面向复杂主题建模的流式层次狄里克雷过程

韩忠明[1),2)] 张梦玫[1)] 李梦琪[1)] 段大高[1)] 陈 谊[1),2)]

[1)](北京工商大学计算机与信息工程学院 北京 100048)

[2)](食品安全大数据技术北京市重点实验室 北京 100048)

摘 要 互联网已经成为真实事件信息的主要来源.针对互联网海量新闻语料的主题挖掘是新闻事件的组织和追踪任务中关键的一环.主题模型已被广泛应用于挖掘和分析新闻等文本语料,LDA(Latent Dirichlet Allocation)是最常见的主题模型,然而现有基于 LDA 的方法没有考虑到主题之间的层次关系,且需要预先提供主题个数.作为 LDA 模型的扩展,层次狄里克雷过程(Hierarchical Dirichlet Process,HDP)是非参数贝叶斯主题模型,HDP 能够自动确定主题个数.对于具有层次等特性的复杂主题,HDP 难以挖掘出隐式层次结构,且容易产生噪音主题.为了解决这个问题,该文提出了基于 HDP 改进的非参数贝叶斯模型:流式层次狄利克雷过程(Flow Hierarchical Dirichlet Process,FHDP),FHDP 通过在 HDP 模型中加入流动操作,加强了对主题之间的同属领域信息的利用,以便于更好的对主题进行层次分析.利用加入了流动操作的中国连锁餐馆模型(Chinese Restaurant Franchise,CRF)对数据进行建模,设计相应的马尔可夫链蒙特卡罗(Markov Chain Monte Carlo,MCMC)采样方法,以推导 FHDP 模型的分布参数分布.FHDP 的主要贡献在于:(1)对含有层次关系的主题建模时,减少了无意义信息.解决了 HDP 得到主题不明确的问题,扩大了 HDP 的应用领域;(2)由于在 FHDP 中加强了对主题隐含领域信息的利用,主题的层次关系变得更加明确.为了客观衡量 FHDP 和 HDP 的性能差异,利用模拟和真实数据进行了大量实验.实验表明,在轮廓系数、主题覆盖度、单字对数似然等指标上,FHDP 模型明显优于 HDP 模型.

关键词 层次狄里克雷过程;主题模型;非参数贝叶斯模型;马尔可夫蒙特卡罗;流式层次狄里克雷过程

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.01539

Flow Hierarchical Dirichlet Process for Complex Topic Modeling

HAN Zhong-Ming[1),2)] ZHANG Meng-Mei[1)] LI Meng-Qi[1)] DUAN Da-Gao[1)] CHEN Yi[1),2)]

[1)](*School of Computer and Information Engineering, Beijing Technology and Business University, Beijing* 100048)

[2)](*Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing* 100048)

Abstract Internet has become defacto information source of real world events. Algorithms to detect the topics of abundant news on the Internet precisely are very important to the organization and tracking of different events. Topic modeling provides users with methods to organize, understand and summarize large collections of textual information on Internet. Latent Dirichlet Allocation (LDA) is one of the commonest topic model methods. In the LDA model, each document is modeled as a mixture of topics that are present in the corpus. However, existing LDA-based methods generate topics without considering hierarchical relation of different topics and furthermore the number of topics need to be provided in these LDA-based methods in advance which brings great challenges to unsupervised topic detection. As an extension of the LDA model, the hierarchical Dirichlet process (HDP) is a nonparametric Bayesian approach to clustering topics in the corpus. HDP uses a Dirichlet process to capture the uncertainty in the number of topics, which means

收稿日期:2017-11-30;在线出版日期:2018-10-16.本课题得到国家自然科学基金(61170112)、北京市自然科学基金(4172016)、北京市科技计划课题(Z161100001616004)资助.**韩忠明**,博士,教授,中国计算机学会(CCF)会员,主要研究方向为互联网挖掘、大数据分析与信息检索.E-mail:hanzm@th.btbu.edu.cn.**张梦玫**,硕士研究生,主要研究方向为互联网数据挖掘.**李梦琪**,硕士研究生,主要研究方向为互联网数据挖掘、社会网络,**段大高**(通信作者),博士,副教授,主要研究方向为社会计算、多媒体信息处理.E-mail:duandg@th.btbu.edu.cn.**陈 谊**,博士,教授,主要研究领域为大数据可视分析与挖掘.

HDP can automatically select the number of topics in the corpus. However, for hierarchical text corpus with highly correlated topics, such as topics share overlapping structure and some topics have sub topics, HDP method has difficulty digging out hidden hierarchical structure and will introduce noises into topics easily. In order to solve these problems, a novel method named Flow Hierarchical Dirichlet Process nonparametric Bayesian model (FHDP for short) is proposed in this paper for modeling topics with hierarchical structure. In order to make better use of words that imply domain-specific information rather than corpus-specific words, we assume that a document is composed of some topics and different topics can flow into other topics. FHDP takes advantage of common information between topics in same domain by applying a vital operation called "flow", which produces better hierarchical analysis of topics in the corpus. We introduce a new concept "room" in Chinese Restaurant Franchise (CRF), which represents domain-specific information in the corpus. Based on the concept of "room", a new flow operation is proposed for CRF process. Based on the standard Gibbs sampling algorithms, we design a new Markov chain Monte Carlo (MCMC) algorithm to implement the sampling for the CRF with flow operation. Finally, we get the distribution parameters of FHDP model. The main contributions of FHDP are: (1) reducing meaningless information when modeling topics with hierarchical relationships. FHDP solves the problem that HDP can obtain ambiguous topic and expands the application field of HDP. (2) enhancing the utilization of implicit fields information in the word co-occurrence, as a result the hierarchical relationship of different topics becomes clearer. To evaluate the performance of FHDP, comprehensive experiments are conducted on synthetic and two real data sets. LDA and HDP are used as comparative methods. We use GridSearch method to determine the number of topics in different data sets with the lowest degree of confusion. The experiment results show that FHDP significantly outperforms HDP and LDA in terms of Silhouette Coefficient, coverages of topics, per word log likelihood and other indicators. FHDP can be used to model real corpus with hidden hierarchical structure.

Keywords hierarchical Dirichlet process; topic model; Bayesian nonparametric model; Markov chain Monte Carlo; flow hierarchical Dirichlet process

1 引 言

随着社会媒体的快速发展,互联网上的文本成为了消息传播的主要载体,每天都有海量的文本信息产生,主题模型是自动发现和探索文档中隐藏主题的重要理论,已被广泛应用于各个领域,尤其在舆情检测[1]、用户特征挖掘[2]、图像[3-4]等领域.

Latent Dirichlet Allocation(LDA)模型[5]是主题模型中被广泛应用的方法,LDA 通过训练语料能够获得相应文档-主题分布和主题-单词分布.但 LDA 需要事先给定主题个数,且得到的主题没有层次性.这为无监督的主题探测、事件发现与跟踪等问题带来了极大挑战.层次狄利克雷(Hierarchical Dirichlet Process,HDP[6])解决了这两个问题.

HDP[6]是非参数贝叶斯主题模型中的一种重要工具.HDP 作为 LDA 模型的扩展,能够发现父子主题之间的关系,依据语料库自动确定主题个数.由于不能对 HDP 模型进行直接采样,需要利用中国餐馆连锁模型(Chinese Restaurant Franchise,CRF[6])对文本进行主题建模后,用马尔可夫链蒙特卡罗方法(Markov Chain Monte Carlo,MCMC)算法对隐含分布进行采样.

互联网文本普遍有来源多、涉及领域广、包含主题丰富的特点,且同一领域内存在着大量主题.同领域的不同主题,隐含了同属领域的高层次信息,这种信息对于主题层次划分十分有用.如果能利用主题隐含的领域信息,则可以更好地对主题进行聚类,甚至挖掘出高层的广义主题(即领域级别的主题).

本文将语料库中一些具有层次关系的主题定义为复杂主题,即语料库中存在一些主题同属于某个领域.对于某个含有复杂主题的语料库,主题中既有

属于各自的主题-单词分布，也有共享的单词分布.例如以下为新闻语料库的部分文档：

文本 1(T1)：{比赛，中超，北京国安，卫冕，队员，时间，记 0 者}；；

文本 2(T2)：{比赛，活动，刘翔，田径，训练，时间，队员，前线，记者，报道}；

文本 3(T3)：{历史，文化，考古，发现，古墓，机会，时间，记者，说}；

文本 4(T4)：{文化，戏剧，精粹，历史，发扬，记者，说}.

如图 1 所示，文档 1、2 对应的主题 T1、T2 属于领域 D1，因此 T1 与 T2 对应文档中都存在着领域 D1 的词，例如示例新闻文档中的“比赛”、“队员”这类体育新闻中常常出现的词，这类词对于划分“体育”这一广义主题具有重要意义.而 T1、T2 和 T3 虽然属于不同领域，却也存在着与语料库特性 C1 相关的词，例如“记者”、“说”这类在新闻体裁中常出现的词，这种词对于该语料库的主题挖掘没有意义.

