doi:10.3772/j.issn.1000-0135.2015.005.010

基于混合模型的文本聚类研究综述[1)]

王 方 成 颖 柯 青

（南京大学信息管理学院，南京 210023）

摘要 相较于其他聚类算法，模型聚类的实证研究结果表现出了独特的优势，越来越受到学界的关注。本文梳理了混合模型文本聚类的相关研究，根据聚类分析的技术路线，主要综述了文本建模、参数建模以及模型推理等三个主要模块，在此基础上总结了特征降维、半监督聚类以及聚类过程的系统整合等不同研究中的共性问题。最后，提出了本领域未来可能的研究方向。

关键词 模型聚类 混合模型 文本聚类

Mixture Model-based Text Clustering: A Review

Wang Fang, Cheng Ying and Ke Qing

(*School of Information Management, Nanjing University, Nanjing* 210023)

Abstract Model-based clustering has attracted more and more attention, and empirical studies also showed distinct advantage. This paper reviews the status of the document clustering based on mixture models. According to the technical routes, it summarizes three main parts, such as document modeling, parameter modeling, and model inference, and analyses the common problems in different researches, including feature reduction, semi-supervised clustering and the integration of clustering process. At last it presents possible future research directions in this field.

Keywords model-based clustering, mixture model, document clustering

1 引 言

聚类方法可以分为判别式与生成式两类[1]。判别式方法基于数据的相似性（或距离）进行聚类，例如k-means以及层次聚类算法；生成式方法通过特定的模型表征簇，使得数据与模型的拟合最优，例如自组织地图（SOM）与混合模型（Mixture Models）聚类。其中混合模型聚类的基本思想是将聚类问题转化为数学建模问题，即利用简单概率分布的组合模拟复杂概率分布的统计建模方法，一个概率分布代表一个簇的特征分布，概率分布的组合即代表整个数据集的特征分布，核心环节是数据建模与模型推理。相对于判别式方法，混合模型聚类有坚实的概率论基础，通过数据分布的形状与结构而非相似性进行类簇的识别，更具灵活性，而且能够从概率统计的视角为每个簇提供直观的解释，具有良好的理论与应用前景。

混合模型研究的早期存在模型推理过于复杂的瓶颈，1977年Dempster等[2]提出的EM算法较好地

收稿日期：2014年8月4日。

作者简介：**王方**，女，1991年生，硕士生，主要研究方向：信息检索。E-mail：frie_wong@gmail.com。**成颖**，男，1971年生，博士，教授，博士生导师，主要研究方向：信息检索、信息行为。E-mail：chengy@nju.edu.cn。**柯青**，女，1979年生，博士，副教授，主要研究方向：人机交互。E-mail：keqing@nju.edu.cn

1） 本文得到国家社会科学基金重大招标项目“面向学科领域的网络信息资源深度聚合与服务研究（12&ZD221）”以及国家自然科学基金项目“融合范式视角下的链接分析理论集成框架及其实证研究（71273125）”的资助。

解决了该问题,促进了该模型在文本处理、生物医学、图像处理以及模式识别等领域的研究与应用,并取得了令人欣喜的成果。基于此,本文聚焦于基于混合模型的文本聚类研究,其流程见图1。混合模型文本聚类的主要模块有:①文本建模。即假定文本集符合某种统计模型,如多元伯努利(Bernoulli Mixture, BM)、多项式混合模型(Multinomial Mixture, MM)或者vMF(von Mises Fisher)混合模型等;②模型推理。即利用基于BM等模型的数据似然(或后验概率)作为优化准则,通过EM算法、变分推理以及马尔可夫链-蒙特卡洛(Markov chain Monte Carlo ,MCMC)方法等推理算法得到模型参数的估计值以完成聚类;③聚类评估,常用指标有准确率、召回率以及F值等。除了这些主要模块之外,部分研究中还涉及参数建模(由于仅在部分研究中出现,图1中呈现为虚线),即对混合模型的参数进行建模,旨在实现模型优化,常用方法有狄利克雷分布以及狄利克雷过程等。

在介绍文献来源之后,本文按照图1的线索组织。混合模型文本聚类的分词、词干提取、同义词归并等文本预处理工作与其他自然语言处理相似,不再赘述。聚类评估的相关指标及应用与判别式聚类相同,也不再展开。本文重点综述文本建模、参数建模以及模型推理的相关研究进展。此外,还对混合模型聚类的特征选取、特征抽取(feature extraction)、特征选择(feature selection)、半监督聚类以及聚类过程的系统整合等共性问题进行了总结与展望。

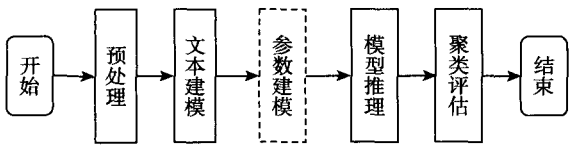

图1 混合模型文本聚类流程图

2 文献来源

2.1 文献选取原则

(1)综述文献的基本设定为:文本集由若干简单概率分布的混合模型生成,每一混合成分对应一个簇,文本属于其中一个(或若干)簇;

(2)具体范围包括混合模型文本聚类中的文本建模、参数建模以及模型推理等研究型文献;

(3)语种为英文与中文;

(4)文献类型包括期刊论文、硕博士论文以及会议论文。

2.2 文献选取过程

(1)基于WOS核心合集(排除化学索引),时间跨度为1900~2014年,检索式为:主题=((" mixture model * " or " mixture distribution * " or mixtures) and cluster *) or " probabilistic cluster * " or " distributional cluster * " or " model-based cluster * ") and (text * or document *),检索结果为416篇文献,根据题名和文摘筛选后得到41篇相关文献。进一步检索:主题=((multinomial or Bernoulli or " von mises fisher " or vmf) and (text * or document *) and cluster *),检索得到63篇文献,补充得到2篇相关文献。

(2)基于Google Scholar搜索引擎,采用model-based、" mixture model" 分别与" document cluster" 相组合,相应的检索结果为487、209,经筛选后得到7篇文献。

(3)在中文文献方面,主要基于CNKI以及万方两个数据库,检索式为:(混合模型+概率模型)*(文本+文档)*聚类,在CNKI得到92篇结果,在万方得到100篇结果,筛选后得到6篇。

(4)在文献阅读过程中根据参考文献不断扩展综述文献的范围,另获得20篇参考文献,最终合计相关文献76篇。根据权威的中国计算机学会推荐的国际学术会议和期刊目录选择了核心文献18篇;在此基础上,根据被引频次、发文量等指标确定了核心作者,如Banerjee A,Zhong S, Meila M等,由此得到核心文献19篇,二者去重后共得到核心文献24篇(附录1),以此为线索,结合另外52篇相关文献完成本文。

3 文本建模

假设文本集$D=\{d_1,d_2,\cdots,d_N\}\in\mathbf{R}^W$,$N$表示$D$中的文本数量,$W$表示文本的特征维数,$d_n=\{d_{n1},d_{n2},\cdots,d_{nw}\}$表示第$n$个文本,$d_{nw}$表示第$n$个文本的第$w$个特征维度。$D$包含$K$个簇,即由$K$个成分的混合模型生成。现有的混合模型都符合朴素贝叶斯假设,即特征维度之间相互独立,文本之间也相互独立。$\{d_1,d_2,\cdots,d_N\}\in\mathbf{R}^W$可以看做是随机向量$d_n\in\mathbf{R}^W$的独立实现,从模型中生成$d_n$以及$D$的概率见公式(1)和公式(2)[3,4]。

$$p(d_n \mid \Theta) = \sum_{K=1}^{K} \pi_k p_k(d_n \mid \theta_k) \quad (1)$$

$$p(D \mid \Theta) = \Pi_{n=1}^{N}(\sum_{K=1}^{K} \pi_k p_k(d_n \mid \theta_k)) \quad (2)$$

