主题模型在文本挖掘中的应用

赵鑫, 李晓明

Peking University, Beijing, China

Technical Report PKU-CS-NCIS-TR2011XX
June 2011
Computer Networks and Distributed Systems Lab
Peking University
Beijing, China 100871

主题模型在文本挖掘中的应用

赵鑫, 李晓明

Computer Networks and Distributed Systems Lab, Peking University
Email: *batmanfly@gmail.com, lxm@pku.edu.cn*

Abstract

主题模型是进些年来文本挖掘中出现的一种概率模型，不再像传统的空间向量模型和语言模型那样，只单纯地考虑文档在词典空间上的维度，而是引入了主题空间，从而实现了文档在主题空间上的表示，每一个主题是一个在词典空间上的概率分布。通过引入主题这个概念，可以带来两个好处：1）实现了文档的低维表示；2）抽取了文档集合上隐含语义的挖掘，即主题。自从David Blei提出Latent Dirichlet Allocation（LDA）后，该论文已经被引用2836多次[1]，主题模型的应用几乎覆盖了文本挖掘和信息处理的所有领域。主题模型背后的数学基础部分较为复杂，本综述尽量回避一些枯燥而又复杂的公式推导，从文本挖掘的角度来分析主题模型，为进一步理解和应用主题模型做一个铺垫工作。

关键词 主题模型，文本挖掘，信息处理

1 引言

出现主题模型前，信息处理和文本挖掘领域在文本表示上主要还停留在1）空间向量模型 [78]；2）统计语言模型 [71]。空间向量模型简单、易懂并且在实际应用中非常有效，自从提出后，得到了极大的应用，目前最为成熟的商业搜索引擎（如google、百度等）和开源的超大规模语言模型工具（如Lucene[2]等）仍然采用其作为检索模型中的文档表示方法。统计语言模型是近些年来随着统计概率的迅速发展而出现的，它有着非常坚实的数学基础作为支持，并且非常容易扩展、融入更多的信息特征。尽管两者在方法论上有着不同，一个是基于线性代数的几何变化，一个是基于统计方法的概率分布，两者拥有着很多共同点，其中一个，也是最重要的就是两者认为文档都是在词典空间上进行表示的，也就是说一个文档会形成一个一对多（文档→词）的映射或者表示。

随着人们对于文本认识的发展，人们开始追求除了文本本身更深的理解，从而使得计算机甚至人们能够更好地“理解”文本。一方面，出现了一些比较深入的文本挖掘或者自然语言任务，如自动人工问答。同时人们也开始追求更富有“表现力”的文本表达方式（如何挖掘文本潜在的“语义”）。其中潜在语义分析(Latent Semantic Analysis) [43] 就是一个早期的代表工作，后来的主题模型实际上也仍然延用了其核心思想。潜在语义分析，打破了以往人们对于文本表示的一个思维定式：文本是表示在词典空间上的。潜在语义分析则创新式地引入了语义维度，语义维度是文档集合上信息的浓缩表示，而文档则是在这些语义维度上的一个表示。形象地说，在以往的文档→词映射表示中，引入了一个语义维度，即文档→语义→词。潜在语义本质想法是考虑词与词在文档中的共现，然后通过线性代数的方法来提取出这些“语义”维度，然后实现文档在语义空间上的低维表示。低维表示或者说降维是传统数据分析中的一个重要问题 [5]，其目的就是想要找到一个对于数据更具有表现力、更压缩的表示方法，低维表示有一些好处，如除去噪音的影响、降低数据表示成本等等。

随着概率统计分析在文本建模应用的不断发展，潜在语义分析从线性代数的分析模式被进一步提升到概率统计的分析模式 [37](pLSI或者pLSA) [3]。之前每个语义维度对应一个特征向量，在概率模型中，每个语义维度t则对应到一个词典上$\mathcal{V}$的概率分布，即$\{Pr(w|t)\}_{w\in\mathcal{V}}$；文档对于每个语义维度的权重，对应到概率模型中，将每个文档d表示成一个语义空间上$\mathcal{T}$的概率分布，即$\{Pr(t|d)\}_{t\in\mathcal{T}}$。早期的概率主题模型被称为“aspect model”。将潜在语义分析的概率拓展的可以带来几个好处：1）可以容易引入更多的信息，如先验信息；2）更方便地对模型进行拓展，如引入作者、时间维度；3）使得很多启发式处理手段可以得到理论上的解释，例如在原始的VSM中的权重设定，在概率表示中，则把其考虑成概率分布的估计（平滑）。尽管得到了概率拓展，pLSA仍然不是一个“完整的”贝叶斯模型，原因是在pLSA中，对于对于文档→主题的分布（即）和主题→词的分布（即）仍然看作是参数而不是随机变量。

主题模型，即topic models，在 [11]第一次被显式提出来，实际上这里的主题就是指的之前潜在语义分析中的语义维度。在这里，我们在主题模型的框架详细地下介绍一下什么是主题（topic）。主题是语

[1]截止到2011年6月6日，根据谷歌搜索统计。

[2]http://lucene.apache.org/

[3]对于这个模型有pLSI和pLSA两种叫法，其中pLSI中的I是“indexing”缩写。因为最早LSI是在检索背景下提出的，所以pLSI沿用了之前的叫法。但是后续工作的开展，其实pLSI已经不仅仅不局限于检索问题，所以pLSA更科学一些。

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Figure 1: 几个主题的例子。

料集合依赖的，也就是说给定不同语料，它们背后隐藏的语义是不同的。主题是语料集合上语义的高度抽象、压缩表示。图 1 给出了几个主题的例子，我们可以看到每一个主题对应一个比较一致的语义。在主题模型中，每个主题被表示成一个多项式分布，在实际应用中，我们往往截取前10个关键字作为结果展示。每个主题相对文档表达的内容更加抽象、更加压缩，通过图 1，我们可以发现每个主题是对于文档内容的潜在模式的挖掘。

由于主题模型的良好的数学基础和灵活拓展性，一经推出，马上得到了来自各个领域学者的关注，然后被广泛地应用各种文本挖掘和信息处理的任务中。由于目前和主题模型应用相关的论文有大概近3000篇，我们不能穷举每篇论文，本文将主要围绕以下任务进行展开：关系网络数据挖掘、情感分析、进化文本流分析、科技论文挖掘、社会媒体挖掘以及一些传统的信息检索任务还有自然语言处理中的一些任务。对于方法理论部分，我们将尽量去叙述一些核心问题，并且忽略繁琐而又枯燥的推导细节。

2 主题模型方法概述

在本章，我们将对于主题模型的一些关键方法和知识进行介绍，这对于理解后续章节的内容将会起到一定的帮助作用。

2.1 基础知识

我们首先介绍以下概念和方法：

- 统计语言模型。首先成功利用数学方法解决自然语言处理问题的是语音和语言处理大师贾里尼克(Fred Jelinek)。当时贾里尼克在IBM 公司做学术休假(Sabbatical Leave)，领导了一批杰出的科学家利用大型计算机来处理人类语言问题。统计语言模型就是在那个时候提出的。统计语言模型是研究一个文本序列的生成概率的问题。例如，给定一个序列$\mathcal{S} = w_1, w_2, ..., w_L$，我们想要估计或者计算$Pr(S)$。

- N元语言模型。统计语言模型希望对于一个给定的序列能够输出一个概率，即$Pr(S)$。一个最常见的模型就是N元语言模型，即$Pr(S) = P(w_1)P(w_2|w_1)...P(w_i|w_{i-1}, w_{i-2}, ..., w_{i-N+1})...$ $P(w_L|w_{L-1}, w_{L-2}, ..., w_{L-N+1})$。理论上，随着$N$的增加，模型的表现能力将不断增加，但是同时模型的复杂度也同时增加，在实际应用中，往往受限于语料的有限性。所以目前最常用的N，仍然是1、2和3。当$N = 1$，这时模型的估计变成$Pr(S) = \prod_{i=1}^{L} P(w_i)$。一元模型由于其简单和有效性得到极为广泛地应用。其中图 2给出了一个一元语言模型的概率图表示方法。一元语言模型一般表示成一个多项式分布。

- 一元混合语言模型。常见的混合模型如高斯混合模型，在高斯混合模型中，每个组件是一个高斯分布。在一元混合语言模型中，每个组件是一个一元语言模型。给定一个文档集合，我们首先假

设它们对应K个堆，然后**整个文档集合有一个对应K个堆的分布**，然后根据这个分布来选择一个堆，每一个堆对应一个一元语言模型。当选定堆了后，然后由该堆对应的一元语言模型生成对应的文档。图 3给出了一个一元混合语言模型的示意图。

- pLSI模型 [37]。pLSI是对于潜在语义分析的概率拓展。它假设整个文档集合对应有K个主题，然后**每个文档有一个文档特定的对于K个主题的分布**，文档中的每个词生成分为两步：a) 首先根据文档的主题分布选择一个主题；b) 然后由该主题对应的一元语言模型生成该词。图 4给出了一个pLSI的语言模型的示意图。

- 概率图模型 [5]。概率图模型是一种将概率和传统的图结构结合起来的表示方法。通过图结构的清晰表示，我们可以更好地理解概率的分布，概率图模型是一个非常理论深入而又内容广泛的学科，在此我们仅仅讲述一些对于理解主题模型有帮助的内容。图 5展示了一个LDA的概率图模型的表示方法。其中圆圈表示变量，如果被涂成灰色则表示该变量已知，否则表示该变量未知。变量间的箭头方向表示概率依存关系。其中的方框表示内部结构的重复，方框右下角的脚标表示重复的次数。实际上我们可以发现LDA体现出一种层级关系，它是一种层级贝叶斯模型。

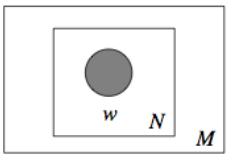

Figure 2: 一元语言模型示意图。

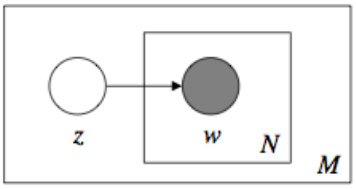

Figure 3: 一元混合语言模型示意图。

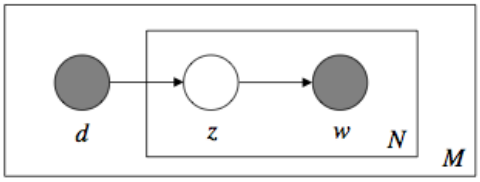

Figure 4: pLSA示意图。

下面我们来比较以下图 2, 3和4：1）一元语言模型假设整个文档集合仅有一个语言模型；2）一元混合模型假设整个文档集合有K个一元语言模型，但是只有一个全局的概率分布用来选择这K个语言模型的某一个；3）pLSI假设整个文档有K个语言模型，对于每一个文档有一个文档特定的概率分布可以为用来选择这K个语言模型的某一个。这三种模型体现了人们对待文档语言建模过程的探索和不断增加模型表现力的过程。实际上，LDA仅仅是pLSI的一个贝叶斯的拓展，其生成过程与pLSI本质上是相同的。

