·方法学探讨·

大众媒体视野下的“健康中国”
——基于2016—2017年部分媒体报道的文本分析

姜浩然[1]* 周 萍[2] 杨肖光[2]
1. 辽宁社会科学院社会学所 辽宁沈阳 110031
2. 复旦大学公共卫生学院 国家卫生健康委卫生技术评估重点实验室 上海 200032

【摘 要】目的：对2016—2017年部分综合性媒体关于健康中国的报道文本进行量化分析，从一个新的视角了解健康中国的政策导向、实施进展和重点领域。方法：采集部分综合性媒体中关于健康中国的公开报道，通过词频分析、主题模型聚类等方法，分析健康中国报道的重点领域及分布规律。结果：健康中国的报道呈现多样化的特点，其报道在时间上的分布规律与国家重要事件（如“两会”、中共十九大）相契合，在内容上与国家政策重点相一致，较为明确的反映出健康中国实施过程中的重点领域及其关注程度。主题模型分析显示，健康中国的报道主要分为医疗卫生、健康生活、健康产业等类别，对于健康中国的政策能够形成一定呼应。结论：健康中国的政策框架与媒体报道的呈现形式具有一致性，提示新闻媒体报道可能成为公共政策分析的潜在证据来源，同时也显示出文本挖掘方法在完成相关政策分析任务上的潜力。

【关键词】健康中国；媒体报道；文本挖掘；主题模型
中图分类号：R197 文献标识码：A doi：10.3969/j.issn.1674-2982.2018.09.014

The presence of “Healthy China” in mass media: A text mining of news reports in 2016—2017

JIANG Hao-ran[1], *ZHOU Ping*[2], *YANG Xiao-guang*[2]
1. *Institute of Sociology, Liaoning Academy of Social Science, Shenyang Liaoning110031, Chia*
2. *School of Public Health, Fudan University, Key Lab of Health Technology Assessment, National Health Planning Commission, Shanghai 200032, China*

【Abstract】 Objective: Analyze the selected news report on Healthy China in the media during 2016—2017, and try to understand the policy focus and progress from a new perspective. Methods: Collect 6999 pieces of news reports on Healthy China and analyze key areas and their distribution patterns through word frequency analysis and topic model clustering. Results: The news report of Healthy China presents highly diversified characteristics. The time distribution of report is in line with national important events (such as the “two sessions” and the 19th CPC National Congress), and it is consistent with the major framework of the national policy on the content. The topicmodel analysis shows that the reports could be divided into categories like health care, life style, health industry, which is also matching the policy focus. Conclusion: The policy framework of Healthy China is consistent with the presentation form of media reports, suggesting that news media may be a potential source of evidence for public policy analysis, and it also shows the potential of text mining methods in accomplishing related policy analysis tasks.

【Key words】 Healthy China; News reports; Text mining; Topic models

“健康中国”是当前中国重点推进的国家级战略。健康中国的内涵极为丰富，涉及从微观层面的健康生活方式、健康服务，到宏观层面的健康保障、健康环境、健康产业以及健康治理体系等各个方

* 基金项目：复旦大学新进职工科研启动项目
作者简介：姜浩然，女（1979年—），硕士，助理研究员，主要研究方向为社会政策研究。E-mail：jianghaoran616@163.com
通讯作者：杨肖光。E-mail：yangxg@fudan.edu.cn

面[1]，带动了全国范围内围绕健康议题而开展的各项政治、经济和社会活动。健康中国建设的进程也为各级各类新闻媒体所持续关注。自十八届五中全会提出“推进健康中国建设”理念，到 2016 年 8 月全国卫生与健康大会召开及中央政治局审议通过《“健康中国 2030”规划》，再到十九大正式提出“实施健康中国战略”，其间累积的大量媒体报道信息，为全景式的认识这一国家重大政策的实施进程提供了潜在的可能性。

