**Collapsed Gibbs Sampling in LDA**

朱佳晖 zhujiahui@whu.edu.cn

# LDA符号

：文档集合（词汇的向量）

：潜在主题的个数

：词汇表中的单词个数

：文档数

：某个文档中的词汇数。具体在使用时往往是，表示第个文档中的词汇数

：文档-主题分布，

：主题-词汇分布，

：某一个文档的主题分配序列。表示第*d*个文档中的第*n*个词所对应的主题编号，编号范围(1到)

：文档-主题分布的Dirichlet先验参数，

：主题-词汇分布的先验参数，

总体目标：根据文档集合以及先验参数、，估计，然后根据得到和。

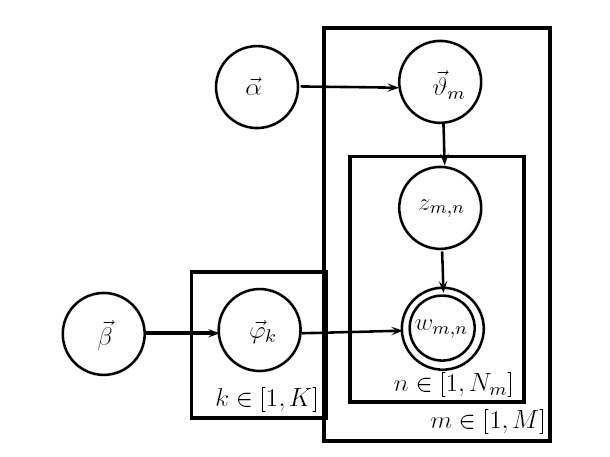


图1

除了文档的词汇向量，以及先验参数，，图中的其余的各个变量都是不可观测的。

# K维Dirichlet分布概率密度函数



文档-主题分布服从参数为的Dirichlet分布。

# 重要的概率公式

* 文档-主题分布，主题-词汇分布，主题分配序列，以及所有文档的联合（生成）概率为：



结合图1，可知表示的生成概率，文档-主题分布服从参数为的Dirichlet分布，表示第个文档的第个词汇（对应的位置上）的主题被分配为的概率，而表示这个该词汇的生成概率。表示该主题-词汇分布服从参数为的Dirichlet分布。

由于带有连乘符号，为了后续的表示方便，将公式(2)写成如下形式：



* 文档集合的生成概率为：



该公式推导自公式(2)，文档集合的生成概率即为(2)中的联合概率的关于的边缘概率(marginal distribution)。那么，只需将公式(2)依次关于，和求积分即可。且由于主题分配序列是一个离散的量，因此积分便是求和。

* 则文档-主题分布，主题-词汇分布以及文档主题分配序列的后验概率(posterior distribution)为：



本质就是贝叶斯公式：



求关于待估计参数的后验概率往往是进行参数估计的第一步，在EM等一系列算法中皆是如此。此类方法便是极大后验估计(MAP)法。因此，需要构造出文档-主题分布，主题词汇分布和文档主题分配序列的联合后验概率。

# Collapsed Gibbs Sampling

## 4.1 直接求解的困难性

文档-主题分布，主题词汇分布和文档主题分配序列的后验概率已经构造完毕，按道理就可以采用相应的参数估计方法估计，和了。估计参数需要极大化后验概率，即求解使得后验概率达到最大值（极大值）情况下的值作为待估计参数的最佳估计值。

但是，在这里，情况不是很妙。因为里面含有积分的形式，导致求解起来很麻烦，是不可手算的。由于无法直接用微积分的技巧求得精确解（解析解），因此不妨求个数值解。求解一般可使用随机算法或者近似算法。在2003年LDA的原版论文中，采用的是变分推断法，是属于近似算法。但变分法公式复杂，推导困难，因此不妨采用随机算法。

随机算法，顾名思义，就是随机采样，当精确地求解总体比较困难的时候，可以通过随机采样的方法，通过样本的相关数字特征求解总体的相关数字特征。

在LDA中，随机采样的算法主要有马尔可夫蒙特卡罗采样(MCMC)，吉布斯采样(Gibbs Sampling, GS)便是属于一类特定的MCMC算法。Gibbs Sampling在LDA推断中广为运用，因为操作简便，可编程性强，小规模数据上具有较高的效率。因此，不少基于LDA的改进模型都是采用Gibbs Sampling方法求解。