2.2 LDA的简要介绍

LDA [11]是Latent Dirichlet Allocation的简称，是一种层次的贝叶斯模型。为了方便以后的叙述，我们下面详细地介绍一些形式化的基础。我们假设整个文档集合一共有T个主题，每个主题z被表示成一个词典$\mathcal{V}$上的一元语言模型θ_z，即词典上的一个多项式分布。我们进一步假设每个文档d对应这T个主题有一个文档特定的多项式分布ϕ_d。一个文档的生成过程如下：

- 采样$\phi_d \sim Dir(\alpha)$；
- 对于文档d中的每一个词w，我们：

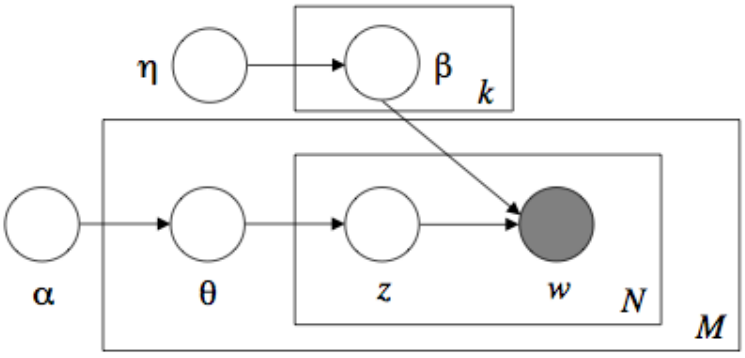

Figure 5: LDA的图模型表示。

– 采样一个主题标签$z \sim$ Multi(ϕ_d)；

– 生成对应的$w \sim$ Multi(θ_z)。

其中$\theta_{(\cdot)} \sim Dir(\beta)$。在主题模型中，我们需要关注两种概率：1）一种是文档~主题的分布；2）一种是主题~词汇的分布。基于这些基础，我们下面讨论一下理解LDA的一些重要知识：

- 贝叶斯层级模型。pLSI和LDA最大的区别就是LDA是"完全的"贝叶斯模型，而pLSI不是。在pLSI中，$\phi_{(\cdot)}$和$\theta_{(\cdot)}$都是参数，他们会随着文档的数目增加和主题数目的增加而增加，而在LDA中，则把$\phi_{(\cdot)}$和$\theta_{(\cdot)}$看做随机变量，他们也是来自于对应的超参数控制的随机变量。
- 可交换性。可交换性指的是，给定一个有限长度的变量序列$z_1, z_2, ..., z_N$，对于该序列的任何一个置换π，有$P(z_1, ..., z_N) = P(z_{\pi(1)}, ..., z_{\pi(N)})$。如果一个序列中的变量满足独立等同分布(i.i.d.)，那么该序列一定是可交换的，但是反之未必成立。根据De Finetti定理[4]，可交换性指的实际是**条件**独立等同分布，即给定决定这些变量分布的参数以及分布函数后，这些变量的分布才满足独立等同分布。
- 隐含变量模型。LDA本质上是一种隐含变量模型，这也是名字中包含的"Latent"主要意思。隐含模型指的是，引入一些中间隐含变量如主题标签，这些变量在实际数据中是虚无的，并且不可见。引入隐含变量的主要目的是使得模型的描述和推导更加清晰、简单。
- 概率共轭。根据贝叶斯定理$P(\theta|\mathcal{X}) \propto P(\mathcal{X}|\theta)P(\theta)$,概率共轭指的是后验概率$P(\theta|\mathcal{X})$和先验概率$P(\theta)$有着同样的概率形式，而此时$P(\mathcal{X}|\theta)$的概率形式和$P(\theta)$的概率形式就叫做一个共轭对。在LDA中，Dirichlet分布和多项式分布就是一个共轭对。

2.2.1 先验设定

在LDA中，有两组先验，一种是文档~主题的先验，来自于一个对称的$Dir(\alpha)$；一种是主题~词汇的先验，来自于一个对称的$Dir(\beta)$。[30]给出了一些经验性α, β取值方法，其中$\alpha = \frac{50.0}{|\mathcal{T}|}, \beta = 0.1$。[29] 证明了实际上pLSI是一种LDA模型的MAP（maximum a posteriori）估计，当先验采用的是对称的Dirichlet概率。实际上由于LDA采用了一个完全的贝叶斯途径，对于未知文档、词汇的估计更加准确，同样pLSI也可以通过采用MAP的估计方法来引入先验。在一般的文本挖掘任务中，这两种模型的实际效果应该接近，但是LDA显得更加灵活、理论基础更坚实，特别是当考虑的文本挖掘问题特别复杂，LDA更加容易实现结合多个模型组件在一个模型中。

[89] 考虑了这两种先验的设定方式：1）对称的；2）非对称的。结合两种先验与两种不同的先验设定方法，我们可以得到以下四种组合：

- AA：文档~主题分布和主题~词汇分布都采用非对称的先验；
- AS：文档~主题分布采用非对称的先验，而主题~词汇分布采用对称的先验；
- SA：文档~主题分布采用对称的先验，而主题~词汇分布采用非对称的先验；
- SS：文档~主题分布和主题~词汇分布都采用对称的先验。

实验发现采用*AS*的方式可以更好地提高LDA对于文本建模的能力。

此外还有一些非参数贝叶斯的方法引入更复杂的先验信息，例如Dirichlet Process [83]和Pitman-Yor process [85]，我们将会在后续章节遇到相关内容时再详细介绍。

2.2.2 模型求解

对于标准的LDA，模型求解是一个非常复杂的最优化问题，很难有精确求解的方法。所以我们采用不精确的模型求解方法，大概有三种不精确的方法进行模型的求解：一种基于Gibbs采样的方法 [30, 34]，另外一种是基于变分法的EM求解 [11]，还有一种是基于期望推进的方法 [62]。一般来说，Gibbs采样的方法推导起来更为简单而且求解效果也不错，所以在本文中，我们将采用其作为主要方法进行介绍。我们首先简要介绍一下Gibbs Sampling。Gibbs sampling是一个Metropolis-Hastings算法[5] [5]的特例。其基本思想是，给定一个多维变量的分布，相比于对于联合分布积分，从条件分布中采样更简单。假设我们想要从一个联合分布概率$P(x_0, x_1, ..., x_n)$中获得K个样本$X = \{x_0, x_1, ..., x_n\}$。该方法的两个通用步骤：

[4] http://en.wikipedia.org/wiki/De_Finetti's_theorem

[5] http://en.wikipedia.org/wiki/Metropolis–Hastings_algorithm

- 随机地初始化每个变量获得$X^{(0)}$；
- 对于每个样本$X^{(i)}, i = 1, ..., K$，对于每一维的变量x_j，我们从条件分布概率$P(x_j^{(i)}|x_1^{(i)}, ..., x_{j-1}^{(i)}, x_{i+1}^{(i-1)}, ..., x_n^{(i-1)})$采样得到$x_j^{(i)}$。

在LDA的基于Gibbs采样的模型求解中，往往采用“Collapsed Gibbs Sampling”方法 [30]。基本思想就是不考虑$\theta_{(\cdot)}$和$\phi_{(\cdot)}$两组随机变量，而只考虑对于每个词的主题标签z的推理求解。这使得计算和推导大大简化。我们在附录中给出一个非常详细得推导供大家学习。

在实际应用中，由于文本数量的巨大、文本流的时序特征，有一些研究开始关注LDA的快速推理算法 [72]、在线学习 [36]、文本流的推理算法 [101, 4]、分布式学习 [65, 49, 100, 3]。这些研究将会使得对于LDA模型求解的效率大大得到改善，同时会将适应文本的时序特征，更符合实际文本流的特征。

2.2.3 模型评估

对于主题模型的评估是一个一直以来被被关注，但是并未被很好解决的问题。主要原因，我认为是由于LDA本身是一种文本表示方法，而往往直接很难评估一个表示方法的好坏。目前有的方法，大概可以分为以下三类：

- 基于Perplexity的方法。Perplexity经常被用在语言模型中，是用来衡量语言模型对于测试语料的建模能力的“好坏”，即似然的大小。形式化，可以根据如下的式子计算测试集合上的Perplexity的大小

$$perplexity(\mathcal{D}_{test}) = \exp\{\frac{-\sum_d \log(P(\mathbf{w}_d))}{\sum_d N_d}\} \tag{1}$$

  当一个新的主题模型被提出后，往往通过和标准的LDA进行测试集合上面的Perplexity的对比，如果得到了更小的perplexity，就认为此模型的建模效果更好一些。但是基于Perplexity比较的一个最大的问题就是，一个可以获得较低Perplexity的模型未必一定可以学习得到人们认为很好的主题词。

- 基于高概率主题词。每一个主题最终的表示形式是一个一元语言模型，我们可以根据每个主题内部词汇概率的高低来进行主题词汇的排序。得到的这些高概率的主题词汇往往可以直接作为输出展示给用户，让用户来进行评估。 [24]第一次系统地构建了语言模型的人工评测方法。主要考虑两个维度：第一个维度是主题内部的一致性，具体方法就是对于一个主题，首先选择具有最高概率的5个主题词，然后随机地添加上一个在当前主题下有着较低概率但是在其他主题内部具有较高概率的词汇；第二个维度是文档内部主题分布的一致性，具体方法就是对于一个文档，我们首先计算得到概率最高的三个主题，然后我们随机地添加其他一个主题。对于这两个任务，都是人工地评估检查出随机添加的词或者主题的难易程度。 [90]重新检查了一些评估主题模型的方法，发现实际上之前的方法是不够准确来评估主题模型的，又提出了两种新的评估方法。 [19] 基于 [90]的工作又重新进行了拓展，提出了新的估计测试文档集合的似然的方法。 [1] 提出了对于同一个主题模型学习得到的主题进行重要度排序，基本思想是不是所有学习得到的主题都是等同的，主题排序可以自动地发现更加“重要”的主题。