新闻媒体是重要的信息载体、意见表达渠道和公共沟通平台。媒体在及时、准确记录事件的发生的同时，也反映了社会对于特定问题的态度。同时，媒体也承载着舆论导向的功能，在推行政策的过程中，政府也会有意识的利用媒体进行宣传和倡导。[2]在互联网与大数据时代，随着文本数据挖掘技术的突破，媒体报道的量化分析已引起研究者的重视，并广泛应用于各个领域，如金融、农业、环境等。[3]然而，在卫生与健康领域，媒体报道相关研究分析多停留在新闻传播学的角度开展的媒体报道内容分析。[4]为数不多的基于量化的舆情分析[5]则以报道频次、时间分布、关键词词频等描述方法为主，对新闻文本信息挖掘的深度有限，也一定程度上影响了分析效果。

本文将利用文本挖掘（text-mining）的手段，对 2016—2017 年部分综合性新闻媒体关于健康中国的报道进行挖掘与分析，探索媒体报道健康中国的内容、领域、总体性特点，进而从一个新的视角了解健康中国的政策导向、实施进展和重点领域，为政府有关部门更好的推进健康中国战略提供参考。

1 资料与方法

1.1 数据来源及预处理

1.1.1 数据采集

利用自编 R 语言程序，从国内有影响力的门户网站、重点报刊数字版等渠道采集部分综合性新闻报道文本。具体来源是：从新浪、搜狐、凤凰、腾讯、网易、人民网、新华网、中国新闻网等门户网站的新闻频道采集时政新闻、社会新闻、财经新闻以及新闻评论栏目的全部新闻；从财新网、新京报网、澎湃新闻网三个重要的综合性媒体网站采集各子栏目新闻；同时采集了人民日报、光明日报、中国青年报三家重点报刊数字版的全部新闻，并去除国际新闻、娱乐新闻、体育新闻、广告等栏目。新闻采集时间范围为 2016 年 1 月 1 日—2017 年 12 月 31 日。共获取新闻文本总数 5 343 966 篇。需指出的是，本文数据来源全部为综合性新闻媒体，并未纳入《健康报》、《健康时报》等专业健康媒体。部分由于网站限制采集原因，同时也考虑到专业健康媒体可能会对数据整体分布造成影响。

1.1.2 数据筛选、过滤与分词

采集到的原始新闻文本保留“标题”、“发布时间”、“来源”和“正文”四个字段作为分析的基础数据。首先以词典规则的方法[6]筛选出与健康中国相关的媒体报道①，具体方法是：

（1）筛选出标题和正文中出现“健康中国、全民健康、健康融入所有政策”中任意一个关键词的报道文本，作为初筛结果，共计 13 630 篇新闻。

（2）根据“健康中国、全民健康、健康融入所有政策”三个关键词在新闻报道中的出现位置，对初筛新闻进行打分。经人工测试后确定的赋值规则为：如果任意关键词出现在标题位置则权重为 6，出现在文本首段权重为 3、非首段的首句权重 2、非首段非首句权重 0.6，按出现次数加权后加总得出主题得分。

（3）由于部分报道可能间或出现上述关键词，但其报道本身与健康领域无关（如财经新闻），故本文拟定了若干健康领域的关键词②，这些领域词表中的任一词在正文中出现一次计 0.05 分，加总后作为领域得分。主题得分与领域得分相加得到文档总分。经作者人工判断并讨论后，确定得分 2.5 分以上的入选，共计 10 308 篇。

（4）由于热点新闻可能会被不同的网站多次转发，故本文利用文本相似度计算的方法[7]，对新闻正文进行了去重处理，剩余新闻 6 999 篇。作为文本分析的数据源。

1.1.3 文本分词及预处理

对于 6 999 篇报道，在保留标题、发布时间、来源字段不变的前提下，利用 R 语言 jiebaR 工具包[8]将新闻正文进行分词处理。分词工具中加入自编词库，避免一些专有词汇（如“健康融入所有政策”）被

① 词典规则法即根据若干关键词在文档中出现的频次与位置赋分，并以特定阈值为限进行文本筛选或归类的方法。

② 健康领域关键词为：医疗、医保、卫生、医药、医院、医生、健康、疾病、治病、医药、医疗保险、医疗保障、健身、健康产业、养老、医改、病人、患者、卫计委、诊疗、医务、医学、寿命、控烟、吸烟、食品安全、残疾、中医、老年、疾控、老龄、慢病、慢性病、疫苗、疫情、用药、防治、保健。

