概率主题模型简介 Introduction to Probabilistic Topic Models

此文为David M. Blei所写的《Introduction to Probabilistic Topic Models》的译文，供大家参考。

摘要：概率主题模型是一系列旨在发现隐藏在大规模文档中的主题结构的算法。本文首先回顾了这一领域的主要思想，接着调研了当前的研究水平，最后展望某些有所希望的方向。从最简单的主题模型——潜在狄立克雷分配（Latent Dirichlet Allocation，LDA）出发，讨论了其与概率建模的联系，描述了用于主题发现的两种算法。主题模型日新月异，被扩展和应用许多领域，其中不乏有趣之处。我们调研发现很多扩展都弱化了LDA的统计假设，加入元数据（meta-data）进行文档分析，使用近似的模型分析如社会网络、图像和基因这类多样化的数据类型。我们在文章的最后给出了主题模型目前还未探索但很重要的方向，包括严格检验数据模型的方法，文本和其它高维数据可视化的新技术，以及如何从传统信息工程中的应用推广到更多科学应用。

1 引言

如今公开的知识日益以新闻、博客、网页、科学论文、书籍、图像、声音、视频和社交网络的形式被数字化存储，巨大的信息量同时也增加了人们寻找和发现自己所需要的知识的难度。人们需要新的计算工具以组织、搜索和理解这些庞大的信息量。现在的在线信息挖掘使用两种主要的工具——搜索和链接。向搜索引擎提交关键词就可以找到相关的文档和其它相链接的文档。这种与在线文档的交互方式虽然有效，但却丢失了某些信息。

假设所要搜索和寻找的文档由各类主题组成。这样，通过对文章进行“放大”和“缩小”就可以得到较具体或者较粗略的主题；在文档中就可以看到这些主题是如何随着时间变化，或者说是如何相互联系的。搜索文档就不只是通过关键词寻找，取而代之的是先找到相关的主题，然后再查找与这一主题相关的文档。

拿纽约时报所记载的历史举例。从较广的层次来看，报纸中的主题就对应着报纸各个版块——对外政策、国内事务、体育，再拿对外政策进行“放大”，就可以得到其不同方面——中国对外政策、中东冲突、英国与俄罗斯的关系。接下来，我们跟踪这些专题是如何随着时间演变的，例如过去50年里的中东冲突。如此这般探索就能找到与主题相关的原始文档。可见，这种主题结构是探索和理解文档的新窗口。

但以这种方法与电子文档进行交互是不现实的，因为随着网上文本的数量越来越多，单单仅靠人力已经无法全部阅读和研究所有的文本。由此，概率主题建模应运而生。机器学习领域的研究人员们开发出了一套旨在发现和标记大规模文档的主题信息的算法。主题建模算法是一种统计方法，它通过分析原文本中的词以发现蕴藏于其中的主题，主题间的联系，以及主题随时间的演变（就比如后面图3，通过分析耶鲁法律找到主题），而且不需要事前对文档进行标记。也就是说，人力所无法完成的文档标记，主题建模算法能够进行组织和归纳。

2 潜在狄立克雷分配

潜在狄立克雷分配（LDA）是最简单的主题模型，其基础是文档是由多个主题构成的。如图1所示，《Seeking Life’s Bare(Genetic) Necessities》是一篇对基因数量进行数据分析的文章（基因是有机体赖以进化的基础）。