- 利用其他任务的效果来间接评估。对于一个语言模型，我们可以通过模型求解，最终可以得到两种概率第一种就是文档~主题的概率，第二种就是主题~词汇的概率。这两种概率都可以在一些其他任务中使用，如文档相似度的计算，主题间相似度的计算等等。通过这些间接任务来比较主题模型的好坏有一个问题就是，对于不同的任务，每个主题模型的优点可能不一样，所以往往一个任务不能衡量出来两个主题模型之间的孰好孰坏。这种方法比较合适在一个特定任务中提出的主题模型的好坏程度，如情感分类。

2.2.4 主题数目的确定

对于基于LDA的主体模型，一个非常重要的问题就是如何确定主题数目。而主题数目的确定实际上是一个模型选择的模型，就像之前的刚刚讨论过的模型评估部分，由于模型评估本身是一个非常困难的问题，所以对于主题模型中的模型选择问题仍然是一个非常困难的问题。目前的方法大概有两种：

- 经验设定。在一些文本挖掘工作，研究人员往往通过反复地调试或者枚举主题的数目来观察实验效果的好坏，例如观察高概率的主题词汇的好坏、语义是否一致等等。这种方法虽然是启发式

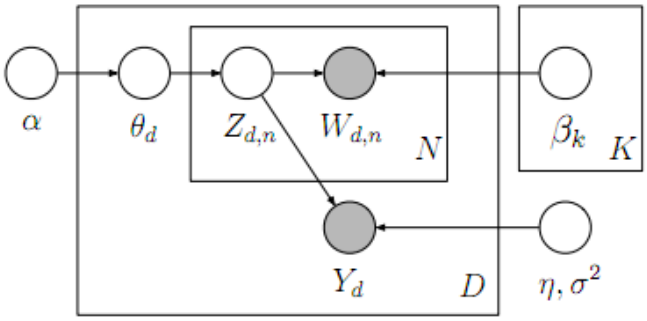

Figure 6: Supervised-LDA的图模型表示。

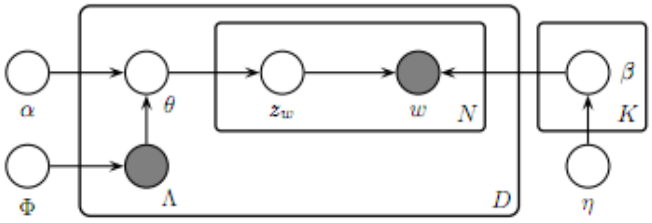

Figure 7: Labeled-LDA的图模型表示。

的，但是往往实际中简单易行，因而是最常用的方法。实际上，对于大部分的研究工作都默认没有很好设定主题数目的方法，而直接采用这种经验性的方法来设定主题数目。

- 基于Perplexity的确定方法。在模型评估中，我们讲到了如果一个主题模型在测试预料集合上获得了较低的Perplexity，那么我们就认为这个主题模型具有更好的模型表示能力。但是还是那个老问题，Perplexity的大小和主题模型在实际中任务中的好坏在理论上并不能有直接的练习。所以在一些具体的文本挖掘工作，这种方法并不被采用，而是直接使用经验设定的方法；但是对于机器学习社区，往往通过此方法来验证一个新提出的主题模型的好坏。
- 使用非参数的贝叶斯方法对主体模型进行拓展。我们将在下一节对于这个方法进行详细地介绍。

2.2.5 基于LDA的主题模型变形

随着LDA的推出，截止到今天，谷歌学术搜索的显示的引用论文数高达近2900次。大批的学者对于基本的LDA进行了各种变形和拓展，还有在各种任务上的应用。在此，我们试图将这些模型的变化和拓展进行一个粗略的总结，借此来更好地了解主题模型的发展。

- 打破原有的可交换的假设。在原始的LDA模型中，可交换的思想应用到两个地方：1）文档集合内部，文档之间是可交换的；2）文档内部，词与词之间是可交换的。但是这些可交换假设只是在建模过程中为了减少模型的复杂度而设立的，实际上这些假设限定了模型的表示能力。为此有一些

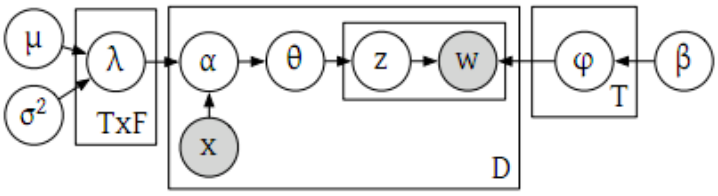

Figure 8: Dirichlet-multinomial Regression (DMR) topic model的图模型表示。

模型进而对于这些可交换假设进行松弛：如 [9, 46]引入了主题之间的关联、 [23] 引入了文档之间的关联、 [96] 考虑了词与词之间的顺序关系。但是一旦打破原有的可交换性之后，模型的复杂度将显著增加，所以必须考虑的一个问题是，增加这些复杂度后，得到的新主题模型能否在实际数据上获得较好的效率。另一个折中的方法，是在设计模型过程，我们并不打破这些可交换假设，而是最后在优化公式基础上增加一些改进。如增加先验信息来区别不同的主题 [51]、增加规则化因子 [56]来引入一些关联信息和先验信息。

- 基于非参数贝叶斯方法的变形。其中比较有代表性的就是基于Dirichlet Process [83]的变形。基于Dirichlet Processd的方法可以自动地学习出来主题的数目 [84]。那么就有这样一个问题，是不是基于Dirichlet Process的方法就可以解决主题模型中的自动确定主题数目这个问题？答案是两面的：1）在一定程度上解决了主题模型中自动确定主题数目这个问题；2）代价是我们必须更小心地去设定、调整一些其他参数的数值，例如超参数的设定；3）基于Dirichlet Process的方法往往在实际中运行复杂度更高，由于在学习过程中总有主题消亡和诞生，这将使得代码变得复杂和难以维护。所以在实际中，往往取一个折中，看看自动确定主题数目这个问题到底对于整个应用问题的需求到底有多严格，如果经验设定就可以满足的话，那么就不用采用基于非参数的方法；但是对于一些由于问题，如为了引入一些先验知识或者结构化信息，在这种情况下，往往非参数的方法是优先选择，例如树状层次的主题模型 [7]和有向无环图结构的主题模型 [46]。还有一些其他非参数的贝叶斯方法，如基于Pitman-Yor Process的方法 [85, 79]。

- 从无结构信息→结构化或者半结构化的信息。标准的主题模型是一种无监督学习的方法，只需要输入主题数目和一个文档集合的所有文档，模型就能够进行主题的自动学习。对于一些特定的应用问题，例如文档分类，当我们已经有了部分训练数据后（例如，文本的类别标签等等），那么如何在主题模型中如何使用这些训练数据的信息。如果再把这个数据的要求进行松弛，那么就是考虑如何融入文档附加信息。纯文本往往把文本直接看做一个个词袋子，除了词袋子，没有任何附加信息。随着文档数据格式的丰富和互联网数据的发展，传统的纯文本观点往往不适合，容易忽略了一些很重要的其他特征，例如时间标签、类别标签、用户提供的标签等等。所以在主题模型中，一个非常重要的方向就是如何融入主体模型中的这些有用的特征。在所有这些特征中，作者实体、时间、网络结构、标签信息等等都是非常典型的特征，得到了学者们的广泛关注。这种信息的融入可以看做某种监督学习或者弱监督学习的方法，实际上是一种如何在主题模型中融入结构化的信息。 [10] 提出同时关联主题学习和响应变量（英语，response variable，它可以是用户对于一个文档的评分、文档的类别标签、或者其他数值变量）的生成。如图 6所示，其基本想法就是关联起每个文档的响应变量和对应该文档的词汇的主题标签向量。这种关联起额外信息和主题的方法在之后的工作被经常使用，例如关联时间和主题。另一个相关工作 [75] 则是直接建立起离散的标签与主题之间的映射关系，对于每个标签，我们限定其有倾向性地映射到一小部分的主题，而不是使用之前每个文档~主题的均匀先验分布。 [60] 是一个比较经典的工作，进一步将 [75]的想法完善。如图 8所示，它通过引入一个log-linear先验在文档~主题分布上面，可以将主题抽取关联到文档的“Metadata”等等多种特征（例如作者、时间等等），如果当我们仅仅考虑相应变量作为该文档的特征时，实际上 [10]可以看做 [75]的一个特例。

  对应于pLSA，当我们拥有了一些结构化的先验信息，我们可以通过使用Dirichlet先验 [51]来加入这些结构化的限制，或者我们可以通过使用规则化的方法对于目标函数施加限制 [56]。对于第一种方法，基于LDA的模型同样可以使用，实际上我们在之前已经叙述过这个问题，基于贝叶斯方法的主题模型可以更灵活地使用多种多样的先验信息。对于第二个方法，由于规则化因子的出现，我们需要使用变分法的EM算法进行求解 [99, 80]，而Gibbs采样的方法往往很难处理这种规则化因子，需要一些启发式的方法 [82]。实际上对于LDA，当我们有足够先验信息来推理一个词汇的主题标签，我们可以在Gibbs采样中直接设定该词的主题标签而不进行标签的采样 [47]。

3 基于主题模型的应用

由于主题模型的方法几乎被应用到了所有的文本挖掘和信息处理领域，在此我们仅仅详细地回顾一下几个受关注特别多的领域：情感分析、时态文本流分析、社会媒体分析、学术文章挖掘、网络结构化的数据挖掘和一些其他应用。

3.1 主题模型在情感分析中的应用

情感分析 [69]（英语为opinion mining、sentiment analysis）是近些年来新出现的一个研究方向。其基本任务就是从用户生成的包含观点和意见的文本中抽取出这些观点和意见，然后生成情感摘要、进行情感

分类、自动构建情感词典等等情感分析任务。本质上来讲，情感分析是一个应用（或者说语料）驱动产生的新方向。随着网络的发展，人们可以很方便地在线购买各种产品、服务，同时又可以随时发表自己对于这些产品或者服务的观点和建议，而随后的用户往往又会参考之前的用户的评论，周而复始，这种消费模式以及用户生成的数据使得情感分析对于工业界、商业界格外重要，对于用户本身也是一个非常有意义的文本挖掘任务。

