PSJ
POLICY STUDIES JOURNAL

政策研究杂志,卷。 49,第1期,2021年

定性政策的主题建模和文本分析研究

卡罗莉娜·伊索阿霍 , 达里亚·格里森科和埃图·马克拉

本文促成了对政策研究有直接影响的关键方法论讨论:如何在不损害科学完整性的情况下将计算方法具体纳入现有的文本分析和解释过程。我们重点关注主题建模的计算方法,并研究它如何与两大类定性方法相互作用:以对作为交流单位的单词的兴趣为特征的内容和分类方法,以及以对交流行为的意义的兴趣为特征的话语和表示方法。

根据对最近使用主题建模进行文本分析的学术出版物的分析,我们的研究结果表明,将主题建模与两组方法相结合时,不同的混合方法研究设计是合适的。我们的主要结论是,主题建模使学者能够将政策理论和概念应用于更大的数据集。也就是说,计算方法的使用需要真正理解这些技术才能获得实质上有意义的结果。我们鼓励政策学者仔细反思方法论问题,并提供简单的启发式方法,以帮助在使用主题建模设计研究时识别和解决关键点。

关键词:主题模型、定性研究、机器学习、大数据、混合方法研究

本文为一项重要方法讨论做出了贡献,并可以直接影响政策研究,即源自不损害科学针对整体情况,将计算方法具体地制定政策文本分析和解读过程。我们将重点放在主题模型（TM)的计算方法,并研究它如何与以下两类确定性方法相互作用:其一为内容和分类方法,其特点是对作为交流单位的词汇感兴趣;其二为文字和表征方法,方法为关注交流行为的意义。通过分析最近使用主题建模进行文本分析的学术文献,发现将TM 与两组方法结合起来,我们应该主题采用不同的混合方法进行研究设计。我们的重点是,主题建模能够帮助研究学者将政策理论并概念性地应用到更大的数据集。尽管如此,使用计算方法时需要真正理解这些方法才能真正有意义的成果。我们鼓励政策更加仔细地思考方法问题,并提供一种简单的式启发算法,以识别和解决在设计主题建模研究时遇到的关键问题。

一、简介

自从 Harold Lasswell 对内容分析技术的方法论贡献以来,将文本作为数据进行分析一直是政策科学的重要组成部分

doi:10.1111/psj.12343

（Lasswell、Lerner 和 de Sola Pool,1952 年）。虽然后来的方法论发展建立在 Lasswell 的定量方法之上,并偏离了 Lasswell 的定量方法,专注于定性内容分析和话语方法来探索政策现象,但当今的政策学者正在经历并不得不对一种新的、有影响力的现象做出反应,即社会科学中所谓的“计算转向”。获得前所未有的规模和范围的数据促使越来越多的学者尝试使用计算算法作为文本分析的主要或补充方法 这是一项由研究人员传统上手动执行的费力活动（Mills,2018 年）。这些所谓的文本即数据方法代表了一大类计算技术（有关概述,请参阅 Grimmer 和 Stewart,2013 年）。近年来,主题建模 (TM) 的计算方法在政策相关研究中获得了关注。根据 Scopus 数据库,TM 目前是专注于政策分析的期刊中使用最广泛的文本挖掘技术（图 1）。由于主题建模日益流行,本文我们对主题建模的使用进行了仔细审查,并检验了其在定性政策研究方面的潜力。

TM 是一系列计算算法的统称,旨在“[发现]遍布大量非结构化文档集合的主题”（Blei,2012 年,第 77 页）。有人认为,TM 分析能够识别大样本中的主题（Murakami、Thompson、Hunston 和 Vajn,2017)并具有“高水平的实质性可解释性”（DiMaggio、Nag 和 Blei,2013）。因此,该方法“阅读”文本的能力在许多情况下被认为是合理的（Mohr & Bogdanov,2013）。

虽然这些特征对学者们很有吸引力,但应用TM的实践却零星出现。文本的自动分析已应用于多种来源,包括来自博客、Facebook 和 Twitter 等社交媒体平台的文本（Dehghani、Sagae、Sachdeva 和 Gratch,2014 年;Kim、Jeong、Kim、Kang 和 Song） ,2016）;新闻媒体（DiMaggio 等,2013; Grimmer,2010;Gritsenko，

图 1.近十年政策相关期刊中使用 TM 的文章数量,2008-2018 年。
注:基于 Scopus 数据库中的以下查询:(ALL [“topic model*” OR lda] AND SRCTITLE [policy]) AND PUBYEAR > 2007 AND PUBYEAR < 2019。[彩色图可在 wileyonlinelibrary.com 上查看]

2016）;政党宣言、演讲、新闻稿、立法提案和立场文件等政治和政策文本（Isoaho, Moilanen, & Toikka, 2019;Munksgaard & Demant, 2016;Quinn, Monroe, Colaresi, Crespin, & Radev, 2006;罗斯和堡盟，2016）。虽然一些学者更加关注 TM 相对于文本分析的方法论维度（Boussalis & Coan, 2016; Bryman, 2006; DiMaggio, 2015; Grimmer & Stewart, 2013），但大多数应用都没有对其应用进行批判性反思。研究设计 从语料库构建到结果解释。据我们所知,目前还没有关于 TM 在政策研究定性研究中的使用的英文摘要说明。因此,不存在既定的惯例,在这种惯例下（如果有的话)TM 可用于理解和解释社会科学研究中的政策文本。

在本文中,我们试图通过研究政策学者如何利用计算方法进行文本分析而不损害科学完整性来弥补这一差距。本文的主要目的是在两组文本分析方法的背景下讨论 TM:内容和分类 (C&C) 方法,其特点是关注单词作为交流单位,话语和表征 (D&R) 方法,其特点是关注交流行为的意义 (Titscher、Meyer、Wodak 和 Vetter,2000)。这两种方法之间的区别对于选择适当的研究设计似乎至关重要,因为它决定了认识论和策略层面的新方法论方法和组合。

本文的结构如下。下一节介绍并讨论了概率 TM 及其变体。随后一节考虑了 TM 在社会科学研究中的应用。它将 TM 与定性文本分析方法进行比较,以探索它们的潜在兼容性。然后,我们讨论了在不同混合方法研究设计中将 TM 与定性文本分析方法一起使用的情况,并举例说明了如何将它们应用于政策研究。最后,我们总结了讨论,并提出了一个简单的启发式方法来确定何时在定性政策分析中使用 TM 是合适的。

2. 概率主题建模

2.1.主题建模简介

主题建模是一系列计算算法的统称,这些算法用于对由较小的主题集产生的文档集合中的文本进行建模。这些算法涉及潜在变量发现、降维以及（软)聚类。由于它们通常不使用人工策划的主题种子,而是纯粹从数据中得出它们,因此它们通常被归类为无监督机器学习方法。然而,主题模型也有使用种子词作为主题的扩展（Jagarlamudi, Daumé, & Udupa, 2012），并且主题模型的机制也被纳入使用教学数据的分类算法中（Mcauliffe & Blei），2008;Ramage、Hall、Nallapati 和 Manning,2009），因此这种分类绝不是定义该方法的核心。

第一个明显属于该家族的算法是潜在语义索引（Deerwester、Dumais、Furnas、Landauer 和 Harshman，1990），它使用奇异值分解的矩阵分解技术从一组文档中导出潜在变量。由于它纯粹基于线性代数变换，LSI 最初对于如何导出潜在主题并没有一个清晰的人类可理解的解释，这阻碍了解释和分析。后来通过用概率术语构建因式分解来纠正这一点，这也导致了概率 LSI 模型的略微改进（Hofmann，1999），后来证明它相当于非负矩阵因式分解的最常见形式，这是另一种线性代数技术（丁、李和彭，2008）。最后，在潜在狄利克雷分配 (Blei, Ng, & Jordan, 2003) 中，pLSI 模型被扩展为一个可理解的、完整的生成概率模型。实际上，这意味着该算法包含一个模型，用于通过根据概率参数从一组主题中随机挑选单词来生成文档。重要的是，可以使用贝叶斯概率推理调整模型的概率参数，以更好地匹配证据（即现有的文档集合）。然后可以将调整后的参数读回作为该集合中主题的描述（Blei，2012）。

