Topic Models

fancyerii fancyerii@gmail.com

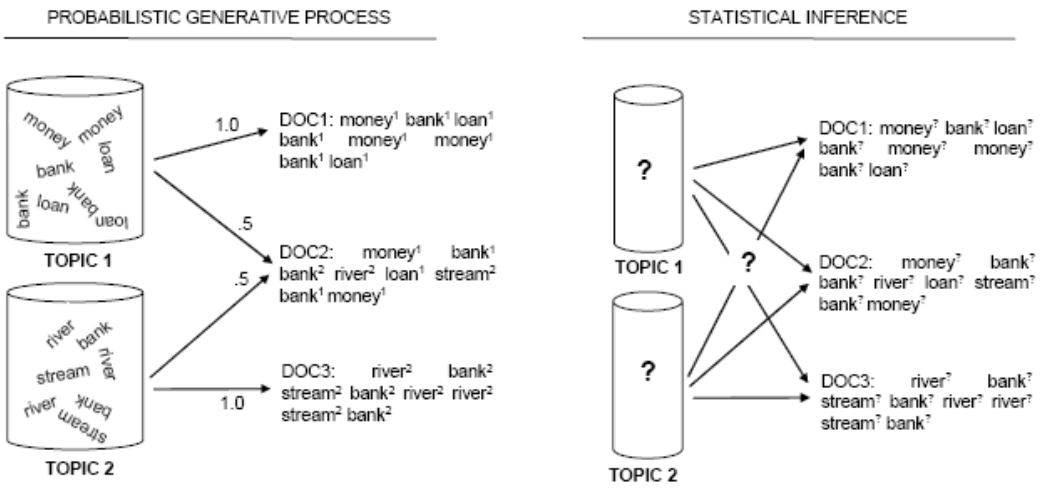
1， Probabilistic latent semantic indexing(1999)

PLSI 模型一开始像 LSA 一样用来对向量空间模型进行降维，把一篇文档降到主题空间上。

PLSI 的思想是：每篇文档是由不同的主题(Topic)混合而成，而每个主题有不同的词分布。

下面的左图是一个生成文档的例子。左边的圆柱代表主题，共有两个主题，分别表示 money 和 river。根据不同的比例而混合出三篇文档。

右图表示由我们观察到的三篇文档做推理(inference)求出 topic 的词分布以及每篇文档的主题分布的过程。



由于有了隐变量 z，本文使用了 EM 算法来估计参数。

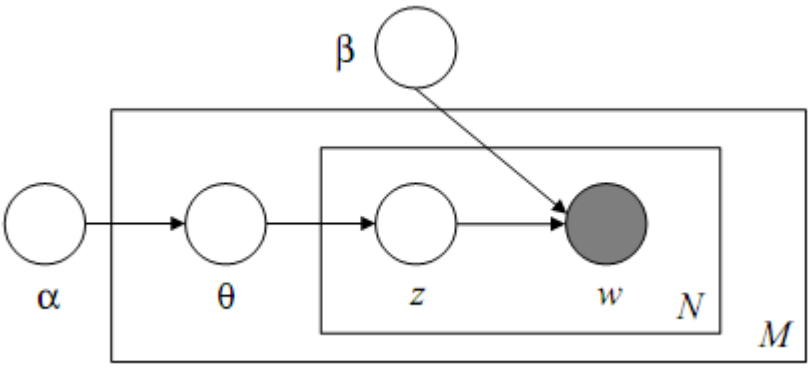
它的一个缺点是：把文档 d 当成多项式(multinomial)分布随机变量，它只能取训练语料中出现的文档，而没有出现的文档 p(d)=0。因此有新的文档加入时要么重新训练，这样主题的词分布可能发生变化；要么使用 fold-in 的方法。

2， Latent Dirichlet allocation(2003)

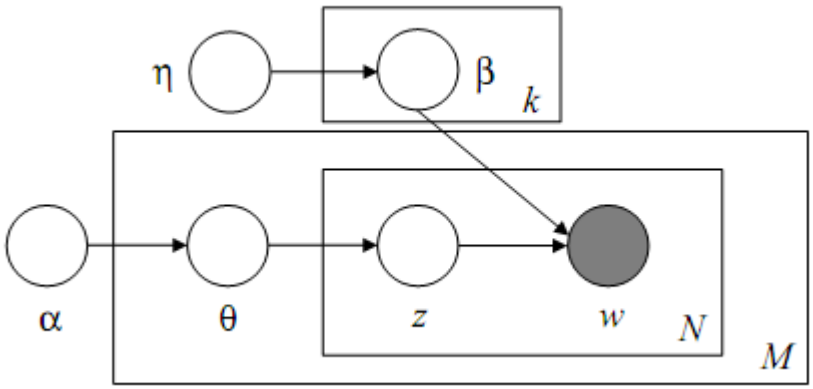
为了克服上面的缺点，需要定义文档的完整生成过程。我们关注的是文档的主题分布，而不是文档本身，所以应该把主题的分布 θ 作为随机变量。为了 θ 定义一个先验分布，就需要给概率测度定义一个分布，由于主题的分布 θ 是离散的，最常见的分布就是 Dirichlet 分布了。它的好处是：与多项式(multinominal)分布是共轭(conjugate)的，这样给推理和参数估计带来很多方便之处。因为这个原因，Dirichlet 分布在离散的贝叶斯网中经常被用来作为先验分布。因此联合概率分布为：

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta^{\alpha_1 - 1} \theta^{\alpha_2 - 1} \ldots \theta^{\alpha_K - 1}$$



为了使数据平滑，是没有在训练样本中出现的词也有一定概率，可以再给 β 加一个先验的分布，得到如下的图模型。



贝叶斯网有两个基本问题，推理和参数学习。推理是指已知联合分布，求条件分布；参数学习是指根据数据对参数做最大似然估计或者贝叶斯估计。精确推理(Exact Inference)需要对联合分布积分或者求和，为了共享中间结果，一般使用 junction tree 算法，在图结构比较复杂的情况下是 NP 问题。近似推理(Approximate Inference)的方法有：MCMC（包括最简单常用的 Gibbs Sampling）；变分法（Variational Methods）；循环传播算法（Loopy Propagation）。

Blei 等使用的推理方法是变分法；由于有隐变量 z 和 θ，为了估计参数 α β 而使用了 EM 算法。Griffiths（2004）提出了使用 Gibbs Sampling 来做推理。因为在贝叶斯估计中，如果数据够多的话，参数取值对计算后验概率影响不大，所以他假设 α β 是已知的，并给出了一个经验值：α =50/T, β =0. 01。Gregor Heinrich 的 *Parameter estimation for text analysis* 对 Gibbs Sampling 的实现做了详细的推导。关于贝叶斯网和推理的参考文献：张连文(2006);M. Jordan(2003a); M. Jordan(2003b); M. Jordan(1999); Bishop(2006)。

作者在 TREC AP Corpus 上做了分支度(Perplexity)的实验，LDA 的分支度比 PLSI 要低。从降维角度考虑，用来作为文本分类的特征，LDA 也比 PLSI 好。

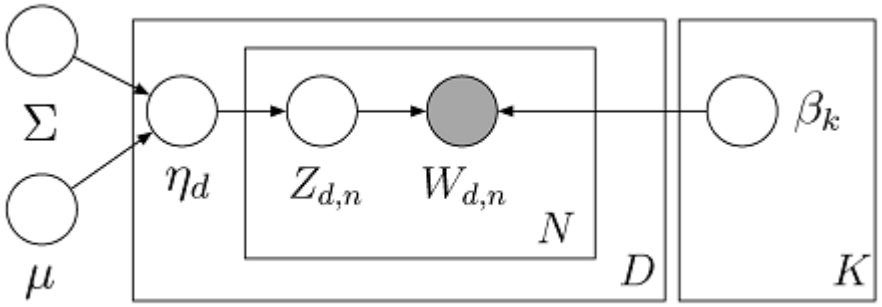
3， Correlated Topic Models(2006)

