基于 RNN 和主题模型的社交网络突发话题发现

石磊，杜军平，梁美玉
（北京邮电大学智能通信软件与多媒体北京市重点实验室，北京 100876）

摘 要：社交网络数据是稀疏和嘈杂的，并伴有大量的无意义话题。传统突发话题发现方法无法解决社交网络短文本稀疏性问题，并需要复杂的后处理过程。为了解决上述问题，提出一种基于循环神经网络（RNN, recurrent neural network）和主题模型的突发话题发现（RTM-SBTD）方法。首先，综合 RNN 和逆序文档频率（IDF, inverse document frequency）构建权重先验来学习词的关系，同时通过构建词对解决短文本稀疏性问题。其次，模型中引入针板先验（spike and slab）来解耦突发话题分布的稀疏和平滑。最后，引入词的突发性来区分建模普通话题和突发话题，实现突发话题自动发现。实验结果表明与现有的主流突发话题发现方法相比，所提 RTM-SBTD 方法在多种评价指标上优于对比算法。

关键词：社交网络；突发话题发现；主题模型；循环神经网络
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018056

Social network bursty topic discovery based on RNN and topic model

SHI Lei, DU Junping, LIANG Meiyu
Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: The data is noisy and diverse, with a large number of meaningless topics in social network. The traditional method of bursty topic discovery cannot solve the sparseness problem in social network, and require complicated post-processing. In order to tackle this problem, a bursty topic discovery method based on recurrent neural network and topic model was proposed. Firstly, the weight prior based on RNN and IDF were constructed to learn the relationship between words. At the same time, the word pairs were constructed to solve the sparseness problem. Secondly, the “spike and slab” prior was introduced to decouple the sparsity and smoothness of the bursty topic distribution. Finally, the burstiness of words were leveraged to model the bursty topic and the common topic, and automatically discover the bursty topics. To evaluate the effectiveness of proposed method, the various experiments were conducted. Both qualitative and quantitative evaluations demonstrate that the proposed RTM-SBTD method outperforms favorably against several state-of-the-art methods.

Key words: social network, bursty topic discovery, topic model, RNN

1 引言

作为一个新形式的社会媒体，社交网络已经成为信息产生和传播的载体。社交网络经常第一时间发布和传播一些如自然灾害、暴恐事件、领导选举等重大的突发话题[1,2]。随着社交网络的日益普及以及它本身所具有的短小便捷、不受时间及地理限制、传播迅速等突出特性，导致社交网络成为新的舆情事件产生和民意反映的集聚地。从社交网络中发现突发话题，有利于引导公众舆论、控制网络谣

收稿日期：2017-11-08；**修回日期**：2018-02-03
通信作者：杜军平，junpingdu@126.com
基金项目：国家自然科学基金资助项目（No.61320106006, No.61532006, No.61772083）
Foundation Item: The National Natural Science Foundation of China (No.61320106006, No.61532006, No.61772083)

言。突发话题发现对于研究信息安全和话题演变也有重要的作用[3,4]。社交网络的文本特殊短小，并呈现稀疏性，如何从海量短文本中精确地提取高质量的话题是一个备受关注的问题；社交网络话题十分嘈杂和稀疏[5]，并且具有大量的日常聊天和无意义的话题，因此，区分突发话题和一般话题是十分必要的。

目前，突发话题发现的研究主要聚焦于利用主题模型对文本建模和利用聚类方法[6]，通过手工标注或对比设定阈值等后处理方式来发现突发话题。在基于主题模型的方法中，一般通过文档—词的共现来发现主题[7]。传统的主题模型方法主要被用于进行长文本建模，将其直接应用于社交网络短文本的突发话题发现中效果较差。尽管一些基于主题模型的变种方法能够解决短文本稀疏性问题[8,9]，但是它们都不能直接应用于突发话题发现，需要进行一些复杂的后处理操作。为了克服这些问题，研究者提出了在线主题模型[10]和时间主题模型[11]等方法，但它们存在手工标注等后处理问题，不仅效果一般，而且需要消耗较长的时间。

另一类方法主要通过突发词聚类来发现突发话题[12,13]。然而这些方法存在一些复杂的启发式后处理过程，且无法解决短文本稀疏性问题。这些方法发现的突发特征是嘈杂和稀疏的[11]，很难发现 2 个几乎同时产生的相似的突发话题。

为了解决上述问题，本文提出一种基于 RNN 和主题模型的社交网络突发话题发现方法（RTM-SBTD），可以适用于多种类型的社交网络平台，它不仅能够有效地发现突发话题，也能够有效地缓解短文本稀疏性问题；可以通过 RNN 网络记录之前观察到的词，有效地反映词关系的紧密性，并通过过滤高频词，构建权重先验来进一步优化主题建模，从而更好地完成突发话题发现；通过在模型中使用“spike and slab”先验来稀疏和解耦突发话题的稀疏和平滑，使发现的突发话题更加聚焦。

2 相关工作

突发话题发现的基础是计算文档之间的相似性。具体方法是预先设置关键词或突发词，通过计算词与词之间的相似度来监测突发话题，文档之间相似性常用的度量方法为夹角余弦。随着社交网络的发展，传统的事件监测方法已经不能很好地适用于社交网络这种特殊场景。当前突发话题发现的主流方法是基于主题模型的方法及其变种和基于聚类的方法等。

2.1 基于主题模型的突发话题发现

在基于主题模型的方法中，传统的主题模型设计的初衷是发现新闻事件的主题，并没有针对社交网络短文本进行考虑，因此，研究人员在传统主题模型的基础上进行了改进。Cheng 等[8]提出了双词话题模型（BTM, biterm topic model），有效地解决了社交网络短文本话题稀疏性问题，该方法将原始主题模型建模文档—主题和主题—词的过程转换为从文档直接建模双词，这种转换在一定程度上缓解了短文本稀疏性问题，同时，Yan 等[12]在此模型的基础上提出了突发话题模型（BBTM, bursty biterm topic model）实现了突发话题的发现。Quan 等[9]提出了自聚合主题模型（SATM, self-aggregation topic model）。SATM 不依赖外部信息，仅靠短文本自身的主题进行短文本自聚合。但是其参数个数随着数据增加而增加，这很容易导致它过拟合。

