数据挖掘导论
Introduction to Data Mining

第五章 文本挖掘

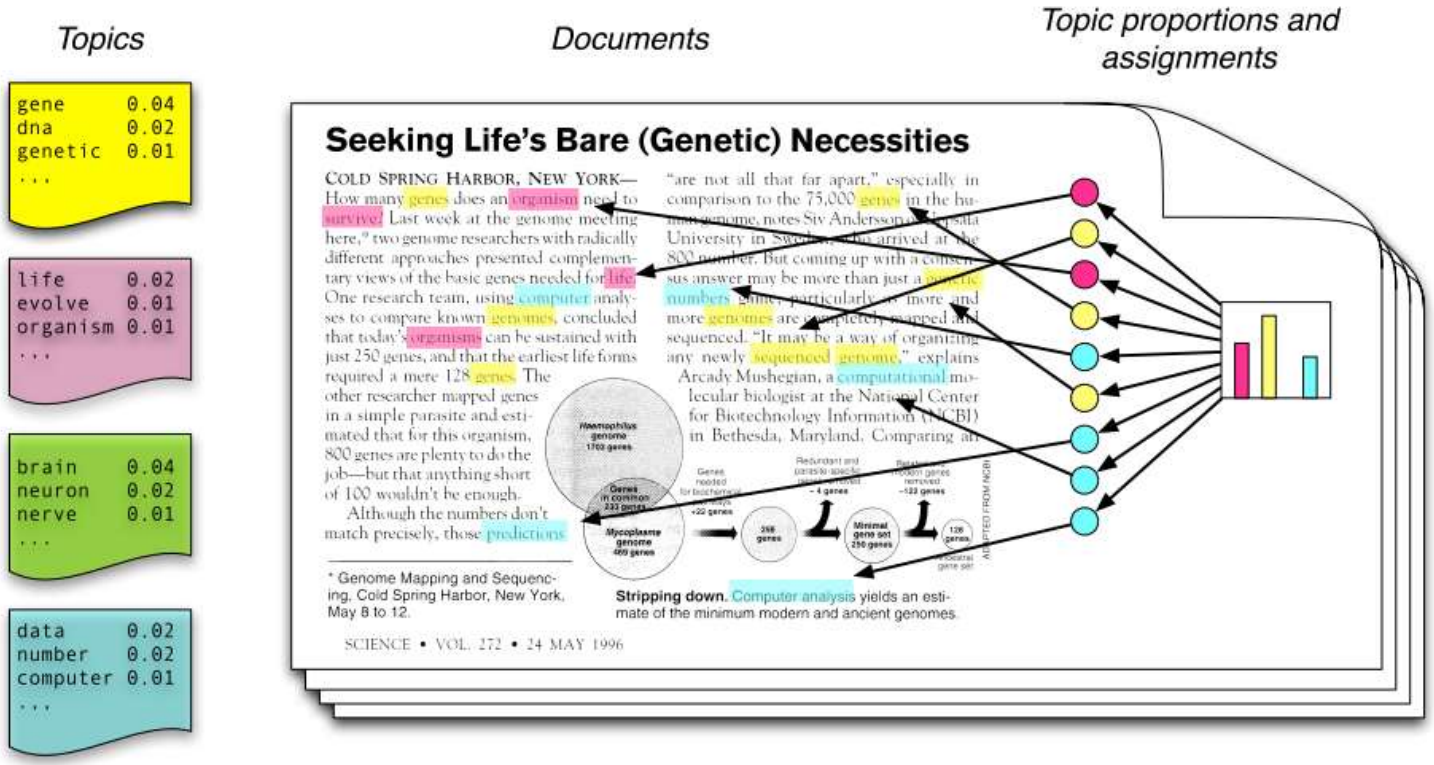
刘 淇
Email: qiliuql@ustc.edu.cn
课程主页：
http://staff.ustc.edu.cn/~qiliuql/DM2017YZ.html

目录

- 传统的文本处理
  - 自然语言处理
  - N-Gram语言模型
- 文本挖掘
  - 文本挖掘简介
  - 特征抽取
  - 特征选择
  - 文本分类
  - 文本聚类
  - 文本挖掘的常用方法
    - 主题模型
    - word2vec

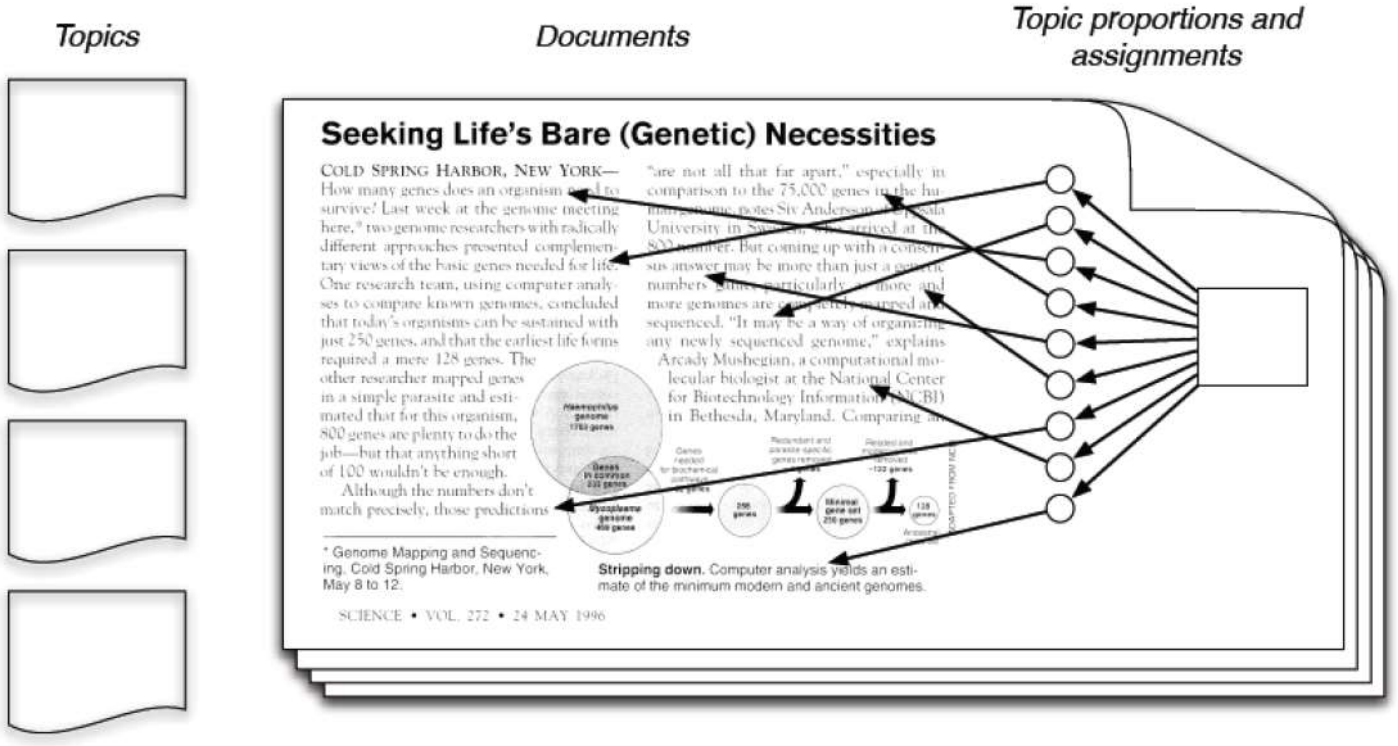
主题模型

- 文档一般涉及多个主题
- 每个主题又是一个关于单词的概率分布



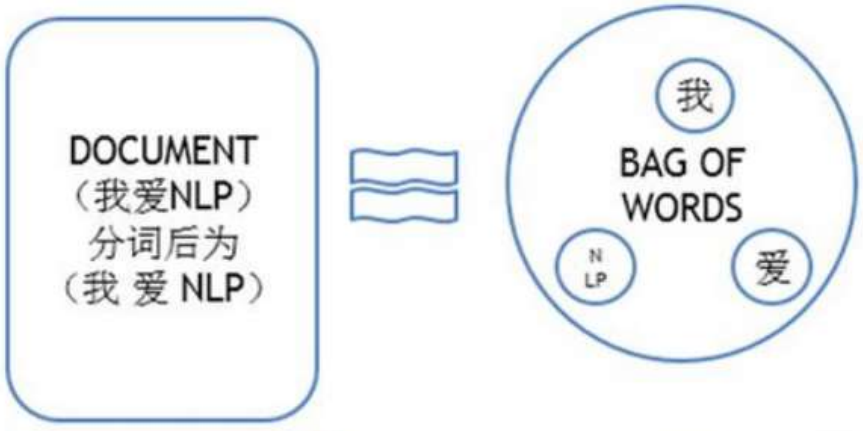
主题模型

- 现实：只观察到一些文档，其他信息都是未知的
- 目标：基于文档集，推测主题分布和单词分布



词袋模型假设

- 词袋模型是在自然语言处理和信息检索中的一种简单假设。
  - 主题模型是一类词袋模型
  - 文档是无序的，不考虑出现的先后顺序
  - 文档中的单词也是无序的，不考虑语法限制

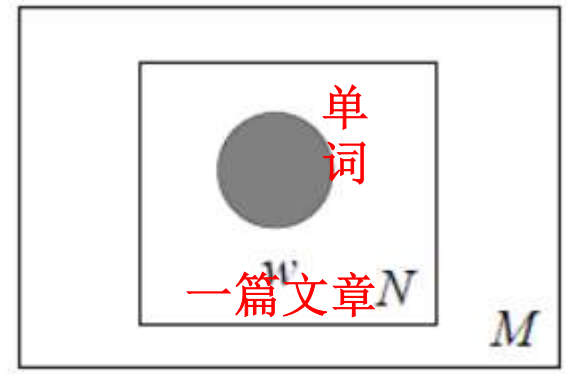


一元模型

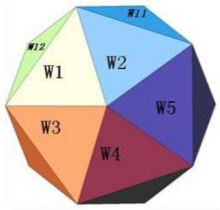
- 一元模型：每个文本的词语都是独立地从一个多项式分布产生

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$

- 简单直观的词频概率模型，没有考虑文本的主题



M篇文章 (a)一元模型



$\vec{p}$

$w \sim \text{Mult}(w|\vec{p})$

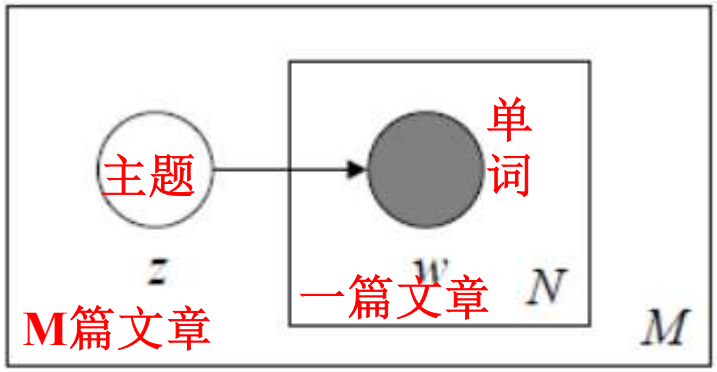
给定文档集，需要求解（学习）的是$p(w_n)$

一元混合模型

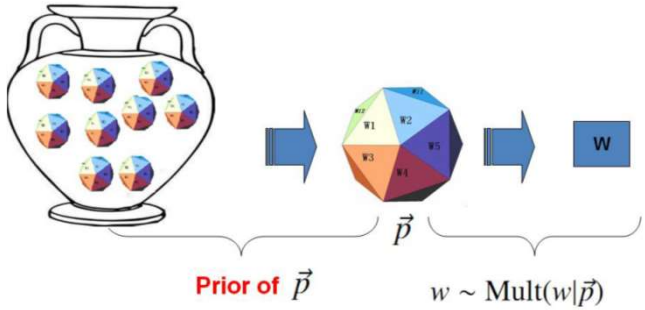
- 如果在一元模型里面加入一个离散的随机主题变量 z，则成为一元混合模型
- 在混合模型里，对于每个文本，首先选择一个主题 z，然后根据条件多项式 p(w|z) 独立地生成该文本的 N 个词语。由此得到每个文本的概率为：

