基于多层次主题模型的科技政策文本量化研究

韩 旭，杨 岩
（中国科学技术信息研究所，北京 100038）

摘 要：政策文本是政府行为的直接记录，对政策文本的量化分析研究不但有利于把握政策的整体走向，还能够为后续的政策延伸提供一定的参考。本文以 2001 年至今科技部发布的中央级别科技政策为研究对象，构建基于词和文档两个层次的主题模型，从整体的科技规划角度，对政策文本进行量化分析，分析我国近些年科技政策的发展，将我国科技发展的演进方式进行总结，为后续科技政策的制定和执行提供数据支撑。

关键词：科技政策；量化分析；主题挖掘；LDA 模型
中图分类号：G322.0 **文献标识码**：A DOI：10.3772/j.issn.1009-8623.2020.11.009

科技创新是引领国家发展的第一动力。政策是科技发展的基本框架，纵观全球科技发展趋势，世界各国均推出一系列政治举措来推动本国的科技发展。同时，美国政府近些年相继推出太空、生物、网络等多项科技战略，力图维持科技领先优势；英国启动“未来领导者研究基金计划”等多项政策，以此激励本国的科技创新及人才培养；德国在 2018 年推出“高技术战略 2025”，为德国在新能源、健康、环保等多个领域建立了科技创新长期目标；日韩两国同样相继出台了多项相关政策，以激励本土的科技创新能力。在全球科技创新快速发展的背景下，我国的科技创新也经历了一系列的相关演变。

政策文本是政府行为的直接记录，在规范化的行政文件系统中，正式政策文件在我国政治社会经济生活中有着重要的引领作用，而中央文件无疑具有最高的权威性和最广泛的指导意义。在科技领域，早在 2006 年的《国家中长期科学和技术发展规划纲要（2006—2020 年）》中，我国就已做出了创新驱动发展的战略部署；党的十八大以来，我国把科技创新摆在国家发展全局的核心位置，高度重视科技创新，围绕实施《国家创新驱动发展战略纲要》，加快推进以科技创新为核心的全面创新。中央的多项举措表明，我国已经把科技创新放在国家战略发展的关键位置。

政策法规原始文本作为一种特殊类型的文献，和期刊论文、专利文献等科技信息资源相比，具有较强的权威性、严肃性及约束力。对科技创新的相关政策进行细化分解和剖析，能够为科研院所、科技企业等单位带来更多方向性的指引。近年来，关于政策文本的量化分析方法成为科研人员的重要研究课题之一。随着数据挖掘及自然语言处理的发展，主题模型已逐渐成为文本类型数据的主要分析手段之一。构建“文本 – 主题”“主题 – 关键词”之间的关系，能够深入分析文本内部的隐藏含义。本文收集整理了自 2001 年以来科技部发布的我国科技相关政策，并进行基于词语和文档两个层次结构的主题模型构建，探索我国科技不同阶段政策的演进趋势、探索政策之间的相关关系和相互作用并

第一作者简介：韩旭（1991—），女，助理研究员，博士，主要研究方向为自然语言处理、政策文本分析。
项目来源：中国科学技术信息研究所创新研究基金青年项目“中国科技创新的政策文本量化分析方法研究”（QN2020–09）；中国科学技术信息研究所重点工作“科技创新大数据决策分析模型构建平台开发”（ZD2020–11）。
收稿日期：2020–09–16

对其进行多角度的分析，为政策的制定和执行提供一定的参考。

1　相关工作

学术界对政策法规的研究由来已久，研究对象涵盖政策法规的制定、执行及反馈等诸多方面，研究涉及了整个政策法规的生命周期。多年来，国外学者从各角度运用政策文本分析方法，取得了丰富的研究成果[1-5]。政策工具视角[6]将公共政策工具定义为政府的行为方式，以及通过某种途径用以调节政府行为的机制。大多数学者建立的政策分析模型均以政策工具为理论依据，提出针对某个主题的政策分析方法，对政策正文根据不同的内容逐条分类，制定政策的主题编码，人工进行政策的整理和信息标注，从不同角度对政策文本进行深度分析。黄萃等[7]依据政策工具视角，对我国高新技术产业税收开展了相关政策研究，并指出现存的问题及治理意见；周京艳等[8]利用政策文本法分析出我国大数据政策应适度提高供给面政策工具的使用；王宏起等[9]基于政策文本定量分析方法，重点对“双创”政策进行细粒度的三维分析；赵润娣[10]通过解析美国、英国以及澳大利亚的相关政策文本，探究构建开放政府数据政策内容框架的可行性；李凡等[11]从政策目标、政策工具、政策执行三个维度建立框架，在政策文本分析的基础上综合运用聚类分析和因子分析方法对金砖国家技术创新政策布局进行比较；黄菁[12]运用定量统计分析和多维尺度分析等方法对239项地方科技成果转化政策的类型、地域以及作用领域的分布情况进行研究；白彬等[13]对创业拉动就业政策进行文本分析研究，提出优化和完善该政策体系的路径和方法；汪涛等[14]基于内容分析法对中长期科技发展规划进行量化分析，从多种角度探讨政策群的协同状况；王立等[15]从新材料领域入手，对我国40年内科技政策体系的资源投入进行深入研究。以上的研究从多个方面和主题对政策文本进行量化，但大多采用政策工具视角及人工分析标注的形式，对政策文本进行分析及专家分析，分析手段较为简约，尚待开发与拓展。