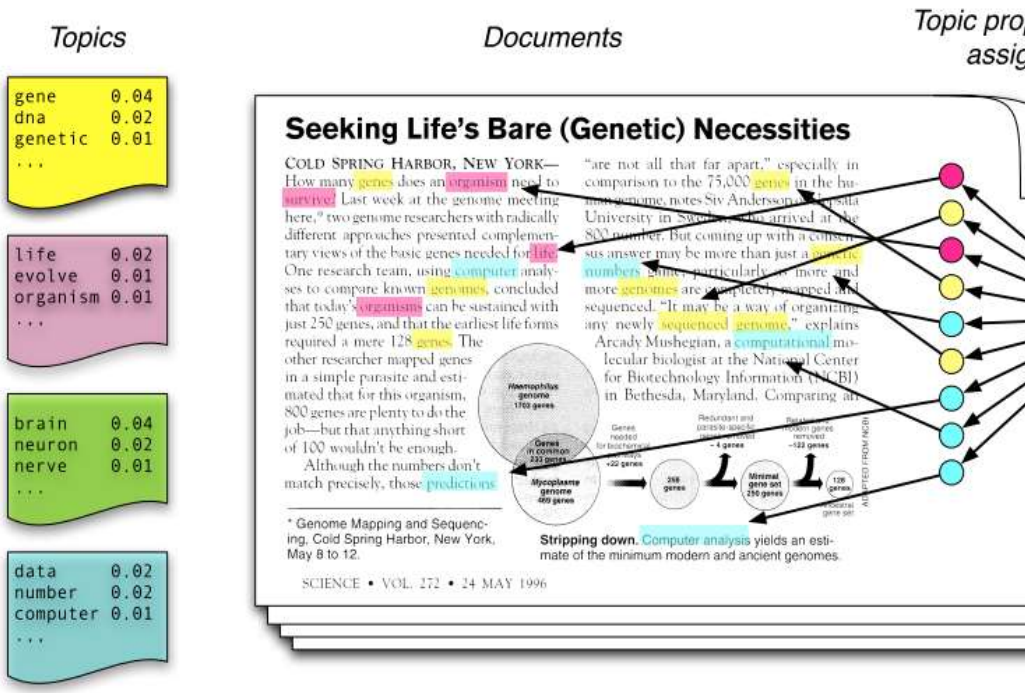


图1：潜在狄立克雷分配的直观现象。如图左所示，假设主题是词语上的概率分布；图右是主题直方图。从直方图到文章的词的过程是这样的：首先随机产生一个主题直方图，然后选择其中一主题，最后从该主题对应的主题分布中选择一个词。这里的主题和主题直方图只作说明之用，与文章其实并不相匹配。相匹配的主题见图2。

文章中不同的词被高亮在不同的颜色。如"computer"和"prediction"之类有关数据分析的词以蓝色标记；如"life"和"organism"之类关于进化生物学的词以粉红色标记；如"sequenced"和"genes"之类有关遗传学的词以黄色标记。将所有词语进行这样的标记，并剔除"and"、"but"和"if"这类包含极少主题内容的词语后可以发现，这篇文章由不同主题以不同的比例组成，更进一步地看，多个主题可以帮助人们在一堆科技论文中发现这篇文章。

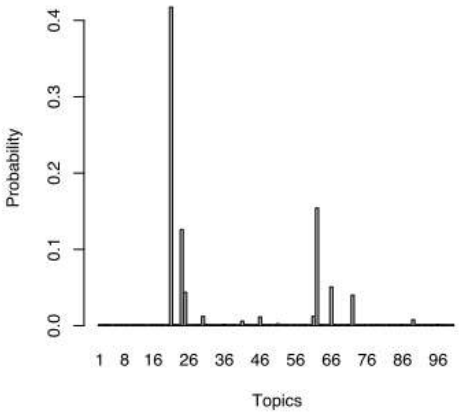
建立在文档集合上的统计模型LDA就试图描述上述直观的现象。LDA可以看作是一个文档产生的过程（2.1节将具体解释概率模型LDA）。形式化地定义主题是固定的词语的概率分布。例如，"遗传学"主题中"genes"的概率就相当高，类似地，"进化生物学"主题中"life"的概率也相对较高。假设所有的主题在文档产生之前就已经产生且指定。生成文档（或者说生成文档中的词）可以看成是如下两个过程：

1. 随机产生一个主题直方图（或者说分布）；
2. 对文档中的每个词：
    1. (a) 从第一步产生的直方图里随机选择一个主题；
    2. (b) 从主题对应的词语的概率分布中随机选择一个词。

从文档产生的过程来看，第一步使得每篇文档由不同主题以不同比例组成。第二步的第二小步（b）使得每篇文档中每个词从一个主题中得来，其中的主题从第一小步（a）得来。实际上，第一步主题直方图（或者说分布）是一个狄立克雷分布（Dirichlet distribution），其作用是将文档中的词分配给不同的主题，那为什么是潜在的呢？且听后面分解。

对图1所示的文章来说，主题直方图中主题"遗传学"、"数据分析"和"进化生物学"都会占一定比例，文章中每个词都由这三个主题中的一个所给出。文档集中也可能会有一篇关于"数据分析"和"神经科学"；其主题直方图中这两个主题都将占有一定的比例；这就是潜在狄立克雷分配的显著特征——集合中所有文档共享同一主题集合，但每个文档中各个主题所占的比例又都各不相同。