为什么 θ 的先验要用 Dirichlet 分布呢？前面说了，是因为它与多项式分布共轭，便于计算。但是它也有缺点，由于它的分量是独立的，不能体现出 topic 之间的关系来。本文作者提出了 Correlated Topic Models（CTM），他把多项式分布写出自然参数(natural parameterization)的形式：

$$p(z \mid \eta) = \exp\{\eta^T z - a(\eta)\}$$

这样不同的实向量 η 就对应不同的多项式分布。然后认为 η 是高斯分布的，这样就引入了协方差的信息，可以挖掘两个不同主题之间的关系。

下面是它的图模型，与 LDA 唯一的区别就是 z 由 η 来表示。

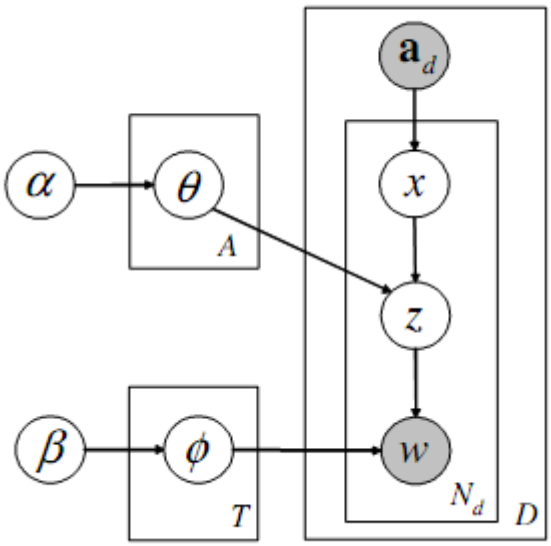


由于 logistic normal distribution 和多项式分布不是共轭的，所以推理会更加困难。本文使用的是变分法来推理。

作者用 Science 的文章做了实验，说明确实能发现一些 topic 之间的关系，并且 CTM 的分支度比 LDA 要低。

4， Author-Topic Models(2004)

在一些应用中，可能需要知道一个 Author 的 Topic 分布。为此，Steyvers 等人提出 Author-Topic(AT)模型。它认为每个作者有一个 topic 的分布，每篇文档的作者是已知的，文档的生成过程是：随机选择一个作者，根据这个作者的 topic 分布，生成一个词。它的图模型如下：

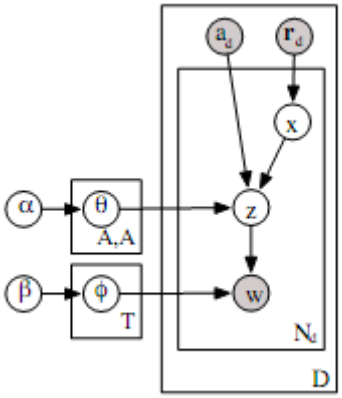


本文作者使用的推理算法是 Gibbs Sampling。

实验部分对 NIPS 和 CiteSeer 中的文档做了分析，发现这个模型确实能够挖掘出作者的 topic 分布。也做了分支度的实验，结果是：当已知词很少时，AT 模型比 LDA 好，但是已知的词比较多时，AT 不如 LDA。原因比较直观——当已知的信息很少时，作者的 topic 分布能够提供有用的信息；但信息足够时，作者的 topic（一般的信息）对于具体一篇文档的 topic 分布反而会是干扰。在 future work 中，作者认为这个模型可以为用户提供一些 Author 的信息，帮助用户了解某个作者的研究领域。另外也认为可能会被用来做 author identification。

Yang Song(2007)利用类似的模型来做科技文献中作者的人名消歧和个人网页的消歧，取得了不错的效果。

但是在 Email 这种文档中，邮件的发件人和收件人地位是不对等的。比如一封经理发给员工们的邮件，和员工发给经理的邮件，使用的 topic 是不一样的。如果简单的把发件人和收件人都当成文档的作者，那么就不能区分发件人和收件人的身份。McCallum 等人为此提出了 Author-Recipient-Topic(ART) 模型，并用这个模型来挖掘作者的 Social Networks 和作者的 Roles。它的图模型如下：



与 AT 模型不同之处在于：主题分布 θ 不是由一个作者决定，而是由发件人和收件人同时确定。也就是说，每两个作者（一个发件人，一个收件人）都有一个词的分布，代表了发件人对收件人说话的 topic。

使用 ART 模型来做 Social Networks 相对与一般的 Social Networks Analysis(SNA)的好处是：普通的 SNA 只考虑 Social Networks 的结构，Entity 之间边的连接，而没有考虑边的内容（如邮件）；而使用 ART 结合了 Topic 和边的信息，能做得更好。

本文使用的推理算法是 Gibbs Sampling。由于 ART 没有显式的对 Role 建模，作者又提出了 Role-Author_Recipient-Topic Models。

5， HMM-LDA Model(2004)

由于 LDA 等 Topic Models 的 Bag-Of-Words 假设，这些模型没有考虑词之间的顺序。

Griffiths（2004）认为：一个词出现在句子之中可能有两种原因，一种是它起到的是句法(Syntatic)的功能，也就是说它是虚词；另一种功能就是表示语义（Sematic）的功能，也就是实词。句法的约束一般是 short-range 的，一般不会超过一个句子；而语义的约束一般是 long-range 的——同一篇文档的不同句子会有相近的内容，使用相似的词。句法的约束一般通过 Hidden Markov Model(HMM)，Probabilistic Context Free Grammar(PCFG)等模型来实现；而语义的约束一般使用 Topic Models 来实现。作者认为 Combine 这两者会得到更好的模型，因此提出了 HMM-LDA 模型。下面的是它的图模型和文档的生成过程：

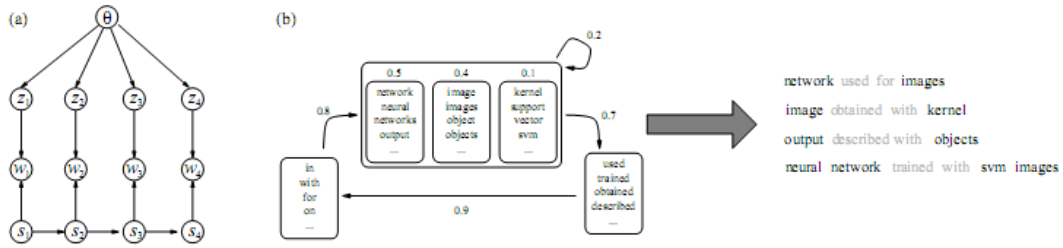


Figure 1: The composite model. (a) Graphical model. (b) Generating phrases.