政策文本大多数为非结构化文本，使用文本挖掘的方法，将其转化为半结构化数据，有助于对政策文本进行深层分析。然而政策本身具有政治性、高时效性、严谨性等特性，需要相关专家的先验知识，以及专业的研究者对其进行深入的研究挖掘，传统文本挖掘方法难以深入其中，获取到更多内涵。很多学者尝试将传统计量学运用在政策文本中，将政策文本看作文献的形式，进行政策之间的引用分析、关键词共现等分析，也取得了一定的成果。张文伟等[16]以纳米材料领域数据为基础，对比了隐狄利克雷分布（Latent Dirichlet Allocation, LDA）和BTM两种主题模型对于主题抽取的效果；杨慧等人[17]构建了基于融合的LDA模型的政策文本量化分析，对国际气候的政策情况进行总结和分析；赵杰等[18]针对京津冀协同发展问题，构建了概率主题模型，并重点从主题强度和主题相似度多个角度分析了主题演化趋势；杨奕等[19]针对共享单车议题，进行基于主题模型的公众反馈意见采纳研究。这些研究都为政策文本从技术角度上的量化提供了一些思路和方法。

本文与其他研究的区别在于：在主题选择上，不局限于某个具体方向，而尝试从宏观视角出发，关注科技部官方发布的全部科技政策；在方法选择上，尽量减少人工干预及主观分析，通过文本挖掘手段，构建无监督的政策主题模型；在时间跨度上，选取从2001年开始，近20年的科技政策进行建模分析，观测我国长时间跨度内的科技政策发展态势。本文通过提取政策文本中的关键问题及实施对象，将政策文本进行上下位关联，挖掘政策的主题路线及发展趋势，从而最终达到辅助决策的目的。

2　模型构建

本文从宏观视角出发，使用文本挖掘方法，对全主题的科技政策文本进行量化，减少人工标注及干预，客观分析我国科技政策发展趋势。本文首先建立科技政策文本量化框架，并依据框架步骤对政策文本进行深入分析。

快速了解一篇文本的内容主要基于主题，除基础的关键词外，分析文本的深层语义信息更能揭示一篇文本的核心内容。主题模型是一种无监督的机器学习算法，其核心是一种对文本中隐含的语义结构进行聚类的统计模型。常见的主题模型有LDA、非负矩阵分解、潜在语义分析等，其中最具

代表性的是 LDA。LDA 模型由 Blei[20] 于 2003 年提出，用来推测文档的主题分布。LDA 可以将文档集中的每篇文档主题以概率分布的形式给出，并给出每个主题的关键词信息，从而通过主题分布进行主题聚类。本文尝试将词表示融合 LDA 模型引入政策分析中，以科技政策为研究主体，结合政策文本的相应结构和特点，对科技政策进行基于内容的主题建模，以及对应政策引用关系的分析，并结合科技领域当前形势，对政策文本进行深入的分析和解读，分析验证文本主题挖掘的效果，最终起到辅助科技决策的作用。

一条完整的政策文本包含有政策发布时间、文号、发布机构、正文等多维特征，在政策的正文中，也会具备政策主体、作用面、具体实施方案等多角度的特征。政策文本是典型的长文本文档，语言风格较为正式客观，情感偏中性，且政策术语较多。另外，政策文本较少存在摘要结构，因此本文以政策文本的题目及正文全文内容作为文本解析主体。本文构架的政策主题模型框架如图 1 所示。其中第一部分为数据获取部分，通过数据爬虫等手段，获取科技部官网发布的自 2001 年起的科技政策，并进行数据的初步清洗；第二部分为数据预处理；第三部分为词语层面的主题分析；第四部分为文本全文的主题模型构建。整体模型基于 Python 语言模块构建。

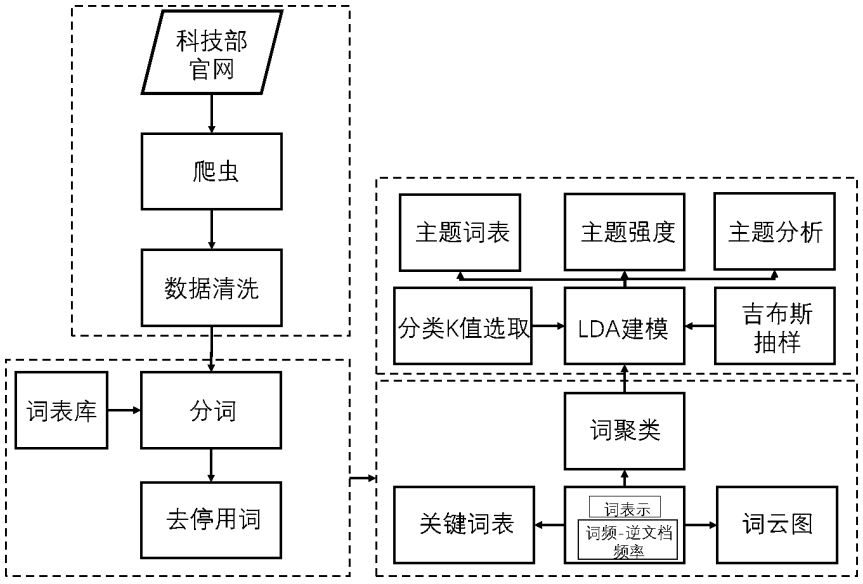

图 1　主题模型框架

2.1　数据来源

本文主要从中央层面的政策着手，仅以官方门户网站发布的政策为准，政策抓取的时间段为自 2001 年以来的科技部官网发布的政策文件，其中根据官网的政策体系分类，分为法律、行政法规、部门规章、规范性文件四个部分，构成我国科技领域政策文本语料库，用于挖掘科技政策文本的研究主题及演化趋势，在原始数据获取之后，进行去重、pdf 文件转文本等数据清洗操作，最终获得科技相关政策数据 1 784 条。

