·研究综述与前沿进展·

政策文本量化研究的综述与展望

郑新曼[1,2] 董 瑜[1,2*]
（1. 中国科学院文献情报中心，北京 100190；
2. 中国科学院大学经济与管理学院图书情报与档案管理系，北京 100049）

摘 要：［目的/意义］政策文本量化是一种新兴的跨学科研究方向。本文从文本数据与分析方法融合的角度，系统梳理了政策文本量化研究的最新进展，以指导政策文本量化分析实践。［方法/过程］基于文本量化的不同方式，将现有研究方法归纳为政策计量分析、内容分析法和效词分析法，分别总结了这些方法的特点、流程及典型应用。［结果/结论］政策文本量化研究近年来发展迅速，集中体现在文本数据类型拓展、多领域方法迁移与应用，其中效词分析法应用逐渐广泛；未来应关注政策文本数据源和语料库建设以及方法的误差研究。

关键词：政策文本；定量分析；文本量化；文本分析
DOI：10.3969/j.issn.1008-0821.2021.02.018
〔中图分类号〕G203 〔文献标识码〕A 〔文章编号〕1008-0821（2021）02-0168-10

Review on Quantitative Analysis of Political Texts

Zheng Xinman[1,2] Dong Yu[1,2*]
（1. National Science Library，Chinese Academy of Sciences，Beijing 100190，China；
2. Department of Library，Information and Archives Management，School of Economics and Management，University of Chinese Academy of Sciences，Beijing 100190，China）

Abstract: ［Purpose/Significance］The quantitative analysis of political texts is an emerging interdisciplinary research direction. From the perspective of the integration of text data and analysis methods，this paper systematically summarizes the latest developments in the quantitative research of political texts，aiming to support practical guidance for future research. ［Method/Process］Three main quantitative methods were summarized in the different ways of how political texts were converted into data，then concluded their process，characteristics and research tasks. ［Result/Conclusion］The quantitative research of political texts has developed rapidly in recent years，which mainly studied the expansion of data resources，application of multi-domain methods，and the tokenization method became increasingly popular. In the future，researchers can pay more attention to adopting more data sources for policy research，the construction of political texts corpus and validation of methods.

Key words: political texts；quantitative analysis；text as data；text analysis

政策文本是政策存在的物理载体，是政府政策行为的反映，也是记述政策意图和政策过程的客观凭证[1]，因此政策文本研究是追溯和观察政策过程的一个重要途径。近些年，随着政府信息公开化以及互联网的蓬勃发展，政策文本研究能够利用前人难以想象的数据，这些数据不仅体量大，而且种类丰富。同时，信息技术的快速发展，文本挖掘、数据分析等领域不断涌现的新方法和新工具，也极

收稿日期：2020-09-28
基金项目：中国科学院文献情报能力建设专项“科技知识服务大数据基础”（项目编号：Y9290002）。
作者简介：郑新曼（1996-），女，硕士研究生，研究方向：情报理论与方法。
通讯作者：董瑜（1971-），女，研究馆员，硕士生导师，研究方向：科技政策战略情报研究。

大地拓展了政策文本研究的范式。目前政策文本研究方法主要包括定性分析和定量分析，其中定性分析高度依赖研究者的实践经验和分析能力，强调从总体和宏观角度把握政策内容的复杂背景和思想结构。随着政策文本研究数据的增长，定性分析的人力成本激增。定量分析是将政策文本中有价值的信息转换成计算机可处理的结构化数据，进而利用数学模型进行分析，这极大降低了大样本量政策文本研究的人力成本，提高了结果的可复制性[2]，因此成为当前政策文本研究的趋势。

目前政策文本研究中所使用的定量方法大多迁移自其他学科，方法多样且近年来发展迅速，有必要及时进行梳理与总结。此外，现有的研究综述多从方法本身的角度进行梳理，如杨正总结了政策计量方法的概念、研究现状及未来趋势[3]，裴雷等从政策文本计算的角度对方法论、应用工具和典型研究议题进行了梳理[4]，但这些综述对研究方法与文本数据的融合分析关注不多，这一定程度上影响了对政策文本分析实践的指导。因此，本文从文本数据与分析方法融合的角度，系统梳理了国内外相关研究成果，归纳研究特点、流程及典型应用，以期为后续研究提供参考。

1 数据来源与处理

政策文本的定量分析融合了管理学、政治学、统计学、信息技术等多个学科。在文献检索过程中，为了尽可能覆盖相关文献，首先根据文献调研，对国内外文章中主题词的描述进行整理，然后结合数据库检索规则，确定检索主题词，如表 1 所示。

表 1 检索主题词集

序号	主 题 词
1	政策文本计算
2	政策文本量化
3	政策文本挖掘
4	政策文本定量分析
5	Text-as-data and Policy (Political)
6	Computerized (Computational) Text Analysis and Policy (Political)
7	Textual Quantitative Analysis and Policy (Political)
8	Textual Analysis and Policy (Political)
9	Quantitative Analysis and Policy (Political)

以 WoS 核心合集、CNKI 等数据库为文献来源，以表 1 的主题词进行检索，检索逻辑是“或含”。仅选择期刊文献。检索结果显示，WoS 核心合集第一篇相关文献发表于 1959 年，CNKI 的第一篇相关文献发表于 2007 年。以 2007 年为起始时间进行发文数量统计，可以看出，2016 年后国内外相关文献均有较为明显地增长。需要注意的是截至论文成稿日，2020 年文献尚未完全收录。