文档中每个词的产生除了依赖与它的 topic 之外，还与当前的状态 s 有关。当要确定一个词是怎么产生的时候，首先根据当前 s 的跳转概率，确定下一个状态是什么。如果下一状态需要一个实词，那么就根据 topic 产生一个实词，否则根据跳转产生对应的虚词。

举个例子，假设有两个 syntatic 类（一个是动词的被动语态，另一个是介词），一个实词类（名词，作者认为表示语义概念的一般是名词，而动词，介词等表示的是句法的约束）。句法的约束是名词+动词被动语态+介词+名词（它的跳转概率概率较大），那么就会产生这样的句子：

Network used for images; image obtained with kernel。

如果换一篇文档（换了 topic），那么可能是：

Variables associated with a distribution。

本文使用的是 Gibbs Sampling 推理算法。

作者做了一些实验来证实自己的想法，可以聚类出虚词类和 topic，

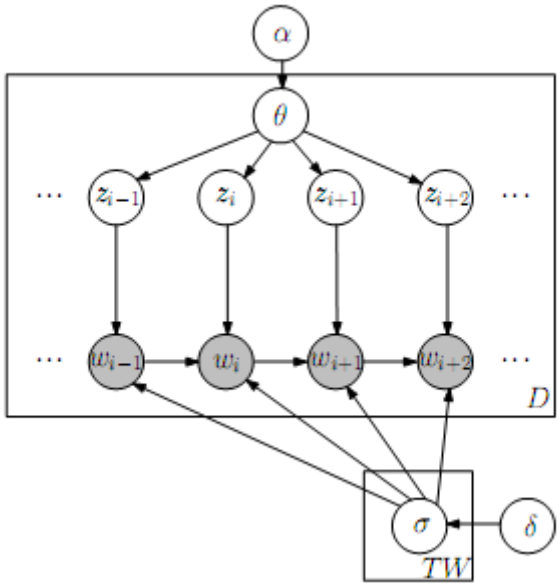
而且这些虚词类的语法功能确实很类似；可以知道句子中的一个词是实词还是虚词；可以做无监督(unsupervised)的词性标注(Part-of-speech tagging)，比其它方法要好一些；做文本分类，比 LDA 要差一下，作者认为可能是表示关系的数据比较少。

6， Topic Modeling: Beyond Bag-of-Words (2006)

传统的语言模型(Language Model)只考虑词之间的跳转概率，比如 Bigram 模型，它由 p(wt|wt-1)来描述。一个词出现的概率取决于前一个词是什么，Hanna Wallach(2006)认为一个词除了与前一个词是什么有关外，还与前一个词的主题有关。比如前一个词是 hard，它有很多意思。作为困难的和硬的词义时，后面接的词的概率肯定是不同的。那怎么来描述前一个词的词义，也就是当前词的上下文(Context)呢？作者认为可以使用 topic model 来描述。下面是联合概率分布：

$$P(\boldsymbol{w},\boldsymbol{z}|\Phi,\Theta)=\prod_i\prod_j\prod_k\prod_d\phi_{i|j,k}^{N_{i|j,k}}\theta_{k|d}^{N_{k|d}},$$

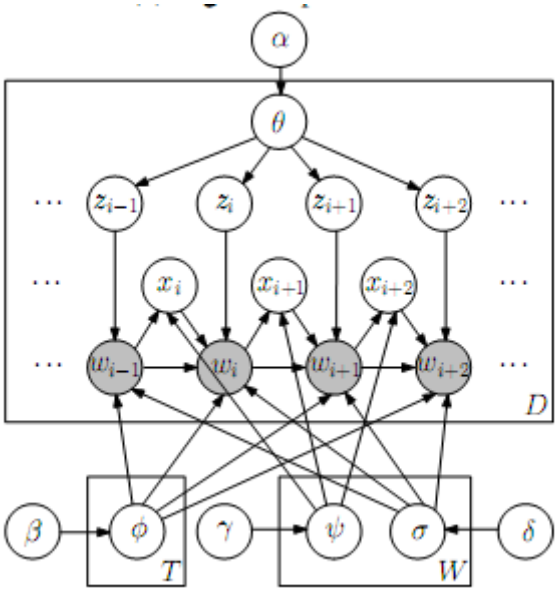
可以拿它与 LDA 和 Bigram 相比较。LDA 中主题的词分布Φ只与 topic k 有关；而 Bigram 的Φ只与前一个词 j 有关；这个模型在 Combine 了它们。



7， Topical N-grams (2005)

与上文不同，McCallum 认为不是任何两个词都可以是一个 phrase，两个词可能没有什么关系，也可能是一个 phrase。

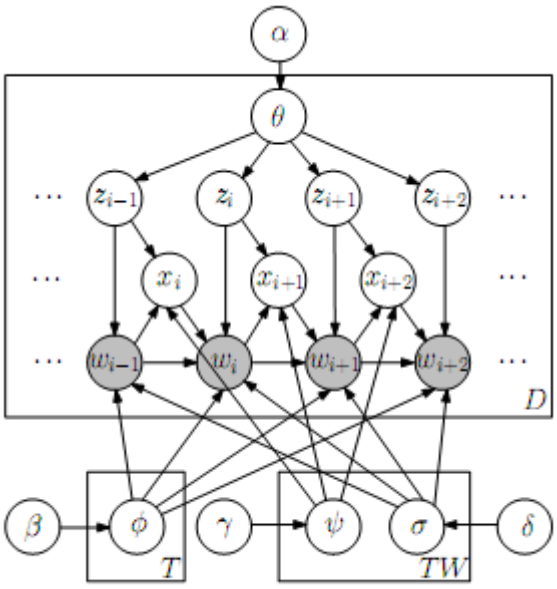
最早出现关于搭配(Collocation)的模型出现在 Griffiths 的 Matlab Topic Modeling Toolbox 中，并没有发表文章。图模型如下：



(b) LDA-Collocation model

生成一个词时，先看看 Bernoulli 分布的随机变量 x 的值，如果是 1 的话，代表是一个搭配，根据前一个词生成当前词；否则不是搭配，之间根据主题 z 生成当前词。

McCallum 认为这个模型的缺点是：它的搭配是固定的，两个词要么就是搭配，要么就不是搭配。他认为两个词是否是搭配还要看它们出现的上下文，比如 white house 在政治相关的上下文中可能是一个特殊用法，是搭配，而在建筑这样的上下文中可能就不是搭配。因此他提出了 Topic N-gram 模型，图模型如下：



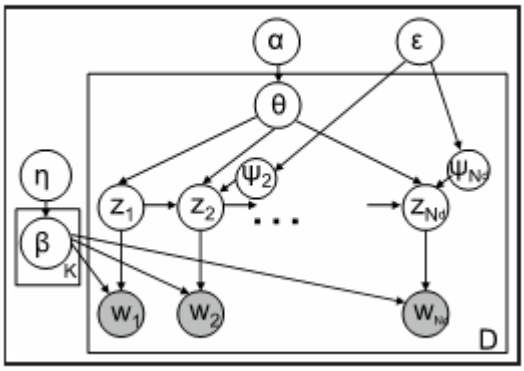
(c) Topical n-gram model

从图中看出，x_i 取决于上下文，也就是前一个词的 topic z_{i-1}。

8， Hidden Topic Markov Model (2007)

上面的文章在解决 bag-of-word 问题时，关注的是词直接的跳转，用途是处理词之间的关系。而本文注意了文档的篇章结构信息——同一个句

子的 topic 应该是一样的，相邻的句子的 topic 可能比较接近。基于以上假设，Grubber(2007)提出了 Hidden Topic Markov Model(HTMM)。图模型如下：



c

生成过程如下：如果不是句首，ϕ=0，否则ϕ服从 Bernoulli 分布。意思是：如果不是句首，主题不变；如果是句首，换了一个句子，那么有一定的概率换主题，但很可能主题不变，这样就符合前面的假设——同一个句子主题一样，相邻的句子主题相似。它的一个特例：每换一个句子，ϕ=1，重新生成一个句子的 topic，这可以看成 bag-of-sentence 的假设，把句子看出最小的单元。

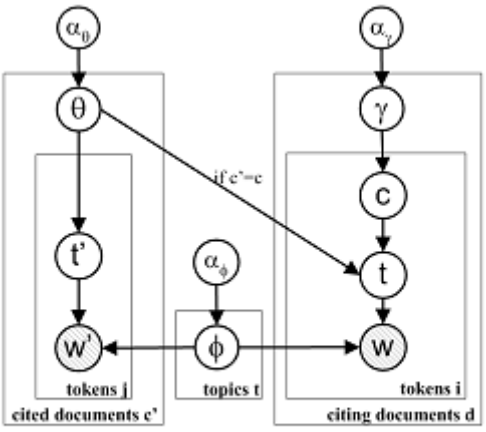
分支度的实验表明 HTMM 比 LDA 要好。在观察的词较少时 bag-of-sentence 的 HTMM 比 LDA 好；词较多时 LDA 较好。

