应用hLDA进行多文档主题建模关键因素研究 *

衡伟[1]，于佳[1]，李蕾[1]

（1.智能科学技术中心 北京邮电大学 计算机学院，北京 100876）

摘要：hLDA（层次潜在狄利克雷分配）在层次主题建模中的良好效果已经得到广泛验证。为了实现半监督或无监督，通常采用交叉验证或抽样超参来确定参数。但由于语料特征、建模需求等不确定因素，参数调节方法、建模效果和效率都是实际应用中的难点。本文首先结合贝叶斯线索和范围线索构成的统一分析框架，研究 hLDA 主题建模中的关键影响因素，然后给出一个切实有效的建模策略及流程，最终结合 ACL MultiLing 2013 多文档摘要语料进行实际建模效果评估。

关键词：层次潜在狄利克雷分配，层次主题建模；统一分析框架

中图分类号：TP391　　**文献标识码：**A

Research on Key Factors in Multi-document Topic Modeling Application with HLDA

Wei Heng[1], Jia Yu[1] ,Lei Li[1]

(1.Center for Intelligence Science and Technology, BUPT Beijing 10086 ,China)

Abstract: The results of hLDA (hierarchical Latent Dirichlet Allocation) in the hierarchical topic modeling have been widely validated. In order to achieve semi-supervised or unsupervised learning, cross-validation or sampling super parameters are usually used to determine the true parameters. However, corpus features, modeling demand and some other factors are uncertain. Hence, parameter adjustment, modeling effectiveness and efficiency are difficulty to achieve in practical applications. This paper builds a unified analytical framework by combining Bayesian theory and boundary information, analyzes the key factors in its topic modeling, then gives a series of practical and effective modeling strategies and processes, and finally evaluates the modeling results with multi-document summary corpus from ACL MultiLing 2013.

Key words: Hierarchical LDA; Hierarchical Topic Modeling; Unified Analytical Framework

1 引言

为了从数据中学习层次信息，Blei 等人[1]提出了基于 nCRP 的层次潜在狄利克雷分配（以下简称 hLDA）非参模型。模拟数据和 JACM 语料的实验评估证明了其具有非常好的效果[2]。Asli 和 Dilek 等人[3]利用采用交叉验证的方法对 hLDA 进行英文多文档摘要建模，效果显著。刘等[4]应用 hLDA 进行中文多文档聚类和摘要的研究，亦取得了非常好的效果。

但 hLDA 的建模效果却因语料和应用建模者的不同而差异巨大。虽然在 Blei 的论文中提出了无监督的、针对先验超参的 MH 抽样方法[5]，在迭代尽可能多的条件下，理论上可以实现完备的后验推理[6]。但是，实际应用中资源有限，我们无法保证进行足够多次的迭代，且不同的语料特征以及建模需求使得迭代次数具有很大的不确定性，加之吉布斯后验推理算法是一种随机算法，每次迭代稳定的状态都不同。再者，应用中对最优树结构的评估方法也有较大的不确定性。一定程度上需要通过多次局部最优的结果来逼近全局最优。Blei 仅仅给出了模型参数本身[2]，却没有详细的分析参数选择过程。同样 Asli 和 Dilek 的交叉

* **收稿日期：**2013.6.1　　**定稿日期：**2013.7.15

基金项目：国家自然科学基金资助项目（71231002，61202247）；北京邮电大学青年科研创新计划专项；北京市科学技术情报研究所项目“科技情报辅助系统”；中央高校基本科研业务费专项资金（2013RC0304）。

作者简介：衡伟（1989——），男，硕士，主要研究方向为统计机器学习；于佳（1986——），女，硕士，主要研究方向为文档摘要；李蕾（1974——），女，副教授，主要研究方向为自然语言处理，机器学习。

检验寻找模型参数方法则过多的限制了模型的泛化效果。本文通过统一分析框架，采用理论分析与实验相结合的方式，对应用 hLDA 到多文档主题建模任务中的关键影响因素进行深入研究，试图寻找优化的建模策略、建模流程以及有效的参数配置方法，能在有限资源、有限迭代、语料多变的情况下，尽可能让建模结果更好地逼近全局最优，从而为 hLDA 实现高效的层次主题建模提供有益的参考。

文章结构安排如下：第二节介绍统一分析框架，从贝叶斯线索和范围线索两个角度分析多种影响因子，同时简要介绍了实验模块；第三节针对 hLDA 模型文档生成过程的超参进行分析和实验；第四节给出了吉布斯抽样算法后验推理的关键影响因素；第五节针对影响建模的全局因子进行了分析；第六节给出一个经验型优化建模流程，并结合最新的 ACL MultiLing 2013 多语言多文档摘要数据进行建模效果实验与效果评估。

2 统一分析框架及实验模块

2.1 统一分析框架

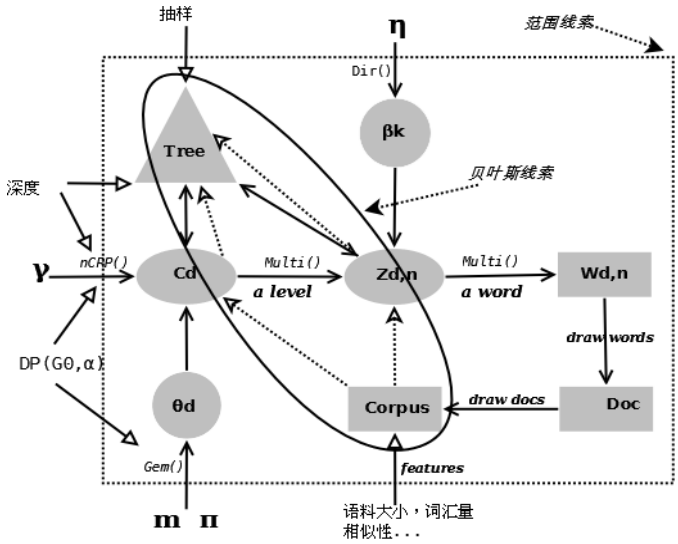

图 1 统一分析框架

如图 1 所示，黑色矩形虚线表示范围线索，在虚线框之外是一些全局的建模影响因子，如抽样与否、树的深度、语料的大小，词汇量等特征。虚线内部是贝叶斯线索的影响因子，如狄利克雷(*Dirichlet*[1]())分布超参η、 *GEM*分布超参m, π、 *nCRP*过程超参γ[2]。黑色椭圆形实线是贝叶斯过程：黑色有向实线箭头所示为基于先验的文档生成过程。逆向的黑色虚线箭头即贝叶斯后验推理：从语料出发[3]，选择路径C_d和词$W_{d,n}$，获得最优树结构。