如前引言所述，主题建模的目的是为了自动地发现文档集中的主题。文档自然是可被观察到的，但主题结构——主题、主题直方图（或者分布）和主题的词分布——却是隐藏的。所以主题建模的中心问题就是利用看到的文档推断出隐藏的主题结构，其实也就是产生文档的逆过程。



"Genetics"	"Evolution"	"Disease"
human	evolution	disease
genome	evolutionary	host
dna	species	bacteria
genetic	organisms	diseases
genes	life	resistance
sequence	origin	bacterial
gene	biology	new
molecular	groups	strains
sequencing	phylogenetic	control
map	living	infectious
information	diversity	malaria
genetics	group	parasite
mapping	new	parasites
project	two	united
sequences	common	tuberculos

图2：图1的LDA。我们从《自然》上的17000篇文章提取100个主题及其相关词，然后对图1所示的文章进行分析，左边是主题所占比例的直方图，右边是文章常见主题的最常出现的前15个词。

如图2所示，就是一个推断图1中文章的例子。使用主题建模算法（假设有100个主题）推断《科学》上17000篇文章的潜在主题结构，然后推断出最能描述图1中示例文章的主题分布（图左）。需要注意的是，尽管主题分布上有无穷个主题，但事实上只有其中的一小部分的概率不为零。进一步地，文章中词可被分主题进行组织，可以看到最常见的主题所包含的概率最大的词。
需要强调的是，算法事先并不知道这些主题，文章也未有关键词或主题标记。计算潜在结构得到的主题分布可以产生所观察到的文档集合（由推断算法产生的主题对所分析的文档集合几乎都具有可解释性，主题似乎与语言的统计结构和LDA的具体概率假设有关）。如图3显示了《Yale Law Journal》中发现的主题（这里设置主题数为20）。主题由基因和数据分析替换为歧视和合同法。主题建模是管理、组织和标记大规模文本的一种算法。推断得到的隐藏结构近似于文档集的主题结构，能标记文档集中各个文档。这代替了痛苦的手工标记，并有助于信息检索，分类和语料库搜索。

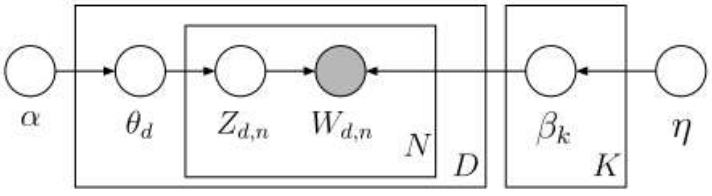
2.1 LDA和概率模型

LDA和其它主题模型都属于概率建模这一更大领域。数据被看作是经过包括隐藏变量在内的生成过程得到的。生成过程定义了观测随机变量和隐藏随机变量的联合概率分布。通过使用联合分布来计算在给定观测变量下隐藏变量的条件分布（后验分布）来进行数据分析。对于LDA来说，观测变量就是文档中的词；隐藏变量就是主题结构；生成过程如之前所述。那么推测从文档中隐藏的主题结构的问题其实就是计算在给定文档下隐藏变量的条件分布（后验分布）。
形式化地定义如下：所有主题为$\beta_{1:K}$，其中βk是第k个主题的词的分布（如图1左部所示）。第d个文档中主题所占的比例为θ_d，其中$\theta_{d,k}$表示第k个主题在第d个文档中的比例（图1右部的直方图）。第d个文档的主题全体为z_d，其中$z_{d,n}$是第d个文档中第n个词的主题（如图1中有颜色的圆圈）。第d个文档中所有词记为w_d，其中$w_{d,n}$是第d个文档中第n个词，每个词都是固定的词汇表中的元素。那么LDA的生成过程对应的观测变量和隐藏变量的联合分布如下：

$$p\left(\beta_{1:K},\theta_{1:D},z_{1:D},w_{1:D}\right)=\prod_{i=1}^{K}p\left(\beta\right)\prod_{d=1}^{D}p\left(\theta_d\right)\left(\prod_{n=1}^{N}p\left(z_{d,n}|\theta_d\right)p\left(w_{d,n}|\beta_{1:K}\right.\right.$$

（1）

这一分布指明变量之间的依赖关系。例如，$z_{d,n}$依赖于θ_d，$w_{d,n}$依赖于$z_{d,n}$和$\beta_{1:K}$（在操作上，先确定$z_{d,n}$指的哪个主题，然后再看$w_{d,n}$在主题中的概率）。正是这些依赖定义了LDA：它们存在于生成过程的统计假设里，在联合分布的特定数学形式里以及LDA的概率图模型里（概率图模型为描述概率分布提供一个图形化的语言，如图4所示。事实上概率图模型是阐明概率独立、图理论和计算概率分布的算法的有力工具）。这三种表现形式在描述LDA的概率假设上是等价的。