错误拆分。分词后的文本去掉“的、我”等单字停用词、数字和英文字母,词语最小长度保留为两字,最终形成用于描述分析和主题模型分析所用的语料数据。

1.2 数据分析

1.2.1 文本词频分布的描述分析

数据分析同样使用 R 语言相关工具包完成。首先描述新闻在月度时间序列的分布情况,以及媒体来源统计,对本文所分析的新闻文本集合进行整体描述。文本关键词及其词频识别与计算是文本挖掘内容的重要方法[9],本文利用词频—逆文档频率(TF-IDF)方法[10]筛选出新闻文本中的高频关键词,并描述高频词的时间序列分布情况,以此发现媒体报道健康中国的聚焦点及其随时间的进展变化。

1.2.2 基于 LDA 主题模型(Topic Model)的文本挖掘

本文运用主题模型(Topic Model)方法对 6 999 份已经分词的文本进行自动分类,尝试发现健康中国相关新闻报道中不同侧重点和方向。主题模型(topic-model)[11]是文本挖掘的重要进展,可以通过无监督类机器学习算法,依据给定的主题数量对文档进行自动分类。该模型假设,整个文档集合中存在若干个主题(topic),每一个特定主题由文档中包含的词汇以不同的概率定义出来,而每一篇特定文档(document)中与某个主题的相关程度也是不一样的。模型拟合的结果之一是展示某一特定主题所关联的高频词及其从属于该主题的概率,通过列举高概率词语组合,可以判断出该主题的内容。[12]此外,主题模型的拟合还可以实现按主题将文档聚类的效果。本文选择主题模型中最为常用的 LDA(Latent Dirichlet allocation)模型[12],利用 R 语言 topicmodels 工具包作为具体工具,对新闻语料进行主题识别。主题数量在运行模型前由研究者自行确定。尽管在理论上可以用 perplexity[13]或 coherence[14]指标评估主题区分效果,进而确定合适的主题数量,但在实际研究中,通常做法是参考相关指标,通过人工审读方式确定主题数量。故本文将在参照 perplexity 指标的基础上,以人工判断的方式,选择分类效果最好的主题数量作为结果,详见结果部分。

2 结果

2.1 报道分布情况

2.1.1 时间趋势分布

图 1 是 2016 年 1 月—2017 年 12 月关于健康中国报道数量的时间分布趋势(以未去重的 13 630 篇新闻计算),从中可以看到,健康中国的报道力度与全国“两会”、全国卫生与健康大会、中共十九大等事件密切相关。

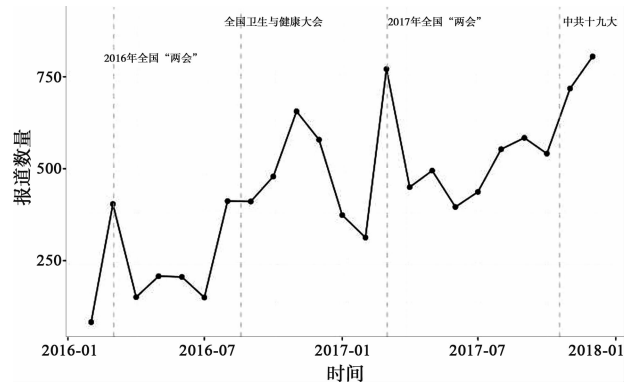

图 1　月度健康中国相关新闻报道量分布

2.1.2 报道来源分布

从媒体来源上看,经过去重的 6 999 篇报道来自超过 500 家国内信息来源,形式以报纸和新闻网站为主,同时也有少量来自政府网站、新媒体、自媒体的信息被报纸和网站所转载。表 1 列举了报道超过 50 篇以上的媒体名称。

2.2 词频分析

本文统计了健康中国报道的关键词及其分布情况,以原始词频和 TF-IDF 加权得分分别统计。原始词频,即特定词语在报道正文中出现的次数,能够在一定程度上表现出新闻报道用语的特点。表 2 是前 32 位原始词频表,图 2 是相应的前 60 位的原始词云图,可以看到,“健康”与“发展”是涉及最多的词语,而“推进”、“建设”、“改革”、“促进”、“实现”、“加快”等表示政府行动的词语也频繁出现。