从图 1 的统一分析框架，我们能够窥见 hLDA 模型算法的全貌，从而有利于深入分析建模影响因子。后面三节便是以这两个线索为纲展开的，并结合具体的实验分析，为此，我们对用于建模分析的实验系统做一个较为全面的介绍。

2.2 实验系统介绍

2.2.1 实验系统框架

实验系统由三个模块组成：预处理、hLDA 建模、结果分析评估。如图 2 所示。预处理

[1] 为了与 Blei 论文中表示一致，后面我们将使用 Dir()表示。

[2]此处表示同 Blei hlda 算法包[http://www.cs.princeton.edu/~blei/topicmodeling.html]中参数命名。

[3]忽略了预处理到生成文档到词的过程，图中的 Doc, $W_{d,n}$ 也应该有相应的黑色虚线表示。

模块对语料进行分句、分词、生成词表、统计词频特征等。生成 hLDA 建模输入文件，同时为 hLDA 超参选择提供分析依据。hLDA 建模是核心模块，其设置文件中的超参选择是建模的重点，主要依据语料本身的特征分析以及对于建模结果分析的反馈。结果分析评估模块是实

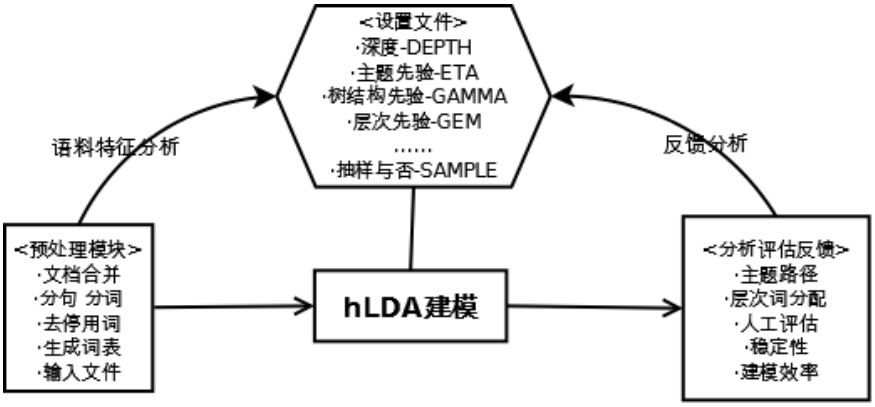

图 2 试验系统框架图

验分析的基础，从而验证各个建模影响因素。

2.2.2 实验语料来源

实验语料来源于两部分：其一收集了国内门户网站新闻报道，共十个话题，每个话题 10 篇相关报道，即 Portal News。便于图表叙述，给出了英文缩写，如甘肃校车事故（SBAG），伊朗制裁（IRSA）等。其二是 ACL MultiLing 2013 多语言多文档摘要评测发布的数据。也由十个话题，每个话题下 10 篇新闻组成：如印度洋海啸（M000），伦敦爆炸案（M001）等。

3 文档生成过程

3.1 嵌套中国餐馆过程 nCRP

nCRP 是 hLDA 的核心，属于贝叶斯非参建模家族，近些年在层次主题建模领域受到了广泛的关注[11][12][13]。*nCRP* 构造了一个树状层次结构先验，超参 γ 决定先验树结构的形状，即每个文档每一层的路径选择。称之为嵌套中国餐馆过程是因为本质上它只是对于每一层都使用中国餐馆模型（即 *CRP*）进行路径选择。

3.1.1 中国餐馆过程 *CRP*

CRP 可被简单表述为如何从以下等式的条件概率函数中选择一个样本所属的类别聚簇：

$$\boldsymbol{P}(C_{N+1} = k \mid \gamma, \boldsymbol{n}) = n_k / (N + \gamma)$$

$$\boldsymbol{P}(C_{N+1} = K + 1 \mid \gamma, \boldsymbol{n}) = \gamma / (N + \gamma) \quad (3\text{-}1)$$

其中，N 表示已有的样本数，C_{N+1} 表示新来的样本，K 表示目前的样本类别数，n_k 表示第 k 个样本类别所含有的样本数目，$\boldsymbol{n}$ 表示所有 n_k 所组成的集合向量。可以看出，某一个类别上的样本越多，则新抽样本属于该类别的概率越大。最终聚簇数的期望如等式 3-2 所示：

$$\mathbf{E}[\ K_n \mid \gamma\ ] = O(\ \gamma \log\ n\ ) \quad (3\text{-}2)$$

在给定 γ 的情况下，占用聚簇数的期望随着样本数 n 呈指数增长，因此，可以通过分析文档数目和期望的聚簇数来反向估计 γ 的范围。这在实际的分析 *nCRP* 超参的过程中亦具有较高的参考意义。

3.1.2 *nCRP* 及 γ 值实验分析

CRP 是一个在整型离散空间上的随机过程，*nCRP* 同样是一个随机过程，但不是在一维的整型空间，而是在树的深度维度上的整型向量空间。因此，当假设每个聚簇上有一个潜在主题变量 β_k 时，某一条聚簇路径上也有一个潜在向量$<\beta_{0,k}\beta_{\cdots,k}\beta_{1,k}>$。*nCRP* 过程指定了文档所属的潜在向量聚簇。对于三层树结构，*nCRP* 过程相当于在一个三维整型空间中去选择聚簇，每个样本则是三维空间中的某一个点。