$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z)$$

- 这个模型只允许一篇文档的一次过程里只有一个主题

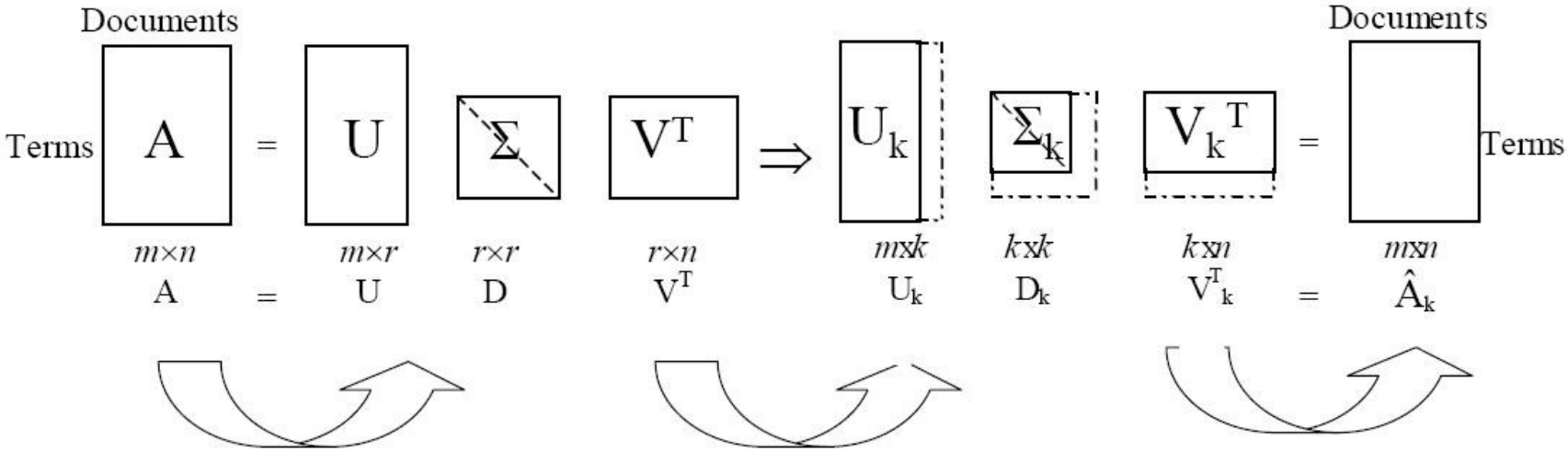


(b)一元混合模型

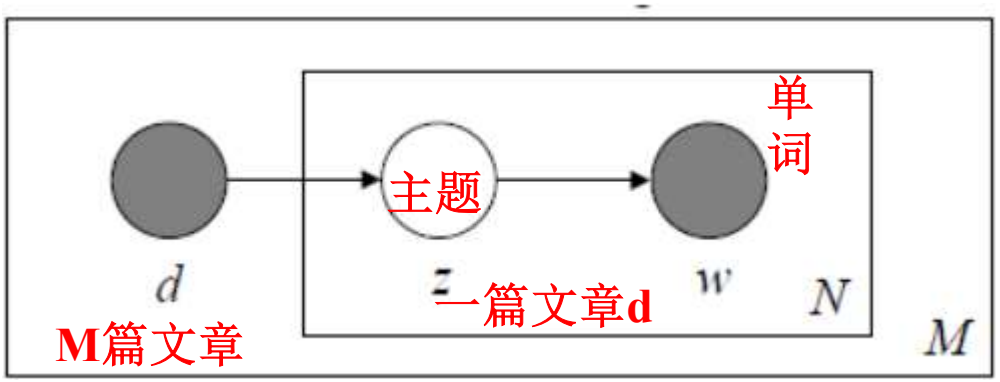


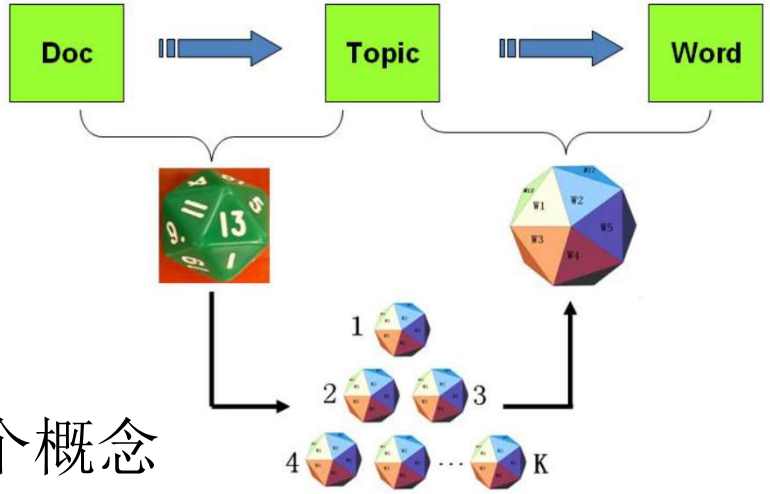
LSI (Latent Semantic Indexing)

- 奇异值分解（SVD）
- 通过维数规约，消除了大部分同义词和多义词的信息冗余，巧妙地把它们映射到中间过渡矩阵的同一个元素当中
  - 每个文档可以视为有r个隐含主题
  - 得到的不是一个概率模型，缺乏统计基础，结果难以直观解释。
    - SVD是个耗时的通用模型，非针对文本类型的数据和生成过程



pLSI模型



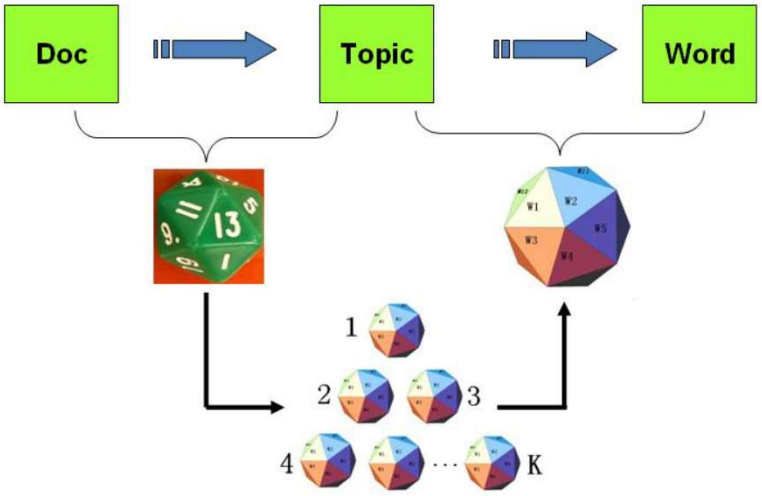


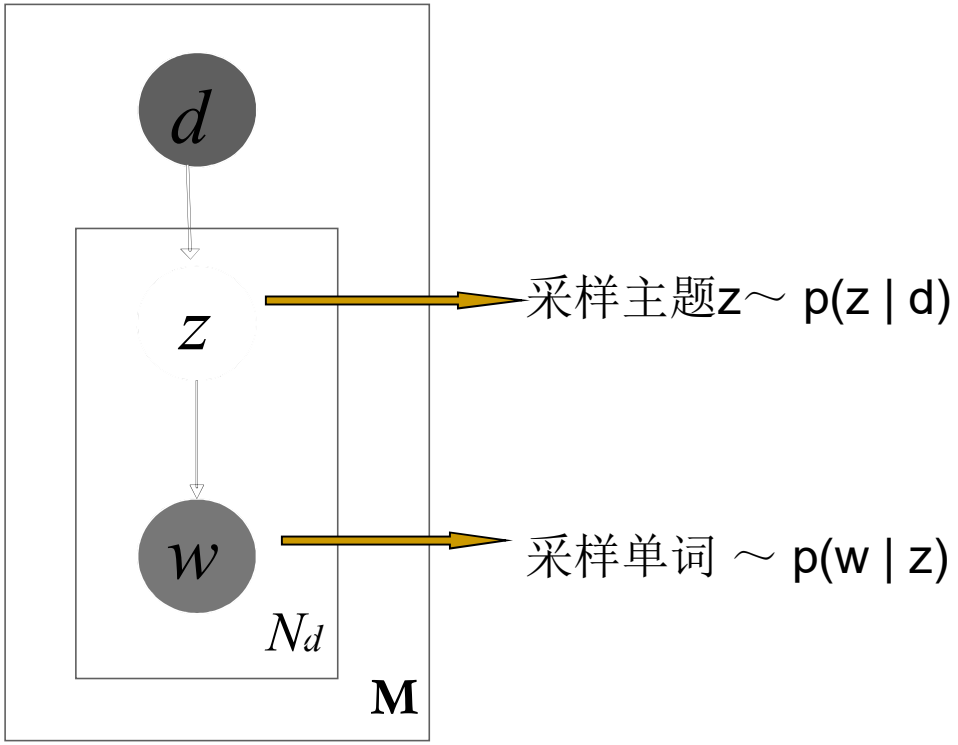
- pLSI 模型引入了“潜在主题”这个概念
- 一篇文档允许同时有多个主题
- pLSI 通过对训练集拟合，给这些潜在变量赋予合理的取值
- 假设文本 d，单词 w，它们与潜在主题 z 的条件概率相互独立，则：

$$p(d,w_n)=p(d)\sum_z p(w_n\mid z)p(z\mid d)$$

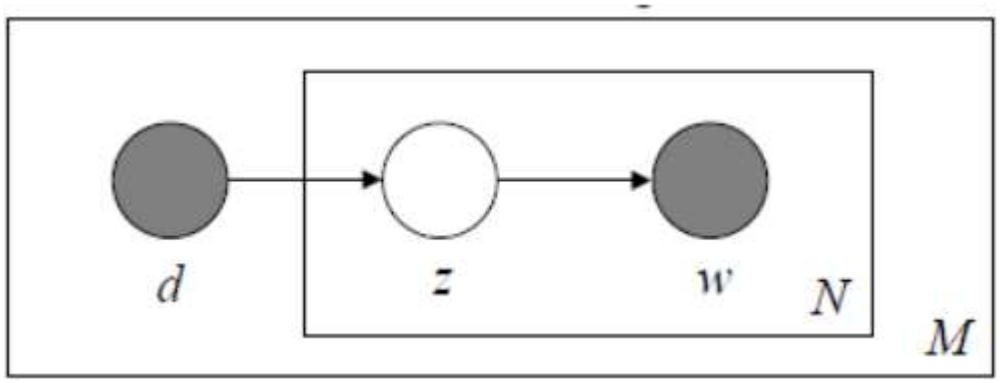
pLSI模型

$$p(d, w_n) = p(d) \sum_z p(w_n \mid z) p(z \mid d)$$



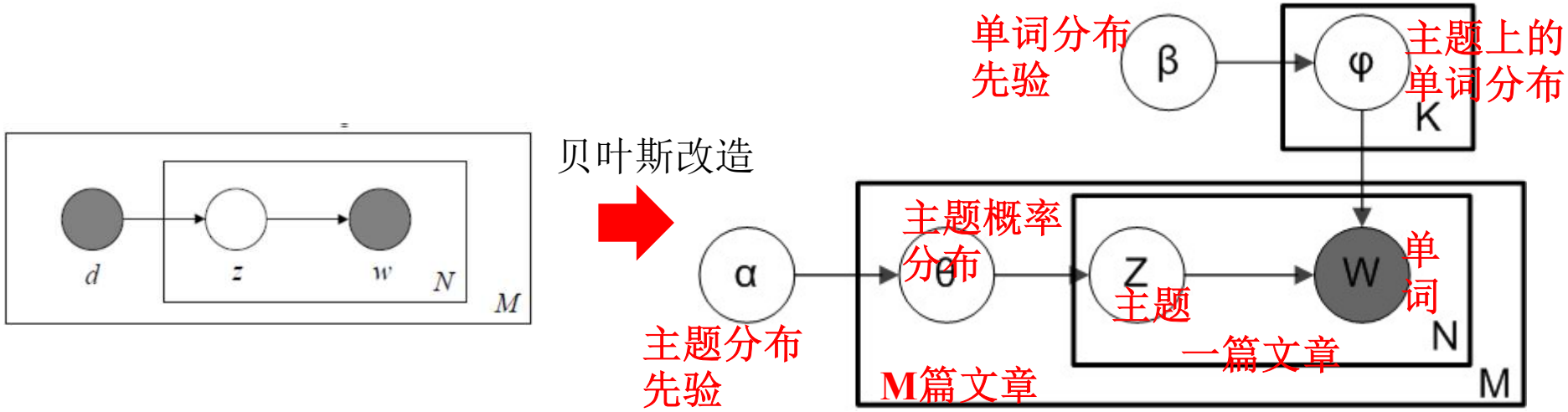


pLSI模型



- pLSI模型存在的问题：
  - pLSI 模型需要得到一个标签（主题）的先验概率，这个概率只建立在已见的训练集的基础上，对于训练集之外的未见文本，并没有一个合适的先验概率来描述；
    - 文档d上的主题z分布
  - 随着训练样本的增加，概率矩阵的大小也在线性地增加，存在着过度拟合问题，即 pLSI 模型当中存在了太多仅仅适用于训练集文本的离散特征，而这些特征无法恰当的描述训练集以外的其他文本。

