经验/政策

社会潮流
2019,第6(4)卷307–318

DOI:10.1177/2329496519846505
Journals.sagepub.com/home/scu

结构主题建模对于社会科学家:A简短案例研究社会运动研究文学,2005-2017

内森·C·林德斯特1

抽象的

社会学家经常在他们的研究中使用语言作为数据,使用的方法包括开放式调查、深度访谈和内容分析。不幸的是,随着与研究过程相关的成本和时间的增加,研究人员分析不断增加的数据的能力下降。主题建模是一种计算机辅助技术,可以帮助社会科学家应对这些数据挑战。尽管语言在社会学研究中发挥着核心作用,但迄今为止,该领域在很大程度上忽视了自动文本分析的前景,转而采用更熟悉和更传统的方法。本文概述了特别适合社会科学研究的主题建模框架。通过使用社会运动研究文献摘要的案例研究,为结构主题建模方法提供了从数据准备到数据分析的简短教程。这个例子展示了文本分析如何应用于社会学研究,并鼓励学者不仅将这些方法视为新颖的工具,而且将其视为可以与现有方法一起工作并增强现有方法的有用补充。

关键词
主题建模、方法论、集体行为和社会运动

社会学的必要条件是试图理解和解释群体行为的复杂性。为此,该领域在很大程度上依赖于通过口头或书面语言传达的信息。

开放式调查、深度访谈和内容分析等研究方法从根本上来说,是为了有序地审查一些基础语言数据而建立的一套实践。鉴于语言作为数据对于大部分研究的重要性,当今社会学家在创新语言及其意义分析方面普遍落后于其他学科,就像语言研究的数量一样可供研究的文本数据的数量正在迅速增长。一些学者已经认识到需要新的工具来更好地处理日益丰富的数据,因为学者们试图扩展和构建社会生活理论(例如,Evans 和 Aceves 2016)。其他学者甚至

1美国华盛顿州立大学,普尔曼

通讯作者:
Nathan C. Lindstedt,华盛顿州立大学社会学系,邮政信箱 644020,普尔曼,WA 99164,美国。

电子邮件: nathan.lindstedt@wsu.edu

宣称经验社会学由于忽视了这些方法论上的进步而濒临“危机”状态（例如,Savage 和 Burrows 2007）。一些作者建议将语言及其含义的研究与基于计算机的文本数据自动分析程序相结合（例如,Bail 2014；Lindstedt 2018）。为了响应这些呼吁,本文为希望扩展其方法论知识以应对即将到来的数据挑战的社会科学家提供了结构主题建模的介绍。1

潜在狄利克雷分配 (LDA) 与社会科学的结合

主题建模是一组归纳技术,用于发现文本数据中包含的隐藏主题。首先由 David M. Blei、Andrew Y. Ng 和 Michael I. 开发。

Jordan (2003) 混合成员主题建模,也称为 LDA,是一种用于识别文档集合中关键主题的无监督方法。 LDA 的底层是一个生成统计模型,该模型假设观察到的文档是由潜在主题的混合产生的。2这些未观察到的主题（其数量由研究人员事先定义）然后根据其概率分布生成关联词。因此,给定一组文档和多个主题,LDA 尝试识别未知主题的哪些组合可以生成这些文档。

结构主题模型 (STM) 将 LDA 框架扩展到社会科学研究的有前途的方向。STM 方法建立在标准主题模型的基础上,但它具有一些吸引社会科学家的特定优势：它可以以协变量的形式提供补充信息,这些信息可以揭示讨论主题的重要方面,或者有助于描述讨论主题的频率 (Roberts 等人,2013)。在前一种情况下,与主题内容有关的协变量可以回答有关讨论主题时使用的语言差异的研究问题（例如,政治意识形态、地理位置、等）,而在后一种情况下,与主题流行度有关的协变量可以回答有关讨论主题的频率差异（例如日期、作者等）的研究问题。以下案例研究介绍了能够结合模型协变量来解决其中一些研究问题的额外好处。

使用社交的案例研究

运动研究文学

对集体行为和社会运动的研究长期以来一直是社会学的核心兴趣。3但是,尽管社会运动研究代表了社会学分析的一个相对成熟的领域,但这并不意味着主体兴趣已经固定。

因此,该地区提供了一个合适的案例来探索这个表面上发达的子领域随着时间的推移发生了什么（如果有的话)重大变化。

多年来,研究项目的兴衰导致我们对社会运动的理解一次又一次地修正。这些更广泛的变化包括从集体行为的“经典”理论（例如,Smelser 1962)到更当代的领域理论（例如,Fligstein 和 McAdam 2012)的进展,其中社会运动是作为对系统压力的反应而出现的。 ）,其中有争议的行动取决于现任者和挑战者之间的竞争过程。随着这些更大的计划变化,研究人员的主题兴趣也发生了变化。其一,时事改变了学者创造知识的环境。社会时刻,例如#BlackLivesMatter 和#MeToo 运动的出现,可能会将研究人员的注意力引向某些研究方向,这最终可能会改变未来学术工作的方向（Moody 和 Light 2006）。