如表 3 所示，实验分析在同一个话题语料[4]下，γ值所引起的聚簇数（即路径数）的变化。

表 3 GAMMA 值对聚簇数和词层分配影响

GAMMA（γ）	主要路径数	主要/总路径数	各层词分配
0.2	34/11/9/6/5/4/3/3/3/	6/54	2116/437/45
1.0	15/12/8/7/5/5/4/3/3	7/65	2117/447/34
8.0	17/9/8/8/7/5/5/4/4/3	9/72	2134/437/27

当 γ=0.2 时，前段主要路径聚集，而路径数却相应的减少。随着 γ 值从 1.0 变化到 8.0，各聚簇分布逐渐趋向平均，且路径数也在相应增加。各层词分配在随机抽样允许的变动范围内，比例基本不变。从原理上分析，如等式 3-1 所示，γ 值增大使得选择新聚簇的可能性增加，在总文档数不变时，原本过于聚集的簇倾向于分散，产生更多新聚簇。而第三层词的分配是随着 γ 的增大而减少，而最后的路径数却呈现增多的趋势，其原因便在于，γ 增大的过程中，从根节点到叶子节点各个层次聚簇数都相应的增加，由公式 3-2 可知，文档数越大，词数越多的情况，聚簇的增加越快，因此相对而言，根节点增加的要比叶子节点快，为了满足这样的先验假设，后验词分配便逐渐的从叶子往根聚集，从而导致叶子节点词的减少。

3.2 折棒构造

折棒构造是狄利克雷过程（以下简称 *DP* 过程）的另一种构造方式，侧重于以最终分布为中心的构造。每次折棒，都会通过 *Beta* 分布得到最终分布比例的一部分，而 *CRP* 每次抽样只是对最终分布比例的一次更新，随着抽样次数的增加进而愈加接近最终分布。关于折棒过程更为详细的叙述，很多论文中皆可参考[3][4][11][14]。既然 *CRP* 和折棒过程都是 *DP* 过程的不同构造方式，对于 *CRP* 的理论分析同样适用于折棒过程，如最终聚簇的期望等。

3.2.1 参数 *m* 和 π 实验分析

参数 *m* 控制着从根节点到叶子节点的分配比例，而 π 则指定该分配比例的严格程度，但相对 *m*，其影响要小。实验首先从 Portal News 语料中随机选取两篇，分析 *m* 从 0.25 到 0.75 变化时词的分配。如图 4 所示，上下各三个饼状比例图，分别表示一个主题在不同的 *m* 值条件下，树中各层词分配比例。饼状图中的黑色部分表示第三层叶子节点所占词的比例，白色部分表示中间层词所占据的比例，而灰色部分则表示根节点词所占的比例。

从第一行三个饼状图的比例变化可以看出，随着 *m* 值的变大，叶子节点中词的比例明显增加，而根节点中词所占的比例则在减少。第二行虽然不同语料差异使得各层词的比例不完全相同，但是这种趋势也非常显著。可以推测，*m* 越大，文档中的词越向叶子节点聚集，越倾向于较为具体的主题，反之亦然。

接着图 4 的分析，利用 ACL MultiLing 2013 中前五个主题的语料，分析确定 m 值情况下，词分配比例的稳定性，如表 5 所示。从原理上来看，折棒构造过程中 m、π 的先验影响效果比较明显，尤其在树的层数小，主题下文档数较少时，贝叶斯后验解释受先验的影响较大。因此，在应用建模时可以根据期望的主题层次分布和抽象具体词的比例来确定 m 值的

[4] ACL MultiLing 2013 中文语料下的 M004 话题。

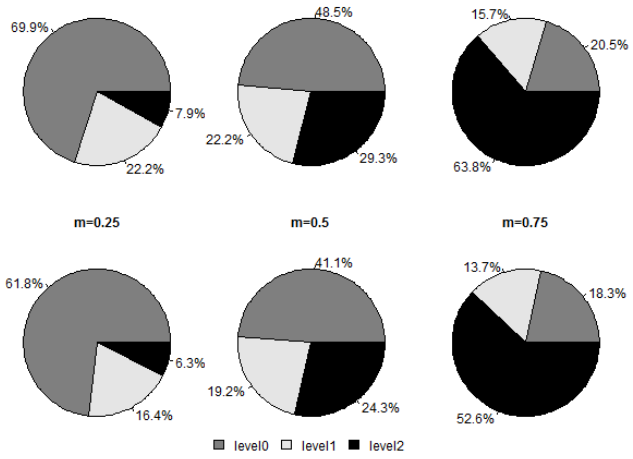

图 4 不同 *m* 下，三层树结构时词的分配比例

范围。以此类推，结合 *CRP* 中的理论分析和实验评估，我们可以给出一个经验化的比例范围，从而有利于我们所期望的更为精确的 *m* 值控制。

表 5 不同语料数据时，确定 m 值下各词层分配的稳定性

m/层次 主题	m=0.25			m=0. 5			0.75		
	Level0	***Level1***	***Level2***	***Level0***	***Level1***	***Level2***	***Level0***	***Level1***	***Level2***
M000	0. 6868	0. 2444	0. 0688	0. 5813	0. 2444	0. 1743	0. 4921	0. 2294	0. 2785
M001	0. 6716	0. 2273	0. 1011	0. 5459	0. 2709	0. 1832	0. 4766	0. 2145	0. 3089
M002	0. 6477	0. 2652	0. 0871	0. 5309	0. 2841	0. 1850	0. 4482	0. 2436	0. 3082
M003	0. 6915	0. 2315	0. 0772	0. 5785	0. 2353	0. 1862	0. 4581	0. 2471	0. 2948
M004	0. 6401	0. 2617	0. 0982	0. 5416	0. 2533	0. 2052	0. 4403	0. 2679	0. 2918

接着前面在 *nCRP* 试验中的分析，当 γ 产生较大变化时，对于路径树和聚簇比例的变化有一定的影响，但是各个层次词的分配比例受到 γ 参数的影响较小。这在一定程度上为我们细化 γ 和 *m* 参数对树结构的调节范围提供了可能。