Hoffman 等[10]提出了在线主题模型技术，引入在线式的变分推断方法，实现了直接对大量在线数据流进行分析处理。Lau 等[13]提出了在线主题模型方法（OnlineLDA, online latent dirichlet allocation）实现话题发现，该方法在每个时间窗口中增量式地更新主题，并利用 Jensen Shannon 散度措施来监测主题的变动程度。Cao 等[14]结合隐狄利克雷分配（LDA, latent Dirichlet allocation）和深度学习方法提出了一个事件监测框架，有效地表示文档和词。Xie 等[15]在原始主题模型的基础上提出了基于简单图主题模型集合的实时突发事件监测算法。Gao 等[16]提出了一种新颖的分层贝叶斯主题模型，该模型采用了 N 元概念层次的潜在主题，通过该方法可以捕捉到一个词在局部语境下出现词的依赖关系，在多文档话题发现上取得了较好的效果。

2.2 基于聚类的突发话题发现

在基于聚类的方法中，通常根据语料的主题相似度把文档被聚集在一起，增量聚类用于新主题检测。Petrović 等[17]提出利用局部敏感散列（LSH)方法检测 Twitter 中的事件，在不失精度的前提下，该方法极大地提升了处理速度。其他类似的研究是利用字典学习方法来发现新主题。Becker 等[18]利用增量聚类方法从社交网络中监测突发事件。McMinn 等[7]利用倒排索引的每个命名实体及其相关近邻的聚类来实现突发事件监测和跟踪命名实体。Li 等[19]

提出了 Twevent 算法，通过 *n*-grams 分析 Twitter 中消息的内容特征，同时借助外部语料统计信息过滤掉不重要的特征。该方法采用外部知识库来过滤微博中大量的垃圾和噪音，是对 TwitterMonitor 框架的扩展。Huang 等[20]提出了一种新兴主题跟踪方法，它将时间观点的新兴词检测与空间视角的共现主题挖掘相结合。然而，这些方法需要复杂的启发式调整和处理，因为检测到的突发特征分散并存在较多噪声，这样不利于其聚类。

3　RTM-SBTD 方法

基于上述社交网络突发话题发现方法存在的问题，本文提出了基于 RNN 和主题模型的社交网络突发话题发现（RTM-SBTD）方法，其框架如图 1 所示。该框架包括 4 个部分，分别是数据预处理、基于 RNN 的先验知识学习、基于“spike and slab”先验的稀疏主题模型构建以及社交网络突发话题发现。假设一个突发性强的词可能是由一个突发话题所产生；相反，一个突发性弱的词更可能是由一个普通话题所产生。当突发话题出现时，与之相关的词可能比平时出现得更加频繁。例如，在新浪微博中“汶川地震”“马航飞机失联”等，在事件发生的时间段里，与之相关的词成了突发的高频词，这种高频词为突发话题的发现提供了至关重要的线索，因此，可以用词的突发性来指导突发话题的发现。

通过 RNN 来学习模型的先验知识，利用 RNN 来学习词之间的关系，并利用 IDF 过滤其中的高频词，利用 IDF 加权的形式结合 RNN 的输出构建模型的先验知识，同时借助基于“spike and slab”先验的弱平滑先验来解耦主题的稀疏和平滑。与传统话题模型通过建模文档的生成过程不同，RTM-SBTD 通过建模文档集合中每个词对的产生过程来学习话题。

本文假设每个词对中的 2 个词都是独立地从同一个话题中产生的，而该话题则是从一个全局的话题分布中产生。

3.1　数据预处理

采用的数据集来自爬取的新浪微博数据，并对数据进行以下处理：去除微博中重复的文档并过滤非中文文档，去除广告和噪音数据，去除了词数小于 3 的文档，去除出现次数小于 8 的词；切分词并去除停用词，按天将其分成 18 个时间段。

3.2　基于 RNN 的先验知识学习

BTM[8]、SATM[9]、BBTM[12]等解决短文本稀疏性的主题模型方法忽视了词之间的量化关系，但是这种量化关系在文本理解中非常重要，如果 2 个词关系紧密，则其出现在相同主题的可能性较大。而 RNN 已经被证明在句子生成过程能够有效地学习文本词之间的关系[21]。受 AMIRI 等[22]利用 RNN 解决短文的本表示问题启发，本文选择利用 Elman RNN 来学习词之间关系，并利用 IDF 过滤高频词。同时基于 RNN 的输出和 IDF 的结果构建加权的先验，加入到稀疏主题模型中，并同时重新构建词对的生成。基于 RNN 的先验知识学习框架如图 2 所示。在图 2 中，$w_t \in \boldsymbol{R}^T$ 表示当前词，T 是向量的长度，$s_t \in \boldsymbol{R}^S$ 是隐藏层，S 是隐藏层的大小，$y_t \in \boldsymbol{R}^N$ 是输出层，t 是当前的输入时间。$x_t = [w_t, s_{t-1}]$ 是输入层，其中，$x_t \in \boldsymbol{R}^{T+S}$，利用 x_t 可以计算隐藏层和输出层。