9， Unsupervised Prediction of Citation Influences (2007)

上面的文章解决了同一篇文章中词的顺序的问题，但是没有考虑文章之间的顺序关系。文章之间的顺序关系有两种：引用关系，时间先后顺序。

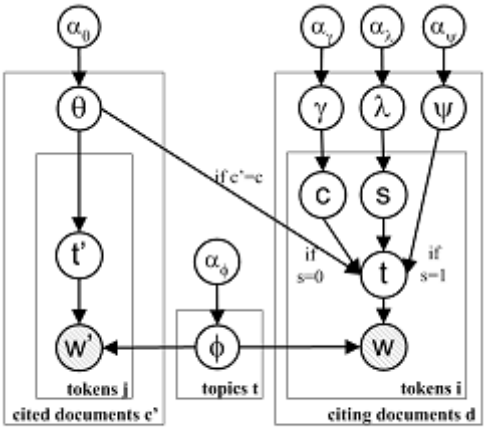
Dietz(2007)提出了两个模型 CopyCat 和 Citation Influence 模型。CopyCat 模型的把文档分成两类，引用的文档和被引用的文档，如果一篇文档同时被引用和引用别的文档，那么把它当成两篇文档。

被引用的文档的生成过程和 LDA 是一样的。引用的文档认为完全是“抄袭”它所引用的文档。所以叫它 Copycat，γ 表示“抄袭”的比例。



b) Copycat Model

如上图所示，当要生成 citing document 中的一个词时，首先根据 γ 来选一篇 cited document 的主题分布 θ，然后根据它来生成一个词。但是认为 citing document 完全是“抄袭”的是不准确的，所以作者又提出了 Citation Influence 模型。



c) Citation Influence Model

这个模型认为文档部分是“抄袭”来的，部分是自己写的。

它使用一个 Bernoulli 分布的变量 λ 来确定是“抄袭”还是原创，如果是原创的话就像 LDA 一样根据 Ψ 来生成词，否则和 Copycat 模型一样。

作者使用这个模型来获取一篇 cited 文档对 citing 文档的影响 (influence)，这与传统的文献计量学不同。传统的方法根据被引用的次数来确定一篇文章是否重要，并认为重要的文章对别的文章的影响比较大。传统的方法没有考虑文章的内容，而且很多引用很可能是出于尊敬等等其它原因。

10， Link-PLSA-LDA (2007)

Cohen (2007)认为，上面的模型只能获得一篇文章对另一篇文章的 general 的影响，而不能知道在某个具体主题下一篇文章对另一篇文章的影响。

最早把文档之间的链接加入模型是 Hofmann(2001)，他们在 PLSA 的基础上加入了链接，除了认为词是由主题生成外，还认为链接(link)也是由主题产生的。它的假设是：如果两篇文章都链接很同一篇文章，那么它的主题也就很相似。Lafferty (2004)用 LDA 做了类似的模型。这两个模型基本类似，除了后者用 LDA 代替了 PLSA 外。下面是 Link-LDA 的图模型：

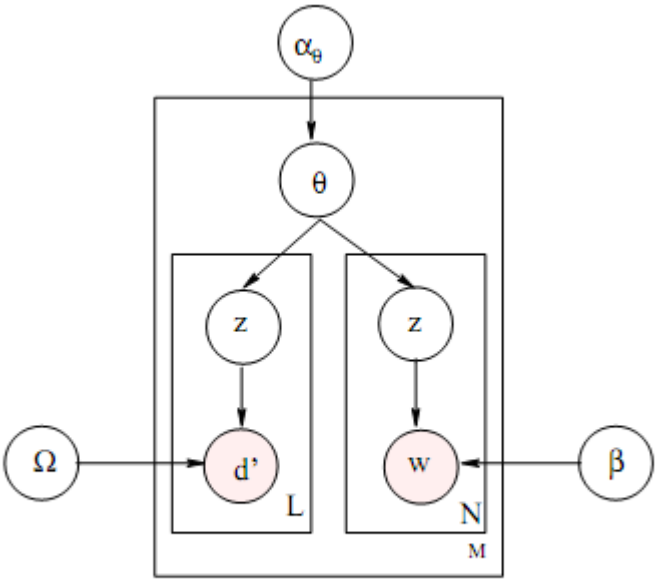


Figure 1: Graphical representation of the Link-LDA model

Cohen (2007)认为 LDA 利用词的共现来发掘主题这个聚类；Link-PLSA 和 Link-LDA 只是利用了引用的共现，而没有显式的建立引用和被引用文档之间的关系。Dietz 的模型虽然显式的建立了这种关系的模型，但是却不能知道在某个主题下一篇文章对另一篇文章的影响。

因此他们提出了 Link-PLSA-LDA 模型。

下面是它的图模型：

它和 Link-PLSA 的区别是：前者的 Ω 与主题相关，也就是说每个主题有一个 link 的分布，引文和被引文是对称的；而新模型的引文和前面一样，被引文都放在一起，相当于先训练出一个关于主题的词分布和 link 分布的模型出来。然后引文利用这个主题的 link 分布，从而显式的建立了引文和被引文的关系。

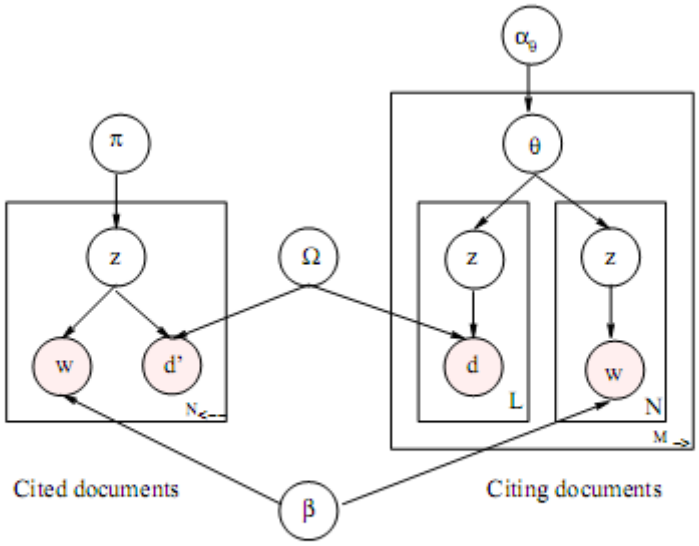


Figure 2: Graphical representation of the Link-PLSA-LDA model

作者使用了这个模型来分析 blog 数据，可以挖掘出主题以及在某个特定主题下一篇文章对另一篇文章的影响的大小。

11， Topics over Time (2006)

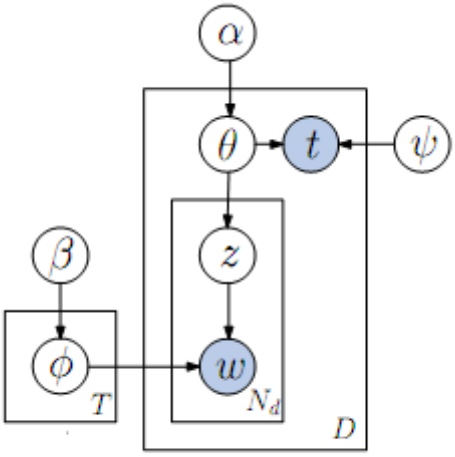
如 9 所述，除了引用关系，文章之间还有时间的先后顺序关系。或者说，Topic 并不是一成不变的，同一个 topic 的词分布是随着时间变化的。最早关注这个问题的是 Griffiths(2004)，作者在文章中用 topic models 来分析热门的(hot)topic 和冷门的(cold)topic，使用的方法很简单，由于他使用的 Corpus 中文章的类别是知道的，所以直接按年计算这一年文章的主题分布 θ，然后把聚类出来的主题与已知的主题人工对应上，这样就可以分析主题随时间的变化趋势了。