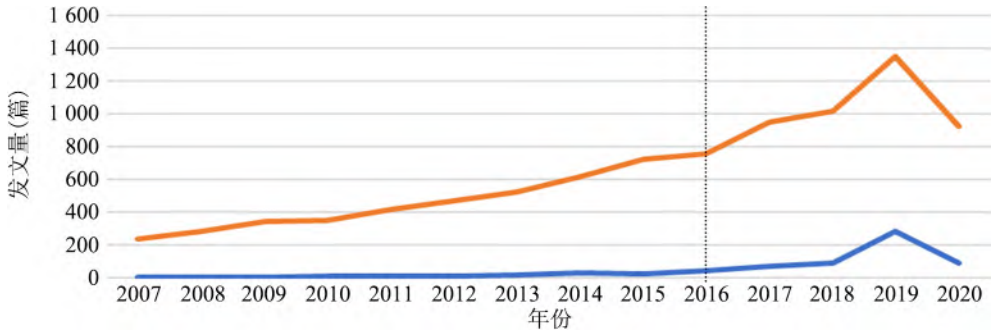

图 1 2007—2020 年 CNKI 和 WoS 相关文献年份分布图

因此，为了总结最新研究进展，本文以近 5 年(2016—2020 年) 发表的文献为切入点，经过主题筛选、文献溯源等方式，获得代表性文献 75 篇，其中中文文献 31 篇，英文文献 38 篇，中英文专著 6 部。基于这些文献，本文对政策文本定量研究的最新进展进行总结和分析。

2 相关概念辨析

2.1 政策文本量化的定义

在政策研究领域，国内外研究者对定量研究政策文本有多种表述，例如政策文献计量[5-6]、政策文本计算[4]、政策内容分析[7]等。政策文献计量主要是对政策文本的结构属性进行量化分析[5]，政

策文本计算强调运用计算机科学、语言学和政治学理论进行政策文本挖掘[4]，政策内容分析突出了对文本内容的定量和定性分析[7]。这些表述虽然侧重点有所不同，但核心思想一致，即区别于定性方法对少量政策文本的解读，展示出对大量政策文本分析方法的关注，探索的是一种新的研究范式。本文将上述表述统称为政策文本量化。

从内涵上看，政策文本量化是通过一系列的转换范式将非结构化政策文本转换成抽象化、特征化的计算机可处理的结构化数据[8]。从研究范围看，政策文本量化是政策研究、计算机科学等领域交叉融合产生的新领域。从研究方法看，政策文本量化是从计算机科学、文献计量学到政策科学的多领域方法迁移。

2.2 政策文本的范围

政策文本是指因政策活动而产生的记录文献，既包括国家、地方各级权力或行政机关以文件形式颁布的法律、法规、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，以及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[4]，如候选人在竞选期间的辩论和政策立场陈述，新闻中有关国际关系的报道等，这些文本都可看作是广义上的政策文本。目前常用于分析的政策文本类型包括立法、决策机关的公文、政党声明、听证会陈述、条约、政治科学论文以及媒体数据等。

可以看出，政策文本类型多样，结构复杂。在实际研究中，获取政策文本的方式主要有：①基于政府部门的主动公开和社交媒体平台的开放性，使用应用程序接口（API）或根据网页特征编写脚本批量获取；②基于研究机构和研究者的收集整理，建立专门的政策文本数据库，比如国内的综合性法律文件检索平台北大法宝、清华大学开发的政策文件数据库 iPolicy[9] 以及中国科协创新战略研究院建立的全球政策法规库等。

3 政策文本量化研究的典型方法

政策文本量化通常分为 4 个步骤[8]：①获取文本数据并进行预处理；②将政策文本表示成计算机可处理的数据形式，即文本量化；③处理数据获得量化结果；④检验方法，并对结果进行描述性分析或因果分析。

在文本量化步骤，目前有多种把政策文本转化成数据的方式，基于这些方式的特点，可以把当前政策文本量化方法分为三大类：政策计量分析、内容分析法和效词分析法，如图 2 所示。其中政策计量分析是通过定义并识别政策文本的结构要素，使其结构化后再进行分析；内容分析法是构建从政策语词到政策语义的分析单元框架，并以编码的方式将文本内容转变成含有语义信息的数字；效词分析法是使用文本表示模型将政策文本分解成含语义特征且可处理的基本单元，从而进行后续的分析。在实践中，文本量化方式的选择需要权衡计算成本和分析效果，即想要获得简单且有意义的数据以精简计算，但不要因此丢失过多的信息影响分析效果。

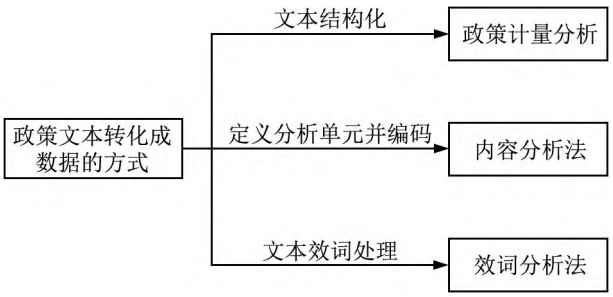

图 2 政策文本量化的 3 类典型方法

在数据处理步骤，模型选择是关键，其有效性直接决定了结果的可靠性。相比其他领域，政策研究对分析结果的可信度要求更高。在结果分析步骤，为了提高易读性，研究者通常利用可视化工具对结果进行可视化。

3.1 政策计量分析

政策计量分析（Policiometrics）是一种分析政策文献体系和结构属性的方法，由文献计量学、统计学、数学等学科有机结合产生，继承并迁移了文献计量学中的三大定律（“洛特卡定律”“布拉德福定律”与“齐夫定律”）。其基本流程主要包括收集政策文本，并根据文本的结构要素将政策文本结构化；借助文献计量学或社会网络分析相关方法进行分析。