概率版本，而LDA是用以解决pLSI的问题，可以看作是对离散数据进行主成分分析。下一章节将详细描述LDA的推断算法。

2.2 LDA后验概率的计算

使用前面的记号，LDA后验概率的公式为

$$p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}\right) = \frac{p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}\right)}{p\left(w_{1:D}\right)}$$

（2）

分子为随机变量的联合分布。对于隐藏变量的任何值来说，联合分布是容易计算的。分母是观测变量的边际概率，是通过观察可见的语料库得到的概率。理论上，可以通过将联合分布对隐藏变量的所有可能值进行累加得到。但其计算量在实际操作中是异常庞大的（对于一个主题，这种累加包括了将每个词的所有可能的主题配置，而且文档集合通常有数量级达百万的词）。就像众多现代概率模型（包括贝叶斯统计）那样，后验概率的分母（即先验概率）往往是无法计算得到的。故而现代概率建模的一个核心研究目标就是尽一切可能接近之。如前图1和图3所述的那样，主题建模算法其实是求得近似后验分布的常用方法的一种变种。

主题建模算法主要有两类：基于采样的算法和变分算法。基于采样的算法通过收集后验分布的样本，以样本的分布求得后验分布的近似。主题建模中最常用的采样算法是吉布斯采样（Gibbs sampling），通过吉布斯采样构造马尔可夫链（Markov chain），而马尔可夫链的极限分布就是后验分布。马尔可夫链是由独立于前一个随机变量的随机变量组成的串。对主题模型来说，随机变量就是定义在一个特定的语料库上的隐藏主题。采样算法从马尔可夫链的极限分布上收集样本，再用这些样本来近似后验分布。通常，只有概率最高的样本会被收集以作为主题结构的近似。文献[33]详细描述了LDA的吉布斯采样，开源社区里有R语言的快速开源实现（http://cran.r-project.org/web/packages/lda/index.html）。

变分算法的确定性要比基于采样算法高上不少。变分算法先假定一族在隐藏结构之上的参数化的分布，再寻找与后验分布最接近的分布（概率分布之间的距离使用信息论的Kullback-Leibler散度度量，）。也就说，推断问题转换为了最优化问题。变分算法的创新之处也正在于此，它将最优化引入了概率建模中。文献[8]介绍了协调上升的变分推断算法；文献[20]介绍了一个更为快速的在线算法（以及开源软件），它能轻松处理上百万文档并能适应文本流的集合。

粗略地讲，这两种算法都在主题结构上进行了搜索，而固定的文档集合提供了搜索的方向。哪种方法更适合取决于所使用的具体的主题模型（下面的章节会介绍除LDA以外的其它主题模型），而这通常是学院派们争论的导火索。文献[1]很好地讨论了这两种方法的优缺点。

3 主题建模的研究进展

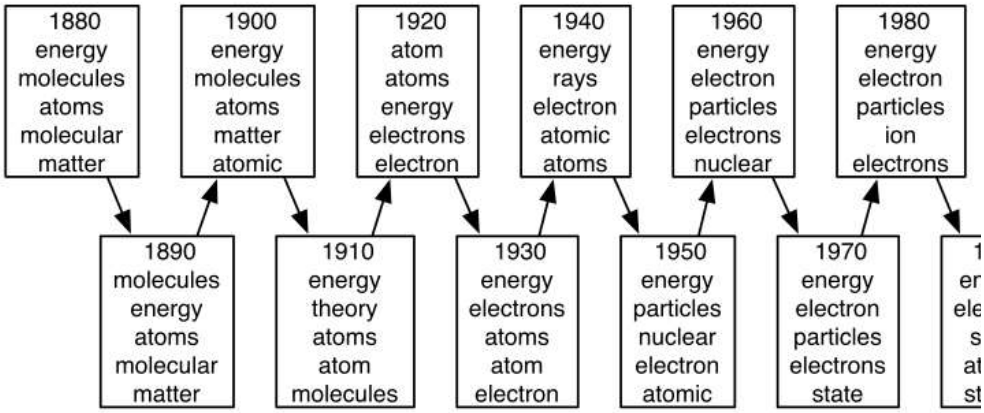
简单的LDA模型提供发现和探索大规模文本中隐藏主题结构的有力工具。LDA形式化为概率模型的一个主要优点在于它可以作为一个模块被其它更复杂的模块应用更为复杂的应用系统中。