2.2　数据预处理

数据预处理包括对文本正文的分词、去停用词等操作。在分词部分，使用分词工具结合词表的方式。政策文本不同于传统文本，其具有一定的官方性和权威性，用词较为正式且专业，本文使用《政务文书档案专业词表》[21]，对政策文本中的词组进行预处理，并进行同义词替换，以此提高后续模型的效率。在去停用词部分，使用 jieba 的停用词表，结合哈工大停用词表、政策特征停用词表进行处理。另外，在统计过程中，单词频次小于 5 次的词语被删除，并删除“科技”“科技部”等对科技政策分析而言信息量较低的高频词汇。

2.3　基于词语的主题分析

本文首先以词为单位，对政策文本进行初步的主题分析。在文本预处理后，使用 TF-IDF 对政策文本的关键词进行提取，生成政策文本的关键词

表，并依据年份生成对应的关键词云图，以此分析科技政策逐年的主题变化趋势。

在政策文本中，两个关键词之间的共现能体现一定的主题关系。传统的词共现生成模型一般只考虑词频和共现关系，未考虑词语的语序关系及词语之间的内部含义。本文从词的深层含义出发，结合词的分布式表示方法，对同义词进行合并，生成词语的主题聚类图，以直观分析科技政策类文本的关键词主题演变趋势。同时，本文使用基于 CBOW 模型的词向量来对词语进行表示。该模型通过输入目标词的上下文词汇，来判断输出目标词的概率。设当前词为 x_i，当前词的向量表示为 w_i，上下文前后各取 n 个字构成，因此输入层向量数为 $2n$。其次在映射层中进行相关的运算，最后在输出层输出预测单词 x_i 出现的概率 $p(w_i|context\ w_i)$。本模型设置窗口大小 n 为 5，进行词表示的训练，并用于后续的词语级别主题分析中，最终生成对应的词聚类云图。

2.4　基于文档的 LDA 主题模型构建

LDA 是一种文档生成模型，其核心作用是将文档集中的每篇文档的主题以概率分布的形式给出。LDA 默认一篇文档具有多个主题，每个主题对应着不同的关键词。一篇文档的构造过程，首先是以一定的概率选择某一些主题，随后在主题下以一定概率选取某一些关键词，构成了整篇文章。而 LDA 模型的使用即为该过程的逆过程，根据现有的文本库，寻找对应的主题。

LDA 主题生成模型的基本步骤如下[20]：（1）按照先验概率 $p(d_i)$ 选择一篇文本 d_i；（2）从 Dirchlet 分布 α 中取样生成文本 d_i 的主题分布 θ_i；（3）从主题的多项式分布 θ_i 中取样生成文本 d_i 的第 j 个词的主题 $z_{i,j}$；（4）从 Dirichlet 分布 β 中取样生成主题 $z_{i,j}$ 对应的词语分布 $\varnothing_{zi,j}$，词语分布 $\varnothing_{zi,j}$ 由参数为 β 的 Dirichlet 分布生成；（5）从词语的多项式分布 $\varnothing_{zi,j}$ 中采样，最终生成词语 $w_{zi,j}$。在本文的模型试验中，LDA 参数估计模型为吉布斯采样法，超参数 α 和 β 分别设置为 50 和 0.2，迭代次数设置为 200。

在 LDA 构建过程中，其主题个数 K 为模型的关键参数，一般是结合先验知识，以及多次实验，得到 K 的最优值。在 LDA 模型中，一般使用困惑度（Perplexity）和 JS 散度（Jensen-Shannon divergence）来对模型的 K 值进行调参。困惑度是一种评估模型泛化能力的参数，其值越低，模型的泛化能力越强。困惑度值计算公式如下：

$$perplexity(D)=\exp\left[\frac{\sum_{d=1}^{D}\log_2 p(w_d)}{\sum_{d=1}^{D}N_d}\right] \quad (1)$$

其中 N_d 表示第 d 个文档的词汇数；D 为文档个数；$p(w_d)$ 为第 d 个文档中词汇的概率分布。

JS 散度公式如下：

$$avg_{sim}(T_i, T_j)=\frac{\sum_{i=1}^{k=1}\sum_{j=i=1}^{k}JS(T_i||T_j)}{K(K-1)/2} \quad (2)$$

其中 T 表示主题，$JS(T_i||T_j)$ 表示主体之间的散度。两个主题之间的差异性越大，证明主题识别效果越强。

在针对科技政策文本的主题分析中，LDA 模型初始 K 值设置为 10，并在滑动范围内进行微调，综合 P 值和 JS 值来确定最终的主题数，使得模型识别出的主题效果最优，本模型 K 的最终取值为 8。

3　实验结果及数据分析

3.1　科技政策文本初步分析

3.1.1　科技政策总量分析

首先，对数据进行整体的时间维度分析，根据政策的颁布时间，进行数量统计，结果如图 2 所示。

从图 2 中可以看出，2005 年之前的政策发布数据相对平稳，年均保持在 50 条左右。从 2006 年起，有比较明显的三个时间变化节点，分别是 2006 年、2011 年和 2017 年。结合我国的国家整体战略规划部属情况，2006 年和 2011 年分别是我国“十一五”和“十二五”规划的首年，国家对未来五年科技发展的规划性政策发布较为密集。而 2016 年是我国“十三五”规划的首年，同年我国又发布了《国家创新驱动发展战略纲要》，以此激励科技创新发展。在这一年，科技政策的出台呈现了滞后性，在 2017 年进入政策发布的高峰年份。通过对政策内容的具体分析，大部分 2017 年发布的新政策都提到了《“十三五”国家科技创新规划》《国家创新驱动发展战略纲要》两个重点政策文件。2016 年 5 月 30 日，习近平主席在全国科技创新大会、两院院士大会、中国科协第九次全国代表大会上讲话，提出“为建设世界科技强国而奋斗”，这

