出的一种基于中间度概念的社团发现算法。由于社会网络的发展，社团结构开始出现彼此包含的关系，一些重叠社团发现算法被提出，例如 Palla 等人[2] 提出了面向重叠社区的发现算法 CPM，该方法认为社团是一个强联通图，这使得该算法在稀疏网络中的效果较差。这些社团发现方法研究的出发点是社会网络中的节点关系属性。该类研究忽略了社会网络中存在的语义信息。2003 年，Blei[3] 提出了 LDA 模型，认为文档是多个主题的概率分布。Steyver 等人[4] 认为主题是多个关键词的概率分布，用户也以某种概率分布对多个主题感兴趣，并提出 AT 模型，发现了用户、文档、主题和关键词之间的关系。Zhou 等人[5] 在 AT 模型中加入了 user 分布取样，提出了 CUT 模型。但是，以上方法只考虑到了文本内容，忽略了用户社会联系的重要性。为此，不少学者提出结合语义信息和社会联系的社团发现方法[6-10]，如 Zhou 等人[9] 提出了一个社团整合模型(COCOMP)，该模型能够识别出与给定用户相关的社团以及社团相关的主题和人群，但是该模型发现的社团只具有主题倾向性，并没有考虑情感信息以及用户的行为模式；Yang 等人[10] 在 Zhou 的模型的基础上增加了情感元素，该模型所划分出来的社团带有情感色彩，但是模型计算过于复杂。

由于微博中含有大量用户与微博文本的交互信息，参与用户具有不同的行为表现，直接应用现有的社团发现方法解决微博话题中传播人群的划分并不合适，需要对微博的特有性质进行进一步考虑并优化算法。

因此，本文针对现有社团发现算法的不足，在研究简单 LDA 模型的基础之上，首先采用基于情感词典的方法对话题集中的微博文本做情感倾向性分析，即给每个微博文本添加情感标签，然后结合微博文本信息、用户在话题传播中的交互以及用户的行为模式，将文本的主题情感元素考虑在内，提出了一个群体划分模型 BP-STG(Sentiment Topic model combine with participants' Behavior Pattern for Group partition)。

## 1 基本定义和假设

### 1.1 基本定义

本节给出与微博话题相关的概念，并对群体的概念进行形式化描述。

**定义 1**(话题集) 话题是指包括一个核心事件或活动，以及所有与之直接相关的事件或活动[11]。一个特定的微博话题通常由多条带有 # 话题名 # 的微博构成。因此，一个话题 $T$ 可以表示成 $T=\{M_1,M_2,\cdots,M_D\}$，其中 $M_D$ 表示带有 # 话题名 # 的一篇微博，$D$ 表示微博总数。

**定义 2**(话题中的微博) 在本模型中，微博由 4 部分组成：微博内容 $t$，用户集合 $u$，微博的类型 $h$(微博分为 3 种类型：原创、转发和评论)以及文本的情感极性 $l$。因此，一篇微博可以表示为 $M=\{h,t,u,l\}$，其中 $u=\{u_p,u_r,u_c\}$。$u_p$ 表示微博的原创账户集，$u_c$ 表示微博的评论账户集合，$u_r$ 表示微博的转发账户集合。

**定义 3**(主题) 话题倾向于描述一个真实存在的具体事件，与话题相比，主题更倾向于描述某一类事件。例如：两会是一个大的话题，在两会话题下，有关于经济、政治、法律、环境等的讨论，这些经济、政治或法律就表示话题中隐含的主题。

**定义 4**(群体) 具有直接或者间接联系同时具有相同主题倾向性的用户团体。

事实上，一些群体中存在这样的子团体，它们在群体中有相同的主题兴趣，但是它们之间并没有真实的联系，这种现象在微博话题中普遍存在。本文即认为这些子群之间存在着虚拟的边。如图 1 所示，用点表示微博账号，边表示微博账号之间的关系，其中实线代表微博账号存在至少一次微博文本共享，虚线代表微博账号之间虽没有存在实际的文本交互，但是却具有相同的主题倾向性，图中存在 3 个群体，有 4 个重叠节点，分别为 7，8，9，10。节点 3 与 5 之间用虚线表示，代表着两个微博账号之间有相同的主题兴趣。

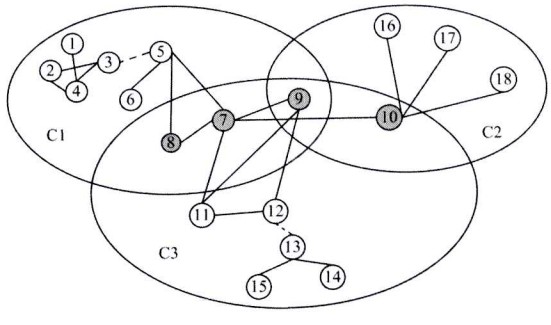



图 1 群体示例图

### 1.2 基本假设

采用 BP-STG 模型进行话题传播群体划分基于如下假设。

假设：话题集中的一条微博是某个群体在话题传播过程中的一次会话，每个参与会话的成员的情感倾向由微博文本内容的情感倾向体现。同时群体中的成员参与话题讨论的行为具有偏好性，不同用户在行为模式上存在一定的差异。

本文的主要任务就是在给定特定话题集 $T$ 后，利用 BP-STG 模型挖掘出在话题传播过程中具有不同主题倾向性的群体以及群体的情感分布、活跃用户以及用户的行为模式。

