基于混合权重合并策略的社交网络用户关注点识别方法

姬建睿 刘业政 姜元春*
(合肥工业大学管理学院 合肥 230009)
(过程优化与智能决策教育部重点实验室 合肥 230009)

摘 要：主题模型是用于识别博客、网络社区、微博等社交网络平台上用户关注点的重要手段。考虑到社交网络平台上短文本主题识别的特殊性，该文根据短文本内容在上下文上的相关性，提出一种基于混合权重合并策略的 AW-LDA 模型。该模型将符合上下文相关条件的短文本进行虚拟合并，并根据上下文相关程度对不同短文本赋予不同的权重，构建了一种新的短文本主题识别方法。通过网络 BBS 社区与微博社区两组数据的实验，该模型能够有效识别不同话题下社交网络用户关注点，为解决短文本主题识别问题提供了新的解决思路。

关键词：社交网络；主题模型；关注点识别；混合权重；“邻近”用户
中图分类号：TP393; TP181 **文献标识码**：A **文章编号**：1009-5896(2017)09-2056-07
DOI: 10.11999/JEIT161348

Recognizing Users' Focuses on Social Network Based on Mixed-weight Combined Strategy

JI Jianrui LIU Yezheng JIANG Yuanchun
(*School of Management*, *Hefei University of Technology*, *Hefei* 230009, *China*)
(*Key Laboratory of Process Optimization and Intelligent Decision-making*, *Ministry of Education*, *Hefei* 230009, *China*)

Abstract: It is an important measure to utilize the topic model to recognize the users' focuses on social networks, such as blog, online community, and microblog. Considering the particularity of topic recognizing of short texts on the social network platform, this paper develops an AW-LDA model based on mixed-weight combined strategy according to the relevance of short texts' context. This model virtually combines short texts, which are in line with contextual-related conditions, and endows different short texts with different weights according to the related extent. It proposes a new method of recognizing short texts' topics. According to the experiments on data of BBS and Weibo communities, the results show that the model can effectively recognize social network users' focuses on different subjects and it proposes a new idea about solving the topic recognition problem of short texts.

Key words: Social network; Topic model; Focus recognition; Mix-weight; “Neighboring” user

1 引言

各类在线社交网络平台已经成为用户内容分享、信息获取，以及观点表达、讨论与传播的重要渠道，很多经济社会领域的热点问题在社交网络平台上很容易形成热门话题。例如“央视 315 晚会曝光江淮汽车生锈质量问题”、“厦门公交纵火案”等在社交网络上都形成了热门话题，并引发了大量的讨论与争论。由于人的认知存在差异，网络用户会从不同的角度看待问题，且随着事件的发展，用户对事件的关注点也逐渐发生了变化。挖掘用户在社交网络上面向话题而创建的内容主题，识别社交网络用户关注点，能够帮助企业或政府组织了解用户对企业或社会热点话题关注的角度与程度，以及随着时间发展用户关注点的变化，对于改善企业或政府组织提高管理决策水平具有重要意义。

在线社交网络内容的主题识别与跟踪一直是研究的热点之一。文献[1]提出了具有背景分布的话题 N 元模型，抽取单个词语作为话题线索标签了解新闻事件各个角度。文献[2]利用与事件有关的位置、日期等信息构建基于标签的主题模型，并在 Twitter

收稿日期：2016-12-09；改回日期：2017-05-12；网络出版：2017-06-14
*通信作者：姜元春 ycjiang@hfut.edu.cn
基金项目：国家自然科学基金(71490725, 71521001, 71371062, 91546114, 71501057)，国家 973 规划项目(2013CB329603)，国家科技支撑计划项目(2015BAH26F00)，教育部人文社会科学研究青年基金(15YJC630111)
Foundation Items: The National Natural Science Foundation of China (71490725, 71521001, 71371062, 91546114, 71501057), The National 973 Program of China (2013CB329603), The National Key Technology Support Program (2015BAH26F00), MOE Project of Humanities and Social Sciences (15YJC630111)

语料中发现子事件。文献[3]使用微博平台文本内容以及用户社交关系信息预测突发性事件并提出一种扩散模型进行事件趋势的预测。上述研究主要用来进行新闻事件的检测与跟踪，对由事件产生的话题以及用户的关注点研究较少。识别文本语料主题内容的一种重要手段是主题模型[4]，而微博、微信、网络社区等在线社交网络平台的一个重要特征是每篇文本的篇幅较短，因此，一些学者通过引入外部信息增强短文本的理解能力。Weng 等人[5]将相同用户的 tweets 文本合并成为长文本；文献[6–8] 等通过维基百科等外部资源实现对短文本的扩充。文献[9]利用迁移学习的方法从外部补充的长文本数据来得到短文本的主题。文献[10]提出一种双词话题模型解决短文本的主题识别问题。文献[11]将若干短文本聚合为伪长文本，并对伪长文本使用钉板先验取得较好效果。Lin 等人[12]将原有文本词特征空间转化依据词共现的伪文本词特征空间，并在该特征空间使用基于词的一致性聚类获得短文本语料的主题。Zhao 等人[13]借助每条短文本仅包含一个主题的假设，构建短文本一元混合模型并与传统模型进行对比，但该假设在某些情景中并不适合，而且会在模型训练中引起过拟合问题[4]。综上可以看出，现有研究经常利用外部知识扩充或增强短文本的表示性，从而解决主题建模中由短文本导致的稀疏性问题。但是在实际应用中，并非所有问题都存在或者易于获得完整的外部信息，这在针对某一具体话题的语料中尤为突出。因此，需要设计有效的方法从已有的语料数据中挖掘短文本的内在信息。