3.1.1 政策文本的结构要素

政策文本具有与科技文献类似的基本结构要素，如发文机关、发布日期等，目前政策计量分析常用的结构要素及计量方法[6,10]，如表 2 所示。

表 1 政策文献与论文的结构要素对比及常用计量方法

政策文献结构要素	论文结构要素	常用计量方法
发文机关	作者、机构	合作网络分析、共引分析、共被引分析
主题分类	学科分类	共词分析、网络分析
主题词（人工标引）	关键词	频次分析、网络分析
文　　种	载　体	频次分析、网络分析
发布日期	发文日期	频次分析
参考的政策文件（人工标引）	参考文献	共引分析、共被引分析

与论文相比，政策文本的结构要素在含义、表征等方面有着自身的特殊性，如政策文献缺乏关键词和参考文献等结构要素。因此研究者常通过人工整理政策的标题和正文获取相关数据。张会平等[11]运用词频分析软件和人工方式提取从政府门户网站收集到的政策文本的关键词。苏竣等[12]整理制定了由 18 类 586 个术语构成的检索词表。

3.1.2 主要的应用

政策计量分析常基于政策文本结构要素的统计来分析政策的分布特征，如时间分布、文种分布、主题词词频分布等，可展现政策文本中隐含的关系网络，如颁布机构的合作关系，政策扩散及主题变迁等。黄萃等[13]对4 707份科技创新政策文本的主题词进行词频统计、共词和聚类分析，揭示出我国不同历史时期科技创新政策主题及其演进的阶段特征。李燕萍等[14]从发布时间、文种、颁发部门及关键词等 4 个方面对我国 488 份科技人才政策文本进行量化，并结合共词网络梳理了我国科技人才政策的整体状况、发展过程及趋势。陈慧茹[15]集成扎根理论、词频分析等，构建了基于政策属性与关键词权重的共词网络模型。张会平等[11]采用 CiteSpace 5.0 分析政策文本的时间分布、关键词共现网络、高频关键词及其共现关系。

3.1.3 存在的问题与对策

政策计量分析融合了文献计量学和社会网络分析法，有助于宏观层面的政策分析，如政策主体的合作模式、政策体系的结构与演进等分析，但也存在较为突出的问题。首先，更为深层的语义角度的分析仍依靠人工解读，同时一些低频重要词汇易被忽视。其次，现有传统指标无法满足复杂的政策研究需求，因此如何构建有效的分析指标是政策计量分析的一个重要方向。一些学者也在积极探索政策计量指标的设计，如张剑等[16]基于政策参照网络和关键词时序分析图谱，设计了强度、广度、速度与方向 4 个维度作为政策扩散的测量指标；刘建华[17]从直接语义关系、直接共现关系、间接共现关联、关联路径衰减指数 4 个维度构建了科技政策实体关联的多指标模型，并结合时间属性揭示了政策演化路径。此外，政策计量分析及应用较为依赖结构化语料库，但现有政策文本语料库在数量、质量以及开放性等方面尚无法完全满足需求，未来应关注政策文本结构化语料库的建设。

3.2 内容分析法

内容分析法（Content Analysis）是当前政策文本研究中常用的分析方法，通过定义能反映政策语义与语词之间映射关系的分析单元进行政策概念的识别和处理，并构建从分析单元到数值的编码标准与从政策文本到政策语义的政策分析框架[18]。具体研究过程包括 4 个步骤：①提出研究问题并抽取政策文献样本；②确定分析单元与编码标准；③对文本内容进行编码并进行百分比、平均值、相关分析、回归分析等统计操作；④解释并检验。

3.2.1 定义分析单元并编码

分析单元是内容分析中最重要、可结构化的元素，能够反映政策语义与政策文本内容之间的关系，可以是单词、符号、主题、以及意义独立的词组、句子或段落等[18]。在政策研究中，通常基于政策工具理论作为定义分析单元的理论依据，如典型的 Rothwell R 等的政策工具分类法[19]。编码是将政策文本中的分析单元转变为数值数据的过程，其关键在于编码标准及编码的可信度[20]。

目前常用的编码方式有人工编码和计算机辅助编码。人工编码包括编码标准构建、编码员培训和编码员间编码可靠性评估等要素[20]。有研究证实

提早确立编码标准有利于辨清和查找文本内容中固有的语义问题[21]，由于当前大多数编码方案是通过阅读文本归纳所得，因此为了确保内容分析法有效，在对政策文本进行编码前，往往需要邀请专家对编码标准进行修订。此外，由于人工编码依赖于人工对文本的理解，因此编码初期需要测度编码员对内容编码的一致性，即信度检验，通常认为 Kappa 系数达到 0.8 以上时编码可靠[22]。

随着计算机技术的进展，计算机辅助编码蓬勃发展。CAQDAS、Code-A-text，Ethnograph、MAX-QDA、Nvivo、QDA Miner、Symphony Content Analysis、ROST Content Mining（ROST CM）、DICTION、AtlasTi 和 ALCESTE 等文本分析工具的出现，帮助了编码人员对大样本文本内容进行编码[20,23-24]。有研究者对比了人工编码和工具编码的结果，发现使用计算机辅助编码一定程度提高了编码效率，降低了编码成本[25]，但仍要注意信度检验。