## 2 基于概率生成模型的微博话题传播群体划分方法

### 2.1 LDA 主题概率生成模型

LDA 主题概率生成模型的基本思想是每一篇文档可以表示成一系列主题的混合分布；同时每个主题是词汇表中所有单词上的概率分布。

LDA 生成文本的方式可以用图 2 中的贝叶斯网络图表示。最开始，LDA 从参数为 $\beta$ 的狄利克雷分布抽取主题与单词的关系 $\varphi$。LDA 生成一个文本时，首先从参数为 $\alpha$ 的狄利克雷分布中抽样出该文本 $d$ 与各个主题之间的关系 $\theta_m$，当有

$T$ 个主题时，$\theta_m$ 是一个 $T$ 维向量，每个元素代表主题在文本中出现的概率，满足 $\sum_K \theta_{m_k}=1$；接着，从参数为 $\theta_m$ 的多项式分布中抽取出当前单词所属的主题 $z_{m,n}$；最后从参数为 $\varphi_{z_{m,n}}$ 的多项式分布中抽取具体单词 $w_{m,n}$。一个文本中所有单词与其所属主题的联合概率分布如式(1)所示：

$$p(w,z|\alpha,\beta)=p(w|z,\beta)p(z|\alpha) \tag{1}$$

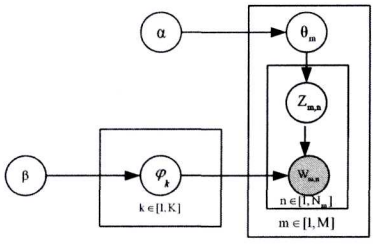



图 2　LDA 的贝叶斯网络图

在 LDA 中，文本的单词是可观测到的数据，而文本的主题是隐式变量。根据文本的生成规则和已知数据，LDA 通过概率推导可以求得文本的主题分布。常用的推导方法有变分贝叶斯、吉布斯抽样、期望传播等。

### 2.2　BP-STG 群体划分模型描述

BP-STG 模型是在 LDA 概率生成模型的基础之上进行拓展的。在该方法中，隐含群体的分配是由主题兴趣分布、参与人员交互以及微博文本的情感极性共同决定的。模型的基本思想是将话题集 $T$ 中的一条微博看作是某个群体的一次会话，因此该微博的主题应从其所属的群体的主题分布中抽取，相关参与用户应当是该群体中相对比较活跃的用户，情感倾向应当从群体的情感倾向分布中抽取。模型的贝叶斯网络如图 3 所示。

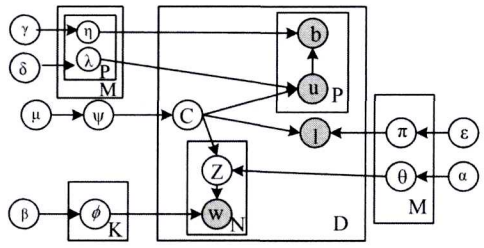



图 3　BP-STG 模型的贝叶斯网络

图模型符号含义：方框表示重复抽样，方框内的数字表示重复次数，灰色节点表示可观测变量，白色节点表示隐含变量或参数，箭头表示条件依赖关系。

(1)模型中的变量

在模型中主要有两种变量，即可观察到的变量和隐含变量。其中隐含变量为群体 $C(1,2,\cdots,M)$ 和话题 $Z(1,2,\cdots,K)$；可观察到的变量为文档中出现的单词 $W$、共享这篇文档的用户集合 $u$、用户与文档的交互方式 $b$、文档的情感极性 $l$。

(2)模型中的超参数

图 3 中，$\gamma$ 是每个群体中成员的行为的多项式分布的狄利克雷超参数；$\delta$ 是每个群体中成员活跃度的多项式分布的狄利克雷超参数；$\mu$ 是群体在话题传播过程中活跃度的多项式分布的狄利克雷超参数；$\alpha$ 是每个群体中主题的多项式分布的狄利克雷超参数；$\beta$ 是每个主题下词的多项式分布的狄利克雷超参数；$\varepsilon$ 是每个群体中情感极性的多项式分布的狄利克雷超参数。

(3)参数

假设有 $K$ 个隐含主题。$Dir(\cdot)$ 表示狄利克雷分布。其中 $\phi_k$ 表示词汇表中分配给主题 $k$ 的单词的概率，服从 $Dir(\beta)$ $(k=1,2,\cdots,K)$。

假设有 $M$ 个群体和 $S$ 个情感极性，每个群体与 4 个参数有关：主题矩阵 $\theta$、用户的参与矩阵 $\lambda$、情感矩阵 $\pi$、用户在群体中行为矩阵 $\eta$。在群体 $m(m=1,2,\cdots,M)$ 中：

