**Host 聚类算实现**

作者：

时间：

* **背景**

该项目的目标是能够自动的把一个具有一定量的HOST归到一定类别当中，并能够自动的加入一定的广告策略当中。

Host聚类所用文本主要是由一些由类别词和关键词组成（Keywords和Description），这些通常变化是很小的；以及一些网内链接的锚文本所组成，这些相对变化较大些，会每天都有一定的更新。所以在对HOST进行聚类时，我们更倾向于选择Keywords和Description作为聚类的基础。

# 算法实现

* **输入**：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| HOST | KEY1 | KEY2 | KEY3 | KEY4 | KEY5 |  |
| Host1 | Key11 | Key12 | Key13 | Key14 | Key15 | ... |
| Host2 | Key21 | Key22 | Key23 | Key24 | Key25 | ... |
| Host3 | Key31 | Key32 | Key33 | Key34 | Key35 | ... |
| Host4 | Key41 | Key42 | Key43 | Key44 | Key45 | ... |
| ... |  |  |  |  |  |  |

* **输出：**

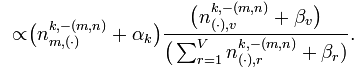
|  |  |  |
| --- | --- | --- |
| Host1 | Predict Type1 | ... |
| Host2 | Predict Type2 | ... |
| Host3 | Predict Type3 | ... |
| ... |  |  |

* **处理过程**：

1. 输入M 个Host, 依次为 Host[1] , Host[2] ... Host[M], Host[m] 为存放该Host关键词的数组
2. 使用LDA(Latent Dirichlet allocation) 对每个关键词 Key[i], , N 为 在Markov chain 上进行抽样确定每个Key[i]的标记Type[j] , 1j, K 为预设的主题个数

Gibbs抽样所用条件概率公式是：



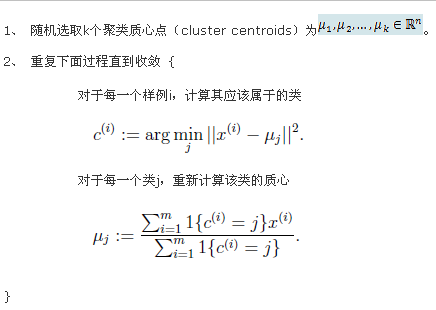


从抽样样本(即对N个key都给定相应的Type标签)计算参数：



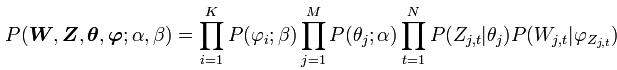
是指文档d属于主题j的概率。可以进行多次抽样，计算多个样本

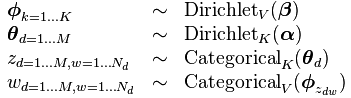
1. 假设, 根据该条件概率构造 矩阵, 作为对角元素构造方阵，按同样的方法从抽样样本计算。矩阵的第i行与第j行的内积作为文档和的相似度。
2. 使用K-Means对文档(在这里指host)聚类。文档的坐标设置为矩阵的第i行：



* **主要函数介绍**
* **LDA所涉及符号的解释及Gibbs抽样所用条件概率公式的推导**

|  |  |  |
| --- | --- | --- |
| **变量** | **类型** | **解释** |
| K | 整数 | 预设主题的个数，例如50 |
|  | 整数 | 词典中单词的个数(无重复) |
|  | 整数 | 文档的个数 |
|  | 整数 | 文档中单词的个数 |
| N | 整数 | 所有文档中单词的个数之和， |
| \alpha_{k=1 \dots K} | 正实数 | 主题在一篇文档中的先验权重；通常队于所有的主题都相同；一般是小于1，例如0.1，倾向于稀疏的主题分布，例如一篇文档只有少量的主题 |
| \boldsymbol\alpha | 正实数的维向量 | \alpha_k值的集合，通常作为一个向量 |
| \beta_{w=1 \dots V} | 正实数 | 单词*w*在一个主题中的先验权重；通常对于所有的单词都相同；一般是一个小于1的数，例如0.001，倾向于稀疏的主题分布，即每个主题有较少的单词 |
| \boldsymbol\beta | *V*维的正实数向量 | \beta_w值的集合，通常作为一个向量 |
| \phi_{k=1 \dots K,w=1 \dots V} | 概率（0和1之间的正实数） | 单词在主题k中的概率 |
| \boldsymbol\phi_{k=1 \dots K} | 维的概率向量，元素加和为1 | 单词在主题k中的分布 |
| \theta_{d=1 \dots M,k=1 \dots K} | 概率（0和1之间的正实数） | 文档d中单词属于主题k的概率 |
| \boldsymbol\theta_{d=1 \dots M} | 维的概率向量，元素加和为1 | 文档d中的主题分布 |
| z_{d=1 \dots M,w=1 \dots N_d} | 1和K之间的整数 | 单词w在文档d中的赋值 |
|  | N维整数向量 | 文档d中所有单词（可重复）的赋值 |
|  | 1和V之间的整数 | 文档d中所有单词w的赋值 |
| \mathbf{W} | N维整数向量 | 所有文档包含所有单词的赋值 |