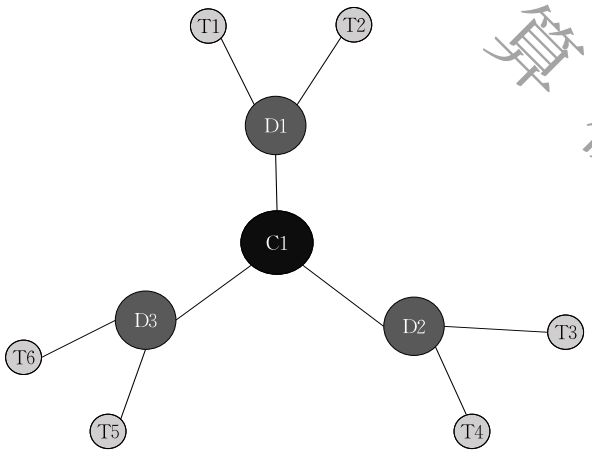

图 1　复杂主题结构图(图中新闻语料库 C1 包含了领域 D1、D2、D3 的新闻，其中每个领域又可分为两个主题，共 6 个主题)

虽然 HDP 可以用于捕获父子主题之间的关系，但由于 HDP 是一种基于词共现的模型，且没有特意利用领域 D1、D2、D3 的信息，用 HDP 对语料库进行主题建模，可能会产生类似于{记者，说，时间}这样的无意义主题，即得到的主题内容不够明确，各主题轮廓不够清晰等.因此本文提出了一个流式 HDP 模型(Flow Hierarchical Dirichlet Process，FHDP)，FHDP 在 HDP 非参数贝叶斯模型的基础上做了改进，加入了流动操作，利用可流动的 CRF(Flow Chinese Restaurant Franchise，FCRF)对数据进行建模，并设计了相应的流动 MCMC 算法，加强了各个主题中同属领域信息的流动，使得同一领域的词集中在一起.大量实验证明，FHDP 在轮廓系数、主题覆盖度、生成主题质量、主题明确性等指标中其结果要优于 HDP.

FHDP 的主要优势在于：(1)对含有层次关系的主题建模时，减少无意义信息.解决了 HDP 得到的主题不明确的问题，扩大了 HDP 的应用领域；(2)由于在 FHDP 中加强了对主题隐含领域信息的利用，因此主题的层次关系变得更加明确.

本文第 1 节简单介绍 FHDP 模型；第 2 节回顾本文方向的相关工作；第 3 节详细介绍 HDP 模型的框架以及构造方法；第 4 节介绍 FHDP 的模型框架以及改进后的 MCMC 算法；第 5 节分析相关实验；第 6 节总结本文工作并对未来进行展望.

2　相关工作

主题模型[7]是一种无监督学习方法，利用观测到的单词的分布规律，反推隐含变量，最终得到的主题由语料库中的共现词项所定义.主题模型主要分为参数贝叶斯模型和非参数贝叶斯模型.被广泛应用的 LDA 就属于参数贝叶斯模型.LDA 是由概率潜在语义索引(probabilistic Latent Semantic Analysis，pLSA[8])变种而得.

LDA 模型不能捕获父子主题之间的关系.为解决这一问题，产生了许多分层主题模型，Hierarchical LDA(HLDA)[9]、HDP、Hierarchical Pachinko Allocation Model(HPAM)[10]、Political Issue Extraction(PIE)[11]和 Guided Hierarchical Topic Model(Guided HTM)[12]等.其中，HDP 作为非参数贝叶斯模型，模型的主题个数不需要提前输入，可以随着文档数目的变化动态调整，因而被为广泛应用.HDP 模型不仅适用于文本主题挖掘，而且面对互联网各种数据类型的要求，HDP 模型对图像内容的挖掘[13]以及音频[14]、视频[15]、学术数据[16]的处理等方面都有扩展.

目前阻碍 HDP 更广泛应用的主要因素为：第一，基于 Dirichlet 过程的方法在算法实现中计算量较大，HDP 相关算法需要的时间限制了算法的应用场景；第二，对于具有层次结构的数据，HDP 表现不够好.近年来很多学者提出了基于 HDP 的扩展模型扩大 HDP 应用场景.

基于非参数贝叶斯方法的实现一般计算量较大，必须依靠近似推理的方法，比如，MCMC 采样[6]、变分推断[17].在 MCMC 采样中，构建目标分

布的马尔可夫链在平稳后用以模拟真实后验.马尔可夫链需要迭代多次才能达到平稳分布,而 MCMC 中的增量 Gibbs 采样器[18]一次只允许改变主题状态里的一个单词,即目标分布的结构不可能产生巨大变化,导致采样器拟合很慢.所以 SMHDP(Split-Merge Hierarchical Dirichlet Process)①在 MCMC 算法中加入分裂合并操作,加速马尔可夫链收敛,提升主题挖掘效果,能够发现两个相似但略有不同的主题.

对于具有层次结构的数据,文献[19]提出了一个基于 HDP 的概率主题模型,利用 HDP 学习与子故事相关的子主题,从而可以处理子故事中的细微变化.进一步地,我们观察到当语料库中存在一些主题同属于某个领域时,利用 HDP 得到的"主题-单词分布"就会出现不均衡的问题.为了解决这个问题,Guided HTM[12]将从其他语料库获得的领域知识作为"概念层"加入主题层次结构,并通过增加一个 Uniform Process 作为先验分布扩展为 DF-LDA[20].但是 GHTM 仍然需要从其他语料库中提取领域知识.受 SMHDP 模型的启发,本文在 HDP 中加入流动操作,扩展中国餐馆连锁模型(CRF[6])为可流动的 CRF(FCRF)来解决"主题-单词分布"不均衡的问题,CRF 将在 3.2 节详细介绍.FCRF 在 CRF 的餐馆中增添了"房间"元素,以便在相应的 MCMC 采样方法中实现流动操作,可导致潜在结构发生巨大变化,加强对同属领域信息的利用.

3 HDP 主题模型

Dirichlet Process(DP)具有良好的聚类性质,HDP 是一种层次化的 DP 分布,可以对多组数据进行聚类分析,本文使用双层的 HDP 模型进行主题建模.

3.1 HDP 模型

如式(1)所示,HDP 生成模型中首先使用基分布 H 和 Concentration 参数 γ 构成一层 DP,并从中抽样出分布 G_0,再使用 G_0 作为二层 DP 的基分布,结合 Concentration 参数 α_0 为每组数据抽样出相应的分布 G_j 作为该组数据所属分布的先验.

$$\begin{aligned} G_0 \mid H,\gamma &: \mathrm{DP}(H,\gamma) \\ G_j \mid G_0,\alpha_0 &: \mathrm{DP}(G_0,\alpha_0) \end{aligned} \tag{1}$$

再利用先验 G_j 构造该组的 DP 混合模型.其中 x_{ji} 为第 j 组中的第 i 个观测数据,函数 $F(\theta_{ji})$ 为给定参数 θ_{ji} 下 x_{ji} 所属的分布,θ_{ji} 由分布 G_j 抽样而得,如式(2).

$$\begin{aligned} \theta_{ji} \mid G_j &: G_j \\ x_{ji} \mid \theta_{ji} &: F(\theta_{ji}) \end{aligned} \tag{2}$$

整个过程的有向图表示如图 2 所示.图中,G_j 被定义在一个连续的空间(例如 simplex 空间),因此每组数据里的先验分布 G_j 共享相同的分布 G_0.

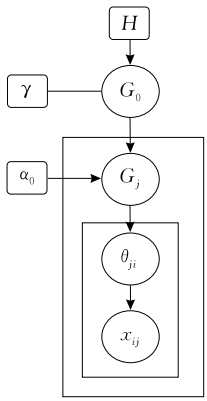

图 2 HDP 的有向图表示

在主题挖掘领域,语料库表示为文本集合 $X=\{x_1,x_2,\cdots,x_N\}$,该集合有 N 篇文档,每篇文档对应 HDP 中的一组数据,其中第 $j(1\leq j\leq N)$ 篇文档中的第 i 个单词为观测值 x_{ji},x_{ji} 符合的分布 F 为多项式分布,多项式分布参数 θ_{ji} 为一组概率值.基分布通常选取词汇表上的对称 Dirichlet 分布 $\mathrm{Dir}(\eta)$.主题 $\boldsymbol{\Phi}=(\phi_k)_{k=1}^{\infty}$ 由 $\phi_k\sim\mathrm{Dirichlet}(\eta)$ 中独立抽取出来.假设文本集合中的文档可以划分为 K 个主题,则主题挖掘的目的是从文本集合 X 中自动发现潜在的主题 $\boldsymbol{\Phi}=(\phi_k)_{k=1}^{K}$.