主题模型在情感分析最主要的任务就是学习出来用户讨论、用户评论中的内容主题。在情感分析中，每一个topic通常被称为aspect，为了叙述的方便和统一，在此我们将沿用主题来作为aspect的翻译。在情感分析分析中，会将词汇区别成为情感词汇（如great, good等等）和主题词汇（如food, drink等等）。对于一些在线论坛网站，往往有一些预设的可“ratable”的主题，并且附带了一些打分机制，如0-1打分（满意、不满意），分级打分（1∼3或者1∼5）。这里为了叙述的方便，仍然沿用“ratable”这个英语单词，不进行翻译。

从用户生成的评论数据中提取主题大概分为三个方法:

- 无监督主题抽取。标准的主题模型如pLSI [37]和LDA [11]都可以直接应用到用户评论中，进行无监督的主题抽取。但是 [87]发现传统的语言模型如LDA从用户评论中学习得到的主题更倾向对应某个品牌而不是对应可“ratable”的主题。 [87]分析一个非常重要的原因就是之前的主题模型考虑的是文档内部的共现，基于文档这个“大背景”(context) 学习得到的主题往往是对应品牌而不是可以“ratable”的主题。为了解决这个问题， [87]考虑了一个多级的背景主体模型：词∼句子∼段落∼文档，通过这样经过试验验证，这种多粒度的主题建模方法可以更好地抽取“ratable”的主题。后续的一些工作也进一步沿用这种思想，但是使用了相对较为简单的模型：1）把单个句子当做一个文档，然后使用标准的LDA [17]；2）对于一个句子采样一个主题标签，而不是对于一个词采样一个主题标签 [40, 44]。这两种方法的思路都是硬性地缩小“背景”，但是在实际应用中取得了不错的效果。

- 弱监督学习。在目前的在线评论网站中，除了用户的文字评论外，往往有很多结构化或者半结构化的信息，例如用户对于一个商品的打分、用户对于商品添加的一些标签。在之前基于用户生成数据中抽取主题的工作中，遇到的一个关键问题就是如何将学习得到的一个主题X对应到实际中对于分析用户情感有帮助的主题Y。一些研究工作人员利用在线评论网站中结构化和半结构化的信息来帮助抽取更具有实际意义的主题。具体方法大概有以下两种：

  - 利用模型的先验信息。无论在pLSI中还是在LDA中，我们都可以通过设定先验信息的方法来帮助抽取主题。其基本思想是，对于一些主题，我们有一些比较准确的先验信息，例如，对于一些汽车产品，我们可以从维基百科中提出它的各个特征的描述，然后将它们训练当做先验信息。例如借助维基百科中结构化的文本来帮助抽取博客中的主题 [51]。
  - 融入结构化特征到主题模型中。另外一种方法，是我们想办法把一些具有结构化信息的特征融入到主题模型中。具体来说，我们同时关联两个生成过程，一个就是文档中词的生成，另一个就是这些结构化特征的生成。通过这种关联，我们可以间接地引入这些结构化信息到主题学习过程中 [86, 16]。

- 联合抽取主题和观点。之前的方法只是从抽取主题的角度去考虑的，在每个主题内部，往往会把主题词和情感词混合在一起，不加区分。有一部分工作考虑区分开情感和主题两种不同类型的词汇，进而同时抽取主题和观点。这些工作大概可以有以下两个途径：

  - 对于所有主题，设定一系列共有的情感语言模型（通常包括褒义、贬义和中性三种情感标签） [57]。也就是说在文档中的每个词首先采样一个二元变量，来决定是主题词汇还是情感词汇。如果是主题词汇，之后的生成过程和在标准的语言模型中一样；如果是情感词汇的话，就从情感语言模型中选择一个合适的情感标签。其中图 9
  - 对于每个主题，设定一个特定的情感模型，在这种情况下，往往不再细分语言模型的极性，而只是学习得到一个主题特定的情感词的语言模型。

  在联合抽取主题和观点的时候，最难的问题就是如何识别情感词汇，换句话说，如何分开主题词汇和情感词汇。第一种方法就是利用一些已经有的情感词典作为先验信息，然后在主题模型训练过程中再发现新的情感词汇；另一种方法就是引入一些监督学习的组件到主题模型中，这样我们可以同时兼有监督学习和主题模型的优点：对于监督学习来说，它比较适合区分主题和情感词汇，但是不适合用来聚类语义相近的词汇；对于主题模型来说，它比较适合聚类语义相近的词汇，但是完全无监督的主体模型不适合用来区分主题词汇和情感词汇。 [107]通过结合最大熵组件和主题模型，使得该模型同时具有了监督学习和主体模型的优点。

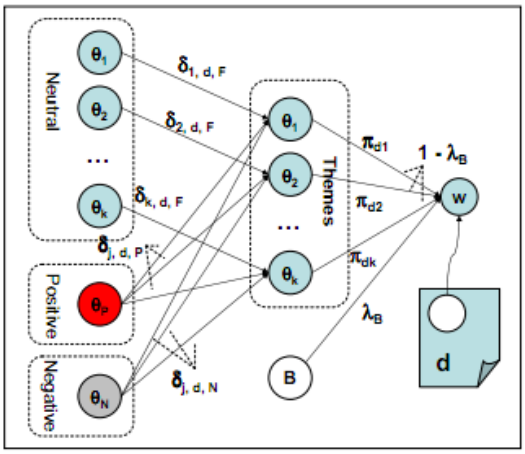

Figure 9: 一个主题-情感联合模型的示意图 [57]。

Service		Room Condition		Ambience		Meal		General
Aspect	Opinion	Aspect	Opinion	Aspect	Opinion	Aspect	Opinion	Opinion
staff	helpful	room	shower	room	quiet	breakfast	good	great
desk	friendly	bathroom	small	floor	open	coffee	fresh	good
hotel	front	bed	clean	hotel	small	fruit	continental	nice
english	polite	air	comfortable	noise	noisy	buffet	included	well
reception	courteous	tv	hot	street	nice	eggs	hot	excellent
help	pleasant	conditioning	large	view	top	pastries	cold	best
service	asked	water	nice	night	lovely	cheese	nice	small
concierge	good	rooms	safe	breakfast	hear	room	great	lovely
room	excellent	beds	double	room	overlooking	tea	delicious	better
restaurant	rude	bath	well	terrace	beautiful	cereal	adequate	fine

Figure 10: 一个基于主题的摘要示意图 [107]。

当我们抽取主题和情感词汇之后，我们通常而可以利用主题模型的输出来进行：

- 文档情感分类。我们可以计算每个文档整体的情感倾向 [57, 47]，由于在主题模型中，无论是主题还是情感，通常都是表示成一元语言模型，因此我们可以使用标准的概率方法来计算一个文档的情感倾向，或者我们直接在模型中设定情感标签 [57, 47]。还有一些模型设定文档特殊的情感标签或者词特定的情感标签，更容易来完成这个任务。
- 生成基于主题的情感摘要。主题模型的结果可以直接用作一个情感语料集合的摘要，目前大部分的工作都是生成基于主题的情感摘要，也就是按照主题组织情感词汇和主题词汇。 [107]进一步提出使用主题特殊的情感词汇进行情感摘要。主题特殊的情感词汇与主题词汇之间有更强的语义关联。图 10 给出了一个基于主题的情感摘要。
- 自动构建情感词典。学习得到的情感词汇可以作为情感词典的候选，我们还可以进一步附加上主题背景信息来丰富这个词典的表达能力。

3.2 学术文章挖掘

学术文章挖掘是主题模型的一个重要应用，通过对于学术文章的挖掘来进一步理解学术文章的发展、进化，这对于了解之前的科技进步、未来的科技发展趋势都是非常有意义的。我们在此对于这些研究进行一个粗略的分类。

3.2.1 对于作者进行建模

Author-Topic Model [77]是从作者的角度来考虑文档中主题的生成的。

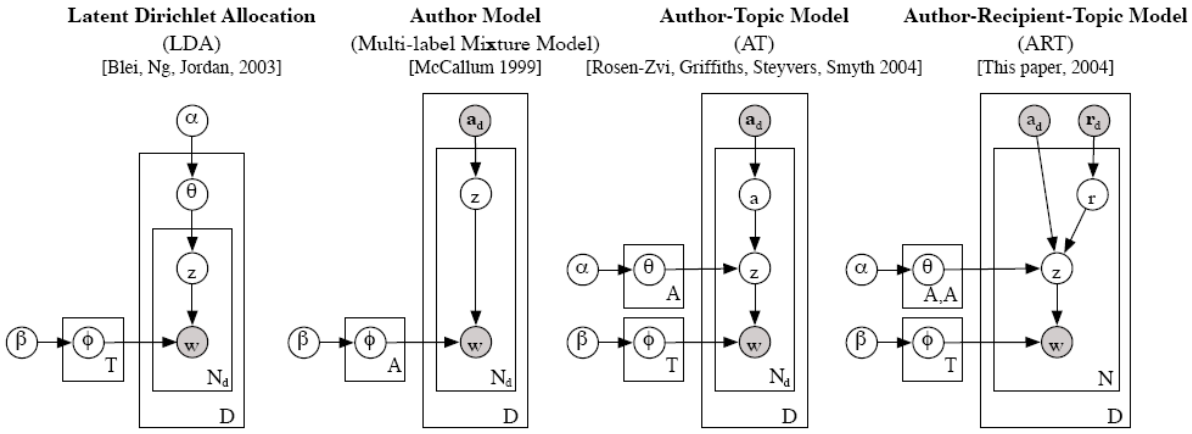

Figure 11: 几种和作者相关的主题模型示意图。

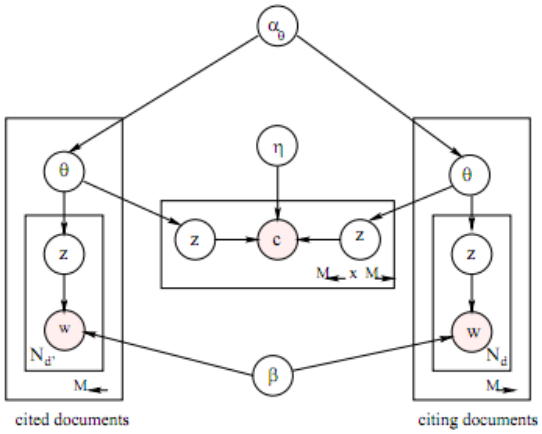

Figure 12: Pairwise Citation LDA model示意图 [64]。

对于学术文章语料集合，主题模型 [11]并没有考虑到文章的作者，实际上是把所有的作者都看做完全等同的；单一主题的作者模型 [53]，任何一个作者都对应着一个特定的语言模型，但是这实际上是一种非常强烈的限制；实际上Author-Topic Model [77]是对上述两种模型进行的折中考虑。对于每一个作者不再限定该作者只能对应一个主题，而是每个作者对应于一个主题上的分布，同时文档~主题的分布随之消失，即被作者~主题的分布而取而代之；同时，我们可以看到所有的作者共享一个主题的集合。