- $\theta_m$ 表示群体 $m$ 的主题分布，代表不同主题在群体 $m$ 中的权重，其中 $\theta_m|\alpha$ 服从参数为 $\alpha$ 的狄利克雷分布，可表示为 $\theta_m|\alpha\sim Dir(\alpha)$。
- $\lambda_m$ 表示不同用户在群体 $m$ 中的活跃度矩阵，其中 $\lambda_m|\delta$ 服从参数为 $\delta$ 的狄利克雷分布，可表示为 $\lambda_m|\delta\sim Dir(\delta)$。
- $\pi_m$ 表示不同情感极性在群体 $m$ 中的权重，其中 $\pi_m|\varepsilon$ 服从参数为 $\varepsilon$ 的狄利克雷分布，可表示为 $\pi_m|\varepsilon\sim Dir(\varepsilon)$。
- $\eta_{m,p}$ 表示群体 $m$ 中成员 $p$ 的行为分布，其中 $\eta_{m,p}|\gamma$ 服从参数为 $\gamma$ 的狄利克雷分布，可表示为 $\eta_{m,p}|\gamma\sim Dir(\gamma)$。
- $\psi$ 表示在整个话题传播中群体的活跃度，其中 $\psi|\mu$ 服从参数为 $\mu$ 的狄利克雷分布，可表示为 $\psi|\mu\sim Dir(\mu)$。

### 2.3　模型的生成过程

整个微博话题集 $T$ 中的每一篇微博 $M_i(i=1,2,\cdots,D)$ 的生成过程如下($Mult(\cdot)$ 表示多项式分布)。

(1)为微博 $M_i$ 选择一个群体 $c_{M_i}$：$c_{M_i}|\psi\sim Mult(\psi)$。

(2)假定有 $P_{M_i}$ 个用户共享微博 $M_i$，设 $p=1,2,\cdots,P_{M_i}$，对于每一个与微博 $M_i$ 交互的用户 $u_{M_i,p}$，生成过程如下：

a)从群体 $c_{M_i}$ 的用户活跃度矩阵中抽取用户 $u_{M_i,p}$：$u_{M_i,p}|\lambda,c_{M_i}\sim Mult(\lambda_{c_{M_i}})$；

b)选择用户 $u_{M_i,p}$ 的行为模式 $b_{u_{M_i,p}}$：$b_{u_{M_i,p}}|\eta,u_{M_i,p},c_{M_i}\sim Mult(\eta_{c_{M_i},u_{M_i,p}})$。

(3)假定一篇微博的文本内容包含 $N_{M_i}$ 个词汇，对于每个词汇 $w_{M_i,n}(n=1,2,\cdots,N_{M_i})$，生成过程如下：

a)从第 $c_{M_i}$ 个群体的主题矩阵中抽取一个主题 $z_{M_i,n}$：$z_{M_i,n}|\theta,c_{M_i}\sim Mult(\theta_{c_{M_i}})$；

b)根据抽取出的主题 $z_{M_i,n}$，从主题单词分布中抽取单词 $w_{M_i,n}$：$w_{M_i,n}|z_{M_i,n},\phi\sim Mult(\phi_{z_{M_i,n}})$。

(4)从 $c_{M_i}$ 群体的情感矩阵中抽取一个情感极性：$l_{M_i}|\pi,c_{M_i}\sim Mult(\pi_{c_{M_i}})$。

一条微博中，其所属群体 $c$、所有参与人员 $u$、所有单词 $w$ 与其所属主题 $z$ 及主题的情感倾向 $l$ 的联合概率分布如式(2)所示。

$$\begin{aligned}&p(u,c,z,l,w|\delta,\mu,\beta,\alpha,\varepsilon)\\&=p(u|c,\lambda)p(c|\psi)p(z|c,\theta)p(l|c,\pi)p(w|z,\phi)p(\lambda|\delta)\\&\quad p(\psi|\mu)p(\theta|\alpha)p(\pi|\gamma)p(\phi|\beta)\end{aligned} \tag{2}$$

### 2.4　模型的推导和参数估算

BP-STG 模型的训练与推导采用吉布斯抽样的方法。吉布斯抽样是一种高效的 MCMC(Markov chain Monte Carlo)抽样方法，通过迭代抽样的方式对复杂的概率分布进行推导，多用于贝叶斯图模型的求解。在这个模型中，一篇微博只能被分配给一个群体。为了方便模型的推导，表 1 列出模型推导公式中涉及的参数。

表 1 符号定义说明

| 参数 | 定义 |
|---|---|
| D | 话题集 T 中的微博总数 |
| S | 情感极性的个数 |
| K | 隐含主题的数目 |
| V | 单词的总数 |
| P | 参与话题讨论的总的人数 |
| $D_m(D_m^{-d})$ | 表示被分配群体 m 的总的微博数(除了微博 d) |
| $n_{m,k}^{-d}$ | 出现在群体 m 的文档中被分配主题 k 的单词总数 |
| $z_d$ | 表示微博 d 的主题集合 |
| $N_{d,k}$ | 微博 d 中分配给主题 k 的单词总数 |
| $D_{m,s}^{-d}$ | 群体 m 中情感极性为 s 的总的微博个数 |
| $l_d$ | 微博 d 的情感极性的集合 |
| $h_{m,p}(h_{m,p}^{-d})$ | 表述 p 参与群体 m 讨论的次数(除了参与微博 d 的讨论次数) |
| $u_d$ | 参与微博 d 讨论的用户集合 |
| $e_d$ | 参与微博 d 讨论的总的用户数 |
| t | 微博 d 中的第 i 个单词 |
| $z_t$ | 微博 d 中第 i 个单词被分配的主题 |
| $n_{k,v}^{-t}$ | 单词 v 分配给主题 K 的次数(除了微博 d 的第 i 个单词) |