LDA模型

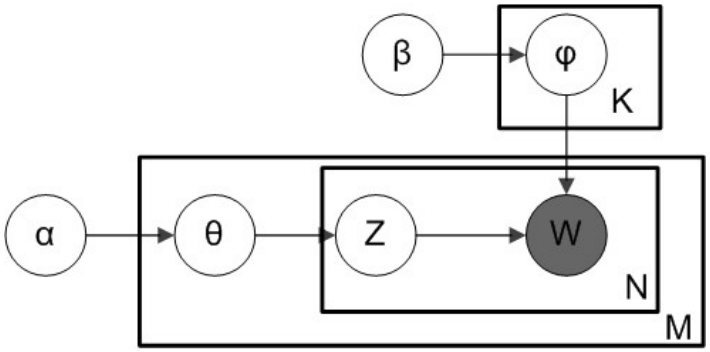


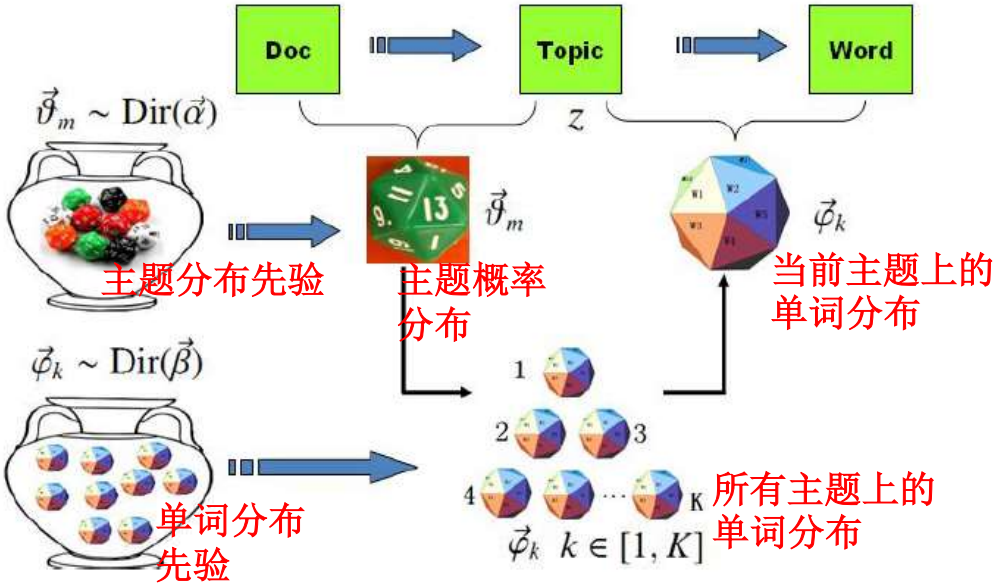
- 主题概率分布和单词概率分布符合多项式分布
  - 1. 从Dirichlet分布β中分别生成K个主题对应的单词概率分布φ
  - 2. 从Dirichlet分布α中分布生成M篇文档的主题概率分布θ
  - 3. 生成一个单词时，每次从主题概率分布θ中得到一个主题Z，再从对应单词概率分布φ中得到单词W，重复N次，得到一个完整的文档

Dirichlet分布是多项式分布的共轭先验分布，采用共轭先验的原因是可以使得先验分布和后验分布的形式相同，这样一方面合符人的直观（它们应该是相同形式的）。另外一方面是可以形成一个先验链，即现在的后验分布可以作为下一次计算的先验分布，如果形式相同，就可以形成一个链条。

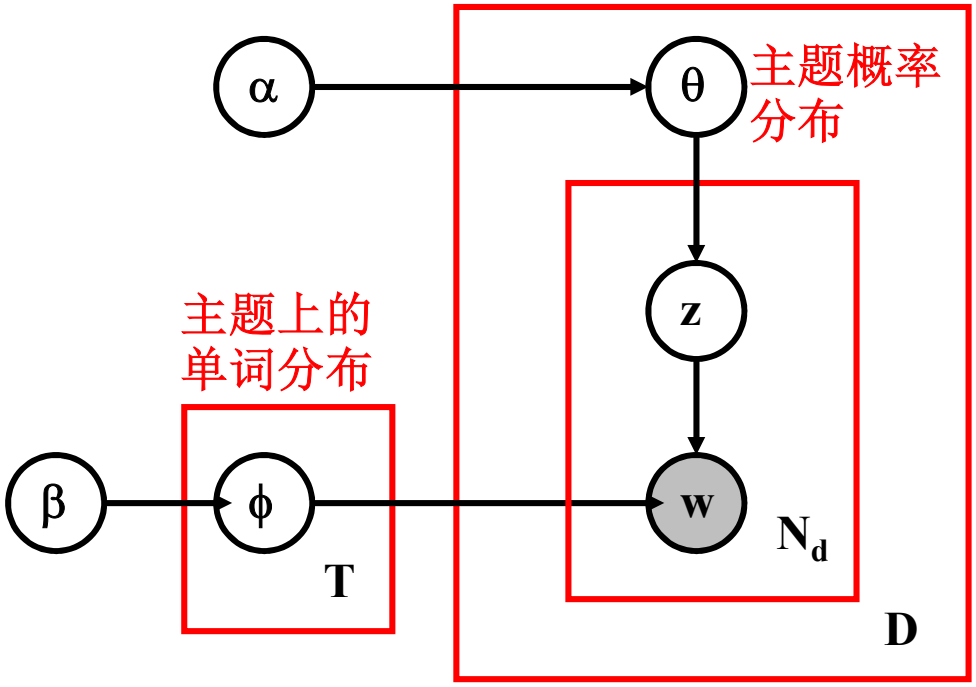
LDA模型

- 每次生成一篇新的文档前，上帝从服从 $\vec{\theta}_m \sim \mathrm{Dir}(\vec{\alpha})$ 的坛子中抽取出一个doc->topic骰子，然后重复以下步骤：
  - i. 投掷这个doc->topic骰子，得到一个topic编号z。
  - ii. 从服从 $\varphi_k \sim \mathrm{Dir}(\vec{\beta})$ 分布的坛子里共K个topic-word骰子中选择编号为z的那个，投掷这枚骰子，于是得到一个词。





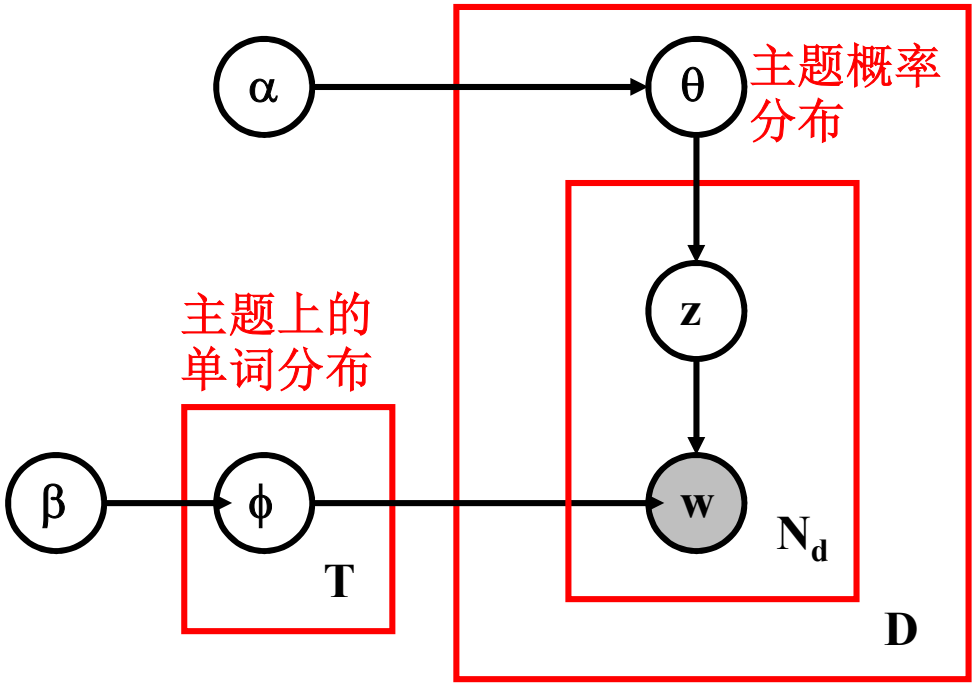
LDA模型



给定先验分布，生成一个主题分布向量θ，以及N个单词（一个文档）：

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

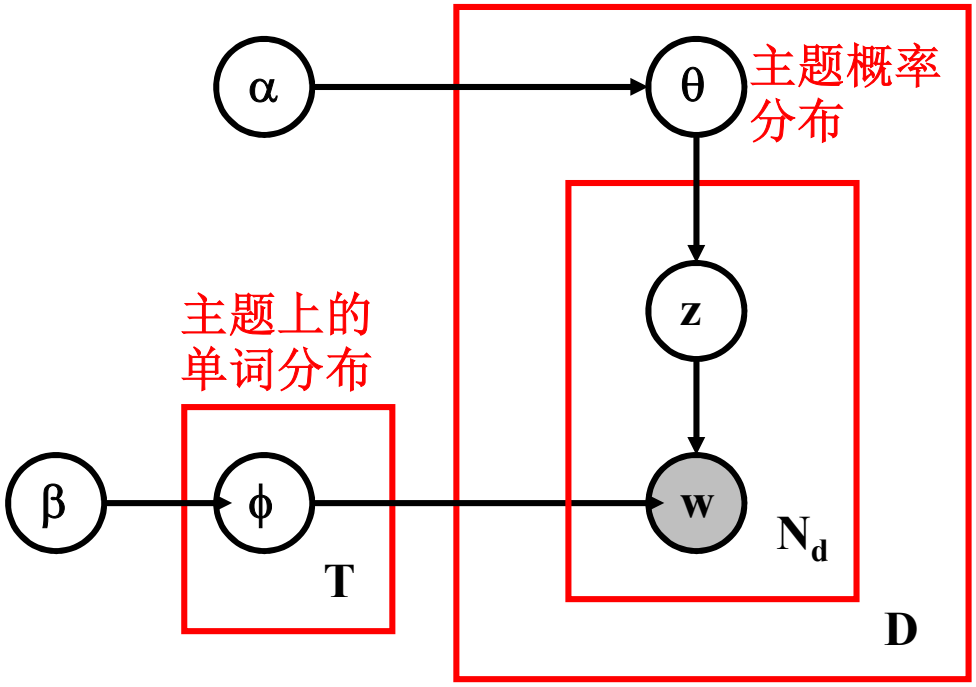
LDA模型



给定先验分布，根据所有主题分布向量θ生成N个单词（一个文档）：

$$p(w|\alpha,\phi)=\int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\phi)\right)d\theta$$

