

·专题 政策文本计算方法研究·

编者按:当 Jim Gray 提出数据密集型科学研究范式, David Lazer 提出计算性社会科学等研究理念后,以数据分析和计算思维为引导的方法理论也逐渐与传统人文社会科学结合,产生了数字人文、社会计算、计算传播学以及相关计算性人文社会科学,并得到广泛的应用和发展。计算分析方法在政策研究和政策分析领域却一直独立发展,虽然在政见分析、比较政党研究、政治演讲分析、政策认同和政策情感研究领域都取得了不错的研究进展,但政策计算分析一直没有作为一个独立术语或研究范畴被提出。事实上,在政策分析领域计算机辅助政策分析和政策文本计算分析拥有悠久的历史,并产生了如内容分析(content analysis)、一致性分析(concordance analysis)、话语分析(conversational analysis)、话语文本分析(discourse analysis)、计算诠释学(computational hermeneutics)、定量文本分析(qualitative text analysis)等相关的研究方法和工具。

本专题三篇研究论文系教育部人文社会科学青年项目“基于概念统计的信息政策文本计算与实证”(项目编号:11YJC870020)与国家社会科学基金青年项目“信息政策扩散与转移研究”(项目编号:12CTQ024)的系列研究成果。具体而言,《政策文本计算:一种新的政策文本解读方式》主要从方法论角度讨论政策文本计算方法的可行性与应用前景,通过梳理计算分析方法在政策分析领域的应用,对政策文本计算的方法论、应用工具和典型研究议题的跟踪,提出了政策文本计算方法的主要特征以及政策计算分析的可能应用前景;《中国信息化政策扩散中政策主题跟踪研究》则以课题组构建的中国信息政策语料库中的国民经济和社会信息化中长期规划为研究样本,从主题承继与主题创新、主题跃迁与主题衰退、政策扩散涟漪效应和漏斗效应等角度研究了政策扩散中的政策主题变化特征,并从信息化政策文本的计算分析中,发掘了信息化政策在地域扩散和历时扩散两个维度的扩散特征;《政策扩散时间滞后效应及其实证评测——以江浙信息化政策实践为例》则选择了从扩散滞后效应测度的角度,通过二维政策主题密度分布中的差异化象限,提出通过政策词频的密度分布可以间接反映政策扩散所处的阶段,从而间接测度扩散滞后性,并针对两省信息化政策文本与信息化政策理论研究的实测,发现了该方法的可行性。

政策文本计算:一种新的政策文本解读方式*

裴 雷 孙建军 周兆韬

(南京大学信息管理学院 江苏南京 210023)

摘 要 政策文本计算是大数据环境下政策分析科学与计算科学交叉融合的产物。文章通过对政策文本计算的方法论、应用工具和典型研究议题的跟踪和梳理,提出了政策文本计算方法的主要特征与不足,并讨论了该方法在精细化政策分析和定量政治研究领域的研究前景。

关键词 政策文本计算;政策诠释;政策分析;方法论

中图分类号:D03 文献标识码:A DOI:10.11968/tisyqb.1003-6938.2016110

Policy Text Computing: A New Methodology of Policy Interpretation

Abstract Policy text computing is a new integrated methodology combined with policy analysis science and computing science in the big data era. This paper reviewed the development of computing methods in political text analysis, summarized the typical research topics, tools and applications in this area, then concluded the main characters and shortcomings of this methodology, and discussed the potential application of policy text computing in meta-policy analysis and qualitative political analysis.

Key words policy text computing; policy interpretation; policy analysis; methodology

* 本文系教育部人文社会科学青年项目“基于概念统计的信息政策文本计算与实证”(项目编号:11YJC870020)与国家社会科学基金青年项目“信息政策扩散与转移研究”(项目编号:12CTQ024)研究成果之一。

收稿日期 2016-10-29,责任编辑 魏志鹏

1 引言

政策文本是指因政策活动而产生的记录文献,既包括政府或国家或地区的各级权力或行政机关以文件形式颁布的法律、法规、部门规章等官方文献,也包括政策制定者或政治领导人在政策制定过程中形成的研究、咨询、听证或决议等公文档案,甚至包括政策活动过程中因辩论、演说、报道、评论等形成的政策舆情文本,历来是政策研究的重要工具和载体^[1]。如在政策研究方法中,Trauth^[2]认为主要有“预测-描述”的诠释范式、“价值批判-价值构建”的价值范式、政策过程范式以及政策评估和绩效范式等主要形式,其中诠释范式又分政策文本分析、政策分类或框架体系、政策生命周期律、政策社会系统等理论。可见,政策文本研究在政策分析研究领域占有重要地位。