考虑到社交网络平台上短文本主题识别的特殊性，本文根据短文本内容在上下文上的相关性，提出了一种基于混合权重合并策略的(Aggregated-Weigh LDA, AW-LDA)，模型。AW-LDA 模型是在标准 LDA 主题模型的基础上，放宽了文本间的独立性假设而形成的。该模型假定针对特定话题，“邻近”用户的关注点具有相似性。所谓“邻近”用户是指在相邻的时间针对特定话题发表帖子或者在主帖上进行跟帖、回复的用户，将“邻近”用户发表的帖子称为上下文相关的帖子。因此通过分析“邻近”用户文本间的上下文关系有助于语料中主题的发现。考虑到不同“邻近”用户的上下文相关性存在差异，AW-LDA 模型采纳了一种混合权重策略为每条待合并的文本赋权，即越“邻近”的用户文本权重越高。该赋权策略可以降低噪音文本的影响。

本文选择汽车之家网络社区中奥迪 A4L 品牌车型的话题帖子与讨论厦门 BRT 纵火案的微博这 2 个话题的短文本进行实验，实验结果表明本文所提模型在基于短文本主题的网络用户关注点识别上效果显著。本文主要贡献在于(1)提出一种短文本的主题建模新思路。(2)该方法作为一种面向特定话题的关注点识别方法，为话题引导、监督提供了理论指导。

2　混合权重合并策略的关注点识别方法

在线社交网络用户关注点的识别即为某话题下文本语料的主题识别。传统的主题识别方法主要应用在长文本语料中，但由于网络社区与微博的特殊性，用户每条文本的发言包含汉字较少，对短文本直接使用主题模型将产生数据稀疏问题。而且针对特定话题的用户关注点识别问题，使用现有研究方法并不适合。本文在标准 LDA 模型的基础，放宽文本间独立性假设，假定针对特定话题“邻近”用户关注的焦点具有相似性。“邻近”用户是指在相邻的时间针对特定话题发表帖子的用户或者在主帖上进行跟帖、回复的用户，将“邻近”用户发表的帖子称为上下文相关的帖子。“邻近”用户相比其他非“邻近”用户更多围绕相似主题发表上下文相关帖子，如针对某话题的微博语料，相邻发表微博时间的用户倾向于发表相同主题的微博，这些用户互称为“邻近”用户，而随着时间推移，用户关注点发生变化，距离发表时间较远的非“邻近”用户发表微博主题有较大差别。又如针对网络 BBS 社区中话题讨论帖语料，同一讨论帖主贴及其对应跟帖或回复的用户都是围绕主贴这一主题发表帖子，本文称为“邻近”用户，而不同话题讨论帖则为不同的“邻近”用户。通过分析“邻近”用户文本间的上下文相似关系能够更有助于语料中主题的发现。本文对符合上述假定的短文本构建混合权重合并模型，即 AW-LDA 模型。

2.1 AW-LDA 模型

根据上述假设，将“邻近”用户的各短文本置于同一区间，在该区间内，每条文本的主题分布较其他区间更为相似，因此将该区间内短文本进行虚拟合并，区间内所有短文本作为合并后一篇新的虚拟长文本的子文本，这些子文本的主题分布彼此条件共享。与已有将短文本合并的方式不同，本文合并后的新文本是一个虚拟文本，新文本内词的生成方式与真实文本生成方式并不相同，文本中的词共现形式与真实文本也不相同。真实文本中词的共现属于直接共现形式，而新虚拟长文本中不同子文本间的词实际并不是来自同一个真实文本，本文把这种共现形式称为条件共现。如果把条件共现近似为直接共现，必然引入共现噪音，可能会使识别效果

下降。为了避免对条件共现的简单近似，同时由于新虚拟长文本内上下文相关程度不同的子文本对当前子文本的词的主题分布贡献也并不相同，本文通过赋予各子文本不同的权重区分贡献率。