## 4.2 推断的核心变量



以上的后验概率表达式中，难求的主要是分子。如果按照LDA变分推断的思路，核心就是解决的近似求解问题。但是如果采用Gibbs 采样，不妨直接近似等式左边的。但是显然，直接Gibbs采样这个概率分布值也是很困难的，因为、、彼此之间有耦合。同时，我们也可以发现，文档-主题分布和主题词汇分布均可由主题分配序列推断得到。也就是说，一旦知道了，就相当于知道了和。由推得的图示原理如图2所示，对于由推得，也是类似的。

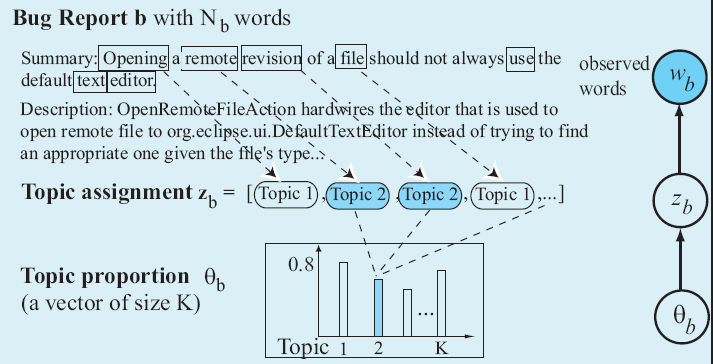


图2

由上面的分析可知，一旦知道了主题分配序列，所有问题都迎刃而解了。因此，最重要的就是近似得到，即求解联合概率等价于求解。是对关于和求积分得到的，通过积分，和都被隐去了。直接对的采用叫Gibbs采样，而对的采样叫坍缩Gibbs采样(Collapsed Gibbs Sampling, CGS)，因为和都被隐去了。一般在LDA中，用的都是坍缩Gibbs采样。

## 4.3 GS和CGS采样原理

采样，顾名思义就是随机选取。Gibbs采样首先基于马尔可夫链，然后基于随机变量的后验条件概率分布来采样一个值。在马尔可夫链的基础上，若干次采样之后便会收敛到一个稳定的状态，这个稳定状态下的值就是最终的采样结果。

Gibbs采样一般使用的场景往往是这样的：需要从联合概率分布中采样某个或某几个随机变量的值，但是该联合概率分布的表达式是未知的或者是很难写出来的，但是关于某些变量的条件概率分布是已知的。

比如：现在要从的联合概率分布中采样的一组值，但是的表达式是未知的，但是存在关于每个变量的条件概率。那么采样的过程大致可以是这样的：

1. 随机初始化一组的值
2. For *t* = 1,2,…, *T*



该过程最终收敛，收敛状态下的那组值便可作为最终的采样结果。

那么什么是坍缩Gibbs采样呢？假设有联合变量a,b,c，那么隐去b之后的坍缩Gibbs采样就是：



这样之后b就直接被无视了。

# LDA Collapsed Gibbs Sampling

## 5.1 采样建模

前面说到，在坍缩Gibbs采样的推断方式下，最重要的就是近似得到，也就是采样一组主题分配序列。采样的分布是，其中。很显然这是一个关于的联合概率分布，而且其具体的分布函数是很难获得的。所以，如果要使用坍缩Gibbs采样，则先要有每一个主题在其余主题下的条件概率，为了方便称呼，称其为主题互补条件概率。

其中表示当前词汇所分配的主题，表示除去当前词汇的其他所有词汇的主题分配序列。

因此，在LDA的坍缩Gibbs采样过程中，最重要的便是计算主题互补条件概率。

## 5.2 公式转化

利用贝叶斯公式将转化：



所以求等价于求。而属于联合概率的边缘概率，因此要得到，只需要将联合概率依次关于和求积分：



进一步展开：



拆分到这里很显然已经差不多了，下一步就要涉及到具体的概率分布计算了。

## 5.3 Dirichlet分布与Δ函数的引入

若要求解公式(8)需要具体展开Dirichlet分布。由于服从Dirichlet分布，因此根据公式(1)有



而服从多项式分布，即



其中表示与该文档的第个词所分配主题(编号)在当前文档-主题分布中的概率分布值。因此有：



令



称带有的该类式子为函数，则



所以有



即



结合公式(10)和公式(13)有



* 第一行是公式展开
* 第二行是代入函数，并移项
* 第三行是分布的叠加，且运用到了公式(13)。其中表示第个文档中的单词被分配给第个主题的次数。。

由以上公式可知对于这个过程的分布为 ，服从多项式分布，(注意：它与是同一个，写成包含了多项式分布的参数，而的形式看不出其多项式分布的参数)，服从Dirichlet分布。所以从变量和参数的角度来看，的是由和叠加而成的。因此这个Dirichlet分布与这个多项式分布共轭，可称为Dirichlet-Multinomial共轭结构，如图3所示：

