基于情感增强非参数模型的社交媒体观点聚类

刘勘
中南财经政法大学
信息与安全工程学院
/ 武汉430073
liukan@zuel.edu.cn

陈昱
中南财经政法大学
信息与安全工程学院
/ 武汉430073
chen997@stu.zuel.edu.cn

何佳瑞
中南财经政法大学
信息与安全工程学院
/ 武汉430073
hejiarui@stu.zuel.edu.cn

摘要

本文旨在使用文本聚类技术，将社交媒体文本根据用户主张的观点汇总，直观呈现网民群体所持有的不同立场。针对社交媒体文本模式复杂与情感丰富等特点，本文提出使用情感分布增强方法改进现有的非参数短文本聚类算法，以高斯分布建模文本情感，捕获文本情感特征的同时能够自动确定聚类簇数量并实现观点聚类。在公开数据集上的实验显示，该方法在多项聚类指标上取得了超越现有模型的聚类表现，并在主观性较强的数据集中具有更显著的优势。

关键词： 观点分析；短文本流聚类；非参数模型；社交媒体

A Sentiment Enhanced Nonparametric Model for Social Media Opinion Clustering

LIU Kan
School of Information
and Safety Engineering,
Zhongnan University of
Economics and Law
/ Wuhan 430073
liukan@zuel.edu.cn

CHEN Yu
School of Information
and Safety Engineering,
Zhongnan University of
Economics and Law
/ Wuhan 430073
chen997@stu.zuel.edu.cn

HE Jiarui
School of Information
and Safety Engineering,
Zhongnan University of
Economics and Law
/ Wuhan 430073
hejiarui@stu.zuel.edu.cn

Abstract

Based on text clustering techniques, this paper aims to aggregate social media texts according to the different opinions claimed by users. To address the characteristics of short length, large number and complex patterns of social media texts, this paper proposed Sentiment Distribution Enhanced (SDE) method to improve the existing nonparametric-based clustering algorithm. We model text sentiment with a Gaussian distribution, the proposed method automatically determines the number of clusters and achieves opinion clustering while capturing sentiment features. Experiments on public datasets show that the method achieves clustering performance that surpasses existing models on multiple clustering metrics, and has more significant advantages in datasets with strong subjectivity.

Keywords: opinion analysis , short text stream clustering , nonparametric model , social media

1 引言

基金项目：国家自然科学基金(72174156)

社交媒体作为自由互联的平台，有着大量的观点从中产生及传播，吸引到了各领域的众多关注。微观层面上，这些观点表现于用户的各类互动行为中，表达了其态度和立场，从宏观的角度看，某事件中的观点集合在一定程度上反映出了舆情。充分挖掘、分析社交媒体平台中丰富的观点信息对网络舆情的引导和治理有着积极的意义。观点分析的研究主要围绕观点倾向性展开，即要求算法能够自动地判别用户发表言论是正面、负面还是中立的观点(Chen et al., 2020)。然而，用户对于事件或话题的观点往往具有更丰富的内容，仅分析观点的倾向不足以全面了解用户的态度和立场。因此，从用户主观言论中总结并提取核心观点成为观点分析亟待解决的问题之一。尤其在社交媒体平台的热点事件或话题中，存在着表达了不同主张的海量用户发言，将其按照观点类别正确划分有助于厘清事件态势，并能从相应的角度直观分析民众诉求和关注点。但是，在归类用户观点的过程中面临着以下困难：（1）用户发言文本长度短，处理长文档的文本挖掘技术难以应用；（2）文本数量动态增长，需要采用能够无限增量处理数据的模型；（3）用户观点模式复杂且无法预先估计，获取实际数据标签成本较大，传统的监督学习方法不再适用。因此，观点挖掘多作为一种聚类问题进行分析。但是普通的文本聚类方法，如Zhao等(2011)基于隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）提出的Twitter-LDA等需要事先设定聚类数量的方法并不符合社交媒体真实情况。

面对上述难点，诸多研究者试图从短文本数据流聚类的方向上解决社交媒体观点聚类问题。其中较经典的研究有Shou等(2013)针对社交媒体的特性设计了Sumblr系统不断监听主题相关推特数据流并进行聚类和摘要，为用户提供主题相关的公众观点总结。还有Yin等(2016)基于非参数贝叶斯模型（Bayesian nonparametric model）提出了GSDPMM，从词的共现特征（co-occurrence）出发，通过概率生成过程灵活地推断聚类数量并完成聚类。

现有的短文本流聚类研究大多仅通过短文本的词特征实现聚类，而社交媒体数据具有更多的特征，比如用户情感就是其中需要重点关注的特征之一。有许多观点分析的研究者假设用户所持看法、立场与其发表文本内容所蕴含的情感具有较强的联系，甚至通过直接或间接的方式将文本情感与观点二者联系起来(O'Connor et al., 2010; 高俊峰and 黄微, 2019)。因此，为了更好地实现社交媒体用户观点聚类，本文使用情感分析的方式挖掘文本观点特征，提出情感分布增强（Sentiment Distribution Enhanced, SDE）方法改进狄利克雷过程混合模型（Dirichlet Process Mixture Model, DPMM）聚类算法。该算法基于非参数模型，使用高斯分布建模文本情感以捕获社交媒体用户言论中的主观信息，融合使用词-情感生成分布实现聚类。

总结本文的贡献有：（1）基于非参数概率模型，提出使用高斯分布建模情感信息以增强聚类算法，解决社交媒体观点聚类难点。（2）在原有的坍缩吉布斯采样（collapsed Gibbs sampling）算法基础上，给出了情感高斯分布的后验参数更新方式以及预测分布（Predictive Distribution）计算。（3）实验证明，SDE方法相比于现有的前沿算法在聚类表现上有所提升，并在主观性较强的社交媒体数据集上取得了更显著的表现。

2 相关研究

观点分析是社交媒体研究领域的经典问题，已有诸多学者对该问题从不同方面进行了探索，本节除了介绍常用的观点分析模型，还重点从短文本数据流的角度介绍社交媒体的用户观点聚类研究。