3.2 Chinese Restaurant Franchise 构造

在 CRF 中,中餐连锁店中每家餐厅分店共享同一份菜单,餐厅中每张餐桌只有一个菜肴,由该餐桌的第一个顾客从共享菜单 $\boldsymbol{\Phi}=(\phi_k)_{k=1}^{K}$ 中点菜产生.对于某家餐厅,第一个顾客坐在第一张桌子上,第二个顾客来了可以选择坐在第一张桌子上,或坐在一张新的桌子上.

假设第 j 家餐厅中,第 i 个顾客到来时有两种选择:

(1)以概率为 $\dfrac{n_{jt}}{i-1+\alpha_0}$ 坐在第 t 个餐桌 φ_{jt} 上(n_{jt} 为第 j 家餐厅中第 t 个餐桌上的人数),即选择已有餐桌的概率与该餐桌上的顾客人数成正比.

① Wang C, Blei D M. A split-merge mcmc algorithm for the hierarchical Dirichlet process. computing research repository. http://arxiv.org/abs/1201.1657, 2012, 9, 24

(2) 以概率为$\frac{\alpha_0}{i-1+\alpha_0}$选取一张新的餐桌坐下. 若选新桌子，则需要为新桌子选菜肴，如果选择菜单中已有的菜肴，则选择第 k 个菜肴 ϕ_k 的概率为$\frac{m_{\cdot k}}{\sum m_{\cdot k}+\gamma}$，或者以$\frac{\gamma}{\sum m_{\cdot k}+\gamma}$的概率选择新菜肴 ϕ_{k+1}，其中 $m_{\cdot k}$ 为所有餐厅中选用了菜肴 ϕ_k 的桌子总数.

CRF 由两层构成，其中高层按照餐桌供应的菜肴不同将餐桌分类，底层按照顾客所在的餐桌不同将一个餐厅里的顾客分类.

主题挖掘中的一篇文档对应其中的一家连锁餐厅，第 j 个文档中第 i 个单词 x_{ji} 为观测值，x_{ji} 的分布参数 θ_{ji} 为顾客，第 k 个主题 ϕ_k 对应菜肴. 顾客 θ_{ji} 所在的餐桌由餐桌索引 t_{ji} 指示，餐桌 t 供应的菜肴的索引为 k_{jt}. 为了方便记忆，定义 $z_{ji}=k_{jt_{ji}}$，z_{ji} 是单词 x_{ji} 所属的主题索引，如式(3).

$$\theta_{ji}=\phi_{z_{ji}},x_{ji}\sim Mult(\theta_{ji}) \tag{3}$$

4　流式 HDP 模型

对于有层次关系的主题，本文提出流式 HDP 模型(FHDP)，利用改进后的 FCRF 构造方法对 FHDP 构造，利用带有流动操作的 MCMC 算法采样.

4.1　FHDP 模型

挖掘得到的第 k 个主题的主题-单词分布为 ϕ_k，每个主题由概率最大的前 T 个单词表示，得到 $K\times T$ 的主题表，其中各个主题出现不同程度的交叠，例如，均包含“比赛”、“记者”.

为了更好地利用隐含同属领域信息的词如“比赛”，而不是具有语料库特点的词汇如“记者”等词，本文假设一个文档可由全部主题构成，并且属于不同主题的部分可以相互流动，为此在 CRF 中加入“房间”的概念. 如图 3 所示，其中长方形框表示餐厅，圆角矩形代表房间，圆形为餐桌，菱形为顾客，五边形为菜肴.

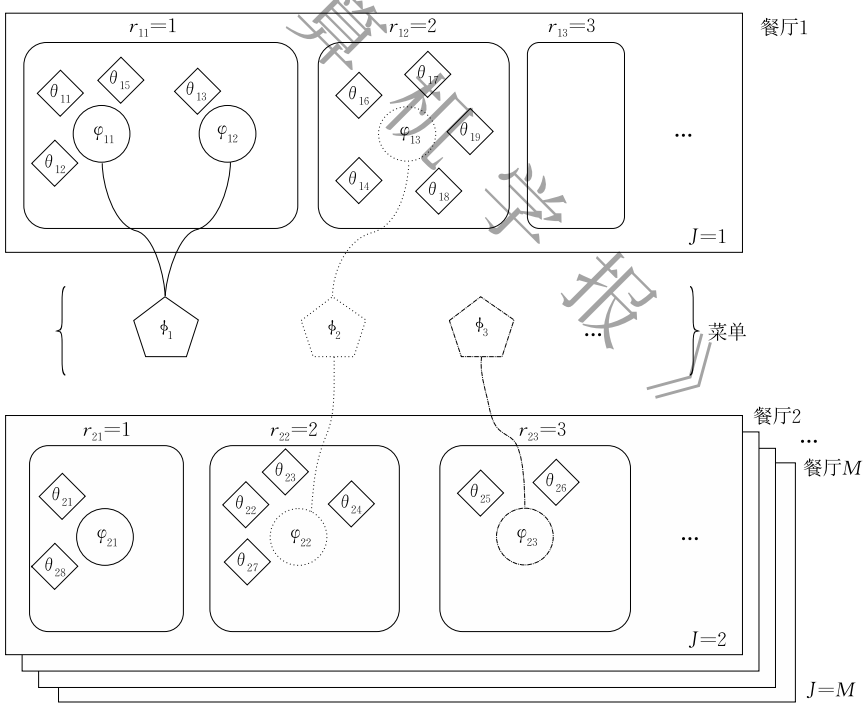

图 3　基于 FCRF 的 FHDP 模型框架图(图中，针对第 1 篇文档的第 6 个词 x_{16}，$x_{16}\sim Mult(\theta_{16})$，$\theta_{16}$ 位于第 3 个餐桌，属于第 2 个主题. 因此相应餐桌索引 $t_{16}=3$，餐桌的主题索引 $k_{13}=2$，θ_{16} 的主题索引 $z_{16}=2$，位于房间 r_{12}. 若主题 ϕ_2 向 ϕ_1 流动，如果 x_{16} 与房间 r_{11} 中的某一顾客指向同一词，则将 x_{16} 移入房间 r_{11}，并重新采样餐桌)

假设一个文档内可以拥有所有主题的内容，因此 FCRF 的每个餐厅 j 内均有 K 个房间，K 为主题 $\boldsymbol{\Phi}$ 的总个数. 餐厅 j 内第 k 个主题对应的房间为 r_{jk}，该房间内包含了 $k_{jt}=k$ 的所有餐桌 t，房间可为空. FHDP 将针对不同“房间”中共现的词施加流动操作(具体流动方法将在 4.2.2 节中详细介绍)，能加强领域信息的集中度.

4.2　流式 MCMC 采样

我们依据 FCRF 的采样需求提出了基于 MCMC 算法的流式 MCMC.

4.2.1 基于 HDP 的 MCMC 采样算法

给定一个集合的文档,后验推断的目的是计算潜在变量文本-主题分布和主题-单词分布的后验分布.在实际运用中通常采用 MCMC 统计模拟的方法估计潜在变量的后验分布.其中 Gibbs 抽样方法是 MCMC 方法中应用最广泛的一种.

本文并不直接对潜在变量进行采样,而是利用 Gibbs 抽样方法迭代采样餐桌索引和主题索引,这样也便于重建模型得到潜在变量.

(1)对餐桌下标的 Gibbs 采样

HDP 的底层是对文档 j 中的顾客按照所在餐桌不同进行了划分,在 Gibbs 采样中对应着 Gibbs 采样器迭代采样文档 j 中每一位顾客的餐桌索引 t.

以概率 $p(t_{ji}=t|\boldsymbol{t}^{-ji},\boldsymbol{k})$ 为顾客 θ_{ji} 采样新的餐桌标号如式(4):

$$p(t_{ji}=t|\boldsymbol{t}^{-ji},\boldsymbol{k})\propto\begin{cases}n_{ji\cdot}^{-ji}f_k^{-x_{ji}}(x_{ji}), & \text{如果 } t\in[1,m_{\cdot}]\\ \alpha_0 p(x_{ji}|\boldsymbol{t}^{-ji},t_{ji}=t',\boldsymbol{k}), & \text{如果 } t=t'\end{cases}\tag{4}$$

$$f_k^{-x_{ji}}(x_{ji})=p(x_{ji}|\boldsymbol{x}^{-ji},\boldsymbol{t},\boldsymbol{k})=\frac{p(\boldsymbol{x}|\boldsymbol{t},\boldsymbol{k})}{p(\boldsymbol{x}^{-ji}|\boldsymbol{t},\boldsymbol{k})}=\frac{\int f(x_{ji}|\varphi_k)\prod_{j'i'\neq ji,z_{j'i'}=k}f(x_{j'i'}|\varphi_k)h(\varphi_k|\eta)d(\varphi_k)}{\int\prod_{j'i'\neq ji,z_{j'i'}=k}f(x_{j'i'}|\varphi_k)h(\varphi_k|\eta)d(\varphi_k)}\tag{5}$$

$t\in[1,m_{\cdot}]$ 为采样到已有的餐桌,其中 $n_{ji\cdot}^{-ji}$ 为文档 j 餐桌 t 中除去单词 x_{ji} 的单词个数,给定主题 ϕ_k 中除去 x_{ji} 以外的所有单词,x_{ji} 的条件概率 $f_k^{-x_{ji}}(x_{ji})$ 由式(5)计算.