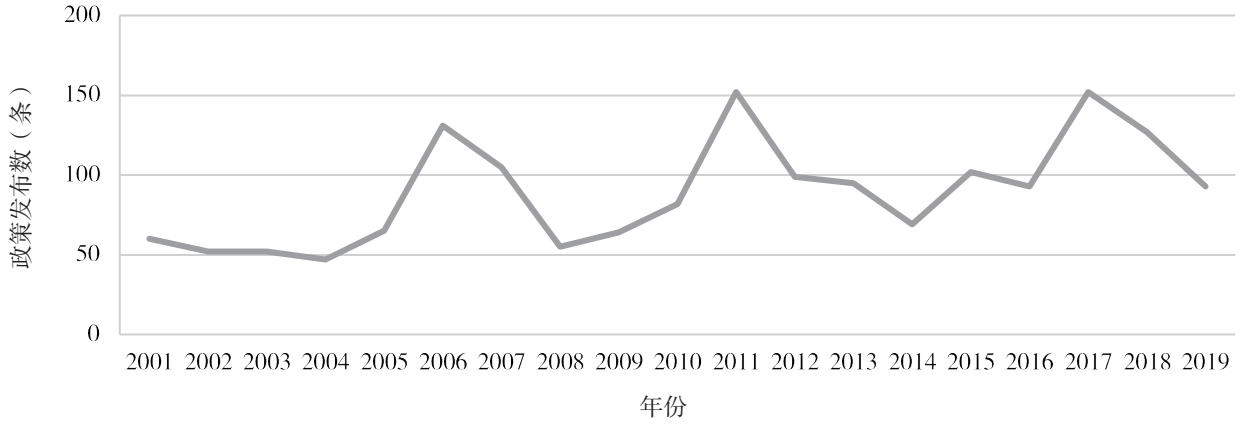

图 2　政策发布数随时间变化趋势

也间接影响了 2017 年科技政策数量的增长。通过政策随年份发布的变化趋势可见，科技政策的颁布与国情及科技相关的重要事件密不可分。

3.1.2　政策被引关系分析

传统的政策文本，尤其是规范性文件中，通常会在政策正文较靠前的部分提到起指导作用的相关政策，政策之间会存在“贯彻实施”“落实”“依据”等参照关系[22]，这里将这种提供参照关系的政策称为上位政策。在政策文本研究中，抽取政策文本相关联的上位政策，并对上位政策进行梳理，整理出自 2001 年起，全部科技政策的上位政策排名前 5 名的政策文本，如表 1 所示。

表 1　政策被引前 5 名

排名	政策标题	提出年份	被引次数
1	《国家中长期科学和技术发展规划纲要（2006—2020 年）》	2006	218
2	《关于深化中央财政科技计划（专项、基金等）管理改革的方案》	2014	67
3	《国家创新驱动发展战略纲要》	2016	47
4	《国家“十二五”科学和技术发展规划》	2011	42
5	《“十三五”国家科技创新规划》	2016	40

整理引用这 5 条政策的相关政策数量，并将其按时间变化进行统计，其统计结果如图 3 所示。可以看出大多数政策在被提出的第二年被引次数最多，从整体来看 2017 年是政策被引高峰，这与图 2 中整体政策数量变化趋势相符。被引次数居首位的是《国家中长期科学和技术发展规划纲要（2006—2020 年）》，自 2006 年颁布后，该政策被众多科技政策参照引用，显示出了极强的政策影响力。该纲要的战略布局横贯 15 年，是我国近些年科技战略部署的主要依据，该政策在发布当年即达到了被引量的高峰，并且在每个国家五年计划的时间节点均被频繁提及。

值得一提的是排在第 3 位的《国家创新驱动发展战略纲要》，该政策的提出时间为 2016 年，和其他政策相比提出时间较晚，但具有较高的被引次数，从该政策提出至今，平均每年有近 15 条政策引用了该政策，充分说明了该政策的重要性。另外，起到战略主导地位的国家“十二五”“十三五”规划也同样具有较高的被引次数，证明我国整体的科技规划是按照既定方向前进的。

3.2　政策文本主题分析

3.2.1　基于词语的主题分析

为统计我国的政策文本随时间变化的主题词趋势，并更加具体地探测我国科技创新政策文本的热点，本文提取了历年来的政策文本关键词，并进行了基于词表示和政务文书词表的同义词归并。从