为了更好地了解子领域中产生的知识的当前状况及其潜在影响,重要的是要盘点过去产生的知识。也就是说,要了解你在哪里,

以及你可能要去哪里,知道你去过哪里是很好的。这个简短的案例研究对应用于当代社会运动研究的文本分析进行了简要介绍。它旨在为社会科学家提供适合研究这些和其他类型的主题趋势的定量技术的实用知识。4

该过程展示了 11 份顶级国家和地区社会学期刊摘要中确定的社会运动奖学金的 24 个关键主题的结果,并跟踪了 2005 年至 2017 年期间这些主题的流行度和影响力的变化。在此演示中,文档集由学术期刊的摘要组成,但它们可以轻松地从其他来源获取,例如报纸文章(例如,DiMaggio、Nag 和 Blei 2013)、开放式调查回复(例如,Tingley 2017)、演讲(例如 Light 和 Cunningham 2016)、公众评论(例如 Levy 和 Franklin 2014)或多个综合来源(例如 Farrell 2016a)来解决各种研究问题。5为了完成此分析,使用由 Margaret E. Roberts、Brandon M. Stewart 和 Dustin Tingley (2018) 开发的 stm R 包。

虽然最近存在使用 STM 的社会学学术研究的其他例子(例如,Almquist 和 Bagozzi 2017;Bohr 和 Dunlap 2017),但没有一个讨论完成此类研究所需做出的许多实际决策。 6这项工作通过揭秘来解决这一差距。该程序用于通过总结结构主题建模实践者迄今为止提出的主要建议来得出易于解释和分析上有用的结果。本文使用来自在线引文索引和开源软件的数据,介绍了 STM 框架内的数据准备、模型选择、估计、诊断评估和数据分析。

数据收集和方法

此分析的数据是使用 Web of Science 的社会科学引文索引 (SSCI) 收集的。其中包括以下期刊的摘要: American Journal of

表 1.社会学期刊摘要的描述性统计。

学术期刊	频率 %	
美国社会学杂志	51	7.22
美国社会学评论	64	9.07
动员	299	42.35
社会力量	54	7.65
社会问题	59	8.36
社会科学季刊	8	1.13
社会学论坛	55	7.79
社会学探究	26	3.68
社会学观点	三十八	5.38
社会学季刊	三十九	5.52
社会学谱系	13	1.84
全部的	706	100

社会学、美国社会学评论、动员、社会力量、社会问题、社会科学季刊、社会学论坛、社会学探究、社会学视角、社会学季刊和社会学谱系(见表1)。大多数期刊被选择纳入,因为它们代表了国家和地区层面上最受大众关注的社会学期刊。还包括专业期刊《动员》因为它与研究的子领域相关。其他期刊被省略,因为它们要么是该领域相对较新的补充,例如Social Currents,要么不包含在 SSCI 中,例如 Research in Social Movements, Conflicts and Change。

除了动员之外,只有包含关键词搜索词“社会运动”或“社会运动”的文章才会被考虑纳入。这种选择大大缩小了所包含文章的范围,同样受限的关键词搜索可能无法为研究人员提供最佳结果,因为研究人员有特定的研究领域。在研究过程的早期阶段,研究人员应该投入大量时间来考虑哪些文件应该和不应该包含在他们的数据集中。在这种情况下,使用“集体行为”、“集体行动”、“激进主义”和“抗议”等密切相关的术语的试验产生了存在大量边界问题的结果。换句话说,

这些连续术语的组合返回的搜索结果包含大量来自定义的兴趣领域之外的文章,这是一个令人烦恼的偏见的潜在来源。

Jeremiah Bohr 和 Riley Dunlap (2017) 指出,由于语言的模糊性,使用 SSCI 等在线引文数据库进行数据收集通常需要在获取明确边界的数据集和捕获每篇相关文章之间进行平衡。最近的工作建议使用计算机辅助技术来改进关键词选择,并警告“研究人员通常以临时方式选择关键词,这种方式远非最佳,而且通常存在偏见”(King、Lam 和 Roberts 2017:1)。不幸的是,对这种从非结构化文本中发现关键字的方法的深入总结超出了本文的范围。然而,一般来说,将文档集限制在明确定义的感兴趣区域的关键字搜索会提高主题建模的推理性能,并将有助于限制误报的数量,尽管忽略相关文档的可能性仍然存在。该数据收集程序总共生成了 706 份摘要。

除了摘要中的文本数据外,SSCI 还包含每篇文章的附加数据,包括出版年份、总引用次数和定期使用次数。

然后,这些所谓的元数据可以用作 STM 框架内的协变量。为了演示如何将额外的主题流行度协变量集成到主题模型中,我们通过将总引用次数除以出版年数,从包含的元数据中构建了一个使用每年平均引用次数的影响力度量。

除了能够解释模型协变量的好处之外,stm R 包还允许估计主题之间的相关性。此属性是可取的,因为如前所述,在主题建模中,文档由主题的混合组成。

因此,研究人员可以使用结构主题建模来观察哪些主题在文档级别上彼此密切相关。

然后使用 Fruchterman-Reingold 算法绘制这些相关性的图表在 igraph R 包中找到(Csardi 和 Nepusz 2006)。

下一节将介绍 24 个主题解决方案的结果及其计算出的相关性。不过,在介绍这些结果之前,有必要讨论一下主题建模的缺陷,尤其是模型选择和模型诊断的缺陷。

正如学者贾斯汀·格里默 (Justin Grimmer) 和布兰登·M. Stewart (2013) 警告称,虽然自动化方法有望大幅降低传统文本数据分析所需的成本和时间,但它们并不能完美地替代人类的解释和特定领域的专业知识。7