对于每一篇微博 $d$,分配给它的群体的后验条件概率为:

$$p(c_d=m|c_{-d},u,z,l,w)$$
$$=\frac{p(c,u,z,l,w|\mu,\delta,\beta,\epsilon,\alpha)}{P(c_{-d},u,z,l,w|\mu,\delta,\beta,\epsilon,\alpha)}\propto\frac{D_m^{-d}+\mu_m}{\sum_{j=1}^{M}\mu_j+D-1}\times$$
$$\frac{\prod_{k\in z_d}\prod_{i=0}^{N_{d,k}}(\alpha_k+n_{m,k}^{-d}+i)}{\prod_{i=0}^{N_d-1}(\sum_{k=1}^{K}\alpha_k+n_{m,k}^{-d}+i)}\times\frac{\prod_{s\in l_d}(D_{m,s}^{-d}+\epsilon_s)}{\sum_{s=1}^{S}\epsilon_s+D_m-1}\times$$
$$\frac{\prod_{p\in u_d}(\delta_p+h_{m,p}^{-d})}{\prod_{i=0}^{e_d-1}(\sum_{p=0}^{P}\delta_p+h_{m,p}^{-d}+i)} \tag{3}$$

假设微博 $d$ 被分配给群体 $c_d$,则微博 $d$ 中第 $i$ 个单词隐含主题 $z_{d,i}$ 的条件后验概率计算公式如下:

$$p(z_t=j|w,z_{-t},c_d)$$
$$=\frac{p(z_t,w,c_d)}{p(z_{-t},w,c_d)}\propto\frac{n_{c_d,k}^{-t}+\alpha_k}{\sum_{k=1}^{K}n_{c_d,k}^{-t}+\alpha_k}\times\frac{n_{k,v}^{-t}+\beta_v}{\sum_{v=1}^{V}n_{k,v}^{-t}+\beta_v} \tag{4}$$

对式(3)和式(4)反复迭代,并首先对所有群体进行抽样,然后根据群体对所有主题进行抽样,最后达到抽样结果稳定。由于群体、参与人员及其行为模式、主题、单词以及微博情感倾向性都满足多项式分布,因此 $\psi_m$,$\lambda_{m,p}$,$\eta_{m,p,b}$,$\theta_{m,k}$,$\pi_{m,s}$ 和 $\phi_{k,v}$ 的迭代结果如下:

$$\psi_m=\frac{D_m+\mu_m}{\sum_{m=1}^{M}\mu_m+D} \tag{5}$$

$$\lambda_{m,p}=\frac{h_{m,p}+\delta_p}{\sum_{p=1}^{P}h_{m,p}+\delta_p} \tag{6}$$

$$\eta_{m,p,b}=\frac{h_{m,p,b}+\gamma_b}{\sum_{p=1}^{P}\sum_{b=1}^{B}h_{m,p,b}+\gamma_b} \tag{7}$$

$$\theta_{m,k}=\frac{n_{m,k}+\alpha_k}{\sum_{k=1}^{K}n_{m,k}+\alpha_k} \tag{8}$$

$$\pi_{m,s}=\frac{D_{m,s}+\epsilon_s}{\sum_{s=1}^{S}\epsilon_s+D_m} \tag{9}$$

$$\phi_{k,v}=\frac{n_{k,v}+\beta_v}{\sum_{v=1}^{V}\beta_v+n_{k,v}} \tag{10}$$

吉布斯抽样过程的算法如表 2 所列。

表 2 模型的吉布斯抽样算法

算法:Gibbs Sampling for BP-STG

输入:话题集 T,超参数 $\gamma,\delta,\mu,\beta,\alpha,\epsilon$,主题数目 K,群体数目 M

输出:群体活跃度 $\psi$,参与成员在每个群体中的活跃度分布 $\lambda$,参与用户在群体中的行为模式分布 $\eta$,每个群体的主题分布 $\theta$,群体的情感极性分布 $\pi$,主题词分布 $\phi$

```
1. for each microblog d=1to D
2.    randomly assign a community to c_d
3.    for each word n=1 to N_d
4.       randomly assign a topic to z_{d,n}
5.    end for
6. end for
7. for each Gibbs Sampling iteration
8.    for each microblog d=1to D
9.       Draw a community c_d according to Eqn2
10.      for each word n=1 to N_d
11.         Draw a topic z_{d,n} according to Eqn3
12.      end for
13.   end for
14. end for
15. Estimate model parameters ψ,λ,η,θ,π,φ according to Eqn5,Eqn6,Eqn7,
    Eqn8,Eqn9,Eqn10,respectly
```

至此,模型通过吉布斯抽样求解出给定话题集 $T$ 中群体在话题传播过程中的活跃度 $\psi$、群体中每个参与成员的活跃度 $\lambda$、每个参与成员在群体中的行为分布 $\eta$、每个群体中的主题分布 $\theta$ 以及群体中的情感极性分布 $\pi$。对 $\psi_m$ 进行比较,可以得到整个话题传播过程中最活跃的群体;对 $\lambda_{m,p}$ 进行排序,可得群体中相对活跃的用户;同时通过计算 $\eta_{m,p}$ 可以挖掘出相对活跃用户在群体中的行为模式。通过 $\theta_{m,k}$ 的计算,可以挖掘出每个群体最感兴趣的主题。