LDA模型

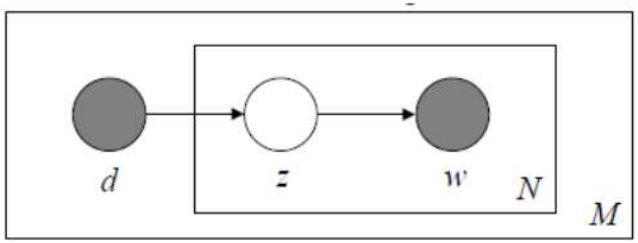


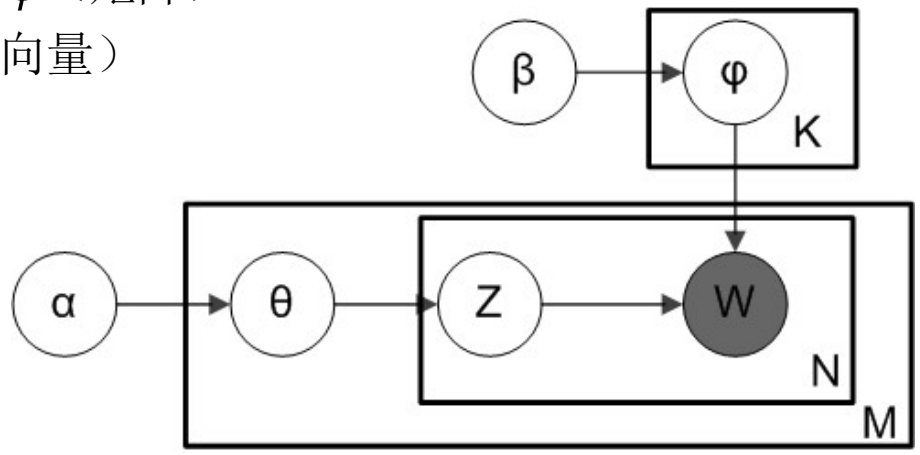
给定先验分布，根据所有主题分布向量θ生成所有文档：

$$p(D|\alpha,\phi)=\prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\phi)\right)d\theta_d$$

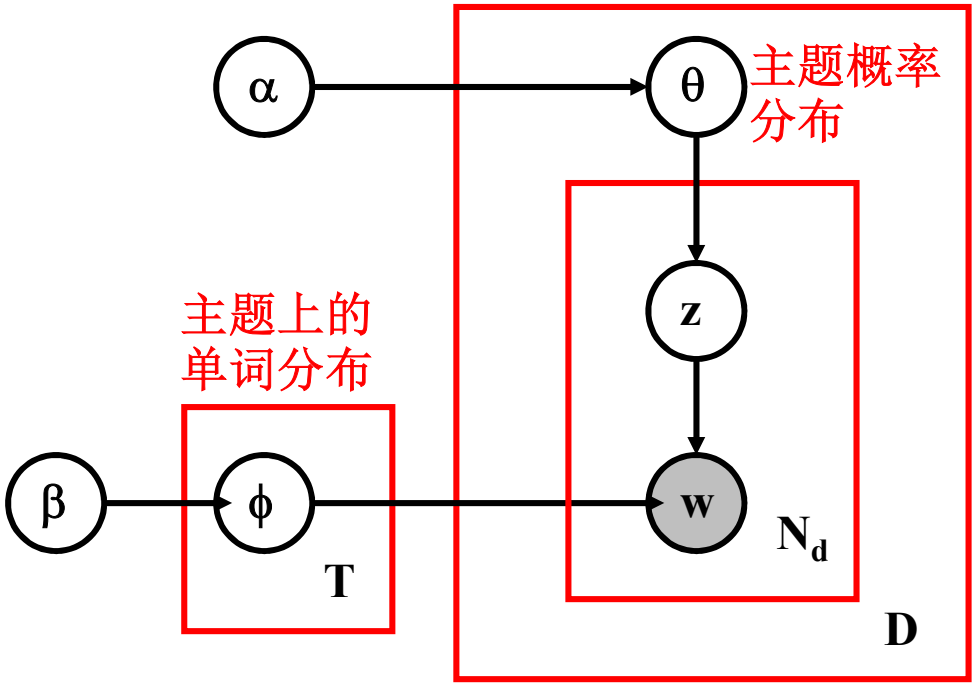
模型求解

- 生成文档(Generation)的反过程------学习模型参数(Inference)
  - 在PLSA中，使用EM算法去估计“主题-词项”矩阵Φ和“文档-主题”矩阵Θ这两个参数，而且这两参数都是个固定的值，只是未知，使用的思想其实就是极大似然估计MLE。
  - 而在LDA中，估计θ、φ这两未知参数可以用变分(Variational inference)-EM算法，也可以用gibbs采样，前者的思想是最大后验估计MAP，后者的思想是贝叶斯估计。
    - K个主题对应的单词概率分布φ（矩阵）
    - M篇文档的主题概率分布θ（向量）





LDA模型



给定先验分布，根据所有主题分布向量θ生成所有文档：

$$p(w|\alpha,\phi)=\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\int\left(\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{V}\left(\theta_i\phi_{ij}\right)^{w_n^j}\right)d\theta$$

LDA模型

- 参数学习过程(文档生成过程的反过程)：Gibbs采样
  - 对于随机变量U_1, U_1,…, U_K,很难从它们的联合分布中抽取一个样本
  - 容易从条件分布Pr(U_j | U_1, U_2,…, U_{j-1}, U_{j+1},…,U_K), j=1,2,…,K 模拟
  - Gibbs Sampling从每个分布轮流地选择一个来模拟，并且当该过程稳定时，提供联合分布的一个样本

抽象算法：

1.取初值 $U_k^{(0)}$ ，k=1,2,…,K

2.对于t=1,2,…重复：

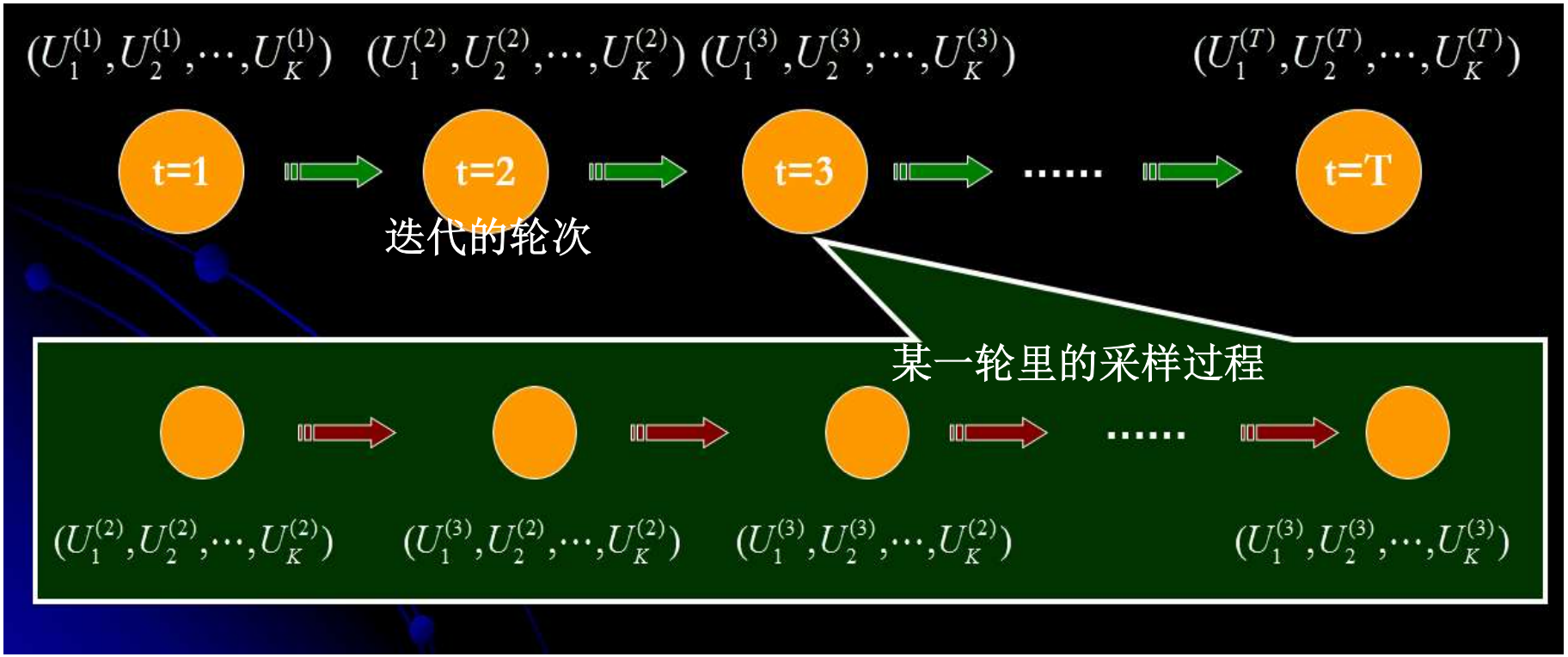
对于k=1,2,…,K，从下式产生 $U_k^{(t)}$：

$$\Pr(U_k^{(t)} \mid U_1^{(t)},\cdots,U_{k-1}^{(t)},U_{k+1}^{(t-1)},\cdots,U_K^{(t-1)})$$