更具体地说，现代 TM 算法中编码的生成模型如下：首先，主题被建模为词袋，每个词袋内的每个单词的数量可变（例如，特定主题可能有许多单词副本，例如“学校”、“老师”和“学位”，但很少出现“树”或“小猫”等其他单词，因此我们可以假设该主题是关于教育的）。集合中的文档也被建模为这样的袋子，保存文档中的所有单词，而不考虑它们出现的顺序。然后，TM 算法尝试通过以下过程重新创建这些文档词袋：首先，从覆盖整个文档集合的主题集中，选择特定文档所属的一定数量的主题，以及它们在文档中的比例-ument（例如，该文件 67% 涉及教育，33% 涉及环境保护）。然后，通过按照先前选择的主题比例从每个主题袋中随机采样单词来重新创建文档词袋（Blei et al., 2003）。

TM算法在开始运行时，将每个主题包所含词汇的比例初始化为随机，每个文档所含主题的比例也初始化为随机，然后经过多轮贝叶斯推理，使这两个比例逐渐改变，使得模型生成的词包尽可能与实际数据中的词包相对应。

2.2.主题输出

TM 的输出包含两个项目：主题词比例和文档主题比例。在训练过程结束时，两个输出项目都可以从模型中读出并进行人工分析。表 1 和表 2 显示了典型的显示，其中主题以一组包含前几个单词的单词列表的形式呈现

表 1.主题建模输出示例,前五名术语。来自 Mallet 软件

话题	5 个热门术语
0	保护品种灰水员工上船
1	Tui AG 发展报告组
2	船舶安全邮轮培训
3	船上员工公司计划浪费
4	管理环境水系统环境
5	公主号乘客搭乘阿拉斯加游轮
6	嘉年华邮轮环保宾客
7	浪费客人的订单意味着经济
8	可持续能源员工社会德语
9	嘉年华英国邮轮公司和冠达邮轮公司

表 2.主题输出示例、主题-文档比例。来自 Mallet 软件。分析基于嘉年华集团和皇家加勒比邮轮公司在 2008-13 年发布的年度可持续发展报告

#doc	主题比例	主题比例	主题比例	主题比例				
0	7	0.47	4	0.28	9	0.23	3	0.01
1	8	0.38	9	0.36	4	0.22	3	0.02
2	9	0.47	3	0.32	4	0.18	0	0.04
3	6	0.42	4	0.29	9	0.27	3	0.02
4	2	0.50	9	0.35	4	0.12	8	0.01
5	9	0.52	0	0.22	4	0.15	3	0.11
6	9	0.41	5	0.34	4	0.17	3	0.05
7	1	0.52	4	0.36	9	0.11	8	0.01

来源:作者。

与每个主题关联的五个单词,通过显示与它们关联的前四个主题及其在文档中的比例来显示文档。

这些衍生主题与研究人员感兴趣的任何现象的对应程度取决于（i)文档集合可以被认为是由上述模型创建的（每个文档以一定比例处理一定数量的主题）;所讨论的主题决定了所使用的词汇）,（ii)上述模型中主题的定义与感兴趣的现象的对应程度,以及（iii)有关进一步定义的确切主题模型变体的各种假设主题如何表现。

作为此类假设效果的示例,基于隐含狄利克雷分配 (LDA) 的原始主题模型从对称狄利克雷概率分布中抽样了主题中的单词比例以及文档的主题比例。文档主题比例的对称狄利克雷分布能够很好地表示:(i) 每篇文档都提到所有主题的文档集合,(ii) 大多数文档提到少数主题的文档集合,以及 (iii) 文档之间提到的主题数量差异很大的文档集合。但是,当某个特定主题持续且广泛存在,而其他主题很少一起出现时（例如,关于欧盟政策各个分支的集合）,它无法很好地对集合进行建模

其中欧盟术语始终存在,但其他主题差异很大)。实际上,对称狄利克雷分布不太适合对行为方式不同的主题选择进行建模。

这并不意味着主题模型对此类材料完全不起作用,而是产生的主题会更糟糕。正因为如此,传统的 LDA 需要大量的预处理来删除一般语言单词(例如“the”、“and”等,通常称为“停用词”),因为这些单词会使模型感到困惑。另一方面,当非对称 LDA(Wallach、Murray、Salakhutdinov 和 Mimno,2009)中对主题表现相同的要求放宽时,模型能够将这些一般语言单词分离为单个主题,而不会对其他主题的质量产生不利影响。

类似地,传统的基于 Dirichlet 的 LDA 模型假设主题出现在文档中彼此独立。由于这一假设显然在实践中的大多数文本集中并不成立,因此开发了 LDA 的扩展。

相关主题模型 (CTM) (Blei & Lafferty, 2006) 用能够捕捉主题之间相关性的先验取代了主题比例先验(因此,例如,主题“地幔同位素地壳板块地球”经常与主题“断层地震数据地震图像”一起出现)。然而,虽然这提高了模型模仿原始文本的能力,并为研究人员提供了主题相关性数字作为可能有用的信息,但据报道,生成的主题本身不太容易被人类解释(Chang、Gerrish、Wang 和 Blei,2009)。这可能是因为关联主题意味着它们不再尽可能独立和独特。

扩展主题模型的另一个方向是将文本外部关联包含到模型中。动态主题模型(Blei & Lafferty,2006)添加了时间作为这种关联,从而能够进行图表和比较,例如,某个主题在不同时间如何在集合中讨论。其他扩展(Mimno & McCallum,2008;Rosen-Zvi、Chemudugunta、Griffiths、Smyth 和 Steyvers,2010)添加了类别关联,允许按作者、群体或政治倾向等进行比较。作为 TM 的最终有用扩展,结构主题模型 (STM)(Roberts 等人,2014)将 CTM 的相关主题与来自 DTM 和其他属性关联模型的文本外部关联相结合。STM 还提供统计模型评估工具,包括用于选择要提取的最佳主题数量的启发式方法。

最后,必须注意的是,在实践中,在运行主题建模算法之前,其输入通常会经过大量的预处理,旨在统一词汇并使其更适合主题推断。除了已经提到的停用词删除之外,通常采取的步骤包括删除所有标点符号、将所有变格词替换为其基本形式、将所有数字替换为单个标签以及统一使用不同大写字母书写的相同单词。虽然严格意义上来说,这些预处理步骤超出了主题建模算法本身的应用范围,但可能会对其结果产生很大影响,因此也应进行严格的检查和验证(Denny & Spirling,2018)。这一点尤其重要,因为最近的研究表明,许多普遍接受的预处理步骤没有甚至产生不利影响

主题建模（Schofield、Magnusson 和 Mimno,2017;Schofield、Magnusson、Thompson 和 Mimno,2017）。

2.3. 模型验证与解释

运行主题模型时,需要三个相互交织的步骤来确保分析的有效性。首先,需要解释模型的输出。

其次,需要验证预处理和建模参数的选择。
最后,主题对所调查现象进行建模的能力需要
有待评估。

不幸的是,目前许多研究放弃了严格的评估,并采用了有缺陷的解释程序。首先,通常的做法是独立于文档并仅基于与集合中与主题最相关的前 5 到 20 个单词来解释主题输出。有人可能会认为这种做法是有缺陷的,因为主题是对文档集合的描述,并且不能脱离文档集合而存在。因此,它们也需要不是孤立地解释,而是在这些词最初出现的文档的上下文中进行解释。此外,仅用前 N 个单词来总结文本材料中出现的所有单词的实际分布可以隐藏重要且有趣的信息。

除了歪曲实际内容外,这种做法还可能隐藏模型参数化方面的问题。例如,如果主题数量或预处理参数设置不正确,生成的主题最终可能会成为多个不同主题的混合体。然而,特别是考虑到人类思维能够在任何地方找到联系的能力,这可能无法仅从合并主题中的前五个单词中看出。简而言之,孤立地解释短单词列表会导致误解所识别的主题在句子中的含义