[54] 进一步拓展了Author-Topic Model，通过引入接收者元素，即任何文本任何一个词都是由一个（作者，接收者）对共同决定的。之前在是每个作者对应一个主题上的多项式分布，现在一个作者，接收者）对对应一个主题上的多项式分布。

图 11对比展示了四种模型，在实际应用中，Author-Topic Model [77]是最常使用的同时对于作者和主题进行建模的方法。

[66] 从另外一个角度来考虑“作者”的角色，即论文评审者的角度，在这个工作，作者假设每个作者是有一些虚拟的“个性化”主题分布，而不是像Author-Topic Model中，每个作者仅仅有一个主题分布。

3.2.2 对于学术引用建模

学术文章不同一般的新闻报道、博客文章等等，一般都需要附带非常详细的参考文献，这些参考文献往往都与原始的论文有着非常紧密的联系。实际上这是一些显式的论文与论文之间的联系。如何对于这些链接关系进行建模，一直是学术文章研究的一个热点和关键问题。

[64] 是第一个提出来同时对于主题和参考引用进行建模的文章。其基本思想就是首先分别对于文本采用之前标准的主题模型的生成方式，例如LDA [11]，然后对于任何一对具有引用关系的文档对，根据文档~主题分布的相似性生成引用链接关系。这种处理方式将文本和链接之间的关联通过主题分布建立

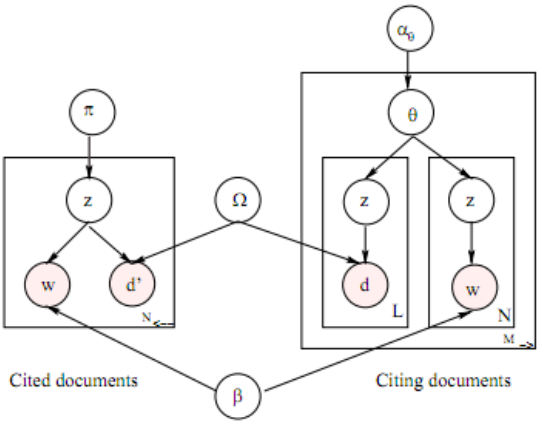

Figure 13: Citation Link-PLSA-LDA 示意图 [64]。

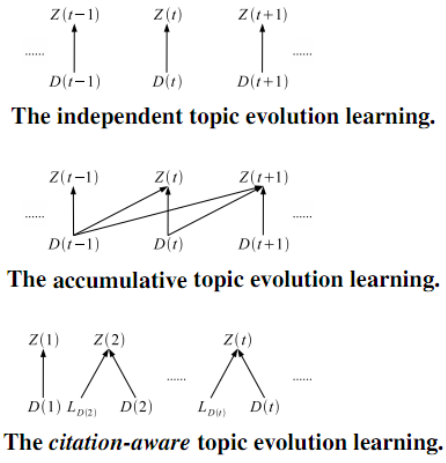

Figure 14: 不同的主题进化建模思路示意图 [33]。

起来。图 12给出了这个模型的示意图。这个模型的复杂度特别高，所以 [64]又进一步提出了一个改进的模型，进一步简化了被引用文档的文本生成模式，见图 13。

自动学习或者发现学术文章中的主题进化是一个非常重要的问题， [33]考虑了如何拓展标准的LDA [11]来加入引用信息，来更好地进行主题进化建模。假设给定一个时间点t，对应有一个主题空间$\mathcal{Z}_t$，主题进化的主要任务就是对于每一个时间点t学习得出对应的主题空间$\mathcal{Z}_t$，同时发现相邻时间点间主题的变化，还有同一个主题在不同时间点内的变化。其中最主要的问题就是如何估计出每个每一个时间点t学习得出对应的主题空间$\mathcal{Z}_t$。 [33] 讨论了三种主要的主题进化建模思路：

- 第一种就是时间独立的主题进化模式。即每个时间点的主题仅仅与当前时间点内部的文章集合$\mathcal{D}_t$有关系，而与周围的没有关系；
- 第二种就是时间累积的主题进化模式。即每个时间点的主题仅仅与当前时间点以及之前时间点之前的所有文章集合有关系；
- 第二种就是引用敏感的主题进化模式。即每个时间点的主题仅仅与当前时间点文章$\mathcal{D}_t$及之前被这些文章引用的文章$\mathcal{L}_{\mathcal{D}_t}$有关系。

[33]通过实验发现这种基于引用的方法可以更好地进行主题进化的学习。

3.2.3 对于学术语料的分析

基于LDA [11]对于学术语料的分析是一个主题模型在学术语料挖掘中经常进行的任务。其中 [30]是最早的一个基于主题模型来进行学术语料分析的工作。它主要分析了那些主题是“热”主题，哪些主题

1. For each topic $t = 1, \ldots, T$,
   (a) draw $\phi^t \sim \mathrm{Dir}(\beta)$
2. For each tweet $s = 1, \ldots, N$,
   (a) draw $\theta^s \sim \mathrm{Dir}(\alpha)$
3. for each tweet $s = 1, \ldots, N_u$
   (a) for each word $n = 1, \ldots, N_{u,s}$
      i. draw $z_{u,s,n} \sim \mathrm{Multi}(\theta^s)$
      ii. draw $w_{u,s,n} \sim \mathrm{Multi}(\phi^{z_{u,s,n}})$

Figure 15: LDA：Tweets的生成过程.

是“冷”主题，然后分析了这些主题随着时间的发展的强度变化，其中主要使用了平均的文档∼主题的分布来计算强度。 [32] 使用了LDA来分析“ACL Anthology”[6] 从1978年到2006年之间的语料，主要分析了以下几个问题：

- 哪些主题变得更加流行。
- 哪些主题热度下降。
- 计算语言学会议是变得越来越应用了吗？
- 国际三大主流自然语言处理大会ACL、EMNLP和COLING之间的差异。

[109] 分析了作者如何影响主题进化的这个问题。作者试图去回答一个问题“给定一个新主题，那么这个新主题从何而来，也就是说从哪些主题进化而来呢？”。 [109]认为主题的进化是作者与作者之间的交互而带来的，提出了一个马尔科夫模型来对于这种基于作者交互的话题交互进行建模，并且应用这个模型分析了一些有意思的问题，如对一个给定的主题进化，是哪些作者主导了这次变化呢。

[52] 提出使用主体模型进行细粒度的文章影响力的分析，主要是提出了一些计算度量来对学术研究进行分析；1）引用数目；2）主题影响因子；3）主题传播和多样性；4）主题优先性；5）主题寿命；6）主题迁移。通过这些度量在30万左右的学术文章语料集合上的使用， [52]得到了一些有意思的结论。

综合上述这些研究，我们发现基于主题模型的方法对于学术语料提供了一种细粒度的分析。主要分析集中在以下几个维度：

- 时间维度。主题如何进化、发展。每个主题随着时间的强度如何变化。
- 内容维度。内容维度主要是考虑相似性和多样性，哪些主题之间比较相似，哪些文章覆盖了较多主题的内容。
- 作者维度。作者对于主题的兴趣、作者如何影响主题进化。

3.3 社会媒体中的主题提取研究

3.3.1 主题模型在Twitter中的初步尝试

Tweets作为一种新型的社会媒体数据，与传统的文档集合（新闻报道和学术文章）有着显著的不同，体现在：噪音大、篇幅短、更新速度极快、数量巨大等等。目前大部分工作都是将已有的主题模型直接应用在Tweets上或者稍加变形。其中最大的难点就是文档长度太短，之前有过相关研究已经表明传统的主题模型（如LDA）在短文档上的效果不好 [51]。

目前在Tweets数据上主要应用的模型是LDA [11, 30]、Author-Topic Model（ATM） [77]和Twitter-LDA [105]，这三个模型非常相似，但都有一些各自的不同，图 15, 16和 17介绍了这三个模型在Tweets上的生成过程[7]。

下面我们来分析三种模型之间的联系和不同。

[6] http://aclweb.org/anthology-new

[7] 为了便于叙述，所有生成过程都用英语叙述。LDA和ATM的生成过程和原始论文有些差异，这里我们只给出了在Tweets上的具体生成过程，而不是通用文档集合上的生成过程

1. For each topic $t = 1, \ldots, T$,
   (a) draw $\phi^t \sim \mathrm{Dir}(\beta)$
2. 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. for each word $n = 1, \ldots, N_{u,s}$
         A. draw $z_{u,s,n} \sim \mathrm{Multi}(\theta^u)$
         B. draw $w_{u,s,n} \sim \mathrm{Multi}(\phi^{z_{u,s,n}})$

Figure 16: ATM：Tweets的生成过程.

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 $y_{u,s,n} = 1$

Figure 17: Twitter-LDA：Tweets的生成过程。

- 首先，我们可以发现实际上ATM是首先将Tweets聚合形成一个大的用户“文档”，本质上仍然是直接应用了LDA[8]。但是ATM通过聚合可以对用户兴趣进行刻画和建模，这一点是LDA直接应用在单一Tweet上不能做到的，ATM刻画的是在用户层面的背景(或者简单理解为“共现”)，而LDA刻画的是则是Tweet层面的背景。

- 其次，通过比较图 16和 17，我们可以看出，Twitter-LDA实际上是基于ATM的变化，主要有：1）引入背景模型（ θ_B ）；2）每一个Tweet内部的所有词汇拥有一个统一的主题标签。

- Twitter-LDA同时对于用户层面的背景和Tweet层面上的背景进行建模。每一个Tweet内部的所有词汇拥有一个统一的主题标签其实是一种很硬性的对于Tweet层面背景进行建模的方法，另外一种方法是同时有用户层面的主题分布和Tweet层面上的主题分布。但是由于Tweet长度过短，直接无监督地学习，很有可能不能得到理想的Tweet主题分布， [38] 在将主体模型得到的结果应用到后续相关的应用中发现了这个问题。

- [105]和 [106]表明Twitter-LDA和LDA和ATM做比较，对于学习主题词汇有着很好的效果。

- 此外Twitter-LDA的一个在后续应用中的优点就是可以很方便地回答Tweet层面上的统计问题，如语料集合中有多少条政治相关的问题，“美国总统”和“奥巴马”在多少条政治相关的Tweet中共现。而LDA和ATM对于这类的统计问题在Gibbs采样后进行最大后验概率计算。