简而言之,自动化方法不是替代而是对多阶段研究过程中研究人员能力的补充。8,9学者们指出,主题建模技术的一个缺点是,没有直接的方法来选择大量主题,这些主题可以产生易于解释和分析有用的结果(例如,Farrell 2016b;Roberts 等人 2014)。问题的关键在于,在 LDA 中,研究人员必须提前选择主题的数量。事实上,由于自动文本分析基于错误的语言模型,因此对于选择主题数量没有一个正确的答案(参见 Grimmer 和 Stewart 2013)。用于从文本数据中提取主题的概率“词袋”语言模型的假设是,可以从文档中术语的共现推断出主题,而单词排序不会影响分析。此外,虽然存在多种评估模型拟合度的方法,但严格遵守这些诊断可能会导致模棱两可的结果。部分原因是,当使用主题模型进行预测时,那些最能预测样本外文档的模型通常与人类判断不一致。因此,基于优化保留可能性度量的模型选择可能会导致不太有洞察力的结果 这种现象被称为预测-可解释性权衡(Chang 等人,2009 年;Wesslen,2018 年)。除了这些警告之外,还有其他有用的指导方针

表 2.结构主题模型诊断表。

主题数（K）	独家性	语义连贯性	保留可能性	剩余的
5	9.310	−71.253	−5.669	1.282
10	9.537	−80.076	−5.645	1.141
15	9.433	−79.853	−5.640	1.068
20	9.559	−83.056	−5.607	1.017
25	9.632	−87.440	−5.634	0.975
三十	9.686	−88.139	−5.621	0.945
35	9.705	−90.539	−5.613	0.920
40	9.729	−90.367	−5.583	0.905
45	9.775	−91.677	−5.587	0.901
50	9.766	−93.159	−5.593	0.891

和诊断方法如果遵循,可以指导研究人员为他们的分析找到良好的起点。

值得庆幸的是,stm R 包的作者就如何处理选择主题数量的棘手问题提供了一些指导,尽管他们承认没有既定的方法可以在不同的文档集中获得一致的结果。初步的、更主观的建议与文档长度、文档焦点和 LDA 模型的性能之间的关系有关。对于较短的、集中的语料库（即,大小从几百到几千个文档的语料库）,最初选择 5 到 50 个主题之间是最好的,而对于较大的、没有重点的语料库（即,范围从数万到数千个文档的语料库）,最好选择 5 到 50 个主题。数十万个或更大的文档）,之前的研究发现 60 到 100 个主题是最好的（Roberts 等人,2018）。

次要的、不太主观的建议涉及检查诊断表以及语义连贯性和排他性计算的图。语义连贯性是一组主题词在同一文档中同时出现的概率的度量。排他性是衡量一个词主要落在单个主题的最高排名中的概率的指标。主题数量的模型选择是沿着语义连贯性排他性“前沿”进行的,其中没有任何模型受任一指标支配（Roberts et al. 2014）。有关诊断输出,请参阅表 2;有关语义连贯性排他性图,请参阅图 1

初步建议的主题编号范围内的模型。通常,研究人员会希望缩小模型选择范围,因为模型在语义连贯性和排他性维度上都具有理想的属性。也就是说,他们将首先选择候选模型,其解决方案使它们更接近绘图的右上象限（Roberts、Stewart 和 Tingley 即将推出）。在此示例中,包含大约 20 个主题的模型似乎在每个指标之间提供了最佳平衡。相反,具有大约 15 个主题的模型在语义一致性方面得分较高,但在排他性方面得分较低,就其可解释性而言可能不是合适的解决方案。

鉴于模型选择的困难以及预测模型和可解释模型之间的权衡,模型选择的最终责任在于研究人员及其明智的判断。

因此,研究人员需要“验证、验证、再验证”他们的研究结果（Grimmer 和 Stewart 2013:5）。这个过程可以通过多种方式完成,但 stm R 包中最有用的验证方法是其内置函数,该函数提供了特定主题最具代表性的文档列表。

另一种方法是通过评估构成每个主题的词簇来评估主题质量（Roberts et al. 2014）。提供了附加的内置功能,可以显示每个主题中密切相关的单词,包括在每个主题或单词中找到的概率最高的单词