文件。

与此同时,有大量的先前研究支持对主题模型输出的严格验证。迪马吉奥等人。 (2013, p. 586) 确定了三种验证形式;统计、语义和预测。其中,统计验证通常从分析模型预测一组文档的能力开始。

通过比较使用不同参数运行的模型之间的统计困惑度估计,可以找到主题数量以及其他参数的更好值。除了寻找最佳参数之外,统计验证还可以用于评估数据与模型中统计假设的吻合程度（Mimno & Blei,2011）。 Tang、Meng、Nguyen、Mei 和Zhang（2014)还使用统计方法和理论分析来探索主题建模在文档数量及其长度方面的局限性。从他们的实验来看,主题建模需要至少 100-200 个单词长度的文档,而文档数量至少需要 1000-2000 个。此外,如果与不同主题相关的词汇集清晰分开,主题建模效果会更好。输出的质量还很大程度上取决于研究人员选择正确数量的主题,这进一步凸显了严格验证该参数值的必要性。

同时,相关主题模型的可解释性较低,尽管困惑度较低（Chang 等人,2009）,但统计指标不应盲目地作为验证模型及其参数选择的唯一方法。相反,正确的验证需要结合统计、语义和外部指标（DiMaggio，2015）。在语义验证中,研究人员反思模型输出以评估其可解释性和合理性。一种严格的方法是手动编码数据样本并将这些结果与模型输出进行比较,Boussalis 和 Coan（2016,第 94 页)将其称为与语义验证不同的并发验证。

最后,在外部验证中,研究人员验证模型是否反映了相关的馆藏外部信息。例如,可以验证主题流行度以响应相关的收集外部事件,例如新闻关注周期或政治辩论时间表。

作为这些不同形式的验证和解释如何相互有益的例子,Boussalis 和 Coan (2016) 使用统计指标检查主题之间的“语义”距离,从而提高了其结果的语义有效性。同样,Mimno 和 Blei (2011) 使用与模型假设的统计偏差度量来帮助解释主题以及各种系数的影响。

这些方法的有用性始终取决于每个单独研究的研究目标,因此研究人员必须评估如何在自己的工作中使用每个验证措施（Boussalis & Coan,2016）。尽管如此,关键的经验法则适用于所有希望在文本分析中使用主题建模的学者：“验证、验证、验证”（Grimmer & Stewart,2013）。用 Grimmer 和 Stewart (2013, p. 5) 的话来说,这句话经常被引用,但值得重复：

> 模型的输出可能会产生误导,甚至完全错误。因此,研究人员有责任验证他们对自动文本分析的使用。（…)因此,应该避免盲目使用任何没有验证步骤的方法。

3. TM在社会科学研究中的应用

在现有文献中,TM 经常被等同于社会科学研究中常用的多种文本分析方法。然而,正如引言中指出的那样,区分主要关注作为交流单位的词语的方法和关注交流行为意义的方法很有用。这种区别使我们能够更仔细地讨论 TM 与 C&C 和 D&R 方法组,同时考虑到它们的认识论和实践考虑。表 3 总结并比较了我们分析中考虑的每种方法的特征。

如表 3 所示,C&C 方法中的“分析流程”更接近 TM 的过程,而 D&R 方法与 TM 的使用有更显著的差异。

表

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible] s/	[illegible]	[illegible]	[illegible]、[illegible]
	[illegible]	[illegible]	[illegible],[illegible]	
	[illegible]	[illegible]-	[illegible]	
	[illegible]	[illegible], [illegible]隐	[illegible]关	
[illegible]		[illegible]标	[illegible]内	[illegible]; [illegible]
	[illegible]	[illegible];	[illegible]、[illegible]	
	[illegible] 析	[illegible]、	[illegible]、	
	[illegible]	[illegible]	[illegible],	

因此,虽然 TM 越来越多地应用于社会科学研究,但目前理解 TM 及其在文本分析中的输出的尝试各不相同,并且存在误导的风险。

尽管 Blei 最初提出 TM 是一种提供“浏览体验”的工具或“一种管理、组织和注释大量文本档案的算法解决方案”(Blei,2012,第 77–79 页),但 TM 越来越多地被用作一种方法来揭示话语环境(DiMaggio 等,2013;Goldstone 和 Underwood,2014;Munksgaard 和 Demant,2016)、语义或主题类别(Jaworska 和 Nanda,2016;Mohr 和 Bogdanov,2013)、问题定义(Nowlin,2016)、叙述(Grubert 和 Algee-Hewitt,2017)、框架(DiMaggio 等,2013)和作者特征(Seroussi、Zukerman 和 Bohnert,2014)。显然,这些解释缺乏共同的连贯性,并提出了一个问题:TM 的输出是否能够代表如此多样化的现象和概念。这是一个关键的方法论问题,因为正如我们之前所解释的那样,除了保证模型的稳健性和有效性之外,使主题模型可解释的决定性要求之一是使主题与感兴趣的现象相对应。

理论驱动的 D&R 方法的输出尤其包括上下文和语义理解。有人认为,数据驱动的 TM 掌握这些方面的能力有限,因为它只分析文档语料库中包含的单词。正如 Klein、Eisenstein、Sun 和 Jacko(2015 年,第 132 页)恰当地指出的那样,

> “除了这些词组经常表现出的暗示性相似性之外,没有任何内在理由相信根据共现统计分组的单词应该真正意味着或证明任何事情。要真正整合主题模型 (…),用户必须能够探究模型提出的语义关联,并寻找模型本身的其他视角。”

事实上,虽然目前正在开发和研究上下文的算法探索,但目前的 TM 并没有适合于检查作为 D&R 方法核心的隐藏权力关系、代理和上下文。

此外,D&R 方法明确承认主观性。文本分析中出现的散漫故事情节、框架或叙事故事总是由研究人员构思的,研究人员心中有一个特定的问题,并且理解给定文本的读者应该知道的隐含语境和背景知识因素(Hajer,1995;Schön & Rein,1994)。

TM 和 C&C 方法均基于实证假设,分析过程以文本为基础,重点关注显性项目(单词、句子、段落、文档)。然而,TM 提供的是生成模型,而内容或主题分析旨在“通过编码和识别主题或模式的系统分类过程对文本数据的内容进行主观解释”(Hsieh & Shannon,2005,第 1278 页)。因此,TM 的潜力可能最好在混合方法设计中得到充分利用,而不是从 TM 分析中直接获得定性价值。

4. 混合方法设计中的主题建模

我们现在讨论两种混合方法,嵌入式设计和顺序设计,并给出将 TM 应用于文本分析的良好实践示例。
为了证实我们的方法论论证,我们报告了最近发表的 25 项研究,这些研究应用 TM 在定性环境中进行文本分析。
这些研究是使用 Scopus 数据库确定的,并由作者审查（支持信息中的附录 2）。表 4 总结了将 TM 与 C&C 和 D&R 方法结合使用的注意事项和限制。

4.1.嵌入式设计

嵌入式设计意味着项目有一个主导方法来指导分析,并有一个辅助方法来增强这一过程（Creswell,2003）。这种设计的好处是,研究人员可以从更广泛的角度看待给定的问题,使用不同的方法来研究设计中的不同层次,或者以不同的方式处理材料的不同方面（参见 Creswell & Clark,2007,第 230 页）。我们认为,通过应用嵌入式设计,内容、主题和词汇分析可以与 TM 相结合（图 2）。

内容分析。TM技术可以对内容分析进行有益的补充。
Baumer、Mimno、Guha、Quan 和 Gay (2017) 认为,TM 与扎根理论（一种流行的内容分析归纳实现)惊人地相似,表明将这两种方法应用于同一研究问题可以获得相似的结果。他们认为,两者的主要相似之处体现在“策略”层面 即基于数据迭代细化的临时理论（就扎根理论而言)或模型（就 TM 而言）。然而,我们认为,GT 与 TM 的不同之处恰恰在于策略层面:GT 旨在提出理论（对某些结果的解释）,而 TM 提供的是模型（对某些变量的检查）。但它们确实有相似之处。首先,这两种方法都是迭代的。定性内容分析中对编码类别的不断比较和修订类似于 TM 中通过贝叶斯信念操作迭代合并的主题。其次,在这两种方法中,同一个词可以包含在多个类别中。然而,与扎根理论不同的是,TM 将文本中的所有单词都考虑在内,从而创建了一种可靠且可复制的方法来解构整个文本集合。