除了这三个模型外， [74] 将之前的Labeled-LDA [75] 直接应用到Tweets上，它把每个HashTag都当作一个标签。这种方法的一个好处是可以利用用户在使用HashTag时形成的隐式模式，但是这个模型的一个缺点是不能直接应用到所有不包含HashTag的Tweet中去。

3.3.2 主题模型在Twitter中的应用挑战

主题模型在Twitter中的应用挑战主要可以分为两大类。第一类是由于目前主体模型研究的局限性带来的，例如主题数目的确定和主题模型的评价。这两部分在前面都有详细的讨论，在此我们略过。

第二类是由于Twitter数据本身带来的挑战：

- 数据的超大规模以及及时更新速度。截止到目前为止，还没有工作完全解决主题模型的大规模部署以及在线学习。

- 噪音大。Tweets往往都是很随意书写的，所以错别字（词）、新生词、网络用语、符号语言等等比起以普通网络文本数据有了很大的增加，这一问题未被很好解决。此外，一些常见的背景词汇，如，“love”，”tomorrow”，由于极高的出现频率也对于文本挖掘任务造成了一定的影响。因此在主题模型设置背景模型是一种降低这些词汇的影响方法。

- 数据的多特征关联性。Twitter本质上是一种社会关系网络，当前的使用的主题模型往往忽视了这方面的考虑；并且即使单单在内容上，Tweets本身也具有很多特点，如不同类型的分类（原创、转载、回复等等），使用Hashtag，还有时间特征等等。所以说Twitter语料集合了多个影响特征。，如何同时将这些特征融入主题模型进行建模仍然是一个开放待解决的问题。

[8]基于用户的Tweet聚合，然后基于聚合后的文档使用LDA，可以理解为LDA或者ATM，但是为了便于叙述逻辑上和比较上的清晰，我们将这种方法叫做ATM而不是LDA。

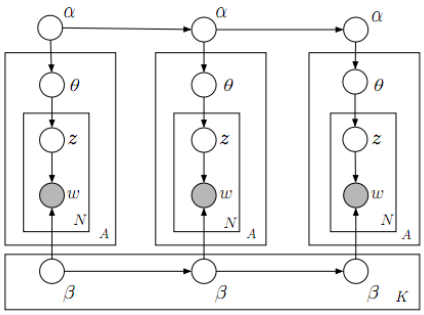

Figure 18: 动态主题模型示意图 [8]。

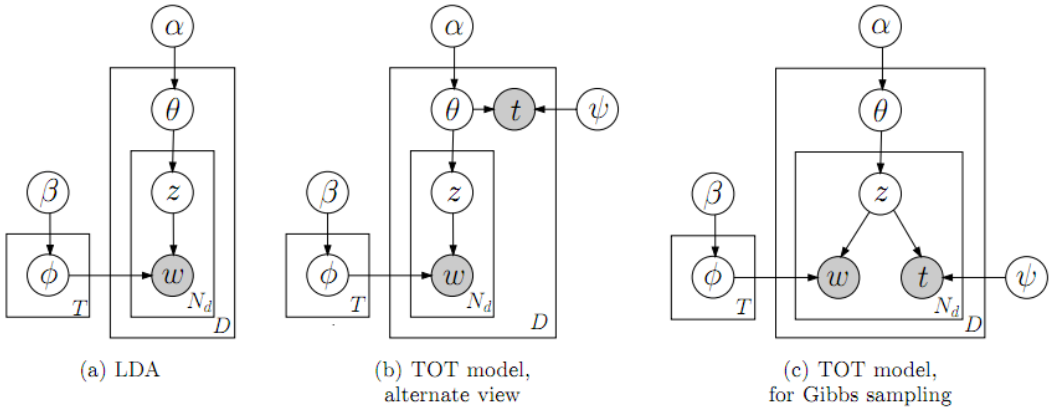

Figure 19: 非马尔科夫连续时间主题模型示意图 [95]。

3.4 时序文本流挖掘

传统的主题模型如 [11, 37] 都是把文档集合看做一个完全静态的集合来处理。随着互联网的不断发展，文档都是动态到来的，换句话说我们面对的是文本流。在文本流中，每个文本除了文本内容特征外，一个非常关键的要素就是时间因素。如何在主题模型中刻画时间因素、如何从主题模型的角度来考虑时序文本流的挖掘工作，这些都是非常关键的问题，在此，我们简要地总结一下：1）主题模型中如何引入时间因子；2）主体模型在时序文本流上的常见任务。

3.4.1 主题模型中引入时间因子

David M. Blei首先提出动态的主题模型 [8]，我们在图 18给出了动态主题模型的一个示意图。对于这个动态主题模型，基本思想是参数化α和β，不是假定一个时间范围的所有时间点都是相同的，形式化地说，我们使用α_t和β_t 然后利用马尔科夫假设，即$\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$和$\beta_t|\beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \delta^2 I)$。通俗地说，当前时间点的信息都依赖于前一个时间点的信息。 [2] 采用了同样的处理手段，但是使用了一个固定长度的滑动时间窗口。本质上这些方法都是将时间离散化，也就是通俗上说的“批处理”的方法。

但是对于一些应用，可能对于时间的粒度要求更高或者很难按照批划分数据。 [95] 提出了一个非马尔科夫连续时间模型，如图 19所示。不像之前的工作使用马尔科夫假设， [95]认为对于一个文档除了文本信息可见以外，时间标签也是可见信息。然后通过主题分布信息来同时关联起来词汇和时间标签。注意，对于 [95]仍然假定主题集合不随着时间变化而变化。 [92] 进一步松弛了这种假设，提出另外一种连续时间主体模型，在这个模型中，主题集合随着时间变化而变化。

3.4.2 时序文本流一些常见的文本挖掘任务

在此我们简要地介绍一些基于主体模型的时序文本流挖掘任务。

- Burst挖掘。侦测Burst首先被Kleinberg提出 [42]，指的是文本流中的词或者短语出现的词频上的突然增加，而这种词频上的显著增加很有可能对应着某些重要的事件，所以基于burst的方法经常被

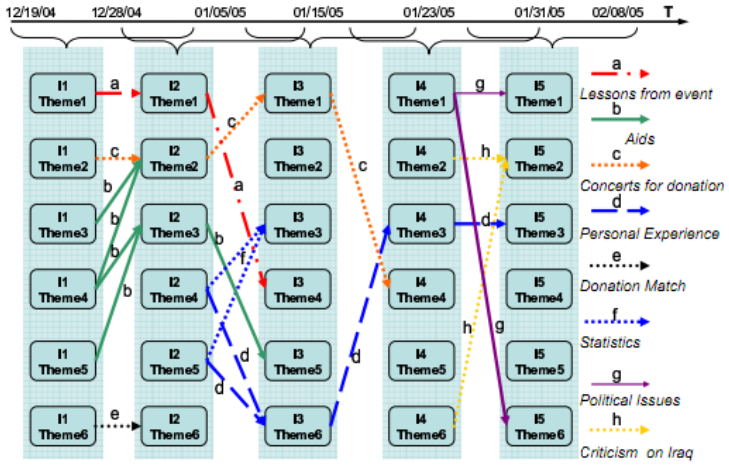

Figure 20: 文本流中的主题时序进化示意图 [59]。

用于事件检测 [28]。而在这里，我们谈论的是“Bursty Topic”，所以不是指单个词汇的burst，而是指某一热点话题。侦测多个流之间的bursty主题首先在 [94]被提出， [94]首先将时间划分成段，然后分别使用pLSA [37]学习得出每个时间片段内部的主题。为了侦测bursty主题， [94]提出了两个方法：1）使用前后时间段内部的主题进行平滑；2）不同流之间的主题互加强。但是 [94]中的方法使用的同步的文本流，也就是说对于同一个主题的报道，多个文本流之间的时间应该比较接近。 [93] 进一步松弛了这个限制，研究从多个异步的文本流挖掘bursty主题。对于一些文本流，例如学术文本流，由于各自研究领域的不同还有参与学者的不同，很有可能对于同一个主题的涉及有着较大差异。

- 文本流中的主题时序进化。这部分内容实际上在之前讨论的问题“主题模型中如何引入时间因子”中已经得到了部分涉及。在此，由于这个问题的重要性，我们将再次从应用的角度来进行讨论。 [59]第一次显示地提出了进行文本流中的主题时序进化分析。从这个问题标题，我们可以看到有两个主要问题：1）第一个是如何挖掘主题；2）第二个是如何进行主题的时序分析。 [59]使用了相对较为简单的方法，也是首先划分数据段，然后分段学习得到主题集合，然后根据在连续的两个时间段内的主题相似度对其建立链接关系。图 20给出了一个事件的主题时序进化的例子。我们可以发现这种展现事件内部主题的进化的方法实际上提供了一种对于文本流的摘要方式。实际上，谷歌的Timeline[9]也采用了相同的模式，只是每个子元素都是一个具体事件而不是一个主题。

3.5 网络结构数据的挖掘

网络结构是除了时间特征之外另一个文本的附加信息，它存在于各种类型的数据集合中：如文献中的引用关系、博客中的好友关系、网页中的链接关系。这些链接可以更好地帮助分析文档的语义含义。实际上对于一些数据类型来说，例如社会关系网络数据，链接关系本身就是一种数据类型，处于和文本同样重要的地位。在此，我们试介绍一下一些常见的将网络结构信息融入主题模型的方法。

利用相似性

我们首先介绍一下Relational topic model(RTM) [23]，图21给出了一个简单生成过程示意图。我们可以看到与标准的LDA [11]相比，RTM多了第二个大步骤，这与之前的讨论的link-plsa-lda [64]的基本思想非常类似，都是试图从主题分布的相似性来考虑进一步生成附加的链接结构。实际上，这是使用了一个隐含的假设：如果两个文档之间有着链接关系，那么他们之间的主题分布应该更为相似。那么给定文档的主题分布之后，如何考虑链接呢？ [23]考虑了两个式子作为链接概率函数:

$$\text{函数1：}\ \phi_\delta(y=1) = \delta(\eta^T(\bar{\mathbf{z}}_d \cdot \bar{\mathbf{z}}_{d'}) + \nu), \quad (2)$$

$$\text{函数2：}\ \phi_e(y=1) = \exp(\eta^T(\bar{\mathbf{z}}_d \cdot \bar{\mathbf{z}}_{d'}) + \nu), \quad (3)$$