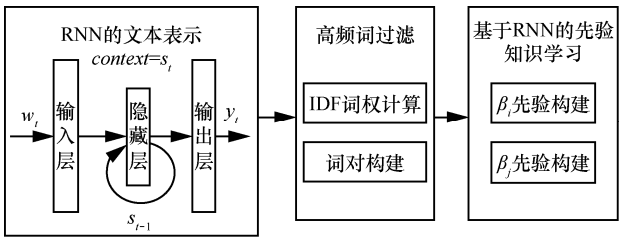


图 2　基于 RNN 的先验知识学习框架

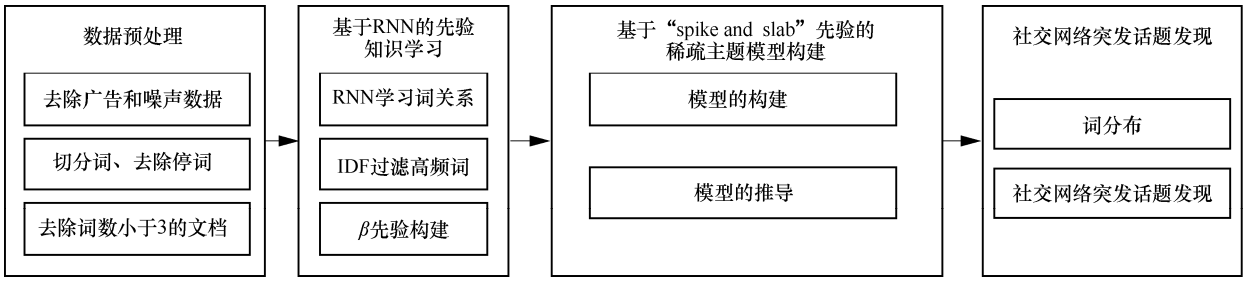


图 1　基于 RNN 和主题模型的社交网络突发话题发现框架

由于隐藏层 s_t 和 s_{t-1} 可以在时间 t 之前记忆所有观察到的词，因此，可以利用 RNN 来学习先前观察到的词和当前词之间的关系。定义 y_i 表示词对 w_i、w_j 的直接关系

$$y_i(j)=P\left(w_j \mid w_i, s_{t-1}\right) \tag{1}$$

其中，$y_i(j)$是 y_i 的第 j 个值，表示给定 w_i 时词 w_j 出现的概率。

为了消除高频词的影响，本文并没有采用类似 Xia 等[23]提出的删除部分常用词对的方法，因为在短文本存在稀疏性问题，删除大量的词对可能会使得文本的语义更加稀疏。本文利用 Lu 等[24]提出的基于逆向文件频率（IDF）的高频词过滤方法

$$IDF\ w_i=\lg\frac{\left|N_D\right|}{\left|d\in D: w_i\in d\right|} \tag{2}$$

其中，$\left|d\in D: w_i\in d\right|$表示词 w_i 出现在文档中的数量，w_i 在文档中出现的次数越多，IDF 的值越小，本文使用一个权值来降低 w_i 生成主题的概率。

定义 w_i 和 w_j 分别基于 RNN 的先验知识为

$$\beta_i=\tau\times y_i(j)\times IDFw_i \tag{3}$$

$$\beta_j=\tau\times y_i(j)\times IDFw_j \tag{4}$$

其中，τ用来避免 β 的值太小。

同时，通过扫描整个数据集来重新构建词对 $\left(w_i, w_j, IDF_{wi}, IDF_{wj}, y_i(j)\right)$。

3.3 基于“spike and slab”先验的稀疏主题模型的建立

利用“spike and slab”先验[25]加入到模型中来解耦突发话题的稀疏和平滑，在模型中引入伯努利变量先验以决定某个变量的开关状态。在所提 RTM-SBTD 方法中，稀疏主题模型中的开关变量指示主题是否是数据集的聚焦主题。

由于“spike and slab”先验可以产生空选择问题，这会引起概率分布的不明确，因此，利用弱平滑先验来解决这个问题[26]，通过该方法避免直接应用先验产生的分布不明确的问题，从而产生了更简单的推理过程，确保了提出的 RTM-SBTD 方法的可扩展性。

3.3.1 基于“spike and slab”先验的稀疏主题模型的定义

假设在一个时间片 T 内词 P 发生 n_w^t 次，由于某个词可能是正常使用，也有可能是突发的，因此，可以把一个正常的词分解成 2 个部分，表示词 W 正常的次数，$n_{w,1}^t$ 为产生突发的次数，因此，根据式 $\hat{n}_{w,1}^t=\max\left[\left(n_w^t-\bar{n}_w^t\right),\varepsilon\right]$ 可以获取到 $n_{w,1}^t$ 的估计值。$n_{w,0}^t+n_{w,1}^t=n_w^t$，其中，$n_{w,0}^t$ 和 $n_{w,1}^t$ 不能被观测到，ε 是一个比较小的正数，用来避免 0 值的出现。利用 w 的时间频率近似估计它们的值，估计方法如式(5)所示。

$$\mu_w^t=\frac{\max\left[\left(n_w^t-\bar{n}_w^t\right),\varepsilon\right]}{n_w^t} \tag{5}$$

其中，μ_w^t 表示词 W 在时间片内突发的概率，它表明词 W 在时间片 T 内出现的比其他时间内更加频繁，更有可能成为突发主题。

定义 1 突发话题和普通话题。突发话题的内容在当前时段内急剧增加，而普通话题的内容则是随时间的变化基本保持不变。

定义 2 主题选择器。给定一个短文本语料 $\boldsymbol{D}=\{d_1,d_2,\cdots,d_{N_d}\}$，一个主题选择器 b_z 是一个二进制开关变量，用来表示主题是否是 D 的聚焦主题，从伯努利分布中采样。

定义 3 平滑先验和弱平滑先验。平滑先验是狄里克雷超参数α用来平滑主题选择器的选择，弱平滑先验是另外一对狄里克雷超参数 $\bar{\alpha}$，用来平滑主题没有被选择。由于 $\bar{\alpha}<<\alpha$，所以超参数 $\bar{\alpha}$ 被称为弱平滑先验。

定义 4 聚焦主题。如果主题选择器 b_z=1，主题 K 是一个聚焦主题。数据集 A_z={$z: b_z=1,\ z\in\{1,\cdots,K\}$}被定义为聚焦主题。

3.3.2 基于“spike and slab”先验的稀疏主题模型推导

词的突发性与话题的突发性存在着直观的联系，由于某个词可以被观察到是正常使用或是突发的，RTM-SBTD 方法通过学习词的突发性来发现突发话题，并利用词的突发性来指导突发话题的发现。基于“spike and slab”先验的稀疏主题模型如图 3 所示。

