Frontiers of Humanities and Social Sciences

Axon Connecting Ideas to Impact.

数字经济政策文本挖掘分析

董长宏

甘肃省通信产业服务有限公司，甘肃兰州，730000

摘　要　数字经济的发展需有效政策的扶持，对现有政策进行文本挖掘，可为后续政策的废改立等提供依据。本文基于 Python 语言，对 2015—2023 年间中央及省级发布的 1641 项政策文本进行自动化采集与清洗，从发文时序、类型分布、主体结构及地域覆盖等维度开展计量分析，以揭示政策演进的宏观态势与结构特征。在此基础上，借助 jieba 分词、TF-IDF 加权与 LDA 主题模型等方法，对政策文本进行深度挖掘，提炼出数字化体系建设、社会服务建设与基础设施建设三类主题簇。研究发现，我国数字经济政策体系呈现“中央—地方”协同推进、“技术—服务—设施”多轮驱动的鲜明特征，为主题聚类清晰、关键词分布集中的结构化政策生态提供了实证依据。

关键词　数字经济、文本挖掘、政策文本、LDA 主题模型

Abstract: The development of the digital economy requires effective policy support. Text mining of existing policies can provide a basis for subsequent policy revisions and reforms. This study utilizes Python to automatically collect and clean 1,641 policy documents issued by central and provincial authorities between 2015 and 2023. Through quantitative analysis of document timelines, category distribution, institutional structures, and geographical coverage, we reveal the macro trends and structural characteristics of policy evolution. Building on this foundation, advanced methods including jieba word segmentation, TF-IDF weighting, and LDA topic modeling were employed to conduct in-depth analysis, identifying three thematic clusters: digital system construction, social service development, and infrastructure development. The research demonstrates that China's digital economy policy framework exhibits distinct features of "central-local" collaborative advancement and multi-phase "technology-service-infrastructure" momentum. These findings provide empirical

Received: January 22, 2026
Revised: February 3, 2026
Accepted: February 5, 2026
Published:February 20, 2026

evidence for a structured policy ecosystem characterized by clear thematic clustering and concentrated keyword distribution.

Keywords: digital economy, text mining, policy texts, LDA topic model

1. 引言

数字经济作为农业经济、工业经济之后的新生产方式，对资源配置、经济效率、社会公平等多方面产生了深刻的影响，并最终影响中国式现代化的历史进程。近年来，我国高度重视发展数字经济，2019 年政府工作报告明确指出要壮大数字经济，2020 年政府工作报告提出打造数字经济新优势，2021 年再次强调，要加快发展数字经济，建设具有中国特色的数字经济新体系。党的二十大报告也多处谈及数字经济，明确指出要着力发展数字经济，促进数字经济与实体经济快速融合，打造数字经济新业态，提升我国数字经济竞争力和国际影响力[1]，建设具有国际化、数字化的产业集群。可见，数字经济的发展是新一轮科技和产业发展的重要方向，是促进我国经济可持续、高质量发展的重要途径，是构建高端技术、绿色环保、新能源、新材料等一批新的经济增长引擎，是提升我国经济发展的韧性与竞争力[2]。据统计显示，目前我国与数字经济相关的政策已有 330 余条，且发文量呈现出东部>中部>西部的趋势。党的十八大以来，党中央积极推进国家治理体系、治理能力现代化[3]，落实对政策的决策与监督，高度重视公共政策的评估作用，所以对已有政策量化评价是非常必要的[4]。一直以来，甘肃省数字经济发展指数持续稳步提升，但发展仍处于中下等水平。《中国上市公司数字经济白皮书》指出，政策的发布数量与当地 GDP 呈现高度正相关，相关研究也表明，数字经济的发展需相关有效政策的扶持[5]。因此，在当前国内外经济形势和“双循环”多重背景下，如何提升数字经济综合实力，打造具有国际竞争优势的数字化产业链，成为目前亟需解决的问题，而提升数字经济综合实力的本质是如何做强、做优、做大我国数字经济。