[9]http://newstimeline.googlelabs.com

1. For each document d:
   (a) Draw topic proportions $\theta_d|\alpha \sim \text{Dir}(\alpha)$.
   (b) For each word $w_{d,n}$:
      i. Draw assignment $z_{d,n}|\theta_d \sim \text{Mult}(\theta_d)$.
      ii. Draw word $w_{d,n}|z_{d,n}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{d,n}})$.
2. For each pair of documents d, d':
   (a) Draw binary link indicator

$$y|\boldsymbol{z}_d, \boldsymbol{z}_{d'} \sim \psi(\cdot|\boldsymbol{z}_d, \boldsymbol{z}_{d'}).$$

Figure 21: Relational topic model的生成过程 [23]。

其中第一个函数是常用的Sigmoid函数，第二个是常用的指数函数。

[48] 拓展了 [23]，通过引入了社会关系在最后的链接关系形成函数中。也就是说，两个结点形成链接关系是由于两个因素：1）第一个是主题分布的相似性；2）另一个因素是社区从属关系的相似性。 [97] 也同时考虑文本和链接的生成，和[23]不同的是它侧重考虑每次链接事件的形成过程：首先，为整个链接事件选择一个主题标签；然后基于该主题为参与到该事件的两个结点进行群组（或者社区）标签的采样；最后基于这两个群组标签来决定一个链接事件形成的概率。

利用规则化的方法

规则化（Regularization）是一种在最优化、统计、机器学习中经常使用的方法[10]。其基本思想是对于模型的最优化函数上添加一些限制，通过这些限制使得模型避免过度拟合等病态学习问题。实际上，对于任何大多数主题模型，例如标准的LDA和pLSA，其求解过程都是一个最优化的过程，其目标函数就是使得该模型在语料集合上的似然最大。 [56, 20] 提出使用网络规则化因子在主题模型。基本思想还是如果两个结点存在链接关系，那么它们之间一定存在着一些度量上的相似性。一些具体的例子，如果两个作者曾经合作发表过论文，那么他们之间的主题兴趣分布应该相似；如果两个临近的地方都有着对于一个事件的报道，那么他们所报道的主题应该相似。

引入社区变量进行隐式聚类

上面所说的方法都是显示地利用网络结构特征进行聚类形成网络子结构方法。另外一种方法是不显式地对于链接进行建模，实际上Author Topic Model [77]可以看做一种隐式聚类的方法，也就是说我们可以按照主题的分布对于作者进行聚类，例如将作者分到他具有最大数值的主题内的群体。 [110] 基于Author-Topic Model [77]提出了两个这样的模型，在主题模型中增加一个“子结构”变量（例如，在社会关系网络中，子结构为社区），所有子结点都和这些“子结构”建立联系，形成了一种类似星形结构，这与前两种方法有着本质的不同。其基本思想是在Author-Topic Model的基础上引入“社区”变量，然后设定每个社区在作者集合上有一个多项式分布。

3.6 一些其他应用

在上述章节，我们详细地介绍了一些主题模型的应用和变形。在这个章节，我们简要地回顾下主题模型的一些其他应用。在这里，我们不再深入细节，而是采用一种宏观的角度来大概说一下这些应用。

自然语言处理

标准的语言模型，如LDA和pLSA，很难直接应用到一些传统的自然语言处理任务中，如词性标注、句法分析等等。人们往往拓展这些基本的主题模型，采用层级贝叶斯模型来专门处理这些任务。如词性标注 [88]、词义消歧 [15, 21, 12]、句(语)法分析 [13, 41, 31]、短语识别 [96, 106]、命名实体分析 [99, 67]、双语挖掘 [14, 39, 61, 68, 104, 103]、摘要生成 [82, 26, 22]等等。在这些应用中，主题模型的应用往往泛化到贝叶斯模型的应用。

[10] http://en.wikipedia.org/wiki/Regularization_(mathematics)

信息检索

其实之前讲过的一些章节，有很多任务都是和信息检索相关的。在此，我们再叙述下其他一些可能应用主题模型的信息检索任务。如基本的检索任务 [98, 108, 45]、专家寻找 [27]、信息抽取 [102, 25]、标注 [108, 58, 80, 16, 50, 91]等等。

4 主题模型工具

在这里，我们简要列一下一些基本的主题模型工具库，仅供之后的学术研究使用。有 [6, 18, 73, 35, 55, 63, 70, 76, 81]。其中最常用的一般是 [70]。一般来说，基于Gibbs采样方法实现的工具速度比较慢，但是实现起来较为简单。

5 总结

主题模型本质上是一种对文本的概率建模的方法，并不是像想象中那样的神秘和高不可攀，也不是像想象中对于所有的任务都是万能的。对于研究工作者，必须根据自己的实际任务的需要和主题模型的特性进行思考，从而判断是不是应该使用主题模型的方法。希望本文对于文本挖掘研究工作人员有着一定的参考作用。

References

[1] Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. Topic significance ranking of lda generative models. In *ECML*, 2009.

[2] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 3–12, Washington, DC, USA, 2008. IEEE Computer Society.

[3] Arthur Asuncion, Padhraic Smyth, and Max Welling. Asynchronous distributed learning of topic models. In *NIPS*, pages 81–88, 2008.

[4] Arindam Banerjee and Sugato Basu. Topic Models over Text Streams: A Study of Batch and Online Unsupervised Learning. In *SDM*. SIAM, 2007.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[6] David M. Blei. lda-c, 2003.

[7] David M. Blei, Thomas Griffiths, Michael Jordan, and Joshua Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2003.

[8] David M. Blei and John D. Lafferty. Dynamic topic models. In *ICML*, 2006.

[9] David M. Blei and John D. Lafferty. A correlated topic model of science. *AAS*, 1(1):17–35, 2007.

[10] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[12] Jordan Boyd-Graber and David M. Blei. Putop: Turning predominant senses into a topic model for wsd. In *SEMEVAL*, 2007.

[13] Jordan Boyd-Graber and David M. Blei. Syntactic topic models. In *NIPS*, 2008.

[14] Jordan Boyd-Graber and David M. Blei. Multilingual topic models for unaligned text. In *UAI*, 2009.

[15] Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. In *EMNLP*, 2007.

[16] S.R.K. Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. Learning document-level semantic properties from free-text annotations. In *Proceedings of ACL-08: HLT*, pages 263–271, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[17] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 804–812, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[18] Wray L. Buntine. Discrete component analysis, 2009.

[19] Wray L. Buntine. Estimating likelihoods for topic models. In *Asian Conference on Machine Learning*, 2009.

[20] Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. Modeling hidden topics on document manifold. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 911–920, New York, NY, USA, 2008. ACM.

[21] Jun Fu Cai, Wee Sun Lee, and Yee Whye Teh. Nus-ml: Improving word sense disambiguation using topic features. In *SEMEVAL*, 2007.

[22] Asli Celikyilmaz and Dilek Hakkani-Tur. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 815–824, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[23] Jonathan Chang and David Blei. Relational topic models for document networks. In *AIStats*, 2009.

[24] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

[25] Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Combining concept hierarchies and statistical topic models. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1469–1470, New York, NY, USA, 2008. ACM.

[26] Hal Daumé, III and Daniel Marcu. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 305–312, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[27] Hui Fang and ChengXiang Zhai. Probabilistic models for expert finding. In *Proceedings of the 29th European conference on IR research*, ECIR'07, pages 418–430, Berlin, Heidelberg, 2007. Springer-Verlag.

[28] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, and Hongjun Lu. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*, 2005.

[29] Mark Girolami and Ata Kabán. On an equivalence between plsi and lda. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 433–434, New York, NY, USA, 2003. ACM.

[30] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[31] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *NIPS*, pages 537–544. 2004.

[32] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 363–371, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[33] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 957–966, New York, NY, USA, 2009. ACM.

[34] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[35] Gregor Heinrich. Infinite lda, 2011.

[36] Matthew Hoffman, David M. Blei, and Francis Bach. Online learning for latent dirichlet allocation. In *NIPS*, 2010.

[37] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42:177–196, January 2001.

[38] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[39] Jagadeesh Jagarlamudi and Hal DauméIII. Extracting multilingual topics from unaligned comparable corpora. pages 444–456, 2010.

[40] Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 815–824, New York, NY, USA, 2011. ACM.

[41] Mark Johnson. Pcfgs, topic models, adaptor grammars, and learning topical collocations and the structure of proper names. 2010.

[42] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 2003.

[43] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, (25):259–284, 1998.

[44] Peng Li, Jing Jiang, and Yinglin Wang. Generating templates of entity summaries with an entity-aspect model and pattern mining. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 640–649, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[45] R.M. Li, R. Kaptein, D. Hiemstra, and J. Kamps. Exploring topic-based language models for effective web information retrieval. In E. Hoenkamp, M. de Cock, and V. Hoste, editors, *Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR 2008)*, pages 65–71, Enschede, April 2008. Neslia Paniculata.

[46] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML*, 2006.

[47] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 375–384, New York, NY, USA, 2009. ACM.

[48] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link lda: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 665–672, New York, NY, USA, 2009. ACM.

[49] Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, and Maosong Sun. Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing. *ACM Trans. Intell. Syst. Technol.*, 2:26:1–26:18, May 2011.

[50] Caimei Lu, Xiaohua Hu, Xin Chen, Jung-Ran Park, TingTing He, and Zhoujun Li. The topic-perspective model for social tagging systems. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 683–692, New York, NY, USA, 2010. ACM.

[51] Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 121–130, New York, NY, USA, 2008. ACM.

[52] Gideon S. Mann, David Mimno, and Andrew McCallum. Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '06, pages 65–74, New York, NY, USA, 2006. ACM.

[53] Andrew McCallum. Multi-label text classification with a mixture model trained by em. In *AAAI'99 Workshop on Text Learning*, 1999.

[54] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Int. Res.*, 30:249–272, October 2007.

[55] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit, 2002.

[56] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 101–110, New York, NY, USA, 2008. ACM.

[57] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 171–180, New York, NY, USA, 2007. ACM.

[58] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *KDD*, pages 490–499, 2007.

[59] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 198–207, New York, NY, USA, 2005. ACM.

[60] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, 2008.

[61] David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *EMNLP*, 2009.

[62] Thomas P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[63] Ramesh Nallapati. multithreaded lda-c, 2010.

[64] Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 542–550, New York, NY, USA, 2008. ACM.

[65] D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed Inference for Latent Dirichlet Allocation. 2007.

[66] David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 680–686, New York, NY, USA, 2006. ACM.