显然上面的方法是有问题的，topic models 得到的 topic 不见得能和人工分类的 topic 一一对应。

McCallum(2006)提出了 Topics Over Time(TOT)模型，他的想法是：不同的时间出现主题的分布是不同的，也就是说，文章的时间与它的主题是相关的。比如某年流行神经网络，而这篇文章的主题关于神经网络的概率比较大，那么它在这年被发表的概率也较大。

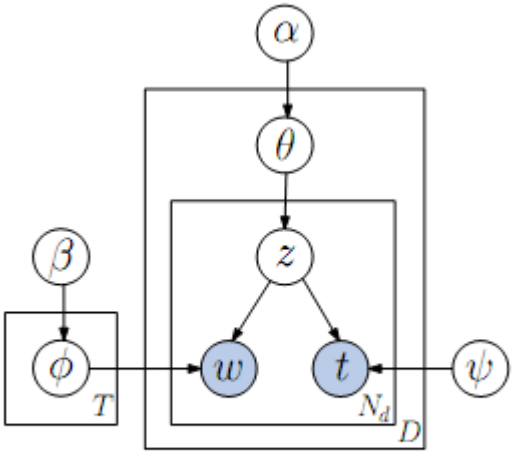
图模型表示如下所示，生成一个词的方式和 LDA 一样的，只不过每篇文章除了生成词外还要生成这篇文章的发表时间。

可以看出，这里把发表时间当成了随机变量，和一般的时间序列分析以及切分时间片的方法是很不一样的。因此这个模型也能估计一篇文章的发表时间。



(b) TOT model,
alternate view

另外为了 Gibbs Sampling 方面，可以把时间与每个词联系起来，如下图所示。如果同一篇文章的 t 是相同的话（实际上我们只能认为是相同的，否则不知道），那么和上面的是非常类似的，只不过下图的 t 依赖与主题的类别 z，这是个有限的离散的量；而上面的 t 依赖与连续的 θ，这个很难计算，除非再为它建一个模型。



(c) TOT model,
for Gibbs sampling

本文作者使用了这个模型来分析主题随时间的变化，以及用它来解决估计一篇文章的发表时间等问题。

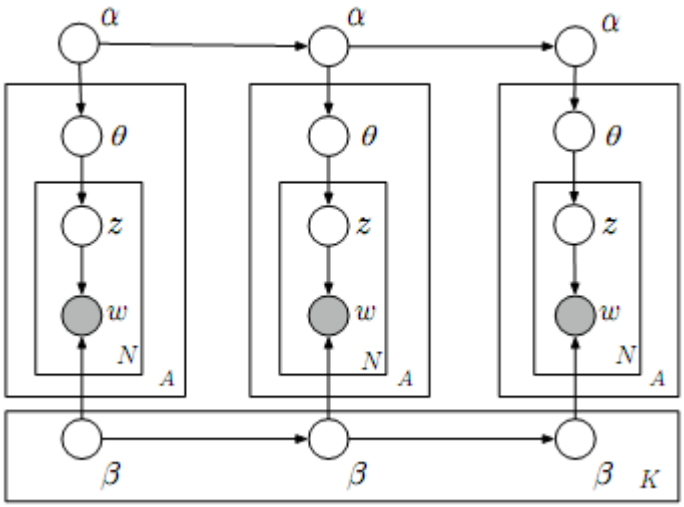
12， Dynamic topic models (2006)

同一年的文章里，Blei 等提出的模型更像时间序列分析的方法，他们认为 topic 会随着时间变化，这种变化满足一阶马尔科夫假设。

图模型如下，可以看到，超参数 α 是随时间变化的，topic 的词分布参数 β 也是随时间变化的，由于多项式分布是离散了，为了把它变成连续的，本文才用了类似与 CTM 模型中的方法，把多项式分布的参数变成自

然参数，然后α β在给定前一个时间片时服从高斯分布的。生成过程为：

1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
   (b) For each word:
      i. Draw $Z \sim Mult(\pi(\eta))$.
      ii. Draw $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$.



由于引入了时间信息，本文采用的是变分 kalman 滤波算法(variational kalman filtering)和更准确的变分小波回归(variational wavelet regression)。

13， Multiscale Topic Tomography (2007)

Nallapati(2007)认为，DTM 使用的是 Logistic Normal 分布，这个分布与多项式分布不是共轭的，这会给推理带来很大的困难。

本文作者提出了 Multiscale Topic Tomography 模型(MTTM)，使用一个 Poison 过程来建模词出现的次数。它把文章按时间顺序划分成 epoch，每个 epoch 的词出现的次数由参数 μ 决定。

1. For each epoch $t = 0, \cdots, 2^S - 1$
2. For each topic $k = 1, \cdots, K$
3. Generate $\theta_{tk} \sim \text{Dir}(\cdot \mid \alpha)$
4. For each document $d = 1, \cdots, M$
5. For each word $w = 1, \cdots, V$
6. Generate $n_{tdw} \sim \text{Poiss}(\cdot \mid \sum_k \theta_{tkd} \mu_{tkw})$

参数 μ 的生成过程如下：

1. For each topic $k = 0, \cdots, K$
2. For each word $w = 1, \cdots, V$
3. Generate $\mu_{0kw}^{(0)} \sim \text{Gamma}(\cdot|\lambda_\mu, \delta_\mu)$
4. For each scale $s = 0, \cdots, S-1$
5. For each epoch $t = 0, \cdots, 2^s - 1$
6. Generate $\beta_{tkw}^{(s)} \sim \text{Beta}(\cdot|\delta_\beta, \delta_\beta)$

14， Hierarchical topic models (2004)

前面的 Topic Models 很少考虑 Topic 之间的关系，忽略这种将导致它们不能发现大量的细粒度的(fine-grained)，紧密相关的(tightly-coherent)的 Topic。

最早的层次模型是 Thomas Hofmann (1999b)，它是一个隐类(Latent Class) 模型。它的主题模型类似 PLSA，它的主题关系(结构)是通过模型选择的方法从数据学习出来的。

另外 CTM 模型也能发现 Topic 之间的关系，但是这种关系是两两之间的关系。

Blei(2004)把它用 LDA 做了扩展，使它成为完整的产生式模型。为了给关系(树结构)提供先验分布，作者使用了 Nested Chinese Restaurant Processes(CRP)——一种对 CRP 的扩展来作为树的先验。这样得到了 Hierarchical LDA(hLDA)模型。

Chinese Restaurant Processes(CRP)是一个整数划分的分布，它是通过模拟 M 个顾客来一个有无限个餐桌的中国餐馆吃饭来实现的。

每个顾客可以坐一个有人的桌子，也可能坐一个新的没人的桌子。他坐有人的桌子时，这个桌子人越多，他坐下的概率就越大，另外还有一定的概率坐新的桌子。

这其实是符合自然界的规律的：物以类聚，越聚越多；但也不能太多，合久必分。

$$p(\text{occupied table } i \mid \text{previous customers}) = \frac{m_i}{\gamma + m - 1}$$
$$p(\text{next unoccupied table} \mid \text{previous customers}) = \frac{\gamma}{\gamma + m - 1}$$