3.继续步骤2，直到 $(U_1^{(t)},U_2^{(t)},\cdots,U_K^{(t)})$ 的联合分布不再改变

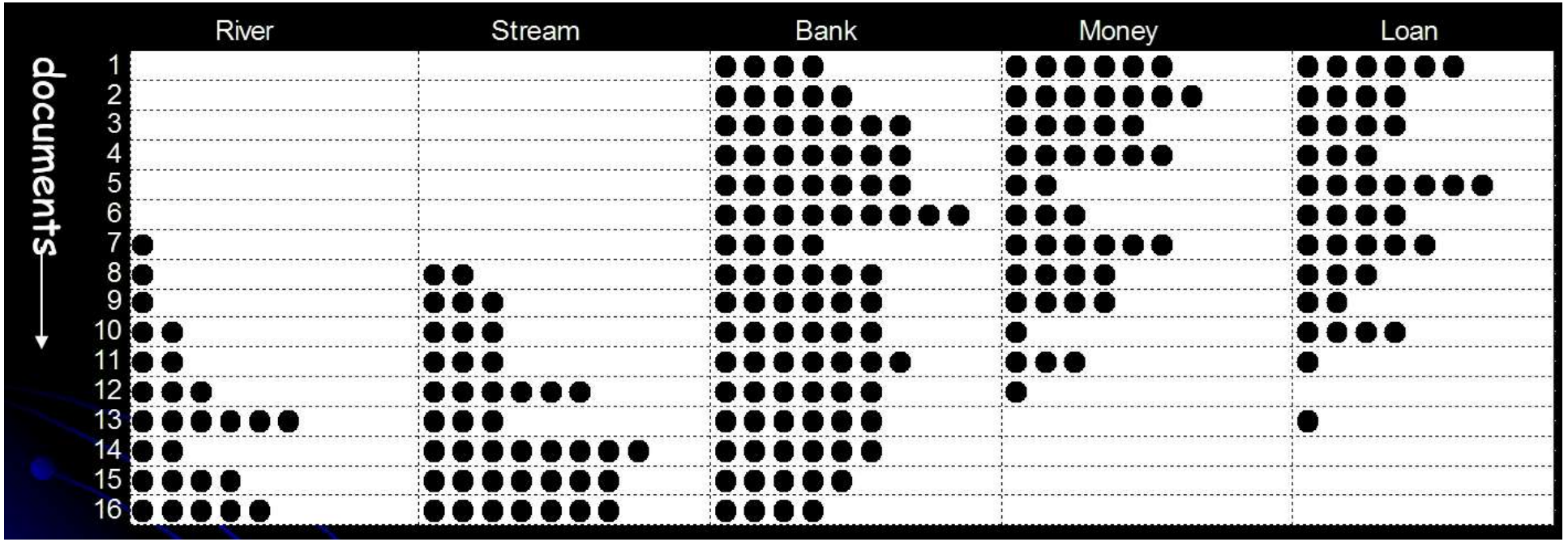
LDA模型

- 参数学习过程(文档生成过程的反过程)：Gibbs采样
- Gibbs Sampling产生一个Markov Chain，其平稳分布是实际的联合分布



LDA模型

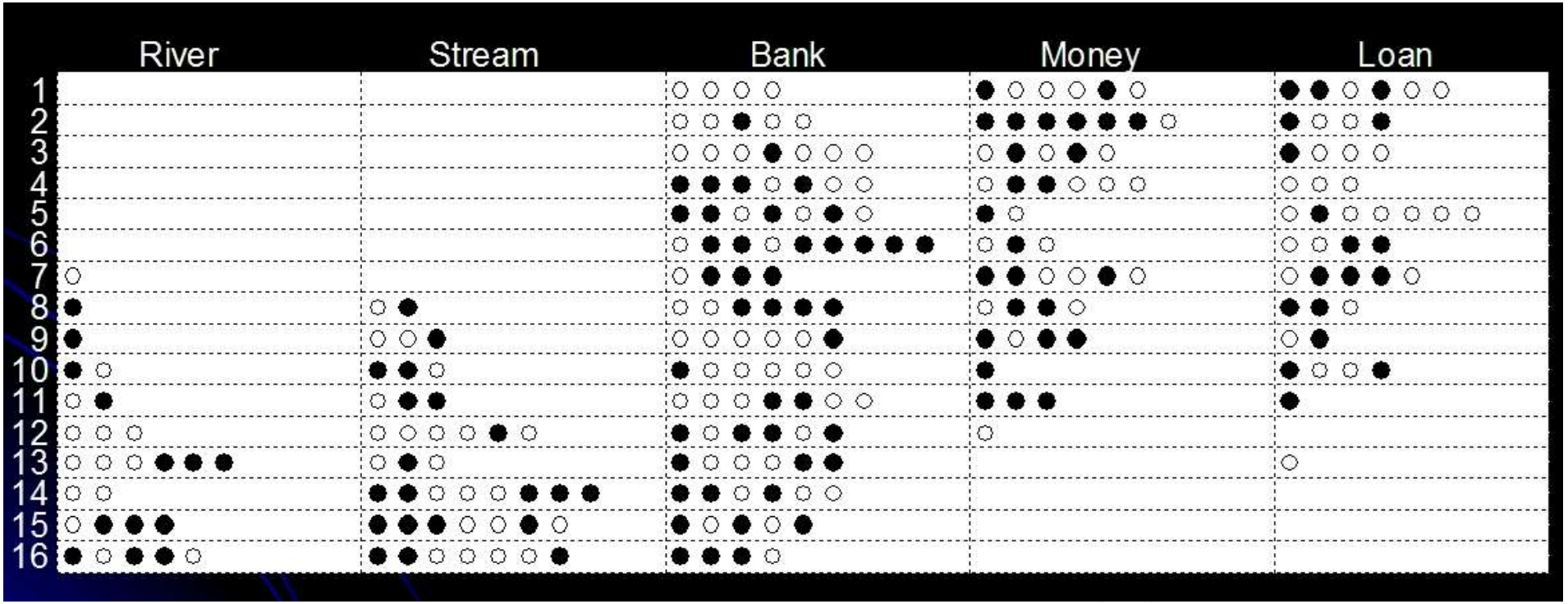
- 参数学习过程(文档生成过程的反过程)：Gibbs采样
  - Can we recover the original topics and topic mixtures from this data?



LDA模型

- 参数学习过程(文档生成过程的反过程)：Gibbs采样
  - Assign word tokens randomly to topics

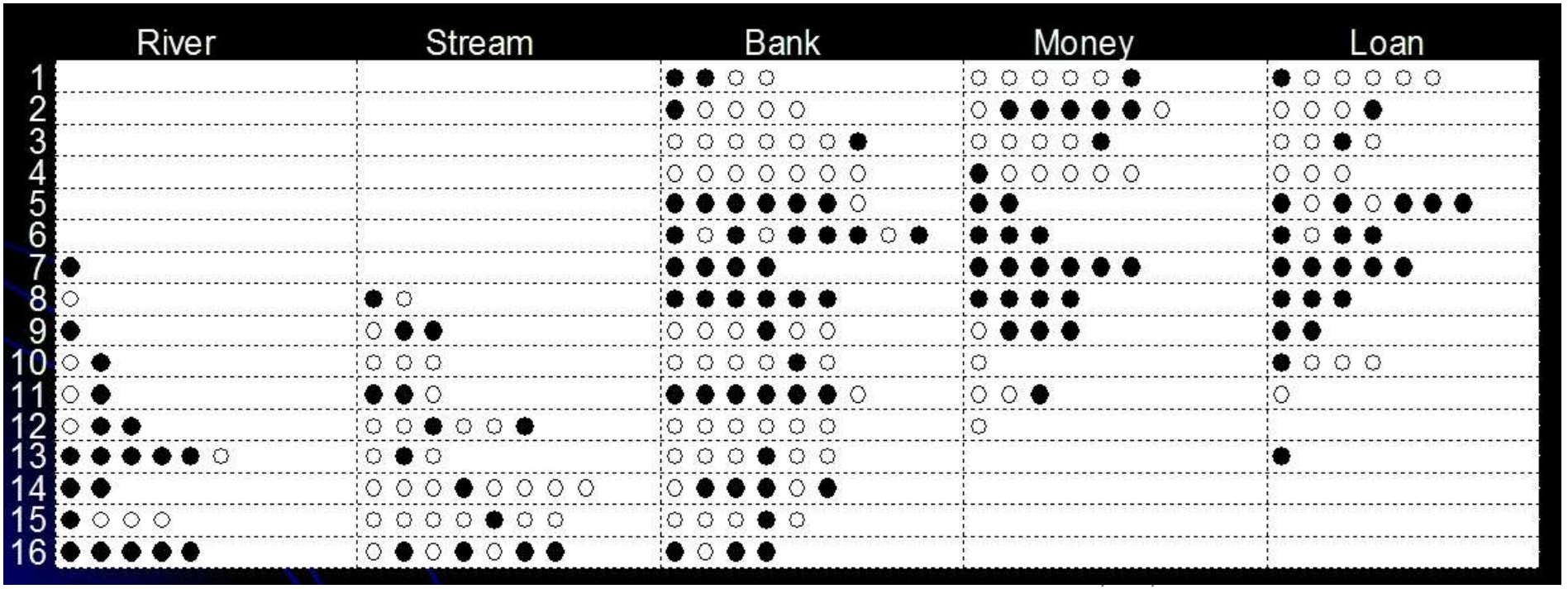
○ **Topic 1**　　● **Topic 2**



LDA模型

- 参数学习过程(文档生成过程的反过程)：Gibbs采样
  - After 1 Iteration: Apply sampling equation to each word token

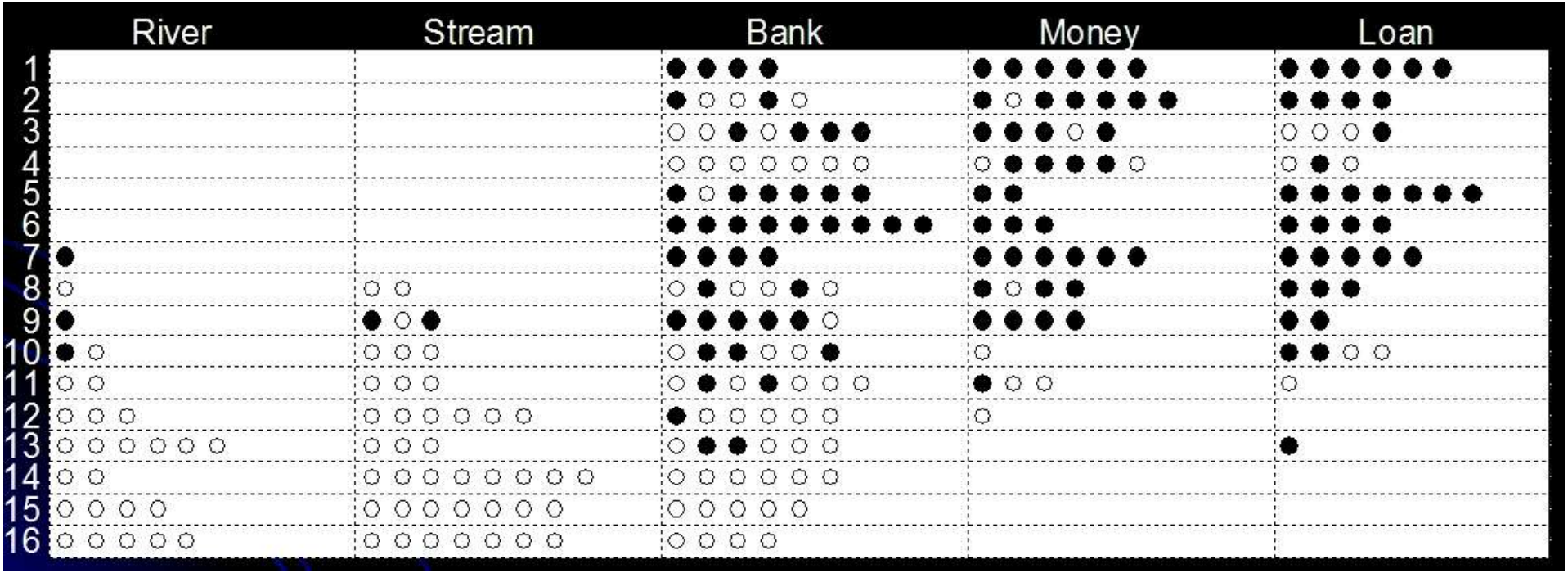
Topic 1　　Topic 2



LDA模型

- 参数学习过程(文档生成过程的反过程)：Gibbs采样
  - After 4 Iterations

Topic 1 **Topic 2**

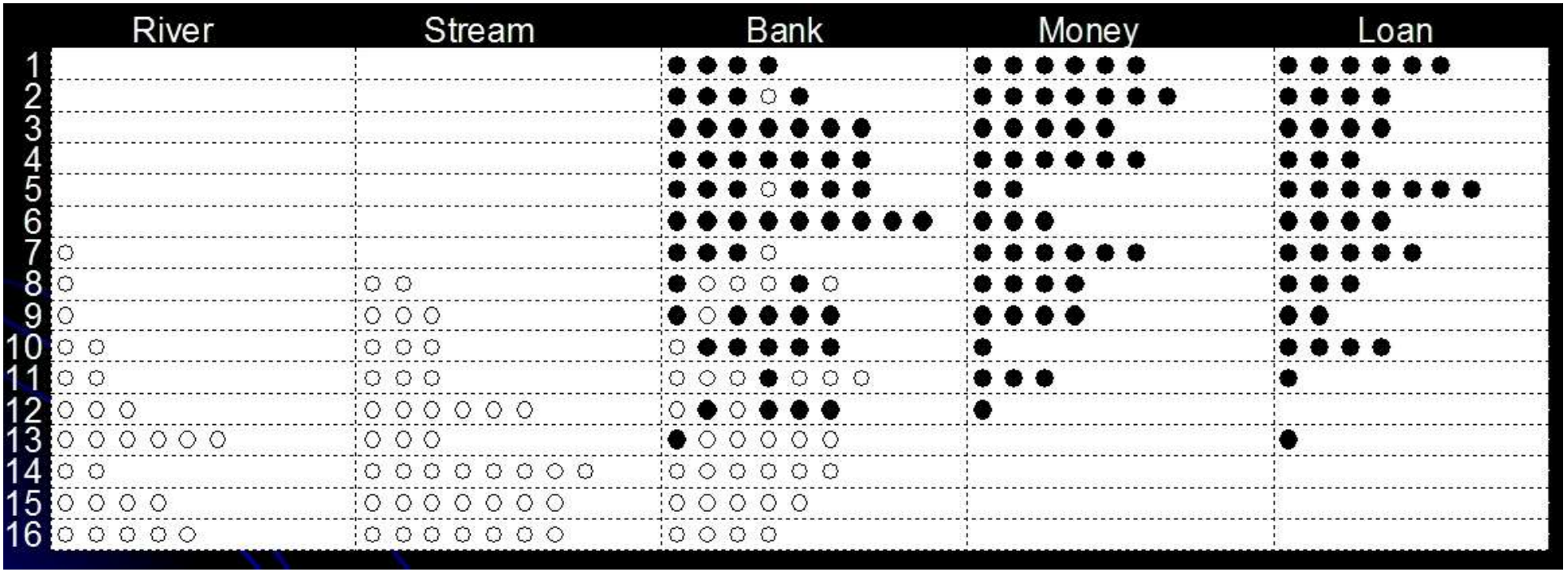


LDA模型

- 参数学习过程(文档生成过程的反过程)：Gibbs采样
  - After 32 Iterations

topic 1	
stream	.40
bank	.35
river	.25

topic 2	
bank	.39
money	.32
loan	.29



LDA模型

- 参数学习过程(文档生成过程的反过程)：Gibbs采样

$$\Pr(U_k^{(t)} \mid U_1^{(t)},\cdots,U_{k-1}^{(t)},U_{k+1}^{(t)},\cdots,U_K^{(t)}):$$

$$P(z_i = j \mid z_{-i}, w_i, d_i, \square) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\beta} \cdot \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT} + T\alpha}$$