而 TF-IDF 得分则能够反映出词语在报道文本中的相对重要程度,能够更好的反映出报道的主题和聚焦点。表 3 是 TF-IDF 得分前 32 位的高频词表,图 3 是与之相应的前 60 位的高频词 TF-IDF 得分的词云图,如医疗卫生方面(中医药、医疗、医院、卫生、患者、医生等),体育健身方面(体育、活动、全民健身、运动等),健康管理方面(健康体检、营养、居民等),以及健康产业方面(健康产业、企业、产业等),更多的反映出健康中国的内容。

表 1　媒体来源与报道数量

媒体来源	报道数量	媒体来源	报道数量	媒体来源	报道数量	媒体来源	报道数量
人民网	567	光明日报	213	经济日报	78	健康报网	58
中国新闻网	484	南方都市报	186	新华网	76	中国网	56
央广网	333	华龙网	107	消费日报	73	北京日报	55
新华社	281	澎湃新闻	107	安徽日报	69	云南网	52
人民日报	272	中国青年报	100	深圳特区报	61	北京青年报	52
齐鲁晚报	258	天津日报	96	长城网	61	中国经济时报	50
亚心网	228	国际在线	79	齐鲁壹点	61	华商报	50

表 2　报道中出现的热点词汇及词频(原始词频)

关键词	词频	关键词	词频	关键词	词频	关键词	词频
健康	46 204	推进	13 639	健康中国	11 054	加强	9 423
发展	30 017	人民	12 940	全面	10 572	开展	9 258
中国	23 794	医院	12 289	提高	10 104	推动	8 798
医疗	17 670	社会	12 288	创新	9 924	促进	8 648
国家	17 549	我们	11 532	全国	9 907	重要	8 572
服务	17 306	改革	11 435	问题	9 892	一个	8 488
建设	17 240	实现	11 163	群众	9 740	实施	8 431
工作	14 438	卫生	11 147	企业	9 610	通过	8 377

表 3　报道关键词及其 TF-IDF 得分

关键词	得分	关键词	得分	关键词	得分	关键词	得分
中医药	45. 14	全民健身	29. 06	论坛	25. 37	我们	24. 03
体育	38. 95	健康	28. 52	健康产业	25. 10	医学	23. 88
体检	37. 87	健身	27. 47	工作	24. 87	企业	23. 84
活动	36. 23	营养	27. 36	群众	24. 78	全民	23. 83
医院	35. 78	健康体检	26. 33	患者	24. 52	产业	23. 65
医疗	35. 60	人民	26. 09	养老	24. 46	居民	23. 17
中国	33. 59	服务	25. 91	运动	24. 43	集团	22. 55
卫生	32. 20	发展	25. 47	医生	24. 33	改革	22. 48

图 2　报道热点词汇词云图(原始词频)

图 3　报道热点词汇词云图(TF-IDF 得分)

2. 3 主题模型分析

根据主题模型分析的一般步骤[12],作者分别设定了 5 ~ 25 个主题数量,分别生成不同的主题分类组合。在对分类结果分别进行人工审阅后发现,主题数量设定为 19 的情况下,分类效果最为明显,能够较好的反映出健康中国报道的不同方面。其中,表 5 中的 6 个主题类别与医疗卫生相关,表 6 中的主题与民众的健康生活相关,表 7 则是健康产业相关主题。另外,还剩余一些相关性不明显的主题,也一并列出。

2. 3. 1 医疗卫生类主题

医疗卫生类主题是健康中国报道中最重要的主

题类别。表5列出了医疗卫生类中不同主题词概率得分在前15位的词,以及该类别下新闻文档的数量。其中主题1是与医药卫生体制改革相关的报道,从中可以看到医疗、医保、家庭医生、分级诊疗等当前国家医改重点推进的政策领域。这一主题类别下的有报道659篇,也是所有主题中最多的。主题2是医疗服务相关的话题,围绕医生、患者、疾病等议题展开。主题3是医学教育和医学人才培养的话题。在当前医学人才需求增加、医患矛盾突出等背景下,这一话题也是媒体报道和讨论的热点。主题4与医学科技创新、国际合作等议题相关。健康中国建设以科技创新为重要推动力,同时也为科技发展和成果转化提供了重要平台。此外,该主题还提示了十九大以来愈加重要的"全球健康"议题。尽管在前15位关键词中体现的不明显,但是该主题的文档集中也纳入了诸如习近平总书记访问世界卫生组织、全球健康促进大会在上海召开、中国与东盟、非洲国家地区的卫生合作等新闻报道。主题5是中医、中药相关的话题,也说明中医药以及中国传统医学文化在健康中国建设中的重要地位。主题6是与公共卫生和疾病控制相关的话题,包括了疾病预防、妇女儿童保健、残疾人、农村地区等公共卫生的重点领域。