$t=t'$ 为采样到新餐桌,作为新餐桌的第一位顾客,以概率 $p(k_{jt'}=k|\boldsymbol{t},\boldsymbol{k}^{-jt'})$ 为餐桌点菜如式(6).其中 $k\in[1,K]$ 为采样到菜单已有主题,$k=k'$ 为采样到一个新主题,$f_{k'}^{-x_{ji}}(x_{ji})$ 由式(7)计算.

$$p(k_{jt'}=k|\boldsymbol{t},\boldsymbol{k}^{-jt'})\propto\begin{cases}m_{\cdot k}f_k^{-x_{ji}}(x_{ji}), & \text{如果 } k\in[1,K]\\ \gamma f_{k'}^{-x_{ji}}(x_{ji}), & \text{如果 } k=k'\end{cases}\tag{6}$$

$$f_{k'}^{-x_{ji}}(x_{ji})=\int f(x_{ji}|\varphi_{k'})h(\varphi_{k'}|\eta)d\varphi_{k'}\tag{7}$$

式(8)中综合了 $t=t'$ 的全过程,可得到式(4)中的 $p(x_{ji}|\boldsymbol{t}^{-ji},t_{ji}=t',\boldsymbol{k})$.

$$p(x_{ji}|\boldsymbol{t}^{-ji},t_{ji}=t',\boldsymbol{k})=\sum_{k=1}^{K}\frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}f_k^{-x_{ji}}(x_{ji})+\frac{\gamma}{m_{\cdot\cdot}+\gamma}f_{k'}^{-x_{ji}}(x_{ji})\tag{8}$$

(2)对主题下标的 Gibbs 采样

高层的 CRP 是对餐桌按照主题不同进行了划分,在 Gibbs 采样中对应着 Gibbs 采样器迭代采样文档 j 中餐桌 t 的主题索引 k.同样由于高层 CRP 也具有可交换性,在高层 CRP 中可以视餐桌 k_{jt} 为最后一个餐桌,因此通过组合除了 x_{jt} 外点菜肴 k_{jt} 的人数和 x_{jt} 的条件概率得到 k_{jt} 的采样式(9).

$$p(k_{jt}=k|\boldsymbol{t},\boldsymbol{k}^{-jt},\boldsymbol{x})\propto\begin{cases}m_{\cdot k}^{-jt}f_k^{-x_{jt}}(\boldsymbol{x}_{jt}), & \text{如果 } k=1,\cdots,K\\ \gamma f_{k'}^{-x_{jt}}(\boldsymbol{x}_{jt}), & \text{如果 } k=k'\end{cases}\tag{9}$$

对于文档 j 中的餐桌 t,按照上式计算 t 属于每个主题 k 的概率,得到主题的条件概率 $\boldsymbol{p}$.最后从 $Mult(\boldsymbol{p})$ 中为餐桌 t 抽样新的主题 k.$f_k^{-x_{jt}}(\boldsymbol{x}_{jt})$ 为文档 j 餐桌 t 下的单词集合 $\boldsymbol{x}_{jt}$ 的先验密度函数,可按照式(11)计算.$f_k^{-x_{ji}}(\boldsymbol{x}_{ji})$ 与 $f_k^{-x_{jt}}(\boldsymbol{x}_{jt})$ 类似,由于基分布 H 和主题的单词分布 F 是狄里克雷-多项分布共轭分布,因此二者可被简化为式(10)、式(11)用于计算.

$$f_k^{-x_{ji}}(\boldsymbol{x}_{ji}=v)=\frac{n_{\cdot\cdot k}^{-ji,v}+\eta}{n_{\cdot\cdot k}^{-ji}+V\eta}\tag{10}$$

$$f_k^{-x_{jt}}(\boldsymbol{x}_{jt})=\frac{\Gamma(n_{\cdot\cdot k}^{-jt}+V\eta)}{\Gamma(n_{\cdot\cdot k}^{-jt}+n^{jt}+V\eta)}\frac{\prod_v\Gamma(n_{\cdot\cdot k}^{-jt,v}+n^{jt,v}+\eta)}{\prod_v\Gamma(n_{\cdot\cdot k}^{-jt,v}+\eta)}\tag{11}$$

式(10)中 $n_{\cdot\cdot k}^{-ji,v}$ 是主题 ϕ_k 中除 x_{ji},单词为 v 的个数.

4.2.2 基于 FHDP 的流式 MCMC 采样算法

流式 MCMC 采样过程中,MCMC 每迭代固定次数后执行一次流动操作.首先,每个餐厅都存在 K 间房间.流动操作首先按照式(12)和式(13)选出主题 ϕ_{k_1} 和 ϕ_{k_2}.

$$p(k_1=k)\propto n_{\cdot\cdot k}^{-ji},\ k=1,\cdots,K\tag{12}$$

$$p(k_2=k)\propto -n_{\cdot\cdot k}^{-ji},\ k=1,\cdots,K\tag{13}$$

如果 k_1 等于 k_2,则重新采样 k_2.对所有文档 j 中属于 k_2 的所有单词的餐桌索引重新采样:

假设文档 j 中的单词 x_{ji_2} 满足既属于房间 r_{jk_2},又使得房间 $r_{j'k_1}$(其中 $j'\in[1,\cdots,N]$)中存在单词 $x_{j'i_1}=x_{ji_2}$ 这个移动条件时,就将单词 x_{ji_2} 移动到餐厅 j 的房间 r_{jk_1} 中.移动到房间 r_{jk_1} 里时,会遇到两种情况:

第 1 种情况.若文档 j 中房间 r_{jk_1} 中无顾客(即也无桌子),则先在房间 r_{jk_1} 中开辟一个新桌子 t',并令新桌子的主题为 k_1.那么就将 r_{jk_2} 中的单词 x_{ji_2} 移

动到餐厅 j 的房间 r_{jk_1} 中新建的餐桌 t' 上.

第二种情况. 若文档 j 中房间 r_{jk_1} 中也已经有了顾客(即也有餐桌),则按照式(14)将 r_{jk_2} 中的单词 x_{ji_2} 在餐桌集合 $T_{r_{jk_1}}$ 重采样餐桌.

$$p(t_{ji}=t)\propto n_{jt}^{-ji},\ t=1,\cdots,m_j,\ t\in T_{r_{jk_1}} \quad (14)$$

当对文档 j 中满足移动条件的所有顾客执行了移动操作后,文档 j 的移动操作完成,当对所有文档执行此移动操作后,就完成了主题 k_2 向 k_1 的流动. 如果在 FHDP t 采样、k 采样的过程中产生了一个新的主题,则所有餐厅均新开辟一个房间. 这样就完成了一次 MCMC 中新增的流动操作.

流式 MCMC 算法的新增流动操作的具体算法如算法 1 所示.

算法 1.　流式 MCMC 算法.

输入:文本集合 x,超参数 η、α_0 和 γ

输出:潜在主题 $\boldsymbol{\Phi}=(\phi_k)_{k=1}^{K}$

假定现在状态为 Gibbs 迭代采样了 T 次后,执行一次流动操作:

按照式(12)和式(13)选出主题 k_1 和 k_2

WHILE($k_1==k_2$)

FOR $j=1$ TO N:

FOR x_{ji_2} IN $\boldsymbol{x}_j$ 中所有单词:

$flag=$False

IF x_{ji_2} 属于房间 r_{jk_2}:

FOR $j'=1$ TO N:

FOR $x_{j'i_1}$ IN $x_{j'}$:

IF $x_{j'i_1}$ 属于房间 $r_{j'k_1}$ AND $x_{j'i_1}=x_{ji_2}$:

$flag=$True

IF($flag$):

IF 房间 r_{jk_1} 中无顾客:

在房间 r_{jk_1} 中开辟 t_{new},令 $k_{t_{\text{new}}}=k_1$,并为 x_{ji_2} 指定餐桌 $t_{ji}=t_{\text{new}}$

ELSE:

依据式(14)为 x_{ji_2} 在房间 r_{jk_1} 中指定餐桌 t_{ji}

继续执行第 $T+1$ 次 Gibbs 迭代采样

5　实验与结果分析

为了检验 FHDP 模型算法的有效性,我们采用一个模拟数据集和两个真实数据集从不同角度进行实验. 实验将验证 FHDP 和 HDP 在分析具有层次关系主题的数据时得到的主题的明确性、覆盖度以及主题之间的层次性等指标.