3.3 狄利克雷分布和 *DP* 过程

狄利克雷分布决定了每个节点上主题先验 β_k。文档生成过程中，首先假设一个无限深度和无限宽度的树结构，树中的每一个节点以超参 η 生成一个主题，以此进行嵌套中国餐馆过程和折棒过程的构造。两种构造却已经不仅仅是由样本数 N 和先验超参 γ 或 *m*、*π* 控制，在单纯的整型空间上的聚簇划分。因为每个节点都有了实际意义，即主题 β_k。于是这两种过程都变成了在潜在主题变量下的混合模型。我们首先分析 *DP* 的形式化定义以及 *DP* 的两种构造过程，以此为切入点来分析狄利克雷分布所确定的主题和这两个构造过程的关系。

3.3.1 从 *DP* 的角度分析文档生成过程

DP 是一种随机概率测量在一个可测量空间上的分布[15]。在生成主题节点值空间的 *Dir()* 分布的基础上，分别把 *nCRP* 和 *GEM* 的构造过程理解成为一种 *DP* 过程。对于 *nCRP* 而言：

$$\beta_{1,\infty} \sim Dir(\eta); \qquad \boldsymbol{L}_{1:D} \sim CRP(\boldsymbol{\gamma}_{0:L-1}); \qquad \boldsymbol{\beta}_{1:D} \sim \sum_{0}^{n-1} Categorical\, \boldsymbol{\beta}_{L_{1:D}} \tag{3-5}$$

$$G_{nCRP} \sim \boldsymbol{DP}_{ncrp}\left(\sum_{0}^{n-1} Dir(\eta),\ \gamma\right); \qquad \boldsymbol{\beta}_{1:D} \sim G_{nCRP} \tag{3-6}$$

等式 3-6 中，$\boldsymbol{L}$ 表示的是嵌套的层数向量所构成的分布矩阵，对于每一层通过 *CRP* 过程，得到一个比例分布，结合层数得到这样一个矩阵结构。接下来对一篇文档从 $\boldsymbol{L}$ 矩阵中的第一行开始一直到结束，逐步选择相应的基分布值，即 **β** 向量，其向量长度等于嵌套的次

数，也即树的深度。*Categorical(K)*分布表示从某一有 *K* 个结果的随机事件中抽样。这便是通过嵌套 *CRP* 方式构造一个 *DP* 过程。而等式 3-6 中，则是直接从 *DP* 定义的概率测量的角度来生成。好处在于能够直接清晰的分析出狄利克雷分布作为基分布在整个 nCRP 过程中的作用。而对于 *GEM* 而言，其基分布则是前面 nCRP 所形成的 **β** 向量的分布，然后对于向量的维度即同样树的深度 L，进行折棒构造，如等式 3-7,3-8 所示：

$$\boldsymbol{\beta}_{\mathrm{d}} \sim G_{nCRP};\ \boldsymbol{l}_{0,n-1} \sim GEM(m,\ \pi);\ \boldsymbol{l}_i \sim Categorical(\beta_{d_{1_i}});\ \mathrm{W}_{\mathrm{d,l}} \sim Categorical(\boldsymbol{l}_i) \quad (3\text{-}7)$$

$$G_{\mathrm{GEM}} \sim \boldsymbol{DP}(GEM(\mathrm{G}_{\mathrm{nCRP}}),\ (1\text{-m})\pi);\quad \boldsymbol{l}_i \sim G_{\mathrm{GEM,d}};\quad \mathrm{W}_{\mathrm{d,l}} \sim Categorical(\boldsymbol{l}_i) \quad (3\text{-}8)$$

从以上的分析我们不难理解，对于基分布狄利克雷而言，*nCRP* 过程类似于一种对每一层取值空间进行了扩展组合，然后在一个高维的更大的空间内进行向量选择，在此基础上，GEM 分布再对已选的向量进行每一维度上的概率选择，从而产生相应的词。

3.3.2 参数 η 实验

如表 6 所示，我们分析在叶子节点上的 η 值（其余两层值分别为 5.2/0.025)变化时，相应的主题路径以及各节点上文档和词的变化：

表 6 η 对树结构的影响

η（叶子）	词分配	总路径	文档/词	--	--	--	--
0.5	1335/901/362	93	6/23	6/17	6/17	4/10	4/13
0.05	1267/749/582	44	48/222	23/110	13/43	12/65	6/19
0.005	1288/746/564	49	30/110	16/75	13/47	8/42	7/28

其中，第一列是叶子节点上的 η 值，第二列是从根节点到叶子节点总的词分配，第三列是总的路径数，剩下的五列表示最主要的五个路径上文档和词分配。当 η 为 0.05 和 0.005 时，叶子节点的词分别为 582 和 564，在一定抽样不确定性允许的情况下，词数是相对稳定的，叶子节点上总路径数变多，各个路径上文档和词的数量则变小。随着叶子节点 η 的迅速变小，其词由上往下流动，前面两层的路径数变小，下层的路径数相对变多，嵌套的效果使得整体路径数变小。因此 η 对于词的分配、路径数有很大影响。且往往在与 GEM_MEAN（m）参数混合作用的情况下，这种影响会导致在实际建模中一些意想不到的问题，最经典的则是 mode.levels 文件的缺失。

4 后验推理

4.1 吉布斯抽样

吉布斯抽样广泛应用于统计推理领域，尤其是贝叶斯后验推理。主要通过构造一个蒙特卡罗马尔科夫链使得其稳定状态分布等于后验分布[2]。实际应用中如何评估马尔科夫链的收敛性往往决定着推理的效果。而迭代次数的设定对链的收敛有着很大的影响。

4.1.1 迭代次数实验

通过 Gibbs 抽样算法，在无限次迭代达到收敛时，可以实现对语料理想的建模。但正如我们在第一节所讨论的，往往受限于实际应用瓶颈。因此，在前面参数调节的基础上，我们首先通过初始参数设定，对目标语料形成了一个较为理想的层次分类，然后分析在增加迭代次数的条件下，各个层次主题的变化。在 m 和 π 参数不变的情况下（m=0.5, π=100），观察迭代次数对词层分配的影响，如表 7 所示：