表5　医疗卫生类相关主题及关键词

	主题1	主题2	主题3	主题4	主题5	主题6
	医疗	医院	医学	医疗	中医药	卫生
	医院	患者	医师	合作	中医	人口
	医保	治疗	教育	国际	中药	预防
	改革	艾滋病	培训	医学	文化	农村
	医改	医生	人才	全球	养生	贫困
	诊疗	医疗	医生	卫生	传承	居民
	公立医院	病人	信息	论坛	中药材	公共卫生
	基层	防治	学生	大会	阿胶	扶贫
	医疗机构	手术	职业	科技	中华	健康服务
	制度	临床	人民网	大数据	国家中医药管理局	儿童
	签约	康复	论坛	应用	药材	计生
	家庭医生	检测	传播	产业	世界	医疗卫生
	医生	肿瘤	协会	世界	种植	防控
	分级	感染	烟草	科技创新	医学	残疾
报道数量	659	398	429	487	310	517

2.3.2 健康生活类主题

表6中的主题与民众的健康生活更加密切。其中主题7是营养与健康生活方式相关的话题,包括饮食、运动、常见疾病知识等。主题8的体育健身也是健康中国的重要内容,其中可以看到从日常锻炼、广场休闲到专业体育赛事等各种类别的体育健身在报道范围中。主题9涉及到健康科普宣传等活动,一定程度上反映了政府和社会开展健康知识宣传、提升民众健康素养的行动。主题10是食品安全相关话题。主题11则是健康中国的另一个重要话题——养老。

表6　社会生活类主题及关键词

主题7	主题8	主题9	主题10	主题11
营养	体育	活动	食品	养老
生活方式	全民健身	公益	食品安全	旅游
膳食	健身	宣传	企业	产业
儿童	运动	现场	消费者	恒大
运动	活动	启动	产品	国际
糖尿病	赛事	主题	监管	老年人

（续）

	主题7	主题8	主题9	主题10	主题11
	居民	项目	仪式	犯罪	养生
	慢性病	比赛	协会	乳业	老年
	高血压	马拉松	本次	品质	健康产业
	口腔	体育产业	大赛	保健食品	医疗
	体重	锻炼	社区	案件	老人
	饮食	场地	科普	营养	项目
	身体	广场	深圳	检察	社区
	食物	健身休闲	基金会	保健品	海南
报道数量	352	533	497	320	331

2.3.3 健康产业类主题

表7中的主题与健康产业相关。主题12首先提及的企业、市场、产品等主要关键词,说明当前健康产业发展的积极态势。也可以看到互联网、(人工)智能等最新的科技进展在健康产业(如健康管理)中的重要作用。主题13和主题14分别代表了健康保险和生物医药这两个健康产业中的重点领域。前者连带着金融、投资等健康产业的拓展领域,而后者则

与上市、集团化等资本运作相关。主题 15 则涉及到市场与投资环境的治理、制度建设等。而农业和农村的话题也在这个主题下出现。主题 16 则提到了边疆和少数民族地区的报道，特别是健康体检相关话题，也显示出健康中国在边疆和少数民族地区实施过程中的特点。

表 7　健康产业类相关主题及关键词

	主题 12	主题 13	主题 14	主题 15	主题 16
	医疗	保险	药品	改革	体检
	企业	市场	医药	企业	健康体检
	行业	公司	企业	产业	青年
	集团	基金	疫苗	制度	全民
	产品	投资	制药	农业	新疆
	市场	行业	医疗器械	市场	免费
	互联网	亿元	研发	试点	工程
	健康产业	板块	流通	农村	自治区
	产业	健康保险	用药	监管	惠民
	品牌	指数	临床	投资	各族
	公司	产品	上市	消费	居民
	智能	数据	集团	就业	村民
	健康管理	健康险	市场	治理	检查
	用户	金融	改革	工程	卫生院
报道数量	473	235	198	126	278