随着计算机方法的引入应用,政策文本分析所能处理的素材量和处理精度得到了大幅提升,并引入了新的方法和理念。尤其是政策文本数据,如文本型数据(Textual Data)、数据文本(Text as Data)、文本数据空间(Text Universe)等相关概念的提出,研究者在政策文本内容分析法的基础上相继提出了政策文本语料库分析和政策文本数据挖掘方法,并利用上述方法解读和获知政策立场、政策倾向、政策价值、政策情感等深层政策内涵以及广义的政策比较分析。我国李江等^[3]提出运用政策计量(Policymetrics)的研究思路来揭示政策引用、主题共现以及机构共现等政策关系。本文通过梳理国内外政策文本内容分析、政策语料库以及政策文本挖掘的相关理论研究进展,探讨了政策文本计算分析的可行框架与应用前景。

2 政策文本计算的方法论解析

政策文本计算是21世纪初 Michchael Laver、Kenneth Benoit 和 Will Lowe 等提出的,运用计算机科学、语言学和政治学的理论建立的海量政策文本挖掘和计算分析框架。政策文本计算主张运用政策编码、政策概念词表或政策与语词之间的映射关系进行政策概念的自动识别和自动处理,最终构建从

政策文本到政策语义的自动解析框架,并在此基础上关注政策文本及其内涵分析。具体到方法论层次,政策文本计算被认为是一种非介入式、非精确性的解析方式,并广泛应用于元政策分析领域。

2.1 政策文本计算是非介入式研究方法

从分析主体看,政策文本计算源自政策话语分析,是作为政策分析的一种非介入式方法引入政策科学领域。在政策分析传统中,一般强调以政策利益相关者的心理或行为假设为出发点,以公共政策绩效或调整结果为评价,并对政策过程、政策工具的可行性进行相关评估研究。因此,不论是运用控制论、运筹学、系统分析或博弈论等过程分析方法,还是运用行为科学、社会心理学、组织理论、权威理论、群体理论等行为解释理论,或是预设一定的分析框架予以验证,都不可避免地要预设政策立场以及政策价值取向,作为政策分析的判断标准。而政策文本分析或政策话语分析(Discourse Analysis)认为政策文本已经蕴含了政策交流系统中的语义与价值情感^[4],研究者无需再设计相应的政策框架,仅需要转述或提取政策文本中蕴含的语义,并有序表达。

非介入式方法的优点是研究结果的中立与客观性,弱化了研究者因政策立场偏见、被调查者(样本)主观偏性而带来的效度瑕疵^[5],并且便于将研究结果复现和应用于大范围尺度和长时间尺度,在宏观政策研究、比较政策研究和非预见性研究中具有广阔应用前景^[6];但不足是文本处理过程效度不够,无法兼顾政策语境的差异性,研究结果的可解释性较弱。

2.2 政策文本计算是非精确性研究方法

从分析方法看,政策文本计算的出发点是政策文本的自然语言处理,即政策的语法解析。虽然众多政策文本计算研究者试图构建语法文本与语义文本、语用文本的映射关系,或依据研究者的理解构建分析词表或抽取若干政策元素或属性,然后以“聚焦”方法跟踪研究。但早期通过这种“重构”或“再塑造”方式建构的政策文本内容分析方法,不仅耗时长、成本高,而且在方法论上形成了研究者事实上的“意识介入”,研究者本身作为研究工具存在于研究过程,其可靠性依然为学界所诟病。

随着政策文本数量的激增和开放获取的便捷

性,基于海量政策文本的语义自动提取方法日益成熟,在显性政策要点、政策情感以及政策立场领域的识别精度越来越高。如 Hjorth 等^[7]对自动文本分析与专家调查分析的对照分析发现,两种自动分析方法和专家分析对 CMP RILE measure 政治演讲语料库的对比分析中,自动分析政策主题排序与专家主题分析排序的 spearman 相关系数(Spearman's ρ)显著优于专家与一般选民识别的 spearman 相关系数。不过,从政策计算的分析结果看,政策文本分析结果仍然是非精确性的。如 Proksch 和 Slapin^[8]认为,现有的政策文本处理的算法缺陷、政策文本的语言特征以及政策文本结构和语境适用性缺失都是政策文本计算分析的致命不足;虽然 Mikhaylov 和 Benoit 等^[9-10]在研究政见语料库时均发现,专家研究的手工编码误差并不比计算机自动编码误差小,因而政策文本计算的分析误差来自编码本身,而非计算机算法或处理误差。而在主流政策分析领域,政策研究者虽认可政策计量在问题识别和政策分析中的价值^[11],但认为政策计算分析的结果仍是非精确性的、参考性的^[12]。Grimmer 和 Stewart^[13]甚至提出政策自动文本分析的“4 原则”:第一,所有的自动文本分析结论都是“错误”的,但可用;第二,自动文本分析永远无法替代政策分析者本身;第三,永远没有最好的文本分析解决方案;第四,连说三遍“研究效度”。因此,研究者普遍认为,加强政策的解释性分析,并融合质性方法的混合方法更具有应用前景^[14]。