首先按照“邻近”用户将短文本数据根据发表时间先后顺序分成若干区间，每一区间内的短文本合并为一篇新虚拟长文本 d_{AR}^{i}，表示为 $d_{\mathrm{AR}}^{i}=\left(\mathrm{w}_{11}^{i},\mathrm{w}_{12}^{i},\mathrm{w}_{13}^{i},\cdots,\mathrm{w}_{h1}^{i},\mathrm{w}_{h2}^{i},\cdots,\mathrm{w}_{H1}^{i},\mathrm{w}_{H2}^{i},\mathrm{w}_{H3}^{i},\cdots\right)$，每篇虚拟长文本包含原始短文本数量为 $d_{\mathrm{AR}}^{i,H}$，$i\in[1,M]$，M 为虚拟长文本的数量，语料集所有短文本数量为 $\mathrm{AM}=\sum_{i=1}^{M}d_{\mathrm{AR}}^{i,H}$。虚拟长文本中每条子文本根据顺序设置对应的位置索引编号，编号距离越大，代表发表时间间隔越大。编号能够反映用户的“邻近”程度。第 i 篇虚拟长文本中第 h 子文本表示为：$d_{h}^{i}=\left(\mathrm{w}_{h1}^{i},\mathrm{w}_{h2}^{i},\cdots,\mathrm{w}_{hN}^{i}\right)$，$h$ 即为其编号，w_{h1}^{i} 为其中的词。

在第 i 篇虚拟长文本中，h 子文本中 w_{h1}^{i} 与 w_{h2}^{i} 等词来自同一真实文本，呈完全直接共现；而与其他非 h 子文本中 $\mathrm{w}_{\neg h1}^{i}$ 等词实际上位于不同的短文本中，因此 w_{h1}^{i} 与 $\mathrm{w}_{\neg h1}^{i}$ 呈现弱共现关系，即条件共现。

AW-LDA 模型的生成过程类似于 LDA，如图 1 所示。与标准 LDA 不同，对于每一新虚拟长文本，由于弱共现关系的存在，其余子文本的主题与词会影响当前子文本的分布。具体地说，在每篇虚拟长文本中，每条子文本不仅由该文本自身的主题分布决定，还受到其对应虚拟长文本中其余“邻近”子文本的主题分布影响；同时，当前子文本的主题分布也影响其余子文本主题与词的生成。在考虑某当前子文本的主题分布时，由于新虚拟长文本内各子文本间隔不同，不同的子文本上下文相关程度不同，对当前子文本的词的主题分布贡献也并不相同。根据每条子文本对其余子文本“邻近”程度的差别，赋予各子文本不同的权重区分贡献率，如距离当前子文本发布时间越接近的子文本影响越大，即距离当前子文本位置编号越近的子文本影响越大。

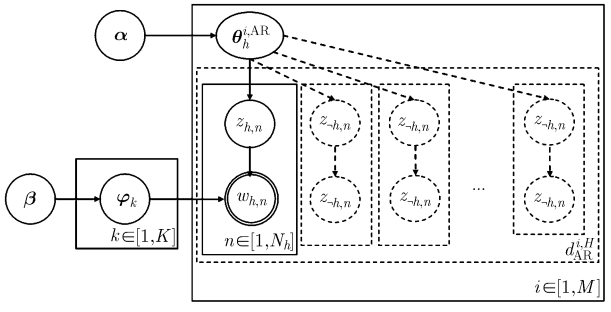

图 1 AW-LDA 模型

本文使用 τ 描述第 l 条子文本对第 h 条子文本的影响权重值，$\tau=\dfrac{\lambda}{|h-l|+1}$。$h$，$l$ 分别为当前子文本 h 以及其余第 l 条子文本在虚拟长文本的索引编号，编号越接近，文本越“邻近”，权重值越大，第 l 子文本对第 h 子文本的主题分布影响越大。λ 为权重的调节参数，λ 取值越大，其余子文本对当前子文本的主题分布影响越大，λ 通过实验获得最优值。

这种影响机制相当于为第 h 子文本增加若干含权重的词。例如：在虚拟长文本 i 内，当计算第 h 子文本的主题分布时，标准 LDA 模型只考虑文本自身词的分布，而 AW-LDA 模型对该子文本进行虚拟赋权的词扩展，即为第 h 子文本词的数量增加 $\sum_{l=1,l\neq h}^{H}\tau w_{\neg lN}^{i}$。此时，该子文本的主题分布为 $\theta_h^{i,\mathrm{AR}}$。$\theta_h^{i,\mathrm{AR}}$ 为考虑虚拟长文本 i 内非 h 子文本对 h 子文本影响下的主题分布。如前所述，由于针对虚拟长文本内不同的 h 子文本，非 h 子文本影响权重也不相同，因此不同 h 子文本对应 $\boldsymbol{\theta}_h^{i,\mathrm{AR}}$ 并不相同。非 h 等子文本为虚拟合并，在样本生成过程只采样 h 子文本词的主题，非 h 子文本词的主题由其对应的 $\boldsymbol{\theta}_{\neg h}^{i,\mathrm{AR}}$ 分布采样。