2.3.4 其他类别主题

此外，模型中还归类了其他 3 个主题，大多为国家领导人讲话或重要政策文件，以及宣传落实党的精神的新闻报道（表 8）。这些政治类的报道大多是综合性的，涉及经济社会各个方面，健康中国有时仅作为一个话题在其中提及，因此在主题关键词上体现的不是很明显。而且由于是无监督的自动机器学习，主题 19 也出现了主题混淆的现象。

表 8　其他主题及关键词

	主题 17	主题 18	主题 19
	社会主义	十九大	全运会
	改革	精神	天津
	制度	报告	建议
	教育	学习	法律
	民生	大家	立法
	总书记	孩子	审议
	精神	护理	规定
	脱贫	一名	全运
	文化	工作者	常委会
	政治	党员	委员
	伟大	基层	意见
	九大	老人	草案
	就业	干部	会议
	新时代	宣讲	习近平
报道数量	473	235	198

3 讨论

3.1 健康中国的媒体呈现特点

3.1.1 健康中国在媒体中占有重要位置

首先，从分析结果上看，健康中国作为国家宏观战略，始终保持着高度的媒体关注度，并且还在持续的上升。健康中国在媒体中的重要性可以从报道的时间与来源分布中凸显出来。从报道的时间分布上看，在媒体报道集中的时间段内（如两会、十九大、全国卫生与健康大会），健康中国的报道也呈现明显的多发趋势，这也在一定程度上反映了媒体对于健康中国议题的关注度。从媒体报道来源可以看到，人民网、中国新闻网、央广网、新华社、人民日报等国家级媒体是健康中国新闻报道的最重要主体，这也充分体现出了当前国家级媒体在宣传健康中国政策过程中的重要作用。

3.1.2 健康中国媒体报道领域广泛、内容丰富

无论是词频分布分析还是主题模型分析，都可以看出健康中国报道分布在不同领域，媒体报道的内容与健康中国的政策要点能够基本呼应。且不同类别中文档的分布数量相对平衡，体现出较好的区分情况。这也说明本研究中的报道文本能够相对全面和完整的覆盖健康中国的各个方面。同时，不同领域也呈现出各自特点，如医改和医疗卫生体制问题作为健康中国建设中的核心问题，仍然受到媒体的大量关注。食品安全主题（主题 10）一方面反映出媒体和公众对于食品安全问题高度的关注程度，另一方面也体现了政府在食品安全监管的重视，以及对相关违法行为的打击。而主题 1 中，养老与“产业”、“项目”等词语关联起来，也反映出当前养老向产业化和社会化方向的发展态势。

3.1.3 健康中国的媒体报道态度趋向正面

虽然本文未做专门的文本情感分析（sentiment analysis），但从关键词的罗列中可以发现，媒体报道的健康中国相对正面和积极，“问题”、“矛盾”等负向的词语几乎没有在高频词中出现。这也说明，健康中国作为一项普惠性的国家政策，并未在媒体和社会中引起太多争议。一方面，国家借助媒体为政策的推进营造良好的舆论氛围，另一方面，媒体也对于健康中国政策持积极态度，这都使得健康中国的报道在态度上较为正面。

3.2 文本挖掘方法在卫生政策研究中具有巨大潜力和价值

从方法学角度看,本研究是利用计算机辅助技术,从大规模非结构化文本中提取健康政策信息的一次尝试,体现出了文本挖掘方法在卫生政策研究中的巨大潜力。文本挖掘方法的价值首先在于海量信息的处理能力。如前所述,健康中国是一个内涵极为丰富的国家战略,相关信息的处理要求已经超出传统的定性内容分析方法的能力范围,而这也恰恰是计算机辅助技术的优势所在。同时,文本挖掘的结果也可以为进一步的研究提供线索。如对于关键词及其时间趋势分布的分析,能够对政策进程中的重点和热点问题起到提示作用,便于进一步探索。主题模型本身在实现主题聚类的同时,也能够有效的实现新闻文本的筛选和分类,有助于开展常规的基于人工阅读与编码的内容分析。