可以使用以下将 TM 嵌入到内容分析中的策略。
研究人员可能会在初始编码过程中考虑 TM 输出,这是一种以系统方式筛选大量数据的程序,通常在扎根理论方法框架内实施（Glaser & Strauss,1967）。事实证明,具有“计算扭曲”的扎根理论可以在探索阶段利用 TM,在模式细化阶段利用深度阅读,在模式确认阶段利用监督机器学习方法,以提供强大、严格且可重复的方法框架（Nelson,2017）。虽然扎根理论明确旨在开发小范围或中范围理论,并持有

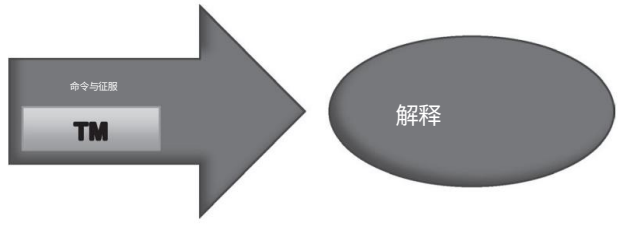

图 2.嵌入式设计中的内容和分类方法以及主题建模。

除了采样和质量控制的处方外,相同的步骤也可用于其他受益于内容分析归纳方法的情况（Boussalis & Coan,2016;D Amato et al.,2017）。

主题分析。主题分析通常被描述为“一种识别、分析和报告数据中模式（主题)的方法”（Braun & Clarke,2006）。它与各种理论兼容,可以是归纳的或演绎的,并且它采用编码作为处理数据的技术程序。主题分析并不试图量化主题、计算其频率或建立任何其他重要性代理（Vaismoradi、Turunen 和 Bondas,2013）,它通常用于具有大型数据集的研究项目的初始阶段,而解释方法随后可以应用于其特定部分并回答更具体的研究问题。

TM 已被用作主题分析的一种形式,以估计多维政策空间内特定问题的显着性（Nowlin, 2016）。克莱因等人。 （2015)在他们的 TOME（交互式主题模型和元数据可视化)模型中结合了 TM（他们将其视为自动主题分析）、意义构建和交互式可视化。他们的工作说明了 TM 对于这项任务的有用性。同样,Greene 和 Cross（2017)使用 TM 来揭示欧盟议会辩论的主题。自然语言的统计规律似乎呈现了相当准确的内容结构图景。尽管如此,需要强调的是,模型生成的主题还可能识别特定的历史事件、显着的文体特征或系统的转录错误,仅举几个非主题主题。

与主题分析相关的主要挑战是定义主题是什么以及给定数据集中有多少个主题。 TM 不仅允许探索更大的语料库,还可以以更系统的方式发现主题。因此,如果研究人员想知道数据集的含义,则可能存在高度收敛,甚至有可能用 TM 替代主题分析（Murakami 等,2017）。这两种方法都以文本为基础,它们的输出是通过将数据集分解为单独的主题而获得的总体描述。他们都不关注语言的使用或代理（即谁在说话）。然而,这种集成要求对主题的解释不仅要考虑前 N 个单词,还要考虑主题中的前 N 个文档。只有通过检查与主题相关的文档,研究人员才能捕获主题,该主题定义为数据集中的潜在含义,捕获与研究问题相关的数据的重要信息（Braun &

Clarke,2006）。虽然 TM 可以嵌入到主题分析中,但后者不能是一种纯粹的数据驱动方法,因为采样、语料库的管理和与语料库相关的主题的解释至关重要。

词汇分析。词汇分析是一组专注于“社会集体常用的词汇系统及其含义”的技术

（Loewenstein,Ocasio,& Jones,2012,第 41 页）。其主要假设是“意义并不存在于词语中,而是存在于词语与其语言环境的关系中,即词语与其他词语的关系中”（Krippendorff,2004,第 290 页）。
最常见的是,词汇分析与单词到单词和单词到示例关系（即文本内的关系网络)的识别相关。

词对词分析（有时称为地图分析或情景概念分析)旨在详细阐述由相互关联的概念组成的网络,而不是频率计数或潜在主题（Carley,1994）。实际上,研究人员试图确定某些单词是否与其他单词相关

在一组文本中。 “文本中概念分布的差异以及它们之间的关系提供了对文本内容和结构的相似性和差异的见解”（Carley,1993,第 92 页）。生成的“地图”中的词与词之间的关系可以通过图形和统计方式进行检查（Carley,1993）。

因此,基于词共现的词对词词汇分析在思想上与TM非常接近,尽管它们的技术实现有所不同。 Carley（1993，p.102)指出，“地图分析技术的一个缺点是它们更难自动化。”在计算机辅助文本挖掘的早期,自动化已经应用于概念编码,但词到词词汇分析所需的关系编码尚未实现自动化。 TM 是一种工具,通过为单词分配属于同一主题的概率,以及提供其他感兴趣的指标（例如主题分布之间的距离）,自动对单词之间的关系进行编码。

词对例分析旨在解决词对词分析在符号基础和意义整体论问题方面的局限性（Loewenstein 等人，2012 年）。这种分析的一种类型 由 Mohr 和 Duquenne（1997 年)提出的实践视角 考察词语和实践的共现。其他学者研究了专有名词（公司名称 [Kennedy,2008]、项目名称 [Nigam & Ocasio,2010])与词汇的关系。这种方法显示了什么最能体现数据集中的某些类别,反之亦然,哪些词汇用于描述某些示例。我们认为 STM 可用于词对例分析,因为它允许使用背景变量来区分主题（Chandelier、Steuckardt、Mathevet、Diwersy 和 Gimenez,2018 年;Lucas 等人,2015 年）。

因此,至少在理论上,TM 可以嵌入词汇分析中,因为它允许将文本集合表示为词汇集合,并且检查包含热门词的文档部分可以提供派生主题的关键词上下文 (KWIC) 视图。动态主题模型甚至提供了研究词汇如何随时间演变的机会 这项任务通过手工编码来执行是出了名的费力。由于我们还没有遇到过这种情况

在撰写本文时,这种混合方法设置还处于早期阶段,设计需要进一步测试。词汇分析和 TM 方法之间的差异在于研究策略层面:虽然执行词汇分析的分析师可能希望专注于特定词汇,但 TM 算法以主题词比例的形式提供大量词汇,这些词汇可能与研究主题相关,也可能不相关。然而,我们认为词汇分析的假设 属于一起的词系统具有特殊属性(即,词是其他词的上下文),允许研究人员弥合词作为显性词汇单位与其隐性含义之间的差距 可以扩展到 TM。这使得使用 TM 增强词汇分析的途径看起来最有希望。

4.2.顺序设计

序贯设计是指首先分析定性数据,然后分析定量数据(反之亦然),以理解给定的研究问题。在此设计中,数据在研究过程的连续阶段中单独分析(Creswell,2003)。虽然定性分析优先,但定量结果对于指导定性分析过程也很重要。顺序设计的好处是它允许两种方法相互补充,并有助于更稳健的数据分析过程。

图 3 演示了如何在序贯设计中将 TM 与 D&R 方法结合起来。鉴于 D&R 方法的应用在实践中很少遵循线性路径,而是包括数据和新兴解释之间的多轮迭代,因此应用 TM 来指导和告知这一过程还有空间。我们现在考虑在顺序设计中使用 TM 和不同的 D&R 方法的方法。

话语分析。话语分析方法阐明了现实是如何通过知识生产和意义建构而进行社会建构的 (Keller, 2013)。
正如我们所解释的,由于话语方法具有强大的理论基础,早期研究人员应该警惕将主题输出直接解释为话语,但主题模型在顺序环境中使用时可以帮助揭示话语的各个方面。