该模型的联合概率分布为

$$\begin{aligned}p(\boldsymbol{z}\mid\boldsymbol{\alpha})&=\int p\left(\boldsymbol{z}\middle|\boldsymbol{\theta}_h^{i,\mathrm{AR}}\right)p\left(\boldsymbol{\theta}_h^{i,\mathrm{AR}}\middle|\boldsymbol{\alpha}\right)\mathrm{d}\boldsymbol{\theta}_h^{i,\mathrm{AR}}\\&=\int\prod_{m=1}^{\mathrm{AM}}\frac{1}{\Delta(\boldsymbol{\alpha})}\prod_{k=1}^{K}\left(\boldsymbol{\theta}_{h,k}^{i,\mathrm{AR}}\right)^{\mathrm{NUM}}\mathrm{d}\boldsymbol{\theta}_h^{i,\mathrm{AR}}\\&=\prod_{m=1}^{\mathrm{AM}}\frac{\Delta\left(\boldsymbol{n}_m^{\mathrm{new}}+\boldsymbol{\alpha}\right)}{\Delta(\boldsymbol{\alpha})}\end{aligned}\tag{1}$$

其中，$\mathrm{NUM}=n_h^k+\sum_{l=1,l\neq h}^{H}\dfrac{\lambda n_l^k}{|h-l|+1}+\alpha_k-1$。

$$\begin{aligned}p(\boldsymbol{w},\boldsymbol{z}\mid\boldsymbol{\alpha},\boldsymbol{\beta})&=p(\boldsymbol{w}\mid\boldsymbol{z},\boldsymbol{\beta})\,p(\boldsymbol{z}\mid\boldsymbol{\alpha})\\&=\prod_{k=1}^{K}\frac{\Delta\left(\boldsymbol{n}_k+\boldsymbol{\beta}\right)}{\Delta(\boldsymbol{\beta})}\prod_{m=1}^{\mathrm{AM}}\frac{\Delta\left(\boldsymbol{n}_m^{\mathrm{new}}+\boldsymbol{\alpha}\right)}{\Delta(\boldsymbol{\alpha})}\end{aligned}\tag{2}$$

基于上述描述，该模型详细生成过程如下：

(1)对于每一个主题 z，采样一个主题对词的分布：$\boldsymbol{\varphi}_z\sim\mathrm{Dir}(\beta)$；

(2)对于第 i 篇新虚拟长文本 d_{AR}^{i} 中一条子文本 d_h^i，采样该文本对主题的分布：$\boldsymbol{\theta}_h^{i,\mathrm{AR}}\sim\mathrm{Dir}(\alpha)$；

(3)对于子文本 d_h^i 中每一个词 w_{hn}^{i}，先采样 $z_{h,n}^{i}\sim\mathrm{Multi}\left(\boldsymbol{\theta}_h^{i,\mathrm{AR}}\right)$；

(4)根据 $z_{h,n}^{i}$，再采样 $\mathrm{w}_{hn}^{i}\sim\mathrm{Multi}(\boldsymbol{\varphi}_z)$。

AW-LDA 模型假设将“邻近”的 H 条文本看成虚拟长文本，并没有发生真实合并，因此式(2)中 m 仍为原语料集第 m 篇文本。

$\boldsymbol{n}_m^{\text{new}}$ 与 LDA 模型 $\boldsymbol{n}_m$ 不同，代表联合考虑当前文本与其对应虚拟文本中其余子文本的词的数量。

2.2 参数估计

本文使用 Gibbs 抽样方法进行参数估计，应用上述联合概率分布以及链式规则，得到如式(3)后验条件概率：

$$p\left(z_o = k \mid \boldsymbol{z}^{-o}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{\left(n_{h,-o}^{k} + \sum\limits_{l=1,l\neq h}^{H} \frac{\lambda n_l^k}{|h-l|+1}\right) + \boldsymbol{\alpha}}{\sum\limits_{k=1}^{K}\left(n_{h,-o}^{k} + \sum\limits_{l=1,l\neq h}^{H} \frac{\lambda n_l^k}{|h-l|+1}\right) + K\boldsymbol{\alpha}} \cdot \frac{n_{k,-o}^{t} + \boldsymbol{\beta}}{\sum\limits_{t=1}^{V} n_{k,-o}^{t} + V\boldsymbol{\beta}} \tag{3}$$

其中，o 为第 h 子文本第 o 个词。从式(3)能够看出，文本中每个词的主题仍由两部分决定，文本的主题分布与主题对应的词的分布。文本的虚拟合并不会改变主题相对词的分布。与以往主题模型不同，当前词的主题不仅依赖于当前子文本主题词的数量，还依赖于新虚拟长文本内不同权重的其余子文本不同主题词的数量。