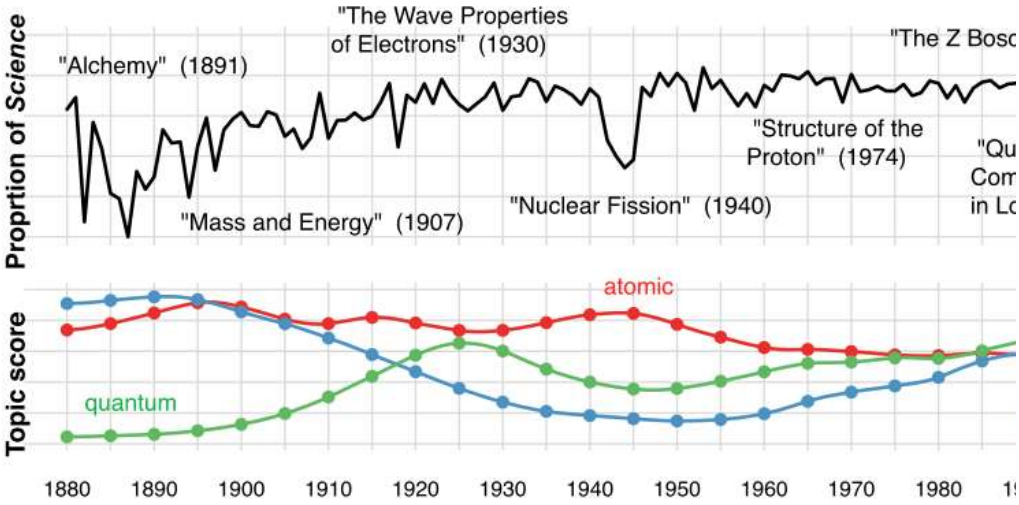
3.1 弱化LDA的假设

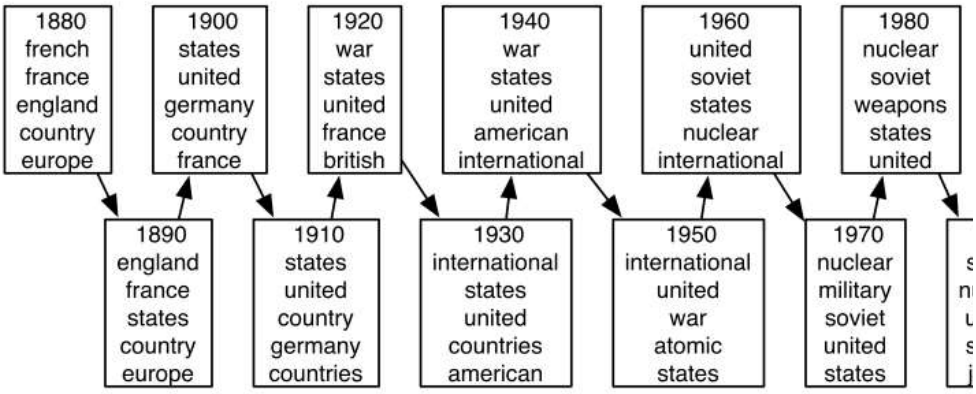
LDA由对语料库的统计假设定义，主题建模研究领域中一个热点就是如何弱化和扩展这些假设以发现文本中更加复杂的结构。