其中，π_k（$\pi_k \geq 0$，且 $\sum_{K=1}^{K} \pi_k = 1$）代表混合成分 k 在混合模型中的比重，$p_k(\cdot)$ 表示混合成分 k 的概率分布，θ_k 为混合成分 k 的参数。模型中的所有未知参数用Θ表示，如果簇的数量 K 确定，则称之为有限混合模型，$\Theta = \{\pi_1, \pi_2, \cdots, \pi_k; \theta_1, \theta_2, \cdots, \theta_k\}$；如果簇的数量无限，则称之为无限混合模型，$\Theta = \{\pi_1, \pi_2, \cdots, \pi_k; \theta_1, \theta_2, \cdots, \theta_k; K\}$，假设 K 值趋近于无穷大[5~7]。

在具体的研究与应用中，需要根据数据集的类型选择特定的概率分布函数 $p(\cdot)$。目前，在混合模型文本聚类中常用的模型有：

3.1 离散型混合模型

该类型主要有 BM 和 MM。其中，BM 最早应用于文本分类的文本建模，并且得到了较好的效果[8,9]。BM 只考虑项在文本中是否出现，每个特征维度的取值 $d_{nw} \in \{0,1\}$，取 1 表示项 w 在 d_n 中出现，否则反之。MM 还考虑了文本中项 w 出现的频次，用 d_{nw} 表示。如果文本集的特征维度为 W，则二者第 k 个混合成分的参数为 $\theta_k = \{\theta_{k1}, \theta_{k2}, \cdots, \theta_{kw}\}$，$\theta_{kw}$ 代表簇 k 中项 w 出现的概率。BM 的文本生成概率见公式(3)[10,11]，将其代入公式(2)即可得到文本集的生成概率(下同)。

$$p_k(d_n \mid \theta_k) = \Pi_{W=1}^{W} \theta_{kw}^{d_{nw}}(1 - \theta_{kw})^{1-d_{nw}} \quad (3)$$

令 l_n 为 d_n 的长度，多项式模型的文本生成概率见公式(4)[6,7,12,13]。

$$p_k(d_n \mid \theta_k) = \frac{l_n!}{\Pi_{W=1}^{W} d_{nw}!} \Pi_{W=1}^{W} \theta_{kw}^{d_{nw}} \quad (4)$$

BM 与 MM 未区分项主题表达能力的不同，即某些项与特定主题密切相关，在不同簇上的分布差异明显；而有些项则概念宽泛，在不同簇上的分布类似。对此，Li 等[14]基于 MM 将 θ_k 细化为 θ_T^k 与 θ_G^k，分别对应与主题内容相关的主题模型，以及与写作知识相关的通用模型，该方法与经典的 MM 相比，在 Marco - F1 指标上提高了 40%；Bouguila[10] 基于 BM，引入参数 λ_w 代替 θ_{kw} 表征所有无关特征项，参数 ρ_w 用于表征项 w 为相关特征项的概率。上述研究归纳起来基本思路都是区分核心(core)项与通用(general)项，研究中具体计算公式虽有差异，但都可以统一于公式(5)。其中，θ_{kw}^b 为通用项的参数，θ_{kw}^t 为核心项的参数，系数 ε 用于平衡两者的效用比重，类似于语言模型中的平滑思想。

$$P(w \mid k) = (1 - \varepsilon)\theta_{kw}^b + \varepsilon\theta_{kw}^t \quad (5)$$

3.2 连续型混合模型

Salton 等[15]的研究表明，对文本向量进行归一化处理能够解决文本长度造成的偏差；Dhillon 等[16]的研究显示，基于余弦相似度而非欧氏距离的 k-means 算法(spk-means)在文本聚类中效果更好。据此，Banerjee 等[17]认为文本向量具有方向性数据(directional data，即 $\|d_n\| = 1$)的特征，并应用 vMF (von Mises Fisher)分布完成文本建模[公式(6)][5,17]。

$$p_k(d_n \mid \theta_k) = c_w(\kappa_k \mu_k^T d_n) \quad (6)$$

其中，$\|d_n\| = 1$，$c_w(\kappa)$ 为标准化常量，参数 $\theta_k = (\mu_k, \kappa_k)$，$\mu_k$ 为均值向量，$\kappa(\kappa >= 0)$ 为聚集参数(concentration parameter)，表示向量 d_n 在 μ_k 周围的集中程度，κ 值越大说明 d_n 在 μ_k 方向上的集中度越高，当 $\kappa = 0$，密度函数降为一个均匀分布，$\kappa \to \infty$，则密度函数趋向于一个点密度。

研究表明 vMF 具有鲜明的优点：Banerjee 等[17]发现 vMF 是余弦相似度基于参数模型的一般化，从而理论上 vMF 模型能在保证聚类质量的同时提升算法效率；Banerjee 等[18]的实验表明该模型能够较好地解决高维空间中的簇重叠、簇的分布偏斜、簇规模过小等问题；Mardia 等[19]还证实 vMF 是一种适用于高维空间方向统计的分布。不过，vMF 也具有参数估计困难等缺点，由于参数 κ 包含贝塞尔方程比值的倒数，这一复杂的非线性公式使得极大似然估计以及 EM 算法中的 M 步难以实现[20,21]。对此，Zhong 等[1]在研究中采用了一种简化的 vMF 模型，即每次迭代中所有混合成分的参数 κ 取值相同，然后逐次增大 κ 值；Banerjee 等[18]则运用 κ 的近似替代 κ，该方法比固定的 κ 值能得到更好的聚类结果，但是运算效率不高。

3.3 一点思考

(1)模型性能。Zhong[1]的对比实验显示，在聚类效果方面，大部分数据集上 vMF 优于 MM，后者又优于 BM，例如在 NG20 数据集上三者的 NMI 指标分别为 0.57、0.54、0.19；在运行时间方面，基于 NG20 数据集在硬分配条件下三者的运行时间分别为 17.5s、36.7s、43s，vMF 呈现出良好的时间效率，但在软分配条件下分别为 76.8s、47.7s、77.8s，显示 vMF 更容易受分配策略的影响，相比之下 MM 更具

稳定的时间效率。总体而言,BM 结构简单,但实验结果不够理想。MM 时间复杂度较低且聚类效果较好,在实践中取得了广泛的应用。vMF 相比前两者模型更加复杂,模型推理的时间复杂度更大,建模效果相对更优,不过尚存部分理论问题有待解决,是目前的研究热点。

(2)文本长度的影响。McCallum 等[9]的研究表明,在 Yahoo 数据集上,MM 在 1000 词时准确率最高,而 BM 在 200 词时准确率最高;在 Industry Sector 数据集上,MM 在 20 000 词时准确率最高,而 BM 在 1000 词时准确率最高;另外三个数据集也得到了相似的结论。由此可见,BM 适合于短文本,而 MM 更适合于长文本的复杂聚类任务;研究还表明,MM 更适合于文本长度变化较大的数据集,而 BM 则更适合于文本长度比较稳定的数据集。Novovičová[22]的研究表明,除非常小的项集之外 MM 的分类准确率总体上优于 BM,MM 最高能达到 94.9%,比 BM 平均高 4%;与 BM 相似,vMF 也在较短的文本上表现优异[1]。