例如,一些研究人员使用 TM 来指导话语分析的第一步,这通常包括确定数据范围并进行第一轮编码以熟悉数据。Törnberg 和 Törnberg (2016a, 2016b) 将 TM 与批判性话语分析相结合,以研究在线讨论中伊斯兰恐惧症和反女权主义之间的话语联系。作者在研究的第一阶段使用 TM 来获得其语料库的归纳经验分类,以便进行后续的话语分析。他们发现这种顺序设计非常丰富,因为 TM 归纳地揭示了存在某些话语领域的段落 (Törnberg & Törnberg, 2016a, p. 133)。同样,Light 和 Cunningham (2016) 和 Lindgren (2018) 使用顺序两阶段方法来研究媒体话语。在这两项研究中,TM 首先用于提供

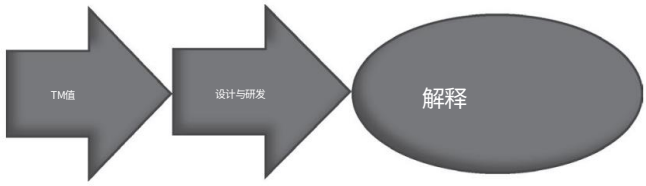

图 3.顺序设计中的话语和表示方法以及主题建模。

语料库概述。 Light 和 Cunningham 使用 TM 来研究几十年来诺贝尔和平奖演讲中如何讨论国际和平运动。 TM 用于定位诺贝尔演讲中的主题,主题输出作为“内容代码”,然后在文本的定性话语分析中打开。因此,TM 可以告知（并且在某种程度上自动化)话语分析的第一步,例如检查某些单词的常见搭配,然后将它们放入主题类别。

除了为研究过程提供信息和指导外,在话语分析中应用 TM 还可以增加分析的严谨性。Törnberg 和 Törnberg (2016a) 以及 Jaworska 和 Nanda (2016) 发现,应用 TM 可以利用与文本编码相关的选择偏差 这是应用这些方法的关键问题之一。TM 的另一个关键优势是它能够列出与某些主题最密切相关的文档。这使得研究人员能够确定一组较小的文档以进行后续的话语分析,从而使他们能够超越主题结构并考虑超出单词分析的文本间和语境因素 (Lindgren, 2018)。总之,作为一种独立方法,TM 在进行话语分析方面受到限制。但是,当单独应用并充分考虑其局限性时,该方法可以提供有用的见解。

框架分析。框架通常被定义为“连贯的解释包”
（Entman,1993;Gamson 和 Modigliani,1989）。按照这种观点,围绕某个主题出现的模式、元素和假设构成了一个框架。一旦识别出来,就可以通过解释其含义或权衡其有效性来进一步分析框架。自动化帧识别的能力让许多学者探索TM是否可以简化帧分析中的分析过程（Jacobi, van Atteveldt, & Welbers, 2016; Pashakhin, 2016）。作者转向TM,认为通常用于框架发现的传统内容和话语分析方法在面对大量文本时受到限制,并且在文本选择和分析方面也容易出现主观性。

然而,虽然 TM 分析中规模和范围的好处显而易见,但将 TM 与框架分析结合起来会带来许多问题。Van Atteveldt、Welbers、Jacobi 和 Vliegenthart（2014 年)以及 Jacobi 等人（2016 年)研究了主题是否可以被视为应用框架,并得出结论,主题输出不符合框架作为连贯解释包的方法论定义。相反,有些主题比其他主题与实质内容联系更紧密,主题解释取决于案例和研究问题。此外,TM 不

掌握框架分析的一个关键部分,即考虑框架中未表示的内容以及在遵循固定框架时文本中遗漏的内容。

尽管如此,TM 可以为框架分析提供有用的经验途径。有些主题“是非常高精度的问题指标”,这“开辟了使用选定主题作为实质性指标的可能策略”(Van Atteveldt 等人,2014 年,第 2 页)。
Jacobi 等人(2016 年)将 TM、情绪分析和定性框架分析结合起来,以提供更好的框架和互文性结果。总之,可以通过应用 TM 来启动框架发现的基础,而 TM 的实证用途在于它允许分析主题的连续性。然而,必须注意的是,TM 仍然是一种不完整的独立框架分析方法。

叙事分析。叙事分析是指将故事作为调查对象的注意力结合起来的技术。叙述可以指生活故事;对生活和事件的扩展描述;以及围绕人物、背景或情节组织的故事(Riessman,2005)。该方法旨在强调文本中的顺序和结果,而不是挑选脱离其更广泛上下文的文本片段(Mishler,1995)。由于研究人员在叙事分析中的作用非常重要,即从文本中提取叙事,建立在隐含的语境和背景知识的基础上,TM 在方法论上似乎与叙事分析相距甚远。

然而,有研究建议在叙事分析中使用 TM。
Jaworska 和 Nanda (2016) 以及 Grubert 和 Algee-Hewitt (2017) 建议将主题输出解释为叙事的信号或叙事的潜在前景。
首先,他们认为 STM 利用文档中的元数据的能力可以使其有助于揭示叙述主题并将其与文档作者的信息联系起来。通过这种方式,STM 可以为揭示潜在叙事以及识别属于叙事故事的角色(代理人)提供一个起点。

此外,诸如 STM 之类的动态主题模型使得追踪语料库中主题的时间演变成为可能。这种分析可用于了解元叙事和叙事序列结构的变化。

STM 分析可以帮助研究人员识别叙述的开头、中间和结尾。随后,STM 输出可用于指导对叙事结构进行更详细的分析。然而,只有在明确确定数据所涵盖的时间范围的情况下,STM 的应用才有用。重要的是,TM 输出不应被解释为连贯的叙述。我们看到 STM 作为叙事方法补充的特殊潜力。

5. 政策研究申请

最后,我们提请大家注意三项将 TM 与政策概念和理论相结合的研究。在此过程中,我们并不试图详尽地介绍这些联系,而是深入了解如何将 TM 应用于政策概念研究。

Nowlin (2016) 是一项受益于嵌入式设计的研究示例,用 TM 代替了主题分析。该论文将 LDA 应用于大量国会听证会,以生成问题定义模型。这个型号

建议问题定义是多维的,即具有多个属性,但“只有问题的某些属性（即维度)可能在任何时间点都是突出的”（Nowlin,2016,第312 页）。当参与者操纵问题定义的维度时,他们会以特定的方式构建（或重新构建)问题。随着时间的推移,这些框架聚集起来并成为集体问题定义。因此,要研究“问题定义”,需要确定问题可能的“维度”、它们何时以及如何部署,以及这些过程如何演变。这可以通过进行主题分析来完成 识别与给定研究问题相关的主题（在这种情况下,定性研究人员会问“哪些特征与所讨论的政策问题相关？”并仅对那些特征进行编码)与回答此问题相关的部分材料）。 Nowlin（2016,p.320)的研究假设主题“与问题的维度同义”。通过文件中主题的流行程度来突出问题的突出性。本研究的实际步骤是:

1. 构建一个处理一个具体问题（废核燃料)的语料库
   1975 年至 2012 年间,
2. 将 LDA 应用于该语料库,
3. 将主题输出解释为与每个维度最相关的术语,
4. 标记每个维度并观察主题随时间的比例（Nowlin,
   2016,第320页）。

作者深入思考了语料库的构建,展示了所选材料（出席国会听证会的证人的开场陈述)如何代表这些参与者对废核燃料问题的看法。结果验证是通过将研究结果置于“外生事件”（即立法变化）的背景下进行的。使用以立法行为作为自变量的简单 OLS 回归来检查语料库中每个主题的平均比例（按年份汇总）。研究得出的结论是,该方法成功地捕捉了不断变化的问题定义。虽然这是一种有趣的方法,但作者反思了它的主要缺点 LDA 无法对基于对称先验假设的一系列主题进行建模,这些主题的行为方式并不相同（请参阅我们之前的讨论）。该研究将这一缺陷称为缺少“与重新构建政策问题的特定尝试相关的一些细微差别”（Nowlin,2016 年,第 326 页）,但正如我们在本文中所讨论的,它代表了方法论问题:感兴趣的现象和 LDA 算法的技术假设。