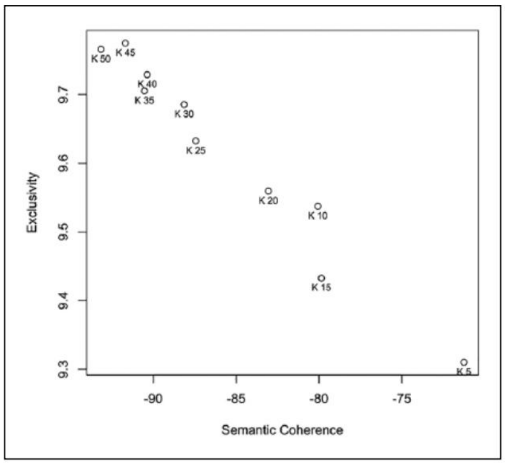

图 1.结构主题模型语义一致性-排他性图。

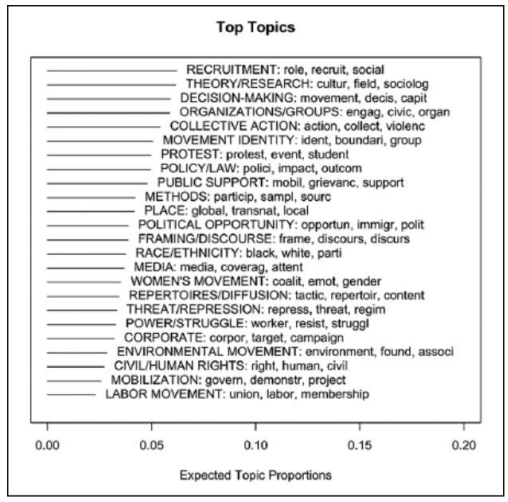

图 2. 24 个主题解决方案的标签。

根据每个主题中的频率和排他性进行加权（Roberts 等人即将出版）。
基于候选模型的语义连贯性排他性诊断建议的大约 20 个主题的初始数量,这些组合的内置函数用于定性评估 20 到 20 个主题中每个可能模型的结果。 25个主题范围得出最终主题数。

关于重新思考反色情运动联盟的文章被退回。当被要求提供有关招募主题的更多文件时,White（2007)关于爱尔兰共和主义招募的方法论文章,Munson（2010)对社交网络配置变化如何解释大学保守派动员的研究,以及 Crossley（2015)的研究在线女权论坛扩大招募基地回归。显然,这些都是与招募、溢出和联盟建设问题有关的文件。

结果

图 2 显示了 24 个主题解决方案的结果。在量表的顶端,占文档总数的约 6%,最常见的主题是“招聘”。在量表的底端,占文档总数的约 2%,最不流行的主题是“劳工运动”。对于社会运动学者来说,组织、运动身份和政治机会等主题在这个排名中占据突出地位可能并不令人意外,因为这些主题反映了该子领域的一些主要理论观点。但是,stm R 包提供的工具包允许研究人员更进一步,通过识别每个主题领域内的原型文档（即,那些单词比例最高的文档,专门用于特定主题）。例如,当查询原型文档的招聘主题时,Whittiers（2014）

但观察主题比例及其代表性文献并不能告诉我们这些主题多年来的趋势如何。

为了更全面地了解子领域产生的知识随时间的变化情况,可以使用出版年份作为主题流行度协变量,然后绘制主题比例变化的估计值。为了演示，“妇女运动”和“种族/

之所以选择“种族”主题,是因为它们突显了观察期内主题比例的急剧逆转（见图 3）。在 2005 年至 2017 年期间,妇女运动主题占报告主题比例的比例从约 4.5% 下降到 2.4%。相比之下，种族/民族主题占报告主题比例的比例从约 2.2% 上升到 5.1%。