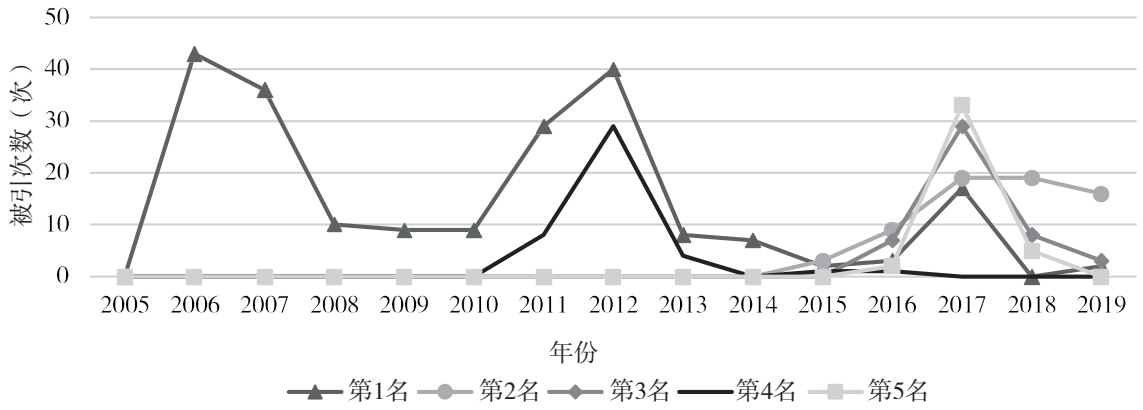

图 3　前 5 政策随时间的被引趋势

2001 年起，选取关键词的前 5 项进行统计，并将 20 年来的政策文本关键词做汇总，统计前 10 项关键词，具体关键词见表 2。对关键词进行分年份的热词云图表示，如图 4 所示。

表 2　历年科技政策关键词

年份	关键词
全部	项目、企业、实验室、建设、课题、高新技术、创新、验收、科普、创业
2020	疫情、防控、新一代人工智能、试验区、创新
2019	申报、项目、负责人、研发、新一代人工智能
2018	项目、开发区、产业、高新技术、建设
2017	专项、创新、申报、十三五、负责人
2016	项目、推荐、实验室、指南、创新
2015	建议、合作、提交、企业、建设
2014	中心、中小企业、创业投资、国家级、引导
2013	计划、农村、产业、认定、创新
2012	十二五、专项规划、行业、落实工作、规划
2011	项目、课题、十二五、预算、验收
2010	计划、实验室、大学、课题、企业
2009	企业、评审、申请、资助、自然科学
2008	企业、 实验室、农村、公司、高新技术
2007	企业、 建设、计划、科技、高新技术
2006	项目、企业、预算、财务、高新技术
2005	建设、软件产业、科学技术、预算、高新技术

续表

年份	关键词
2004	科技进步、技术创新、示范、高新技术、试点工作
2003	实验室、认定、基地、建设、科学技术
2002	课题、科研、高新技术、试点工作、预算
2001	十五、企业、科学、出口、农业

可以看到，政策文本中的整体关注重点为以下几部分：首先，在基础科研方面，主要关键词是“项目”“实验室”“课题”等，其相关政策正文旨在通过项目和课题的研究，加快我国科技基础研究的发展；其次，在行业领域，关键词包括“企业”“行业”“产业”等，且出现频率较高的年份在 2007 年到 2011 年之间，证明我国在十一五期间，科技政策规划向企业行业倾斜；在技术领域，“新一代人工智能”这一关键词在 2020 年高频出现，代表我国近年来面向科技前沿开始逐步关注技术层面的具体科技发展规划及部署。从表 2 和图 4 中可以看出，从 2016 年开始，“创新”占据了较为主导的位置。另外，由于 2020 年新冠疫情的爆发，“疫情”“防控”等主题词也出现在高频词汇中，体现了科技政策紧跟时代、国情，贴近人民生活需求，保障人民生活与健康，服务国计民生。我国每五年规划作为我国国民经济和社会发展的重要部分，也贯穿了我国科技政策的关键词表。

目前我国正处于科技创新快速发展的时期，从 2001 年以来的关键词展示可以看出，我国整体的科技发展战略部署，从基础研究向高新技术创新转

变，从重视宏观管理向兼顾重点领域转变，同时又牢牢把握科技管理中的日常工作，如项目的统计验收、科学技术普及等。对近 20 年来整体的关键词云展示如图 5 所示。

图 4　2001—2020 年政策关键词云图

图 5　总体热词云图

从图 5 中可知，近年来我国科技方面关键的研究主要集中在对实验室和企业的科技基础研究建设中，再次突显了二者所代表的基础研究与应用转化对于科技创新的决定性作用。另外政策对于科技创新、万众创业方面给予了足够的重视，科技与经济的结合是科技政策制定的重要抓手。同时，在维系国计民生的农业方面，科技政策也给予了较大的关注，整体而言我国的科技政策热点词汇与我国的基本国策及政策导向相吻合。

除分析政策关键词外，在政策文本热点中，对关键词的聚类同样能够反映出主题的变化趋势。本文使用了 VOSVIEWER 软件，对政策文本的关键词进行词频及词共现的分析，生成的词聚类如图 6 所示。

在政策文本的关键词聚类中，通过对聚类的参数调整，最终共聚集 8 个类别，其中较为突出的几类可以表明科技政策的主题关联规律。首先，我国中央级别的科技政策，整体主导了项目的申报及研究、科研机构的经费和管理等科技管理工作，这类关键词聚类占整体关键词的 1/3；其次是有关基础研究的主题聚类，该类主题重点关注我国的实验室建设以及持续的基础研究基地建设，相关主题关联显示其政策影响已下沉到我国的各个省份；另外，高新技术、产业、园区、创业等词语在同一类别中，体现了我国对科技与经济结合的重视；创新、技术、科技、转化等在同一类别，证明我国科技政策较为关注创新的转化与效率，并将创新的成果转化和产出作为一个重要的衡量标准来执行。

3.2.2　基于文档的主题分析

本文对科技政策的全文内容进行基于主题的分析，使用上文提出的融合 LDA 算法进行主题模型构建，得到 8 个主题和每个主题对应的关键词。每篇政策文本由 8 个主题的多项式分布表示。表 3 显示每个主题的前 10 个高频词，根据高频政策文本以及高频词汇的信息，可以归纳总结出近 10 年科技文本中的研究主题。