因此,根据TM的多方面有效性标准,用于政策分析的嵌入式主题建模设计（例如上述研究)将受益于依赖原始文档来解释主题输出。

此外,由于 LDA 假设整个语料库中存在相同的主题,因此研究人员通常应该谨慎使用该方法来捕获新主题的出现,而可以使用具有宽松假设的算法（例如 STM)来利用新主题。有趣的是,Gilardi、Shipan 和 Wueest (2017) 的工作提供了使用 STM 调查问题定义的示例。在研究美国各州对吸烟的限制时,作者使用了主题

一个国家内的流行程度作为问题定义的衡量标准。他们首先利用包含 49 个州反吸烟法的超过 300 万个段落的数据集来确定主题及其随时间的分布,然后对主题进行四次有效性检查（其中包括情绪分析并考虑主题如何影响吸烟）。与州一级通过禁烟令的时间相关）。

最后,作者进行了概念讨论,认为“主题随着扩散网络中各州先前采用政策的频率而变化”且“主题流行度与一州扩散网络中先前的政策采用相关”（Gilardi 等人,2017 年,第 13 页）,从而阐明政策制定过程的问题定义阶段是否以及如何与先前的采用相关。虽然该设置与 Nowlin 的工作无法直接比较,但它显示了使用 STM 而非 LDA 的一个明显优势。Gilardi 等人在他们的模型中加入了几个协变量:这使他们能够检查哪些主题出现最频繁,哪些主题随时间保持稳定,以及主题流行度的其他变化。但是,要使这种方法发挥作用,政策必须广泛实施,以便创建足够数量的语料库。

另一个例子是Fawcett、Jensen、Ransan-Cooper 和 Duus (2018) 在帧分析中顺序应用 TM。这项研究的目的是解释为什么问题会从问题流中“消失”,从而提高我们对“问题窗口”如何形成和变化的认识。作者将 TM 与传统的手工编码一起应用来跟踪框架和框架发起者（即影响问题化的强大精英参与者)的共同进化。为此,他们使用了序贯设计,步骤如下:

1. 为主题建模捕获并准备相关数据;
2. 进行 TM 分析,作者从该模型中选择了一个 15 个主题模型和一个政策相关主题;
3. 从所选主题中确定了四个问题窗口;
4. 框架和框架赞助商的存在是在步骤 1 中确定的四个问题窗口期间发布的新闻文章中手工编码的（Fawcett 等人,2018 年,第 7 页）。

通过这种顺序设计,Fawcett 等人。我们能够识别相关时间段并放大这些时间段,然后对特定时间发布的新闻文章进行进一步的定性分析。他们令人信服地认为,这种研究设计比使用“文章频率计数作为‘放大’特定时间段的理由”的传统方法有好处（Fawcett 等人,2018 年,第 6 页）。TM 允许他们从实际问题化时期选择文章,而不是那些源于“其他类型的无关报告”的文章（Fawcett 等人,2018 年,第 6 页）。

六,结论

本文旨在讨论 TM 在定性环境中对文本分析的贡献。其目的是向政策学者揭示 TM 背后的逻辑,并强调制定稳健研究设计的学术能力

TM 取决于研究人员对技术、技术假设的熟悉程度以及对所研究现象的良好了解。

本文提出的关于TM的第一个要点是它的输出包含两项:主题词比例（通常以一组词列表的形式呈现)和文档主题比例。我们强调,这些衍生主题与研究人员感兴趣的任何现象的对应程度取决于：(i)文档集合可以被认为是由 TM 底层的生成模型创建的，(ii))该模型中主题的定义与感兴趣的现象的对应程度如何,以及 (iii)有关所使用的确切主题模型变体的各种假设,这些假设进一步定义了主题的行为方式。在 TM 分析可信之前,必须了解、理解和验证上述所有内容。

根据上述定义,我们讨论了 TM 以及常用的 C&C 和 D&R 方法的分析过程。关于C&C方法,我们认为TM算法可以自动化该组的分析过程。根据方法的不同,它可以完全（如主题分析)或部分（如归纳内容分析和词汇方法)替代以前“手工”执行的程序。这之所以成为可能,是因为 TM 共享相似的实证主义假设、以数据为基础,并专注于文本作为明确含义的体现。然而,在没有严格考虑方法学目标的情况下,不应进行此类替代。我们鼓励学者进一步对 TM 进行实证实验,以测试 TM 能够在多大程度上自动化 C&C 流程。 TM 方法的扩展克服了原始 LDA 算法的一些限制,似乎特别适合此目的。

关于TM和D&R方法,第一个决定性的发现是TM的输出不应等同于话语、框架或叙述。后面的概念高度依赖于理论,传达上下文、表示和语义含义,而这些目前尚不在基于贝叶斯概率的算法的范围内。因此,我们的启发表明,由于认识论方面的考虑不同,TM 在分析程序方面不能替代 D&R 方法的任何部分。尽管如此,虽然 TM 的未来发展可能允许训练算法来提供有关话语、框架和叙述的信息,但我们认为,在目前的状态下,TM 作为D&R 方法的补充可以增加价值。 TM 可以作为混合方法设计的一部分顺序集成到分析中。在这种情况下,TM 不是用来代替“手工”分析,而是作为一种补充:例如,在执行给定定性方法的所有步骤之前识别代表性文本或检查叙述结构。

最后,我们得出结论,在文本分析中使用 TM 对政策研究有直接影响。TM 最明显的好处是它有助于繁重的文本分析工作。因此,现有的政策概念和理论可以应用于庞大的数据集。除了规模和范围的好处之外,应用 TM 还为新方法论和组合提供了途径,使政策研究人员能够以新的方式处理政策概念,例如政策问题定义和问题窗口。特别是,TM 的特征

主题的时间性和连续性可以提供有用的解释途径。
然而,由于主题解释在很大程度上取决于研究的案例和背景,因此需要谨慎应用不同的协同作用。我们建议研究人员在设计使用 TM 的研究时考虑以下问题:

1. 不同TM变体的技术假设如何与感兴趣现象的具体情况相一致,以及如何最好地设置预处理和建模参数;

2. 所编制的语料库能够可靠地回答哪些问题以及如何回答
   脓液的大小和处理会影响潜在的结果;
3. 如何考虑输出的两个部分(单词/主题和主题/文档)
   参与评估感兴趣的现象;
4. 如何加强、验证和批评对主题输出的解释
   根据文献集合进行计算。

我们希望这种启发式方法能够帮助识别和解决关键点,从而帮助研究人员开发新颖的混合方法设计,在不影响方法稳健性的情况下释放TM在定性政策研究中的潜力。

Karoliina Isoaho是赫尔辛基大学(芬兰)社会科学学院的博士生。

Daria Gritsenko是赫尔辛基大学(芬兰)亚历山大研究所和赫尔辛基数字人文中心的助理教授。

Eetu Mäkelä是赫尔辛基大学(芬兰)赫尔辛基数字人文中心的助理教授。

笔记

1. 2008 年至 2018 年间,使用流行文本挖掘技术(情绪分析、词嵌入、监督学习、文本聚类、文本分析或 lda/非负矩阵分解/主题模型)发表的 160 篇文章中,49% 使用了主题建模。该数据基于作者在 Scopus 数据库中的关键词搜索(有关完整搜索查询,请参阅支持信息中的附录 1)。

2. 本文并非旨在全面或详尽地列出所有文本分析方法。有关可用于内容和分类或话语和表示的其他文本分析方法的详细信息,请参阅 Denzin 和 Lincoln (2007)、Kuckartz 和 McWhertor (2014) 或 Silverman (2016)。

3. TITLE-ABS-KEY (“主题模型*” OR “LDA” AND 文本*) AND (LIMIT-TO (DOCTYPE, “ar”) OR LIMIT-TO (DOCTYPE, “ip”)) AND (LIMIT-TO (SUBJAREA,“SOCI”))并精心挑选与政策研究的相关性。