2.1 观点分析

观点分析，又称观点挖掘或文本意见挖掘，对于社交媒体这一关键的网络舆论平台有重要的现实意义，因而受到了众多研究者的关注。现有研究大多旨在判别用户所发表言论的观点极性(Benkhelifa and Laallam, 2018; Wu et al., 2020)。对观点进行极性分析，其优点是可以转化为分类问题从而采用有监督学习模型，准确率较高，但缺点在于仅从态度倾向或情感标签的角度分析过于简单，难以呈现观点的丰富内涵。由此，有学者使用文本聚类的方法挖掘同类观点，分析民众对于热点事件或话题的各种态度(Ni et al., 2018)。李秀霞等人(2016)针对此问题采取"密度-距离"的快速搜索聚类算法进行共词聚类，该算法可以使数据自动确定聚类中心和数目而不需要人工设定。然而，当新数据到来时此算法也必须重新在完整数据集上运行，而且聚类过程受事先设定的密度阈值影响较大，无法很好地适应社交媒体数据持续增长、分布复杂的实际情况。

2.2 短文本流聚类

面对社交媒体用户观点聚类中的挑战，短文本数据流聚类方法显示出更强的现实场景适应性(Aggarwal, 2013; Nguyen et al., 2015)，其中的工作可分为基于相似度和基于模型的方法。

2.2.1 基于相似度的方法

基于相似度的短文本数据流聚类原理是将文本根据特征映射到向量空间中，并设计向量的相似性度量，在扫描文本数据时计算向量相似度，当对应的相似度大于设置阈值时聚为一类。这种自聚合的聚类机制无需人工设定聚类数量，如Geng等(2020)选择将社交媒体短文本按照其词频、长度等统计信息表示为向量，再计算相应的文本与聚类之间相似度，最终实现聚类。Rakib 等人(2021)改进了此类聚类算法的计算过程，在聚类时通过动态采样一部分聚类计算相似度而无须遍历所有聚类，一定程度上减少了计算成本。

这类基于向量和阈值的方法虽运算较快，但需要在不同数据集上搜索得出最佳阈值，且难以适应数据分布变化较大的数据集。因而该类方法在时间跨度较大的动态数据集中表现不佳。

2.2.2 基于模型的方法

基于模型的短文本数据流聚类方法假设短文本是由概率模型所生成的，文本组成的词汇由聚类相应的词分布抽样得来。其中具有代表性的是Yin等(2014)首先提出的基于狄利克雷过程的混合模型GSDPMM，该模型利用短文本的单词共现信息完成聚类，假设文本之间包含相同的词越多则越可能属于同一类。

基于狄利克雷过程混合模型的短文本聚类方法进一步地克服了固定阈值这一缺陷，能够根据数据分布情况灵活推断聚类数量并完成聚类。很多学者都是在狄利克雷过程混合模型的框架上不断创新，如Yin等(2018)改进了推断算法，使其能以“一遍扫描”（one-pass）的方式处理数据，无需多次迭代使聚类算法收敛，相较于需要迭代求解的GSDPMM更符合社交媒体的观点聚类。Kumar等(2020)深入挖掘了文本特征，在词共现关系之外寻得词的重要性特征，提升了聚类算法的表现。Li等(2016)利用词嵌入（word embedding）表征文本之间的深层相似关系。

短文本流聚类主要以文本与词的统计信息来衡量文本与聚类的相似性，而实际在社交媒体的观点表达中还包含了用户丰富的情感信息。因此本文提出了情感分布增强方法，为社交媒体短文本情感建立概率分布并将其加入生成模型中，与聚类-词的多项式分布共同作为文本的联合生成分布，在推断混合模型参数的同时得到用户观点聚类。

3 问题描述

3.1 概念定义

本文的研究旨在设计一种短文本数据流聚类算法将社交媒体中表达相似观点的用户言论聚合为一类，以直观地分析总结网民看待事件的不同角度。从形式上定义短文本数据流，其中$D=\{d_1,d_2,\ldots,d_\infty\}$是无限长度的数据序列，正如社交媒体平台上不断增加的海量用户发言。其中d_i是长度为$|d_i|$的用户发表文本，由其所包含的词$W^{(i)}=\left\{w_1^{(i)},w_2^{(i)},\ldots,w_{|d_i|}^{(i)}\right\}$组成。同时，每条言论$d_i$都有唯一对应的聚类（即观点）编号$c_i$。本文的最终目标是不断地将社交媒体用户言论分配到对应的聚类$z_k=\{d_{k1},d_{k2},\ldots,d_{k\infty}\}$，使得持同一种观点的文本汇总到一起，即$z_k$内文本聚类编号满足$c_{k1}=c_{k2}=\cdots=k$。此时观点聚类数量是无限的，这是因为随着相关事件或话题发展，新的观点有可能不断出现，但通常观点数量远小于评论数量，有$|Z|\ll|D|$。同时，假设用户评论为短文本，一条文本仅属于一个聚类簇，当$i\neq j$时有$z_i\cap z_j=\emptyset$，若是实际评论长度较长，可以通过拆分等方式处理。

3.2 狄利克雷过程混合模型文本聚类

狄利克雷过程（Dirichlet Process, DP）是一种随机过程，它每次抽样的结果都是一个概率分布。狄利克雷过程在非参数贝叶斯模型中被广泛运用，常作为混合模型的先验，形成狄利克雷过程混合模型(Li et al., 2019)。将混合模型应用于聚类过程中就可以自动地从数据推断聚类簇的数量，无需人工指定类别。

假设观测数据$X=x_1,x_2,\ldots,x_n$相互独立，来自一个具有K个未知形式成分的混合分布，记为$\mathrm{F}(\Phi)$，其中$\Phi=\phi_1,\phi_2,\ldots,\phi_k$是各成分的参数集合，因此有$p\left(X\mid\phi_1,\cdots,\phi_K;\pi_1,\cdots,\pi_K\right)=\sum_{i=1}^{K}\pi_i\,\mathrm{F}\left(X\mid\phi_i\right)$，$\pi_i$代表第i个分布在混合模型中的权重，