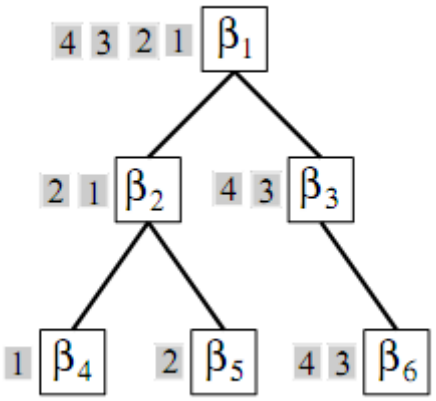
M 个顾客的每一种坐法就代表了对 M 个数据的一个划分，而每种坐法的概率不是显式的，而是通过上面的公式隐式的表示。这样不同参数 γ 就代表了不同的分布，一个 CRP 就是一个对整数划分的概率分布。CRP 和 Dirichlet Process(DP)的样本有相同的形式，它是 DP 的一种表示。后面 HDP 里会涉及到。由于它的聚类性质以及是划分的分布，所以它经常用来作为聚类(分量，Component)数未知的混合模型的先验(比如 Dirichlet Process Mixture Models)。

但是现在我们的混合模型是层次化的(Hierarchical)，比如说是一棵树。所以这个过程并能提供我们需要的先验。

Blei 等人把 CRP 扩展到了层次化的，提出了 Nested CRP。Nested

CRP 可以用这样的场景来定义：中国餐馆还是和前面的一样，有无数个桌子；某个城市又有无数家这样的中国餐馆。其中一家餐馆是 root，它的每个桌子是其它餐馆的名字(索引)，其它餐馆的每一个桌子也是另外一些餐馆的名字，这个结构是无穷递归的。一个顾客在这个城市里要吃 L 次饭，第一次只能在 root 吃，他在 root 根据 CRP 中的那个公式选一个桌子(另一个餐馆的名字)，然后第二次去这个餐馆吃饭，同样的方法选一个桌子，…。

这样每一次 Nested CRP 的模拟过程都得到如下的一棵树，因此 Nested CRP 给每一棵树都定义了一个概率值，这样 Nested CRP 就可以作为树这个结构的先验分布。



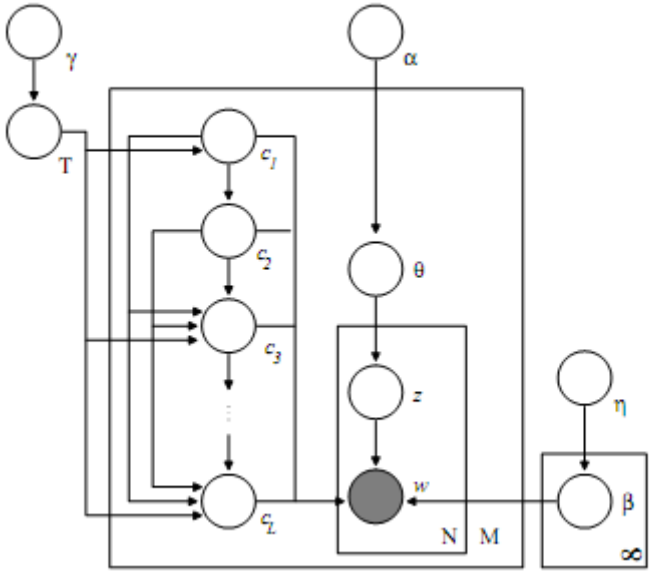
生成一篇文档的过程如下，参考上下两个图：

1，对应每一篇文档根据 Dirichlet 先验生成 θ，比如是 0.2,0.3,0.5

2，对于每一词根据 Nested CRP 生成一条从 root 叶子的一条路径，比如 1->2->4

3，根据 θ 选择一个 topic，比如是 4

4，然后根据选择的 topic 生成词。

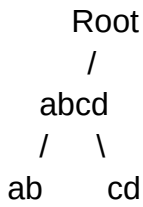


与 LDA 不同的是，主题是无穷多个，根据数据做最大似然估计可以

求出最合适的主题个数；LDA 中的 θ 是一篇文档关于所有 Topic 的分布，而这里的 θ 是关于主题关系的比例，比如上面的例子 θ 用来确定是从根产生词的概率是 0.2，从第二层 topic 的概率是 0.3，从叶子 topic 产生词的概率是.5。为每一个词选一个主题的任务由 θ 转交给了 Nested CRP 了，Nested CRP 选一条路径其实就隐含了为词选主题的过程。

考虑一个极端的情况：树只有两层，而且 θ 不是变量，而是常量[0,1]，也就是说根主题不产生任何词，只有叶子节点产生词。那么这个 Nested CRP 就退化成了 CRP，这样就是用一个 CRP 来作为主题（数目是无穷）分布的先验。其实就是把 LDA 的推广到 Nonparametric 去了。和后面的 HDP 很类似。

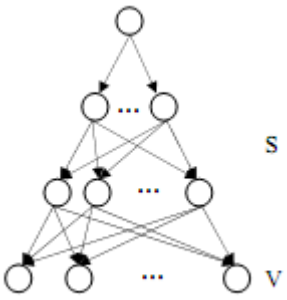
这个模型为什么能够产生层次化的 Topic 呢？假设有 a,b,c,d 4 个 topic，它们都属于一个 topic 的子 topic，然后 a，b 比较接近，c，d 比较接近。根据 Nested CRP 的过程，在第二层的 topic 中，因为 abcd 比较接近，它们的词经常共现，所以代表它们的一些词会聚成一个类。每层的 topic 都有一定概率来生成词，所以在第三层中，ab 共现的比 ac，ad 更频繁，所以它们又聚成了一类。可以看出 θ 的作用，它表示一篇文章抽象的 topic(如 abcd)和具体的 topic(如 ab)的比例。我觉得其实从 θ 可以发掘一篇文章到底是在类似于综述的泛泛而谈还是具体的研究某个细节。



15， Pachinko Allocation Model (2006)

McCallum(2006)提出了另一种层次化建模的方法。他们的想法很简单，topic 不仅仅是词的分布，也可以是其它 topic 的分布。这样他们把主题组织成一个有向无环图(DAG)，每个叶子节点是一个词，每个内部节点是一个 topic。如果 topic 的所以孩子都是词的话，这就是 LDA 里的 topic，否则 topic 的孩子也是 topic(孩子既有词又有 topic？好像没什么意义)。

如果一个 topic 有 K 个孩子，那么到底取那个孩子的值应该是一个 K 维的多项式分布，每次都不一样，所以还有给多项式分布的随机变量再定义一个分布。当然可以定义很多种分布，但是最简单的就是 Dirichlet 分布了。孩子全是叶子节点的 topic 和 LDA 中的 topic 是一样的，它的这个多项式分布是确定的(因为假设一个确定主题它的词分布是确定的)，所以不用定义 Dirichlet 分布了。当然也可以看出 Dirichlet 分布的特例——它只在 Simplex 的某个点上概率为 1，其它的点概率是 0.