表 7 迭代次数对于词层次分布的影响

层数	Iter10000	Iter30000	Iter50000	Iter80000

0	1217	1152	1131	1175
1	666	757	733	718
2	715	689	734	705

在此迭代条件下，可以发现迭代次数对于词分布影响很小，马尔科夫链已经达到稳定的局部收敛状态，因此，我们可以从词层的稳定性上来判断链的收敛情况。但是，对于不同特征的语料，不同迭代次数下迭代收敛和路径数却是不同的，如图 8 所示：

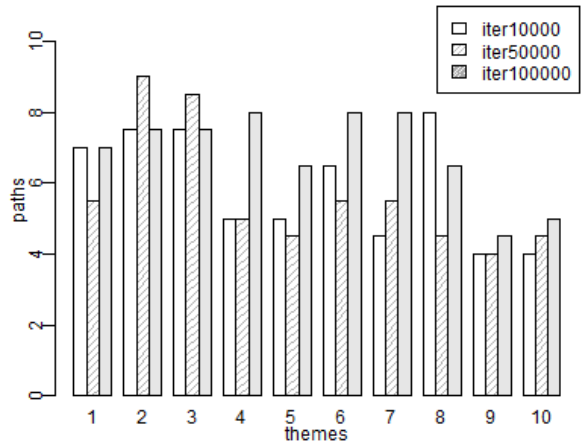

图 8 不同迭代次数下路径数比较

横坐标为话题，纵坐标为最终的路径数。在其他参数一定的情况下，迭代次数越大，得分最高的 mode 所形成的最优路径便会越逼近实际主题中的真实路径。从图中可以看出，在考虑到随机算法的不确定性影响的情况下，这种变化趋势基本上是保持一致的，如主题 2,3,7,9,10。但也有较为不一致的情况，如主题 8，由于其语料特征的差异使得在实验的 10000 到 100000 次的迭代范围之内，路径数不稳定，并没有达到一个较为稳定的状态，因而三组实验时路径数变化很大。

还有一种情况，在一定的迭代范围内，路径树已经趋向于稳定状态，但可能陷入一种局部最优的稳定状态。对此，我们不可能通过无限制的增加迭代次数来最优化，通常通过多次重启抽样，或是改变改变抽样中的随机延迟值 [1](SHUFFLE_LAG)或是抽样延迟值 [1](SAMPLE_LAG)。模型本身的超参很多，加之随着文档数的增加带来链上变量的急速增长，随机条件下通过有限的迭代，很难有较好的效果。因此，相比于这种概率随机条件，启发式的逼近调节效果往往更为显著。

5 全局建模策略

5.1 树的深度

深度假设是 hLDA 一个最基本的假设，也反映了主题建模粒度的期望。我们给出了不同树深度条件下，Portal News 中的 8 个话题平均路径数变化，如表 9 所示：

表 9 不同深度情况下的平均路径树变化

话题\深度	3	4	5
XTUS	4.9	59.7	78.7
IRSA	3.6	35.8	46.5
LTFC	7.1	54.2	74.8
EUDC	4.1	59.2	83.5
GBAB	3.1	47	66.2
ROHN	7.2	49.6	69.7

SBAG	7.7	53.9	75.2
MACO	6.7	44.2	56.3

随着树深度的增长，路径数呈现快速增长的趋势。与其他超参对于路径数的影响比较，其增长趋势是最快的。由此分析，在设定超参时，树的路径树是我们首先需要考虑的参数。结合原理分析，我们知道 *hLDA* 的核心是基于 *nCRP* 的先验树结构，不管是对于文档生成的路径选择还是每个节点的主题分配，首先需要对深度做假设，树越深则 *CRP* 的嵌套效果越明显，*GEM* 分布层次的后验性越强。同时主题层次越多，每次运行的稳定性越差。

5.2 超参抽样与否

超参抽样的目的是为了尽可能的减少手工设定超参对于最终文档树结构的影响[16]，使得实验结果来源于语料本身特征。但是抽样也同样存在一些缺点，首先其限制了我们对于超参更为灵活的、目的性的调整；其次，从算法的效率考虑，抽样情况下的时间复杂度要高出许多。抽样超参的选择主要集中在主题 β_k 的超参 η，词分配超参 m 和 π。我们从语料中随机选择四个主题，进行抽样影响因素的分析，如图 10 所示：

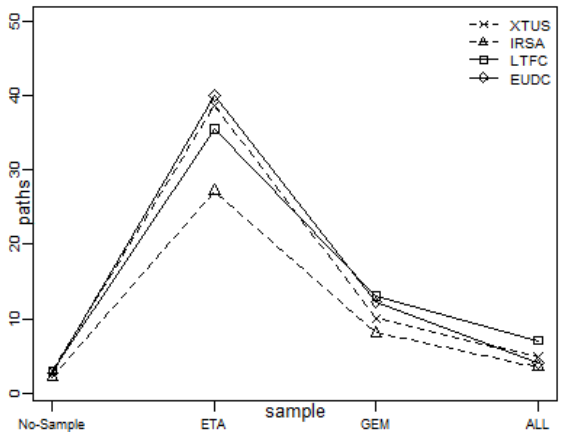

图 10 不同超参的抽样情况下的路径树变化

图中分析了不抽样，抽样 η（*ETA*），抽样 m, π *(GEM)*和两者都抽样时，四个话题路径数的变化。一方面对于不同的抽样选择，四个话题最终的文档路径数变化趋势是一致的，对于抽样 η 和不抽样 η 的情况，路径数变化尤其的大，相比之下 *GEM* 参数的抽样要小点。结合 3.3.1 节的分析，η 是 *nCRP* 过程中的基分布，控制着节点的主题，最直接的反映了文档的后验解释。另一方面，在对其选择抽样时，相比于 *GEM* 参数 m 的 0-1 的取值区间，其可以在整个实数范围内取值，因此随机化的区间更大，在一定的迭代次数下，并不一定能保证逼近最优值，因此对于整个路径数的影响比 *GEM* 的变化更大。