3.2.2 主要的应用

当前许多国内学者采用内容分析法进行政策发展演化研究，从政策工具[26-27]、政策主体[28]、政策作用场域[25]等方面分析并总结了某领域政策演化的阶段性特征及路径趋势[29-31]。黄新平等[25]对我国 72 份科技金融发展政策进行编码分类，总结出其政策工具体系、结构及运行特征。黄如花等[2]对我国政府数据开放共享政策文本进行编码，并基于编码结果进行描述性统计和分析。Huang C 等[9]从“政策目标—政策工具”角度对中国核能政策进行编码，通过计算“政策目标—政策工具”网络节点的特征向量中心性，确定了不同时期的主要“政策目标”和主要“政策工具”，梳理了我国核能政策的演变过程。裴雷等[32]对我国智慧城市政策文本进行编码和扎根统计，分析了我国智慧城市建设的现状。程瑨[33]综合利用内容分析法和计量分析法，探究了政策转移与政策协同的关系。

3.2.3 目前的局限与发展方向

内容分析法从语义的角度对政策文本进行编码，并使用统计学指标、PMC 指数模型[34]等对编码结果进行计算分析，具有可操作性强、适用性广等特点。内容分析法在具体应用方面存在一些需要注意的地方，包括清楚地说明抽样依据，分析框架设计的合理性，编码方案等。此外，随着样本数量和分析视角的增加，内容分析法的人力成本和使用难度将成倍增长，如何提高内容分析法的效率是亟待解决的问题，其中立足政策文本特征制定通用的分析单元体系和框架，基于当前通用的文本分析工具研制专用于政策文本分析的工具等将是有效的解决途径。

3.3 效词分析法

效词分析（Tokenization）源于自然语言处理（NLP），是指通过文本表示模型将文本分解成可处理的基本单元[35]。该方法使用文本表示模型表征政策文本中有意义的内容，旨在最大程度实现自动化文本分析，这也是该方法与前两类方法的最大区别。在政策研究领域，国外学者将此类方法统称为“Text as Data”[36]，以区别于自然语言处理、文本挖掘等较为通用的表示。该方法集成了机器学习、自然语言处理、文本挖掘等技术[23,37-39]，可以从语义角度量化和分析大型文本集，为进行大样本量政策文本的深入分析提供了机会。如 Haeder S F 等[38]利用 Heckman 选择模型和自动内容分析软件 WCopyfind 分析了美国管理和预算办公室（OMB）大量政策法规的变化。

3.3.1 分析流程

效词分析法的流程由 Grimmer J 等[36]最早提出，通常包含以下步骤：①获取文本，并对文本进行效词处理；②根据研究问题选择合适的算法进行计算和分析；③对建立的模型进行评估并验证；④结合实际问题对结果进行实质性解析[40]。

3.3.2 文本效词处理

根据语义表达粒度不同，常将文本表示分为词语级、句子级和篇章级，其中词语是语义最细粒度的表达。结合政策文本研究对语义的关注，效词分析法通常基于词袋模型表示文本，如 TF-IDF 算法；为了提高语义表示的精度，研究者也使用文本分布式表示方法，如 Skip-Gram 模型等。

1）词袋模型

词袋模型（Bag of Words Model）是常用的文本表示方法。其特点在于不考虑词语在文档中出现的顺序，将文档表示成一系列不同词语的组合，即所谓的单词袋，并计算文本中出现的不同单词的频

率。Linguistic Inquiry Word Count（LIWC）是基于词袋模型构建的自动词分析工具[41]，有研究者使用该工具对政策文本进行量化，把 LIWC 生成的单词作为语言变量进行统计学角度的研究[42]。基于词袋模型的常见工具还包括 WordScores[7] 和 Word-Fish[43]，这两个工具常用于从政策文本中提取观点意图[44]。WordScores 是有监督学习模型，依赖于带有标签的文档样本，例如有专家注释的政党宣言，以带标签文档中单词出现的相对概率作为每个单词的分数，并将此分数视为相应意图的指标，之后将分数应用于新文档中找到的单词，以此估计新文档的政治立场类型。Daigneault P M 等[45] 使用 WordScores 定量分析半结构化访谈出版物，研究表明 Wordscores 在补充定性分析方面具有较大潜力。WordFish 是无监督学习模型，所需的唯一输入是列出了所有文档中每个单词频率的单词频率矩阵，因此避免了 WordScores 对专家注释的依赖和某些单词语义受表达习惯和语义环境影响等问题。

词袋模型考虑到了用单词来映射文本语义，但未考虑词法和语序的问题，如仅关注一个单词在文档中是否出现和其出现次数，而忽略其上下文关系，这往往会丢失一部分文本的语义。从理论上讲，WordScores 和 WordFish 都可以扩展为一个以上单词的序列，但这会增加计算成本[46]。有学者针对此问题进行了探索，如 Alschner W 等[47] 使用由文本中连续 5 个字符组成的“词”的词袋模型[48] 来表示双边投资条约（BIT）文本。总体而言，词袋模型能从语义角度表示文本且易于解释，但其依赖单词词频，在实践中还需考虑单词和上下文的关系以及处理高维度变量时权衡计算成本等问题。

2）分布式表示

分布式表示（Distributed Representation）通常也被称为基于神经网络的分布表示、词嵌入或词向量（Word Embedding）。相比词袋模型离散、高维且稀疏的表示，分布式表示将词表示成一个低维且连续的稠密向量。Word2vec 是常用的工具，首先输入文本语料作为训练集，根据训练文本数据构建词汇表，并学习单词的向量表示，最后生成低维连续的实数向量并输出。Rodman E[49] 使用 Word2vec 对单词含义随时间变化的过程进行分析，显示出词嵌入模型在政策研究领域的应用潜力。Rheault L 等[46] 介绍了词嵌入模型在分析议会演讲文本方面的应用；Jentsch C 等[50] 为了研究政党立场，提出了一种词典可随时间变化的新模型。词向量认为相同上下文语境的词具有相似的含义，因而能够通过共现来发现和表示单词之间的关系，在一定程度上解决了仅依赖单词词频方法导致的语义不足问题。但是，词向量的使用效果非常依赖训练用的文本语料库，语料越多效果越好。