使用该条件概率即可对每个词的主题进行抽样。抽样过程如下：

(1)算法输入关注点数量 K，超参数 $\boldsymbol{\alpha}, \boldsymbol{\beta}$；

(2)算法输出主题与词的多项式分布参数 $\boldsymbol{\varphi}$，文本与主题分布参数 $\boldsymbol{\theta}$；

(3)初始化，给每个文本中词的指派一个主题；

(4)迭代 N_{IR} 次，对于每次迭代：根据式(3)，抽样当前词的主题，更新参数 n_h^k, n_k^t；

(5)计算参数 $\boldsymbol{\varphi}, \boldsymbol{\theta}$：

$$\varphi_{k,t} = \frac{n_k^t + \boldsymbol{\beta}}{\sum\limits_{t=1}^{V} n_k^t + V\boldsymbol{\beta}}, \quad \theta_{m,k} = \frac{n_m^k + \boldsymbol{\alpha}}{\sum\limits_{k=1}^{K} n_m^k + K\boldsymbol{\alpha}} \tag{4}$$

由式(4) $\theta_{m,k}$ 可知，与 $\boldsymbol{\theta}_h^{i,\text{AR}}$ 对应的短文本的虚拟主题分布不同，$\theta_{m,k}$ 为语料集第 m 条短文本的实际主题分布，该分布只与该短文本的实际词的主题有关。

3　实验分析与讨论

3.1 关注点识别与一致性检验

本文使用 2 个数据集进行模型测试。

(1)新浪微博关于“厦门 BRT 爆炸”的微博数据集。数据集时间跨度从 2013 年 6 月 7 日第 1 条相关微博出现截止 2014 年 6 月 30 日，经去重，过滤无关微博，共收集 43403 条有效微博。首先去除每条微博中所含话题标签、网址链接等无关信息，再通过 NLPIR 汉语分词系统[14]进行分词，过滤停用词后，将不满足 4 个词数的微博去除。经过上述预处理，共保留 35619 条。针对本微博数据，按照微博发布时间顺序，将每 50 条微博作为“邻近”用户微博文本进行虚拟合并。

(2)奥迪 A4L 论坛数据集。选自汽车之家从 2015 年 5 月 28 日到 2015 年 9 月 11 日共计 7194 话题讨论帖子，经过相同预处理得到 3824 个讨论帖，累计回复帖 115762 个。针对论坛数据，本文将同一话题的主帖以及对应回复帖进行虚拟合并。

本文选择标准 LDA 模型、A-LDA 模型、以及一元混合模型进行对比试验。A-LDA 模型的思路是直接将短文本合并为长文本，如将同一作者的文本合并为一起。选择 A-LDA 模型建模，数据处理与 AW-LDA 模型不同，将“邻近”短文本直接合并为一篇长文本。每篇长文本包含原始文本数量为 $d_{\text{AR}}^{i,H}$ 条。参数估计的条件概率为

$$p\left(z_i = k \mid \boldsymbol{z}^{-i}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{\left(n_{h,-i}^{k} + \sum\limits_{l=1,l\neq h}^{H} n_l^k\right) + \boldsymbol{\alpha}}{\sum\limits_{k=1}^{K}\left(n_{h,-i}^{k} + \sum\limits_{l=1,l\neq h}^{H} n_l^k\right) + K\boldsymbol{\alpha}} \frac{n_{k,-i}^{t} + \boldsymbol{\beta}}{\sum\limits_{t=1}^{V} n_{k,-i}^{t} + V\boldsymbol{\beta}} \tag{5}$$

对各模型识别的关注点进行一致性检验。评价指标[15]为

$$c\left(t; \boldsymbol{V}^{(t)}\right) = \sum_{\text{nw}=2}^{\text{NW}} \sum_{l=1}^{\text{nw}-1} \lg \frac{D\left(v_{\text{nw}}^{(t)}, v_l^{(t)}\right) + 1}{D\left(v_l^{(t)}\right)} \tag{6}$$

$D(v)$ 代表词 v 的文本频率；$D(v, v')$ 代表词 v 与 v' 共现的文本频率；$\boldsymbol{V}^{(t)} = \left(v_1^{(t)}, v_2^{(t)}, \cdots, v_{\text{NW}}^{(t)}\right)$ 代表主题 t 中最可能的代表词集。该指标值越大，说明主题识别效果越好。

本文的实验环境为 Intel(R) Core(TM)2 Duo E7500 2.93 GHz 的 CPU, 8G 的内存，500G 硬盘的 PC 机。操作系统为 WIN 7，实验工具为 jdk 为 1.6.0。经过多次取值实验比较，最终在“厦门 BRT 爆炸”微博语料中设置 λ 为 5，在奥迪 A4L 论坛语料设置 λ 为 4。

表 1 列出使用 AW-LDA 模型识别出“厦门 BRT 爆炸”数据集中民众重点关注的 10 个关注点。从中可以看出关注点 5, 7, 8 描述了爆炸事件发生后，民众重点关注事件发生详情以及伤亡情况；关注点 3, 4 重点表达了民众对伤员救治和伤亡人员情况的了解与祈福；注意到关注点 1, 6, 9 已经转移注意力到事件发生后的补偿与反思，如受伤高考考生考试事宜以及犯罪嫌疑人作案动机等。4 个模型识别的关注点一致性检验在表 2 与图 2 呈现。