5.3 语料特征

如表 11 所示，仍然从 Portal News 语料中随机选择四个话题，统计每个话题下的句子数、总词数、词表大小，以及相应的人工专家进行主题摘要归纳的主题数：

表 11 语料大小与词表统计

话题\规模	主题数	句子数	总词数	词表
XTUS	4.9	242	4360	1563
IRSA	3.6	96	1807	679
HCDH	3.1	194	2792	825
MACO	6.1	191	2760	1055

对表 11 中词频特征按照文档中词的出现顺序进行了统计，以尽可能保证文档中词分布特点的同时，保留住其出现的上下文特征[5]，如图 12 所示：

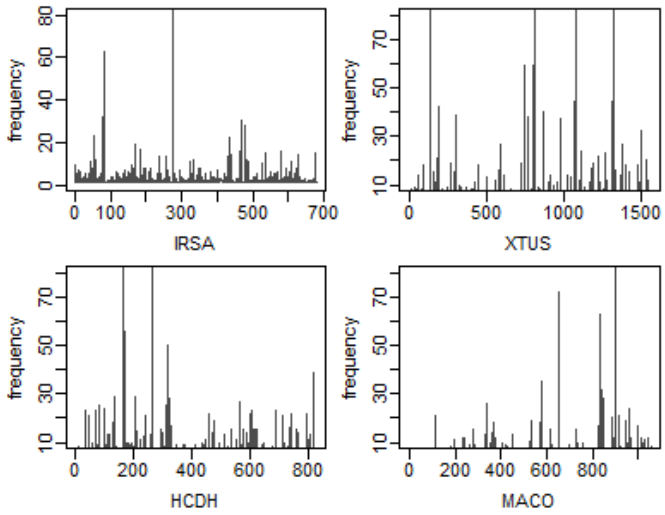

图 12 四个话题词频分布特征

结合表 11 和图 12 来分析语料特征因素对潜在主题数目的影响。如第二个 IRSA 话题，词汇量、总词数相对较少，占据词汇量大部分的主题出现的便会少；再根据图 12 中的词频分布情况，只有三个较为明显的突出部分。与之作为对比的则可以看话题 HCDH，首先句子数目，词汇量以及词表都比较大，但是我们发现其主题也比较少，结合词频特征，其词频高的也比较多，但为何主题数比像 IRSA 中的还要少，原因在于很多词的词频都比较高，但是这些较高的词频往往同时出现，且是关于某个特定的话题，即词之间上下文相关度比较高。

6 建模流程及效果评估

6.1 经验化建模流程

基于全局和局部因子的统一分析，本文给出一个实际建模应用中的经验化建模流程。

1. 产生 hLDA 模型的输入文件以及分析语料中的特征信息：
   做必要的预处理工作产生 hLDA 模型的输入文件，同时分析语料的特征信息，如每个话题下文档的大小、词汇量、词频分布、关联度等统计特征。
2. 评估待建模树结构的深度：
   结合语料规模、高频词语义相似度，以及建模目标等，来最终确定主题建模的深度。一般而言，树的深度至少为三层，且树层数越深，后验推理越复杂，所需的迭代次数也越多，在这最终得到最优结果的稳定性也越差。
3. 是否选择抽样超参：
   后验推理的核心过程便是迭代最优化，因此在足够多次迭代下，往往抽样是较好的选择，但对于 *hLDA* 抽样初始值的选择还没有较为成熟的算法指导，对于一般建模者而言，随机初始化抽样往往不能取得较好的效果。经验表明，一般在两种情况下，我们采取抽样超参的策略。首先，在人工设定超参的情况下，如果建模这对于各个因素的影响因素不清楚。其二，对于运行结果我们不满意，可以通过抽样来确定一个近似的范围，其后在进行人工设置。在抽样超参时应当竟可能增加超参的迭代次数。
4. 每一层的主题参数 η：
   我们注意到如果 η 太大了（如 $\eta > 8.0$），后验的节点聚集便会很大，相应的路径数便会

[5] 主要是词的前后关联顺序不变，有利于我们分析相同词频下相似词的聚集。

变得非常少，反之亦然。同时，我们还应该考虑到最后马尔科夫链的收敛性，对于 η 的先验评估应该尽可能的与下面 *GEM* 参数的调节趋势一致，否则可能导致在迭代次数内评估最优 mode 的失败。

5. 路径词分配的 m，π 参数：

   后验解释倾向于把一般的词放在根节点，具体的词在叶子节点。因此，根据树的深度对 m 进行设置，一般三层或四层的情况如图 4 和表 5 中所示的那样，0.75 已经是很大的值了，其将直接影响词的层次分配和前面 η 参数的调节效果。

6. 非叶子层上的 *nCRP* 参数 γ：

   由公式 3-3 我们知道随着语料数量大小的增加，每一层聚簇数目的期望是呈现 log 增长的趋势，同时在表 3 中我们给出了聚簇数和 γ 之间的关系。我们可以再此基础上相应的较为准确的评估 γ 的超参设置。

7. 树结构先验的参数：

   一个重要的参数便是 SCALING_SHAPE，其直接影响着树的形状。通过对它的调节来对抽样的效果进行修正。参数 SCALING_SCAL 控制着树的规模比例大小。通常我们在对其形状先验预设的基础上，再来调节它。

基于以上建模流程，我们参照 2.2.1 节的框架图进行具体语料超参的设置，顺序建模和部分的循环修正，最终实现最优效果。

6.2 实验及效果评估

我们对 Portal News 语料下的十个话题进行了实验，在三次修正后，对 hLDA 建模结果和人工总结的结果进行了比较，实验中树的层次为 3，如下表 13 所示：