(c) Four-Level PAM

为了生成一个词，首先从根节点根据 Dirichlet 分布产生一个子节点，然后再根据上一步选择的节点的 Dirichlet 分布产生下一个子节点，直到某个节点是词的 Topic，然后根据这个 Topic 的词分布产生一个词。作者认为这个过程很像日本的一种游戏 Pachinko Machines——把一个金属球从上滚下，中间会遇到很多 pin，谈来谈去最后到底部的一个游戏，所以把这个模型叫做 Pachinko Allocation Model(PAM)。

下图是 LDA 和 4-level PAM 的图模型。

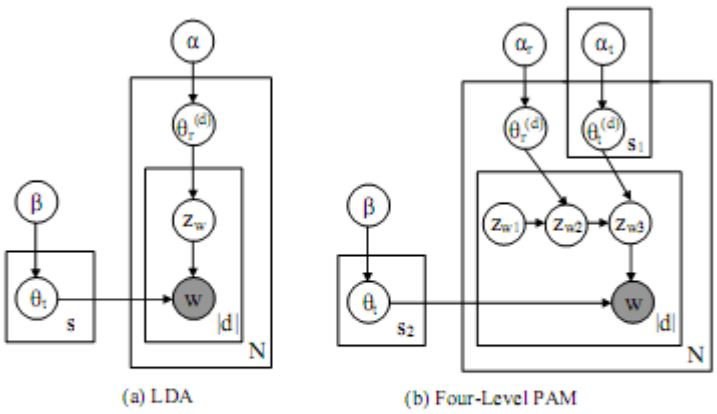


Figure 2. Graphical models for (a) LDA and (b) four-level PAM

PAM 中主题的词分布和 LDA 一样，都是 Corpus 级的，对于一个 Corpus 只生成一次。第二层的 topic 的分布是 Document 级的，每个文档生成一次，意思是每篇文档都有各自的顶级 topic 分布，而第三层的 topic 分布是 Corpus 级的，因为每一个第二层的 topic 是第三层 topic 的混合，而且混合的比例是不变的（混合比例一变，对应的词分布也变化，那就不是一个 topic 了）。

当然 PAM 可以不只是 4 层的，可以扩展到任意层，也可以不是树的结构，只要是任意 DAG 就行。也可以不事先知道模型的结构，通过模型选择从数据中学习出结构来。

Andrew McCallum(2007)又把 hLDA 和 PAM Combine 在了一起提出了新的 hPAM 模型。

16， Hierarchical Dirichlet processes (2006)

LDA 还有一个比较大的问题就是主题数目如何确定的问题。从分支度的实验来看，主题数目不同，分支度差别是很大的。这其实是一个模型选择的问题。目前主流的方法是使用贝叶斯非参数(Bayesian Nonparametrics)的模型来解决。贝叶斯非参数化的方法最近几年在 NIPS，UAI 和 ICML 等会议上开始流行起来。

参数化的方法是预先选择一个模型，然后根据数据使用最大似然估计或者最大似然估计或者贝叶斯估计选择一个最合适的参数。但是模型的选择是一个棘手的问题，如果选得不好，即使参数估计完全准确也无法很好的是模型适应(fit)数据。因此在传统的学习问题中模型选择是个很麻烦的事情，解决的方法从最简单的根据人的经验，到使用交叉验证(Cross Validate)不断尝试，到 AIC，BIC，MDL 等模型选择方法。这些方法都是近似的估计风险的一个上界，估计不是很准确，而且计算量一般也很大。

贝叶斯非参数化的方法并不提供一个模型(一族分布)，而是服从一般的分布，然后给这个分布再定义一个分布。下面用一个例子来说明贝叶斯非参数化方法。

假设给定一些数据，需要我们进行聚类。参数化的方法是怎么做的呢？首先根据经验，觉得这些数据看起来符合某种分布，比如高斯分布。

假设 X～N(μ，1)，频率学派就对数据做最大似然估计，估计似然度最大的μ；而贝叶斯学派认为μ本身也是一个随机变量，然后给它也定义一个分布，比如μ～N(0，a)。这样就可以把人的先验知识加进去。于是根据数据求出μ的后验分布为：

μ|X1，X2,…,Xn～N(b，1/(n+1))

其中$b=\frac{n}{n+1}(\frac{1}{n}\sum_{i=1}^{n}X_i)+\frac{1}{n+1}a$

这就是贝叶斯估计。当然μ是一个随机变量，为了求出一个值来代表它，自然的会选择概率最大的值，这个值就是最大后验概率估计(MAP)。

上面的方法是问题是万一数据不是高斯分布的呢？简单的解决方法是是用高斯混合模型(Mixtures of Gaussians)，从理论上讲，高斯混合模型可以逼近任何分布。但是又有一个新问题——怎么确定高斯分量的个数？

非参数化的方法为什么能解决这个问题呢？它的思想其实很简单，如下面的对比图，假设 X 服从某个分布 P，采用贝叶斯学派的思想，认为分布 P 仍然是一个随机变量，然后给概率测度(分布)P 在定义一个先验的分布，然后根据观察 X，求出 P 的后验分布。

因为把非参数化和贝叶斯结合了起来，所以叫做非参数化贝叶斯方法。

Parametric model:	Nonparametric model:
$X \sim N(\mu, 1)$	$X \sim P$
$\mu \sim N(a, 1)$	$P \sim Q_0$
$\mu \mid X_1, \ldots, X_n \sim N(b, 1/(n+1))$	$P \mid X_1, \ldots, X_n \sim Q_n$

因此问题的关键是怎么给一个概率测度(分布)定义一个分布？

参数化的方法其实也会给概率测度定义分布。比如：

X 服从 Bernoulli(p)，X 服从 Bernoulli 分布，参数是 p。

怎么给 p 定义分布呢？因为 p 在 0，1 之间，最简单的分布就是 Beta 分布了。Beta 分布的另一个好处是它与 Bernoulli 分布是共轭的，也就是说，X 的后验分布也是 Beta 分布的，这样计算上会带来很多方便。

推广到 X 服从 Multinomial(p1,…,pk)，其中 pi>=0,p1+…+pk=1

给 p1,…,pk 定义分布，最简单的就是 Dirichlet 分布了。而且它是与多项式分布共轭的分布。

现在的问题是怎么给概率测度 P 定义一个分布，而且这个分布最好也类似于 Dirichlet 分布。也就是怎么把 Dirichlet 分布推广的无穷维。

满足这样要求的就是 Dirichlet Process(DP)。DP 的定义是：

P 服从 DP(αH)，如果它：对于空间 X 的任意划分(Partition)(A1,…,Ar)都满足（P(A1),…,P(Ar)）服从 Dirichlet(αH(A1),…, αH(Ar))。

从定义上看，P 是概率分布(后面可以证明是离散的)的分布，也就是说它是一个随机概率测度，它把一个集合(事件)映射到一个随机变量。

我们关注的是既然 X 服从 P，P 又是 DP(αH)，那么我们能不能不生成 P 而直接得到样本 X 呢？答案是肯定的，可以有两种方法——Chinese Restaurant Process(CRP)和 Polya Urn Scheme。

CRP 前面讲过了，这里不重复了，它的一个重要性质就是聚类特性(rich get richer)，Polya Urn Scheme 只是用来证明后面的结论，所以也跳过。

DP 的一个重要性质：它是离散概率测度的分布。因为是离散的，所以它可以作为无穷但可数个的高斯分布的先验。

Ferguson 在 1973 年使用 DeFinetti's representation theorem 通过 Polya Urn 证明了 DP 是以概率 1 离散的。Sethuraman 在 1994 年显式的构造出了这个过程，也就是 Stick-breaking construction。

$$\pi_k = \pi'_k \prod_{l=1}^{k-1} (1 - \pi'_l) \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k},$$

其中：

$$\pi'_k \mid \alpha_0, G_0 \sim \mathrm{Beta}(1, \alpha_0) \qquad \phi_k \mid \alpha_0, G_0 \sim G_0 .$$