满足$\sum_{i=1}^{K}\pi_i=1$。然而在社交媒体等许多现实情况中，只存在用户言论可以作为观测数据，却无从预测可能的分布数量K并推断分布参数和权重。为此，需要建模成分数量无限的混合模型，引入狄利克雷过程作为参数Φ 的先验，定义为$G\sim \mathrm{DP}(\alpha,H)$ 。其中G是由集中参数为α，基分布为H的狄利克雷过程抽样得到的分布。

通常将狄利克雷过程的常见构造方式比喻为“折棍构造法”(stick-breaking construction) (Zhou et al., 2011)。对长度为1 的棍子按比例ξ_1 折断，保留比例为$1-\xi_1$ 的部分并记被切割的长度为π_1，而后对剩余的棍子切除比例为ξ_2 的部分，记其长度为$\pi_2\ldots$ 如此重复以获得一系列长度为π_i 的木棍。利用贝塔分布(Beta Distribution)的性质，以$\xi_i\sim \mathrm{Beta}(1,\alpha)$ 的方式抽样切割比例并保证$0<\xi_i<1$ 。这个过程就是折棍构造，即$\pi_i\sim \mathrm{GEM}(\alpha)$ (GEM 代表Griffiths, Engen和McCloskey)，可以得到$\pi_i=\xi_i\prod_{i=0}^{i-1}(1-\xi_i)$，满足$\sum_{i=1}^{\infty}\pi_i=1$。通过折棍构造有式(1)，其中，当$x=0$ 时有$\delta(x)=1$，其他情况下$\delta(x)=0$ 。

$$G(\phi)=\sum_{k=1}^{\infty}\pi_k\delta\left(\phi-\phi_k\right),\phi_k\sim H \tag{1}$$

根据折棍构造，从连续的基分布H中抽样得到相同的参数是可行的，这代表着狄利克雷过程可以生成无限成分的混合分布，且不同观测数据对应的分布参数可能相同，因此其具备聚类的能力，并能随着数据增长创建新的聚类。

同时，聚类需要考虑到文本间的相似度，因而引入文本建模中常见的狄利克雷-多项式共轭（conjugate）关系。使用多项式分布建模文本数据D，此时F(Φ)就被确定为多项式分布，假设文本由聚类对应的多项式词分布独立生成，即$p\left(d\mid\theta_c\right)=\prod_{w\in d}\mathrm{Mult}\left(w\mid\theta_c\right)$，所以这里的混合分布参数集合$\phi$仅包含多项式分布参数$\theta$ 。这样, 词共现关系较强的文本更可能被聚为一类。综上，用于短文本聚类的狄利克雷过程混合模型生成过程如下式所示(Yin et al., 2018) 。

$$\pi\mid\alpha\sim\mathrm{GEM}(1,\alpha) \tag{2}$$

$$c_i\mid\pi\sim\mathrm{Mult}(\pi)\quad i=1,\ldots,\infty \tag{3}$$

$$\theta_k\mid\beta\sim\mathrm{Dir}(\beta)\quad k=1,\ldots,\infty \tag{4}$$

$$d_i\mid c_i,\{\theta_k\}_{k=1}^{\infty}\sim p\left(d_i\mid\theta_{c_i}\right)=\prod_{w\in d}\mathrm{Mult}\left(w\mid\theta_{c_i}\right) \tag{5}$$

其中，α和β均为超参数，Mult($\cdot$)和Dir($\cdot$)分别代表了多项式分布和狄利克雷分布（Dirichlet distribution）。狄利克雷过程混合模型文本聚类算法概率图如图1所示。

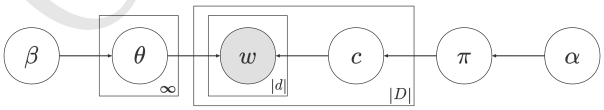


图 1: 狄利克雷过程混合模型概率图

4 情感增强的狄利克雷过程混合模型

4.1 模型结构

情感分布增强方法的目标是在前文所述的狄利克雷过程混合模型中以概率分布形式整合文本情感与词特征，联合学习词-情感分布参数，从而更好地表征用户观点并实现聚类。其中的核心问题为情感值计算、模型结构设计和参数推断过程。

社交媒体情感分析的研究认为，评论等带有情感色彩的主观性文本表现了用户在发出该言论时的情绪状态，其中可能蕴含了用户的个人态度。为了挖掘与表示文本的情感信息，学者们采用了不同的技术将自然语言量化为情感。常见的情感表示法有两类：（1）将情感按照极性分

类为积极、消极和中性等，用标签代表文本对应的情感(Liu, 2012)。该方法将情感看作离散的随机变量，可以视为服从多项式分布。（2）使用实数值表示文本情感的强度，实数的正负分别代表着情感积极或消极的倾向(Zadeh, 2015)。此处情感作为连续的随机变量，依据中心极限定理使用高斯分布建模是合理的。同时，高斯分布的概率密度集中于均值附近，该性质也符合聚类用户言论中相似情感的需求。考虑到连续的情感强度值可以利用在观点情感时序分析等下游任务中，本文选择使用情感值作为文本的情感特征，并使用高斯分布建模。

基于概率生成模型的聚类方法通常假设文本由分布所生成，通过计算文本数据从分布产生的概率进行聚类。情感增强的狄利克雷过程混合模型使用词分布和情感分布联合作用，同时衡量文本的词共现关系和情感相似性完成聚类，其工作过程如图2所示。