## 3 实验

### 3.1 实验准备

#### 3.1.1 数据集

本文采用新浪微博的数据作为实验数据,通过新浪微博提供的官方 API 并结合网络爬虫技术获取了两个话题的微博、微博转发信息和微博评论信息,详细信息如表 3 所列。

表 3 数据信息

| 话题 | 原创微博数 | 转发数 | 评论数 | 用户数 |
|---|---|---|---|---|
| 两会 | 7728 | 83609 | 48846 | 69849 |
| 庆安枪击案 | 1696 | 77823 | 91005 | 105322 |

#### 3.1.2 数据预处理

数据集本身包含的是原始微博数据,在使用 BP-STG 模型分析之前先进行一些数据预处理的工作,主要包括 3 部分工作。

(1)去除无用的 HTML 标签以及文本分词。本实验利用正则表达式去除文本中的 HTML 标签,然后利用中文分词系统对文本进行分词。

(2)为文本添加情感标签。本文是基于台湾大学整理发布的 NTUSD 情感词典对微博文本进行情感极性分析[13]。由于 NTUSD 并不是专门针对微博构建的情感词典,而微博文本中常常会出现一些表情符,以表达博主的情感。例如:👍表示赞,是积极情感的表达;😒表示鄙视,是消极情感的表达。因此,本文在 NTUSD 情感词典的基础之上添加了一些微博表情符,将微博的表情符转化成对应的情感语义词。

(3)去除停用词。停用词主要是指代词和表示时间的常用词等,本文采用基于停用词字典的方法将停用词去除。如

果去除停用词后，微博内容为空，且该微博的类型是转发或者评论，则将用原创微博的内容进行替换。

根据用户与文本交互方式的不同，将微博相对应的用户划分为3类：原创用户集、转发用户集、评论用户集。

### 3.2 评价指标

(1)模型建模质量的评价指标

*perplexity* 能够在给定参数环境下预测单词在文本中出现的概率，常被用来度量概率模型在语料库上建模的好坏，其值越小表明模型的建模能力越好。

在BP-STG模型中，*perplexity* 的定义如式(11)所示。

$$perplexity(W)=\exp\{\frac{\sum_{m}\ln(w_m|w)}{\sum_{m}N_m}\} \tag{11}$$

其中，$N_m$ 表示分配给群体 $m$ 中的单词的总数目，$w$ 为测试集，$w_m$ 为观测到被分配给群体 $m$ 的单词。

(2)群体的可区分度指标

由群体的定义可知，群体的可区分度的比较应当从两个方面进行考虑，即参与人员的可区分度和群体的主题兴趣分布的可区分度。本文将群体的可区分度转化为群体的相似度计算问题。利用余弦相似度计算任意两个群体的主题兴趣分布相似度和群体-参与人员活跃度的相似度。两个群体的综合相似度越小，表明群体的可区分度越大。用 $C_{sim}$ 表示任意两个群体 $i$，$j$ 且 $i$ 不等于 $j$ 的相似度。

$$C_{sim}(c_i,c_j)=0.5*\cos(\lambda_i,\lambda_j)+0.5*\cos(\theta_i,\theta_j) \tag{12}$$

整个模型划分出的群体平均可区分度用 $C_d$ 表示，如式(13)所示：

$$C_d=\frac{2\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}(1-\cos(c_i,c_j))}{M\times(M-1)} \tag{13}$$

其中，$M$ 表示模型划分出的群体数目，$C_d$ 越大表明划分出的群体平均可区分度越大，模型划分出的群体的质量越好。

### 3.3 实验结果与分析

#### 3.3.1 参数 $K$ 和参数 $M$ 的取值分析

在BP-STG模型中，超参数的设置参考文献[12]中的方法，分别设置为 $\alpha=\frac{50}{K}$，$\beta=\delta=\gamma=\mu=0.1$。其中主题数目 $K$ 和群体数目 $C$ 的设置是根据文献[3]的方法即计算 *perplexity* 的值来确定的，根据式(11)，*perplexity* 值越小表明选取的 $K$ 值越优。图4展示了在庆安枪击案话题数据集上总的划分的群体的数目 $C$ 和主题个数 $K$ 的不同取值，以及所对应的 *perplexity* 值的变化。

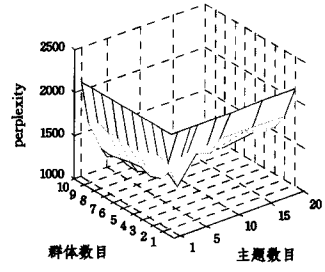



图4 不同主题与群体数目下的 *perplexity* 值

从图4中可以直观地分析出，当 $C$ 取9、$K$ 取8时，*perplexity* 的值最小，模型在这个话题集上建模可达到最佳效果；同理，也可分析出在两会话题集中，当 $C$ 取10、$K$ 取25时，模型建模性能达到最佳。

#### 3.3.2 实验结果展示与分析

基于BP-STG模型进行划分的群体，不仅具有主题倾向性，同时也能够挖掘出群体的情感极性的分布以及参与群体的人员的活跃度及其在群体上的行为分布。表4展示了BP-STG模型在两个话题数据集上挖掘出的主要群体的相关信息。

