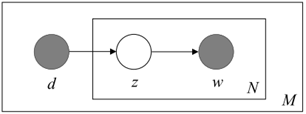
**概率潜在语义分析**(probabilistic latent semantic analysis，PLSA)是一种利用概率生成模型对文本集合进行话题分析的无监督学习方法。

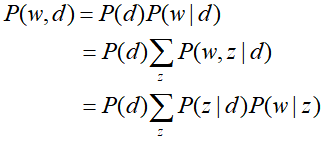
跟潜在语义分析相似，而其特点是基于概率模型，用隐变量表示话题；整个模型表示文本生成话题，话题生成单词，从而得到单词-文本共现数据的过程。

**生成模型**



从数据生成过程可推出文本-单词共现数据T的生成概率为所有单词-文本对（w，d）的生成概率乘积，IMG_260

而每个单词-文本对（w，d）的生成概率由以下公式决定：



**共现模型**

文本-单词共现数据T的生成概率为所有单词-文本对（w，d）的生成概率乘积：

IMG_262

但每个单词-文本对（w，d）的生成概率由以下公式决定：

IMG_263

共现模型假设在话题z给定的条件下，单词w与文本d是条件独立的，即IMG_264

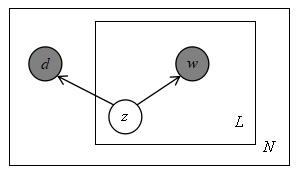


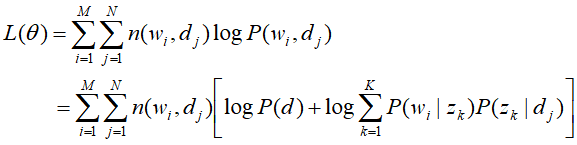
图1-2 概率潜在语义模型的共现模型

虽然生成模型与共现模型在概率公式的意义上是等价的，但拥有不同的性质。生成模型刻画文本-单词共现数据生成的过程，且单词变量w与文本变量d是不对称的，而共现模型描述文本-单词共现数据的拥有过模式，单词变量w与文本变量d是对称的。

生成模型的EM算法

概率潜在语义分析模型是含有隐变量的模型，其学习通常使用EM算法。

目标是估计概率潜在语义分析生成模型的参数IMG_266和IMG_267，使用极大似然估计，在其对数似然函数是：



其中P(d)与需要估计的参数无关，将其舍去也不影响结果，那么代求问题也将变为舍去无关项后对数似然函数的极大问题，

M步求最大化，即最大化Q函数以估计我们所需的参数。然后根据新的参数求p(z|d,w)，再写出新的下界函数，再最大化…反复迭代直至收敛。这就是将EM算法应用到PLSA的求解中。