2.3 政策文本计算聚焦于元政策分析

在政策分析中,元政策一般是“政策的政策”,是从现有政策中抽象出的理念或方法,其关注的是整个政策系统及其改进,涉及公共政策的指导思想、价值标准、行为准则、程序步骤、方式方法等^[15]。而从分析对象看,政策文本计算处理对象多为政策语词、政策概念(主题)、政策义素等显性政策功能词,或政策立场、意识形态、政策倾向、政策情感、政策价值、政策态度等元政策领域。

究其原因:首先,元政策分析的非精确编码属性与政策计算分析的非精确性具有很好的契合度,具备了元政策计算分析的方法论基础;其次,元政策脱离了政策工具、政策区域以及政策地域的语境影响,

一是形成了最大可能的频次聚焦,二是具备了跨区域政策比较的可能性;最后,元政策具有非显在性,无法通过简单观察获知,而借助计量或计算方法的元政策识别机制能为研究者所接受。

3 政策文本计算分析的典型方法与议题

政策文本计算既是一种政策分析研究理念和研究框架,也是完整的政策分析流程。从分析方法角度看,Wiedemann 将政策文本计算,或称为计算机辅助文本分析(Computer Assisted Text Analysis,CATA)分为文本内容分析、文本数据分析和文本挖掘三个研究层次,并先后经历了计算化内容分析(Computational Content Analysis,CCA)、计算机辅助定量数据分析(Computer-Assisted Qualitative Data Analysis,CAQDA)以及语料计算学(Lexicometrics for Corpus Exploration)等不同发展阶段^[16];从分析流程角度看,Grimmer 和 Stewart^[13]将政策计算分为政策文本获取(Acquire Documents)、政策文本处理(Process)和政策文本分析三个典型阶段(见表1)。两者均认为政策文本处理和文本挖掘方法是政策文本计算分析的核心,本文则从政策文本内容分析、政策文本计量分析、政策文本数据分析和政策文本挖掘四个方面考察政策文本计算的典型方法。

表1 Grimmer&Stewart 政策文本计算分析方法

研究主题	典型议题与方法
政策文本获取	政策语料库、政策数据库、开放政策源、政策文本采集
政策文本处理	政策词表、分词、同根词合并、停用词表、文本术语矩阵(DTM)、特征词、语词加权、词义距离
政策文本分析	政策文本分类、基于词表方法、基于概率分析方法、无监督学习、监督学习、类别识别与主题识别、意识形态测度、政策角色识别

3.1 政策文本内容分析方法

政策文本内容分析是一种介于定性与定量之间的半定量研究方法,与之类似的还有一致性分析(Concordance Analysis)、话语分析(Conversational Analysis)、话语文本分析(Discourse Analysis)、计算诠释学(Computational Hermeneutics)、定量文本分析(Qualitative Text Analysis)等研究方法。从20世纪80年代开始业内就陆续研制了相关的文本分析软件用于文本标记、文本编码和相应的编码管理工具,如

Atlas.ti、MAXQDA、QDA Miner、NVivo、SPSS Text Analytics for Surveys、QCAmap、CATMA、LibreQDA、MONK Project 等文本数据管理软件工具。虽然引入了计算机软件对政策文本进行概念抽取和量化统计,并具有文本数据的自动统计和关系识别方法,但其概念抽取方法仍采用传统的文本分析方法和流程,在数据处理环节仍主要依赖研究者的人工提取,体现为一种半计算化分析工具。

因此,这类计算处理方法能够处理的政策文本数据有限,一般处理政策样本集(Sample, $n \leq 200$),最多通过协作方式处理政策主题集(Subsets, $N \approx 1000$)范畴的政策文本集,而对政策语料库(Corpus, $N \geq 10000$)基本上无法处理。因而,这类研究方法的研究议题也主要沿袭了政治学和诠释学中的政治话语研究和政治文本内容分析框架中的符号论和政治语词解读(政策主题识别与比较)的研究传统。

3.2 政策文本计量分析方法

政策文本计量分析主要是采用文本计量分析的基本理论与方法,通过对已有政策文本数据库或政策文本语料库在政策主题分布、政策发布时间序列分布、政策引证以及政策主体关系等要素进行计量分析^[3]。在 Grimmer 的政策计算分析框架中,政策文本主要来自政策数据库和已有语料库、网络政策文本和非电子化政策文本。因此,政策文本计量分析的主要方法和工具也主要有三种类型:一是政策文本数据库自有的文本计量分析方法与工具,如 Lexis Nexis、ProQuest、Westlaw、HeinOnline、北大法宝和 CNKI 政府公报数据库等政策或法律文本数据库,利用数据库自带的字段设定结合政策主题、类型、时间、地域等进行政策统计或计量分析,或应用共词或共现分析,能有效分析政策文献增长、扩散、流变等变化规律;二是利用网络分析和替代计量学(Altimetrics)方法和工具进行网络政策文本分析^[17],如 Wiley、NPG 和 PLOS One 等开始提供 Altimetric 服务,Altimetric 也可以对国内新浪微博进行追踪,因而对社会媒体中的政策文本以及跟踪研究也成为可能,如匹兹堡大学创建的 MPQA 政策辩论语料和卡内基梅隆大学 Sailing 实验室 Jacob Eisenstein 和 Eric Xing 创建的政治博客文本集语料;三是通过政策文