3.3.3 效词分析

通过文本挖掘工具或方法对文本效词处理后[51]，需要根据研究问题选择合适的算法进行计算和分析，如主题模型[52-53]、文本相似度算法[54]、循环神经网络[55] 等。Hollibaugh G E 等[40] 使用结构主题模型（Structural Topic Model）将现实事件与主题变化相对应，以分析政策文本中的主题变化。Alschner W 等[47] 使用 Jaccard Distance 来计算全球主要国家的双边投资条约的文本相似度，研究追溯了全球主要国家投资条约文本中的一致性和创新性。杨锐等[56] 通过高频词识别、共词分析及关键词聚类等方法探索了不同阶段科研诚信政策的主题演变。

3.3.4 主要的应用

效词分析法目前在国内外应用逐渐广泛，不仅促进了政策文本研究方法的创新，而且也为定量分析政策文本提供了新的视角。首先，该方法可用来深度挖掘大样本量或时间跨度较长的政策文本。盛东方等[57] 使用 LDA 主题模型分析了 401 份文本，以研究突发公共事件下中小企业扶持政策的供需匹配问题。张宝建等[58] 采用 K-means 聚类算法分析了我国 1996—2017 年 57 份国家科技创新政策典型文本，揭示出不同政策在科技创新发展各个阶段表现出的差异性和失配特征。其次，效词分析法常用以分析政党立场、政治倾向等隐含知识[59-60]。Windsor L 等[42] 将领导人发言中的单词变成数据进行研究，提供了对某些政治现象的解释依据。Shaffer R[61] 利用国会会议笔录的原始数据集来测量个人关注度的多样性。Rheault L 等[46] 通过对议会演讲文本的分析，估算了政党立场。利用政策文本研究政治主体的政治倾向，一定程度弥补了传统

访谈法样本小，受时间、人力限制[50]以及访谈过程中访谈对象易受影响等不足。

此外，效词分析法能广泛利用非传统政策文本分析公共政策问题，如社交媒体数据。Meng Q 等[62]使用社交网络分析工具 PKUVIS 对政府在微博上发布的消息进行深入分析，探讨了政府在社交媒体上如何治理突发公共事件。Barberá P 等[63]分析了数百万 Twitter 用户的社交媒体数据用以估算用户的政治倾向。Chang W H 等[64]通过分析大量异构社交媒体数据，研究了政治立场和政治策略等问题。Casas A 等[65]分析了上万条美国共和党人的推文（Twitter）发现，强大的政党形象可以帮助候选人保持或获得多数控制权。Miller C[66]通过对来自澳大利亚两个最盛行的反伊斯兰团体在社交媒体上的所有公开帖子进行分析，调查了这些团体的关注点。

3.3.5 优势与不足

效词分析法集成了自然语言处理中的文本表示模型以及文本挖掘中的多项技术，能够借助计算机快速处理大样本量的政策文本，人力成本低，结果可复制性强，一定程度弥补了政策计量分析深度不够和内容分析方法无法快速分析大样本文本数据的不足。但该方法仍有需要改进的地方，如为了获得较好的分析结果，易出现模型过拟合等问题[67]；由于不同领域政策研究问题的侧重点与分析精度需求不同，加之政策文本类型多样，仅通过已有模型或单一数据源进行政策研究也是不够的。此外，相较前两类方法，效词分析法存在一定的技术门槛，虽然目前有不少开源工具包可使用[68-69]，但在实践中仍要求具备一定的技术能力。因此，效词分析法未来需结合政策研究实践中的真实问题，不断设计和开发符合政策研究特点的分析框架、方法与工具等。

4 总结、讨论与展望

4.1 总结与讨论

政策文本量化作为一种新兴的跨学科研究方向，建立在多学科研究知识和技术专长之上。近年来研究人员在方法融合、数据拓展、实践应用等方面进行了积极的探索和研究，取得了一定的研究成果。在研究方法上，不断融合机器学习、自然语言处理、文本挖掘等新技术和新工具，逐步向大规模政策文本分析发展；在工具方面，关注并探索政策文本专用分析工具的研制，如政策分析专有词表；在分析深度上，从基于政策文本外部结构属性的分析逐步深化至对文本内部语义特征的挖掘；在数据类型上，积极推动专业语料库的建设和跨语料库的分析，如多语种分析、视频、音频等多语料分析等；在应用方面，从传统的政策演化、协同、扩散等研究，拓展至政策认同、政治倾向、政治策略、政党竞争与合作以及选情预测等研究。

需要指出的是，在政策文本量化研究中，国内外学者在研究方法、关注点等方面存在较为明显的差异。从方法看，国内常使用政策文本计量和内容分析法；国外倾向于效词分析等新方法的探索与改进。从研究问题看，国内关注领域政策发展演化的路径，以及政策效力等问题；国外多进行比较政治学、政治倾向判断等研究。从数据类型看，国外学者所分析的文本数据种类丰富，形式多样；国内研究多以官方政策文件为主，近年来开始关注社交媒体数据分析。