定义一个二进制开关变量π来表示话题的生成情况。其中，π=0 表示词对 p 是从正常内容生成，而π=1 表示从突发主题生成。因此，可以利用概率先验知识编码突发话题的概率，利用参数为 μ_w^t 的 0~1 分布作为π的先验分布。$\boldsymbol{\theta}$ 表示语料上 K 个突发话题的分布，$\boldsymbol{\phi}_k$ 表示 K 个突发话题中词的分布，$\boldsymbol{\phi}_c$ 表示常态词分布（表示普通话题），并利用平滑

先验和弱平滑先验来聚焦主题。特别地，给定一个短文本数据 $\boldsymbol{D}=\{d_1,d_2,\cdots,d_{N_d}\}$，其对应的词对集合为 $\boldsymbol{P}=\{p_1,p_2,\cdots,p_{N_p}\}$，其中，$p_i=(w_{i,1},w_{i,2})$。基于“spike and slab”先验的稀疏主题模型的符号及其描述如表 1 所示。

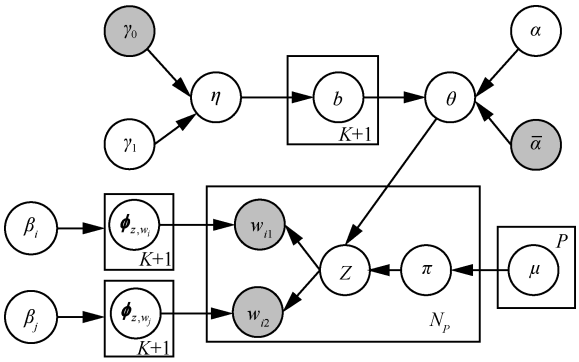


图 3　基于“spike and slab”先验的稀疏主题模型

表 1　　**符号及其描述**

符号	意义
D、N_p	短文本数据集、词对数量
K、 p	主题数量、词对集合
$\boldsymbol{\phi_c}$、$\boldsymbol{\theta_k}$	常态词分布、突发话题分布
b_z、 μ_w^t	主题选择器、词对的突发概率
Z、π	主题分配、二值开关变量
α、$\bar{\alpha}$	平滑先验、弱平滑先验
γ_0、γ_1	超参数
A_z	聚焦主题集合
$I[\cdot]$	示性函数

由于参数之间存在耦合关系，无法通过最大似然估计精确求解这 2 个参数，本文利用 Collapsed Gibbs 采样算法[27]来近似求解，其主要思想是交替地去对待估计的随机变量进行后验采样，每次对随机变量进行采样基于其他随机变量的赋值。模型的生成过程如下所示。

1) 对整个语料：采样 $\eta\sim Beta(\gamma_0,\gamma_1)$，采样主题选择器 $b_z\sim Bernoulli(\eta),\vec{b}=\{b_z\}_{k=0}^{K}$，采样突发话题分布 $\boldsymbol{\theta}\sim Dir(\alpha\vec{b}+\bar{\alpha}\vec{1})$。

2) 对每个突发话题：采样词分布 $\phi_{k,1}\sim Dir(\beta_i)$，采样词分布 $\phi_{k,2}\sim Dir(\beta_j)$，采样常态词分布 $\phi_{c,1}\sim Dir(\beta_i)$；$\phi_{c,2}\sim Dir(\beta_j)$。

3) 对每个词对：采样类别标识 $\pi\sim Bernoulli(\mu_w)$，若 π=0，采样词 $w_{i,1}\sim Multi(\phi_{c,1})$、$w_{i,2}\sim Multi(\phi_{c,2})$，若 π=1，采样一个突发话题 $z\sim Multi(\theta)$，采样词 $w_{i,1}\sim Multi(\phi_{z,1})$、$w_{i,2}\sim Multi(\phi_{z,2})$。

具体来说，在 RTM-SBTD 方法中需要对每个词对采样一个话题，采样一个话题开关变量 π，其中 $\boldsymbol{\theta}$ 是从“spike-and-slab”稀疏先验采样得到的。Dirichlet 超参数 α，Beta 分布超参数 γ_1 通过采样得到，$\bar{\alpha}$ 设置为 10^{-7}，γ_0 设置为 1。条件概率分布分别如式(6)和式(7)所示。

$$P(\pi=0\mid rest)\propto(1-\mu_i)\frac{n_{0,w_{i,1}}^{\neg i}+\beta_i}{n_{0,\cdot}^{\neg i}+W\beta}\frac{n_{0,w_{i,2}}^{\neg i}+\beta_j}{n_{0,\cdot}^{\neg i}+1+W\beta} \tag{6}$$

$$n_{0,w}\ P(\pi=1,z_i=k\mid rest)\propto\mu_i\frac{n_{\cdot}^{\neg i}+|A_z|\alpha+K\bar{\alpha}}{n_{\cdot}^{\neg i}+|A_z|\alpha+K\bar{\alpha}}\cdot \frac{n_{k,w_{i,1}}^{\neg i}+\beta_i}{n_{k,\cdot}^{\neg i}+W\beta}\frac{n_{k,w_{i,2}}^{\neg i}+\beta_j}{n_{k,\cdot}^{\neg i}+1+W\beta} \tag{7}$$

其中，$\pi=\{\pi_i\}_{i=0}^{N_p},Z=\{z_i\}_{i=0}^{N_p},\mu=\{\mu_i\}_{i=0}^{N_p}$，是词对分配给常态词分布的次数，$n_{0,\cdot}=\sum_{k=1}^{W}n_{0,w}$ 是分配给背景词分布的词总数，n_k 表示分配到突发话题的词对的总数，$A_z=\{z:b_z=1,z\in\{1,\cdots,K\}\}$ 是 $\vec{b}$ 状态为开的集合，$|A_z|$ 是 A_z 的基数，$n.=\sum_{k=1}^{K}n_k$ 表示分配到突发话题词的总数量，α是平滑先验，$\bar{\alpha}$ 表示弱平滑先验，$n_{k,w}$ 表示词 w 分配给突发主题的次数，$n_{k,\cdot}=\sum_{w=1}^{W}n_{k,w}$ 是分配给突发主题 K 的词的总数，$\neg i$ 表示排除词对 p。