表 13 建模结果与评估得分

theme	level#1	level#2	hLDA#1	hLDA#2	score
XTUS	4	9	5.3	10	4
EUDC	4	5	5.2	9	3
IRSA	2	5	3.9	8	4
HCDH	5	7	4.8	10	4
CQWF	4	9	5.3	10	5
SBAG	4	10	6	10	5
GBAB	6	8	4.8	11	3
LTFC	6	14	7.2	12	5
MACO	4	7	6.7	8	4
ROHN	5	9	7	9	4

其中分数主要分为五个等级，从 1(差)到 5（非常好）。从十个话题的实际建模效果来看，平均都在 4（好）等级左右。接下来我们又选择了 ACL MultiLing 2013 语料下的巴厘气候会议（M004）话题，对比抽样建模（10 万次迭代）、随机建模结果以及本文提出的基于分析框架下的建模，给出一个可视化的建模树结构，每个树节点就是一个主题，我们选取了每个主题上高频词来反映这个主题的特征，如果某个主题节点上词数太少则为了树结构的展现效果，我们会用其父节点上的词填充。

如图 14 所示，整体上来看，两者树结构显得比较单一、少分支，这反映出了建模聚类结构过分的聚集在前面主要路径上，这不符合我们实际语料中子主题的特点。对于抽样建模（左）情况下，中间层词几乎和其父节点一致，根据前面所说，其表示中间层次的词分配极少，大部分词集中在根节点，这种抽样结果显然不能够很好的解释语料特点。对于随机抽样情况，虽然具有一定的层次树结构，但是各层词明显缺乏主题意义上的聚集。

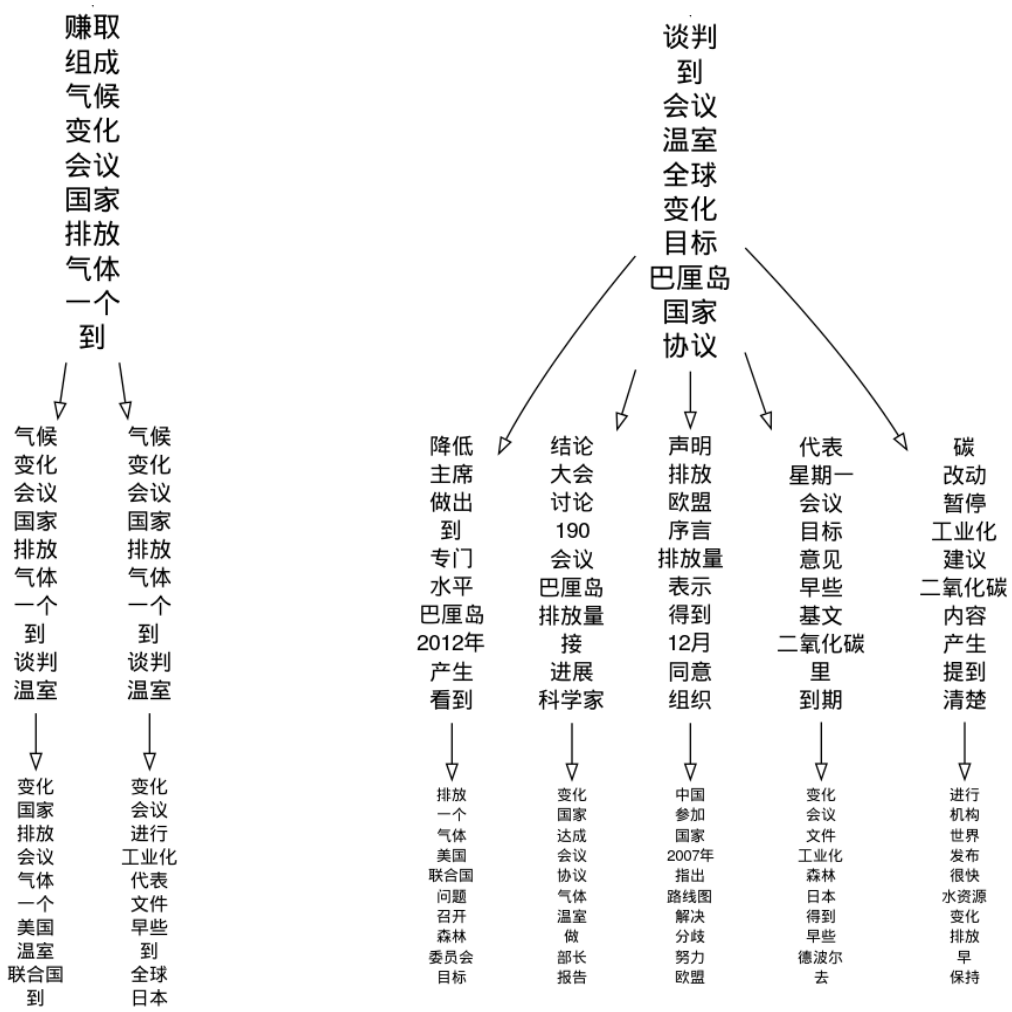

图 14 有限次抽样迭代（左）和随机参数选择（右）结构树

图 15 则是经验化建模流程指导下的层次树结构。如根节点展示了这个主题的一个概括性话题主旨，[巴厘]、[大会]、上关于[全球]气体[排放量]的[协议]问题。接下来在第二层的左边第一个节点显示的是各个参与国家[美国]、[联合国]、[欧盟]等关于[温室气体]、[排放]的谈判。第二层左边第二个节点显示关于[同意]、[接受]大会上设定的一些[决议]等。如此分析接下来分别是美国，联合国其他国家关于[京都][议定书]上结果的意见；关于温室气体排放引发的一些列讨论；[中国]和一些[发展中国家]以及欧盟对[线路图]的立场以及时间规划[2007 年、2020 年]等等。此处由于文档形成的树结构很大，因此我们只选取了几个主要的节点路径上的主要的一些词。同样和人工总结的子主题进行比较发现，其效果是非常好的。

7 小结

我们针对在实际主题建模过程中的建模效果较差，也大多缺乏具体可依据的建模策略的问题，提出了基于关键因素分析的统一分析建模框架，并在此框架基础上，提出了一个统一的建模流程，实验表明取得了很好的效果。但我们也采用了人工评估的方法进行建模效果的评估，这在一定程度上受个人主观性所限。未来仍然后很多值得努力的方向，如关键因子启发式的自调节，如如何自动对建模结果进行合理评估。