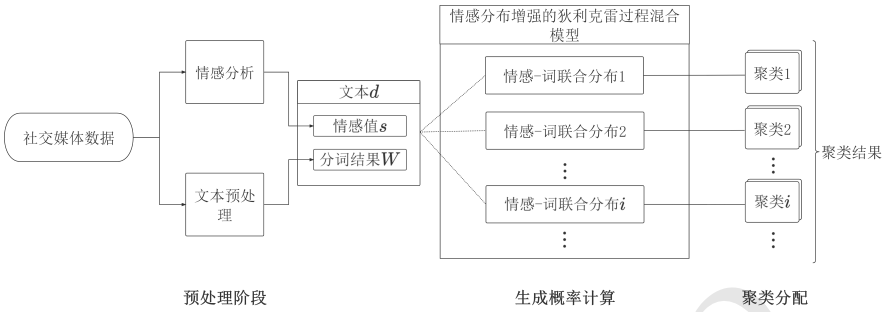


图 2: 基于情感增强非参数模型的观点聚类流程

设每一条短文本d_i 都有唯一对应的情感强度值s_i，其中实数值s_i 越接近正/负无穷则表示文本在积极/消极的情感上拥有越高的强度。因此定义文本情感s_i 由一个高斯分布生成。结合狄利克雷过程混合模型文本聚类框架和情感分布，得到文本生成的似然函数如式(6)。

$$\mathrm{F}\left(d_i \mid \phi_k\right)=\operatorname{Mult}\left(W^{(i)} \mid \theta_k\right) \mathcal{N}\left(s_i \mid \mu_k^s, \sigma_k^s\right) \tag{6}$$

在该似然函数中，此时$\phi_k=\left(\theta_k, \mu_k^s, \sigma_k^s\right)$ 为第k 个聚类(观点) 对应概率分布的参数集合。$\operatorname{Mult}\left(W^{(i)} \mid \theta_k\right)=)=\prod_{w \in d} \operatorname{Mult}\left(w \mid \theta_k\right)$表示在观点聚类$z_k$ 中，每条用户言论中所包含的词都是从聚类相应的多项式分布中依据词袋假设(bag-of-word) 独立生成的。多项式分布的参数θ 其实是一个长度为$|V|$ 的向量，其中V 代表的是算法已处理的所有词汇集合。高斯分布$\mathcal{N}\left(s_i \mid \mu_k^s, \sigma_k^s\right)$ 则拟合了聚类中的文本情感信息，μ_k^s 和σ_k^s 分别是高斯分布的均值和标准差。

接下来需要为上述模型参数设置先验。根据贝叶斯理论，若构造的先验与似然是共轭分布，则它们对应的后验分布与先验将是同一类型的分布。此性质能在狄利克雷过程混合模型参数推断阶段时简化积分式的计算。现已知多项式分布的共轭先验是狄利克雷分布$\mathrm{Dir}(\cdot)$，而高斯分布有多种的共轭先验。在分布参数μ 与σ 都未知的情况下，单维高斯分布的共轭先验可以为高斯逆卡方分布(Normal-inverse-chi-squared Distribution)，记作$\mathrm{Ni}\,\chi^2(\cdot)$ 。本文采用二者的结合设置似然函数的共轭先验，将其定义为式(7)。

$$G_0\left(\Phi_k\right)=\operatorname{Dir}\left(\theta_k \mid \beta_0\right) \mathrm{Ni}\,\chi^2\left(\mu_k^s, \sigma_k^s \mid \Psi_0\right) \tag{7}$$

其中β_0 和$\Psi_0=\left(m_0, \lambda_0, v_0, \epsilon_0^2\right)$ 是先验组成部分中狄利克雷分布与高斯逆卡方分布的参数。在先验分布G_0 中，这些参数均为超参数。基于以上过程图3展示了情感分布增强狄利克雷过程混合模型的概率图，被虚线框起的部分即为文本情感由分布生成的过程。

4.2 参数推断

本文参考经由Yin 等(2018)改进的坍缩吉布斯采样算法，进行混合模型的参数推断。实际上，混合模型的参数推断过程复杂且计算量大，不过在聚类过程中仅需要关注文本d_i所属类别编号c_i，因而可以简化计算，无需求解所有参数。

文本聚类编号的分配由后验预测分布$p\left(c_i=k \mid c_{-i}, d, \Phi\right)$ 确定，它可以根据贝叶斯法则被表示为先验分布与似然函数的乘积，如式(8)。其中，将d展开为d_i和d_{-i}以适应预测分布的形

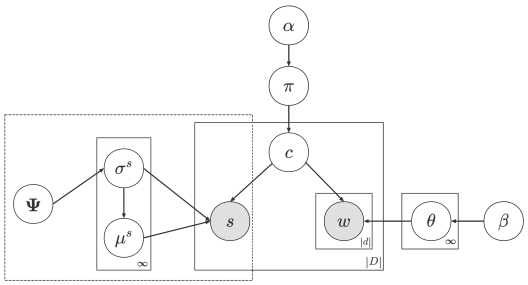


图 3: 情感分布增强狄利克雷过程混合模型概率图

式，并根据条件概率公式从原式变换为第二项，利用D分离（D-Separation）的性质得到最后的结果(Bishop and Nasrabadi, 2006)。

$$p\left(c_i \mid c_{-i}, d, \Phi\right) \propto p\left(c_i \mid c_{-i}, d_{-i}, \Phi\right) p\left(d_i \mid c, d_{-i}, \Phi\right) \propto \left(c_i \mid c_{-i}, \Phi\right) p\left(d_i \mid d_{-i}, c, \Phi\right) \tag{8}$$

经典的中餐馆过程（Chinese Restaurant Process，CRP）(Ferguson, 1973)直观地描述了聚类的过程，该过程说明了假若已知其他评论文本的类别，那么文本i所属观点类别的条件概率$p\left(c_i \mid c_{-i}, \Phi\right)$如式(9)，此条件概率即为式（8）中第一项的展开。

$$p\left(c_i \mid c_{-i}, \Phi\right) = p\left(c_i \mid c_{-i}, \alpha\right) \quad = \begin{cases} \frac{\left|z_{k_i - i}\right|}{|d| + \alpha|d| - 1} \text{(属于已有类的概率)} \\ \frac{\alpha|d|}{|d| + \alpha|d| - 1} \text{(属于新类的概率)} \end{cases} \tag{9}$$

其中$|d|$ 是包含第i 条数据在内，已输入算法的所有文本数量。而$|z_{k,-i}|$ 是第k 个聚类中除去第i 条数据的成员数量。与静态的文本聚类不同，动态的数据流聚类需要配合变量$|d|$，使用$\alpha|d|$ 代替式(2)中的α 作为混合模型成分权重的生成参数。