5.1　实验准备

我们在 3 个数据集上进行实验:

(1) 模拟数据集(Synthetic dataSet). 基于对搜狗数据文本的统计以及词频符合的幂律分布,同时为了使得模拟数据中主题核心内容间部分交叠,本文设计了特殊的多项式文本生成方法,该方法生成的词频特征既符合真实分布,又符合复杂主题的定义.

(2) 搜狗新闻语料库. 经过搜狗实验室手工整理的各类新闻语料. 语料库中的新闻分为了财经、军事、健康、互联网、教育、旅游、体育、文化、招聘这 9 个类别. 本文将该语料库随机分为 3 个数据集:SouGou1(sg_1)、SouGou2(sg_2)、SouGou3(sg_3).

(3) 食品安全新闻语料库. 北京工商大学食品安全数据挖掘研究室实现的食品安全知识图谱中的食品安全事件新闻语料(http://180.76.178.210:8000/index/),来源是从各大新闻网站爬取的食品安全事件新闻报导,一共主要有 57 个带标注的新闻事件.

本文将原始数据集进行分词、去除停用词等预处理操作后结果如表 1 所示.

表 1　实验数据集描述

数据集名称	主题个数	主题内容	各主题对应文本数	总文档数	无重复词个数	总词个数
Synthetic dataSet($sd_1\sim sd_6$)	6	sd_1,sd_2,sd_3,sd_4,sd_5,sd_6	100	600	12	36×600
SouGou1(sg_1)	3	财经,军事,健康	1991,1988,1990	5969	111807	2021151
SouGou2(sg_2)	3	互联网,教育,旅游	1990,1991,1988	5969	124089	1885970
SouGou3(sg_3)	3	体育,文化,招聘	1970,1990,1989	5949	162911	2183506
食品安全数据	57	食品安全事件	其中有 36 个主题文本数大于 100,其余文本数均小于 100	7579	54126	1854598

算法最终得到 $K\times V$ 维的主题-词矩阵,将主题按各个主题下的词频总和降序排序,对于其中每个主题,只保留主题内词频最大的 10 个词作为"代表词集",最终得到 $K\times 10$ 维的主题表.

涉及的指标主要有:

(1) 主题明确性主要采用主题词分布和轮廓系数评价,其目标是衡量不同模型得到的主题是否明确,不同主题的主题词是否能够明确的表达主题的含义. 轮廓系数由 Rousseeuw 提出,它由内聚度和分离度两部分构成. 内聚度和分离度可以

分析主题效果，轮廓系数是结合了两者的优点的指标.

(2) 主题覆盖度采用训练所得主题和真实标注的主题计算主题覆盖的比率，其主要目标是衡量不同模型能否准确发现数据中真实的主题.

(3) 主题层次性主要通过实验结果中主题之间的关系以及主题种类进行分析，主要目标是衡量不同模型是否能发现潜在层次关系.

实验指标的具体计算公式详见附录.

本文中利用了 LDA 和 HDP 作为对比方法. 其中我们按照 Blei 等人提到的方法[5]，用 GridSearch 选出了困惑度最低时的主题个数进行实验. 在 HDP 以及 FHDP 中，超参数 γ 和 α 服从 gamma 先验分布 $\Gamma(1.,1.)$，基分布 H 的超参数 η 人工选择的最优值为 0.5，迭代次数均为 500.

实验环境为 ubuntu 16.04 LTS 操作系统、16 GB 内存、Intel Xeon(R) CPU E5-1620 v4 3.50GHz×8 处理器.

5.2 模拟数据实验与结果分析

本文假设在模拟数据集的主题中存在 4 种词：

(1) 在整体语料库中词频高. 模拟了幂律分布突起的头部，并不具有主题区分能力，例如图 4 中词 12.

(2) 在两个主题中词频高，在其余主题中词频低. 模拟了幂律分布突起的头部，且属于两个主题核心内容交叠的部分，如图 4 中词 2 在主题 1、2 中词频高.

(3) 仅在该主题下词频高，在其余主题中词频低. 模拟了幂律分布的突起的头部，且仅用于区分该主题，如图 4 中词 3 仅在主题 1 中词频高.

(4) 在整体语料库中词频低. 例如图 4 中的词 8，在各个主题下的词频都不高.

依据这样的原则设计的多项式分布函数，产生的数据分布如图 4 所示.

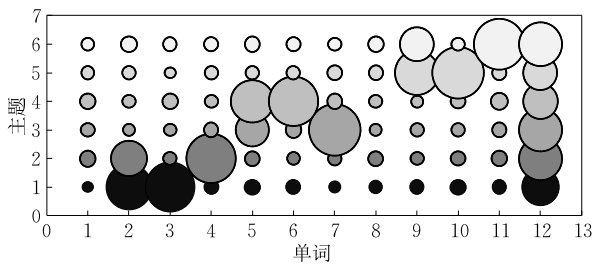

图 4 模拟数据集真实分布(图中圆圈大小代表词在某个主题的文档中出现的频数)

HDP 建模结果有两个问题：(1) HDP 产生了更多的多余主题；(2) 原本分布于各个主题的低频词被聚集在了个别主题中. 如图 5(b) 中的词 6，在各个主题上的分布很不均匀，原本分布在其余主题中的部分集中在了主题 4 中. 这代表 HDP 会将分布在各个主题中的词聚集，导致主题-单词分布不均衡. 如图 5(a) 所示，FHDP 在一定程度上解决了这两个问题，因为 FHDP 将两个主题下共现的词更大概率地流动到父主题中，削弱了无意义词的集中性，更贴近于真实分布.

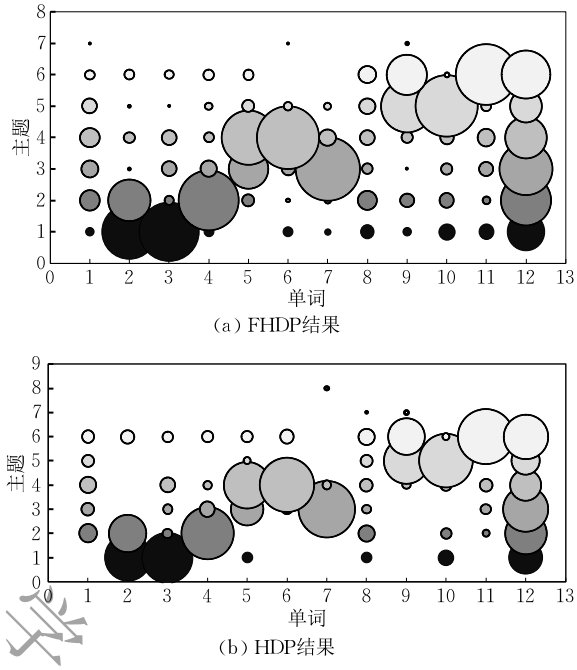

图 5 模拟数据集分别在 FHDP 和 HDP 下得到的主题-单词分布(图中圆大小代表词在某个主题的文档中出现的频数)

5.3 搜狗新闻语料库实验与结果分析

在实际应用中，常利用 HDP 非参数贝叶斯方法挖掘新闻等语料的潜在主题. 而新闻语料具有层次化的特点，即一个大的主题(领域)下包括很多子主题，这些子主题有属于各自的单词分布，也有所属主题的共同的单词分布. 对于每组数据集，我们对得到的 $K\times 10$ 维的主题表进行分析，发现每组数据集前三个主题所占比例均超过了 70%(具体计算方法详见附录)，出现这个现象的原因可能有两个：(1) 每组数据集都包括了三个领域的新闻；(2) HDP 方法具有一定聚集性，容易聚集产生大主题. 因此本文在搜狗新闻语料库中选取前三个主题作为父主题，其余主题作为子主题. 为了验证 FHDP 方法对 HDP 方法的改进，本文从不同角度分析实验结果.

5.3.1 主题明确性

用轮廓系数衡量主题明确性时，主题 ϕ_k 的代表

词集为一簇，单词作为数据点可用向量表示，向量值为单词在真实主题下的 tf-idf 值. 向量代表了该单词在不同主题之间出现的模式，此处采用两个单词向量的余弦相似度计算单词之间的距离. 最终用主题词表中的 Top10 作为代表词计算该主题的平均轮廓系数，最终所得结果如图 6 所示. 其中在语料库子集 sg_1 和 sg_3 下 FHDP 的平均轮廓系数都要大于 HDP，尤其是 sg_3 下 FHDP 的结果明显优于 HDP，但 sg_2 效果提升不明显.