表4 BP-STG模型在两个话题集上挖掘的群体的相关信息

(a)两会话题数据集中的主要群体($C=10$，$K=25$)

| 群体编号 | | 主题 | | 成员 | | 标签 | 情感 |
|---|---|---|---|---|---|---|---|
| 1(0.1676) | 7 | 退休，延迟，方案，推出，上海 | 0.3057 | 天狗商标达人 | 0.098 | 工作 | Pos(0.6662) |
| | 20 | 记者，发展，建议，改革，政府 | 0.4664 | 央视新闻 | 0.067 | | Neu(0.1099) |
| | | | | 上海长宁区_Koman | 0.026 | | Neg(0.2237) |
| 5(0.1581) | 3 | 工作，腐败，检查，人大常委，新闻 | 0.4082 | 正义网 | 0.063 | 反腐 | Pos(0.1080) |
| | 20 | 记者，发展，建议，改革，政府 | 0.3927 | 央视新闻 | 0.058 | | Neu(0.5729) |
| | | | | 邪恶衙门 | 0.044 | | Neg(0.3189) |
| 10(0.1558) | 14 | 建议，医院，思想，政治，财政 | 0.4842 | 幸福快乐一生688 | 0.090 | 医疗 | Pos(0.5320) |
| | 20 | 记者，发展，建议，改革，政府 | 0.3252 | 好人穷追不舍 | 0.028 | | Neu(0.1346) |
| | | | | 央视新闻 | 0.030 | | Neg(0.3333) |
| 2(0.1245) | 6 | 立法，法律，修改，李克强，议案 | 0.3606 | 正义网 | 0.073 | 司法 | Pos(0.7599) |
| | 20 | 记者，发展，建议，改革，政府 | 0.4664 | 要求祝 | 0.053 | | Neu(0.0740) |
| | | | | 央视新闻 | 0.040 | | Neg(0.1659) |

(b)庆安枪击案话题数据集的主要社团($C=9$，$K=8$)

| 群体编号 | | 主题 | | 成员 | | 标签 | 情感 |
|---|---|---|---|---|---|---|---|
| 7(0.2526) | 4 | 死，女儿，开枪，摔，孩子 | 0.5932 | 新浪视频 | 0.428 | 事发描述 | Pos(0.0039) |
| | 5 | 视频，警察，庆安，律师，公布 | 0.3694 | 评论杨涛 | 0.099 | | Neu(0.9581) |
| | | | | 人民日报 | 0.091 | | Neg(0.0379) |
| 3(0.1945) | 5 | 视频，警察，庆安，律师，公布 | 0.3237 | 头条新闻 | 0.104 | 事件调查 | Pos(0.0026) |
| | 7 | 民警，开枪，调查，铁路，公安部 | 0.6558 | 公安部打四黑除四害 | 0.055 | | Neu(0.9774) |
| | | | | 全球歌舞精选 | 0.044 | | Neg(0.0198) |
| 9(0.1540) | 2 | 律师，徐纯合，黑社会，北京，深夜 | 0.8441 | 黄思敏律师 | 0.194 | 律师被打 | Pos(0.0004) |
| | 5 | 视频，警察，庆安，律师，公布 | 0.1274 | 崔小平律师 | 0.069 | | Neu(0.8849) |
| | | | | 头条新闻 | 0.069 | | Neg(0.1147) |
| 6(0.1087) | 3 | 李乐斌，枪击，真相，法律，垃圾 | 0.6829 | 人民日报 | 0.079 | 要求公开视频 | Pos(0.0924) |
| | 5 | 视频，警察，庆安，律师，公布 | 0.2551 | 梧桐客 | 0.034 | | Neu(0.7582) |
| | | | | 侯虹斌 | 0.033 | | Neg(0.1493) |

由表 4(a)可知,在新浪平台上两会期间用户讨论比较热烈的是关于退休年龄延迟、反腐工作、医疗保险以及立法修改等方面的主题。同时可以观察到关于这些主题的讨论的情感倾向,积极的要高于消极的。由于央视新闻是微博看两会话题的主持人,央视新闻应当参与了每个具有不同主题倾向性的群体的讨论,且在群体中是相对比较活跃的。由此验证了提出的 BP-STG 模型的正确性。由表 4(b)可知,庆安枪击案被曝光以后,用户对事情的经过、调查结果、公布的视频的真实性等进行了激烈的讨论,且用户的中立和消极的情感倾向远高于积极的情感态度。人民日报与头条新闻同时参与了多个群体的讨论。从表 4(a)中可以看到主题 20 包含在每个群体中,原因是主题 20 中所包含的词汇是两会话题讨论中经常出现的词汇。同理,表 4(b)中主题 5 包含庆安枪击案的通用词汇,故每个群体中都包含主题 5。

模型除挖掘群体兴趣主题以及群体活跃人物之外,还能够挖掘出参与人员在群体中的行为模式。图 5 以两会数据集中划分出的群体 10 和群体 2 为例,展示出在群体中相对比较活跃的用户的行为分布。

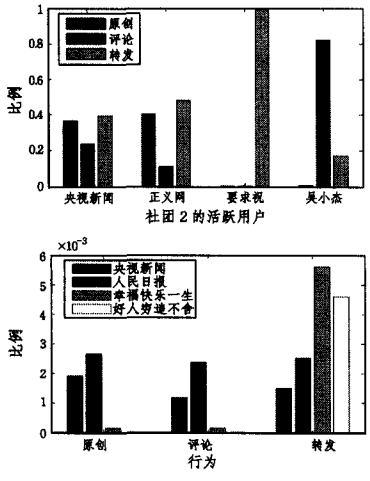
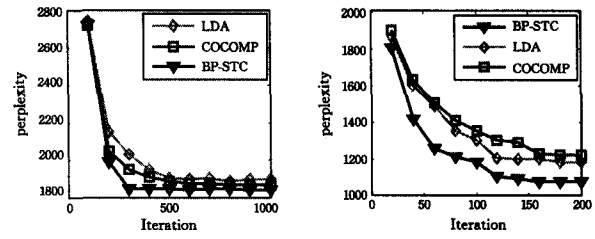