在计算文本由各聚类生成的概率时，为了便于推断需要假设变量之间的条件独立性，进而$p\left(d_i \mid d_{-i}, c, \Phi\right)$ 可以被分解为式(10) 。

$$p\left(d_i \mid d_{-i}, c, \Phi\right) \quad \propto p\left(W^{(i)} \mid d_{-i}, c, \Phi\right) p\left(s_i \mid d_{-i}, c, \Phi\right) \tag{10}$$

因此式(10)可以由混合模型中的词分布与情感分布分别计算。对于$p\left(s_i \mid d_{-i}, c, \Phi\right)$，它作为建模文本情感强度的后验预测分布，被高斯分布和其共轭先验所确定。根据Murphy (2007)的推导，该共轭关系在积分式中可以简化，计算方式详见式(11) 。

$$p\left(s_i \mid d_{-i}, c, \Phi\right) = \iint \mathcal{N}\left(s_i \mid \mu_k^s, \sigma_k^s\right) Ni\chi_2\left(\mu_k^s, \sigma_k^s \mid \Psi_{k,-i}\right) d\mu_k^s d\sigma_k^s = \mathrm{t}_{v_n}\left(m_n, \frac{\left(1+\lambda_n\right)\epsilon_n^2}{\lambda_n}\right) \tag{11}$$

式(11)中$\mathrm{t}_{v_n}(\cdot)$ 是自由度为v_n 的t 分布概率密度函数，根据该函数就可以计算得到文本情感由各聚类生成的概率，达到为文本分配聚类的目的。在推导过程中，出现的变量$\Psi_{k,-i} = \left(m_n, \lambda_n, v_n, \epsilon_n^2\right)$ 其实是对应于聚类z_k 中逆卡方分布的后验分布参数，在每次成功为文本d_i 指定聚类编号$c_i = k$ 后都需要使用该文本的情感强度值更新先验，即不断通过数据所确定的后验以修正聚类先验。文本情感分布的先验逆卡方分布参数更新方式(Murphy, 2012) 如式(12)，其中n代表用于更新后验的数据量，$\bar{s}$ 是当前聚类内情感均值。

$$\begin{aligned} \lambda_n &= \lambda_0 + n, \quad m_n = \frac{\lambda_0 m_0 + n\bar{s}}{\lambda_n} \\ v_n &= v_0 + n, \quad v_n \epsilon_n^2 = \left(v_0 \epsilon_0^2 + \sum_i \left(s_i - \bar{s}\right) + \frac{n\lambda_0}{\lambda_0 + n}\left(m_0 - \bar{s}\right)^2\right) \end{aligned} \tag{12}$$

有关聚类概率表达式$p\left(d_i \mid d_{-i}, c, \Phi\right)$ 的另一部分，即词分布的后验预测分布$p\left(W^{(i)} \mid d_{-i}, c, \Phi\right)$，同样利用多项式分布-狄利克雷分布共轭的性质推导得到文本词汇由各聚类生成

的概率(Yin and Wang, 2016; Xu et al., 2021), 得到式(13) 。

$$p\left(W^{(i)} \mid d_{-i}, c, \Phi\right) = \frac{\prod_{w \in d_i} \prod_{j=1}^{(w)} \left(f_{z_k,-d_i}^{(w)} + \beta + j - 1\right)}{\prod_{l=1}^{d_i|} \left(|v|_{z_k,-d_i} + |V|\beta + l - 1\right)} \tag{13}$$

$f_{d_i}^{(w)}$ 和$f_{z_k,-d_i}^{(w)}$ 分别指的是词w 在文本d_i 中的词频和在聚类z_k 里出现的次数，$|v|_{z_k,-d_i}$ 是聚类内现有词的数量。综上，包括模型参数推断在内的完整算法过程如表1 所示。本文所使用的自定义数学符号及其意义在表2中列出。

算法1: SDE算法

输入：逆卡方分布参数$m_0, \lambda_0, v_0, \epsilon_0$;文本$D = \{d_1, d_2, \ldots, d_\infty\}$

输出：文本对应的聚类编号$C = \{c_1, c_2, \ldots, c_\infty\}$

For d_i in D do

$s_i = get - sentiment - of(d_i)$//使用情感分析获取文本情感值

计算$p\left(s_i \mid d_{-i}, c, \Phi\right)$见式(11) //情感生成概率

计算$p\left(W^{(i)} \mid d_{-i}, c, \Phi\right)$见式(13) //词生成概率

$p\left(d_i \mid d_{-i}, c, \Phi\right) = p\left(W^{(i)} \mid d_{-i}, c, \Phi\right) p\left(s_i \mid d_{-i}, c, \Phi\right)$见式(10)

$p_k = \frac{|z_{k_i - i}|}{|d| + \alpha|d| - 1} \cdot p\left(d_i \mid d_{-i}, c, \Phi\right)$ //文本由已有类生成的概率

$p_{k+1} = \frac{\alpha|d|}{|d| + \alpha|d| - 1} \cdot p\left(d_i \mid d_{-i}, c, \Phi\right)$ //文本属于新产生的类

$c_i = argmax(p_k)$ //为文本分配生成概率最大的聚类

更新d_i所属聚类z_{c_i}的情感后验分布参数见式(12)

表 1: SDE算法过程

数学符号	含义	数学符号	含义
下标i	向量中的第i个元素	π	混合模型成分权重
$\lvert\cdot\rvert$	向量中的元素数量	α	狄利克雷过程集中参数
$-i$	向量中除第i个元素外所有元素	β	狄利克雷分布参数
$\bar{\cdot}$	向量均值	ξ	贝塔分布抽样结果
D, d	社交媒体文本流、单条文本	θ	词多项式分布参数
$w^{(i)}$	第i条文本的组成词	$\mu^{(s)}, \sigma^{(s)}$	情感分布均值、标准差
c	算法已分配的所有聚类编号	V, v	输入算法的词汇集、V的任意子集
s	文本情感值	$f^{(w)}$	词w 的词频
z	聚类（观点）	n	样本数据数量
Φ, ϕ	混合模型所有成分的参数集合、混合模型成分参数	$\Psi = (m, \lambda, v, \epsilon^2)$	高斯逆卡方分布参数