本采集与语料库构建并提出新的统计口径和研究方法,如苏竣和黄萃等对中国科技政策的类型统计分析^[18]以及卡内基梅隆大学 Wilson 等对网站隐私政策的主题解析分析^[19]。

3.3 政策文本数据处理方法

从政策文本的范围看,政策文本结构性差异很大:既有政府的政策文本、法律档案(听证会材料、判例),也有政策新闻、媒体数据和政策研究文献;既有总统竞选纲领、演说文本集,也有社交媒体的公众政治言论和政治评论。而通过自然语言处理将政策文本解析为结构化文本数据(Textual Data),并构建语词、语义或情感等特殊对象,不仅能形成对大规模政策文本语料的系统化处理,而且能在不同的政策文本集中进行比较分析和一致性分析,推动政策文本融合分析。结合政策文本分析的应用,典型的研究方法和工具有政策文本自然语言处理和语法计量分析、政策文本处理以及政策语义分析(见表2)。

在政策文本数据处理过程中,政策文本或语料集适用于通用的自然语言处理方法和文本数据处理方法,政策语词分析和政策语义分析在政策主题统计(聚类)、政策热点识别、政策意见分析中应用较多^[20-21]。目前,在政策文本处理领域最受关注的议题:一是语料库尺度的政策内容分析^[22-24],主要是对政策语料库的统计和计量分析,识别政策语境中的热点议题^[25],关注政策议题的扩散或影响^[26-27],尤其是政治演说语料库、政见语料库、政治纲领语料库分析;二是政党和选举研究中的政策立场分析和政策倾向研究,政策文本计算的概念本身即为比较政见研究(CMP)的 Michchael Laver 提出,而基于先验词权(Reference Score)的 WordScore 和无先验词权的 WordFish 也是政策文本计算分析中应用最广泛的分析软件,CMP 以及后续研究项目(MARPOR)提供的政见语料库也是采纳率最广的语料库。

3.4 政策文本数据挖掘方法

文本挖掘,又称为文本数据挖掘或文本知识发现,是指在大规模文本集合中发现隐含的、以前未知的、潜在有用的模式的过程^[28],涉及数据挖掘、机器学习、统计学、自然语言处理、可视化技术、数据库技术等多个学科领域的知识和技术^[29]。与政策文本处

表2 政策文本数据处理典型方法与工具

研究主题	典型议题与方法	典型工具
政策文本自然语言处理(NLP)和语法计量分析(Lexicometrics)	文本向量空间(VSM)、分词(Stemming/Tokenization)、词性标注(POS)、功能词提取、剪枝(Pruning)\ DTM 矩阵、Key Word in Context (KWIC)、术语识别、词频统计、词频分布、政策词表、共词分析、多维分析、网络分析	OpenNLP、Natural Language Toolkit (NLTK)、WordSmith、WordStat、RSegment、SVMTool、BrillWin、Wmatrix、ICTCLAS、SnowNLP、Lexical Freenet、Lexicoder、CoreLex、UCInet、MAXDictio、NetDraw
政策文本处理	政策文本自动分类/聚类、自动编码、自动摘要	RapidMiner、Carrot2、ReadMe、PolyAnalyst、LIWC
政策语义分析	政策立场分析、政策倾向研究、政策主题发现与跟踪(TDT)、事实数据抽取(Event Data Extraction)	Gensim、Stanbol、WordScore、WordFish、ManifestoR、AUSTIN、D-NOMINATE、ReadMe

理更注重政策语词或语义分析相比,政策文本数据挖掘更注重在大量文本数据集中发现分类/聚类特征、发现关联知识或规则,并注重深层潜在语义的知识发现。因此,政策情感分析、政策意见分析、政府行为预测等典型方法得到政策研究领域的广泛关注,如 Saremento 等对用户评论的政策倾向分析^[30]、Hopkins 和 King^[31]对博客政策意见的分析。政策情感分析在西方国家选情预测中尤为关注,包括政治领导人的政策情感倾向^[32]、选民的情感反馈与倾向^[33-34]以及整体选情预测^[35-37];在政策意见分析中,公众意见收集和意见追踪也是常见的研究主题,并将公众政策意见与其政治立场和政党支持度关联,建立了计算化的政党舆情监测、政党竞争或政党派系识别以及政策结果评估的分析方法^[38-39];政府行为预测体现了政策预测分析的方法和思路,通过对政府领导人、政党的竞选纲领或关键政策文本的分析,挖掘潜在的政策热点或发展轨迹。国内研究者也利用数据挖掘方法对政策热点^[40]以及政策价值^[41]进行了分析,或系统利用文本挖掘方法对政策文本的内部结构关系进行了主题识别或关联分析^[42-44],但总体上缺乏系统性和连续性。