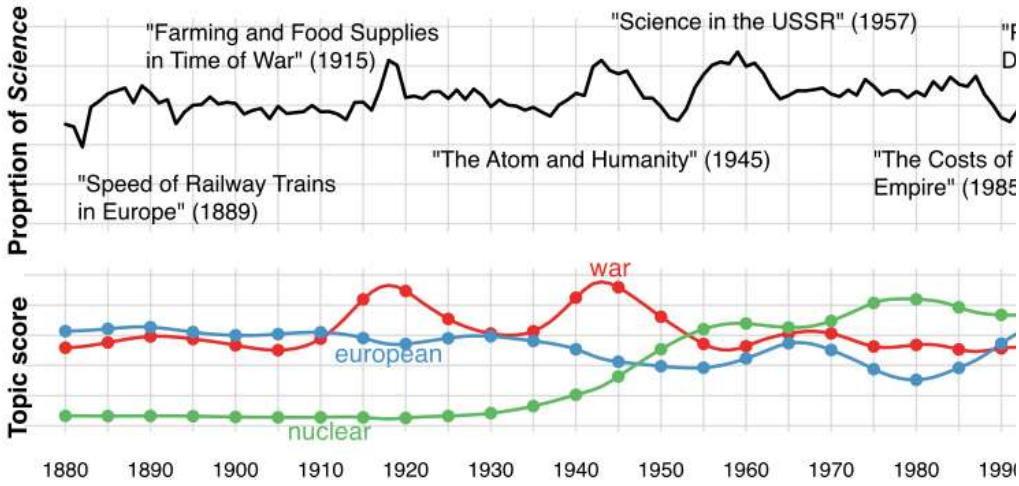
如果不考虑词在文档中的顺序，那么一个文档就是一个词袋。“词袋”就是LDA的一个假设（对于文档中词的任意排列，式（1）都不变）。这个假设当然不符合现实，对于复杂的诸如语言生成之类的目标显然是不合适的，但对发现文本语义结构是有理可循的（缓慢移动图1文章中的词，尽管词移动了位置，文章仍然与遗传学相关）。LDA上模型的不可交换的词也有诸多扩展。例如，文献[36]弱化了词袋模型，假设主题生成词时以前一个词作为条件；文献[18]提出了在LDA和隐马尔可夫模型之间进行切换的主题模型。这些模型显著地扩展了参数空间，并且显示了语言建模带来的性能提升。

LDA的另一个假设是文档的顺序与LDA无关（对于文档的任意顺序的排列，式（1）同样不变）。但当文档集所跨越的时间有几年或几个世纪时，这个假设可能就不合理了。当遇到这样的集合时，通常假设主题是随着时间而发生变化的。这样，主题就是动态的[5]。动态主题模型考虑了文档的先后顺序，并给出了内涵比LDA更丰富后验主题结构。图5展示了使用动态主题模型分析所有科学杂志得到的一个主题。这个主题就不只是词的单一分布，而是词的一连串分布。这样，潜在的主题就可以被发现和跟踪。


1880 energy molecules atoms molecular matter
1890 molecules energy atoms molecular matter
1900 energy molecules atoms matter atomic
1910 energy theory atoms atom molecules
1920 atom atoms energy electrons electron
1930 energy electrons atoms atom electron
1940 energy rays electron atomic atoms
1950 energy particles nuclear electron atomic
1960 energy electron particles electrons nuclear
1970 energy electron particles electrons state
1980 energy electron particles ion electrons


Proprtion of Science
"Alchemy" (1891)
"The Wave Properties of Electrons" (1930)
"Mass and Energy" (1907)
"Nuclear Fission" (1940)
"Structure of the Proton" (1974)
Topic score
atomic
quantum
1880 1890 1900 1910 1920 1930 1940 1950 1960 1970 1980


1880 french france england country europe
1890 england france states country europe
1900 states united germany country france
1910 states united country germany countries
1920 war states united france british
1930 international states united countries american
1940 war states united american international
1950 international united war atomic states
1960 united soviet states nuclear international
1970 nuclear military soviet united states
1980 nuclear soviet weapons states united


Proprtion of Science
"Farming and Food Supplies in Time of War" (1915)
"Speed of Railway Trains in Europe" (1889)
"The Atom and Humanity" (1945)
"Science in the USSR" (1957)
Topic score
war
european
nuclear
1880 1890 1900 1910 1920 1930 1940 1950 1960 1970 1980

LDA还有弱化其它假设的扩展。相关主题模型[6]和弹球分配机器[24]将同时出现的主题视作相关（例如与地理有关的文档可能运动相关，但它更可能与化学相关）；球状主题模型[28]允许词不太可能在主题中出现（例如，“扭伤”显然不太可能出现有关“猫”的主题里）；稀疏主题模型进一步强化了主题分布的结构[37]；而“稠密”主题模型则是词数的一个更符合实际的模型[15]。

3.2 结合元数据

在文本分析配置中，文档通常包含些诸如作者、题目、地理位置、链接等其它额外信息。这些信息可以被用于适配主题模型。目前如何结合这些元数据是百家争鸣。

作者主题模型[29]是较早成功的例子。每个作者拥有一个主题直方图；多个作者的论文中的词由其中一个作者的主题直方图决定。作者主题模型允许从作者或文档进行推断。Rosen-Zvi等人在论文中展示利用作者的主题直方图计算作者间的相似性的例子，而LDA是无法胜任这一工作的。又比如，由于许多文档集合通常是相互链接的（例如科技论文相互引用或者网页相互链接），一些主题模型就考虑将那些链接用以估计主题。关系主题模型[13]假设所有文档都由LDA生成，文档间的链接取决于它们主题直方图的距离。关系主题模型不仅是新的主题模型，而且是新的网络模型，其与传统网络统计模型不同之处在于，它将用于为链接建模的节点属性（文档的词）考虑在内。