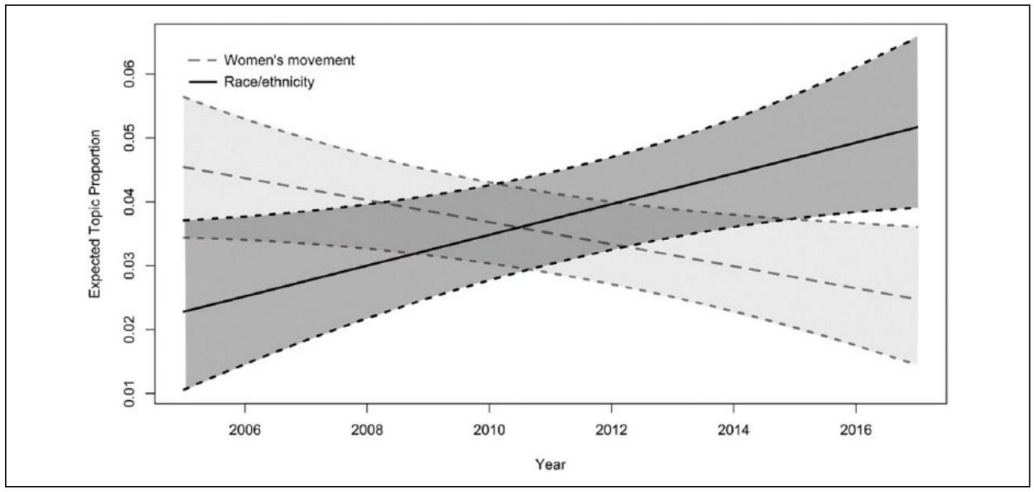

图 3.按年份划分的“妇女运动”和“种族/族裔”主题流行度（95% 置信区间）。

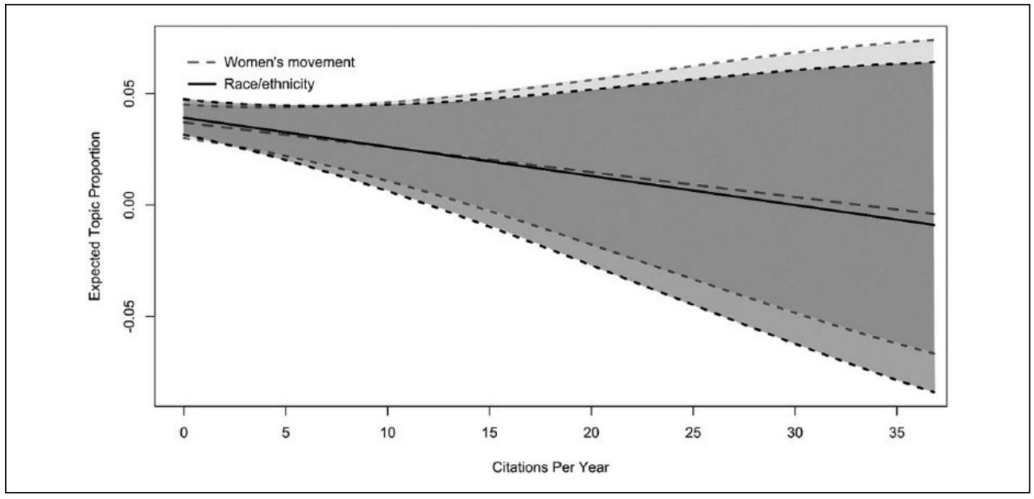

图 4. “妇女运动”和“种族/民族”主题的流行程度,按每年的引用量计算（95% 置信区间）。

为了结束对某些问题的讨论

通过使用协变量可以获得额外的数据分析选项,我们将重新审视每年平均引用的变量。尽管多年来妇女运动和种族/民族主题的流行程度发生了明显变化,但这并不能说明这些主题在不同影响水平上的流行程度。为此,每年平均引用的度量也可以作为模型协变量引入。

从最低的每年零引用到最高的每年 36.82 次引用 平均值为每年 2.04 次引用,中值为每年 1.17 次引用,标准差为每年 3.08 次引用。

请注意,尽管它们的频率随着时间的推移而发生逆转,但每个主题都有相似的影响（见图 4）。也就是说,在每年平均引用平均值的两个标准偏差范围内，“妇女运动”和“种族/民族”主题的频率保持在可比较的水平。相比之下,图5