表 1 “厦门 BRT 爆炸”话题的 10 个重点关注点

关注点	内容
关注点 1	陈水总、警方、 嫌疑人、社会、无辜、犯罪、纵火、微博、政府……
关注点 2	纵火案、安全、47、专题、他们、同学、逃生、中国、社会、如何……
关注点 3	受伤、现场、 高考、考生、 起火、目前、事故、燃烧、重伤、轻伤、 消息、原因、伤员、记者、 医院……
关注点 4	骆琪琳、陈雅慧、戴学彬、陈伟、洪菁菁、蜡烛、事故、医院、联系、 死亡、42、帮忙、造成、今天、33、家人、朋友、伤者……
关注点 5	蜡烛、生命、怎么、6 月、发生、事件、遇难者、7 日、报道、遇难、需要……
关注点 6	觉得、考生、合理、受伤、直接、大学、他们、考试、公平、公布、成绩、全部……
关注点 7	厦门市、死、30、快速、傍晚、昨日、政府、着火、高架路、案件、47、受伤……
关注点 8	蜡烛、死亡、起火、造成、30、事故、愿、安息、逝者、获悉、政府、公车、发生、值班室、希望……
关注点 9	嫌疑人、起火案、工作、疑似、陈水总、希望、自白、生活、一个、厦门市、曝光、案件……
关注点 10	发生、地址、安全、寻人、案件、进展、微博、失去、事件、死伤……

如图 2 所示，AW-LDA 模型在各个关注点数量的一致性检验指标均大于其余模型的一致性指标，说明本文所提模型相较其余模型发现的关注点更便于对文本集的理解。一元混合模型虽然不如 AW-LDA 模型效果好，但仍然优于直接对文本集使用主题模型或者简单合并为长本文再进行主题分析。

表 3，表 4 与图 3 列出了汽车之家网络社区奥迪 A4L 板块用户讨论的部分关注点与主题一致性检验分对比析。与微博数据集类似，AW-LDA 模型在网络社区中的效果也优于其余 3 个模型。而 A-LDA 模型虽不及所提模型，但也有较好效果，一致性检验指标优于 LDA 模型与一元混合模型。

从图 2，图 3 对比发现，A-LDA 和一元混合模型在微博语料和论坛语料的实验效果出现反转，主要原因是由于两个模型不同的假设前提在处理针对上述两种不同类型的短文本语料数据而产生反转效果。A-LDA 模型的思路是直接将“邻近”用户发表的短文本合并为长文本，该模型的假设前提为，“邻近”用户的短文本来自相同的主题分布，将短文本合并为长文本既不影响各短文本的主题分布，又能避免单独建模短文本时产生的稀疏性，因此短文本描述的主题越相似，将短文本合并为长文本越合理，建模效果越好。一元混合模型假设每条短文本只有一个主题，因此该模型针对只包含一个主题的短文本语料数据建模效果良好，而针对包含几个主题的特殊语料时，假设较为强烈，主题建模效果较差。从本文使用语料数据来看，针对“厦门 BRT 爆炸”微博语料，每条微博只表达用户对该事件的一个关注点，而奥迪 A4L 论坛语料，每个主贴及其回复帖的用户会涉及该车型的几个方面，因此一元混合模型的关注点识别效果在针对特定话题的微博语料较好，而针对论坛数据则为最差模型。在短文本相似性方面，针对论坛语料主贴及回复贴都围绕主贴关注点进行阐述，因此短文本具有较强相似性，而微博语料“邻近”用户短文本主题也相似，但由于发表微博用户相互独立，短文本相似性较论坛语料弱，

表 2 “厦门 BRT 爆炸”话题各模型关注点一致性检验

关注点数量	3	4	5	6	7	8	9	10
LDA	−88.73	−89.49	−92.76	−93.31	−94.03	−96.86	−99.69	−101.36
A-LDA	−76.07	−78.54	−86.15	−88.35	−87.49	−90.07	−96.68	−97.32
AW-LDA	−61.20	−65.14	−68.15	−75.63	−83.80	−87.75	−88.16	−91.02
一元混合模型	−70.39	−76.38	−82.72	−85.01	−87.10	−88.01	−92.10	−93.48