4 政策文本计算应用研究进展

4.1 政策文本语料库建设

政策语料库以及语料库语言分析是政策文本计算分析的基础。早期的政策语料库一般针对政府出版物或公开政治文本进行采集加工,如政策条文、相关政策解释、政治人物传记、语录或新闻纪录等;现在则扩展到更加多样化的语料来源。除了 Lexis Nexis、北大法宝等传统的法律信息服务提供商,目前比较典型的政策语料库有:

(1)德国柏林社会科学研究比较政见研究

项目政见文本语料库 (MRG / CMP / MARPOR)^[45]。Manifesto 语料库是目前政策分析领域加工最为成熟的开放政策语料,包括 1945-2015 年 70 年跨度,涉及所有欧洲国家和少数英美联邦国家(美国、加拿大、澳大利亚、南非、新西兰)总计超过 50 个国家的 4051 个政见语料集,涵盖了 1979-1989 年政见研究组 MRG(Manifesto Research Group)、1989-2009 年比较政见研究 CMP (Comparative Manifestos Project)以及当前基于政治表达的政见研究 MARPOR (Manifesto Research on Political Representation)持续研究的政策语料。在语料分析工具包中,既包括手工编码的政策术语编码手册(Code Book),也包括 794,536 个跨语种的机器识别政策术语、短语或词条;既包括软件版本的 WordScore 分析工具,也包括 R 语言的分析包 ManifestoR。因此,Manifesto 语料库和 WordScore 分析软件是目前政见分析和政策文本计算领域引用率最广的语料库,尤其在政策立场和政策倾向研究中。

(2)美国康奈尔大学政策文本语料库(Corpus of political discourse)^[46],它是康奈尔大学计算机系庞大的语料集中的一个子集,主要是由 Matt Thomas、Bo Pang 和 Lillian Lee 整理的总统国会演讲数据集(Congressional speech data),同时因 Lillian Lee 设计开发了相应的情感开发工具 ReadMe,因此在严肃政策文本的政策情感研究领域受关注度较高,目前共有 22 篇研究文献利用或援引了该数据集。

(3)美国匹兹堡大学计算机系的 MPQA Opinion Corpus 语料库 (Multi-Perspective Question Answer, MPQA)^[47],主要是新闻报纸素材的语料,包含 4 个子库、4 个词表和基于语料库分析技术开发的 Opinion-Finder 系统(目前提供 2.0 版本下载),其中有一个专门子库为政策辩论数据库(Political Debate Data)。同

时,因其情感标注系统比较出色,因而也是博客、评论等开源语料政策情感分析的主要素材和工具。

(4)卡内基梅隆大学计算机系 Sailing 实验室的政治博客语料库^[48]。由 Jacob Eisenstein 和 Eric Xing 整理开发,主要采集了 2008 年 6 个博客平台的 13246 个政治博客文本记录,并且通过意识形态的分层抽样,也是政治博客研究比较重要的语料资源。类似的语料集还有美国海军学院 Twitter 政策语料集。

(5)香港浸会大学整理开发的政治演讲语料集 (Corpus of Political Speeches-HKBU Library)^[49]。目前主要包括 4 个部分:美国历届总统演说语料文本集和多媒体文本(1789-2015)(约 443 万字)、历届香港总督或特首施政报告语料集(1984-1996,1997-2015,约 43 万字)、历届中国台湾地区领导人新年致辞和双十演讲语料集以及中国历届政府总理施政报告语料集,是比较完整的中文政策语料集之一。

此外,德国柏林 Brandenburg 科学研究院的阿德莱登·巴拉巴西提供的德国政策语料集^[50]则结合了政策语料分析与可视化研究,利用这个政策语料集可进行总统演讲频率、演讲主题和演讲所涉及的政策语言的可视化分析,网站提供粗语料、分词后的语料以及标引后的语料等不同版本的语料。

4.2 政策文本分析工具研制

因语境意义对政策文本分析的现实意义更大,当前政策文本计算比较注重政策词典和政策文本分析专用工具的研制。目前,主要有两类研究方法:

第一,测试通用文本分析工具在政策文本分析中的适用性。典型如政策情感分析领域,Lori Young 等^[51]对 DICTION、LIWC、RID、TAS/C、ANEW、DAL、WNA、PMI 以及 LSD 等众多情感分析词典的对比研究发现,LSD 在选民情绪跟踪研究和对比研究中具有明显优势;Bei Yu 等^[32]则发现政策评论或政策演说文本中,情感词汇的使用频率明显低于普通文本,并且不同于一般情感分析主要负载于谓词描述,大量政策情感负载于名词性的体词描述中,需要结合上下文才能完全识别,因此在政策文本分类的算法中(SVM、NB),训练文本需更充足。

第二,研制政策分析专有词表和分析工具。典型如政策立场和政见研究中的 WordScore 算法和

WordFish 算法。两种方法都注重政策语词对政策内涵表达的影响权重差异,WordScore 方法通过专家判定的参考文本作为政策语词权重依据,从而生成政策分析文本中政策内涵的表达效果,其实质是对词频结果进行语义加权处理,类似一种基于动态“词典”的分类算法;WordFish 算法认为政策文本具有不同的政策特征向量,在某一特征中政策语词的概率分布符合泊松分布,因此可以通过一种类似非监督学习的方式对政策文本所蕴含的“政策立场”进行分类。由于 WordScore 算法的分类效果和可解释性优于 WordFish,但分类效果受参考文本的影响大,在历时分析或跨文化环境的比较参考分析中效度不高。此外,政策文本计算因德语或北欧国家特有的构词方式而具有一定研究效度,而在英语地区却并不显著,这也是当前政策文本计算研究兴盛于德国和北欧,而英美地区进展缓慢的主要原因。

因此,政策文本分析词表、文本分析效度改进工具和跨语言政策文本分析工具都是目前政策文本分析工具研究的热点问题。

5 政策文本计算的应用前景与障碍

政策文本计算方法是大数据环境下政策分析科学与计算科学交叉融合的产物,目前已经形成了较为稳定的研究议题和研究队伍。随着政策文本资料的日益丰裕以及大数据分析方法的日益为社会科学研究者所采纳,可以预见未来政策文本计算在精细化政策分析和定量政治研究领域具有广阔的研究前景。

5.1 政策文本计算的应用前景

就政策文本计算的应用领域而言,精细化政策分析主要体现在政策预测、政策冲突分析与政策辅助决策、元政策评价与政策比较等研究领域,定量政治研究则体现为政党研究、政治立场、政治态度、政策认同、政治联盟以及选举、外交等政治活动领域。

第一,政策文本计算在精细化政策分析领域已经具有研究基础,尤其在语料库政策语言分析中形成了相对成熟的研究框架。首先,计算方法的引入提供了跨语料分析和实时语料分析的研究可能,对政策预测的时效性和精确度都将大大提升;其次,计算方法的引入将改进政策分析的精度和深度,在政策

制定中不同政策源的立场识别和主题识别可以避免显性的政策条款冲突,同时对政策主题关系识别也能评判政策相似度或政策形式质量预判,辅助政策制定决策;再次,通过政策文本与政策语义的对应关联,能够挖掘政策的潜在语义和元政策要素,从政策价值、政策倾向、政策工具、意识形态等高度评价或比较不同时期、不同地域甚至不同国别的政策差异,更好地跟踪政策扩散过程,促进政策学习与创新。

第二,定量政治研究则融合了政治学、媒介理论以及政党研究的理论视角,能通过泛在的政策文本载体,识别公众的政治态度、政治立场以及不同主体之间的政治互动关系,进一步通过政治文本解析框架可以分析政治立场、政治距离和政治关系紧密度,从而发现政党合作、国际合作的潜在空间;另一方面,通过不同政治参与主体的互动机制,可以在政策认同、政党监督、政党竞争以及选情预测等领域进行有效分析。

5.2 政策文本计算的应用障碍

正如国内外学者对人文社会科学计算方法的担忧^[13,52-53],政策文本计算不论从方法论本身,还是从应用场景的研究效度看,其只能作为决策分析工具,

而无法替代政策分析者本身。究其原因,首先,政策文本语料库的局限。语料库具有一定时效性与完备性限制,而语料库规模和多样性是政策文本计算分析效度的关键,但语料库构建成本和可用技术的限制使得语料库很难完全满足政策分析者的需要;其次,文本挖掘和相关计算分析方法的局限。文本挖掘结果的呈现是抽象的或数据化的,只有结合相关的应用背景才能完全理解相关内涵;文本挖掘或计算分析注重研究创新点的突破,很难兼顾整体研究面的覆盖,因而其结论往往是片面的、非系统的;文本计算分析方法是探索性分析方法,其研究结论是非可预期的、不确定的,而文本语料库建设是高成本的,政策文本计算具有一定的研究风险;第三,政策文本计算是跨学科研究方法,需要政策研究和计算机研究学者的紧密配合,而实际研究过程中很难兼顾二者。