图 5 群体 2 和群体 10 中相对活跃的用户行为比较

图 5 中的参与用户分为两种类型:媒体用户,如央视新闻、正义网、人民日报;普通用户,如要求祝、吴小杰、幸福快乐一生、好人穷追不舍。可以观察到媒体用户在群体中的行为多为原创,而普通用户则多为转发与评论,同时媒体用户行为分布相对比较均匀,而普通用户则行为相对比较极端。在两会话题的讨论中,信息源头是媒体用户,普通用户对话题信息进行了传播。为了进一步验证模型的正确性,人工观察了群体 2 中要求祝(微博昵称)在两会期间的微博,发现其在两会期间之所以在群体 2 中转发大量关于法律方面的微博,是因为其妻子意外死亡而肇事者没有赔偿,其希望利用两会话题讨论能引起社会关注,得到法律保护。

### 3.3.3 对比实验

由于所提出的群体划分模型是在 LDA 概率生成模型基础之上进行改进的,因此为了验证模型在数据集上建模能力的好坏,将其与 LDA 概率生成模型以及结合文本内容及用户的社会联系进行群体挖掘的 COCOMP[9] 模型进行比较。实验中 LDA 模型的参数 $K$ 和 BP-STG 模型的参数 $K$ 设置为相同数值。COCOMP 模型中的参数 $K$ 和参数 $C$ 的设置与 BP-STG 相同。图 6 示出了 3 种模型在两个数据集下的 *perplexity* 的值。




(a)两会数据集下的 *perplexity* 值 (b)庆安枪击案数据集下的 *perplexity* 值

图 6 3 种模型在两个数据集下的 *perplexity* 值

由图 6 可知,BP-STG 模型的 *perplexity* 值均小于 LDA 和 COCOMP 模型的,证明了 BP-STG 模型利用话题传播中用户的交互和文本内容的关联关系及文本的情感倾向对微博进行分析,确实能够提高模型在特定微博话题数据集上的性能。

同时实验还对 COCOMP 模型与 BP-STG 模型在划分出的群体的可区分度方面进行了比较。为了降低计算的时间与空间复杂度,实验中均提取出每个群体中活跃度排名前 100 的成员,通过计算式(13)进行比较,结果如表 5 所列。

表 5 两种模型所划分出的群体的可区分度效果评价

| 模型 \ 数据集 | 两会 | 庆安枪击案 |
|---|---|---|
| COCOMP | 0.8621 | 0.7443 |
| BP-STG | 0.8723 | 0.8543 |

由表 5 可知,根据 BP-STG 模型所挖掘出来的群体的可区分度并不低于根据 COCOMP 模型所挖掘的群体的可区分度。

**结束语** 本文针对微博特殊的文本结构,综合考虑了微博文本的语义信息、用户在话题传播中的交互、用户与文本的交互行为以及文本的情感极性,提出了一个适合于对微博话题传播中群体划分方法的 BP-STG 模型。该模型不仅能够挖掘出微博话题传播中的具有相同主题倾向性的群体,还能有效地挖掘出群体的情感极性分布、参与人员在群体中的活跃度及其行为表现。

BP-STG 模型不仅可以应用于微博数据的群体分析,还能扩展到许多带有社交网络性质的媒体中,比如 Email 数据的社团发现:Email 中的回复关系可以看作评论关系,Email 中收件人和抄送人可以看作是微博中被提及的人,Email 中的转发就为微博中的转发关系;其他例如论文合作网,也可以借鉴 BP-STG 模型进行群体划分。

今后的研究工作中将继续优化 BP-STG 模型的效率,探索能够实现根据用户的行为表现来判定参与用户对主题的情感倾向性的方法;同时设计一个动态的群体划分模型,以实时监测话题中的群体的主题以及情感变化。

## 参 考 文 献


[1] Girvan M, Newman M. Community structure in social and biological networks[J]. PNAS, 2002, 99(12): 7821-78266

[2] Palla G, Barabasi A, Vicsek T. Quantifying social group evolution[J]. Nature, 2007, 446(7136): 664-667

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022


(下转第 239 页)

党兰学,侯彦娥,孔云峰. 基于时空相关的混载校车路径问题邻域搜索[J]. 计算机科学,2015,42(4):221-225

[8] Hoff A, Anderson H, Christiansen M, et al. Industrial aspects and literature survey: Fleet composition and routing [J]. Computers and Operations Research, 2012, 37(12): 2041-2061

[9] Golden B, Assad A, Levy L, et al. The fleet size and mix vehicle routing problem [J]. Computers and Operations Research, 1984, 11(1): 49-66

[10] Taillaid E D. A heuristic column generation method for the heterogeneous fleet VRP [J]. RAIRO, 1999, 33: 1-34

[11] Subramanian A, Penna P H V, Uchoa E, et al. A hybrid algorithm for the Heterogeneous Fleet Vehicle Routing Problem [J]. European Journal of Operational Research, 2012, 221(2): 285-295