证明过程不管，重要的是一个结论：DP 是离散的，因此可以作为 Infinite Mixture Models 的无穷个 Component 的先验。

这样得到的模型就叫 Dirichlet Process Mixture Model。

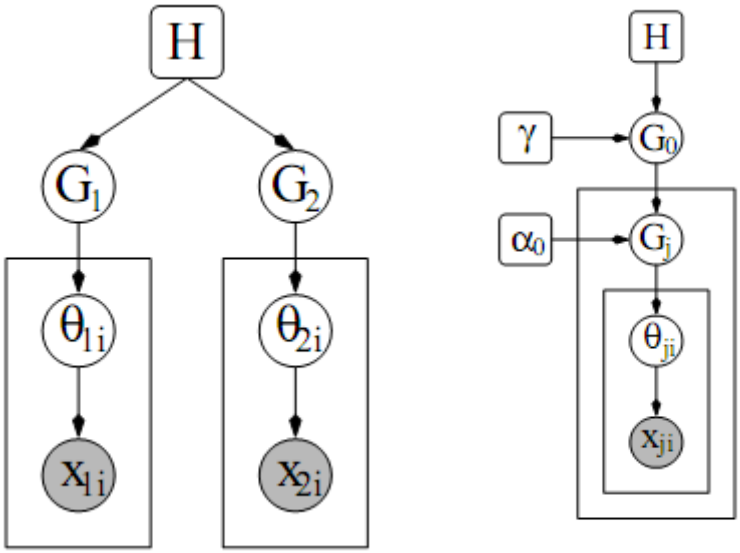
当然除了 Dirichlet Processes 外还要很多其它过程，如 Pitman-Yor Processes，Indian Buffet Processes ，Polya Tree Processes，Tail-free Processes 等等，可能会有不同的性质。

也有很多的应用，比如在NLP领域的HDP-HMM，HDP-PCFG等等，在图像等其它很多领域也有应用，下面的网站上列举了常见的应用和参考文献 http://npbayes.wikidot.com/references 。

上面的例子来自 Stefan Harmeling 的 Bayesian Nonparametrics- A brief introduction to Dirichlet processes，其它参考文献包括 Percy Liang (2007)在 ACL 上的 tutorial，Yee Whye The(2007)的 talk。

现在回到 Topic Models，怎么确定 LDA 主题的数目。

根据前面的结论，我们可以直接给每一篇文档使用一个 DP Mixture，如下图所示的 DP Mixture Models，



但由于基测度 H 是连续的，不同文档之间的 topic 不能共享(share)。为了使 H 是离散的，Y. W. The(2006)提出了给 H 再加一个先验的模型，这就是 Hierarchical DP(HDP)。

HDP 也有类似与 DP 的 CRP 过程——Chinese Restaurant Franchise，类似的 Stick-breaking Construction。

推理可以是最早文章里提出的 Gibbs Sampling，也可以使用变分法等等。

除了把 LDA 推广到 Bayesian Nonparametric 外，也可以把其它模型如 PAM 推广的非参数化的 PAM。

Topic Models 还有很多应用，具体到某个领域也会有很多变化。

ICML2008 和 UAI2008 上也有不少 Topic Models 的文章，还没来得及看。

参考文献

Hofmann, T. (1999). Probabilistic latent semantic indexing. Proceedings of the 22nd. Annual International SIGIR Conference (M. Hearst, F. Gey and R. Tong, eds.). New York: ACM Press, 50--57.

Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum

D. Blei, A. Ng, and M. Jordan (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993-1022

张连文 (2006). 贝叶斯网引论. 科学出版社.

C. Andrieu, N. de Freitas, A. Doucet and M. I. Jordan (2003a). An introduction to MCMC for machine learning. Machine Learning, 50, 5-43

M. J. Wainwright and M. I. Jordan (2003b).Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), Learning in Graphical Models, Cambridge: MIT Press.

Christopher M. Bishop (2006). Pattern Recognition and Machine Learning, Springer

Gregor Heinrich. Parameter estimation for text analysis. http://www.arbylon.net/publications/text-est.pdf

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101, 5228-5235.

Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), Handbook of Latent Semantic Analysis. Hillsdale, NJ: Erlbaum.

Blei, D. M. and Lafferty, J. D. (2006). Correlated topic models. In Weiss, Y., Sch?lkopf, B., and Platt, J., editors, Advances in Neural Information Processing Systems 18. MIT Press, Cambridge, MA.

D. Blei and J. Lafferty. A correlated topic model of Science. Annals of Applied Statistics. 1:1 17–35.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic Author-Topic Models for Information Discovery. The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington.

Rosen-Zvi, M., Griffiths T., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. In 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada

M. Rosen-Zvi, T. Griffiths, P. Smyth, M. Steyvers . Learning author-topic models from text corpora.

Yang Song, Jian Huang, Isaac G. Councill, Jia Li, C. Lee Giles. (2007) Efficient topic-based unsupervised name disambiguation. ACM/IEEE Joint Conference on Digital Libraries (JCDL 2007): 342-351, 2007.

Andrew McCallum, Andres Corrada-Emmanuel, Xuerui Wang. (2004) The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email.Technical Report UM-CS-2004-096, 2004.

Andrew McCallum, Andres Corrada-Emmanuel and Xuerui Wang. (2005) Topic and Role Discovery in Social Networks. IJCAI, 2005.

Andrew McCallum, Xuerui Wang and Andres Corrada-Emmanuel. (2007) Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. Journal of Artificial Intelligence Research (JAIR), 2007.

Griffiths, T.L., & Steyvers, M., Blei, D.M., & Tenenbaum, J.B. (2004). Integrating Topics and Syntax. In: Advances in Neural Information Processing Systems, 17.

Hanna Wallach (2006) Topic Modeling: Beyond Bag-of-Words. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, U.S., 2006.

Xuerui Wang and Andrew McCallum. (2005) A Note on Topical N-grams. University of Massachusetts Technical Report UM-CS-2005-071, 2005.

Xuerui Wang, Andrew McCallum and Xing Wei (2007) Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval. Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), 2007.

Amit Gruber, Michal Rosen-Zvi and Yair Weiss (2007) Hidden Topic Markov Model. Artificial Intelligence and Statistics (AISTATS), San Juan, Puerto Rico, March 2007.

Dietz, Laura; Bickel, Steffen;Scheffer Tobias (2007) Unsupervised Prediction of Citation Influences. In: Proceedings of the 24th International Conference on Machine Learning. Corvallis, Oregon, USA, June 2007

Nallapati, R. and Cohen W. (2007) Link-PLSA-LDA: A new unsupervised model for topics and influence in blogs. International Conference for Weblogs and Social Media, 2008

Nallapati, R., Ahmed, A., Xing, E.P., Cohen, W. (2008) Joint Topic Models for Text and Citations. submitted to KDD 2008

David Cohn and Thomas Hofmann (2001). The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity, in T. Leen et al., eds., Advances in Neural Information Processing Systems 13.

Elena Erosheva, Steve Fienberg, and John Lafferty (2004) Mixed membership models of scientific publications. Proceedings of the National Academy of Sciences, Vol. 101, Suppl. 1, April 6, 2004

Xuerui Wang and Andrew McCallum (2006). Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. Conference on Knowledge Discovery and Data Mining (KDD) 2006.

D. Blei and J. Lafferty. (2006) Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, 2006.

Nallapati, R., Cohen, W., Ditmore, S., Lafferty, J. (2007) Multiscale Topic Tomography, Ung, K., ACM-KDD, 2007.

D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. (2004). Hierarchical topic models and the nested Chinese restaurant process. In Neural Information Processing Systems (NIPS) 16, 2003

Thomas Hofmann (1999b). The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data. Proceedings of the International Joint Conference in Artificial Intelligence, 1999

Wei Li, and Andrew McCallum. (2006). Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations. ICML, 2006.

David Mimno, Wei Li and Andrew McCallum. (2007). Mixtures of Hierarchical Topics with Pachinko Allocation. ICML, 2007

Y. W. Teh, M. I. Jordan, M. J. Beal and D. M. Blei. (2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101, 1566-1581, 2006.

Stefan Harmeling (2006). Bayesian Nonparametrics-A brief introduction to Dirichlet processes. IDA retreat, May 2006

Yee Whye Teh. (2007) Dirichlet Processes and Hierarchical Dirichlet Processes. Talks of Machine Learning @ CUED http://talks.cam.ac.uk/talk/index/6582

Percy Liang, Dan Klein. (2007). Structured Bayesian nonparametric models with variational inference (tutorial). Association for Computational Linguistics (ACL), 2007