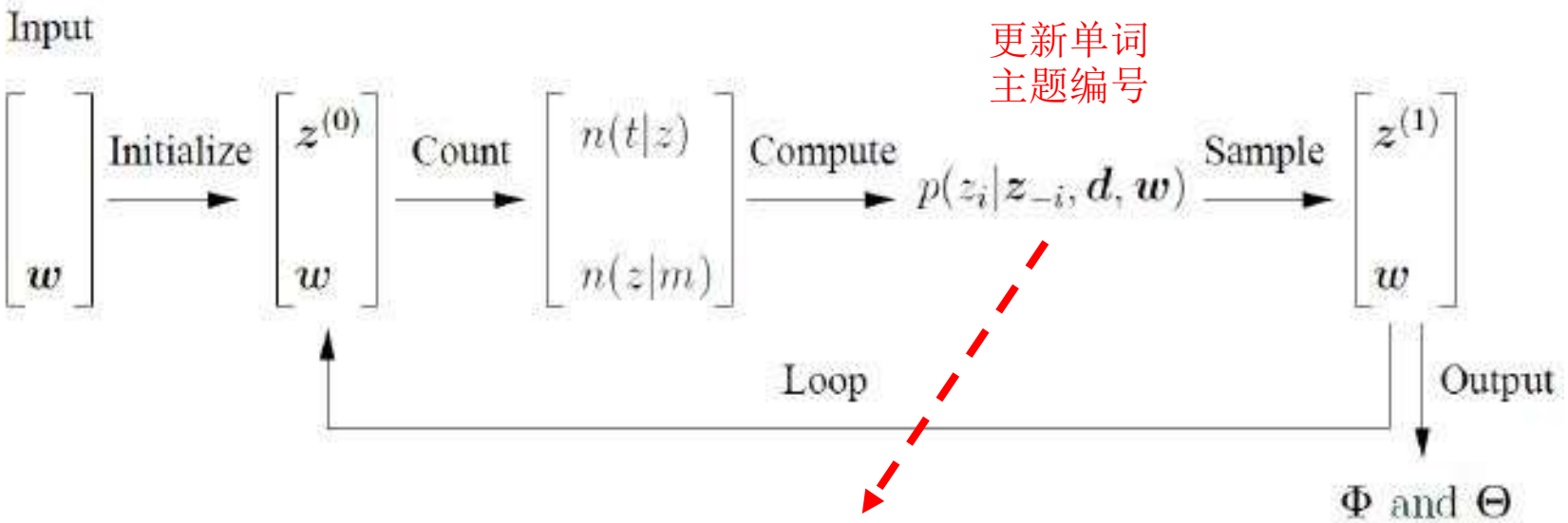
该位置(token)被赋予topic i的概率乘以从topic i里抽出单词j的概率

$$\phi_i^{(w)} = \frac{C_{wi}^{WT} + \beta}{\sum_{k=1}^{W} C_{ki}^{WT} + W\beta}$$

$$\theta_i = \frac{\sum_{d=1}^{D} C_{di}^{DT} + \alpha}{\sum_{k=1}^{T}\sum_{d=1}^{D} C_{dk}^{DT} + T\alpha}$$

LDA模型

- 参数学习过程(文档生成过程的反过程)：Gibbs采样

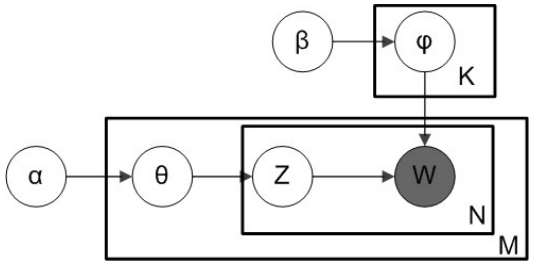


$$p(z_i = k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}) \propto \frac{n^{(k)}_{m,\neg i} + \alpha_k}{\sum_{k=1}^{K}(n^{(t)}_{m,\neg i} + \alpha_k)} \cdot \frac{n^{(t)}_{k,\neg i} + \beta_t}{\sum_{t=1}^{V}(n^{(t)}_{k,\neg i} + \beta_t)}$$

$$= \quad p(topic|doc) \cdot p(word|topic)$$

LDA模型

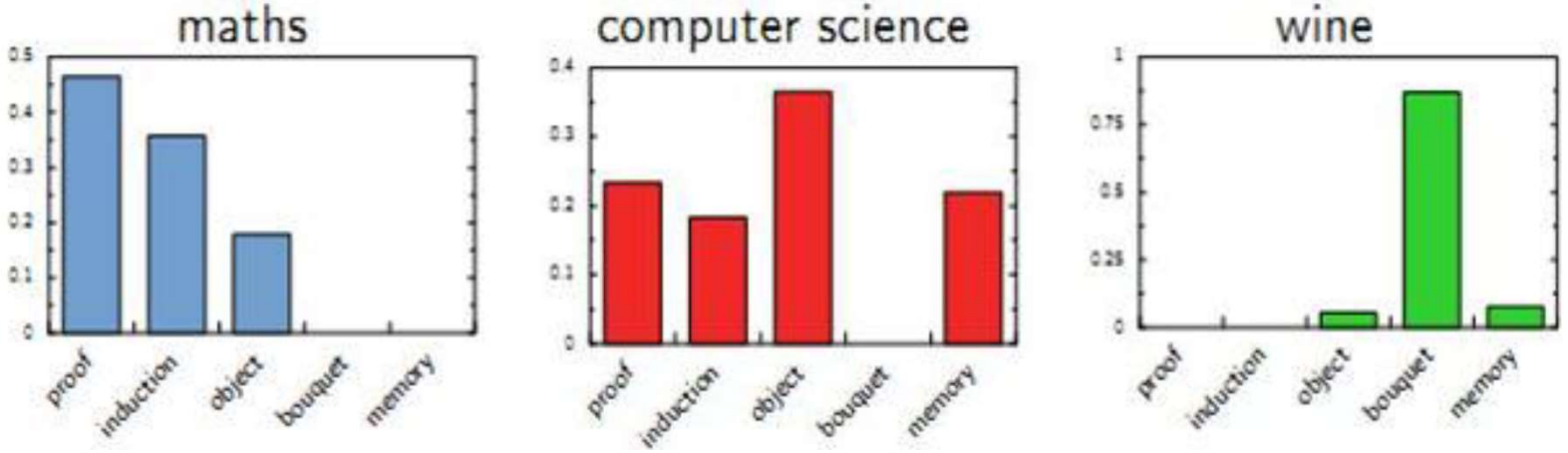
- 估计主体概率分布θ和单词概率分布φ



$$\hat{\theta}_{mk} = \frac{n^{(k)}_{m,\neg i} + \alpha_k}{\sum_{k=1}^{K}(n^{(t)|}_{m,\neg i} + \alpha_k)}$$

$$\hat{\varphi}_{kt} = \frac{n^{(t)}_{k,\neg i} + \beta_t}{\sum_{t=1}^{V}(n^{(t)}_{k,\neg i} + \beta_t)}$$

- 结果示例



LDA模型

- The top 15 most frequent words from the most frequent topics found in the 17,000 articles from the journal Science

"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

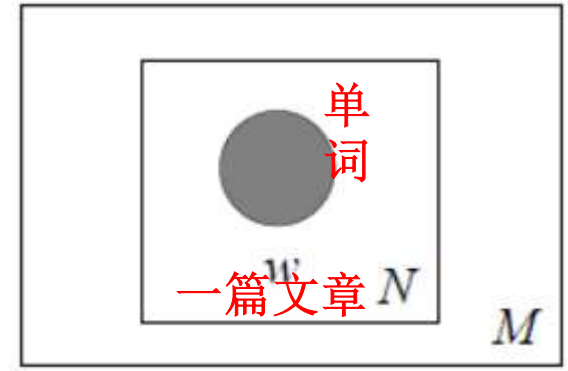
LDA模型应用

- 相似文档发现
  - 主题分布之间的距离

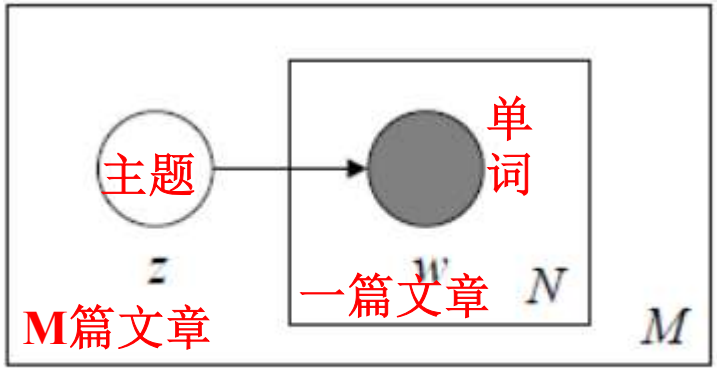
$$doc\_similarity(d,f)=\sum_{i=0}^{K-1}(\sqrt{\theta_{d,i}}-\sqrt{\theta_{f,i}})^2$$

- 文档自动打标签
  - 该文档最大主题概率编号下的概率最大的n个word词
- 新闻个性化推荐系统
  - 根据用户点击过的新闻更新用户兴趣
  - 下次访问时推荐与用户兴趣最接近的文档