当然,本文只是从文本挖掘的角度,从媒体报道的视角展示健康中国的整体进展。这当然无法反映健康中国的全貌,也不涉及效果评估或经验总结。但分析结果也提示媒体报道能够及时的反映出健康中国政策的内容及其进展,进而成为认识和解读这一国家政策的潜在且有效的证据来源。随着数据的积累、方法的进步,相关研究工作将具有很好的政策价值与前景。

3.3 本研究的局限

作为一种新的尝试,本研究也存在一定的不足,主要体现在研究方法的精细度方面。新闻文本属于高度非结构化的数据,固然 TF-IDF、LDA 主题模型等机器学习方法在挖掘文本信息方面较传统的基于统计规则的方法有所深入,但分析结果的呈现仍相对简单。特别是新闻背景、报道时间、新闻类别、来源分类等重要的文本属性信息也并未在分析中体现。近年来,在文本挖掘的前沿研究中,文本属性信息纳入主题模型分析已经有了很多进展[15],而以词向量方法为代表的深度学习方法在自然语言处理领域的突破[16],也使得文本内在的语义关系分析成为可能。这些技术方法与新闻文本数据的进一步结合,也将进一步增强基于海量数据进行卫生政策研究的能力。

作者声明本文无实际或潜在的利益冲突。

参 考 文 献

[1] 李滔, 王秀峰. 健康中国的内涵与实现路径[J]. 卫生经济研究, 2016(1): 4-10.

[2] 郑成武. 政府与大众媒体公共关系模式[J]. 中国行政管理, 2007(11): 81-82.

[3] 喻国明. 大数据方法与新闻传播创新:从理论定义到操作路线[J]. 江淮论坛, 2014, 266(4): 5-7.

[4] 党静萍, 张美, 庞瑞,等. 陕西省神木县"全民免费医疗"媒体报道及影响分析[J]. 中国卫生政策研究, 2010, 3(8): 11-14.

[5] 尹孔阳, 丁一磊, 朱大伟, 等. 基于2015-2017年舆情监测数据的基层医疗卫生改革舆情评价[J]. 中华医学图书情报杂志, 2017, 26(8): 28-33.

[6] 朱恒民, 马静, 黄卫东. 基于领域本体的中文 Web 文本主题特征抽取方法[J]. 情报理论与实践, 2008, 31(2): 286-288.

[7] Loo M P J V D. Thestringdist package for approximate string matching[J]. R Journal, 2014, 6(1): 111-122.

[8] QinWenfeng and the authors of CppJieba for the included version of CppJieba, jiebaR: Chinese Text Segmentation. R package version 0.9.1 [EB/OL] https://CRAN.R-project.org/package=jiebaR

[9] 陈晓云. 文本挖掘若干关键技术研究[D]. 上海:复旦大学, 2005.

[10] 樊梦佳, 段东圣, 杜翠兰, 等. 统计与规则相融合的领域术语抽取算法[J]. 计算机应用研究, 2016, 33(8): 2282-2285.

[11] 姚全珠, 宋志理, 彭程. 基于 LDA 模型的文本分类研究[J]. 计算机工程与应用, 2011, 47(13): 150-153.

[12] 胡吉明, 陈果. 基于动态 LDA 主题模型的内容主题挖掘与演化[J]. 图书情报工作, 2014, 58(2): 138-142.

[13] 曹娟, 张勇东, 李锦涛, 等. 一种基于密度的自适应最优 LDA 模型选择方法[J]. 计算机学报, 2008, 31(10): 1780-1787.

[14] Stevens K, Kegelmeyer P, Andrzejewski D, et al. Exploring topic coherence over many models and many topics[C] Conference on Empirical Methods in Natural Language Processing, 2013.

[15] Roberts M E, Stewart B M, Airoldi E M. A Model of Text for Experimentation in the Social Sciences[J]. Publications of the American Statistical Association, 2016, 111(515): 988-1003.

[16] 林奕欧, 雷航, 李晓瑜, 等. 自然语言处理中的深度学习: 方法及应用[J]. 电子科技大学学报, 2017, 46(6): 19-24.

[收稿日期: 2018-03-30 修回日期: 2018-07-10]

(编辑 刘博)