对于整个文档集,取值范围

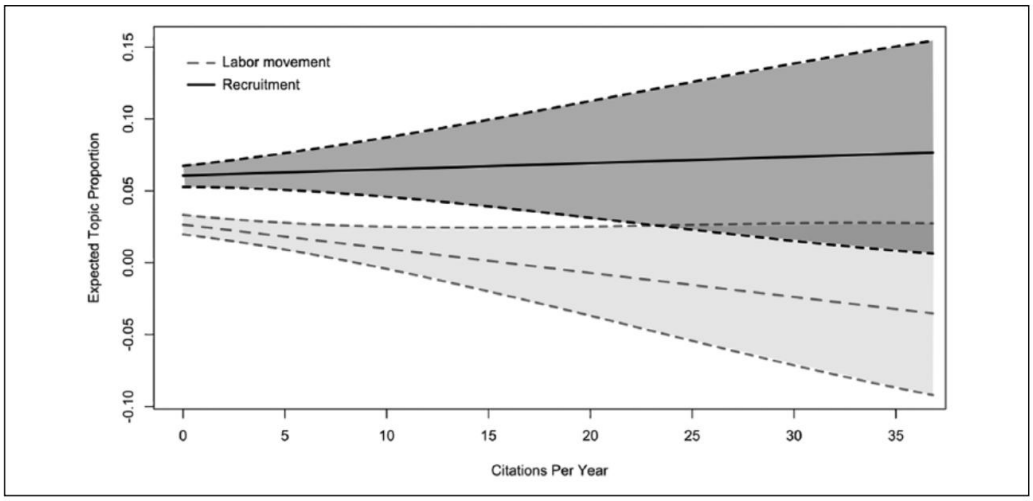

图 5. “劳工运动”和“招聘”的主题流行度（按每年引用次数计算）（置信区间为 95%）。

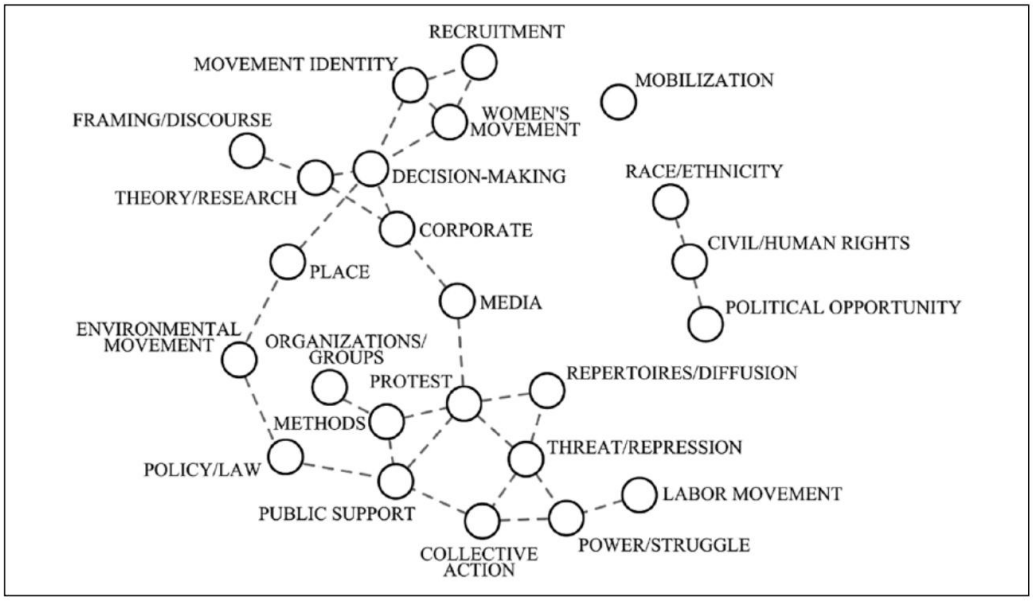

图 6.主题相关图（联系强度 > .01）。

举了一个例子，其中影响力对主题流行度的影响有较大程度的差异。在均值两个标准差以内的文章中，“招聘”主题的文章所占比例明显远高于“工人运动”主题的文章。

配对感兴趣的主题时，“妇女运动”和“种族/族裔”以及“劳工运动”和“招聘”与大于 0.01 水平的关系强度均不显着相关。高于此阈值的正相关表明配对主题可能会在同一文档中进行讨论。因此，这些对似乎并不代表密切相关的主题。但精明

图 6 展示了文档级别的主题相关性图。请注意

社会运动学者可能会在图表中发现该子领域的其他重要特征。这些由诸如“妇女运动”、“招募”和“运动身份”等主题之间形成的二元和三元关系所表征；“妇女运动”、“决策”和“运动身份”；“种族/民族”和“公民/人权”;以及“劳工运动”和“权力/斗争”。对相关主题之间这些联系的一种实质性解释是,这种联系的存在代表了邻近的研究领域,这些领域相互影响的倾向更大,而缺乏这种联系代表了远程研究领域,这些领域的相互影响程度较低。相互影响的倾向（Fligstein 和 McAdam 2012）。

到目前为止,对于处理各种文本数据源的研究人员来说,这些工具的一些潜在用途应该是显而易见的。STM 主题建模扩展不仅允许研究人员完成标准 LDA 框架内已经可以完成的工作,即识别潜在主题,还允许他们解决与社会科学家特别相关的其他问题,即如何使用其他协变量构建文本数据。

考虑到这些好处,结果说明了这种技术对社会科学研究的实用性。虽然不是详尽的资源,但本文展示了结构主题建模在文本数据分析中的一些可能用途。其他工作可能会考虑主题内容的协变量（例如意识形态观点或地理区域)如何影响用于讨论这些主题的语言。虽然结构主题建模可以实现的功能存在重大限制,但它也为那些愿意使用该方法的研究人员开辟了额外的研究方向。

结论

社会学家经常使用语言作为数据。然而,他们在利用日益增多的此类数据方面已经落后于其他学科。本文提出了一种对于希望扩展其方法库以更好地应对这些即将到来的数据挑战的研究人员来说,这是一个指导性教程。通过使用结构主题建模,学者可以总结内容并检查大量文档中主题的结构,而这些方式是传统分析所无法做到的。

事实上,STM 框架可以与其他方法（包括开放式调查、深度访谈和内容分析)结合使用,以实现以前因成本过高或耗时过长而无法进行的分析,同时使学者能够获得对重要研究问题的影响力。也就是说,它可以用于多阶段研究过程,以加快识别大量文本数据集中的主要主题,从而促进后续更密切的定性分析。在包含的案例研究中,STM 框架的额外好处（例如模型协变量的使用)展示了该技术在未来学术中的一些预期用途。总之,本文向研究人员介绍了结构主题建模,并鼓励他们使用文本分析工具来帮助他们超越更熟悉的方法论领域进行创新。

致谢

作者要感谢 Erik W. Johnson、匿名审稿人和编辑对本文早期草稿提供的有益评论和建议。

利益冲突声明

作者声明与本文的研究、作者身份和/或出版不存在潜在的利益冲突。

资金

作者没有获得本文的研究、作者身份和/或出版的任何经济支持。

笔记

1. 本文描述的用于完成自动文本分析的 R 脚本和数据以补充文件的形式在线提供,以便读者可以跟进并复制详细的研究结果。

2. “生成”模型是一种统计模型,它根据观察到的输出生成未观察到的输入。这与“判别”模型相反,后者根据观察到的输入推断未观察到的输出。

潜在狄利克雷分配 (LDA) 是生成模型的一个例子,而逻辑回归是判别模型的一个例子。有关所用生成过程的更多详细信息,Blei (2012) 使用车牌符号对 LDA 的描述为理解工作中的一般原理提供了工作基础。