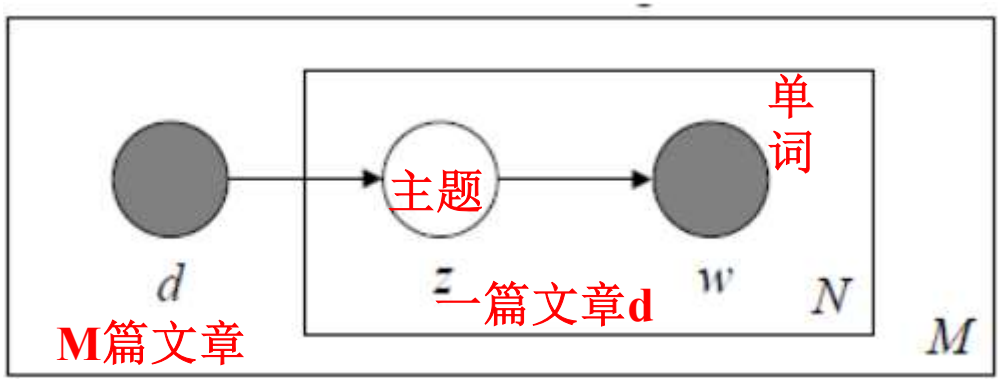
回顾



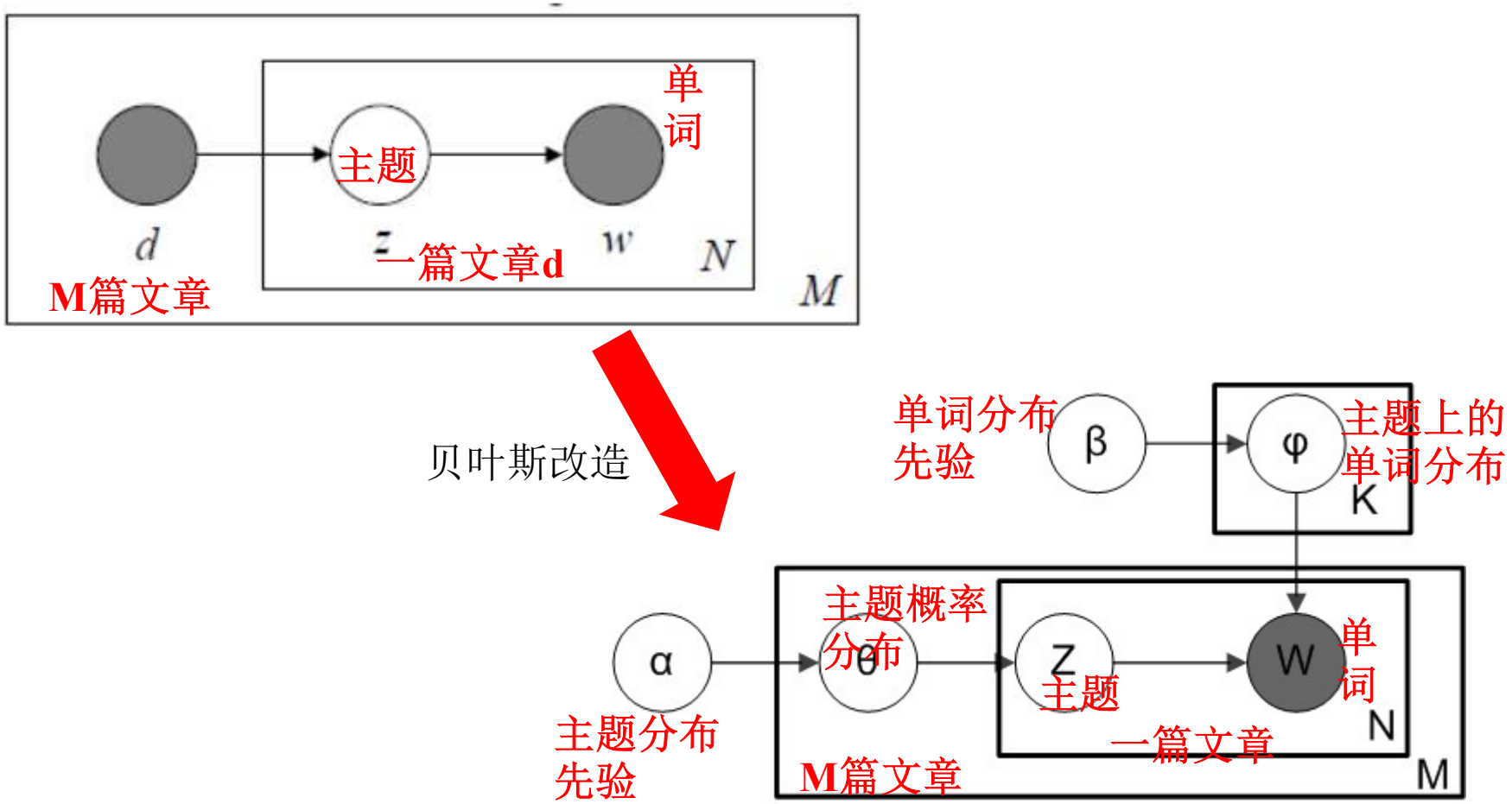
(a)一元模型



(b)一元混合模型



回顾


单词
主题
d
z
w
N
M
M篇文章
一篇文章d
贝叶斯改造
单词分布
先验
β
φ
主题上的
单词分布
K
主题概率
分布
α
θ
Z
W
单
词
主题
N
主题分布
先验
一篇文章
M篇文章
M

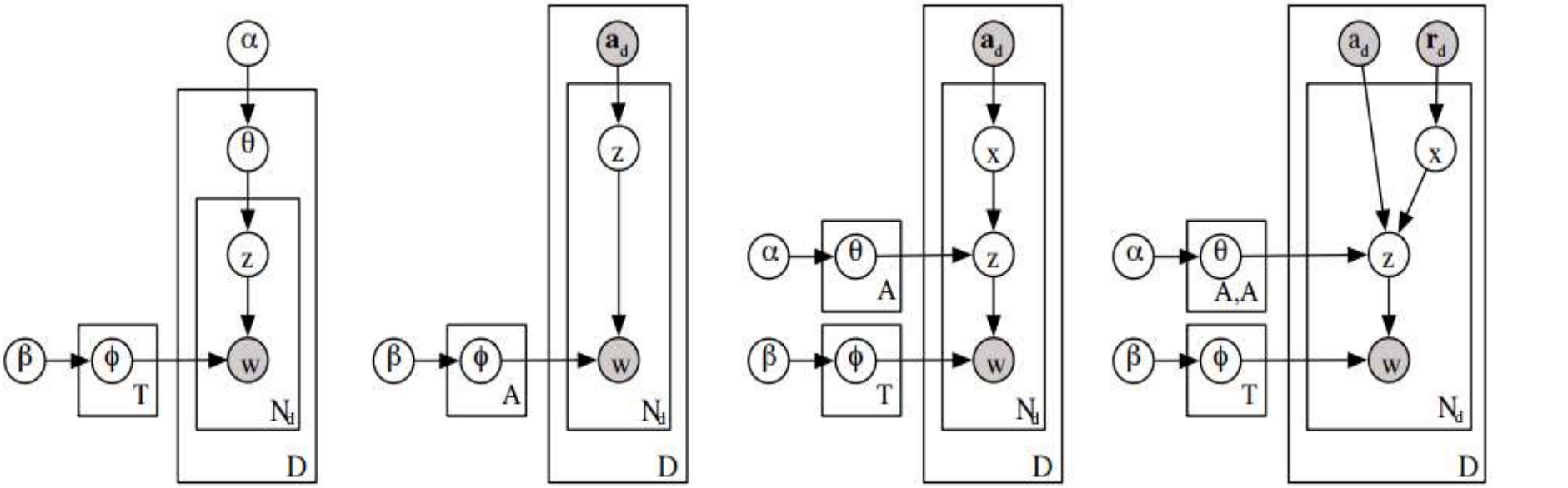
LDA模型变形

Latent Dirichlet Allocation
(LDA)
[Blei, Ng, Jordan, 2003]

Author Model
(Multi-label Mixture Model)
[McCallum 1999]

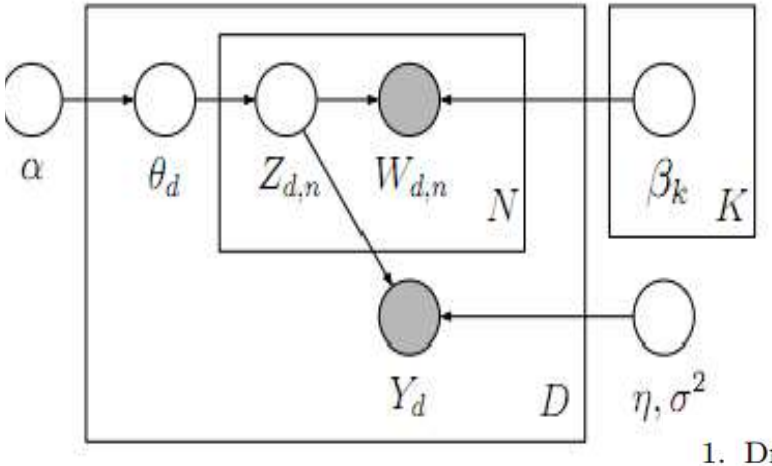
Author-Topic Model
(AT)
[Rosen-Zvi, Griffiths, Steyvers, Smyth 2004]

Author-Recipient-Topic Model
(ART)
[This paper]



Note：LDA（Latent Dirichlet Allocation）
NOT：线性判别分析(Linear Discriminant Analysis)

LDA模型变形



Supervised LDA

1. Draw $\phi^{\mathcal{B}} \sim \mathrm{Dir}(\beta), \pi \sim \mathrm{Dir}(\gamma)$
2. For each topic $t = 1, \ldots, T$,
   (a) draw $\phi^t \sim \mathrm{Dir}(\beta)$
3. For each user $u = 1, \ldots, U$,
   (a) draw $\theta^u \sim \mathrm{Dir}(\alpha)$
   (b) for each tweet $s = 1, \ldots, N_u$
      i. draw $z_{u,s} \sim \mathrm{Multi}(\theta^u)$
      ii. for each word $n = 1, \ldots, N_{u,s}$
         A. draw $y_{u,s,n} \sim \mathrm{Multi}(\pi)$
         B. draw $w_{u,s,n} \sim \mathrm{Multi}(\phi^{\mathcal{B}})$ if $y_{u,s,n} = 0$ and $w_{u,s,n} \sim \mathrm{Multi}(\phi^{z_{u,s}})$ if $y_{u,s,n} = 1$

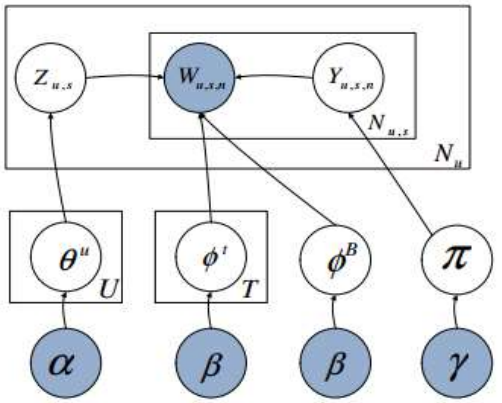
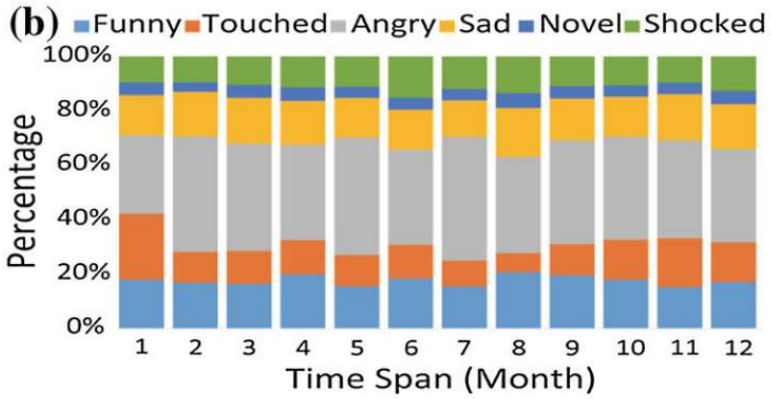
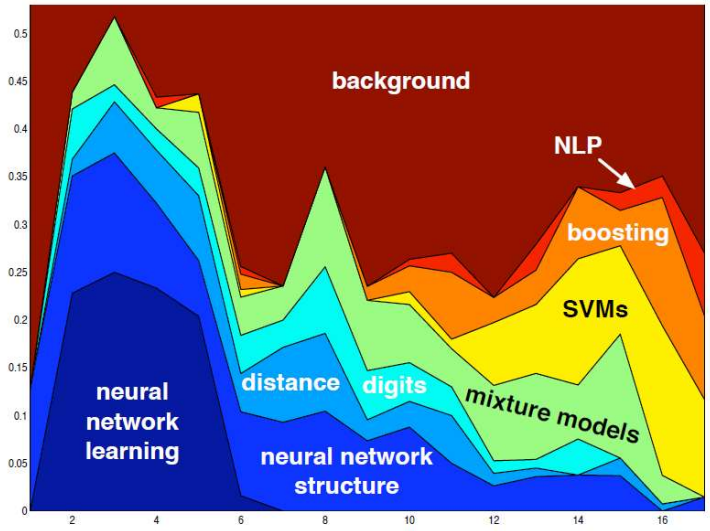


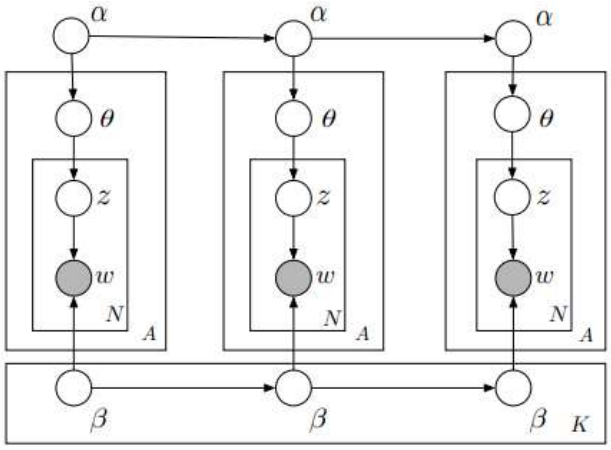
Fig. 1. The generation process of tweets. **Fig. 2.** Plate notation of our Twitter-LDA.