表 2: 本文所用数学符号及其意义对照

5 实验

5.1 数据集及评价指标

5.1.1 数据集介绍

为了便于对比，本文采用在短文本数据流聚类研究领域中广泛应用的Tweets数据集和Google-News数据集作为实验数据来源。

（1）Tweets数据集由2011至2015年TREC（Text Retrieval Conference）会议提供的推特数据构成[0] ，这些来自社交媒体平台Tweets的文本共30322条，并依据其所讨论的内容被标注

[0]http://trec.nist.gov/data/microblog

为269个不同的主题。该数据集被使用于诸多经典的文本聚类研究(Yin and Wang, 2016; Yin et al., 2018; Kumar et al., 2020; Xu et al., 2021; Chen et al., 2019)。Tweets数据集的平均文本长度为7.97个单词，较为符合社交媒体用户观点聚类的场景。

（2）Google-News数据集收集了11109篇新闻文章，合并同一事件的相关报道，共整理得到152个聚类，最终通过提取新闻标题建立数据集(Yin and Wang, 2014)。Google-News数据集的平均文本长度是6.23单词。相较于Tweets数据集，Google-News数据集作为新闻标题，其文本情感特征并不显著，因此对本文提出的情感分布增强模型更具挑战性。

经由预处理后两个数据集的统计信息如表3所示。之后本文将已集成于自然语言处理工具包[1]（Natural Language Toolkit）中的vader情感分析工具(Hutto and Gilbert, 2014)应用于预处理后的原始数据集上，以还原文本中的情感信息。

数据集	文本数量	聚类数量	词汇数量	平均长度
Tweets	30322	269	12301	7.97
Google-News	11109	152	8110	6.23

表 3: 实验数据集统计信息

5.1.2 评价指标介绍

（1）标准化互信息(Strehl and Ghosh, 2002)（Normalized Mutual Information, NMI）是评估聚类算法的最常见指标之一，若是模型结果越接近真实的聚类情况则NMI越近于1，否则NMI越近于0。

（2）聚类准确度（Accuracy，Acc）更直接地比较算法得到的文本类别标签与真实标签。

（3）聚类同质性（Homogeneity，Ho.）和聚类完整度（Completeness，Cp.）是两个不同的聚类目标(Rosenberg and Hirschberg, 2007)。同质性希望每个聚类中只包含该聚类的成员，而完整度指同一个类别的数据应当归属于同样的聚类簇。这两个指标有助于细致地衡量各算法在不同角度的表现。

（4）FMI（Fowlkes-Mallows Index）是聚类精度和召回的几何平均(Fowlkes and Mallows, 1983)。

实验结果将重点关注NMI、Accuracy与FMI，将三者作为考察不同算法表现的主要指标。

5.2 实验设计

5.2.1 对比方法

对比实验将选取短文本数据流聚类研究领域中最前沿的算法，选择依据以下三个条件：第一，聚类方式基于狄利克雷过程混合模型，才能由此比较加入情感分布后的模型表现。第二，能够以“一遍扫描”的方式处理流式文本数据，符合本文设想的社交媒体场景。第三，算法表现优于其他同类算法，且通过实验能复现原论文中的效果。最终，本文选取了下列两个模型：

（1）OSDM(Kumar et al., 2020)：OSDM是一个基于狄利克雷过程混合模型的短文本流聚类算法，仅支持通过“一遍扫描”的方式处理文本数据，它在Yin等人(2018)工作的基础上增强了词共现关系，并加入了词重要性特征，利用语义信息提升了模型表现。在真实数据集上的实验结果表明，OSDM在各个聚类评价指标中的表现都比较优秀。

（2）DP-BMM(Chen et al., 2020)：DP-BMM不同于其他研究，使用了文本中词对（Biterm）的共现关系来代替原有的单词共现特征。相比于单词，词对代表着更丰富的信息，进一步提升了聚类的表现。作为短文本聚类算法，它不仅能够以一遍扫描的方式实现聚类，也可以通过基于批处理（batch-based）的方式多次迭代处理数据。

5.2.2 模型超参数设置

对于模型OSDM与DP-BMM，本文使用原文献中提供的超参数设定以求重现算法的最优结果。对于OSDM，在两个数据集上采用相同的超参数：$\alpha = 0.002$，$\beta = 0.0004$。对于DP-BMM，在Tweets数据集上令$\alpha = 0.3$，$\beta = 0.02$，$batchsize = 1$，而在Google-News数据集上令$\alpha = 0.6$，$\beta = 0.02$，$batchsize = 1$。本文提出的文本情感分布超参数

[1]https://www.nltk.org/

为$\Psi_0 = (m_0, \lambda_0, v_0, \epsilon_0^2)$，即先验逆卡方分布的参数。通过实验，将OSDM-SDE的情感分布超参数设置为：$m_0 = 0$，$\lambda_0 = 1$，$v_0 = 1$，$\epsilon_0 = 0.01$，在Tweets数据集上将DP-BMM-SDE的超参数设置为：$m_0 = 0$，$\lambda_0 = 0.8$，$v_0 = 1$，$\epsilon_0 = 0.03$，而在Google-News数据集上令$m_0 = 0$，$\epsilon_0 = 0.1$，$v_0 = 1$，$\epsilon_0 = 0.001$。

5.3 实验结果及分析

5.3.1 对比实验结果

实验结果包含了各模型在NMI、Accuracy、Homogeneity、Completeness和FMI等五个指标下的聚类表现，总体的实验结果如表4所示，其中加粗字符表示在该指标上更优的结果。