因此,在未来的政策文本计算研究实践中,一是需要加强学科合作,推动专业化的政策语料库的建设,开发适用于政策文本分析的工具;二是政策文本计算研究具有良好的中立性与客观性,国家应该在智库建设和国际政策比较研究中更加重视政策量化和定量政治研究。

参考文献:

- [1] Chilton P A, Schäffner C. Politics as text and talk: analytic approaches to political discourse [M]. John Benjamins Publishing, 2002.
- [2] E. M. Trauth. An integrative approach to information policy research [J]. Telecommunications Policy, 1986, 10(1): 41-50.
- [3] 李江, 刘源浩, 黄萃, 等. 用文献计量研究重塑政策文本数据分析——政策文献计量的起源、迁移与方法创新 [J]. 公共管理学报, 2015(2): 138-144.
- [4] 杨正联. 公共政策文本分析: 一个理论框架 [J]. 理论与改革, 2006(1): 24-26.
- [5] 黄萃, 任弢, 张剑. 政策文献量化研究: 公共政策研究的新方向 [J]. 公共管理学报, 2015(2): 129-137.
- [6] Beauchamp N, Laver M, Nagler J, et al. Using Text to Scale Legislatures with Uninformative Voting [EB/OL]. [2016-09-20]. http://nickbeauchamp.com/work/Beauchamp_scaling_current.pdf.
- [7] Hjorth F, Klemmensen R, Hobolt S, et al. Computers, coders, and voters: Comparing automated methods for estimating party positions [J]. Research & Politics, 2015, 2(2): 1-9.
- [8] Sven-Oliver Proksch, Jonathan B. Slapin. How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany [J]. German Politics, 2009, 18(18): 323-344.
- [9] Mikhaylov S, Laver M, Benoit K R. Coder reliability and misclassification in the human coding of party manifestos [J]. Political Analysis, 2010, 20(1): 78-91.
- [10] Benoit K, Laver M. Estimating party policy positions: Comparing expert surveys and hand-coded content analysis [J]. Electoral Studies, 2007, 26(1): 90-107.

- [11] Hansen, Ejnar M. Back to the Archives? A Critique of the Danish Part of the Manifesto Dataset [J]. *Scandinavian Political Studies*, 2008, 31(2):201-216.
- [12] Benoit K, Laver M, Mikhaylov S. Treating words as data with error: Uncertainty in text statements of policy positions [J]. *American Journal of Political Science*, 2009, 53(2):495-513.
- [13] Grimmer J, Stewart B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts [J]. *Political Analysis*, 2013, 21(3):267-297.
- [14] 孙建军. 大数据使社科研究不再“望数兴叹”[N]. 人民日报, 2016-02-18(7).
- [15] 李民, 肖旭东. 元政策视角下科学发展观的价值分析[J]. 江汉论坛, 2009(11):17-20.
- [16] Wiedemann G. Computer-Assisted Text Analysis in the Social Sciences [M]. *Text Mining for Qualitative Data Analysis in the Social Sciences*. Springer Fachmedien Wiesbaden, 2016:17-53.
- [17] Piwowar H. Altmetrics: Value all research products [J]. *Nature*, 2013, 493(7431):159.
- [18] 苏竣, 黄萃. 中国科技政策要目概览 [M]. 北京: 科学技术文献出版社, 2012.
- [19] Wilson S, Schaub F, Ramanath R, et al. Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work? [C]. *International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016.
- [20] Simon B A F, Xeon M. Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis [J]. *Political Analysis*, 2010, 12(1):63-75.
- [21] Klebanov B B, Beigman E. Lexical Cohesion Analysis of Political Speech [J]. *Political Analysis*, 2008, 16(4):447-463.
- [22] Ädel, Annelie. How to Use Corpus Linguistics in the Study of Political Discourse [M]. Anne O'Keeffe and Michael McCarthy. *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, 2010.
- [23] Rowe C. Politics as Text and Talk: Analytic Approaches to Political Discourse, by Paul A. Chilton; Christina Schäffner [J]. *International Politics*, 2004, 41(2):286-287.
- [24] 涂端午. 政策生产: 价值的权威控制及其演变——1979-1998 年中国高等教育政策文本分析 [J]. *比较教育研究*, 2009(11):95-96.
- [25] Laver M, Benoit K. Locating TDs in Policy Spaces: The Computational Text Analysis of Dáil Speeches [J]. *Irish Political Studies*, 2010, 17(1):59-73.
- [26] Budge I, Pennings P. Do they work? Validating computerised word frequency estimates against policy series [J]. *Electoral Studies*, 2007, 26(1):121-129.
- [27] Monroe B L. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict [J]. *Political Analysis*, 2008, 16(4):372-403.
- [28] 谌志群, 张国煌. 文本挖掘研究进展 [J]. *模式识别与人工智能*, 2005, 18(1):65-74.
- [29] 郭金龙, 许鑫, 陆宇杰. 人文社会科学研究中文本挖掘技术应用进展 [J]. *图书情报工作*, 2012, 56(8):10-17.
- [30] Sarmento, Lu, Carvalho P, Silva, M, et al. Automatic creation of a reference corpus for political opinion mining in user-generated content [C]. *International CIKM Workshop on Topic-Sentiment Analysis for MASS Opinion*. ACM, 2009:29-36.
- [31] Hopkins D J, King G. A Method of Automated Nonparametric Content Analysis for Social Science [J]. *American Journal of Political Science*, 2010, 54(1):229-247.
- [32] Yu B, Kaufmann S, Diermeier D. Classifying Party Affiliation from Political Speech [J]. *Journal of Information Technology & Politics*, 2008, 5(1):33-48.
- [33] Ceron A, Curini L, Iacus S M, et al. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France [J]. *New Media & Society*, 2014, 16(2):340-358.
- [34] Gerber E R, Lewis J B. Beyond the Median: Voter Preferences, District Heterogeneity, and Political Representation [J]. *Journal of Political Economy*, 2004, 112(6):1364-1383.
- [35] Choy M, Cheong M L F, Ma N L, et al. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction [EB/OL]. [2016-09-20]. http://ink.library.smu.edu.sg/sis_research/1436.