(3)混合模型的综合应用。Zhu 等[11]提出每个域都由一个最适合的概率模型独立生成的建模思想,构建了域独立聚类模型(Field Independent Clustering Model, FICM),分别采用伯努利、多项式分布对不同文本域联合构建混合模型,相比于单一的 BM、MM,三者在 TREC 数据集上的平均 NMI 分别为 0.736、0.714、0.712,从而提示混合模型的综合应用是该领域值得探讨的研究方向。

除了上述三种模型之外,早期对于混合模型的研究都是以高斯混合模型(Gaussian Mixture Model, GMM)进行建模,目前仍然应用广泛[23,24]。不过,因为文本的高维与稀疏特征不符合高斯分布,除了 Liu 等[25]利用 GMM 进行的探索性工作,在文本聚类中鲜见该模型。

4 参数建模

MM 模型难以识别项爆发(burstiness),对此,Madsen 等[26]通过参数的先验分布对模型的参数进行建模,提高了建模效果。狄利克雷过程的引入成功地解决了 MM 以及 vMF 模型需要预先设定簇数量 K 的不足。

4.1 狄利克雷分布

狄利克雷分布是多项式分布的共轭分布,通常作为多项式参数的先验分布应用在贝叶斯估计中,在文本建模中已有较多应用[27]。Madsen 等[26]采用 DCM(Dirichlet Compound Multinomial)[28]对文本建模,即通过狄利克雷分布抽取多项式分布[公式(7)],然后基于此生成文本[公式(8)]。

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{w=1}^{W}\alpha_w)}{\prod_{w=1}^{w}\Gamma(a_w)}\prod_{w=1}^{W}\theta_w^{\alpha_w-1} \tag{7}$$

$$p(d_n \mid \alpha) = \int p(d_n \mid \theta)p(\theta \mid \alpha)d\theta = \frac{l_n!}{\prod_{w=1}^{W}d_{nw}!}\frac{\Gamma(\sum_{w=1}^{W}\alpha_w)}{\Gamma(\sum_{w=1}^{W}d_{nw}+\alpha_w)}\prod_{w=1}^{W}\frac{\Gamma(d_{nw}+\alpha_w)}{\Gamma(\alpha_w)} \tag{8}$$

其中,$\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_w\}$ 是狄利克雷分布的参数,该模型的优点弥补了 MM 无法识别项爆发(burstiness)的缺陷,即项在文本中出现一次后趋向于再次出现,参数 α_w 的值越小,项越趋向于爆发,该模型在文本分类的研究中呈现了较好的效果。Elkan[29]鉴于文本的稀疏特性对公式(8)进行了简化,采用 $\Gamma(\alpha_w)$ 代替 $\frac{\Gamma(d_{nw}+\alpha_w)}{\Gamma(\alpha_w)}$ 提出了一种近似的 EDCM 并应用于文本聚类,在 NIPS 数据集的对比实验中,EDCM 与 DCM 的 MI 分别为 0.84364、0.83705,运行时间分别为 6.61s、751.3s,在聚类效果相差无几的情况下,EDCM 的时间复杂度远低于 DCM。Bouguila 基于广义狄利克雷(generalized Dirichlet)[30]提出了 MGDD(Multinomial Generalized Dirichlet Distribution)[31],该方法与 DCM 以及 MM 在 WEBKB 数据集上的聚类准确率分别为 87.12、84.25、81.16,t 检验表明 MGDD 对聚类的改善效果具有统计上的显著性。该结果提示,广义狄利克雷分布在混合模型文本聚类中的研究与应用是一个可选的方向。

4.2 狄利克雷过程

DP 是狄利克雷分布在连续空间上的扩展,通常表示为:$G \sim DP(\alpha, G_0)$,其中 G_0 是基分布,$\alpha(\alpha>0)$ 是集中度参数,G 表示在 G_0 和 α 上产生的随机分布。DP 主要应用于概率模型中作为先验分布以构建狄利克雷过程混合模型(Dirichlet Process Mixture, DPM)[32]。DPM 在文本聚类中有 stick-breaking 和中国餐馆过程(Chinese Restaurant Process, CRP)两种不同的模型构建视角,聚类思想见公式(9)和公式(10),即文本 d_n 以正比于簇规模

的概率属于一个既有的簇，或以正比于 α 的概率属于一个新簇，最终得到的参数集 $\{\theta_1,\theta_2,\cdots,\theta_N\}$ 提供了聚类的依据[33]。

$$\theta_{n+1} \mid \theta_1,\cdots,\theta_n,\alpha,G_0 \sim \frac{1}{n+\alpha}\sum_{k=1}^{K} t_k\delta_{\theta_k^*} + \frac{\alpha}{n+\alpha}G_0 \tag{9}$$

$$z_{n+1} \mid z_1,\cdots,z_n,\alpha,G_0 \sim \frac{1}{n+\alpha}\sum_{k=1}^{K} t_k\delta_{\theta_k^*} + \frac{\alpha}{n+\alpha}G_0 \tag{10}$$

其中，$\delta(\cdot)$ 为狄拉克 δ 函数，新参数 θ_{n+1} 的条件分布服从 polya 分布。Huang 等[6,7]从 CRP 的角度运用 DP，基于多项式模型先后提出了 DPMFS 和 DPMFP 模型，还采用简化的 DMA 模型[34]解决了 DPM 无法快速估计参数的问题。DPMFP[7] 与 DMA、EM-MN、K-MEANS、LDA 以及 EDCM 等模型在簇数量 K 值准确的前提下，NMI 分别为 0.534、0.512、0.531、0.229、0.559、0.510；当 K 值不准确，如设 $K=10$，NMI 分别为 0.534、0.512、0.486、0.184、0.482、0.401。从而显示，DPMFP 在 K 值未知的条件下更为稳健。Nguyen 等[5]采用 stick-breaking 基于 vMF 提出了 DPMvMF 模型，即假设混合系数 $\pi=\{\pi_k\}_{k=1}^{\infty}$ 是一个无穷序列：

$$v_k \sim \text{Beat}(1,\alpha),\pi_k(v)=\pi_k\Pi_{j=1}^{k-1}(1-v_j),$$

$$G=\sum_{k=1}^{\infty}\pi_k\delta_{\theta_k},\theta_n \sim G。$$

$\theta_1^*,\theta_2^*,\cdots,\theta_k^*$ 是参数集 $\{\theta_1,\theta_2,\cdots,\theta_n\}$ 的不重复序列，t_k 表示参数 θ_k^* 在 $\{\theta_1,\theta_2,\cdots,\theta_n\}$ 中出现的频次。DPMvMF 在与基于 MM 的聚类算法、基于 vMF 的软分配算法、基于 vMF 的 DA 算法、CLUTO 等四个模型的对比实验中也得到了与 Huang 等[6,7]类似的结果。

4.3 一点思考

DPM 作为一种非参数贝叶斯模型，具有无需预先给定簇数目 K 以及建模更具灵活性与适应性等优点，在模式识别领域已有较多的研究与应用，但在自然语言处理领域仍处于起步阶段，尤其在混合模型文本聚类中还有一些有待解决的理论与应用问题。