主题 1 科技服务与管理，该分类下的政策一般是科技计划项目中涉及的相关项目审计、项目预算等科技管理类政策，体现了我国在科技项目的服务和管理过程中建立了完善的制度体系，并能够随科技发展现状及时对管理方式和手段进行调整和更新；主题 2 科技活动及科学技术普及，该分类下的科技政策一般为科普及相关的科技活动的通知，重

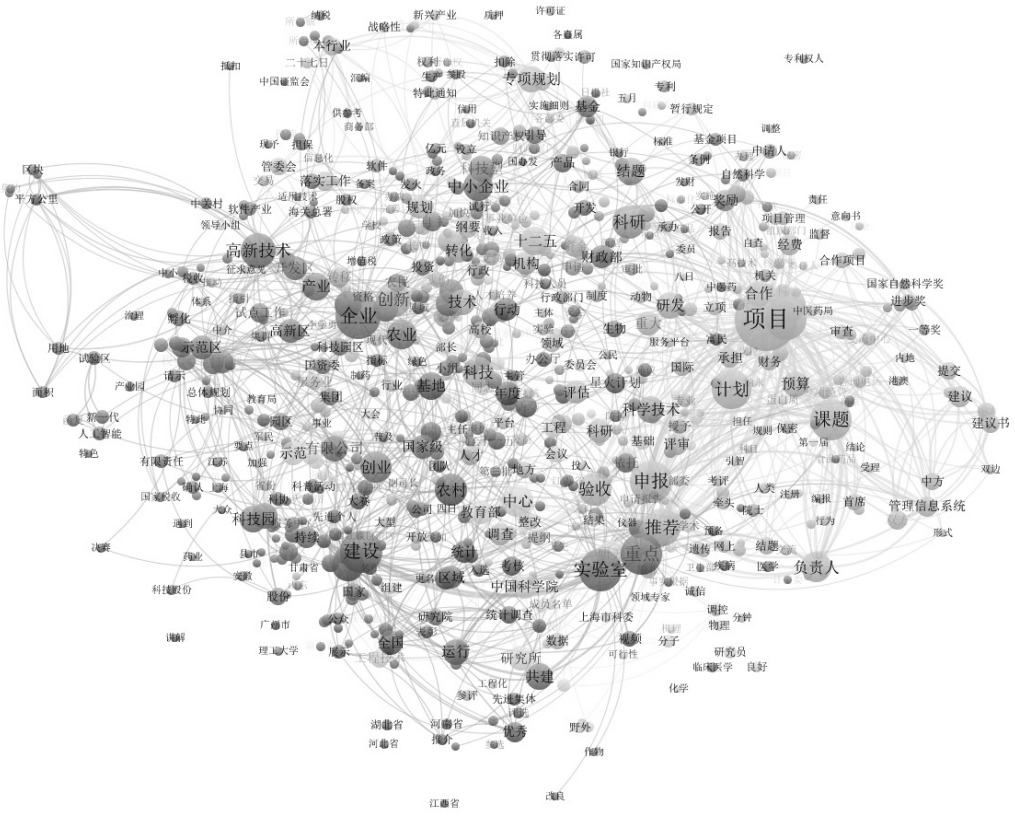

图 6　关键词聚类

点关注的地区集中在农村、试验区等，体现了我国长期关注国民科学素养，重视农业农村科技发展的政策导向；主题 3 科研基地建设与管理，其中重点的关键词体现了国家实验室建设以及基础科研基地建设等相关内容；主题 4 科技统计调查及管理，主要关键词包括统计调查、问卷、进步等，该分类下的政策主要围绕科技管理与评价背景下的科技统计、项目管理等主题，该主题体现了较强的管理特性；主题 5 科研项目申报，其对应的关键词重点为申报、计划等，政策内容大多与科技项目的立项相关，该类主题是科技部官网发布文件中数量最多的一类，也体现科技部的主要职能之一，即制定和发布科技类相关的项目与课题；主题 6 科技资源与基础研究，该分类下的政策主要面向科技发展中相关资源的利用方式、技术研究与服务模式等问题；主题 7 科技金融与发展，主要针对我国在科技金融领域的相关问题颁布政策，面向科技成果向产业转化这一任务，积极从政策层面提供相关保障，为科技型企业排忧解困、提供支持，以此鼓励该领域的发展；主题 8 企业技术进步与高新技术产业化，重点从战略层面，面向企业这一科技创新与经济发展主体，面向高新技术产业发展，针对近些年的高新技术产业化等关注点颁布相对应的指导性政策文件。

3.2.3　基于文档的主题强度分析

本文使用主题强度来描述主题在语料中的受关注程度。主题强度的计算公式如下：

$$P_k=\frac{\sum_i^N \theta_{ki}}{N} \qquad (3)$$

其中 θ_{ki} 是文档 i 的主题的概率分布，N 为文档的数量。图 7 采用所有政策文档在该主题上的概率分布值的平均值表示该主题的强度。

根据图 7 显示，企业技术进步与高新技术产业化是我国科技领域政策中强度最高的主题，这也较为符合我国当前科技发展规划和远期目标。另外，主题 6 科技资源与基础研究主题强度也较为明显，