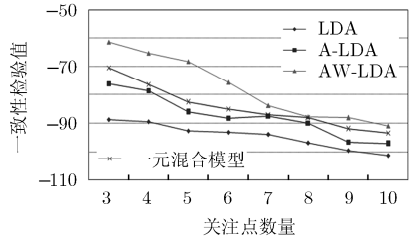

图 2 “厦门 BRT 爆炸”话题各模型关注点一致性检验

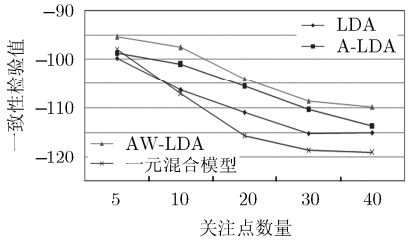

图 3 奥迪 A4L 讨论帖各模型关注点一致性检验

表 3 奥迪 A4L 讨论帖重点关注点

关注点	内容
关注点 1	奥迪、A4L、设计、新、车型、加速、公里……
关注点 2	检查、发动机、更换、建议、是否、系统、故障、正常、出现、启动、原因、工作、需要、造成……
关注点 3	喜欢、外观、比较、感觉、设计、空间、方面、大气、满意、够、外形、时尚、做工、好看……
关注点 4	性价比、配置、舒适、装、操控、比较、倒车、觉得、运动、方向盘、感觉、动力、A4L、多……
关注点 5	问题、档、声音、感觉、正常、踩、刹车、变速箱、解决、挡、油门……
关注点 6	导航、升级、膜、记录仪、轮毂、轮胎、胎、公里……
关注点 7	新款、优惠、提车、时候、卖、销售、配置、肯定、上市……
关注点 8	塑料、发动机、国内、事故、车辆、大众、公司、汽车、对方……
关注点 9	自动、灯、钥匙、大灯、功能、后视镜、手动、打开……

表 4 奥迪 A4L 讨论帖各模型关注点一致性检验

关注点数量	5	10	20	30	40
LDA	−99.76	−106.28	−110.90	−115.13	−115.07
A-LDA	−98.68	−101.04	−105.59	−110.26	−113.63
AW-LDA	−95.39	−97.53	−103.99	−108.59	−109.79
一元混合模型	−97.85	−107.03	−115.72	−118.53	−119.07

因此 A-LDA 模型建模奥迪 A4L 论坛语料效果优于针对特定话题的微博语料。

本文使用 SPSS 对关注点一致性结果进行显著性检验，由于一致性结果受到使用模型与关注点数量两方面影响，因此选用多因素方差分析，结果如表 5，表 6 所示。针对两组语料集选择关注点一致性指标效果最好的 2 个模型进行显著性分析，若通过检验，则剩余两模型也能通过显著性检验。针对微博语料选择本文所提 AW-LDA 模型与一元混合模型，针对论坛语料选择本文模型与 A-LDA 模型进行显著性检验。

从表 5，表 6 看出，校正模型显著性水平即 P 值均小于 0.05，表明两表中系数具有统计学意义。针对微博语料，模型变量的显著性水平为 0.006；针对论坛语料，模型变量的显著性水平为 0.004，两者都小于 0.05，说明所提 AW-LDA 模型与对比 Baseline 模型有显著性差异，因此能够接受本文所提模型效果优于其他模型的结论。

3.2 敏感度分析

为了验证 AW-LDA 模型的的鲁棒性，本文对参数 λ 进行了敏感度分析。图 4 呈现了奥迪 A4L 数据集 AW-LDA 模型不同 λ 设置的关注点一致性检验分析。如图 4 所示，本试验对 5 组不同主题数量进行一致性检验，选用模型为 AW-LDA 模型与 A-LDA 模型，每一组用相同颜色标出。权重影响力参数 λ 取值从 3~10 差值为 0.5 的等差数列，用横坐标标出。因为 A-LDA 模型没有权重设置项，因此该模型一致性检验值在图中呈现 5 条水平直线。

从图 4 可以看出，5 组不同主题数量的一致性检验，不论 λ 取何值，AW-LDA 模型的一致性检验曲线基本都高于对应的 A-LDA 一致性检验曲线，说明 AW-LDA 模型能够获得比 A-LDA 模型更好的主

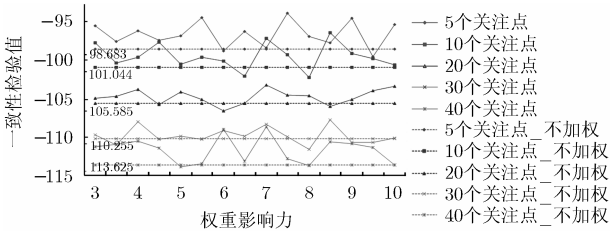

图 4 不同权重影响力设置的主题一致性检验对比分析

表 5 “厦门 BRT 爆炸”话题不同模型关注点一致性显著性检验

源	III 型平方和	自由度	均方	F 统计量	显著性水平
校正模型	1473.640[a)]	8	184.205	14.783	0.001
	104982.124	1	104982.124	8425.198	0.000
关注点数量	1289.035	7	184.148	14.779	0.001
所用模型	184.605	1	184.605	14.815	0.006
误差	87.223	7	12.460		
总计	106542.987	16			
校正的总计	1560.864	15			

a) R 方=0.944 (调整 R 方=0.880)