参考

Baumer, Eric PS, David Mimno, Shion Guha, Emily Quan 和 Geri K. Gay。2017 年。“比较扎根理论和主题模型:极端发散还是不太可能收敛?”信息科学与技术协会杂志68 (6): 1397–410。

布莱,大卫·M.2012。 “概率主题模型。” ACM 55 通讯:77-84。

Blei, David M. 和 John D. Lafferty。2006 年。“相关主题模型。”神经信息处理系统进展18:147–54。

大卫·M·布莱 (Blei)、安德鲁·Y·吴 (Andrew Y. Ng) 和迈克尔·乔丹 (Michael I. Jordan)。 2003.“潜在狄利克雷分配”。杂志
机器学习研究3:993–1022。

Boussalis, Constantine 和 Travis G. Coan。2016 年。“通过文本挖掘气候变化疑虑信号。”
全球环境变化36:89-100。

Braun, Virginia 和 Victoria Clarke。2006 年。“在心理学中使用主题分析。”定性
心理学研究3(2):77-101。

布赖曼、艾伦. 2006年。 整合定量和定性研究:如何完成? ”定性
研究6 (1): 97–113。

Carley, Kathleen。1993 年。“文本分析的编码选择:内容分析与地图分析的比较。”社会学方法论23:75。

————。 1994。“通过文本分析提取文化。”诗学22 (4): 291–312。

枝形吊灯、玛丽、Agnès Steuckardt、Raphaël Mathevet、Sascha Diwersy 和 Olivier Gimenez。2018 年。
“使用结构主题模型对法国报纸上有关狼重新定居的报道进行内容分析。” 《生物保护》 220（1月）:254-61。

Chang, Jonathan、Sean Gerrish、Chong Wang 和 David M. Blei。2009 年。“解读茶叶:人类如何解读主题模型。”神经信
息处理系统进展22:288–96。

Creswell, John W. 2003.研究设计定性定量和混合方法。千
加利福尼亚州奥克斯:SAGE。

约翰·W·克雷斯韦尔 (John W. Creswell) 和维基·L·普莱诺·克拉克 (Vicki L. Plano Clark)。 2007 年。设计和进行混合方法研究，
伦敦:Sage Publications。

D Amato、Dalia、Nils Droste、Ben Allen、Marianne Kettunen、Katja Lähtinen、Jaana E. Korhonen、Pekka
Leskinen、Brent D. Matthies 和 Anne Toppinen。 2017年。 “绿色、循环、生物经济:可持续发展概念的
比较分析。”清洁生产杂志168:716–34。

斯科特·迪尔韦斯特、苏珊·T·杜迈斯、乔治·W·弗纳斯、托马斯·K·兰道尔和理查德·哈什曼。
1990。“通过潜在语义分析进行索引。”美国信息科学学会杂志
41（6）:391-407。

Dehghani、Morteza、Kenji Sagae、Sonya Sachdeva 和 Jonathan Gratch。 2014年。“分析与‘归零地清真寺’建设相
关的保守派和自由派博客中的政治言论。”信息技术与政治杂志11 (1): 1–14。

丹尼（Denny），马修·J. 和亚瑟·斯皮林（Arthur Spirling）。2018 年。 “无监督学习的文本预处理:为何重要、何时会产生误导以及如何应
对。” 《政治分析》 26 (2): 168–89。

诺曼·K·登津和伊冯娜·林肯。 2007 年。收集和解释定性材料,第三版。
加利福尼亚州千橡市:SAGE。

DiMaggio, Paul。2015 年。“将计算文本分析应用于社会科学（反之亦然）。” Big
数据与社会2 (2):1–5。

迪马吉奥、保罗、曼尼什·纳格和大卫·布莱。 2013年。“利用主题建模和文化社会学视角之间的亲和力:美国政府艺术资助报纸
报道的应用。”诗学41 (6): 570–606。

Ding, Chris、Tao Li 和 Wei Peng。2008 年。“论非负矩阵分解与概率潜在语义索引之间的等价性。”计算统计与数据分析
52（8）:3913-27。

Entman,Robert M.1993 年。 “框架:澄清破碎的范式。” 《
通信43 (4): 51-58。

保罗·福西特、迈克尔·詹森、海达·兰桑-库珀和索尼娅·杜斯。 2018. “解释问题流的‘潮起潮
落’:澳大利亚煤层气（‘水力压裂’）未来的框架冲突。”公共政策杂志1-21。 https://doi.org/10.1017/
S0143814X18000132。

威廉·A·甘森 (Gamson) 和安德烈·莫迪利亚尼 (Andre Modigliani)。 1989。“关于核电的媒体话语和公众舆论:一种建构主
义方法。”美国社会学杂志95 (1): 1–37。

Gilardi、Fabrizio、Charles R. Shipan 和 Bruno Wueest。 2017年。 “政策扩散:问题定义http://fabriziogilardi.org/resources/papers/diffusion-policy-frames.pdf。 已进入阶段”。 2019 年 1 月 5 日。

Glaser, Barney G. 和 Anselm L. Strauss。1967 年。《扎根理论的发现:定性研究策略》。芝加哥:Aldine。

Goldstone, Andrew 和 Ted Underwood。2014 年。“文学研究的静悄悄的转变:一万三千名学者能告诉我们什么。” 《新文学史:理论与解释杂志》 45 (3): 359–84。

Greene, Derek 和 James P. Cross。2017 年。“探索欧洲议会的政治议程使用动态主题建模方法。”政治分析25 (1): 77–94。

格里默,贾斯汀。 2010。“政治文本的贝叶斯层次主题模型:衡量参议院新闻稿中表达的议程。”政治分析18 (1):1-35。

贾斯汀·格里默和布兰登·M·斯图尔特。 2013.“文本作为数据:自动的承诺和陷阱政治文本的内容分析方法。”政治分析21 (3): 267–97。

格里森科、达丽娅. 2016.“冰伏特加?揭示俄罗斯媒体对北极的看法。”活力研究与社会科学16:8-12。

格鲁伯特、艾米丽和马克·阿尔吉-休伊特。 2017.《恶棍还是英勇?美国小说和非小说叙事中对石油和煤炭的描述。”能源研究和社会科学31:100–10。

Hajer, Maarten A. 1995. 《环境话语的政治学》。纽约:牛津大学出版社。

Hofmann, Thomas。1999 年。“概率潜在语义分析”。《人工智能中的不确定性》,UAI 99,斯德哥尔摩。

Hsieh, Hsiu-Fang 和 Sarah E. Shannon。2005 年。“定性内容分析的三种方法。”《质性健康研究》 15（9）:1277-88。

伊索霍、卡罗利纳、范妮·莫伊拉宁和阿霍·托伊卡。 2019年。“欧洲能源联盟的大数据视角:从‘浮动标志’转变为脱碳的积极驱动力?”政治与治理7（1）:28。

雅各比、卡琳娜、沃特·范·阿特维尔特和卡斯帕·韦尔伯斯。 2016。“使用主题建模对大量新闻文本进行定量分析。”数字新闻4 (1):89–106。

贾加拉姆迪、贾加迪什、哈尔·道梅三世和拉加文德拉·乌杜帕。 2012年。 “将词汇先验纳入主题模型。”计算语言学协会欧洲分会第 13 届会议论文集,EACL 12, 204–13,法国阿维尼翁。

贾沃斯卡、西尔维娅和阿努帕姆·南达。 2016年。 “通过说好话来做好事:基于主题建模辅助的企业社会责任话语研究。”应用语言学39 (3): 373–99。

Keller, Reiner 2013。进行话语研究。伦敦:圣人。

肯尼迪,马克·托马斯。 2008 年。“计数:市场、媒体和现实。”美国社会学评论73 (2): 270–95。

Kim、Erin Hea-Jin、Yoo Kyung Jeong、Yuyoung Kim、Keun Young Kang 和 Min Song。 2016。“Twitter 和新闻中埃博拉病毒基于主题的内容和情绪分析。”信息科学杂志42 (6): 763–81。