其它结合元数据的主题模型有语言结构模型[10]，关注语料库间的距离的模型[38]，命名实体模型[26]。更一般的方法包括狄立克雷多项式回归模型[25]和监督主题模型[7]。

3.3 其他类型的数据

在LDA中，主题是词上的离散分布，并用于产生文档中的词（观测值）。LDA的一个优势在于其主题参数和数据生成所用的分布，它们经过微调就可以适配于其它类型的观测值所对应的推断算法。LDA作为典型的主题模型，可以看作是分组数据的成员混合模型（mixed-membership model），而不只是将一组文档（观测值）与一个主题（部件）相关。每组文档都以不同的比例包含着不同的主题。为了适配诸如调查数据、用户偏好、声频和音乐、计算机代码、网络日志和社交网络这些多种多样的数据，LDA衍生出众多模型来处理和分析之。下面介绍两个成员混合模型已取得显著成功的领域。

在群体遗传学中，研究人员也独立地开发出了相同的概率模型，用以在个体采样得到的基因中寻找人类祖先（例如，人类从非洲、欧洲或中东等地起源）[27]。基本原理是每个个体的基因型是由一个或多个祖先群体遗传的。生物学家们通过与LDA非常相似的模型，描述了在这些人群中的基因模式（即“主题”），并辨认出单个个体的基因组成（即“主题直方图”）。这一模型如此有效的原因就在于即使具有“纯种”祖先基因的个体不存在，其基因模式依然可以假设，并通过实验得到。

LDA模型的推断算法还可用于自然图像的检索、分类和组织，因此LDA也被广泛地应用于计算机视觉中。研究者们已经从图像到文档做了一个直接的类比。在文档分析的假设中，每个文档包含多个主题，文档集中的所有文档共享同一个主题集。在图像分析的假设中，每副图像是多个视觉模式的组合，同一个视觉模式在图像集中不断重现（预处理阶段会分析图像以得到视觉模式（或者“视觉单词”）的集合）。主题模型在计算机视觉中被用于图像分类[16]，关联图像和字幕[4]，建立图像层次[2,23,31]等。

4 展望

主题模型是机器学习的新兴领域，有很多新方向亟待探索。

评价和模型验证 主题模型的评测和有效性脱节。一般的评价过程如下，首先取一部分语料做为测试集，然后从剩下的语料中训练不同的主题模型，并在测试集上度量其近似性（例如概率），最后选择性能最好的模型。但主题模型通常是用于组织、总结和帮助研究者探索大规模语料，技术上无法保证，准确性越高，组织性就越好或者解释得就越简单。主题建模的一个开放课题是与算法使用相匹配的评测方法。那么如何基于主题的解释性来比较主题模型呢？这就是模型验证问题，当面对一个新语料和新问题时，应该如何选择主题模型呢？哪些建模假设对问题是重要的，哪些是不重要的？该如何试验众多已经开发的主题模型呢？这些问题引起了统计学家的兴趣[9,30]，但他们对机器学习处理的问题的规模认识不足。这些计算问题的新答案将是对主题模型的重要贡献。

可视化和用户接口 主题模型另一个充满希望的未来方向是开发与主题和语料库交互的新方法。主题模型提供了探索大规模文本的新结构，那么如何使用这一结构呢？一个问题就是如何展示主题。主题一般通过列举其最常出现的词来展示（如图2），但选择不同的词展示或者以不同的方式来标记主题，可能会更有效。更进一步，如何更好地展示一个文档中的主题模型呢？从文档上来看，主题模型提供了文档结构的潜在的有用信息。结合有效的主题标记，读者可以辨认出文档中最感兴趣的部分。此外，隐藏的主题直方图隐式地将各个文档相互连接（考虑文档直方图的距离）。如何显示这些连接？整个语料与其推断的主题结构的有效接口是什么？

户接　　建模　常　　建模算法很有希望提　大规模　档的有意义的　　结构

本文调研了处理大规模文档的一套统计模型——概率主题模型。近年来，随着可扩展部件建模、后验推断的可扩展算法和大数据集的日益增多等非监督机器学习的有力支持，主题模型有希望成为总结和理解人们日益增长的数字化信息档案的重要部件。