表 6 奥迪 A4L 讨论帖不同模型关注点一致性显著性检验

源	III 型平方和	自由度	均方	F 统计量	显著性水平
校正模型	337.594[a)]	5	67.519	118.355	0
	109091.027	1	109091.027	191228.525	0
所用模型	19.377	1	19.377	33.966	0.004
关注点数量	318.217	4	79.554	139.453	0
误差	2.282	4	0.570		
总计	109430.902	10			
校正的总计	339.875	9			

a) R 方=0.993(调整 R 方=0.985)

题识别效果，且具有一定的稳定性。同时也说明了直接合并短文本必然会引入一定的噪音，导致降低识别的效果。

4 结束语

研究社交网络特定话题用户关注点的识别具有重要理论价值与实践意义。本文提出一种基于混合权重合并策略的社交网络用户关注点识别方法,该方法将符合上下文相关条件的短文本进行虚拟合并，并根据其相关程度对不同短文本赋予不同的权重，能够针对特定话题的文本语料中识别不同的关注点。通过在 2 组数据集实验取得较好识别效果，并在关注点一致性检验中优于其他基准算法，同时也为短文本的主题建模提供了一种新的解决思路。今后的研究将面向特定话题关注点识别方法进行优化，并分析不同关注点的原因及其演化规律。

参 考 文 献

[1] YAN Zehua and LI Fang. News thread extraction based on topical n-gram model with a background distribution[C]. International Conference on Neural Information Processing, Berlin, 2011: 416–424. doi: 10.1007/978-3-642-24958-7_49.

[2] XING Chen, WANG Yuan, LIU Jie, *et al*. Hash tag-based sub- event discovery using mutually generative LDA in Twitter[C]. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, 2016: 2666–2672.

[3] ZHANG Xiaoming, CHEN Xiaoming, CHEN Yan, *et al*. Event detection and popularity prediction in microblogging [J]. *Neurocomputing*, 2015, 149(3): 1469–1480. doi: 10.1016/j.neucom.2014.08.045.

[4] BLEI D, NG A, and JORDAN M. Latent dirichlet allocation [J]. *Journal of Machine Learning Research*, 2003, (3): 993–1022.

[5] WENG Jianshu, LIM E, JIANG Jing, *et al*. Twitterrank: Finding topic-sensitive influential twitterers[C]. Proceedings of the Third ACM International Conference on Web Search and Data Mining, New York, 2010: 261–270. doi: 10.1145/1718487.1718520.

[6] PHAN X, NGUYEN L, and HORIGUCHI S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]. Proceedings of the 17th International Conference on World Wide Web, Beijing, 2008: 91–100. doi: 10.1145/1367497.1367510.

[7] ZHANG Heng and ZHONG Guoqiang. Improving short text classification by learning vector representations of both words and hidden topics[J]. *Knowledge-Based Systems*, 2016, 102(12): 76–86. doi: 10.1016/j.knosys.2016.03.027.

[8] VO D and OCK C. Learning to classify short text from scientific documents using topic models with various types of knowledge[J]. *Expert Systems with Applications*, 2015, 42(3): 1684–1698. doi: 10.1016/j.eswa.2014.09.031.

[9] JIN O, LIU N, ZHAO Kai, *et al*. Transferring topical knowledge from auxiliary long texts for short text clustering [C]. Proceedings of the 20th ACM International Conference on Information and Knowledge Management, New York, 2011: 775–784. doi: 10.1145/2063576.2063689.

[10] CHENG Xueqi, YAN Xiaohui, LAN Yanyan, *et al*. Btm: Topic modeling over short texts[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(12): 2928–2941. doi: 10.1109/TKDE.2014.2313872.

[11] ZUO Yuan, WU Junjie, ZHANG Hui, *et al*. Topic modeling of short texts: A pseudo-document view[C]. Proceedings of the 22nd ACM international Conference on Knowledge Discovery and Data Mining, San Francisco, 2016: 2105–2114. doi: 10.1145/2939672.2939880.

[12] LIN Hao, SUN Bo, WU Junjie, *et al*. Topic detection from short text: A term-based consensus clustering method[C]. Proceedings of the 13th International Conference on Service Systems and Service Management, Kunming, 2016: 1–6. doi: 10.1109/ICSSSM.2016.7538624.

[13] ZHAO Waynexin, JIANG Jing, WENG Jianshu, *et al*. Comparing twitter and traditional media using topic models[C]. Proceedings of the 33rd European Conference on Information Retrieval, Dublin, 2011: 338–349. doi: 10.1007/978-3-642-20161-5_34.

[14] 张华平. NLPIR 汉语分词系统[OL]. http://ictclas.nlpir.org/, 2016.3.

[15] MIMNO D, WALLACH H, TALLEY E, *et al*. Optimizing semantic coherence in topic models[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, 2011: 262–272.

姬建睿： 男，1986 年生，博士生，研究方向为社交网络分析、文本挖掘.

刘业政： 男，1965 年生，教授，博士生导师，主要研究方向为电子商务与商务智能、决策理论与方法、社会技术系统下的组织行为.

姜元春： 男，1980 年生，副研究员，硕士生导师，主要研究方向为电子商务与商务智能、个性化营销.