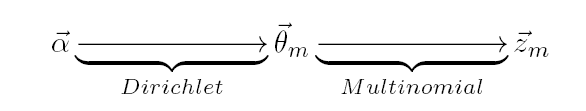


图3

同理，对于这个过程，也是一个Dirichlet-Multinomial共轭结构。



* 第一行是公式的展开
* 第二行代入函数
* 第三行移项
* 第四行分布叠加，再运用公式(13)化成函数。其中表示第个单词被分配给第个主题的次数，。

关于Dirichlet-Multinomial共轭的详细证明参见公式(21)和公式(23)

将公式(14)和(15)代入公式(8)，可得



## 5.4 采样结果表达式

其实最终的结果已经呼之欲出，但是计算才刚刚开始。结合公式(6)和公式(16)，主题互补条件概率的计算如下：



* 第二行是代入函数并移项
* 第三行是将连乘拆项，其中表示排除当前文本(按顺序应该叫“第个文本”)，表示排除当前文本的当前词汇(按顺序应该叫“第个词汇”)。
* 第四行是移项，将“排除当前文本和当前词汇”的式子与“包含当前文本和当前词汇”的式子分开
* 第五行表示只跟包含“当前文本和当前词汇”的式子正相关。为什么呢？因为对于“当前文本和当前词汇”来说，“排除当前文本和当前词汇”的式子无关的
* 最后一行再将无关(恒定)的式子略去。

公式(17)仍可继续计算：



* 第二行表示排除当前主题的情况
* 从第二行第三行是个关键。由到是相等的，由到是近似的。对于主题-词汇部分也是这样的
* 第四行是拆项，利用这个性质。对于主题-词汇部分也是这样的
* 第五行是利用第四行拆分的结果合并归约
* 第六行将所有的连乘式子略去，作为近似。为什么可以略去？因为他们的值对于当前的状况(当前的状况指的是“当前在第个文档的第个词汇”)是无关的

坍缩Gibbs采样的最终结果：



其中表示“排除当前文本(应该叫第个文本)和当前词汇(应该叫第个词汇)”的情况下，第个文档中的词汇被分配给第个主题(也就是当前主题)的次数。

其中表示“排除当前文本(应该叫第个文本)和当前词汇(应该叫第个词汇)”的情况下，第个词汇被分配给第个主题(也就是当前主题)的次数。

## 5.5 计算文档-主题分布和主题-词汇分布

基于公式(19)进行吉布斯采样，最终收敛时的结果便是主题分配序列，接下来便可基于求解文档-主题分布和主题-词汇分布。

在得到最终结果之前，先了解一下Dirichlet分布的期望，对于公式(1)的Dirichlet分布，其期望为：





* 第一行是贝叶斯公式的运用
* 第二行是公式的展开
* 第三行用到了公式(14)的结论

由以上结果可知服从参数为的Dirichlet分布，所以可用该Dirichlet分布的期望来估计的值：



同样地



服从参数为的Dirichlet分布，所以的估计值为：



# LDA Collapsed Gibbs Sampling框架

坍缩Gibbs采样的结果表达式(公式(19)，公式(22)和公式(24))：



下面给出LDA 坍缩Gibbs采样的算法

|  |
| --- |
| 算法：LDA Collapsed Gibbs Sampling (LDA CGS) |
| 输入：所有文档的词汇表示，主题个数，Dirichlet先验参数和 |
| 输出：主题分配序列，文档-主题分布，主题-词汇分布 |
| 初始化：文档-主题计数，文档-主题总和计数，  主题-词汇计数，主题-词汇总和计数 |
|  |
|  |
| 随机初始化当前词汇的主题分配，随机初始化可采用均匀分布 |
| ，，， |
| // iterations为坍缩Gibbs采样的迭代次数 |
|  |
|  |
| ，，， |
| 依据公式(19)计算 |
| 为当前词汇采样新的主题分配 |
| ，，， |
| 依据公式(22)和公式(24)计算文档-主题分布和主题-词汇分布 |