HDP。DP 可以对一组(group)文本进行聚类与分析，但在聚类多组文本时则显得力不从心，Teh 等[35]提出的层次 Dirichlet 过程(Hierarchical Dirichlet Processe，HDP)较好地解决了该问题。Hu 等[36]采用 IHDP-HMM 模型结合主题模型的思想进行了文本聚类的研究，与谱聚类、k-means、SOM 等算法的聚类准确率分别为 94.6、92.9、89.5、90.8，结果显示该算法具有一定的优势，因此，有必要加强其在混合模型文本聚类中的研究与应用。

分配策略。DPM 聚类算法属于硬分配算法，无法处理簇重合的情况。Griffiths 等[37]提出的 Indian Buffet Processes(IBP)能实现实例同时属于多个簇，而且更适用于稀疏矩阵。该模型在图像处理、网络分析、特征抽取等领域已有一定的应用，相应推理算法的研究已取得较为丰富的成果，但是在自然语言处理领域研究与应用甚少[38]。吸收 IBP 的研究成果将其应用于混合模型文本聚类，从而实现 DPM 的软聚类将是一个重要的研究方向。

层次语义聚类。经典的 DPM 为扁平式方法，《中图法》等诸多分类法的类目体系为层次语义关系，因此，聚类算法需要揭示文本集的层次主题结构。Neal 等[39]提出的 Dirichlet Diffusion Trees (DDT)基本思想是：数据点根据先前数据点的路径运行，根据风险函数形成分叉，最终生成二叉树。基于 Pitman Yor process，Knowles 等[40]提出的 Pitman-Yor diffusion tree(PYDT)基本思想是：新的实例在分支点可以遵循已有分支中的任意一支，也可以开始新的分支，从而可以实现聚类结果的多叉树展示。目前，基于 DPM 的层次语义聚类研究与应用尚少，有待进一步深入。

模型整合。DP 与其他模型和算法的整合研究与应用已经取得了较好的成效，比如 Liang 等[41]将 HDP 与 PCFG 相结合实现了句法分析中语法标签数量与过拟合间的平衡；Goldwater 等[42]将 DP 与一元语言模型结合用于语音分段，避免了需要预先设定词形(word type)数量的问题，还将 HDP 与二元语言模型结合，实现上下文依赖的语音分段，在与 NGS 的对比实验中，分段准确率分别为 72.3、68.3，词形准确率分别为 59.1、55.7；在 n-gram 语言模型中，n 越大则模型的表征能力越强，但也更容易出现过度拟合，Teh[43]基于 Hierarchical Pitman-Yor Processes 提出的 HPYCV 方法取得了较好的效果。DP 与语言模型相结合研究的成果理应为混合模型文本聚类提供新的视角。

5 模型推理

5.1 EM 算法

在公式(2)的基础上，定义隐性变量 $Z=\{z_1,$

$z_2,\cdots,z_N\}$，$z_n\in\{1,2,\cdots,K\}$，$z_n=k$ 表示文本 d_n 所对应的簇为 k，$\{(d_n,z_n)\}_{n=1}^{N}$ 代表完整的数据集，基于完整的对数极大似然公式（或是最大后验概率，公式 11）[24] 即可进行 EM 的迭代过程，EM 算法的具体描述参见文献[1,2]。在混合模型文本聚类中 EM 算法多用于有限混合模型的推理，研究主要涉及以下几个方面。

$$L(\Theta\mid D,z_{nk})=\sum_{n=1}^{N}\log\sum_{k=1}^{K}z_{nk}\pi_k p_k(d_n\mid\theta_k)p(\theta_k) \tag{11}$$

分配策略的影响。混合模型文本聚类中使用最多的分配策略是软分配、硬分配（CEM）以及决定性退火技术（Deterministic Annealing，DA），分配策略的差异对混合模型存在不同的影响。Zhong 等[1] 对三者进行了实验比较，结果发现大部分数据集中 DA 优于软分配，后者又优于硬分配，其中 DA 对 vMF 的改善更为明显，但在 BM 中反而有所下降。Meila 等[44] 比较了基于 MM 的软分配与硬分配，得到了与 Zhong 等[1] 类似的结果。据此，可以得到聚类算法的质量与分配的“柔软度（softness）”相关的推论。在时间复杂度上，Zhong[1] 对三种分配策略测试的结果显示 DA 的时间复杂度远高于另外两者，硬分配最低，对 vMF 的影响尤甚。目前，z_{nk} 值的计算大多通过当前参数获得，改进的计算方法则可以通过当前的条件概率密度模拟得到，如随机 EM（stochastic EM，SEM）[45]、模拟退火 EM 与蒙特卡洛 EM[46] 等。

K 值。K 值是混合模型聚类和传统分割式聚类都需面对的问题，已经形成了丰富的研究成果[47]。在混合模型文本聚类中通常也沿用以往的方法对 K 值进行估计，但为了更好地解决该问题，学界做出了三个方向的探索：一是在模型推理中引入簇的修剪机制，即初始化一个较大的 K，经过不断修剪得到最终合适的值；比如，Zeng 等[3] 在迭代中将 π_k 值低于最小阈值的成分从模型中剔除，最终得到一个成分数量较为合理的簇结构，而且避免了传统 EM 的稀疏簇问题。Figueiredo 等[48] 基于类似的机制改变了传统 EM 算法的极大似然估计或最大后验概率估计，尝试采用最小信息长度（MML）准则来平衡模型复杂度与拟合质量，也得到了较好的结果。该方法能够自动选择簇的数目，且不需要精确的初始化，同时避免了参数空间的边际问题。二是采用基于混合模型的凝聚式聚类[21,44] 或分割式聚类[49]，通过层次式方法避免 K 值确定问题；三是将有限混合模型扩展为无限，使用 DP 对文本建模，K 值作为一个未知参数存在，直到算法结束 K 值才最终确定，从而将 K 值转化为非参数估计问题。

特征空间的影响。Rigouste 等[13] 的实验表明，特征维度的减少能降低模型推理结果的困惑度（Perplexity），并且能明显增强参数估计的稳定性，减少初始化条件对算法的影响。基于此，Rigouste 等提出了两种策略：一是根据词频对词表进行分区，从最小分区开始在迭代过程中逐步增大词表规模；二是舍弃稀疏词，通过多次实验选择最佳结果所对应的词表。在后续研究中 Rigouste 等[50] 进一步尝试将两种策略相结合，得到了更优的结果。

5.2 吉布斯抽样

作为马尔可夫链－蒙特卡洛方法族的一员，吉布斯抽样为 DPM 混合模型的推理提供了非常有效的方法[51]，研究表明该方法在优化模型聚类效果方面优于 EM 算法[52,53]，算法的具体描述参见文献[54]。Rigouste 等[13] 在基于多项式混合模型的文本聚类中运用了吉布斯抽样以及 Rao-Blackwellized 吉布斯抽样；结果证明，Rao-Blackwellized 抽样效果优于吉布斯，且收敛更快；吉布斯抽样与 EM 算法相近，但后者更加依赖于初始化。Huang 等[7] 在基于多项式分布的 DPM 文本聚类中运用了 Blocked 吉布斯抽样，该方法将特征组成若干块（Block），在迭代过程中每次更新一个块中的特征，与吉布斯相比收敛速度有所提高，当特征间相互依赖时，该方法更有效率。