参考文献

[1]Asuncion, M. Welling, P. Smyth, and Y. Teh. On smoothing and inference for topic models. In Uncertainty in Artificial Intelligence, 2009.
[2]E. Bart, M. Welling, and P. Perona. Unsupervised organization of image collections: Unsupervised organization of image collections: Taxonomies and beyond. Transactions on Pattern Recognition and Machine Intelligence, 2010.
[3] D. Blei, T. Griffths, and M. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. Journal of the ACM, 57(2):1–30, 2010.
[4] D. Blei and M. Jordan. Modeling annotated data. In Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 127–134. ACM Press, 2003.
[5] D. Blei and J. Lafferty. Dynamic topic models. In International Conference on Machine Learning, pages 113–120, New York, NY, USA, 2006. ACM.
[6] D. Blei and J. Lafferty. A correlated topic model of Science. Annals of Applied Statistics, 1(1):17–35, 2007.
[7] D. Blei and J. McAuliffe. Supervised topic models. In Neural Information Processing Systems, 2007.
[8] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003.
[9] G. Box. Sampling and Bayes' inference in scientific modeling and robustness. Journal of the Royal Statistical Society, Series A, 143(4):383–430, 1980.
[10] J. Boyd-Graber and D. Blei. Syntactic topic models. In Neural Information Processing Systems, 2009.
[11] W. Buntine. Variational extentions to EM and multinomial PCA. In European Conference on Machine Learning, 2002.
[12] W. Buntine and A. Jakulin. Discrete component analysis. In Subspace, Latent Structure and Feature Selection. Springer, 2006.
[13] J. Chang and D. Blei. Hierarchical relational models for document networks. Annals of Applied Statistics, 4(1), 2010.
[14] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990.
[15] G. Doyle and C. Elkan. Accounting for burstiness in topic models. In International Conference on Machine Learning, pages 281–288. ACM, 2009.
[16] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. IEEE Computer Vision and Pattern Recognition, pages 524–531, 2005.
[17] S. Gerrish and D. Blei. A language-based approach to measuring scholarly impact. In International Conference on Machine Learning, 2010.
[18] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In L. K. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing Systems 17, pages 537–544, Cambridge, MA, 2005. MIT Press.
[19] J. Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. Political Analysis, 18(1):1, 2010.
[20] M. Hoffman, D. Blei, and F. Bach. On-line learning for latent Dirichlet allocation. In Neural Information Processing Systems, 2010.
[21] T. Hofmann. Probabilistic latent semantic analysis. In Uncertainty in Artificial Intelligence (UAI), 1999.
[22] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. Machine Learning, 37:183–233, 1999.
[23] J. Li, C. Wang, Y. Lim, D. Blei, and L. Fei-Fei. Building and using a semantivisual

authors and documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pages 487–494. AUAI Press, 2004.
[30] D. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics, 12(4):1151–1172, 1984.
[31] J. Sivic, B. Russell, A. Zisserman, W. Freeman, and A. Efros. Unsupervised discovery of visual object class hierarchies. In Conference on Computer Vision and Pattern Recognition, 2008.
[32] R. Socher, S. Gershman, A. Perotte, P. Sederberg, D. Blei, and K. Norman. A Bayesian analysis of dynamics in free recall. In Neural Information Processing Systems, 2009.
[33] M. Steyvers and T. Griffths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2006.
[34] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581, 2006.
[35] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1(1–2):1–305, 2008.
[36] H. Wallach. Topic modeling: Beyond bag of words. In Proceedings of the 23rd International Conference on Machine Learning, 2006.
[37] C. Wang and D. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, Advances in Neural Information Processing Systems 22, pages 1982–1989. 2009.
[38] C. Wang, B. Thiesson, C. Meek, and D. Blei. Markov topic models. In Artificial Intelligence and Statistics, 2009.

标签：自然语言处理

眺望海接天
关注 - 4
粉丝 - 7
+加关注

2 0

« 上一篇： 使用MathJax在博客园里添加数学公式
» 下一篇： 转导推理——Transductive Learning

posted @ 2013-01-30 08:41 眺望海接天 阅读(6968) 评论(1) 编辑 收藏

评论列表

#1楼 2013-04-19 20:16 无脚的鸟
翻译不错。辛苦辛苦

支持(0) 反对(0)