LDA模型变形

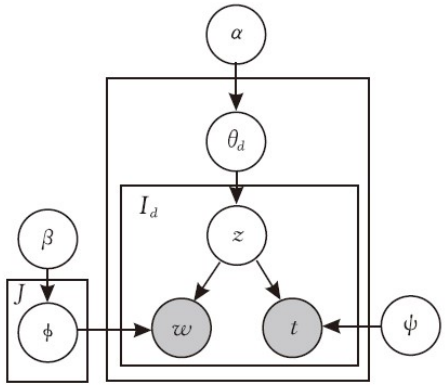
- 动态LDA模型



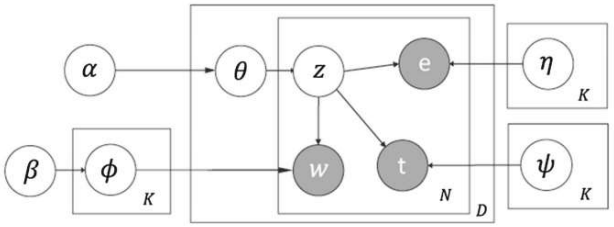




Dynamic Topic Models (DTM)



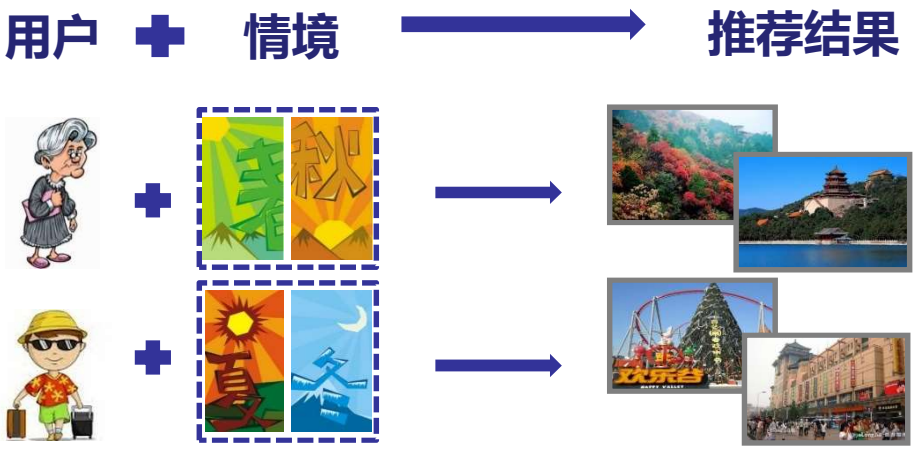
Topics over Time (TOT)



Emotion-topic over time (eTOT)

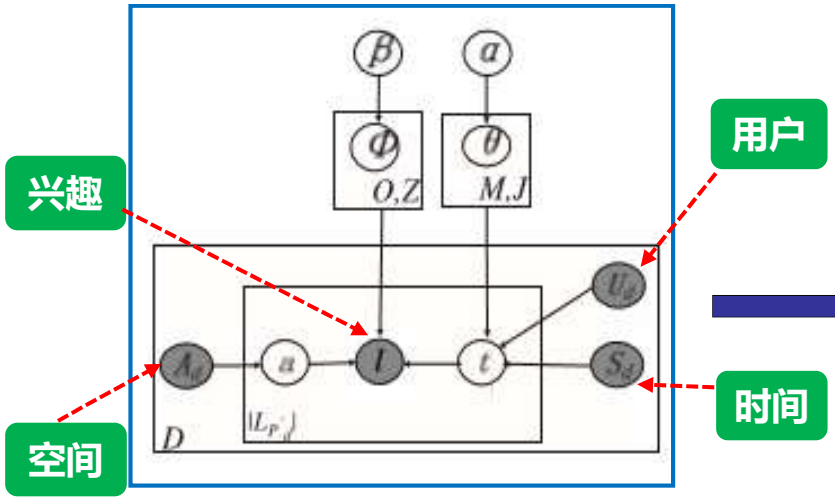
主题模型其他应用（1）

- 融合情境信息的个性化旅游套餐推荐
  - 旅游数据非常稀疏
  - 旅游套餐（文档）推荐有比较强的时间依赖性
  - 每个景点(单词)都位于不同的地区且适合不一样的季节
  - 旅游套餐的价格也会影响游客的选择
  - 很少有游客愿意主动给旅游套餐评分



主题模型其他应用（1）

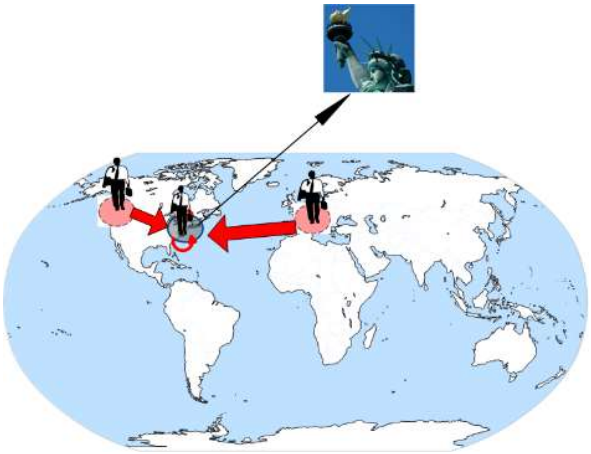
- 建立了用户、情境、推荐对象多边关系的概率图模型
  - **将相同时空情境下喜欢相似推荐对象的用户投影到相近隐空间**
  - **以概率分布形式表示情境感知的用户兴趣模型，提高个性化推荐方法的精度**



情境感知的用户兴趣挖掘模型

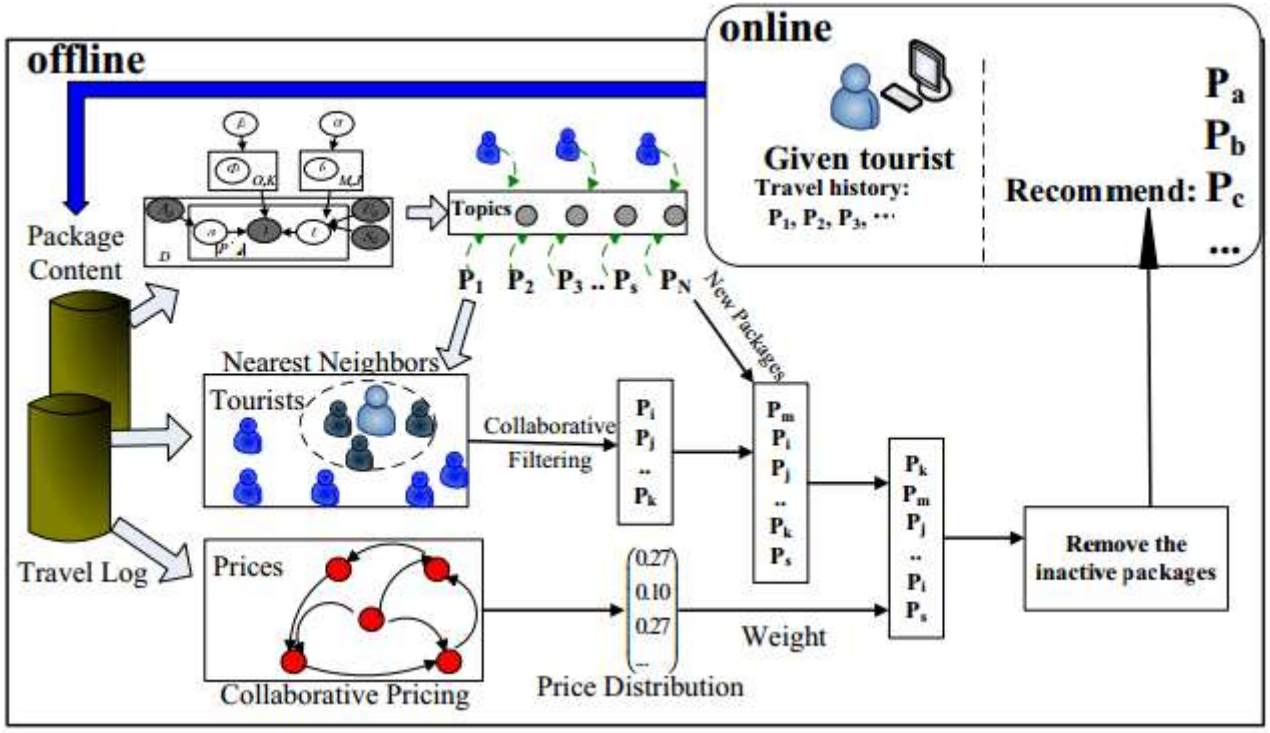
$$\begin{matrix} U_1 & U_2 \\ 0.27 & 0.05 \\ 0.10 & 0.02 \\ 0.05 & 0.27 \\ 0.27 & 0.02 \\ 0.02 & 0.27 \\ \dots & \dots \end{matrix}$$

用户兴趣表示



主题模型其他应用（1）

- 建立了用户、情境、推荐对象多边关系的概率图模型

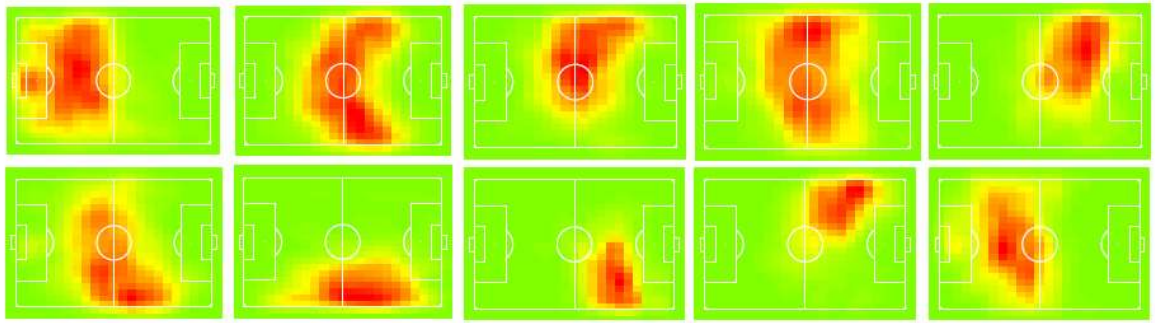


与经典推荐方法UCF和SVD相比，推荐算法的平均推荐精度提升幅度超过11%（*IEEE TKDE* 2014）

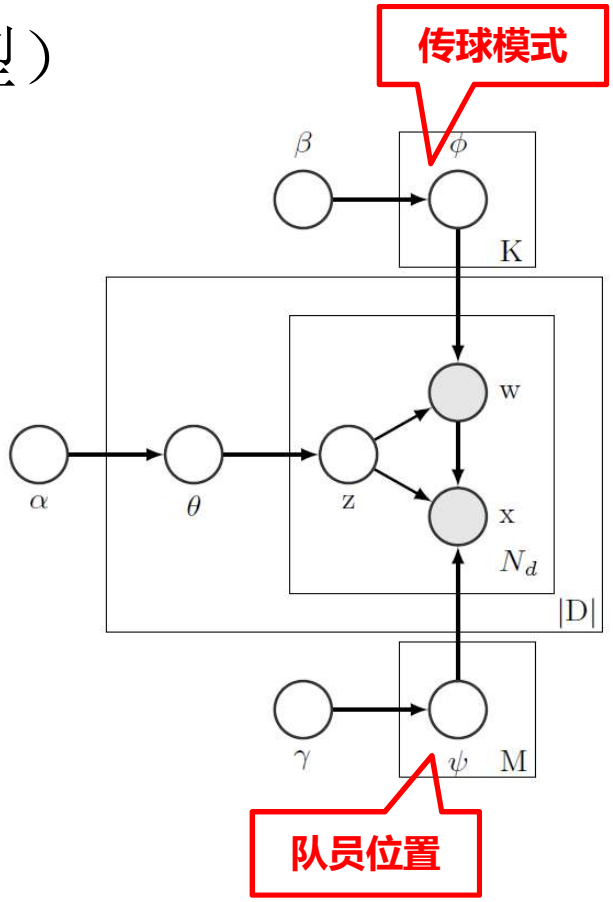
Qi Liu, Yong Ge, Zhongmou li, **Enhong Chen***, Hui Xiong*, **Personalized Travel Package Recommendation**. *The 11th IEEE International Conference on Data Mining(ICDM'2011):407-416* **(Best Research Paper)** Vancouver, Canada, December 11-14, 2011.

主题模型其他应用（2）

- 职业足球队战术策略分析（T3M模型）
  - 每个足球队有多个战术策略
  - 每个战术策略包括
    - 传球模式（谁传谁）
    - 队员位置（在场地的哪个部位接球）



T3M模型发现的10个战术策略



Qing Wang, **Hengshu Zhu**, Wei Hu, Zhiyong Shen, Yuan Yao, **Discerning Tactical Patterns for Professional Soccer Teams: An Enhanced Topic Model with Applications**, *The 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* ***(KDD-2015)***, Sydney, Australia, August 10-13, 2015.

主题模型

- 学习资料
  - 马晨. LDA漫游指南
  - 靳志辉. LDA数学八卦
  - Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
  - Gregor Heinrich. Parameter estimation for text analysis.