采样主题选择器 b_z：为了采样 b_z，通过利用 η 作为一个辅助变量，式(8)为其联合分布。

$$P(\eta,\vec{b}_z\mid rest)\propto\prod_z P(b_z\mid\eta)P(\eta\mid\gamma_0,\gamma_1)\cdot \frac{I[B_l]\Gamma(|A_z|\alpha+K\bar{\alpha})}{\Gamma(n.+|A_z|\alpha+K\bar{\alpha})} \tag{8}$$

有了上述的联合条件分布，可以交替地根据 η 采样 $\vec{b}_z$ 以及根据 $\vec{b}_z$ 采样 η，最终得到 $\vec{b}_z$ 的结果。利用 Lin 等[26]提出的方法积分掉 η 来采样 $\vec{b}_z$，并用 Metropolis-Hastings 方法采样。对于参数 γ_1 采用基于 Gamma 先验的超参数采样方法获得，其中，$I[\cdot]$是一个示性函数。

3.4 社交网络突发话题发现

为了发现社交网络短文本中的突发话题，需要计算社交网络突发话题分布和词分布。随机地分配一个话题给每个词对作为初始状态。在每次的迭代过程中，根据式(6)~式(8)计算条件概率进行采样，经过充分多的迭代次数之后，开始收集统计量，逐个更新每个词对的话题类型标示变量与话题赋值。利用这些统计量可以估计各个参数，当迭代多次趋于稳定后，利用学习到的参数均值作为参数估计，最终得到社交网络突发话题分布和词分布结果，如式(9)~式(11)所示。

$$\theta_k = \frac{n_k^{-i} + b_z\alpha + \bar{\alpha}}{n_{\cdot}^{-i} + |A_z|\alpha + K\bar{\alpha}} \tag{9}$$

$$\phi_{k,w_i} = \frac{n_{k,w_i} + \beta_i}{n_{k,\cdot} + W\beta} \tag{10}$$

$$\phi_{k,w_j} = \frac{n_{k,w_j} + \beta_j}{n_{k,\cdot} + W\beta} \tag{11}$$

根据式(9)可以计算社交网络突发话题分布，根据式(10)和式(11)可以计算整个词分布 $\boldsymbol{\phi}_{k,w} = [\phi_{k,w_1}, \phi_{k,w_2}, \cdots, \phi_{k,w_n}]$。利用得到的社交网络话题分布和词分布可实现社交网络突发话题的发现。

3.5 RTM-SBTD 的实现步骤

RTM-SBTD 方法流程如算法 1 所示。

算法 1 RTM-SBTD 方法

输入 微博文本

输出 突发话题分布和词分布

步骤 1) 数据预处理（分词、去停词，去除字数少于 3 的文档）。

步骤 2) 利用 RNN 学习词的关系，得到式(1)。

步骤 3) 利用 IDF 过滤高频词，构建词对 $(w_i, w_j, IDF_{wi}, IDF_{wj}, y_i(j))$。

步骤 4) 利用式(3)和式(4)构建先验。

步骤 5) 通过稀疏主题模型建模区分建模普通话题和突发话题。

步骤 6) 根据式(9)到式(11)计算突发话题分布和词分布。

利用输入社交网络短文本数据，经过分词、去停用词、去除噪声数据预处理等操作，输入到 RNN 网络来学习词的关联关系，利用经典的 IDF 过滤高频词。结合 RNN 和 IDF 的输出构建词对，同时，构建加权的稀疏主题模型的先验 β_i 和 β_j，得到突发话题的分布和词分布。

4 实验结果与分析

本文利用新浪微博数据进行实验，采用的对比算法分别为 OnlineLDA、Twevent、SATM 和 BBTM。对本文提出的 RTM-SBTD 与对比算法在话题发现准确性、主题一致性、突发话题新颖度及突发话题成分判断等评价指标上进行比较，实验结果取其平均值。

4.1 数据集

从新浪微博爬取 2014 年 2 月 26 日到 2014 年 3 月 15 日的 40 多万条微博数据，其中包含了“马航飞机失联”等多个突发话题。

4.2 对比算法

OnlineLDA：一种典型的基于话题学习的突发话题发现方法[13]，该方法计算 2 个时间段内对应话题的词分布的 Jensen-Shannon 差异，若 Jensen-Shannon 差异大于一个阈值，则认为是一个突发话题。

Twevent：是目前主流的一个基于特征的突发事件检测方法[19]。该方法对微博进行切分，提取切分后的片段作为特征，计算特征的突发性，对突发性强的特征进行聚类，利用 Wikipedia 来过滤一些话题。

SATM：一种自聚合的短文本建模方法[9]。为了将其用于话题发现，利用 SATM 模型生成隐主题，根据余弦相似度贪婪地主题匹配，根据 Jensen-Shannon 差异标识主题词对的差异。

BBTM：在 BTM 模型的基础提出的一个突发话题发现模型[12]，引入二进制的开关变量，根据词的突发性来确定话题是否是突发话题。

4.3 参数设置

在本文实验中，时间片的长度设置为 1 天，$\alpha = 0.1$，$\bar{\alpha} = 10^{-12}$，$\beta = 0.01$，$\gamma_0 = 0.1$，Gibbs 采样过程中的迭代次数设为 500 次。突发话题 K 的数量设置为从 10 到 50（在话题发现质量实验中设置为 5 到 30）。

4.4 RTM-SBTD 与对比算法在话题发现准确度上的比较与分析

为了保证评判的无偏性，将各方法发现的突发话题混合，随机进行人工标注。同时邀请 6 个志愿者对该实验结果进行标注。对于每个突发话题提供的信息包括：日期、概率最大的 10 个词及在该时间段内与该突发话题最相关的 50 条微博。标注者

可以使用 Google、百度、新浪微博搜索等工具来辅助判断。突发话题的认定标准是：当有一半以上志愿者都认为该话题是一个突发话题时，则认为该话题是一个准确的突发话题。计算不同的突发话题数目 K 对应的精度（P@K）来评价各方法对突发话题发现的准确度。实验结果如表 2 所示。