吉布斯抽样由于灵活、简单且易于实现等特点被广泛应用，但也存在 MCMC 方法两个主要的不足：①收敛速度缓慢，计算花销大；②收敛状态难以确定[55]。尤其在处理大规模文本时算法时间复杂度更高，因而 DPM 的进一步推广有赖于推理算法的改进或者模型的简化。相关的探索有：Liu[56] 提出的 Collapsed 吉布斯抽样，是对一部分随机特征进行积分从而加快收敛速度。Gilks 等[57] 提出的基于 DPM 的 split-merge 抽样算法实现了在一个步骤中批量更新多个变量，但该方法在更新变量的过程中并未使用马氏链中的所有信息，导致了收敛速度慢以及碎片类的问题。基于 Collapsed 吉布斯，丁轶群[58] 提出了自适应分裂合并抽样算法（Adaptive split-merge sampler，ASM）并将其运用于文本主题建模，实证表明 ASM 能有效提高收敛速度，解决了 split-merge 抽样算法的碎片类问题，并且得到更优

的聚类结果。

5.3 变分推理

相对于EM、吉布斯抽样，变分推理是一种确定性近似推理方法，在混合模型文本聚类中应用较多的是均值场变分推理，算法的具体描述参见文献[59]。

DPM中的应用。Huang等[7]在多项式分布的DPM文本聚类中将均值场变分[60]与Blocked吉布斯做了对比，聚类结果的NMI分别为0.945、0.979，变分方法准确率略低于抽样方法，时间维度吉布斯方法运行20次需要3小时，而变分方法只需10分钟，后者远优于前者；Nguyen等[5]在DPM文本聚类中采用了一种改良的均值场变分方法，也得到了较好结果。变分推理除了时间方面的优势之外，计算也更为方便[61]，在各种应用中表现出了较好的泛化性能，即无论参数估计还是非参数估计都能得到满意的结果[59]。相关研究的例子如，Fan等[32]基于狄利克雷分布的DPM（即无限狄利克雷混合模型）从stick-breaking模型的视角进行建模，提出运用变分方法进行模型推理，在仿真数据、图像检测、视频分类等数据集中的试验都得到了较好效果。Ma等[62]把伽马分布作为狄利克雷分布的参数先验，用若干相互独立的伽马分布近似参数的先验分布与后验分布，在仿生数据实验中该方法优于EM算法。

在线变分推理。在线变分推理是变分方法的一个研究方向，最早用于LDA以及HDP主题模型[63,64]。在大规模流数据应用情境，学术界采用了随机最优化方法以降低算法的时间复杂度，即通过不断重复子抽样，仅根据获得的子数据集调整变分参数，从而避免在每次迭代过程中遍历所有数据点。Bryant等[65]进一步提出了split-merge在线变分推理，使得截断水平（truncation level）可以动态变化；Fan等[66]将特征选择技术[67]引入狄利克雷混合模型（DM）的在线变分推理，在文本聚类与图像聚类的实验中都取得了较好的效果。

非共轭模型。均值场变分方法存在一些潜在的不足[59,68]：①无法识别隐藏变量间的关系；②后验方差被低估；③难以适用于非共轭模型。针对第三点，Wang等[69]针对特定非共轭模型提出了两种变分均值场近似，即Laplace变分推理和delta方法变分推理；Knowles等[70]通过推导出更低的边界以近似所需的期望，提出了一种变分信息传播（variational message passing）算法；Paisley等[71]使用由蒙特卡洛积分得到的具有较低边界倾斜度特征的随机近似完成参数推理，同时使用基于控制变量的方差衰减方法以减少所需用于构建随机搜寻方向的样本。

5.4 一点思考

推理性能。由于EM操作简单且效果较好，在混合模型推理中取得了广泛的应用。在DPM推理中则多使用吉布斯抽样与变分推理。就性能而言，多数实验结果表明，吉布斯抽样与变分推理优于EM；当测试条件为小样本时，吉布斯抽样的推理效果优于变分推理，但面对大样本时，Gao等[72]的结果表明变分方法更能适应大规模数据集的情况；在时间复杂度层面，变分方法远远优于吉布斯抽样。

非批量式算法。经典的推理算法均为批量式方法，即在每次迭代过程中都需要遍历完整数据集对模型参数进行更新，当面对大规模高维度数据集时，时间复杂度较高，且推理效果没有必然优势。Zhong[73]认为在线参数更新能够改善模型效果，且更适合于数据流环境，Nigam等[74]研究表明增量式方法比批量式方法效果更好，Rigouste等[13]的特征增量式EM算法也证明了这一点。这些研究启示可以从增量式视角改进混合模型聚类研究。

初始化。推理算法首先需要对模型参数赋予初值，不同的初始化可能导致不同的结果，尤其对依赖初始化条件的EM。目前仍没有普遍有效的初始化方法，Meila等[44]对比了随机方法、边际似然对数以及凝聚聚类初始化方法在EM文本聚类中的效果：在仿真数据中，随机方法劣于另外两种；在真实数据上，三者无明显差异；凝聚聚类与边际似然对数方法性能相仿，但后者更有效率，故边际似然对数是最优的选择。Zhong[75]对比了随机方法、边际似然对数、PERTUB以及KKZ初始化方法，结论显示边际似然对数和PERTUB与随机方法性能伯仲，KKZ优于前面三者。不过，多数实验研究中仍主要采取随机方法，具体做法是在相同条件下以不同的初始化值反复实验，最终取结果的均值或最优结果。

6 总 结

从文本建模、参数建模、模型推理的研究文献中可以发现：参数建模中的狄利克雷过程可以应用于连续型以及离散型混合模型，模型推理中的EM算法、变分推理以及马尔可夫链－蒙特卡洛方法也可

以应用于两类文本建模方法之中，从而展现出三者间交叉组合的技术特点，也突显出混合模型聚类的灵活性。Zhong[1]比较了 EM 算法在不同模型中的性能，Fraley 等[52,53]比较了推理算法在单一文本建模方法中的性能，三种推理算法在两类文本建模中的性能比较尚缺乏实证证据，将成为本文进一步的研究方向。此外，在文本建模、参数建模以及模型推理研究中尚存在一些共性的问题。

(1)特征选取

在自然语言处理的早期，文本特征通常表示为词袋模型，即将其看作是一系列独立项的集合，忽略其词序、语法与句法，项相互独立，即一元语言模型。研究者已经发现一元项对主题的表征能力有限，吴夙慧等对此进行了系统的梳理[76]。在混合模型文本聚类中，多数研究者仍主要采取一元模型，文本表示的相关探索甚少，少量的研究如 Liu 等[25]基于 GMM 同时使用项、命名实体以及词组构建文本特征向量，相比使用单一的项聚类效果有所提高。该研究提示，在混合模型文本聚类中可以借鉴传统聚类通过短语、项共现、主题等从语义层面揭示文本内容、提高文本表征准度的做法。在文本内容特征之外，也可以考虑诸如用户标注[77]、链接或引用等相关信息，从多特征的视角选取更丰富的文本表征信息。