总体来看，相较国外学者丰富多样的政策文本量化研究，国内学者大多集中在借助成熟工具或单一方法进行不同领域政策文本的分析，呈现应用研究有余、深层次创新不足的态势，出现这种现象的原因在于国内相关学科研究人员之间缺乏有效地合作和交流。政策文本量化研究的生命力在于应对真实的决策需求，需要研究者兼具政策研究的问题意识和扎实的技术分析能力。此外，大数据时代的决策环境不仅主张科学性的研究方式，还要实现兼具一定深度和广度的实时分析。因此，未来应鼓励我国政策研究人员、领域专家、技术人员积极进行跨界合作，充分发挥多方知识与经验，立足我国决策需求，开展以政策研究中实际问题或需求反哺方法创新等深层次研究，这对改善当前国内研究的整体态势有重要的意义。

4.2 未来展望

对于政策研究而言，大规模文本量化分析既是机遇也是挑战。其具备的快速处理、细粒度分析、操作透明可重现等特点能够满足政策研究的实际需求，同时也极大地促进了政策研究方法的创新，更

拓展了深层次的政策内涵发现，为政策研究提供了新的视角。随着数据类型增多、体量增长以及研究问题的深入，实践中还有许多问题尚待解决：能否集成多种数据资源进行政策研究？如何进行结构化语料库的建设和维护？如何在保证方法使用效果的同时，最简化处理过程？如何提高方法的可操作性、适用性？基于此，本文认为政策文本量化研究还需关注以下几个方面：

1）积极拓展政策文本分析的数据源

政策文本是政策文本量化的数据源，法律条款、政党宣言、政府信息公开年度报告等政府公开数据构成了政策量化研究不断发展的数据基础。目前政府公开数据的质量和数量还有待提升[70]，因此积极推动政府数据开放[71]是拓展政策文本分析数据源的途径之一。此外，还应关注同样具有重要研究价值的非文本形式的数据，如政治会议，政治家电视讲话和访谈节目等音频、视频。已有研究展示了政治演讲音频生成政策文本的潜力[72]。未来可以积极拓展多种形式的数据源，并将其及时转录、存储，以推动政策文本量化研究的广度。

2）构建公开规范的结构化政策文本语料库

大型文本语料库的发展有利于开展政策文本量化研究[73]。尽管大部分机构会有专门的网站来颁布政策文件或发表相关的信息，但研究者在收集特定领域政策文件时仍需要付出大量的时间和精力。构建公开、不加密、可访问、规范的政策文本语料库能降低政策文本量化分析的成本；语料库建设的规范包括数据的结构化格式等，更重要的是数据构建的过程透明、规范、可重复[74]。作为政策文本量化研究的基础架构，公开、规范的结构化政策文本语料库不仅为当下政策文本研究带来极大的便利，也将促进相关学科的深度融合，加速实现自动高效分析政策文本的目标。

3）了解方法局限，重视误差研究

在政策文本量化的数据处理阶段，研究人员目前使用的方法多种多样，但本质上都是统计模型。因此可借助当前常用的数据分析工具进行分析，如 R 的 Topicmodels 包和 Stm 包[75]，Python 的 Gensim、Scikit-learn 和 Stm 包[40]等。但需要注意的是，将非结构化文本转化成结构化数据必然会使部分文本信息被遗漏，如常见的保留高频词而忽略低频词的操作。由于方法上存在局限性，因此需要对量化分析中的度量标准和量化结果进行严格检验以控制误差，如对内容有效性、外部有效性和预测有效性等相关标准的有效性进行测试[23]。

4）基于方法的不确定性制定实践手册

政策文本量化提供了广泛的工具来解决各种不同类型的研究问题，但任何一种方法的效果都取决于具体的研究问题。识别共性问题可以使研究人员共享解决问题的方法，对方法的思考反过来又能够帮助更好地解决研究问题。目前国内外学者积极开展新方法的探索，主要是将文本挖掘、机器学习等领域的方法进行迁移。但方法的迁移往往较为复杂，为特定应用选择最佳方法时，需要进行仔细思考和推理。因此，进一步夯实理论基础，制定高效计算框架，公开详细的操作步骤，并制成实践手册或准则将有助于方法的验证或改进。

参 考 文 献

[1] 李钢. 公共政策内容分析方法：理论与应用 [M]. 重庆：重庆大学出版社，2007.

[2] Denny M J，Spirling A. Text Preprocessing for Unsupervised Learning：Why it Matters，When it Misleads，and What to Do About It [J]. Political Analysis，2018，26 (2)：168-89.

[3] 杨正. 政策计量的应用：概念界限、取向与趋向 [J]. 情报杂志，2019，38 (4)：60-5，51.

[4] 裴雷，孙建军，周兆韬. 政策文本计算：一种新的政策文本解读方式 [J]. 图书与情报，2016，(6)：47-55.

[5] 苏竣. 公共科技政策导论 [M]. 北京：科学出版社，2014.

[6] 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，2015，12 (2)：138-44，59.

[7] Laver M，Benoit K，Garry J. Extracting Policy Positions from Political Texts Using Words as Data [J]. American Political Science Review，2003，97 (2)：311-31.

[8] Gentzkow M，Kelly B，Taddy M. Text as Data [J]. Journal of Economic Literature，2019，57 (3)：535-74.

[9] Huang C，Yang C，Su J. Policy Change Analysis Based on "Policy Target-Policy Instrument" Patterns：A Case Study of China's Nuclear Energy Policy [J]. Scientometrics，2018，117 (2)：1081-114.

[10] 宋伟，夏辉. 地方政府人工智能产业政策文本量化研究 [J]. 科技管理研究，2019，39 (10)：192-9.

[11] 张会平，郭宁，汤玺楷. 推进逻辑与未来进路：我国政务大

数据政策的文本分析 [J]. 情报杂志，2018，37 (3)：152-7，92.