FHDP 通过流动操作提高了比赛、球队、直播、联赛等具有分类能力的主题词的集中度，进而减少了主题词中辨识度低的单词个数，如表 2 所示. 分析 sg_2 提升效果不明显的原因，由表 3 可以看出，sg_2 下 HDP 的第二个主题并不具有实际意义，但是这些单词在三类主题中的分布向量更相似，即计算出的内聚度更高，导致该主题的代表词集内词语轮廓系数的平均值高于 sg_3 中有意义主题. 然而这些单词却并不能覆盖 sg_2 中原主题{互联网，教育，旅游}中的任意一个. 因此需要结合其他指标(如主题覆盖度)来进一步衡量.

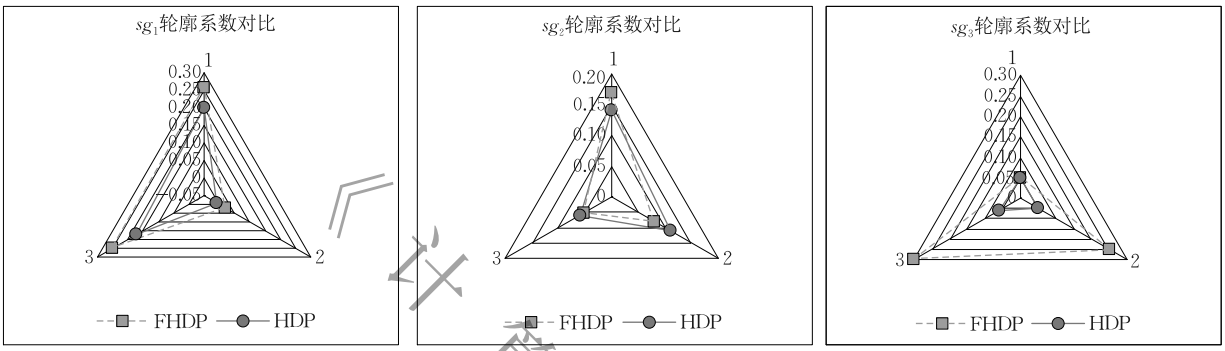

图 6　三个搜狗语料库子集下父主题的轮廓系数

表 2　sg_3 下 FHDP 和 HDP 所得父主题按照单词轮廓系数排序所得的代表词集

FHDP						HDP					
主题 1		主题 2		主题 3		主题 1		主题 2		主题 3	
文化	0.27	直播员	0.37	职场	0.31	朋友	0.10	记者	0.12	历史	0.13
说	0.24	球队	0.37	面试	0.31	做	0.10	前	0.11	文化	0.10
中国	0.23	球员	0.37	招聘	0.31	事情	0.08	机会	0.09	传统	0.10
年	0.22	联赛	0.36	员工	0.31	事	0.08	月	0.08	中国	0.09
时	0.15	比赛	0.36	毕业生	0.30	想	0.07	时间	0.07	精神	0.08
中	0.11	队员	0.36	网页	0.30	走	0.04	中	0.05	世界	0.08
月	0.04	搜狐	0.36	企业	0.29	生活	0.03	时	−0.02	社会	0.02
做	−0.08	日	0.13	就业	0.29	里	0.03	新	−0.06	一种	0.01
工作	−0.29	月	−0.08	老板	0.28	女人	0.01	说	−0.06	中	−0.07
公司	−0.31	中	−0.15	大学生	0.27	说	−0.01	年	−0.09	发展	−0.14

表 3　sg_2 下第二主题代表词的轮廓系数

sg_2 第二个主题的代表词	轮廓系数
里	0.190347563
走	0.186261550
东西	0.120415571
一种	0.111718547
生活	0.103643417
中	0.102863597
说	0.100508490
感觉	0.094623284
喜欢	0.081898953
一位	0.079690415

5.3.2　主题覆盖度

由于轮廓系数还不能够完全描述得到父主题的质量. 我们还需要分析得到的父主题是否能否覆盖真实的主题. 主题覆盖度为模型得到的父主题中真实主题所占比例. 图 7 给出了 sg_1 ～ sg_3 三个语料库子集下 HDP 和 FHDP 的主题覆盖度. 从中可以看出，FHDP 所得父主题对于原有主题的覆盖度要比 HDP 高. 这是因为基于 HDP 方法得到的父主题代表词中低辨识度词语占比较大，导致得到的真实主

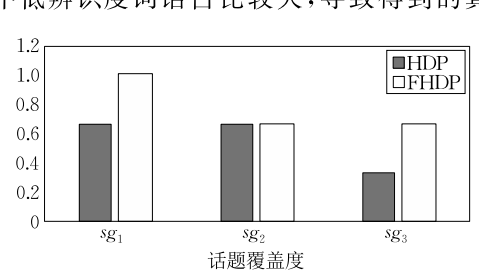

图 7　搜狗语料库子集的主题覆盖度

题无法被覆盖.

5.3.3 主题层次性

主题层次性主要体现在子主题指代的事件是否具体明确,这个可以直接观察到.如表 4、表 5 中分别选取了 sg_3 下 LDA,HDP 和 FHDP 所得主题表的前十个主题.

表 4 sg_3 下 LDA 前 10 主题

LDA 结果
{工作,公司,职业,企业,做,员工,老板,中,能力,发展}
{说,里,做,想,生活,时,男人,女人,喜欢,朋友}
{中国,文化,中,一种,社会,历史,说,世界,精神,传统}
{中国,年,市场,人才,发展,经济,行业,中,城市,北京}
{比赛,中,球队,主场,队员,进球,联赛,申花,对手,一场}
{中国,美国,日本,国家,年,月,日,朝鲜,中,世界}
{比赛,中,中国队,中国,决赛,冠军,年,选手,林丹,张宁}
{招聘,面试,毕业生,大学生,学生,简历,就业,企业,求职,专业}
{革命,中,政治,中国,民主,国家,组织,社会,年,国民党}
{教育,年,学生,学校,大学,学习,英语,工作,教授,说}

对于 LDA,sg_3 下我们选用经典方法 Grid-Search 选取总体困惑度(perlexity)最低的主题个数.如表 4 所示,LDA 主题表的前十个主题没有层次性,挖掘出的主题粒度相近,仍然存在噪音主题.这是因为 LDA 不能发掘数据的层次关系,且必须提前给定合适的主题个数作为参数.

FHDP 得到子主题指向的事件集更具体,而且大多数是"文化"专题下的主题.这是因为 sg_3 中主题"文化"涵盖领域广、共现词较少,子主题中有很多属于"文化"这一主题,比如"历史"、"考古"、"影视"、"戏剧"等子主题."彩票赌博"这一子主题则是属于"体育"类别.HDP 结果层次性就不够明显.在表 5 中,我们的方法挖掘出的父子主题含义清晰,子主题均指向一类具体内涵,能够明确的判定出子主题归属,HDP 平均只有 55% 的子主题有对应父主题,FHDP 有 67%,相比提升了 12%.

表 5 sg_3 父主题与子主题的代表词集

sg_3 主题	HDP 结果	FHDP 结果
父主题	{说,里,想,做,生活,女人,走,事情,事,朋友}	{说,中国,中,工作,年,公司,做,时,月,文化}
	{中,时,说,记者,年,时间,月,机会,前,新}	{比赛,搜狐,中,直播员,球队,日,月,球员,队员,联赛}
	{中国,文化,中,社会,一种,发展,世界,精神,传统,历史}	{企业,招聘,员工,老板,面试,网页,毕业生,大学生,就业,职场}
子主题	{企业,招聘,人才,行业,公司,发展,职业,员工,市场,简历}	{皇帝,项羽,公主,儿子,刘邦,刘备,说,皇后,曹操,刘秀}
	{工作,公司,老板,做,第页,面试,毕业生,找到,学生,职业}	{发现,里,地方,庐山,别墅,建筑,吃,考古,独目,长江}
	{比赛,中,中国队,球队,对手,队员,决赛,冠军,俱乐部}	{工程,年代,年,公元前,断代,夏商周,在位,研究,二里头,夏}
	{中,历史,年,皇帝,时,死,说,天下,人物,杀}	{作品,电影,艺术,影片,导演,画,创作,摄影,艺术品,画家}
	{年,经济,城市,中国,北京,国家,就业,发展,社会}	{张学良,赵一荻,阴阳,父亲,费席尔,赵庆华,天津,东北,韩德彩,症状}
	{中国,美国,日本,年,国家,月,朝鲜,日,印度}	{演出,剧团,宋颖仪,陆平,演员,潮剧,观众,京剧,茶馆,戏}
	{文学,作品,年,小说,电影,作家,中,写,中国,出版}	{赌博,第期,彩票,网络,网站,全场,警方,预测,赌场,彩民}

5.3.4 父子主题的重复词个数

HDP 实验所得主题表中父主题和子主题部分代表词相同,使得 HDP 父子主题内容混淆,层次不明确.为了客观度量层次主题混淆的程度,本文计算父子主题的重复单词个数来检查父子主题之间的代表词重叠情况.从图 8 中可以看到,FHDP 方法得到的父子主题重复词个数比 HDP 少.