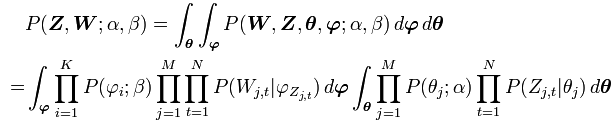
随机变量所服从的分布



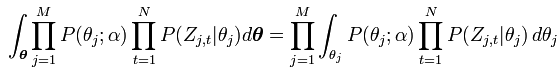
* 和的联合概率分布是：

[collapsed Gibs sampling](http://en.wikipedia.org/wiki/Collapsed_Gibbs_sampling)

* 和\boldsymbol{\theta}积分出来：

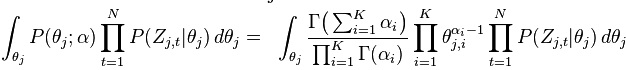


* 和\theta相互独立，分别处理：

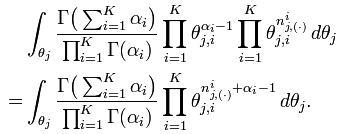


给定\theta文档相互独立，所以可以只考虑文档

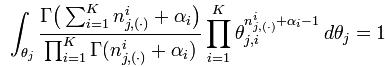




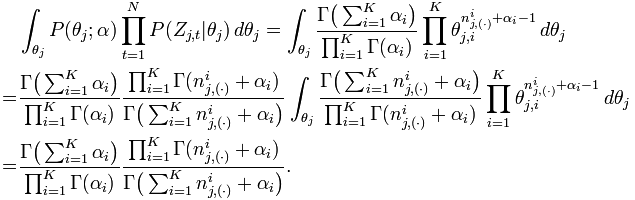




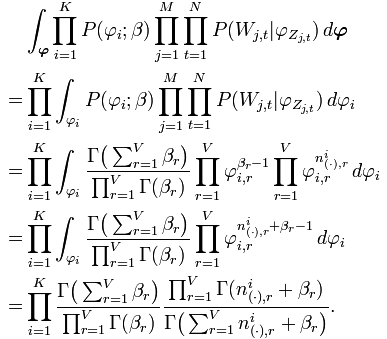
因为：



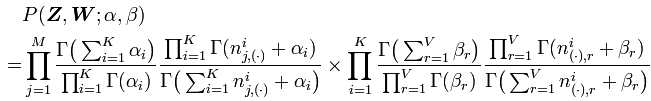
所以：



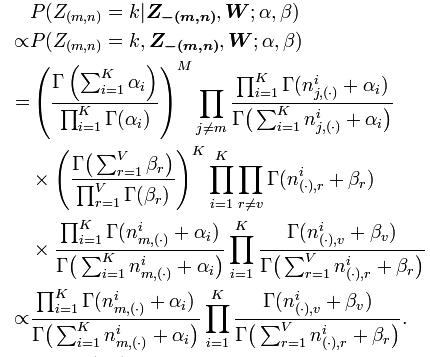
同样的道理：



合并\boldsymbol{\varphi}和\boldsymbol{\theta}部分

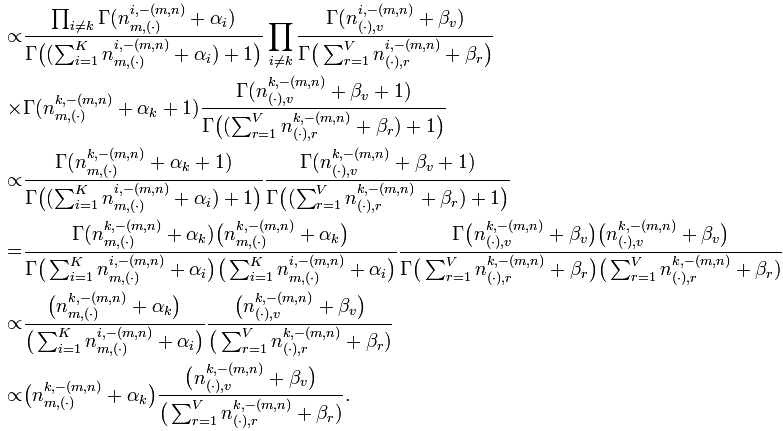








因为是确定的。



公式的推导