[12] 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.

[13] 黄萃，赵培强，李江. 基于共词分析的中国科技创新政策变迁量化分析 [J]. 中国行政管理，2015，(9)：115-22.

[14] 李燕萍，刘金璐，洪江鹏，等. 我国改革开放 40 年来科技人才政策演变、趋势与展望——基于共词分析法 [J]. 科技进步与对策，2019，36 (10)：108-17.

[15] 陈慧茹. 基于扎根理论的国家自主创新示范区科技创新政策共词网络研究 [D]. 合肥：中国科学技术大学，2017.

[16] 张剑，黄萃，叶选挺，等. 中国公共政策扩散的文献量化研究——以科技成果转化政策为例 [J]. 中国软科学，2016，(2)：145-55.

[17] 刘建华. 基于实体及实体间关系的科技政策演化揭示方法研究 [D]. 北京：中国科学院大学，2017.

[18] 邱均平，邹菲. 关于内容分析法的研究 [J]. 中国图书馆学报，2004，30 (2)：12-7.

[19] Rothwell R，Zegveld W. Reindusdalization and Technology [J]. Logman Group Limited，1985：83-104.

[20] Neuendorf K A，Kumar A. Content Analysis [J]. The International Encyclopedia of Political Communication，2015：1-10.

[21] Berelson B. Content Analysis in Communication Research [J]. 1952.

[22] 黄如花，温芳芳. 我国政府数据开放共享的政策框架与内容：国家层面政策文本的内容分析 [J]. 图书情报工作，2017，61 (20)：12-25.

[23] Pandey S，Pandey S K，Miller L. Measuring Innovativeness of Public Organizations：Using Natural Language Processing Techniques in Computer-Aided Textual Analysis [J]. International Public Management Journal，2017，20 (1)：78-107.

[24] 李燕萍，吴绍棠，郜斐，等. 改革开放以来我国科研经费管理政策的变迁、评介与走向——基于政策文本的内容分析 [J]. 科学学研究，2009，27 (10)：1441-7，53.

[25] 黄新平，黄萃，苏竣. 基于政策工具的我国科技金融发展政策文本量化研究 [J]. 情报杂志，2020，39 (1)：130-7.

[26] 谭春辉，谢荣，刘倩. 政策工具视角下的我国科技评估政策文本量化研究 [J]. 情报杂志，2020，39 (10)：181-90.

[27] 李健，荣幸. “放管服”改革背景下社会组织发展的政策工具选择——基于 2004—2016 年省级政策文本的量化分析 [J]. 国家行政学院学报，2017，(4)：73-8，146-7.

[28] 许斌丰. 技术创新链视角下长三角三省一市区域创新系统协同研究 [D]. 合肥：中国科学技术大学，2018.

[29] 杨艳，郭俊华，余晓燕. 政策工具视角下的上海市人才政策协同研究 [J]. 中国科技论坛，2018，(4)：148-56.

[30] 刘红波，林彬. 中国人工智能发展的价值取向、议题建构与路径选择——基于政策文本的量化研究 [J]. 电子政务，2018，(11)：47-58.

[31] 范利君. 2006—2014 年间中国创新政策协同演变研究 [D]. 成都：电子科技大学，2016.

[32] 裴雷，周兆韬，孙建军. 政策计量视角的中国智慧城市建设实践与应用 [J]. 图书与情报，2016，(6)：41-6.

[33] 程瑨. 国家自主创新示范区政策转移量化与协同研究 [D]. 合肥：中国科学技术大学，2017.

[34] 张永安，郗海拓. 国务院创新政策量化评价——基于 PMC 指数模型 [J]. 科技进步与对策，2017，34 (17)：127-36.

[35] Webster J J，Kit C. Tokenization As The Initial Phase In NLP，F，1992 [C].

[36] 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)：267-97.

[37] Marvel J D，Mcgrath R J. Congress as Manager：Oversight Hearings and Agency Morale [J]. Journal of Public Policy，2016，36 (3)：489-520.

[38] Haeder S F，Yackee S W. Influence and the Administrative Process：Lobbying the U. S. President's Office of Management and Budget [J]. American Political ence Review，2015，109 (3)：507-22.

[39] Baker S R，Bloom N，Davis S J. Measuring Economic Policy Uncertainty [J]. The Quarterly Journal of Economics，2016，131 (4)：1593-636.

[40] Hollibaugh G E，J R. The Use of Text as Data Methods in Public Administration：A Review and an Application to Agency Priorities [J]. Journal of Public Administration Research and Theory，2019，29 (3)：474-90.

[41] Tausczik Y R，Pennebaker J W. The Psychological Meaning of Words：LIWC and Computerized Text Analysis Methods [J]. Journal of Language and Social Psychology，2010，29 (1)：24-54.

[42] Windsor L，Dowell N，Windsor A，et al. Leader Language and Political Survival Strategies [J]. International Interactions，2018，44 (2)：321-36.

[43] Slapin J B，Proksch S O. A Scaling Model for Estimating Time-series Party Positions from Texts [J]. American Journal of Political Science，2008，52 (3)：705-22.

[44] 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).

[45] Daigneault P M，Duval D，Imbeau L M. Supervised Scaling of Semi-structured Interview Transcripts to Characterize the Ideology of a Social Policy Reform [J]. Quality & Quantity，2018，52 (5)：2151-62.

[46] Rheault L，Cochrane C. Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora [J]. Political Analysis，2020，28 (1)：112-33.

[47] Alschner W，Skougarevskiy D. Mapping the Universe of International Investment Agreements [J]. Journal of International Economic Law，2016，19 (3)：561-88.