3. Oberschall（1973)在他的《社会冲突和社会运动》一书的开头指出，“写出社会冲突和社会运动的宏观理论的前身,就等于写出社会学思想本身的历史,从马克思和马克思开始。托克维尔”（第 1 页）。

4. Ignatow 和 Mihalcea (2017, 2018) 发布了一系列书籍,这些书籍也旨在为社会科学家提供本文中描述的文本分析方法的介绍。

5. DiMaggio、Nag 和 Blei (2013) 使用报纸文章来确定媒体报道 1986 年至 1997 年间支持艺术的政府拨款的主要主题。Tingley (2017) 使用开放式调查回答来探讨在国际关系中实力下降或实力上升的情况下个人所表达的逻辑。 Light 和 Cunningham（2016)利用诺贝尔奖获得者在和平领域的演讲来揭示主题,然后利用结果来推动他们的定性分析。 Levy 和 Franklin（2014)利用公众对拟议的卡车运输法规的评论来了解个人和组织利益相关者如何就卡车司机工作时间的电子监控进行政策辩论。 Farrell（2016a)使用了多种综合来源,包括书面和口头文本,表明企业对气候变化反运动组织的资助影响了他们两极分化话语的主题内容。

6. 其他学术成果,例如 Mohr 和 Bogdanov (2013),提供了主题建模的非技术性介绍,但没有提供本文所关注的结构化主题建模变体。

7. 为了缓解这一问题,研究表明,集中于相对较少主题的文档数量越多，这些重点文档越长,LDA 的性能就越好（Tang et al. 2014）。

涵盖一系列主题的单卷（例如一本书)或特别短的文档（例如推文)提供的数据集对于主题建模来说并不理想。在这些情况下,研究人员使用了替代技术,例如将书籍分成多个连贯的文档（例如,Mimno 和 McCallum 2007)或根据共享属性聚合推文（例如,Hong 和 Davison 2010）。

8. Nelson（2017)提出了一个计算扎根理论框架,该框架可应用于文本数据,包括开放式调查、深度访谈和内容分析。在第一阶段,基于计算机的文本分析技术可帮助研究人员识别以前未考虑的潜在模式,同时扎根于数据。在第二阶段,研究人员对文本子集进行研究,以确认已识别的潜在模式的可信度,解释结果,并调整计算模型以增强其解释能力。在第三阶段,研究人员测试早期阶段的结果是否具有普遍性,并最终检查计算扎根理论过程的可靠性。

9. 例如,Valdez、Pickett 和 Goodson（2018)认为主题建模既可以用于编码目的,也可以作为确认已生成代码的手段。在第一种情况下,研究人员使用主题建模来发现主题并研究它们的结构。在第二种情况下,研究人员在对定性数据进行编码后采用主题建模来作为所确定代码的额外可靠性检查。

补充材料

本文的补充材料可以在线获取。

ORCID编号

内森·C·林德斯特 https://orcid.org/0000-0002-5263-5687

参考

扎克·W·阿尔姆奎斯特 (Almquist) 和本杰明·E·巴戈齐 (Benjamin E. Bagozzi)。2017 年，“利用激进环保主义文本揭示网络结构和网络特征”。社会学方法与研究。

11 月 16 日以电子版形式出版。doi:
10.1177/0049124117729696。

Bail, Christopher A. 2014.“文化环境:用大数据衡量文化。”理论与社会43(3–4):465–82。

Blei, David M. 2012.“概率主题模型。”
ACM通讯55(4):77-84。

Blei,David M.、Andrew Y. Ng 和 Michael I.
乔丹。2003 年。“潜在狄利克雷分配。”
机器学习研究杂志3:993-
1022.

玻尔、耶利米和莱利·E·邓拉普。 2017年。“环境社会学的关键主题,
1990-2014年:计算文本分析的结果。”
环境社会学4(2):181-95。

Chang, Jonathan、Sean Gerrish、Chong Wang、Jordan L. Boyd-Graber 和 David M. Blei。2009 年。
“阅读茶叶:人类如何解释主题模型。”页码。 《神经信息处理系统进展》第 288-96 页,由 Y. Bengio、D. Schuurmans、JD Lafferty、CKI Williams 和 A. Culotta 编辑。第 22 版。纽约州雷德胡克:Curran Associates。

克罗斯利,艾莉森·达尔。 2015。“Facebook 女权主义:社交媒体、博客和当代美国女权主义的新技术。”动员:国际季刊20(2):253–68。

卡萨迪、加博尔和塔马斯·内普斯。 2006.“用于复杂网络研究的
Igraph 软件包。”期刊间复杂系统
1695:1-9。

迪马吉奥、保罗、曼尼什·纳格和大卫·布莱。
2013 年。“利用主题建模与文化社会学视角之间的亲和力:应用于美国政府艺术资助的报纸报道。”诗学
41(6):570-606。

Evans, James A. 和 Pedro Aceves。2016 年。“机器翻译:挖掘文本以了解社会理论。”
社会学年鉴42(1):21-50。

Farrell, Justin。2016a。“企业资助和气候变化的意识形态两极分化。” 《美国国家科学院院刊》
113(1):92–97。

Farrell, Justin。2016b。“企业资金和气候变化方面的意识形态两极分化。”
补充信息。检索于 2018 年 9 月 4 日(http://www.pnas.org/content/pnas/suppl
/2015/11/18/1509433112.DC补充/pnas
.1509433112.sapp.pdf)。