表 2　RTM-SBTD 与对比算法在话题发现准确度上的比较

方法	P@10	P@30	P@50
RTM-SBTD	0.803	0.822	0.829
OnlineLDA	0.228	0.213	0.186
BBTM	0.720	0.732	0.724
SATM	0.563	0.478	0.449
Twevent	0.711	0.725	0.689

表 2 给出了不同主题数下发现突发话题的准确度，从实验结果可以看出：本文所提 RTM-SBTD 方法的准确度大于 0.8，明显优于其他方法，说明了所提方法能够比较准确地发现突发话题。

对比不同突发话题 K 下的实验结果，发现 RTM-SBTD 在 K=10 的时候话题发现准确度效果稍差，主要是因为主题数目太少，导致主题较为分散。BBTM 的效果优于其他 3 种基准方法，但相比本文提出的 RTM-SBTD 方法，BBTM 相对较差，这说明 RTM-SBTD 通过 RNN 建模词对之间的关系以及利用“spike and slab”来解耦主题的稀疏和平滑，对突发话题发现的准确度有很大提高。Twevent 效果优于 OnlineLDA 和 SATM，这是由于 Twevent 方法考虑了突发事件的特征，同时对其进行聚类，使得突发事件主题比较集中。

2 种采用普通话题模型的方法 OnlineLDA 和 SATM 的话题发现准确度都不高，这是因为普通话题模型没有直接考虑话题的突发性，不能较好地区别普通话题和突发话题。SATM 的话题发现准确度要稍高于 OnlineLDA，原因是 SATM 在通过自聚合的方式缓解了短文本稀疏性问题，故其建模短文本的性能及话题发现的能力要好于 OnlineLDA。

4.5　RTM-SBTD 与对比算法在话题发现的新颖度上的比较与分析

在社交网络上突发话题是持续变化的，因此，本文引入新颖度（Novelty）[12]作为评价指标，用以评估不同算法发现突发性话题的敏感性和新颖性。从主题 Z 中搜集 T 个可能的突发话题的主题词，在每个时间片内构建主题词集合，$W^{(t)}$ 和 $W^{(t-1)}$ 表示 2 个相邻时间片内的突发词的集合。

$$Novelty\left(Z^{(t)}\right)=\frac{\left|W^{(t)}\right|-\left|W^{(t)}\cap W^{(t-1)}\right|}{T\cdot K} \tag{12}$$

其中，$|\cdot|$ 表示集合中元素的数量，T 表示每个主题中包含词的数量。

图 4 展示了不同算法突发话题发现的话题发现的新颖度的结果。随着突发主题数的变化，突发话题的新颖度的变化较为明显。所提 RTM-SBTD 方法和 BBTM 方法都明显优于 OnlineLDA 和 SATM，尤其是当 K 值较大时表现较好，这表明 RTM-SBTD 方法在微博突发话题发现的能力优于传统的时间主题模型。

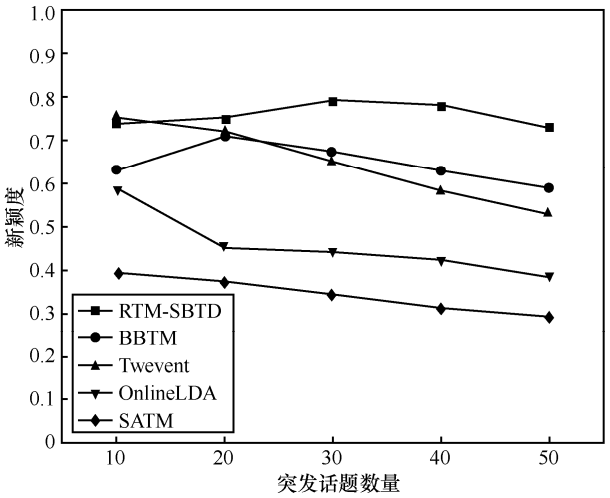


图 4　RTM-SBTD 与对比算法在话题发现的新颖度的比较

当 K 值较小时 Twevent 获得了较好的效果，因为它只通过突发词的变化来监测突发话题。然而，随着主题数 K 不断增加时，效果下降明显，这是因为该方法存在较多小的聚类词，使其效果下降。BBTM 的效果优于 Twevent，这是因为 BBTM 模型通过建模词对有效地提高了模型处理短文本和发现话题的能力，因此，取得了较好的效果。OnlineLDA 的话题发现的新颖度优于 SATM，这是因为 OnlineLDA 是一种基于时间的话题模型，对于监测话题的变化比较敏感，而 SATM 只是一种基于自聚合的短文本话题模型，不能敏锐地监测话题的变化。

4.6　RTM-SBTD 与对比算法在主题一致性上的比较与分析

为了验证话题的一致性，本文使用点对互信息

（PMI, pointwise mutual information）方法来评估提出方法的主题一致性[28]。同时列举概率最大的前 10 个词来定性分析主题的一致性。

通过给定一个主题 z，选择前 N 个可能的词 w_1，w_2,…,w_N，计算每个词的 PMI。通过借用大规模外部数据来计算一个话题中概率最大的前几个词相互之间的平均 PMI。PMI 越高，说明词越相关，可解释性越好，如式(13)所示。

$$\phi_{z,w_i}\ PMI(z)=\frac{2}{N(N-1)}\sum_{1\leqslant i\leqslant j\leqslant N}\mathrm{lb}\frac{p(w_i,w_j)}{p(w_i)p(w_j)} \quad (13)$$

其中，$p(w_i,w_j)$ 是词对 w_i 和 w_j 在相同滑动窗口同时出现的联合分布，$p(w_i)$ 是词 w_i 出现在滑动窗口内的边缘概率分布。

为进一步评价这些突发话题的 PMI，采用从维基百科官网下载的中文维基百科文章作为辅助语料。计算每个突发话题中前 10 个词的平均 PMI 分数。图 5 展示了各突发话题发现方法学习到的突发话题的主题一致性结果。