从表4中可以看出，在整合了文本情感分布后，原有基于狄利克雷过程混合模型的短文本聚类算法在不同程度上均有所提高。情感分布改进后的模型聚类效果在NMI、Accuracy这两个需要关注的指标上都超过了原模型，通过FMI指标也可以看出本文提出的算法在类别分配结果上相较于原模型也更合理。同质性和完整度互相具备冲突的性质，本文提出的算法偏向聚类的完整性，不过二者的调和平均在数值上实际与NMI接近，因此可以看出总体上本文提出的算法更能兼顾二者。

模型	数据集									
	Tweets					Google-News				
	NMI	Acc	FMI	Ho.	Cp.	NMI	Acc	FMI	Ho.	Cp.
OSDM	0.847	0.613	0.583	**0.905**	0.796	0.812	0.617	0.535	**0.829**	0.796
OSDM-SDE	**0.852**	**0.632**	**0.637**	0.902	**0.807**	**0.815**	**0.618**	**0.558**	0.822	**0.807**
DP-BMM	0.799	0.614	0.569	0.773	**0.827**	0.838	0.684	0.635	0.825	0.851
DP-BMM-SDE	**0.801**	**0.629**	**0.578**	**0.778**	0.826	**0.840**	**0.700**	**0.643**	**0.829**	**0.853**

表 4: 对比试验结果

如表5所示，通过对比模型在两个数据集中NMI、Accuracy、FMI评价指标的表现，还可以看出经由文本情感分布增强的聚类算法在社交媒体数据集（Tweets）上的表现提升比在新闻标题数据集（Google-News）上的提升要更加显著。以OSDM为例，情感分布增强后的模型在Tweets数据集上NMI、准确度、FMI分别提升了0.59%、3.1%、9.26%，但是在News数据集上各自只提升了0.37%、0.16%、4.3%。这主要是由于社交媒体平台用户倾向于自由地发表个人言论，因此文本中所蕴含的主观情感色彩较为浓郁。而在新闻标题中则恰恰相反，报道者旨在客观提炼事件信息，用词较为中立。这验证了本文提出的文本情感分布增强方法能够更好地处理情感特征，并在社交媒体用户观点聚类的场景下提升聚类算法的表现。

模型	NMI提升		准确度提升		FMI提升	
	Tweets	Google-News	Tweets	Google-News	Tweets	Google-News
OSDM-SDE	**0.59%**	0.37%	**3.10%**	0.16%	**9.26%**	4.30%
DP-BMM-SDE	**0.25%**	0.23%	**2.44%**	2.34%	**1.58%**	1.26%

表 5: SDE改进模型在不同数据集上的提升对比

5.3.2 数据规模对算法的影响

将Tweets数据集按照3000的步长逐点输出聚类结果，观察数据规模对情感分布增强方法的影响。以OSDM为例，实验结果如表6所示。以NMI、聚类准确度、FMI指标作为评判标准，使用情感分布增强的聚类算法在各数据测试点上取得了超越原模型的表现。观察数据量较少时模型的聚类结果，可以看出SDE方法在小数据量时也一定程度上提升了聚类准确度、NMI与FMI。这项实验结果说明在结合了文本的词特征与情感特征后，能够改进模型在数据较稀疏时的聚类能力，有效利用了社交媒体文本潜藏的用户主观信息。

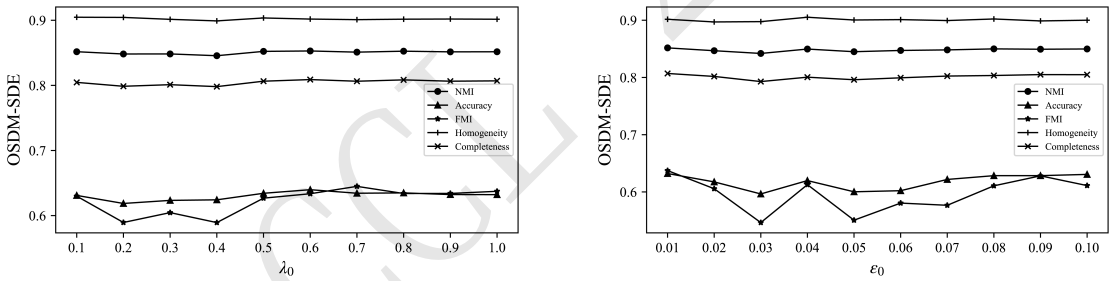
指标	模型	3000	6000	9000	12000	15000	18000	21000	24000	27000	30000
Acc	OSDM	0.601	0.59	0.594	0.621	0.629	0.635	0.637	0.647	0.63	0.614
	+SDE	**0.631**	**0.597**	**0.602**	**0.634**	**0.643**	**0.649**	**0.654**	**0.659**	**0.647**	**0.641**
NMI	OSDM	0.814	0.803	0.807	0.832	0.841	0.848	0.853	0.857	0.852	0.848
	+SDE	**0.818**	**0.803**	**0.81**	**0.834**	**0.844**	**0.851**	**0.857**	**0.86**	**0.856**	**0.853**
FMI	OSDM	0.629	0.585	0.578	0.598	0.629	0.639	0.644	0.655	0.625	0.586
	+SDE	**0.641**	**0.589**	**0.591**	**0.621**	**0.654**	**0.655**	**0.661**	**0.67**	**0.653**	**0.633**

表 6: 不同数据规模的准确度、NMI、FMI(OSDM-SDE)

5.3.3 参数敏感性分析

在使用逆卡方分布Ni$\chi^2(\cdot)$ 作为情感分布$\mathcal{N}(s \mid \mu^s, \sigma^s)$ 的先验时，需要输入其先验逆卡方分布的参数$\Psi_0 = \left(m_0, \lambda_0, v_0, \epsilon_0^2\right)$ 。根据分布定义Ni$\chi^2\left(m_0, \lambda_0, v_0, \epsilon_0^2\right) = \mathcal{N}\left(\mu^s \mid m_0, \sigma^s/\ \lambda_0\right) \times \chi^{-2}\left(\sigma^s \mid v_0, \epsilon_0^2\right)$，可得超参数$\Psi_0$ 所分别对应的解释含义为：m_0 是高斯分布参数μ^s 的先验均值，ϵ_0^2 是参数σ^s 先验分布的缩放参数, λ_0 和v_0 表示了对先验的信任程度。实际应用中通常采取弱信息先验假设，将先验的信任程度设定为一个较小的值(Chipman et al., 2001), 因此本文令λ_0 不大于1 。而v_0 则常被设置为与数据变量维度相同，赋值$v_0 = 1$ 。此外，在建模高斯变量时的一个常见选择是将其均值设为0 ，即$(m_0 = 0)$ 。综上，输入参数中需要人为设定的仅有λ_0 与ϵ_0 两个变量。本小节以OSDM为例，在Tweets数据集下进行敏感性分析。