弗利格斯坦、尼尔和道格·麦克亚当。 2012。场论。纽约:牛津大学出版社。

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。

洪良杰和布莱恩·戴维森。 2010.“Twitter 中主题建模的实证研究”。页码。首届社交媒体分析研讨会论文集,SOMA 10,第80–88 页。纽约:ACM。

Ignatow,Gabe 和 Rada Mihalcea。 2017.文本挖掘:社会科学指南。
加利福尼亚州洛杉矶:Sage Publications。

伊格纳托、加布和拉达·米哈尔恰。 2018。文本挖掘简介:研究设计、数据收集和分析。加利福尼亚州千橡市:Sage Publications。

King, Gary、Patrick Lam 和 Margaret E. Roberts。2017 年。
“从非结构化文本中进行计算机辅助关键字和文档集发现。”美国政治学杂志61(4):971–88。

利维、凯伦·EC 和迈克尔·富兰克林。 2014年。
“推动监管:使用主题模型来研究美国的政治争论
卡车运输业。”社会科学计算机评论32(2):182-94。

赖恩·坎宁安和珍妮·坎宁安。 2016年。
“和平的预言:主题建模、文化机遇和诺贝尔和平奖,1902 年
2012。”动员:国际季刊
21(1):43–64。

Lindstedt,Nathan。2018 年。“转变框架:从对话和关系视角看集体行动框架。” 《社会学指南针》 12(1):1-12。

米诺、大卫和安德鲁·麦卡勒姆。 2007年。
“组织 OCA:从数字图书馆学习多面主题。”
第 7 届 ACM 论文集第 376–385 页/
IEEE-CS 数字图书馆联合会议, JCDL 07。加拿大不列颠哥伦比亚省温哥华:ACM。

Mohr, John W. 和 Petko Bogdanov。2013 年。
“简介 主题模型:它们是什么以及为什么重要。”诗学
41(6):545–69。

穆迪,詹姆斯和瑞安·莱特。2006 年。“从上方看:不断发展的社会学景观。”
美国社会学家37(2):67-86。

蒙森、齐亚德. 2010年。“校园动员:保守派运动和当今的大学生”。社会学论坛25(4):769–86。

Nelson, Laura K. 2017. “计算扎根理论:一种方法论框架。”社会学方法与研究。11月 21 日以电子版形式出版。doi:10.1177/0049124117729703。

奥伯绍尔、安东尼. 1973.社会冲突和社会运动。新泽西州恩格尔伍德悬崖:Prentice-Hall。

Roberts, Margaret E.、Brandon M. Stewart 和 Dustin Tingley。2018 年。“包‘stm’。”R

软件包版本 1.3.3。检索于 2018 年 9 月 4 日(https://cran.r-project.org/web/packages/stm/stm.pdf）。

玛格丽特·E·罗伯茨、布兰登·M·斯图尔特和达斯汀·廷利。即将推出。“stm：用于结构主题模型的 R 包。”《统计软件杂志》1：12。2018 年 9 月 4 日检索 (https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf)。

Roberts, Margaret E.、Brandon M. Stewart、Dustin Tingley 和 Edoardo M. Airoldi。2013 年。“结构主题模型和应用社会科学。”发表于神经信息处理系统进展研讨会主题模型：计算、应用和评估，12 月 24 日，马萨诸塞州剑桥。

罗伯茨、玛格丽特·E.、布兰登·M·斯图尔特、达斯汀·廷利、克里斯托弗·卢卡斯、杰特森·莱德-路易斯、莎娜·库什纳·加达里安、贝瑟尼·艾伯森和大卫·G·兰德。 2014。“开放式调查响应的结构主题模型。”

美国政治科学杂志58(4)：1064–82。

萨维奇、迈克和罗杰·布罗斯。 2007年。“经验社会学即将到来的危机。”

社会学41（5）：885-99。

Smelser, Neil. 1962.集体行为理论。纽约：自由出版社。

唐建、孟肇世、阮宣龙、梅巧竹、张明。 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 ICML 14第 190-8 页，由 EP Xing 和 T. Jebara 编辑。中国北京：机器学习研究杂志。

廷利，达斯汀。 2017年。“心灵力量的崛起。”国际组织21：S165–S88。

Valdez, Danny、Andrew C. Pickett 和 Patricia Goodson。2018 年。“主题建模：社会科学的潜在语义分析。”

社会科学季刊99(5)：1665–79。

韦斯伦、瑞安. 2018.“社会科学的计算机辅助文本分析：主题模型及其他。”北卡罗来纳州夏洛特：北卡罗来纳大学夏洛特分校。

White, Robert。2007 年。“‘我不太确定我上次告诉了你什么’：爱尔兰共和运动高风险活动分子叙述的方法论注释。”《动员：国际季刊》12(3)：287–305。

Whittier, Nancy。2014 年。“重新思考联盟：反色情女权主义者、保守派以及合作对抗运动之间的关系。”《社会问题》61(2)：175–93。