克莱因、劳伦·F.、雅各布·爱森斯坦、艾里斯·孙和 JA·杰克科。 2015年。 “数字化档案馆藏的探索性主题分析”。人文数字学术30（补充1）:i130–41。

克里彭多夫、克劳斯. 2004。内容分析:方法论简介。加利福尼亚州千橡市:智者。

库卡茨、乌多和安妮·麦克沃托。 2014.定性文本分析:方法、实践和使用指南软件。伦敦:SAGE。

拉斯韦尔,哈罗德·德怀特,丹尼尔·勒纳和伊蒂尔·德·索拉·普尔。1952 年。《符号的比较研究》:一个介绍。加利福尼亚州斯坦福:斯坦福大学出版社。

Light, Ryan 和 Jeanine Cunningham。2016 年。“和平的预言:主题建模、文化机遇和诺贝尔和平奖,1902-2012 年* 。” 《动员:国际季刊》 21 (1): 43–64。

林格伦、西蒙. 2018 年。“机器中的幽灵:追踪‘数字’在网络受害的话语过程中的作用。”话语与交流12 (5): 517–34。 https://doi.org/10.1177/1750481318766936

Loewenstein, Jeffrey、William Ocasio 和 Candace Jones。2012 年。“词汇和词汇结构:一种连接类别、实践和机构的新方法。”管理学院年鉴
6（1）:41-86。

卢卡斯、克里斯托弗、理查德 A. 尼尔森、玛格丽特 E. 罗伯茨、布兰登 M. 斯图尔特、亚历克斯·斯托勒和达斯汀·廷利。 2015。“比较政治的计算机辅助文本分析。”政治分析
23（2）:254-77。

乔恩·D·麦考利夫和大卫·M·布莱。 2008.监督主题模型。见:神经信息处理系统的进展,第 121-128 页。 https://arxiv.org/abs/1003.0783。访问日期:2019 年 6 月 11 日。

Mills, Kathy A. 2018。“大数据对定性研究的威胁和潜力是什么？”
定性研究18 (6): 591–603。

Mimno,David 和 David M. Blei。2011 年。“主题模型的贝叶斯检查。”在EMNLP 11 自然语言处理经验方法会议论文集,爱丁堡,227-37。

Mimno,David 和 Andrew McCallum。2008 年。“基于狄利克雷多项回归的任意特征主题模型。”载于UAI 08 第二十四届人工智能不确定性会议论文集,赫尔辛基,411-18。

Mishler, Elliot G. 1995。“叙事分析模型:类型学”。叙事与生活史杂志
5（2）:87-123。

莫尔、约翰·W. 和佩特科·博格丹诺夫。 2013.“介绍主题模型:它们是什么以及它们为何重要。”诗学41 (6): 545–69。

莫尔 (John W. Mohr) 和 V. 杜肯 (V. Duquene)。 1997.“文化与实践的双重性:纽约的扶贫工作
城市,1888-1917。” 《理论与社会》 26 (2): 305-56。

芒克斯加德、拉斯穆斯和雅各布·德曼。 2016。“政治与犯罪的混合 加密市场上政治话语的盛行与衰落。”国际毒品政策杂志35：77-83。

村上隆、阿基拉、保罗·汤普森、苏珊·亨斯顿和多米尼克·瓦因。 2017.“‘这个语料库是关于什么的？’:使用主题建模探索专业语料库。”语料库12 (2):243–77。

Nelson, Laura K. 2017。“计算基础理论:方法论框架。”社会学方法与研究。 https://doi.org/10.1177/0049124117729700。

尼加姆、阿米特和威廉·奥卡西奥。 2010年。“事件关注、环境意义建构和制度逻辑的变化:公众关注对克林顿医疗改革倡议影响的归纳分析”。组织科学21 (4): 823–41。

Nowlin, Matthew C. 2016.“使用定量文本分析对问题定义进行建模。”政策研究杂志44 (3): 309–31。

帕沙欣,谢尔盖。 2016.“新闻媒体框架分析的主题建模。”在AINL 的诉讼中
FRUCT， 103-05。

Quinn、Kevin M.、Burt L. Monroe、Michael Colaresi、Michael H. Crespin 和 Dragomir R. Radev。2006 年。“一种随时间推移对立法演讲进行主题编码的自动化方法,应用于第 105 至 108 届美国参议院。”中西部政治学协会会议,纽约,1-61。

丹尼尔·拉米奇、大卫·霍尔、拉梅什·纳拉帕蒂和克里斯托弗·D·曼宁。 2009.“标记LDA:多标记语料库中信用归因的监督主题模型。”2009 年自然语言处理经验方法会议论文集1（8 月）:248-56。

里斯曼,凯瑟琳·科勒。 2005年。“叙事分析。”在叙事、记忆和日常生活中,编辑。
凯瑟琳·科勒·里斯曼。英国哈德斯菲尔德:哈德斯菲尔德大学,1-8。

Roberts, Margaret E.、Brandon M. Stewart、Dustin Tingley、Christopher Lucas、Jetson Leder-Luis、Shana Kushner Gadarian、Bethany Albertson 和 David G. Rand。2014 年。“开放式调查回复的结构主题模型。”《美国政治科学杂志》58 (4): 1064–82。

Rosen-Zvi、Michal、Chaitanya Chemudugunta、Thomas Griffiths、Padhraic Smyth 和 Mark Steyvers。2010 年。“从文本语料库中学习作者-主题模型。” ACM 信息系统学报
28（1）:1-38。

Roth, Merrill C. 和 Eric PS Baumer。2016 年。“这就是定义,笨蛋!互联网治理论坛辩论中的在线隐私框架。”信息政策杂志4 (2014): 144–72。

Schofield, Alexandra、Måns Magnusson 和 David Mimno。2017 年。“消除停用词:重新思考主题模型的停用词删除”。计算语言学协会欧洲分会第 15 届会议,第 2 卷。计算语言学协会,第 432–36 页。

Schofield, Alexandra、Måns Magnusson、Laure Thompson 和 David Mimno。2017 年。了解潜在狄利克雷分配的文本预处理。http ://www.cs.cornell.edu/~xanda/winlp2017.pdf。

Schön, Donald A. 和 Martin Rein。1994 年。《框架反思:解决棘手政策问题》争议。纽约:BasicBooks。

塞鲁西、雅尼尔、英格丽德·祖克曼和法比安·博内特。 2014.“主题作者归属模型。”计算语言学40 (2): 269–310。

Silverman, David。2016 年。《定性研究》。伦敦:SAGE。

Tang, Jian, Zhaoshi Meng, Xuan Long Nguyen, Qiaozhu Mei 和 Ming Zhang。2014 年。“通过后验收缩分析了解主题建模的限制因素。”第 31 届国际机器学习会议论文集,第 32 卷,北京,中国。

蒂切尔、斯特凡、迈克尔·迈耶、露丝·沃达克和伊娃·维特。 2000。文本和话语分析方法。寻找意义。伦敦:圣人。

Törnberg, Anton 和 Petter Törnberg。2016a。“结合 CDA 和主题建模:分析在线论坛上伊斯兰恐惧症和反女权主义之间的话语联系。”话语与社会27 (4): 401–22。

————。 2016b. “社交媒体话语中的穆斯林:结合主题建模和批判性话语分析。”话语、语境与媒体13:132–42。

Vaismoradi、Mojtaba、Hannele Turunen 和 Terese Bondas。2013 年。“内容分析和主题分析:开展定性描述性研究的启示。”护理与健康科学15（3）:398-405。

Van Atteveldt、Wouter、Kasper Welbers、Carina Jacobi 和 Rens Vliegenthart。 2014.LDA模型主题.....但是什么是“主题”?格拉斯哥（英国）:格拉斯哥社会科学大数据研讨会。

Wallach, Hanna M.、Iain Murray、Ruslan Salakhutdinov 和 David Mimno。2009 年。主题模型的评估方法。第 26 届ACM 机器学习国际会议论文集。

支持信息

其他支持信息可以在出版商网站上的本文在线版本中找到。