Latent Dirichlet Allocation

（隐狄利克雷分配模型）

**摘 要：**我们把LDA模型描述为一种用于离散数据(如文本语料库)集合的生成概率模型。LDA是一个三层的层次贝叶斯模型，其中集合的每一项都被建模为一组潜在主题的有限混合。每个主题依次被建模为一组潜在主题概率的无限混合。在文本建模中，主题概率提供了文档的显式表达。我们提出了一种有效的基于变分方法的近似推理技术和一种EM算法应用于经验贝叶斯参数估计。我们展示了文档建模、文本分类和协同过滤的结果，并与一元混合模型和概率LSI模型的混合进行了比较。

# 介绍

本文研究了文本语料库和其它离散数据集合的建模问题。其目标是找到对集合成员的简短描述，以便对大型集合进行有效的处理，同时保留基本的统计关系，这些统计关系对分类、异常检测、摘要、相似性和相关性判断等基本任务都非常有用。

信息检索(IR)领域的研究人员(Baeza-Yates和Ribeiro-Neto, 1999)在这一问题上取得了重大进展。文本语料库的信息检索研究人员提出的基本方法——一种成功应用于现代互联网搜索引擎的方法——将语料库中的每个文档简化为实数向量，每个实数向量表示（词汇的）计数比率。在流行的tf-idf模式中(Salton和McGill, 1983)，对于文集中的每个文档选择“词”或“术语”作为基本单位，对每个单词的出现次数进行计数。经过适当的标准化之后，将这个词频计数与反向文档频率计数进行比较，后者度量一个词在整个语料库中出现的次数(通常以对数形式表达，然后再次进行适当的标准化)。最终结果是一个文档单词的矩阵，其列包含语料库中每个文档的tf-idf值。因此tf-idf方案将任意长度的文档缩减为固定长度的数字列表。

虽然tf-idf模式具有一些吸引人的特性——特别是在对能够区分集合中文档的词集的基本识别中，但是该方法并没有减少对文档的描述长度，而且没有揭示出文档内部或文档之间的统计结构。为了解决这些缺点，信息检索的研究人员提出了几个其他的降维技术，其中最著名的是隐语义索引(LSI) (Deerwester等，1990)。LSI使用X矩阵的奇异值分解来识别tf-idf特征空间中的一个线性子空间，该子空间捕获集合中的大部分变异数。这种方法可以在大型集合中实现显著的压缩。此外，Deerwester等人认为，LSI的衍生特征，即原始tf-idf特征的线性组合，可以捕捉一些基本语言概念，如同义词和多义词。

为了证实关于LSI的断言，并研究其相对的优势和劣势，开发文本语料库的生成概率模型和研究LSI从数据中恢复生成模型的能力是相当有帮助的(Papadimitriou等.，1998)。然而，在文本的生成模型上，我们并不清楚为什么应该采用LSI方法——我们可以尝试更直接的方法，使用极大似然或贝叶斯方法将模型与数据进行拟合。

Hofmann(1999)在这方面迈出了重要的一步，他提出了概率LSI (pLSI)模型，也称为特征模型（aspect model），作为LSI的替代品。我们将在4.3节中详细描述pLSI方法，它将文档中的每个单词建模为混合模型的样本，其中混合成分是多项式随机变量，可以将其视为“主题”的表示。因此，每个单词都是由一个主题生成的，文档中的不同单词可能是由不同的主题生成的。每个文档都可以表示为这些混合成分的混合比例列表，从而简化为一组固定主题集上的概率分布。这个分布是与文档相关联的“简化描述”。

虽然Hofmann的工作是迈向文本概率建模的重要一步，但它并不完整，因为它没有提供在文档层面上的概率模型。在pLSI中，每个文档都表示为数字列表(主题的混合比例)，而没有这些数字的生成概率模型。这导致了几个问题:（1）模型中参数的数量随着语料库的大小线性增长，导致严重的过拟合问题（2）不清楚如何将概率分配给训练集之外的文档。

要了解如何超越pLSI，让我们考虑包括LSI和pLSI在内的一类降维方法的基本概率假设。所有这些方法都是基于“词袋”的假设，即可以忽略文档中单词的顺序。在概率论的语境中，这是一个关于文档中词语的可交换性（exchangeability）的假设(Aldous, 1985)。此外，尽管没有正是说明，这些方法也假定文档是可交换的；语料库中文档的特定顺序可以忽略。

De Finetti(1990)提出了一个经典的表示定理，即任何可交换随机变量的集合都有一个混合分布的表示——通常是一个无限的混合。因此，如果我们考虑文档和单词的可交换表示，我们需要考虑能够捕获单词和文档的可交换性的混合模型。这一思路引出了我们在本文中提出的隐狄利克雷分配(LDA)模型。

需要强调的是，可交换性假设并不等于假设随机变量是独立且同分布的。相反，可交换性本质上可以解释为“条件独立且分布相同”，即条件是相对于概率分布的潜在参数而言的。在一定的条件下，随机变量的联合分布是简单的，而在潜在参数上的联合分布可能非常复杂。因此,虽然可交换性的假设是文本建模的一个主要的简化假设，而且它也提高了计算效率，但可交换性的假设并不一定意味着可以进行简单的频率计数或线性操作。我们将在论文中证明，通过论证de Finetti定理，我们可以通过混合分布捕获重要的文档内统计结构。

同样值得注意的是，对可交换性的基本概念已有大量的总结概括，包括各种形式的部分交换性，并且上面提到的表示定理也可用于部分可交换的情况（Diaconis，1988）。因此，虽然我们在当前的论文中讨论的工作集中在简单的“词袋”模型上（表现为单个单词(unigram)的混合分布），我们的方法也适用于涉及到更大结构单元的更丰富的模型，如n-gram或段落。

本文的结构如下：在第2节中，我们介绍基本的符号和术语。 第3节介绍LDA模型，并与第4节中相关的隐变量模型进行比较。我们在第5节讨论LDA的推导和参数估计。第6节提供了使用LDA拟合数据的一个例子。第7姐展示文本建模，文本分类和协作过滤的实证结果。最后，第8节给出我们的结论。

# 符号和术语

我们在整篇论文中都使用文本集合的说法，指的是诸如“单词”，“文档”和“语料库”这样的实体时。这有助于引导直觉，特别是当我们引入旨在捕获如主题等抽象概念的隐变量时。然而，需要注意的是，LDA模型不一定与文本绑定，它还可以应用于涉及数据集合的其他问题，包括来自协同过滤、基于内容的图像检索和生物信息学等领域的数据。事实上，在7.3节中，我们展示了协同过滤领域的实验结果。

我们定义以下术语：

* 单词是离散数据的基本单位，来自以作为索引的词汇表。我们用单位基向量来表示单词，它有一个分量等于1，其他分量都等于0。因此，使用上标来表示，词汇表中的第个单词在维向量中表示为：。