[67] David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. Statistical entity-topic models. In *KDD*, 2006.

[68] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. Mining multilingual topics from wikipedia. In *WWW*, 2009.

[69] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, January 2008.

[70] Xuan-Hieu Phan and Cam-Tu Nguyen. Gibbslda++, 2007.

[71] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.

[72] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 569–577, New York, NY, USA, 2008. ACM.

[73] Radim. gensim, 2009.

[74] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[75] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, 2009.

[76] Daniel Ramage and Evan Rosen. Stanford topic modeling toolbox, 2009.

[77] Michal Rosen-Zvi, Tom Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *UAI*, 2004.

[78] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[79] Issei Sato and Hiroshi Nakagawa. Topic models with power-law using pitman-yor process. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 673–682, New York, NY, USA, 2010. ACM.

[80] Yuanlong Shao, Yuan Zhou, Xiaofei He, Deng Cai, and Hujun Bao. Semi-supervised topic modeling for image annotation. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 521–524, New York, NY, USA, 2009. ACM.

[81] Mark Steyvers and Tom Griffiths. Matlab topic modeling toolbox, 2005.

[82] Jie Tang, Limin Yao, and Dewei Chen. Multi-topic based query-oriented summarization. In *SDM*, pages 1147–1158, 2009.

[83] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

[84] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[85] Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 985–992, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[86] Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[87] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 111–120, New York, NY, USA, 2008. ACM.

[88] Kristina Toutanova and Mark Johnson. A bayesian lda-based model for semi-supervised part-of-speech tagging. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1521–1528. MIT Press, Cambridge, MA, 2008.

[89] Hanna Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[90] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *ICML*, 2009.

[91] Chong Wang, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *CVPR*, 2009.

[92] Chong Wang, David M. Blei, and David Heckerman. Continuous time dynamic topic models. In *UAI*, 2008.

[93] Xiang Wang, Kai Zhang, Xiaoming Jin, and Dou Shen. Mining common topics from multiple asynchronous text streams. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 192–201, New York, NY, USA, 2009. ACM.

[94] Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007.

[95] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 424–433, New York, NY, USA, 2006. ACM.

[96] Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 697–702, Washington, DC, USA, 2007. IEEE Computer Society.

[97] Xuerui Wang, Natasha Mohanty, and Andrew McCallum. Group and topic discovery from relations and text. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 28–35, New York, NY, USA, 2005. ACM.

[98] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM.

[99] Gu Xu, Shuang-Hong Yang, and Hang Li. Named entity mining from click-through data using weakly supervised latent dirichlet allocation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1365–1374, New York, NY, USA, 2009. ACM.

[100] Feng Yan, Ningyi Xu, and Yuan Qi. Parallel inference for latent dirichlet allocation on graphics processing units. In *NIPS*, 2009.

[101] Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 937–946, New York, NY, USA, 2009. ACM.

[102] Huibin Zhang, Mingjie Zhu, Shuming Shi, and Ji-Rong Wen. Employing topic models for pattern-based semantic class discovery. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 459–467, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[103] Bin Zhao and Eric P. Xing. Hm-bitam: Bilingual topic exploration, word alignment, and translation. In *NIPS*, 2007.

[104] Bing Zhao and Eric P. Xing. Bitam: Bilingual topic admixture models for word alignment. In *ACL*, 2006.

[105] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *ECIR*, pages 338–349, 2011.

[106] Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achanauparp, Ee-Peng Lim, and Xiaoming Li. Topical keyphrase extraction from twitter. In *ACL-HLT*, 2011.

[107] Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65, Cambridge, MA, October 2010. Association for Computational Linguistics.

[108] Ding Zhou, Jiang Bian, Shuyi Zheng, Giles Lee, and Hongyuan Zha. Exploring social annotations for information retrieval. In *Proceedings of the 17th International World Wide Web Conference*, Beijing, Peking, 2008.

[109] Ding Zhou, Xiang Ji, Hongyuan Zha, and C. Lee Giles. Topic evolution and social interactions: how authors effect research. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 248–257, New York, NY, USA, 2006. ACM.

[110] Ding Zhou, Eren Manavoglu, Jia Li, C. Lee Giles, and Hongyuan Zha. Probabilistic models for discovering e-communities. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 173–182, New York, NY, USA, 2006. ACM.

Appendix

该部分给出LDA的完整的参数推导过程，为了清晰可见，我们将全部使用英语来进行描述和推导。

Let $\alpha = (\alpha_1, .., \alpha_K)$ be the hyper-parameters for document-topic distribution(i.e. $\theta_d = \{p(z|d)\}_z$). Let $\beta = (\beta_1, .., \beta_V)$ be the hyper-parameters for topic-word distribution(i.e. $\phi_z = \{p(w|z)\}_w$). The prior probabilities on parameters are defined in a Dirichlet distribution

$$\begin{aligned} p(\theta|\alpha) &= \prod_{m=1}^{D} \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k - 1} \right), \\ p(\phi|\beta) &= \prod_{k=1}^{K} \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{k,v}^{\beta_v - 1} \right). \end{aligned} \tag{4}$$

The probabilities of all hidden variables given parameters $\{\theta_d\}_d$ is defined in a multinomial distribution

$$p(\mathbf{z}|\theta) = \prod_{m=1}^{D} \prod_{k=1}^{K} \theta_{m,k}^{n_{m,k}}, \tag{5}$$

where $n_{m,k}$ denotes the count of words assigned to topic k in document m.

The probabilities of all observed variables(words) given parameters $\{\phi z\}_z$ and all the hidden variables is defined in a multinomial distribution

$$p(\mathbf{w}|\mathbf{z}, \phi) = \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{k,v}}, \tag{6}$$

where $n_{k,v}$ denotes the count of word v assigned to topic k in corpus.

Noting that, in these two formulas, we assume that given parameters all the variables are indepedent of hyper-parameters.

Given the hyper-parameters, the joint probability of all the variables (including parameters) is defined

$$p(\mathbf{w}, \mathbf{z}, \theta, \phi|\alpha, \beta) = p(\mathbf{w}|\mathbf{z}, \theta) p(\mathbf{z}|\phi) p(\phi|\beta) p(\theta|\alpha). \tag{7}$$

So after deriving the joint probability, we integrate all the parameters

$$\begin{aligned} p(\mathbf{w}, \mathbf{z}|\alpha, \beta) &= \int_\theta \int_\phi p(\mathbf{w}|\mathbf{z}, \theta) p(\mathbf{z}|\phi) p(\phi|\beta) p(\theta|\alpha) d\phi d\theta, \\ &= \int_\theta \int_\phi \left(\prod_{m=1}^{D} \prod_{k=1}^{K} \theta_{m,k}^{n_{m,k}} \right) \left(\prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{k,v}} \right) \\ &\quad \left(\prod_{k=1}^{K} \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{k,v}^{\beta_v - 1} \right) \right) \left(\prod_{m=1}^{D} \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k - 1} \right) \right) d\phi d\theta, \\ &= \int_\theta \left(\left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right)^D \prod_{m=1}^{D} \prod_{k=1}^{K} \theta_{m,k}^{n_{m,k} + \alpha_k - 1} \right) d\theta \\ &\quad \int_\phi \left(\left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \right)^K \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{k,v} + \phi_v - 1} \right) d\phi \\ &= \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right)^D \prod_{m=1}^{D} \int_\theta \left(\prod_{k=1}^{K} \theta_{m,k}^{n_{m,k} + \alpha_k - 1} \right) d\theta \\ &\quad \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \right)^K \prod_{k=1}^{K} \int_\phi \left(\prod_{v=1}^{V} \phi_{k,v}^{n_{k,v} + \phi_v - 1} \right) d\phi \\ &= \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right)^D \prod_{m=1}^{D} \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + n_{m,k})}{\Gamma(\sum_{k=1}^{K} \alpha_k + n_{m,k})} \\ &\quad \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \right)^K \prod_{k=1}^{K} \frac{\prod_{v=1}^{V} \Gamma(\beta_v + n_{k,v})}{\Gamma(\sum_{v=1}^{V} \beta_v + n_{k,v})} \end{aligned} \tag{8}$$

$$
\begin{aligned}
p(z_i = t|\mathbf{w}, \mathbf{z}_{\neg i}, \alpha, \beta) &= \frac{p(z_i, \mathbf{w}, \mathbf{z}_{\neg i}, \alpha, \beta)}{p(\mathbf{w}, \mathbf{z}_{\neg i}, \alpha, \beta)} \\
&= \frac{p(\mathbf{w}, \mathbf{z}, \alpha, \beta)}{p(\mathbf{w}_{\neg i}, w_i, \mathbf{z}_{\neg i}, \alpha, \beta)} \\
&= \frac{p(\mathbf{w}, \mathbf{z}, \alpha, \beta)}{p(\mathbf{w}_{\neg i}, \mathbf{z}_{\neg i}, \alpha, \beta)p(w_i|\mathbf{w}_{\neg i}, \mathbf{z}_{\neg i}, \alpha, \beta)} \\
&\propto \frac{p(\mathbf{w}, \mathbf{z}|\alpha, \beta)}{p(\mathbf{w}_{\neg i}, \mathbf{z}_{\neg i}|\alpha, \beta)} \\
&= \frac{\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)^D \prod_{m=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+n_{m,k})}{\Gamma(\sum_{k=1}^{K}\alpha_k+n_{m,k})} \left(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\right)^K \prod_{k=1}^{K}\frac{\prod_{v=1}^{V}\Gamma(\beta_v+n_{k,v})}{\Gamma(\sum_{v=1}^{V}\beta_v+n_{k,v})}}{\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)^D \prod_{m=1}^{D}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+n_{m,k}^{-i})}{\Gamma(\sum_{k=1}^{K}\alpha_k+n_{m,k}^{-i})} \left(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\right)^K \prod_{k=1}^{K}\frac{\prod_{v=1}^{V}\Gamma(\beta_v+n_{k,v}^{-i})}{\Gamma(\sum_{v=1}^{V}\beta_v+n_{k,v}^{-i})}} \\
&= \frac{\alpha_t + n_{m,t}^{-i}}{\sum_{k=1}^{K}\alpha_k + n_{m,k}^{-i}} \frac{\beta_v + n_{t,v}^{-i}}{\sum_{v=1}^{V}\beta_v + n_{t,v}^{-i}}.
\end{aligned} \tag{9}
$$

The derivations above used the results in Equation 8 and also one important property of Gamma function $\Gamma(x+1) = x\Gamma(x)$.