2. 研究综述

政策评价是由特定主体针对一项政策的各个方面，按照一定标准所进行的各种评价，其目的是发现政策制定过程中的偏差、明确政策的可行程度及对政策资源的重新配置，进而检验政策制定和执行的实际效果[6]。通过对已有文献梳理，政策评价方法大致分为六类，最早是以“五类评估法”和“三 E 评价法”为代表的经验判断法。其次是因果分析法，Wollmann 采用因果

分析法对政策进行量化评价[7]。学者发现，因果分析法只适合对政策实施效果进行分析，并提出了定性分析方法，主要以问卷调查、案例、专家评估、政策文本分析等方法为代表，马雨萌等采用文本分析法对海量科技政策进行分析[8]；谢青和田志龙运用政策文本分析法，借助政策工具，以不同时期、不同试点城市新能源汽车产业为研究对象，研究政策对新能源汽车产业的作用机理[9]。后期学者提出以文献分析、投入产出模型、模糊集合评价法等方法为代表的定量分析法，李靖华和常晓然采用模糊数学法对我国 2001-2012 年流通产业创新政策量化评价[10]；李晓圆和陈颖等采用模糊综合评价方法，结合层次分析法对易地扶贫搬迁政策进行量化评价[11]；研究学者认为，定量分析法和定性分析法虽然具有科学性，但主观性较强，不适合应用于政策文本量化[12]，进而提出一种定量与定性相结合的内容分析法，孙岩等采用内容分析法，借助政策工具，对 32 份上海生活垃圾分类政策量化评价[13]；谭春辉等结合内容分析法和政策工具，对 2009-2020 年具有代表性的 46 份政策文本进行量化分析[14]；栾丽霞和张梦圆利用内容分析法对 2010-2020 年我国 29 份体育公共服务政策进行量化评价[15]。

3. 数据收集与清洗

在收集数字经济政策文本时，考虑到仅以“数字经济”为关键词进行检索，可能会遗漏相关政策。因此，为了保证政策文本的完整性、科学性和全面性，本研究以“数字经济”“数字化”“数字政府”“大数据”等为关键词，收集我国 2015-2023 年间中央各部委和省级行政区颁布的数字经济政策。首先，在国脉电子政务网、中国知网等官方网站进行检索，借助 Python 编程语言，采用 Selenium 自动化工具模拟浏览器的方式打开政策文本网页，定位并爬取政策文本，将其保存到指定的本地文件中。其次，登录国务院和人民政府官网，进一步检索数字经济相关的政策法规，如规划、方案、通知、法律法规、纲要等政策法规文件；最后，利用北大法宝、百度搜索引擎、知乎和 Google 搜索引擎等工具进行查找和补充，得到 1983 项有效数字经济政策文本。

政策筛选是指在众多的政策文件或文本中，通过一系列的标准和程序，对其进行审查、评估和选择，旨在排除与研究不相关、已废除或重复的政策，确保所使用的政策文本具有可靠性和实用性。筛选过程中应遵循以下基本原则：

（1）公开性：所选取的政策文本可以公开获取和使用，在合法范围内被使用和引用，且不涉及国家秘密和机密，提高政策筛选的合法性和可追溯性。

（2）权威性：确保政策的权威性是至关重要的。一方面，所有政策必须来自国家或省市级官方网站，避免使用非权威来源或可能存在误导性的非正式文件，以提高政策文本的可信度和可靠性。另一方面，国家级政策作为总体性文件，具有全国性的影响力和指导作用，对于各级政府和相关部门制定具体政策具有重要的引领作用，是省市级政策制定的重要基础，也是省市级政策的指导和制约，而县级政策（或规范条例）往往是省市级政策的延伸和补充。因此，在筛选过程中，重点关注国家级和省市级数字经济政策文本。

在遵循以上原则的基础上，对政策文本进行三次筛选。首先，根据政策文本的相似性，剔除重复的、已废除的政策文本。其次，对政策文本进行文种过滤，即删除回复函、请示类、公告类、报告类、通告类等政策文件。最后，通过人工审查和研判，剔除与研究主题关联度不大的政策文本，避免研究过程中的信息干扰和误导。