(2)特征降维

文本集大多是稀疏的大规模、高维度数据，混合模型文本聚类在高维空间的表现往往并不理想[78]，在实际应用中需要借助特征选择与特征抽取等降维技术减少特征维度。

特征抽取。在文本聚类领域广泛使用的特征抽取方法包括 LSA、PLSA、LDA 等。Masada 等[79]比较了 LDA、PLSA 以及随机映射的降维效果，实验表明 LDA 与 PLSA 比随机映射提供了更好的聚类结果，但二者之间效果差异不大，LDA 模型推理的时间复杂度较大。Nguyen 等[80]提出了基于项聚类的降维思路，该方法与文本频率(DF)、项共现(TC)、LSA 等方法比较显示：其效果远优于 DF 和 TC，与 LSA 相差不大，且在部分试验中优于 LSA；该方法的不足是对初始化较为敏感，与 LSA 的共同点是子主题的数量需事先确定。Pessiot 等[77]认为在同一上下文中出现相同频次的共现项则语义相关，以此对特征项进行分类实现降维，基于该方法的 MM 与原始 MM、基于 PLSA 的 MM、PLSA 以及 LDA 在 NG20 测试集上平均准确率与运行时间分别为 0.4、20min，0.35、10min，0.4、5min，0.4、3min，0.32、30min，该方法与 PLSA 聚类准确性相当，但不具备时间优势，从而提示 PLSA 的特征抽取效果总体最优。

特征选择。常见的特征选择方法有文本频率(DF)、项共现(TC)、项强度(TS)以及信息熵(En)等。与传统聚类算法不同，基于混合模型的聚类可以把特征选择转换为模型推理问题，即引入一个隐性二分向量 $\gamma=\{\gamma_1,\gamma_2,\cdots,\gamma_w\}$ 用于区分相关特征与非相关特征。Huang 等[6,7]在多项式混合文本聚类中假设判别词与非判别词之间不相关，然后分别对其进行特征建模，实证结果显示采用该特征选择方法的聚类结果在大部分数据集上优于 EDCM 以及 MM。Huang 等[6,7]的方法都默认不同的簇中特征的显著程度均一致，即对所有簇都选择相同的特征子集。但因为每个簇实际的主题差异，其中的特征分布不尽相同，相关的特征子集也不同，因而需要采用子空间聚类。Li[12]基于 MM 将指示变量 γ_w 细化为 γ_{wk}，对于每个 $w=1,2,\cdots,W$，$\gamma_{wk}=\begin{cases}1,\text{第 } w \text{ 个特征在簇 } k \text{ 中为相关特征}\\0,\text{第 } w \text{ 个特征在簇 } k \text{ 中为不相关特征}\end{cases}$。结果显示，其聚类效果显著优于 K-means 以及谱聚类。

上述研究可见，特征选择与特征抽取方法在实现降维的同时都不同程度地改进了聚类效果；特征抽取的相关研究结果显示 PLSA 以及 LSA 的聚类效果更值得期待；Bouveyron 等[24]研究指出“希望读者确信，使用简约模型、子空间聚类或者变量选择方法而非用降维进行预处理”，从而提示应优先选择特征选择而非特征抽取方法。

(3)半监督聚类

文本往往包含部分标签，或是其他约束条件，这类附加信息在半监督聚类中可以作为种子(即初始化条件)、限制(聚类过程中标签不变)或是反馈(聚类结束后根据标签进行调整)等用于改善聚类质量。Basu 等[81]基于 skmeans 的对比实验发现，限制方法与种子方法效果相当；Zhong[75]基于 MM 的实验表明，在标签完整的情况下，限制方法优于其他两种；当标签不完整，甚至无法包括所有簇的情况下，反馈方法聚类效果更优。半监督聚类的重点在于发现未标签文本中的新簇，Zhong[75]提出了采用多元聚类或是条件信息瓶颈方法的思想。两项研究中使用的特定种子、限制或是反馈方法只是众多可能的设计之一，后续的研究中可以考虑采用其他特定方法或是将多种方法融合应用。

(4)聚类过程的系统整合

相关研究表明,特征选择、K值确定以及模型推理等几个步骤之间存在相互影响[82],因而,将特征选择、K值确定融合到模型推理过程中,在模型迭代推理的同时特征空间以及K值也相应地调整变化,将模型推理的过程整合为一个动态系统。Law等[4]、Kersten等[83]将K值确定、特征选择融入到了EM算法中,在与SFFS-EM[84]、RSEM[85]等方法的比较中,其运行时间以及聚类质量都有较好的表现;Constantinopoulos等[86]在变分推理中将模型选择与特征选择相结合,同时得到聚类个数与特征显著性程度,实验结果表明该方法在高维稀疏数据集上更加稳健。变分方法在模型推理中融入了模型选择与特征选择,例如Fan等[87]把基于簇分割的K值选择[88]融合到DM的变分推理中,在文本聚类的实验中该方法的聚类准确率优于其他。上述研究显示聚类过程的系统整合能够有效的提高聚类效果,不过该领域的研究与实验尚少,有待深入探讨。

总而言之,混合模型文本聚类中使用的基础模型主要是BM、MM以及vMF,在研究与应用中为了提高建模的精度或增加模型功能,需要对基础模型进行优化。具体的研究路径可以是模型的复杂化,例如引入新参数用于特征选择,增加模型层次进行参数建模,或是不同模型的综合应用等。目前,混合模型都受到严格的独立性假设的制约,对簇之间、文本之间以及特征之间的依赖关系建模是进一步的研究方向。另一条研究路径是模型的简化,随着模型越来越"精密",推理计算的可行性以及易行性是必须面对的问题,例如vMF中参数κ以及DPM,在提高计算效率的同时,增加模型的稳定性。此外,更加开拓性的方向是尝试新的混合模型,根据文本的性质,在vMF的基础上进一步对球形分布[89]进行拓展研究。

参 考 文 献

[1] Zhong S, Ghosh J. Generative model-based document clustering: A comparative study [J]. Knowledge and Information Systems, 2005,8(3):374-384.

[2] Dempster A P,Laird N M,Rubin D B. Maximum likelihood from incomplete data via the EM algorithm (with discussion) [J]. Journal of the Royal Statistical Society, y. SeriesB (Methodological), 1977,39(1):1-38.

[3] Zeng Hong, Cheung Yiu-ming. Learning a mixture model for clustering with the completed likelihood minimum message length criterion[J]. Pattern Recognition, 2014, 47(5): 2011-2030.

[4] Law M H C,Figueiredo M A T, Jain A K, Simultaneous feature selection and clustering using mixture models[J]. IEEE Trans. Pattern Anal. Mach. Intell, 2004,26(9): 1154-1166.

[5] Nguyen KimAnh,Nguyen The Tam,Ngo Van Linh. Document Clustering using Dirichlet Process Mixture Model of von Mises-Fisher Distributions [C]//Proceedings of the Fourth Symposium on Information and Communication Technology. New York, USA,2013:131-138.

[6] Yu G,Huang R,Wang Z. Document Clustering viaDirichlet Process Mixture Model with Feature Selection [C]// *Proceedings of the 16th ACM SIGKDD International* Conference on Knowledge Discovery and Data Mining. New York,USA, 2010:763-772.

[7] Huang Ruizhang, Yu Guan, Wang Zhaojun, et al. Dirichlet process mixture model for document clustering with feature partition [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25 (8): 1748-1759.

[8] Nigam K,McCallum A K,Thrun S,et al. Text classification from labeled and unlabeled documents using EM [J]. Machine Learning,2000,39(2-3):103-134.

[9] Andrew McCallum,Kamal Nigam. A comparsion of event models for naive Bayes text classification[C]//AAAI-98 Workshop On Learning For Text Categorization,Madison, Wisconsin,USA, 1998:41-48.

[10] Bouguila N. On multivariate binary data clustering and feature weighting[J]. Computational Statistics and Data Analysis, 2010,54(1):120-134.

[11] Zhu S,Takigawa I,Zhang S,et al. A probabilistic model for clustering text documents with multiple fields[M]. Berlin Heidelberg: Springer, 2007.