表 3　主题词及主题描述

序号	关键词	主题描述
1	规定、管理、课题、计划、评审、管理、申请、经费、机构、预算	科技服务与管理
2	科普、农村、发展、作品、天地、科学技术、试验区、活动、中心	科技活动及科学技术普及
3	国家、实验室、建设、重点、科研、基地、评估、计划、创新、成果	科研基地建设与管理
4	组织、进步、社会、政风、问卷、统计调查、有关部门、发展、情况、管理	科技统计调查及管理
5	项目、申报、单位、计划、专项、推荐、国家、课题、重点、研究	科研项目申报
6	研究、技术、发展、资源、创新、基础、国家、科学、服务、基础	科技资源与基础研究
7	金融、发展、创新、投融资、企业、民营、市场经济、建设、产业、创新	科技金融与发展
8	进步、有限公司、战略、产业化、科技园、高新技术、动物、股份、中国科学院、集团	企业技术进步与高新技术产业化

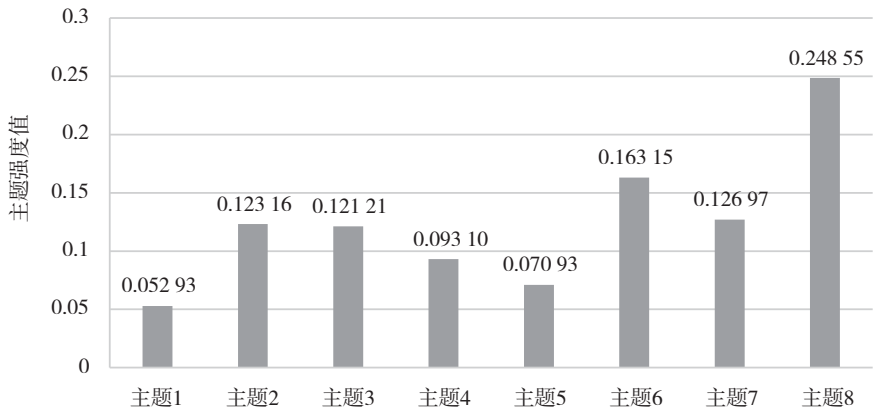

图 7　全局主题强度

这体现了我国一直重视基础研究的态度。紧随其后的主题 2 科技活动及科学技术普及、主题 3 科研基地建设与管理，突出体现了我国对科研基础的重视程度，主题 7 科技金融与发展也具有较为突出的主题强度，体现了科技与国民经济的结合。

根据年份计算每年全部政策文本分别在 8 个主题下的概率分布平均值，来描述分年份的主题强度。图 8 是根据年度评估主题强度，从图中可以直观地看到主题随年份的变化趋势。

从主题强度的总体变化趋势可以看出，近 20 年来，主题强度最为突出的是企业技术进步与高新技术产业化主题，并且在 2012 年达到峰值，主要原因是我国在 2011 年的国家“十二五”规划中提出要大力发展战略性新兴产业，并对高新技术企业有一系列的优惠政策。高新技术产业以科技创新为驱动力，政策的鼓励和推动为企业的高新技术发展提供了持续动力。主题 7 科技金融与发展在 2015 年达到峰值。主题 4 科技统计调查及管理的主题强度呈现周期性变化，可以体现我国科技领域对于科技统计与项目管理的常态化管理趋势。主题 6 科技资源与基础研究，在 2010 年之前主题强度较为突出，之后渐渐趋于平稳，体现了我国对基础研究的持续关注。主题 2 科技活动及科学技术普及，主题强度稳步上升，证明在社会经济稳步发展的条件下，政府越来越重视国民的基本科学素养。

4　总结

本文以从 2001 年起科技部发布的科技政策文本为研究对象，重点对 1 784 条科技政策文本进行了基于多层次主题模型的政策文本量化研究，构建

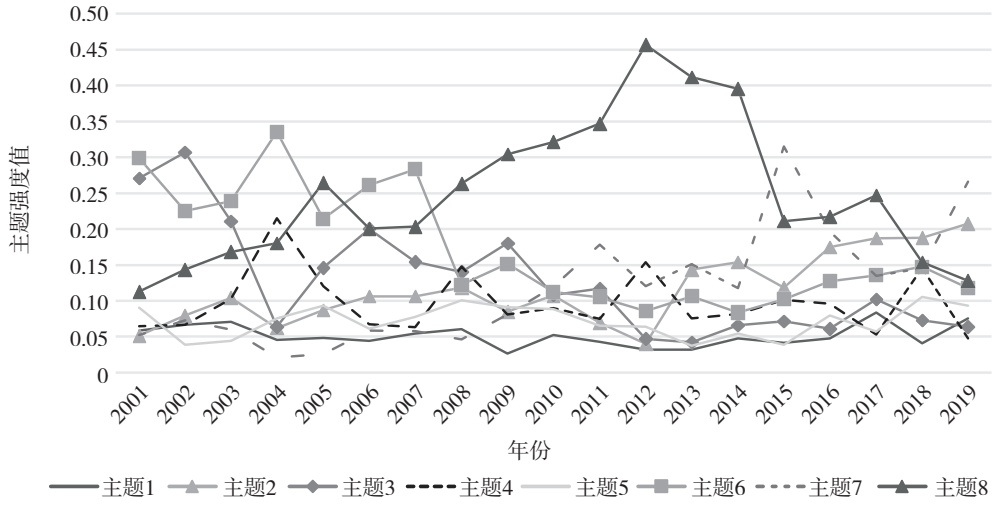

图 8　主题强度随年份变化趋势

了基于概率主题模型的科技政策文本分析算法，以词和文档两个层次，对政策文本进行了基于语义的量化分析。使用词聚类、文本聚类等文本挖掘算法，着重从主题及强度等角度分析了我国近 20 年来的科技政策法规。本文针对科技政策主题模型的构建，较好地分析了我国目前科技政策的战略路线及发展形势，并得出了以下结论。