4. 政策文本计量分析

4.1. 政策发文时间分析

政策发文是指政府或相关机构发布的一系列正式文件、通知或法规，是政府行政管理的一项重要手段，其目的是传达政府的政策意图，以明确、规范和指导某一特定领域的政策内容和执行细则，指导社会各界的行动，推动社会经济的发展和稳定。通过对国家级和省市级的 1641 项政策发文时间统计分析，以揭示出政策制定的发展趋势和规律，从而深入理解政策制定的背景和态势。政策发文时序如图 4.1 所示。

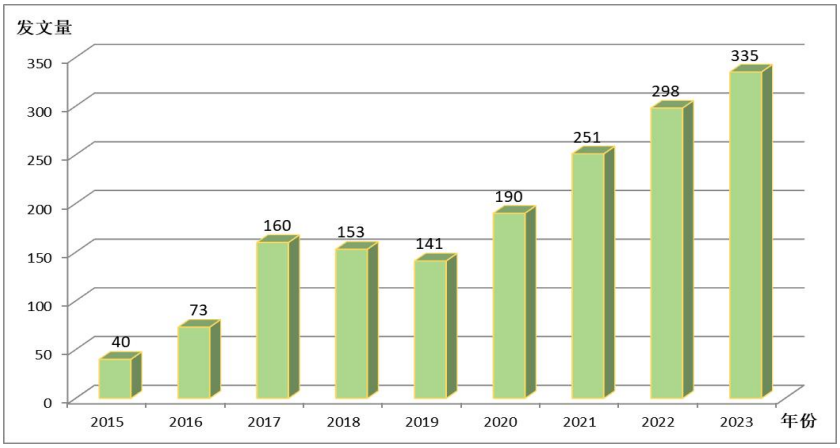

图 4.1　政策发文时序图

4.2. 政策发文类型分析

政策发文类型涵盖政府或相关机构发布的各种形式的文件，旨在规范和指导特定领域的政策内容和执行细则，不同类型的政策发文根据其性质和目的的不同，具有不同的法律效力和约束力，为政府治理和社会管理提供了重要的法律依据和操作指南。本研究根据政策发文标题和具体内容对 1641 项政策进行分类和归纳，将其划分为意见、通知、方案、条例和规划等类型。

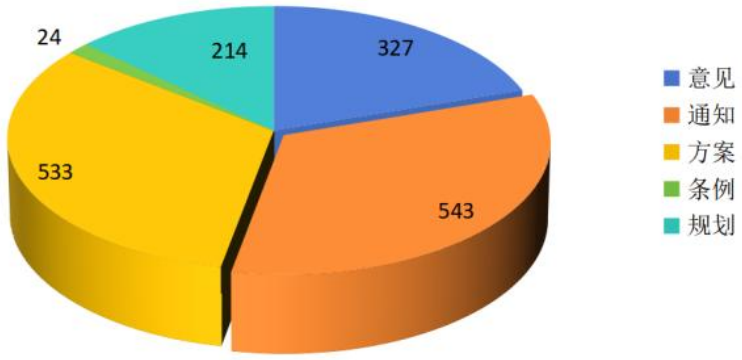

图 4.2　政策发文类型图

4.3. 政策发文主体分析

政策发文主体负责政策的制定、发布和执行等职责，对于确保政策的有效实施、最大程度地实现政策目标具有重要作用。为进一步厘清我国在相关领域政策推动中的主导力量与区域协同态势，本文对各政策发文主体的发文频次进行了统计。如图 4.3 所示，江西省人民政府办公厅、江苏省人民政府办公厅、天津市人民政府办公厅等在发文数量上位居前列，反映出地方政府在政策落地与区域实践中扮演着关键角色；工业和信息化部、国家发改委办公厅、国务院办公厅等中央部委及机构的频繁发文，亦体现了国家层面在战略引导与顶层设计方面的持续发力。这一分布格局既凸显了政策推进中“中央—地方”双向联动的结构特点，也为后续分析政策资源的集聚态势与协同机制提供了重要依据。

