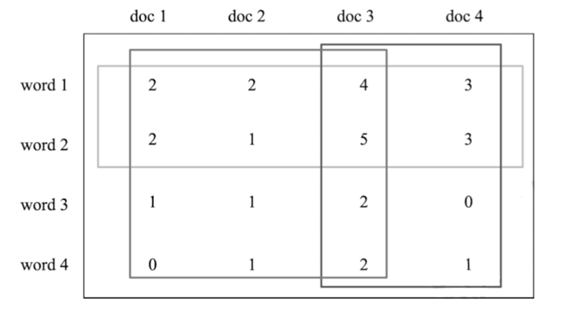
**读书报告**

61518424 王贵涛

1. **问题与解答**

我提出问题：

1. 为什么每个话题也可以由若干个文本表示？



讨论结果：不同的文本可以是属于同一个话题的，比如都属于体育这个话题，也就是对于不同文本可以归属到几个话题之中。

1. 共现模型和生成模型各自的优势是什么？

讨论结果：生成模型是文本到话题到单词。而贡献模型是由话题到单词，话题到文本，学习思路不同。

别人提出的问题：

1. 概率潜在语义分析模型的共现模型与17章中非负矩阵分解有什么关系？非负矩阵分解得到结果均非负，是不是能用这种方法求解贡献模型的参数？

我的解答：这两个结果应该不同，不过应该也有相似的地方，可以将非负矩阵分解的结果作为EM算法的初始值来加快求解速度。

1. PLSA为什么可以防止过拟合？

我的解答：PLSA采用了最大似然估计算法，将期望最大化，PLSA中训练参数的值会随着文档的数目线性递增。

1. **下周计划安排**

看完第十九章并参加讨论。

1. **读书收获**

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。模型的最大特点是用隐变量表示话题；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假设每个文本由一个话题分布决定，每个话题由一个单词分布决定。 概率潜在语义分析受潜在语义分析的启发，前者基于概率模型，后者基于非概率模型。

给定一个文本集合，每个文本讨论若干个话题，每个话题由若干个单词表示。对文本集合进行概率潜在语义分析，就能够发现每个文本的话题，以及每个话题的单词。 话题是不能从数据中直接观察到的，是潜在的。文本集合转换为文本-单词共现数据，具体表现为单词-文本矩阵

概率潜在语义分析就是发现由隐变量表示的话题，即潜在语义。直观上，语义相近的单词、语义相近的文本会被聚到相同的“软的类别”中，而话题所表示的就是这样的软的类别。

虽然生成模型与共现模型在概率公式意义上是等价的，但是拥有不同的性质。

生成模型：

1. 刻画文本-单词共现数据生成的过程
2. 单词变量w与文本变量d是非对称的
3. 非对称模型

共现模型：

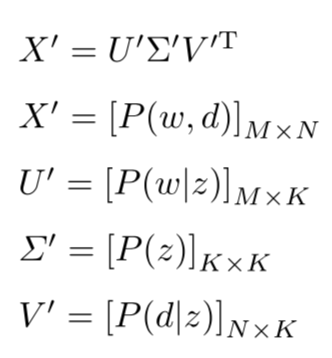
1. 描述文本-单词共现数据拥有的模式
2. 单词变量w与文本变量d是对称的
3. 对称模型

概率潜在语义分析模型（共现模型）可以在潜在语义分析模型的框架下描述。

对单词-文本矩阵进行奇异值分解得到：



共现模型也可以表示为三个矩阵乘积的形式：



概率潜在语义分析模型中的矩阵U’和V’是非负的、规范化的，表示条件概率分布，潜在语义分析模型中的矩阵U和V是正交的，未必非负，并不表示概率分布。

概率潜在语义模型的参数估计采用EM算法。