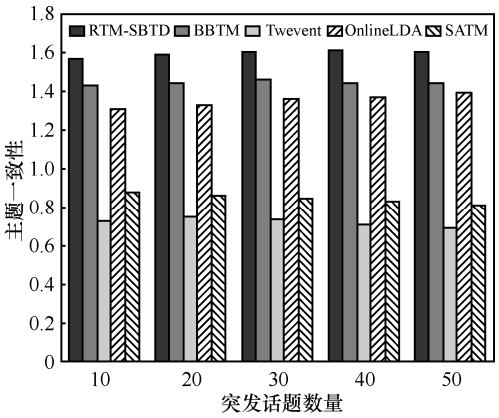


图 5 RTM-SBTD 与对比算法在主题一致性上的比较

可以发现 RTM-SBTD 的主题一致性实验结果明显优于其他对比方法，BBTM 也取得了较好的效果，但比本文所提算法的效果稍差。RTM-SBTD 的主题一致性实验结果显著高于 OnlineLDA 和 Twevent，说明 RTM-SBTD 发现突发话题的质量较高。BBTM 和 SATM 较高是因为可以通过不同方式解决短文本的稀疏性问题。但是由于 SATM 的设计是基于短文本建模，并未考虑突发话题，所以发现的话题可能不是突发话题。另外，还发现 Twevent 发现的突发话题的可读性随着突发主题数的增加显著下降，这是由于 Twevent 中排在后面的突发话题中的词数越来越少，所以聚类效果相对较差。

为了定性评估主题一致性，本文随机挑选了一个突发性较强且高频的话题标签（hashtag），即“#马航飞机失联#”，其中“#马航飞机失联#”对应于 2014 年 3 月 8 日发生的事件。抽取包含该 hashtag 的微博，并统计出其中的词频并归一化，视为一个 hashtag 对应的话题，从每个方法的发现突发话题中找出对应概率最大的前 10 个词。表 3 列出了各方法中概率最大的 10 个词。

表 3 基于#马航飞机失联#发现的概率最大的前 10 个词

算法	前 10 个词
RTM-SBTD	飞机、乘客、马航、失联、MH370、遇难、客机、平安、声明、祈祷
BBTM	客机、击落、飞机、坠毁、马航、服务、俄罗斯、乘客、中国、平安
Twevent	马来西亚、乌克兰、恐怖、贵宾厅、航班、天气、公司、护照、艾滋病、绝望
OnlineLDA	北京、入境处、乘务员、MH370、护照、消息、日本、马航、报道、事件
SATM	祈福、飞机、安息、手机、购物、旅游、华为、北京、马航、机场

从表 3 的实验结果可以看出，RTM-SBTD 中的词和该 hashtag 的事件的关键词较为相关。BBTM 发现的话题关键词也十分相关，但也有若干不相关的词。Twevent 中词比较大众化词，同时还包含一些不相关的词，说明突发词聚类对噪音比较敏感。OnlineLDA 中的词以常见词为主，而且只是部分词与 hashtag 相关，因此，相似度最低。SATM 的结果和 OnlineLDA 类似，包含了较多的常见词，表明基本主题模型不能较好地区分出突发话题和普通话题。

4.7 RTM-SBTD 与对比算法在话题发现质量上的比较与分析

利用聚类纯度和熵作为评价指标来判断突发话题发现的质量。纯度和熵是聚类的 2 种标准评价方法。纯度计算每个类中占主导地位的类别的比率，其中较大的值意味着更好的性能。熵用于测量一组数据中的混沌，因此，较小的熵值表示更好的性能。

在实验中，首先从第 2 天~18 天的微博中选择每天出现次数超过平均每天出现次数的 2 倍的 hashtag。选择 6 个高频且意义比较明确的 hashtag 作为测试集中话题的类别标签。随机抽取 $\frac{1}{10}$ 数据删除标签作为测试集。对于 OnlinLDA、BBTM 和

SATM，把每个突发话题设为一个类，把每条微博 d 赋予给 P(π==1|d)的类；对于 Twevent，按类与消息之间的 Jaccard 系数把话题赋给最相似的类。

表 4 和表 5 展示了不同主题数下发现突发话题聚类的实验结果。设置 K 值从 5~30 变化。可以看出提出的 RTM-SBTD 方法在聚类纯度和熵指标上优于其他对比算法。BBTM 也获得了较好的效果，但比本文所提 RTM-SBTD 方法稍差，这是因为 RTM-SBTD 利用 RNN 能够提前学习到词之间的关系，并通过过滤高频词，降低高频词对突发话题发现的影响，应用弱平滑先验能够使主题更加聚焦。所有对比方法中 Twevent 的效果最差，这是因为该方法只利用突发词来表达突发话题，难以全面准确地判断突发话题和整个话题之间的相似度。

表 4　RTM-SBTD 与对比算法在话题发现聚类纯度上的比较

算法	K=5	K=10	K=15	K=20	K=25	K=30
RTM-SBTD	0.289	0.412	0.489	0.520	0.553	0.559
OnlineLDA	0.264	0.363	0.422	0.461	0.475	0.486
BBTM	0.271	0.387	0.457	0.478	0.489	0.515
Twevent	0.173	0.184	0.187	0.226	0.213	0.209
SATM	0.232	0.359	0.438	0.459	0.467	0.471

表 5　RTM-SBTD 与对比算法在话题发现聚类熵上的比较

算法	K=5	K=10	K=15	K=20	K=25	K=30
RTM-SBTD	0.109	0.092	0.089	0.088	0.085	0.084
OnlineLDA	0.111	0.097	0.096	0.094	0.093	0.094
BBTM	0.110	0.094	0.093	0.090	0.088	0.087
Twevent	0.122	0.124	0.125	0.126	0.128	0.131
SATM	0.113	0.099	0.097	0.095	0.094	0.095

5　结束语

由于社交网络内容具有天然的稀疏性且伴有大量的无意义的话题，从这种场景中发现突发话题是一个挑战。为了解决这个问题，本文提出了基于 RNN 和主题模型的社交网络突发话题发现方法，利用 RNN 来学习词之间的关系，并作为主题模型的先验知识，同时在主题模型的基础上利用“spike and slab”先验来解耦主题的稀疏和平滑，消除不相关主题，使主题更加聚焦，并用词的频率变化作为先验来指导突发话题的发现。RTM-SBTD 方法不仅可以有效地解决社交网络短文本主题建模的数据稀疏问题，而且还可以有效地发现突发话题。实验结果表明 RTM-SBTD 方法的突发话题发现能力显著优于基准方法。但是社交网络中还存在大量的图像内容，而 RTM-SBTD 方法还不能有效地建模多模态数据，在下一步工作中将进一步引入社交网络中的视觉信息，实现基于跨模态主题模型的突发话题发现。

参考文献：

[1] 方滨兴, 贾焰, 韩毅. 社交网络分析核心科学问题、研究现状及未来展望[J]. 中国科学院院刊, 2015(2):187-199.
FANG B X, JIA Y, HAN Y. Social network analysis-key research problems, related work, and future prospects[J]. Bulletin of Chinese Academy of Sciences, 2015(2):187-199.