[48] Spirling A. U. S. Treaty Making with American Indians：Institutional Change and Relative Power，1784 - 1911 [J]. American Journal of Political Science，2012，56 (1)：84-97.

[49] Rodman E. A Timely Intervention：Tracking the Changing Meanings of Political Concepts with Word Vectors [J]. Political Analysis，2020，28 (1)：87-111.

[50] Jentsch C，Lee E R，Mammen E. Time-dependent Poisson Reduced Rank Models for Political Text Data Analysis [J]. Computational Statistics & Data Analysis，2020，142.

[51] 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)：44-9.

[52] Blaydes L，Grimmer J，Mcqueen A. Mirrors for Princes and Sultans：Advice on the Art of Governance in the Medieval Christian and Islamic Worlds [J]. Journal of Politics，2018，80 (4)：1150-67.

[53] 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.

[54] 张涛，马海群，易扬. 文本相似度视角下我国大数据政策比较研究 [J]. 图书情报工作，2020，64 (12)：26-37.

[55] Alschner W，Skougarevskiy D. Can Robots Write Treaties? Using Recurrent Neural Networks to Draft International Investment Agreements [M]. Bex F，Villata S. Legal Knowledge and Information Systems，2016b：119-24.

[56] 杨锐，杨亮，李良强，等. 我国科研诚信政策特征及演化逻辑——基于文本挖掘法 [J]. 科技进步与对策，1-10.

[57] 盛东方，尹航. 基于政策文本计算的突发公共事件下中小企业扶持政策供需匹配研究——以新冠肺炎疫情为例 [J]. 现代情报，2020，40 (8)：10-9.

[58] 张宝建，李鹏利，陈劲，等. 国家科技创新政策的主题分析与演化过程——基于文本挖掘的视角 [J]. 科学学与科学技术管理，2019，40 (11)：15-31.

[59] Gentzkow M，Shapiro J M，Taddy M. Measuring Group Differences in High - Dimensional Choices：Method and Application to Congressional Speech [J]. Econometrica，2019，87 (4)：1307-40.

[60] Laver M，Garry J. Estimating Policy Positions from Political Texts [J]. American Journal of Political Science，2000，44 (3)：619-34.

[61] Shaffer R. Cognitive Load and Issue Engagement in Congressional Discourse [J]. Cognitive Systems Research，2017，44：89-99.

[62] Meng Q，Zhang N，Zhao X，et al. The Governance Strategies for Public Emergencies on Social Media and Their Effects：A Case Study Based on the Microblog Data [J]. Electronic Markets，2016，26 (1)：15-29.

[63] Barberá P，Jost J T，Nagler J，et al. Tweeting from Left to Right：Is Online Political Communication More Than an Echo Chamber? [J]. Psychological Science，2015，26 (10)：1531-42.

[64] Chang W H，Li J L，Lee C C，et al. Learning Semantic-Preserving Space Using User Profile and Multimodal Media Content from Political Social Network [M]. 2019 Ieee International Conference on Acoustics，Speech and Signal Processing，2019：3990-4.

[65] Casas A，Wilkerson J. A Delicate Balance：Party Branding During the 2013 Government Shutdown [J]. American Politics Research，2017，45 (5)：790-812.

[66] Miller C. Australia's Anti-Islam Right in Their Own Words. Text as Data Analysis of Social Media Content [J]. Australian Journal of Political Science，2017，52 (3)：383-401.

[67] Egami N，Fong C J，Grimmer J，et al. How to Make Causal Inferences Using Texts [J]. arXiv Preprint arXiv：180202163，2018，

[68] Wilkerson J，Casas A，Annual R. Large - Scale Computerized Text Analysis in Political Science：Opportunities and Challenges [M]. Annual Review of Political Science，2017，20：529-44.

[69] Benoit K，Watanabe K，Wang H，et al. Quanteda：An R Package for the Quantitative Analysis of Textual Data [J]. Journal of Open Source Software，2018，3 (30)：774.

[70] 段尧清，周密，尚婷. 我国政府信息公开态势及其调控策略研究——基于 2008—2018 年国务院部门政府信息公开年报分析 [J]. 现代情报，2020，40 (8)：121-8，77.

[71] 陈玲，段尧清. 我国政府开放数据政策的实施现状和特点研究：基于政府公报文本的量化分析 [J]. 情报学报，2020，39 (7)：698-709.

[72] Proksch S-O，Wratil C，Waeckerle J. Testing the Validity of Automatic Speech Recognition for Political Text Analysis [J]. Political Analysis，2019，27 (3)：339-59.

[73] Alschner W，Seiermann J，Skougarevskiy D. Text of Trade Agreements (ToTA) A Structured Corpus for the Text-as-Data Analysis of Preferential Trade Agreements [J]. Journal of Empirical Legal Studies，2018，15 (3)：648-66.

[74] Benoit K，Conway D，Lauderdale B E，et al. Crowd-sourced Text Analysis：Reproducible and Agile Production of Political Data [J]. American Political Science Review，2016，110 (2)：278-95.

[75] Roberts M E，Stewart B M，Tingley D. Stm：An R Package for Structural Topic Models [J]. Journal of Statistical Software，2019，91 (1)：1-40.

(责任编辑：郭沫含)

JOURNAL OF MODERN INFORMATION
现代情报
第41卷 第2期
二〇二一年二月

国内统一连续出版物号：CN 22-1182/G3　邮发代号：12-124　定价：28.00元
国际标准连续出版物号：ISSN 1008-0821　广告经营许可证2200004300033

ISSN 1008-0821
02
9 771008 082213