[12] *Li Minqiang,Zhang Liang. Multinomial mixture model with* feature selection for text clustering [J]. Knowledge-Based Systems, 2008,21(7):704-708.

[13] Rigouste L,Capp'e O,Yvon F. Evaluation of a Probabilistic Method for Unsupervised Text Clustering [C]// International Symposium on Applied Stochastic Models and Data Analysis. Brest, France,2005:114-123.

[14] Li X, Yu G, Wang D. MMPClust: A skew prevention algorithm for model-based document clustering [C]// Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2005: 536-547.

[15] Salton G,McGill M J. Introduction to Modern Retrieval [M]. 2nd. New York:McGraw-Hill Book Company,1983.

[16] Dhillon I S, Modha D S. Concept decompositions for large sparse text data using clustering [J]. Machine Learning, 2001,42(1):143-175.

[17] Banerjee A, Ghosh J. Frequency sensitive competitive learning for clustering on high-dimensional hyperspheres [J]. IEEE Transactions on Neural Networks, 2004, 15 (3):702-719.

[18] Banerjee A, Dhillon I, Ghosh J, et al. Generative Model-based Clustering of Directional Data[C]//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discoveryanddata mining. New York, USA, 2003:19-28.

[19] Mardia K V, Jupp P E. Directional statistics[M]. New York: John Wiley & Sons, 2009.

[20] Suvrit Sra. A short note on parameter approximation for von Mises-Fisher distributions:and a fast implementation of Is(x) [J]. Comput Stat,2012, 27(1):177-190.

[21] Vaithyanathan S,Dom B. Model-based hierarchical clustering [C]//Proc. 16th Conf. Uncertainty in Artificial Intelligence. San Francisco, CA, USA ,2000:599-608.

[22] Novovičová J, Malík A. Application of multinomial mixture model to text classification [M]//Pattern Recognition and Image Analysis. Berlin Heidelberg: Springer, 2003: 646-653.

[23] Volodymyr Melnykov . Finite mixture models and model-based clustering [J]. Statistics Surveys, 2010, 4: 80-116.

[24] Bouveyron Charles, Camille Brunet-Saumard. Model-based clustering of high-dimensional data _ A review [J]. Computational Statistics and Data Analysis, 2014, 71: 52-78

[25] Liu X, Gong Y, Xu W, et al. Document clustering with cluster refinement and model selection capabilities [C]//Proc. 25th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval. New York, USA,2002:191-198.

[26] Madsen R, Kauchak D, Elkan C. Modeling Word Burstiness Using the Dirichlet Distribution[C]//Proc. Int'l Conf. Machine Learning. New York, USA,2005: 545-552.

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

[28] Minka T. Estimating a Dirichlet distribution[OL]. http://research. microsoft. com/en-us/um/people/minka/papers/dirichlet/. [2014-7-22].

[29] Elkan C. Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution[C]//Proceedings of the 23rd international conferenceonMachinelearning. New York, USA,2006: 289-296.

[30] Bouguila N,Ziou D. High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007, 29(10): 1716-1731.

[31] Bouguila N. Clustering of Count Data Using Generalized Dirichlet Multinomial Distributions [J]. IEEE transactions on knowledge and data engineering,,2008, 20(4):462-474.

[32] Fan Wentao, Bouguila Nizar. Variational learning for Dirichlet process mixtures of Dirichlet distributions and applications[J]. Multimed Tools Appl,2014, 70(3): 1685-1702.

[33] Blackwell D, MacQueen J. Ferguson distribution via Polya urn schemes[J]. The Annals of Statistics, 1973,1(2): 353-355.

[34] Green P J, Richardson S. Modelling Heterogeneity with and without the DirichletProcess[J]. Scandinavian J. Statistics, 2001, 28(2):355-377.

[35] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical Dirichlet processes [J]. Journal of the American Statistical Association, 2006,101(476):1566-1581.

[36] Hu Weiming, Tian Guodong, Li Xi, et al. An Improved Hierarchical Dirichlet Process-Hidden Markov Model and Its Application to Trajectory Modeling and Retrieval [J]. Int J Comput Vis, 2013,105:246-268.

[37] Griffiths T L, Ghahramani Z. Infinite latent feature models and the indian buffet process [R]. London: Gatsby Computational Neuroscience Unit,2005.

[38] Griffiths T L, Ghahramani Z. The indian buffet process: an introduction and review [J]. Journal of Machine Learning Research, 2011,12:1185-1224.

[39] Neal R M. Density modeling and clustering using Dirichlet diffusion trees [J]. Bayesian Statistics, 2003, 7: 619-629.

[40] Knowles D A, Ghahramani Z. Pitman-Yor diffusion trees [J]. arXiv preprint arXiv:1106.2494, 2011.

[41] Liang P, Petrov S, Jordan M I, et al. The Infinite PCFG Using Hierarchical Dirichlet Processes [C]//EMNLP-CoNLL. Prague, Czech Republic, 2007: 688-697.

[42] Sharon G, Griffiths T L, Johnson M. A Bayesian framework for word segmentation: Exploring the effects of context [J]. Cognition,2009,112: 21-54.

[43] Teh Y W. A hierarchical Bayesian language model based on Pitman-Yor processes[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. USA: Association for

Computational Linguistics, 2006: 985-992.

[44] Meila M, Heckerman D. An experimental comparison of model-based clustering methods[J]. Machine Learning, 2001,42(1-2):9-29.

[45] Bordes L, Chauveau D, Vandekerhove P. A stochastic EM algorithm for a semiparametric mixture model[J]. Comput. Stat. Data An. 2007, 51(11): 5429-5443.

[46] Biernacki C, Celleux G, Govaert G, et al. Model-based cluster and discriminant analysis with the MIXMOD software[J]. Comput. Stat. Data An, 2006, 51(2): 587-600.

[47] 吴夙慧，成颖，郑彦宁，等. K-means算法研究综述[J]. 现代图书情报技术,2012,5:28-35.

[48] Figueiredo M A T, Jain A K. Unsupervised learning of finite mixture models[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2002,24(3):381-396.

[49] Ververidis D, Kotropoulos C. Gaussian mixture modeling by exploiting the Mahalanobis-distance [J]. IEEE Trans. Signal Process. 2008,56(7):2797-2811

[50] Rigouste L, Cappe' O, Yvon F. Inference for probabilistic unsupervised text clustering [C]//Proceedings of the IEEE Workshop on Statistical Signal Processing. Bordeaux, France,2005,387-392.

[51] Chen Liyuan, Brown S D. Bayesian estimation of membership uncertainty in model-based clustering [J]. Journal of Chemometrics,2014,28(5): 358-369.

[52] Fraley C, Raftery A E. Model-based clustering, discriminant analysis and density estimation [J]. J. Am. Stat. Assoc, 2002,97(458): 611-631.

[53] Bensmail H, Meulman J J. Model-based clustering with noise: Bayesian inference and estimation [J]. J. Classif., 2003,20(1):49-76.

[54] Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images [J]. IEEE Transactions on Pattern Analysis and Machine, 1984,6(6):721-741.

[55] 周建英，王飞跃，曾大军. 分层 Dirichlet 过程及其应用综述[J]. 自动化学报,2011,37(4):389-404.