[2] 贾焰, 甘亮, 李爱平. 社交网络智慧搜索研究进展与发展趋势[J]. 通信学报, 2015, 36(12):9-16.
JIA Y, GAN L, LI A P. Research progress and development trend of online social network smart search[J]. Journal on Communications, 2015, 36(12):9-16.

[3] 王晓阳, 郑骁庆, 肖仰华. 智慧搜索中的实体与关联关系建模与挖掘[J]. 通信学报, 2015, 36(12):17-27.
WANG X Y, ZHENG X Q, XIAO Y H. Entity-relation modeling and discovery for smart search[J]. Journal on Communications, 2015, 36(12):17-27.

[4] 黄河燕. 在线社交网络的可视化分析[J]. 中国科学院院刊,2015,(2):229-237.
HUANG H Y. Visual analysis of online social networks[J]. Bulletin of Chinese Academy of Sciences, 2015, (2):229-237.

[5] 唐杰, 陈文光. 面向大社交数据的深度分析与挖掘[J]. 科学通报, 2015, 60(5):509-519.
TANG J, CHEN W G. Deep analytics and mining for big social data[J]. Chinese Science Bulletin, 2015, 60(5):509-519.

[6] WANG Y, LIU J, HUANG Y. Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(7): 1919-1933.

[7] MCMINN A J, JOSE J M., Real-time entity-based event detection for Twitter[C]//International Conference of the Cross-Language Evaluation Forum for European Languages. 2015: 65-77.

[8] CHENG X, YAN X, LAN Y. BTM: topic modeling over short texts[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(12):2928-2941.

[9] QUAN X, KIT C, GE Y. Short and sparse text topic modeling via self-aggregation[C]//International Conference on Artificial Intelligence. 2015: 2270-2276.

[10] HOFFMAN M D, BLEI D M, BACH F. Online learning for latent dirichlet allocation[C]//International Conference on Neural Information Processing Systems. 2010:856-864.

[11] STILO G, VELARDI P. Efficient temporal mining of micro-blog texts and its application to event discovery[J]. Data Mining and Knowledge Discovery, 2016, 30(2): 372-402.

[12] YAN X, GUO J, LAN Y. A probabilistic model for bursty topic discovery in microblogs[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence.2015:353-359.

[13] LAU J H, COLLIER N, BALDWIN T. On-line trend analysis with topic models: # twitter trends detection topic model online[C]// COLING. 2012:1519-1534.

[14] CAO Z, LI S, LIU Y. A novel neural topic model and its supervised extension[C]//29th AAAI Conference on Artificial Intelligence. 2015: 2210-2216.

[15] XIE W, ZHU F, JIANG J. Topicsketch: real-time bursty topic detection from twitter[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(8): 2216-2229.

[16] GAO Y, WEN D, CHEN NS. A novel contextual topic model for multi-document summarization[J]. Expert Systems with Applications, 2015, 42(3): 1340-1352.

[17] PETROVIĆ S, OSBORNE M, LAVRENKO V. Streaming first story detection with application to twitter[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010:181-189.

[18] BECKER H, NAAMAN M, GRAVANO L. Beyond trending topics: real-world event identification on twitter[J]. ICWSM, 2011(11): 438-441.

[19] LI C, SUN A, DATTA A. Twevent: segment-based event detection from tweets[C]//ACM International Conference on Information and Knowledge Management.2012:155-164.

[20] HUANG J, PENG M, WANG H. A probabilistic method for emerging topic tracking in Microblog stream[J]. World Wide Web-Internet & Web Information Systems, 2016, 20(2):1-26.

[21] SUTSKEVER I, MARTENS J, HINTON G E. Generating text with recurrent neural networks[C]//International Conference on Machine Learning.2011:1017-1024.

[22] AMIRI H, DAUMÉ III H. Short text representation for detecting churn in microblogs[C]//AAAI.2016:2566-2572.

[23] XIA Y, TANG N, HUSSAIN A. Discriminative bi-term topic model for headline-based social news clustering[C]//FLAIRS Conference. 2015: 311-316.

[24] LU H, XIE L Y, KANG N. Don't forget the quantifiable relationship between words: using recurrent neural network for short text topic discovery[C]//AAAI.2017:1192-1198.

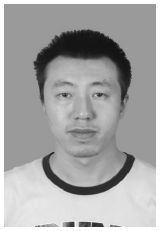
[25] WANG C, BLEI D M. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process[C]//Advances in neural information processing systems.2009:1982-1989.

[26] LIN T, TIAN W, MEI Q. The dual-sparse topic model:mining focused topics and focused terms in short text[C]//International Conference on World Wide Web. 2014: 539-550.

[27] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. The National Academy of Sciences, 2004, 101(1): 5228-5235.

[28] NEWMAN D, LAU J H, GRIESER K. Automatic evaluation of topic coherence[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.2010:100-108.

[作者简介]



石磊（1986-），男，内蒙古突泉人，北京邮电大学博士生，主要研究方向为人工智能、数据挖掘、社交网络搜索。

杜军平（1963-），女，北京人，博士，北京邮电大学教授、博士生导师，主要研究方向为人工智能和数据挖掘。

梁美玉（1985-），女，山东泰安人，北京邮电大学副教授、硕士生导师，主要研究方向为信息搜索、数据挖掘、智能信息处理和计算机视觉。