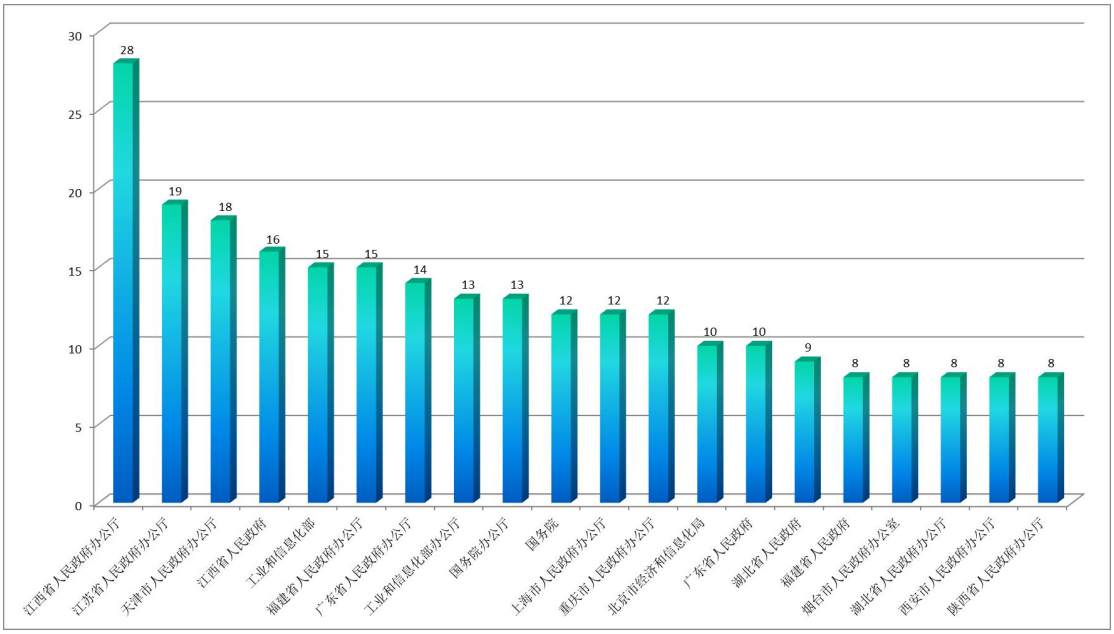

图 4.3　政策发文主体统计图

5. 政策文本挖掘分析

5.1. 政策文本分词

分词是按照一定的规则，将政策文件中的文字进行切分，得到短小的词语或短语的过程，有助于实现信息提取、统计分析和文本挖掘等。本研究首先编写文本文件自动合并代码，将 1641 项数字经济政策合并到一个文档。其次采用 pandas 将文档读入数据库，并定义一个停用词表，其中包含常用的停用词，如“的”“了”“在”和其他符号等，也包含对研究无明显意义的通用词、副词和虚词，如省份名称、人民政府、通知、规划、提升、丰富、非常等频数高，但没有实际研究意义的干扰词汇。最后，使用 jieba 分词库对文档进行切分，得到分词后的文档，

5.2. 政策关键词分析

本研究利用 jieba 分词工具，采用精确模式对政策文本进行分词处理。首先，编写文本文件自动合并代码，通过多线程方式将 1641 项数字经济政策合并到一个文档，减少人工操作失误，提高合并效率。其次，采用 pandas 将文档读入数据库，并定义一个停用词表，其中包含常用的停用词，如“的”“了”“在”和其他符号等，也包含对研究无明显意义的通用词、副词和虚词，如省份名称、人民政府、通知、规划、提升、丰富、非常等频数高，但没有实际研究意义的干扰词汇。最后，使用 jieba 分词工具对文档进行切分，并采用词频-逆文档频率算法（TF-IDF）计算词频，计算步骤如下：

Step 1：计算词频（TF）。TF 是指某个词在文档中出现的频率，用于衡量一个词在一篇文档中的重要性指标，如果一个词在文档中出现频次越少，则 TF 值越小，该词对文本的重要性就越低。反之，出现频次越高，则该词对文档的表达能力越强。