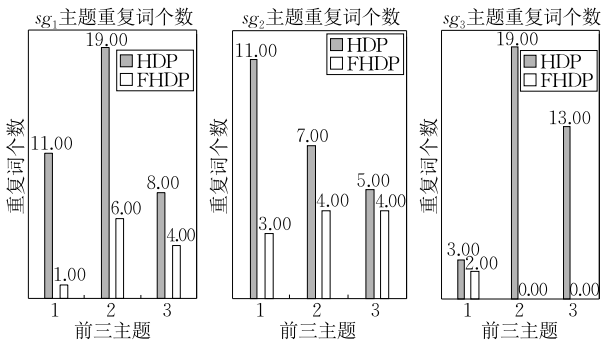

图 8 搜狗语料库子集的主题重复词个数

总体分析,FHDP 在 sg_3、sg_1 上的效果优于 sg_2.由 HDP 处理结果可以看出 sg_2 下"互联网、教育、旅游"的新闻下有效词较为分散,无效词在三个主题下共现次数较多.而 FHDP 本质上是将有效词集中,这样提升主题词质量,所以在主题词较为分散的 sg_2 下提升效果有所下降.

FHDP 方法在主题明确性、主题覆盖度、主题层次性、父子主题重复词个数这四个指标上比 HDP 有所提升,同时,FHDP 比起 LDA 方法,不用设置主题个数,产生了层次性主题.由于 HDP 方法无参考指标,学术界多采用困惑度或单字对数似然(per word log likelihood)来量化误差.如图 9 所示,FHDP 的

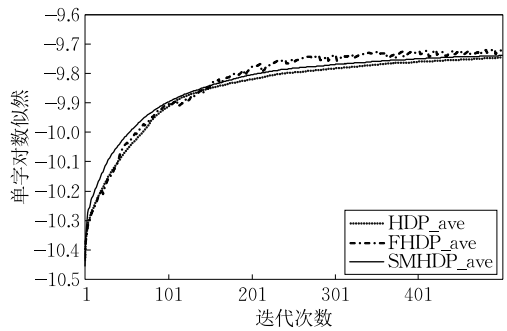

图 9 模型的单字对数似然(per word log likelihood)

per word log likelihood 值提升效果最好.

5.4　食品安全事件挖掘实验结果与分析

食品安全新闻数据集是由北京工商大学食品安全数据挖掘中心整理的一个有标注的数据集，由于相比搜狗新闻语料，标注的粒度更小，精确到具体的事件，因此本文在利用 HDP 和 FHDP 方法挖掘事件以及事件之间的关系时将选择 $K\times 30$ 的词集. 本文用主题描述词中的前 10 个词作为该主题的中心词，分别计算中心词与其余 30 个词的距离，采用基于词共现的归一化谷歌距离[21](Normalized Google Distance，NGD)衡量，选取距离小于 0.5 的词作为描述词，描述词与中心词共同构成一个“事件项”，例如{小龙虾：横纹肌，症，溶解，重金属，患者，症状}. 其中，“小龙虾”为中心词，“横纹肌”到“症状”属于“小龙虾”的描述词，且距离逐渐增大，但不超过 0.5. 最终保留描述词数不小于 2 的事件项. 去除指代内容重复的事件项，即可得到 FHDP 和 HDP 的事件集合，详见附录.

一个主题可以有多个事件项. HDP 方法本质基于词共现，分到同一主题的事件间均隐藏着不同程度的关联.

在得到的结果中 FHDP 的主题更明确，无意义聚类较少. 比如 FHDP 将“水果类”的“芒果催熟”、“柑橘生蛆”和“橙子染色”的事件聚集在了一起，而 HDP 聚出了很多无意义主题，例如将毫无关联的鱼翅事件和红薯粉条事件聚在一个主题下. 这主要是由于 HDP 没有挖掘出领域特性，将无关联的事件混合在了一起. 且 FHDP 得到的主题轮廓也更清晰，主题内平均 NGD 为 0.2441，HDP 为 0.2783，即 FHDP 所得主题的内聚性更强.

同时，FHDP 提升了主题的覆盖度，HDP 覆盖度为 0.4167，FHDP 覆盖度为 0.6667，提升了 60.0%，而且 FHDP 正确挖掘出 27 个事件，HDP 则为 25 个. 且 FHDP 得到的主题层次性更明显，高层广义主题包含事件数和只包含一个事件的主题数都要高于 HDP. 而 HDP 更多的是{造假葡萄酒事件，毒生姜事件}这类无关联的两个事件构成主题.

总之，FHDP 在主题明确性、主题覆盖度、主题层次性上效果明显优于 HDP.

5.5　实验参数分析

作为非参数贝叶斯模型，HDP 不需要事先指定主题个数 K，但需要提前指定超参数 η、α、γ，以及迭代次数. 为了客观分析超参数和迭代次数的影响，本文在模拟数据集上进行了相关实验.

5.5.1　超参数

η：主题 $\boldsymbol{\Phi}=(\phi_k)_{k=1}^{\infty}$ 的先验符合参数为 η 的狄里克雷先验：$\phi_k\sim$Dirichlet(η). 在狄里克雷分布中，参数越小先验越倾向于稀疏的多项分布. 因此 η 与菜肴个数(主题个数)成反比(如图 10 所示). HDP 中通常超参数 η 越小，收敛后得到的主题个数越多.

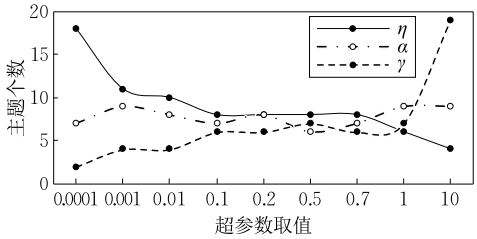

图 10　超参数对主题个数的影响

α：在中国连锁餐馆过程(CRF)中，第 i 个顾客以概率为 $\frac{\alpha_0}{i-1+\alpha_0}$ 选取一张新的餐桌坐下. 因此 α 与餐桌个数成正比(如图 11 所示)，与菜肴个数(主题个数)无关(如图 10 所示).

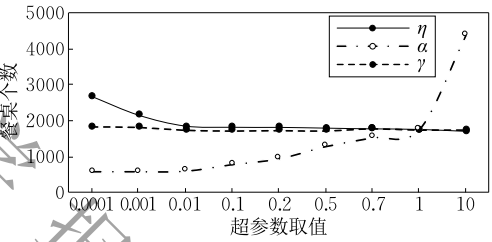

图 11　超参数对餐桌个数的影响

γ：在中国连锁餐馆过程(CRF)中，一个坐在新桌子上的顾客以 $\frac{\gamma}{\sum m_{\cdot k}+\gamma}$ 的概率选择新菜肴 ϕ_{k+1}，其中 $m_{\cdot k}$ 为所有餐厅中选用了菜肴 ϕ_k 的桌子总数. 因此 γ 与菜肴个数(主题个数)成正比(如图 10 所示)，与餐桌个数无关(如图 11 所示).

5.5.2　迭代次数

本文 FHDP 和 HDP 方法随着程序的迭代逐渐收敛，最终稳定在一个固定值. 如图 12 为 sg_2 下 FHD、HDP 分别在迭代次数为 0.1，0.2，0.5 时主

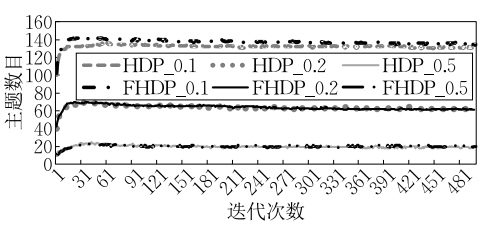

图 12　sg_2 下不同超参数时迭代次数对主题数目的影响

题个数与迭代次数的关系,可以看出,主题个数均在30次迭代前后开始收敛,FHDP与HDP收敛后的主题个数相近.可见对主题个数影响最大的为超参数,迭代次数和FHDP的流动操作影响较小.

至于模型质量,本文用per word log likelihood值衡量,从图9中可以看出,大约在200次迭代后,FHDP和HDP的per word log likelihood值开始收敛,代表模型质量趋于稳定.

6 结论与未来工作

本文针对具有层次结构的文本进行建模,提出具有流动性的双层HDP非参数贝叶斯模型,并设计了相应的流动MCMC算法,成功解决了HDP处理层次结构文本效果不佳的问题.基于模拟和真实数据的实验结果表明,FHDP能够挖掘出复杂主题的层次联系,性能表现优于HDP.