[12] Brandão J. A deterministic tabu search algorithm for the fleet size and mix vehicle routing problem [J]. European Journal of Operational Research, 2009, 195(3): 716-728

[13] Liu S. A hybrid population heuristic for the heterogeneous vehicle routing problems[J]. Transportation Research Part E, 2013, 54: 67-78

[14] Matei O, Pop P C, Sas J L, et al. An improved immigration memetic algorithm for solving the heterogeneous fixed fleet vehicle routing problem [J]. Neurocomputing, 2015, 150: 58-66

[15] Li X, Leung S C H, Tian P. A multistart adaptive memory-based tabu search algorithm for the heterogeneous fixed fleet open vehicle routing problem [J]. Expert Systems with Applications, 2012(39): 365-374

[16] Liu F H, Shen S Y. The fleet size and mix vehicle routing problem with time windows [J]. Journal of the Operational Research Society, 1999, 50(7): 721-732

[17] Jiang J, Ng K M, Poh K L, et al. Vehicle routing problem with a heterogeneous fleet and time windows[J]. Expert Systems with Applications, 2014, 41: 3748-3760

[18] Li L Y O, Fu Z. The School Bus Routing Problem: A Case Study [J]. The Journal of the Operational and Research Society, 2002, 53(5): 552-558

[19] Fu Zhuo. The Open Vehicle Routing Problems and Their Applications[D]. Changsha: Central South University, 2003 (in Chinese)

符卓. 开放式车辆路径问题及其应用研究[D]. 长沙:中南大学,2003

[20] Park J, Tae H, Kim B I. A Post-improvement Procedure for the Mixed Load School Bus Routing Problem [J]. European Journal of Operational Research, 2012, 217(1): 204-213

[21] Ripplinger D. Rural school vehicle routing problem [J]. Transportation Research Record: Journal of the Transportation Research Board, 2005, 1992: 105-110

[22] Ke X. School bus selection, routing and Scheduling[D]. Canada, Windsor: University of Windsor, 2005

[23] Thangiah S R, Fergany A, Wilson B, et al. School Bus Routing in Rural School Districts[J]. Computer-Aided Systems in Public Transport, 2008, 600: 209-232

[24] Luzia V D S, Paulo H S. Heuristic Methods Applied to the Optimization School Bus Transportation Routes-A Real Case[C]// 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE). 2010, 6093: 247-256

[25] Braca J, Bramel J, Posner B, et al. A computerized approach to the New Yoke City school bus routing problem[J]. IIE Transactions, 1997, 29: 693-702

[26] Gendreau M, Potvin Jean-Yves. Handbook of Metaheuristics (2nd Edition)[M]. New York: Springer-Verlag New York Inc., 2010

[27] Martí R, Resende M G C, Ribeiro C C. Multi-start methods for combinatorial optimization [J]. European Journal of Operational Research, 2013, 226: 1-8

[28] Prais M, Celso R. Reactive GRASP: An Application to a Matrix Decomposition Problem in TDMA Traffic Assignment [J]. INFORMS Journal on Computing, 2000, 12(3): 164-176

[29] Alvarez-Valdes R, Parreño F, Tamarit J M. Reactive GRASP for the strip-packing problem [J]. Computers & Operations Research, 2008, 35(4): 1065-1083

[30] Hansen P, Mladenović N. Variable Neighborhood Search: Principles and Applications [J]. European Journal of Operations Research, 2001, 130: 449-467

[31] Vidal T, Crainic T G, Gendreau M, et al. Heuristics for multi-attribute vehicle routing problems: A survey [J]. European Journal of Operational Research, 2013, 231(1): 1-21

(上接第 228 页)

[4] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery[C]// Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2004: 306-315

[5] Zhou D, Manavoglu E, Li J, et al. Probabilistic models for discoveringe-communities[C]// WWW. 2006: 173-182

[6] Pathak N, DeLong C, Banerjee A, et al. Social topic models for community extraction[C]// The 2nd SNA-KDD Workshop. 2008

[7] Sachan M, Contractor D, Faruquie T, et al. Using content and interactions for discovering communities in social networks[C]// WWW. 2012: 331-340

[8] Yang T, Jin R, Chi Y, et al. Combining link and content for community detection: discriminative approach[C]// KDD. 2009: 927-936

[9] Zhou W, Jin H, Liu Y. Community discovery and profiling with social messages[C]// KDD. 2012: 388-396

[10] Yang B, Manandhar S. Stc: A joint sentiment-topic model for community identification [M] // Trends and Applications in Knowledge Discovery and Data Mining. Springer International Publishing, 2014: 535-548

[11] Ding Zhao-yun, Jia Yan, Zhou Bin. Survey of Data Mining for Microblogs[J]. Computer Research and Development, 2014, 51(4): 691-706 (in Chinese)

丁兆云,贾焰,周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展,2014,51(4):691-706

[12] Steyvers M, Griffiths T. Probabilistic topic models[J]. Handbook of Latent Semantic Analysis, 2007, 427(7): 424-440

[13] Chen Xiao-dong. Research and Sentiment Dictionary based Emotional Tendency Analysis of Chinese Microblog[D]. Wuhan: Huazhong University of Science & Technology, 2012 (in Chinese)

陈晓东. 基于情感词典的中文微博情感倾向分析研究[D]. 武汉:华中科技大学,2012