（1）我国科技政策从数量上来看，呈现五年一周期的变化规律，这种规律符合我国五年规划的战略部署；（2）从上下位政策关系的分析可知，我国整体的科技发展规划主要依据《国家中长期科学和技术发展规划纲要（2006—2020 年）》，多年来一直按照既定方向发展，体现了我国科技领域规划的一致性、权威性和前瞻性；（3）本文依据主题模型对我国的科技政策进行基于词语和文档的内容分类，分类结果显示出我国科技持续聚焦基础研究，并重点关注基础科研基地的建设，在科技发展过程中，我国科技政策的制定同样关注科技项目管理及调查统计等管理职能；（4）通过主题强度逐年变化趋势可以看出，我国近 20 年科技发展关注点已从面向基础研究转向同时关注基础研究、成果转化、企业创新、科普素养等方面，科技政策关注点逐渐覆盖整个创新生态系统。

科技发展是国之重点，世界各国均施行基于本国国情的相关政策法规以推进科技进步发展。通过文本挖掘对科技政策进行文本量化的研究需要持续地进行下去，以便累积历史数据，厘清政策发展脉络，供决策者有效地了解政策发展及执行结果，以数据分析为支撑，更好地进行我国科技政策的长期规划。■

参考文献：

[1] Rothwell R, Zegveld W. Reindusdalization and technology [M]. London: Logman Group Limited, 1985: 83-104.

[2] Schneider A, Ingram H. Behavioral assumptions of policy tools[J]. Journal of Politics, 52(2): 510-529.

[3] Bandara A. How effective are countercyclical policy tools in mitigating the impact of financial and economic crises in Africa?[J]. Journal of Policy Modeling, 2014, 36(5): 840-854.

[4] Schauffler H H. Policy tools for building health education and preventive counseling into managed care[J]. American Journal of Preventive Medicine, 1999, 17(4): 309-314.

[5] Robert B. Policy tools theory and implementation networks: Understanding state enterprise zone partnerships[J]. Journal of Public Administration Research & Theory, 2002(2): 2.

[6] 欧文·E·休斯 . 公共管理导论 [J]. 领导决策信息，2002（15）：4-15.

[7] 黄萃，苏竣，施丽萍，等 . 中国高新技术产业税收优惠政策文本量化研究 [J]. 科研管理 , 2011, 32（10）：46-54，96.

[8] 周京艳，张惠娜，黄裕荣，等 . 政策工具视角下我国大数据政策的文本量化分析 [J]. 情报探索，2016（12）：

7-10.
[9] 王宏起，李婧媛，李玥 . 基于政策文本的“双创”政策量化研究 [J]. 情报杂志，2018，37（1）：59-65.
[10] 赵润娣 . 国外开放政府数据政策：一个先导性研究 [J]. 情报理论与实践，2016，39（1）：44-48.
[11] 李凡，章东明，刘沛罡，等 . 技术创新政策比较研究框架构建及应用——基于金砖国家政策文本的分析 [J]. 科学学与科学技术管理，2016，37（3）：3-12.
[12] 黄菁 . 我国地方科技成果转化政策发展研究——基于 239 份政策文本的量化分析 [J]. 科技进步与对策，2014（13）：103-108.
[13] 白彬，张再生 . 基于政策工具视角的以创业拉动就业政策分析——基于政策文本的内容分析和定量分析 [J]. 科学学与科学技术管理，2016，37（12）：92-100.
[14] 汪涛，谢宁宁 . 基于内容分析法的科技创新政策协同研究 [J]. 技术经济，2013，32（9）：22-28.
[15] 王立，王健美 . 1978—2018 年我国科技政策体系对新材料领域的政策资源投入研究 [J]. 全球科技经济瞭望，2020，35（1）：60-67，76.
[16] 张文伟，赵辉 .LDA 与 BTM 概率主题模型抽取科学主题效果比较研究 [J]. 情报工程，2020，6（2）：66-77.
[17] 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.
[18] 赵杰，李海峰，李纯果 . 基于概率主题模型的京津冀协同发展研究主题演化分析 [J]. 科学技术与工程，2019，19（36）：225-234.
[19] 杨奕，张毅，李梅，等 . 基于 LDA 模型的公众反馈意见采纳研究——共享单车政策修订与数据挖掘的对比分析 [J]. 情报科学，2019，37（1）：86-93.
[20] Blei D M, Ng A Y, Jordan M I, et al. Latent dirichlet allocation[J]. J. Mach. Learn. Res, 2012, 3(4-5): 993-1022.
[21] 政务文书档案专业词表编写组 . 政务文书档案专业词表 [M]. 北京：科学技术文献出版社，2019：3-232.
[22] 黄萃 . 政策文献量化研究 [M]. 北京：科学出版社，2016：170-171.

Quantitative Research on the Science and Technology Policies Based on Multi-Hierarchy Topic Modeling

HAN Xu, YANG Yan

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: Policy is the direct record of government behavior. The quantitative analysis of policy is not only helpful to grasp the overall trend of the policy, but also can offer reference for further government decision-making. Taking the policy texts since 2001 from the official website of the Ministry of Science and Technology (MOST) as sample, this paper constructs topic models based on words and documents. This paper makes quantitative analysis of the policy from the perspective of overall science and technology planning, analyzes and summarizes the development of domestic science and technology policy in recent years, and forecasts the future policy trend.

Key words: science and technology policy; quantitative analysis; topics mining; LDA model