对于大规模文本发现,目前MCMC采样算法效率比较低下,后续我们将侧重于研究面对海量文本时的高效MCMC采样算法.

致 谢 感谢Chong WANG提供的HDP实验程序以及评审专家对本文提出的宝贵修改意见!

参 考 文 献

[1] Rao Y. Contextual sentiment topic model for adaptive social emotion classification. IEEE Intelligent Systems, 2016, 31(1): 41-47

[2] Han J, Lee H. Characterizing the interests of social media users: Refinement of a topic model for incorporating heterogeneous media. Information Sciences, 2016, 358(C): 112-128

[3] Niu Z, Hua G, Wang L. Knowledge based topic model for unsupervised object discovery and localization. IEEE Transactions on Image Processing, 2018, 27(1): 50-63

[4] Hoo W L, Chan C S. Zero-shot object recognition system based on topic model. IEEE Transactions on Human-Machine Systems, 2017, 45(4): 518-525

[5] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. The Journal of Machine Learning Research, 2003, 3(3): 993-1022

[6] Teh Y, Jordan M, Beal M, Blei D. Hierarchical Dirichlet process. Journal of the American Statistical Association, 2006, 101(476): 1566-1581

[7] Xu Ge, Wang Hou-Feng. The development of topic models in natural language processing. Chinese Journal of Computers, 2011, 34(8): 1423-1436(in Chinese)
(徐戈,王厚峰.自然语言处理中主题模型的发展.计算机学报,2011,34(8):1423-1436)

[8] Hofmann T. Probabilistic latent semantic indexing//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information. Berkeley, USA, 1999: 50-57

[9] Blei D M, Jordan M I, Griffiths T L, et al. Hierarchical topic models and the nested Chinese restaurant process//Proceedings of the 16th International Conference on Neural Information Processing Systems. Whistler, Canada, 2003: 17-24

[10] Mimno D, Li W, Mccallum A. Mixtures of hierarchical topics with pachinko allocation//Proceedings of the 24th International Conference on Machine Learning. Corvalis, USA, 2007: 633-640

[11] Joshi A, Bhattacharyya P, Carman M. Political issue extraction model: A novel hierarchical topic model that uses tweets by political and non-political authors//Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. San Diego, USA, 2016: 82-90

[12] Shin S J, Moon I C. Guided HTM: Hierarchical topic model with Dirichlet forest priors. IEEE Educational Activities Department, 2017, 29(2): 330-343

[13] Li L J, L Fei-Fei. OPTIMOL: Automatic online picture collection via incremental model learning. International Journal of Computer Visio, 2010, 88(2): 147-168

[14] Hoffman M, Blei D, Cook P. Content-based musical similarity computation using the hierarchical Dirichlet process//Proceedings of the 9th International Society for Music Information Retrieval Conference. Utrecht, Netherlands, 2008: 349-354

[15] Emonet R, Varadarajan J, Odobez J M. Extracting and locating temporal motifs in video scenes using a hierarchical non-parametric Bayesian model//Proceedings of the IEEE Computer Vision and Pattern Recognition. Colorado, USA, 2011: 3233-3240

[16] Umar H, Arandjelović O. Learning nuanced cross-disciplinary citation metric normalization using the hierarchical Dirichlet process on big scholarly data//Proceedings of the Symposium on Applied Computing. Marrakech, Morocco, 2017: 1842-1847

[17] Teh Y, Kurihara K, Welling M. Collapsed variational inference for HDP//Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2007: 1481-1488

[18] Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6(6): 721-741

[19] Srijith P K, Hepple M, Bontcheva K, et al. Sub-story detection in twitter with hierarchical Dirichlet processes. Information Processing and Management, 2017, 53(4): 989-1003

[20] Andrzejewski D, Zhu X, Craven M. Incorporating domain knowledge into topic modeling via Dirichlet forest priors// Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 25-32

[21] Cilibrasi R L, Vitanyi P M B. The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383

附　录.

父子主题划分阈值 70%：

统计语料中每个词的主题编号得到 topic-word 频率矩阵 $\boldsymbol{\Psi}$，当主题表维度为 $K\times 10$ 维时，主题 k 的大小定义为 $\pi_k=\sum_{i=1}^{10}\Psi_{ki}$，并按照主题由大至小重排 $\boldsymbol{\Psi}$ 后得 $\boldsymbol{\Psi}'$. 实验发现 $\sum_{k=1}^{3}\sum_{i=1}^{10}\Psi'_{ki}\Big/\sum_{k=1}^{10}\sum_{i=1}^{10}\Psi'_{ki}$ 均超过了 70%.

内聚度：

内聚度和分离度可以分析主题效果，轮廓系数是结合了两者的优点的指标.

设单词 $x\in C_i(1\leq i\leq k)$，那么 x 的内聚度 $a(x)$ 和分离度 $b(x)$ 计算公式如下：

$$a(x)=\frac{\sum_{x'\in C_i,x\neq x'}dist(x,x')}{|C_i|-1},$$

$$b(x)=\min_{C_j:1\leq j\leq k,j\neq i}\left\{\frac{\sum_{x'\in C_j}dist(x,x')}{|C_j|}\right\}.$$

x 的轮廓系数的定义结合了 $a(x)$ 和 $b(x)$：

$$s(x)=\frac{b(x)-a(x)}{\max\{a(x),b(x)\}}.$$

$s(x)$ 的值在[−1,1]之间，距离 +1 越近代表该主题下的代表词 x 与该主题下的其他词语内聚性越好、与其他主题的词语分离程度越高.

主题覆盖度：

主题覆盖度采用训练所得主题和真实标注的主题计算主题覆盖的比率，其主要目标是衡量不同模型能否准确发现数据中真实的主题. 设集合 T_k 和集合 R_k 分别为训练和标注的“主题表”第 k 个主题的代表词集合. 则主题覆盖度由 $\frac{|R_k\cap T_k|}{|R_k|}$ 计算而得，其中，$|R_k|$ 为 R_k 集合大小，此处 $|R_k|=10$.

父子主题的重复词个数：

父子主题的重复词个数计算中，子主题个数 K_{sub} 等于 HDP 和 FHDP 两种方法的剩余主题数 K_{sub1}、K_{sub2} 中最小值. 选取主题表中第 4 个到 $K_{\text{sub}}+3$ 个 主题作为子主题. 则该语料库下父主题与子主题的重复词个数计算式为

$$rep(Z^x)=\sum_{i=1}^{3}\sum_{j=4}^{K_{\text{sub}}}Z_i^x\cap Z_j^x.$$

式中 Z^x 代表数据集 x 所得的主题表，$rep(Z^x)$ 代表该数据集下“父子主题的重复词个数”，Z_i^x 代表主题表中第 i 个主题的代表词集合.

Normalized Google Distance(NGD)：

NGD 是一种语义相似性度量方法，在自然语言意义上有相同或类似含义的关键词往往在 NGD 单元倾向于“紧密”，而有不同含义的词汇则往往距离较远.

NGD 的计算公式为

$$NGD(x,y)=\frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log M-\min\{\log f(x),\log f(y)\}}.$$

M 是数据集中的总文档数，$f(x)$ 和 $f(y)$ 分别是出现了词 x 和 y 的文档数，$f(x,y)$ 是同时出现 x 和 y 的文档数. 如果两个词从未一起出现在同一文档中，则它们之间的 NGD 是无穷. 如果两个词总是同时出现，则它们的 NGD 是 0.

HAN Zhong-Ming, Ph.D., professor. His current research interests include web data mining, big data and information retrieval.

ZHANG Meng-Mei, M.S. candidate. Her current research interests focus on web data mining.

LI Meng-Qi, M.S. candidate. Her current research interests focus on web data mining and social network.

DUAN Da-Gao, Ph.D., associate professor. His current research interests include social computing and multimedia information processing.

CHEN Yi, Ph.D., professor. Her current research interests focus on big data visual analytics and mining.

Background

With the rapid development of social media, the text on the internet has become a main media of message communication. The topic model is a popular tool for automatically discovering and exploring hidden topics in the document. Hierarchical Dirichlet Process (HDP for short) is a widely used method in research filed of topic model, which does not require a given number of topics in advance. But HDP still faces many challenges in analyzing large-scale internet texts because some features of them, such as multiple sources of internet texts, unclear topic boundary, the nesting of topics, and so on.

The features of FHDP are as follows: (1) Solving the problem that the topic is unclear when using the HDP to model corpus with complex topic structures, and enlarging the application domains of HDP; (2) The meaningless word in the topic of FHDP is reduced, so the hierarchy of the topic becomes more explicit, and the hierarchical relationship among complex hierarchical topics can be analyzed using FHDP.

This work is supported by the National Natural Science Foundation of China under Grant No. 61170112, the Beijing Natural Science Foundation (4172016) and the Beijing Science and Technology Project (Z161100001616004).