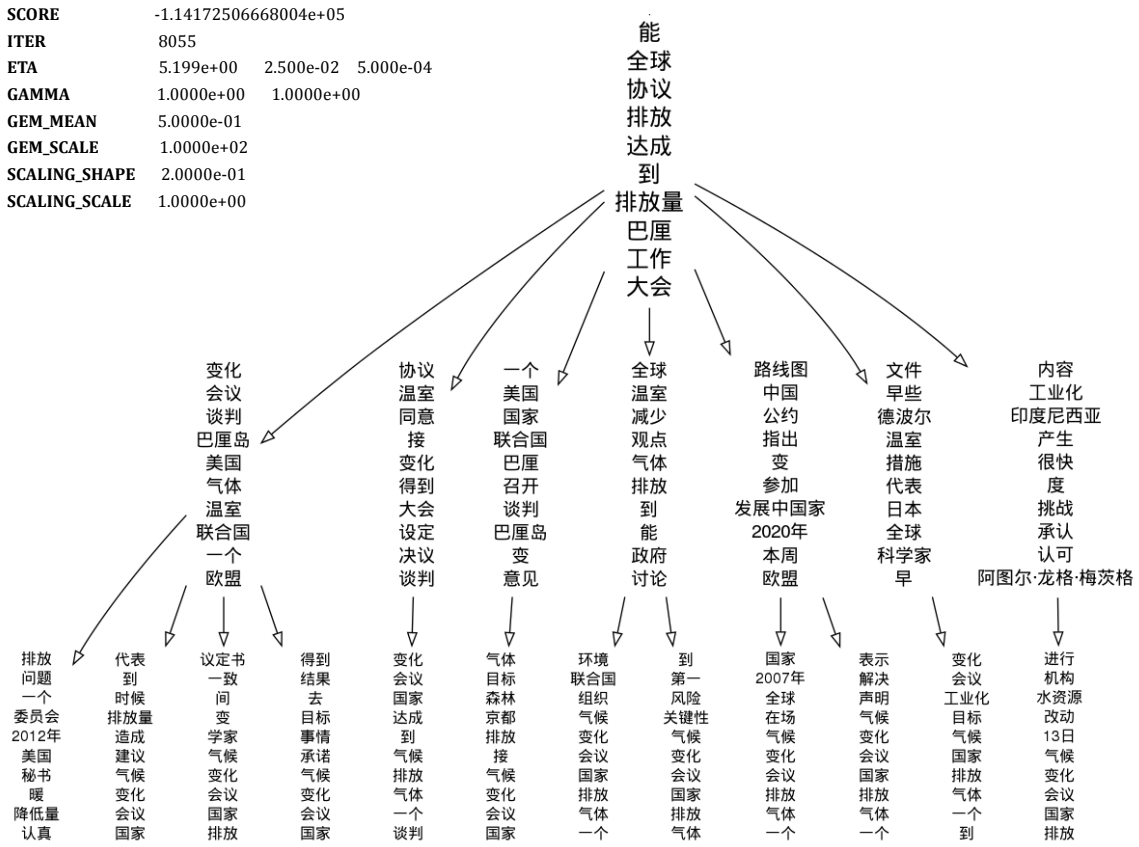

图 15 建模结果树状结构图

参考文献：

[1] Blei D M, Griffiths T L, Jordan M I, Tenenbaum J B. Hierarchical topic models and the nested Chinese restaurant process. Advances in Neural Information Processing Systems 16.Cambridge, MA: MIT Press,2004.

[2] Blei, D, Griffiths, T, Jordan, M. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. J. ACM 57, 2 (2010), 1–30.

[3] Asli C and Dilek H. A hybrid hierarchical model for multi-document summarization. ACL 10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010:815-824.

[4] 刘平安. 基于 HLDA 模型的中文多文档摘要技术研究.MS thesis.北京邮电大学,2012.

[5] GEMAN, S., AND GEMAN, D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Patt. Anal. Mach. Intell. 6, 721–741.

[6] Smith, Adrian FM, and Gareth O. Roberts. "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods." Journal of the Royal Statistical Society. Series B (Methodological) (1993): 3-23.

[7] Blei, David M. "Probabilistic topic models." Communications of the ACM 55.4 (2012): 77-84.

[8] Blei, D., Lafferty, J. Dynamic topic models. In International Conference on Machine Learning (2006). ACM, New York, NY, USA, 113–120.

[9] Blei, David, and John Lafferty. Correlated topic models. Advances in neural information processing systems 18 (2006): 147.

[10] Krestel, Ralf, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. Proceedings of the third ACM conference on Recommender systems. ACM, 2009.

[11] Joon H K, Dong W K, Suin K, Alice Oh. Modeling topic hierarchies with the recursive Chinese restaurant process. Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, New York，2012:783-792.

[12] JH Kim, D Kim, S Kim, A Oh. Modeling Topic Hierarchies with the Recursive Chinese

Restaurant Process. International Conference on Information and Knowledge Management (CIKM), 2012.
[13] Paisley, J., Wang, C., Blei, D. M., & Jordan, M. I. (2012). Nested Hierarchical Dirichlet Processes. arXiv preprint arXiv:1210.6738.
[14] Rodriguez, Abel, and David B. Dunson. "Nonparametric Bayesian models through probit stick-breaking processes." Bayesian Analysis 6.1 (2011): 145-177.
[15] Ferguson, Thomas S. "A Bayesian analysis of some nonparametric problems." The annals of statistics (1973): 209-230.
[16] Bernardo, José M., and Adrian FM Smith. Bayesian theory. Vol. 405. Wiley, 2009.

作者联系方式：
北京市海淀区西土城路 10 号　　100876
衡伟　　18810542083　　hengweipublic@yeah.net
于佳　　13811170303　　yujia920@126.com
李蕾　　13810927112　　leili@bupt.edu.cn