固定其他参数，分别使λ_0 和ϵ_0在对应取值范围变化，参数敏感性分析实验结果如图4所示。随着λ_0 的变动可以观察到，表示聚类算法表现的NMI、准确度、同质性和完整度指标都相对稳定，仅能在FMI 指标曲线上能观测到一些波动。这主要是因为在弱信息先验假设下，λ_0 的值在较小的区间内变动，对聚类过程的作用并不明显。与λ_0 类似, ϵ_0 的变动也没有对本方法的最终表现造成较大波动，NMI、聚类准确度等指标依然稳定。



(a) (b)

图 4: λ_0(a)，ϵ_0(b)在Tweets上的敏感性分析(OSDM-SDE)

6 结论

本文提出了使用文本情感分布增强（SDE）的方法改进短文本数据流聚类模型，旨在解决社交媒体用户观点聚类中文本长度短、数据量未知，以及缺乏观点数量先验知识等难题，利用文本数据中含有的主观情感信息更有效地实现观点聚类。本文将情感量化为数值，并将其建模为高斯分布，加入狄利克雷过程混合模型中作为文本生成的联合分布，用坍缩吉布斯采样算法推断混合模型参数，同时推导了情感分布的参数更新方式。在真实数据集上的实验结果表明，本文提出的SDE聚类方法在NMI、FMI、聚类准确度等方面均超越了现有模型，验证了SDE方法的合理性和有效性。通过对比社交媒体数据集与新闻数据集上的聚类结果，可以发现使用文本情感分布增强不仅提升了现有模型的聚类效果，还能显著地增进模型在具有较强情感色彩数据集上的表现，符合本文假设的社交媒体用户观点聚类场景。未来的研究方向包括利用社交媒体流式数据的时间顺序信息提升模型的表现以及在聚类结果的基础上文本摘要的自动抽取等。

参考文献

李秀霞and 邵作运. 2016. “密度-距离” 快速搜索聚类算法及其在共词聚类中的应用. 情报学报, 35(4):380–388.

高俊峰and 黄微. 2019. 网络舆情场中观点簇丛的情感极化度测算. 图书情报工作, 63(10):106.

Charu C Aggarwal. 2013. A survey of stream clustering algorithms.

Randa Benkhelifa and Fatima Zohra Laallam. 2018. Opinion extraction and classification of real-time youtube cooking recipes comments. In *International Conference on Advanced Machine Learning Technologies and Applications*, pages 395–404. Springer.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Junyang Chen, Zhiguo Gong, and Weiwen Liu. 2019. A nonparametric model for online topic discovery with word embeddings. *Information Sciences*, 504:32–47.

Junyang Chen, Zhiguo Gong, and Weiwen Liu. 2020. A dirichlet process biterm-based mixture model for short text stream clustering. *Applied Intelligence*, 50(5):1609–1619.

Hugh Chipman, Edward I George, Robert E McCulloch, Merlise Clyde, Dean P Foster, and Robert A Stine. 2001. The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134.

Thomas S Ferguson. 1973. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Edward B Fowlkes and Colin L Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

Fei Geng, Qilie Liu, and Ping Zhang. 2020. A time-aware query-focused summarization of an evolving microblogging stream via sentence extraction. *Digital Communications and Networks*, 6(3):389–397.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Jay Kumar, Junming Shao, Salah Uddin, and Wazir Ali. 2020. An online semantic-enhanced dirichlet model for short text stream clustering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 766–776.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174.

Yuelin Li, Elizabeth Schofield, and Mithat Gönen. 2019. A tutorial on dirichlet process mixture modeling. *Journal of mathematical psychology*, 91:128–144.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Kevin P Murphy. 2007. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Hai-Long Nguyen, Yew-Kwong Woon, and Wee-Keong Ng. 2015. A survey on data stream clustering and classification. *Knowledge and information systems*, 45(3):535–569.

Ningning Ni, Caili Guo, and Zhimin Zeng. 2018. Public opinion clustering for hot event based on br-lda model. In *International Conference on Intelligent Information Processing*, pages 3–11. Springer.

Brendan O’Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth international AAAI conference on weblogs and social media*.

Md Rashadul Hasan Rakib, Norbert Zeh, and Evangelos Milios. 2021. Efficient clustering of short text streams using online-offline clustering. In *Proceedings of the 21st ACM Symposium on Document Engineering*, pages 1–10.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. 2013. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 533–542.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

Peng Wu, Xiaotong Li, Si Shen, and Daqing He. 2020. Social media opinion summarization using emotion cognition and convolutional neural networks. *International Journal of Information Management*, 51:101978.

Wanyin Xu, Yun Li, and Jipeng Qiang. 2021. Dynamic clustering for short text stream based on dirichlet process. *Applied Intelligence*, pages 1–12.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242.

Jianhua Yin and Jianyong Wang. 2016. A model-based approach for text clustering with outlier detection. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 625–636. IEEE.

Jianhua Yin, Daren Chao, Zhongkun Liu, Wei Zhang, Xiaohui Yu, and Jianyong Wang. 2018. Model-based clustering of short text streams. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2634–2642.

Amir Zadeh. 2015. Micro-opinion sentiment intensity analysis and summarization in online videos. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 587–591.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pages 338–349. Springer.

Jian Ying Zhou, Fei Yue Wang, and Da Jun Zeng. 2011. Hierarchical dirichlet processes and their applications: a survey. *Zidonghua Xuebao/Acta Automatica Sinica*, 37(4):389–407.