[56] Liu J S. The collapsed gibbs sampler in Bayesian computations with applications to a gene regulation Problems [J]. Journal of the American Statistical Association, 1994,89(427):958-966

[57] Gilks W R, Gareth R O. Strategies for improving mcmc. Markov Chain Monte Carlo in Practice[M]. London: Chapman & Hall, 1996:89-114.

[58] 丁轶群. 基于概率生成模型的文本主题建模及其应用[D]. 浙江大学,2010

[59] Sun Shiliang. A review of deterministic approximate inference techniques for Bayesian machine learning[J]. Neural Comput & Applic, 2013, 23(7-8):2039-2050

[60] Blei D, Jordan M. Variational Inference for Dirichlet Process Mixtures[J]. Bayesian Analysis, 2006, 1(1): 121-144.

[61] Ma Z, Leijon A. Bayesian estimation of beta mixture models with variational inference [J]. IEEE Trans Pattern Anal Mach Intell, 2011, 33(11):2160-2173.

[62] Ma Z, Rana P K, Taghia J, et al. Bayesian estimation of Dirichlet mixture model with variational inference[J]. Pattern Recognition, 2014, 47(9):3143-3157.

[63] Hoffman M, Blei D, Wang C, et al. Stochastic variational inference[J]. J Mach Learn Res, 2013, 14: 1303-1347.

[64] Wang C, Paisley J, Blei D. Online variational inference for the hierarchical Dirichlet process[C]//Proceedings of the 14th international conference on artificial intelligence and statistics, Fort Lauderdale, USA, 2011, 15:752-760.

[65] Bryant M, Sudderth E. Truly nonparametric online variational inference for hierarchical Dirichlet processes[C]//Adv Neural Inf Process Syst. Cambridge: MIT Press, 2012, 25:2708-2716

[66] Fan Wentao, Bouguila Nizar. Online variational learning of generalized Dirichlet mixture models with feature selection[J]. Neurocomputing, 2014, 126: 166-17.

[67] Boutemedjet S, Bouguila N, Ziou D. A hybrid feature extraction selection approach for high-dimensional non-Gaussian data clustering [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2009, 31(8): 1429-1443.

[68] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nestedLaplace approximation[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2009,71:319-392

[69] Wang C, Blei D. Variational inference in nonconjugate models[J]. J Mach Learn Res, 2013, 14: 1005-1031.

[70] Knowles D A, Minka T. Non-conjugate variational message passing for multinomial and binary regression [C]// Advances in Neural Information Processing Systems. Association for Computational Linguistics Stroudsburg, PA, USA, 2011: 1701-1709.

[71] Paisley J, Blei D, Jordan M. Variational Bayesian inference with stochastic search [C]//Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland, UK, 2012:1-8

[72] Gao Jianfeng, Mark Johnson. A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS

taggers[C]//Pro of the 2008 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics Stroudsburg, PA, USA, 2008:344-352.

[73] Zhong S. Efficient online sphericalk-means clustering [C]//Proc. IEEE Int. Joint Conf. Neural Networks. Montreal, Canada,2005: 3180-3185.

[74] Nigam K, Ghani R. Understanding the behavior of co-training[C]//Proceedings of KDD-2000 Workshop on Text Mining. Boston, Massachusetts, USA, 2000: 15-17.

[75] Zhong Shi. Semi-supervised model-based document clustering: A comparative study [J]. Mach Learn, 2006,65(1):3-29

[76] 吴夙慧,成颖,郑彦宁,等.文本聚类中文本表示和相似度计算研究综述[J].情报学报,2012,30(4):622-627.

[77] Pessiot Jean-François, Kim Young-Min, Amini M R, et al. Improving document clustering in a learned concept space [J]. Information Processing and Management,2010,46: 180-192.

[78] Charles Bouveyron,Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review [J]. Computational Statistics and Data Analysis, 2014, 71: 52-78

[79] Masada T, Kiyasu S, Miyahara S. Comparing LDA with pLSI as a dimensionality reduction method in document clustering [M]//Large-Scale Knowledge Resources. Construction and Application. Springer Berlin Heidelberg, 2008: 13-26.

[80] Nguyen DucThang, Chen Lihui, Chan CheeKeong. Feature Reduction using Mixture Model of Directional Distribution[C]//Proc. 10th Intl. Conf. on Control, Automation, Robotics. Hanoi, Vietnam, 2008:17-20.

[81] Basu S,Banerjee A,Mooney R. Semi-supervised clustering by seeding [C]//Proc. 19th Int. Conf. Machine Learning. Sydney, Australia,2002:19-26.

[82] Dy J G, Brodley C E. Feature subset selection and order identification for unsupervised learning [C]// Proceedings of the 17th International Conference on Machine Learning. Burlington: Morgan Kaufmann, 2003: 247-254.

[83] Kersten Jens. Simultaneous feature selection and Gaussian mixture model estimation for supervised classification problems [J]. Pattern Recognition, 2014, 47 (8): 2582-2595.

[84] Kersten J. Ein Rahmenwerk zur interaktiven Klassifikation hochaufloesender optischer Satellitenbilder mittels graphenbasierter Bildmodellierung [D]. Technical University Berlin, 2011

[85] Zhao Q, Hautamäki V, Kärkkäinen I, et al, Random swap {EM} algorithm for gaussian mixture models[J]. Pattern Recognit. Lett., 2012, 33 (16): 2120-2126.

[86] Constantinopoulos C, Titsias M K, Likas A. Bayesian feature and modelselection for Gaus sian mixture models [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2006,28(6):1013-118

[87] Fan Wentao, Bouguila Nizar, Ziou Djemel. Variational learning of finite Dirichlet mixture models using component splitting[J]. Neurocomputing, 2014, 129: 3-16.

[88] Corduneanu A, Bishop C M. Variational Bayesian model selection for mixture distributions [C]//Artificial intelligence and Statistics. Waltham, MA: Morgan Kaufmann, 2001, 2001: 27-34.

[89] Mardia K V. Statistics of directional data[J]. J. Roy. Statist. Soc. Ser. B (Methodological), 1975,37(3): 349-393.

附录1 核心文献

	被引频次	类别	期刊/会议	参考文献编号	时间
核心期刊	230	A	International Conference on Machine Learning	83	2000
	114	A	ACM SIGIR Conf. on Research and Development in Information Retrieval	25	2002
	93	A	International Conference on Machine Learning	36	2006
	75	A	ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	17	2003
	20	A	ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	6	2010
	40	A	IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING	39	2008
	7	A	IEEE Transactions on Knowledge and Data Engineering	7	2013
	70	B	Conf. Uncertainty in Artificial Intelligence	20	2000
	60	B	Machine Learning	75	2006
	172	B	Machine Learning	87	2001
	137	B	Knowledge and Information Systems	1	2005
	42	B	IEEE TRANSACTIONS ON NEURAL NETWORKS	16	2004
	34	B	Information Processing and Management	11	2007
	9	B	Information Processing and Management	76	2010
	16	C	Knowledge-Based Systems	10	2008
	3	C	Advances in Information Retrieval	22	2007
	3	C	NeuralComput & Applic	59	2013
高被引	0	C	Intelligent Data Analysis	81	2014
	2352		AAAI-98 Workshop on Learning for Text Categorization	9	1998
	1821		Journal of the·American Statistical Association	52	2011
	38		Large-Scale Knowledge Resources. Construction and Application	79	2008
	22		Computational Statistics and Data Analysis	78	2014
	17		Pattern Recognition and Image Analysis	21	2003
	10		Computational Statistics and Data Analysis	13	2010

（责任编辑 赵 康）