- [36] O'Connor B, Balasubramanyan R, Routledge B R, et al. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series[C]. International Conference on Weblogs and Social Media, 2010.
- [37] Sudhahar S, Veltri G A, Cristianini N. Automated analysis of the US presidential elections using Big Data and network analysis[J]. Big Data & Society, 2015, 2(1):1-28.
- [38] Hobolt S B, Klemmensen R. Government Responsiveness and Political Competition in Comparative Perspective[J]. Comparative Political Studies, 2008, 41(3):309-337.
- [39] Laver M, Benoit K, Sauger N. Policy competition in the 2002 French legislative and presidential elections [J]. European Journal of Political Research, 2006, 45(4):667-697.
- [40] 杨慧, 杨建林. 融合 LDA 模型的政策文本量化分析——基于国际气候领域的实证[J]. 现代情报, 2016, 36(5):71-81.
- [41] 张惠, 王冰. 基于文本挖掘的政府公共价值测度与比较[J]. 安徽理工大学学报(社会科学版), 2015, 17(1):35-39.
- [42] 张永安, 闫瑾. 基于文本挖掘的科技成果转化政策内部结构关系与宏观布局研究[J]. 情报杂志, 2016, 35(2):44-49.
- [43] 胡嫣然. 基于文本挖掘的中国铁路运输企业财税支持政策研究[D]. 北京: 北京交通大学, 2016.
- [44] 程婷. 基于文本挖掘的中国环境保护政策文本量化研究[D]. 武汉: 华中科技大学, 2014.
- [45] Volkens A, Lehmann P, Matthie T, et al. The Manifesto Data Collection. Manifesto Project (MRG / CMP / MARPOR) [EB/OL]. [2016-10-20]. https://visuals.manifesto-project.wzb.eu/mpdb-shiny/cmp_dashboard_dataset/.
- [46] Corpus of political discourse in Cornell University [EB/OL]. [2016-10-20]. <http://www.cs.cornell.edu/home/llee/data/>.
- [47] MPQA Opinion Corpus [EB/OL]. [2016-10-20]. http://mpqa.cs.pitt.edu/corpora/political_debates/.
- [48] Eisenstein J, Xing E. The CMU 2008 Political Blog Corpus. 2010 [EB/OL]. [2016-10-20]. http://www.sailing.cs.cmu.edu/main/?page_id=713.
- [49] Ahrens, ed. Corpus of Political Speeches. Hong Kong Baptist University Library, Retrieved date of access (2015) [EB/OL]. [2016-10-20]. <http://digital.lib.hkbu.edu.hk/corpus/>.
- [50] Barbaresi A. German Political Speeches, Corpus and Visualization (2012) [EB/OL]. [2016-10-20]. <http://adrien.barbaresi.eu/corpora/speeches/>.
- [51] Young L, Soroka S. Affective News: The Automated Coding of Sentiment in Political Texts [J]. Political Communication, 2012, 29(29):205-231.
- [52] 陆宇杰, 许鑫, 郭金龙. 文本挖掘在人文社会科学研究中的典型应用述评[J]. 图书情报工作, 2012, 56(8):18-25.
- [53] Benoit K, Laver M, Mikhaylov S. Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions [J]. American Journal of Political Science, 2009, 53(2):495-513.

作者简介 裴雷,男,南京大学信息管理学院副教授,研究方向:信息政策分析与信息资源管理;孙建军,男,南京大学信息管理学院教授,研究方向:大数据分析与人文学科、网络信息计量与网络信息资源管理;周兆韬,女,南京大学信息管理学院研究生,研究方向:政策语料库分析(CAPS)。