Step 2：计算逆文档频率（IDF）。IDF 用于衡量一个词在整个文档集合中普遍程度的指标。若 IDF 的值越高，则该词在整个文档集合中的普遍性越低，对于区分文档的能力越强。

Step 3：词频-逆文档频率算法（TF-IDF）。TF-IDF 用于衡量一个词对一个文档在整个文档集合中的重要性。若 TF-IDF 的值越高，则该词在当前文档中的重要性越高，且在整个文档集合中的普遍性越低。

通过上述计算，整理出前 40 个关键词，如表 5.1 所示。词云图是根据词语在文本中出现的频率，自动调整词语的大小，使频率较高的词语更加突出。为了直观地展示出政策文本的关键词和热点词，采用 WordCloud 模块生成词云图。数字经济政策词云图如图 5.1 所示。

表 5.1 政策文本高频词统计

词汇	词频	词汇	词频	词汇	词频	词汇	词频
建设	0.2750	管理	0.1343	基础	0.1030	水平	0.0708
数据	0.2526	数字化	0.1313	中心	0.0968	融合	0.0702
数字	0.2473	技术	0.1245	创新	0.0959	智能化	0.0699
服务	0.2471	政务	0.1201	系统	0.0915	社会	0.0694
发展	0.2352	安全	0.1186	监管	0.0862	协同	0.0690
应用	0.1844	资源	0.1139	共享	0.0844	信息化	0.0670
平台	0.1638	工业	0.1130	智慧	0.0826	设施	0.0667
企业	0.1543	体系	0.1121	领域	0.0747	农业	0.0661
互联网	0.1431	能力	0.1118	智能	0.0714	公共	0.0655
经济	0.1425	政府	0.1092	治理	0.0710	部门	0.0649

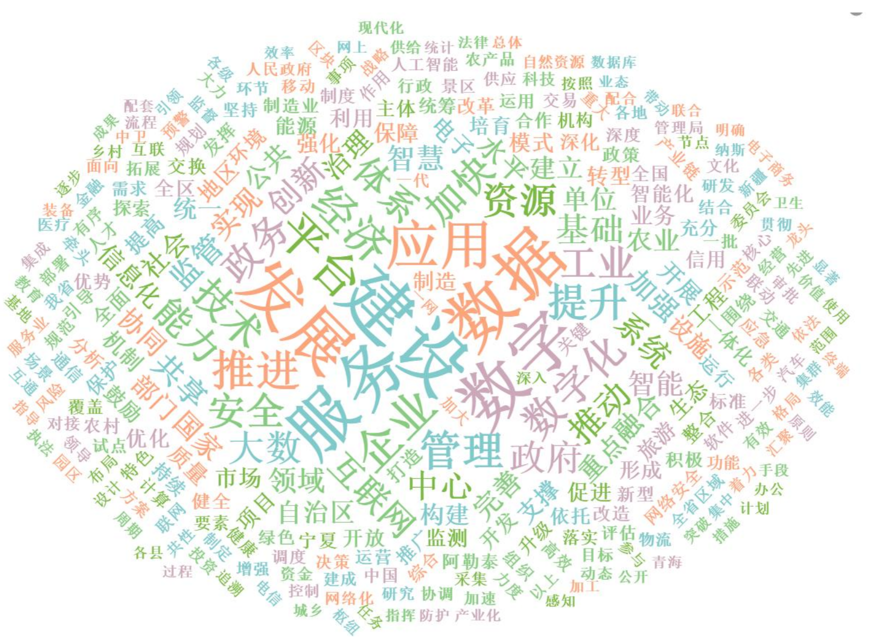

图 5.1 数字经济政策词云图

为进一步揭示政策高频词之间的关联关系，借助 ROSTCM6 软件生成高频词共现矩阵，并将该矩阵导入 Gephi 工具中，利用其强大的网络分析和可视化功能构建共现网络图谱，如图 5.2 所示。

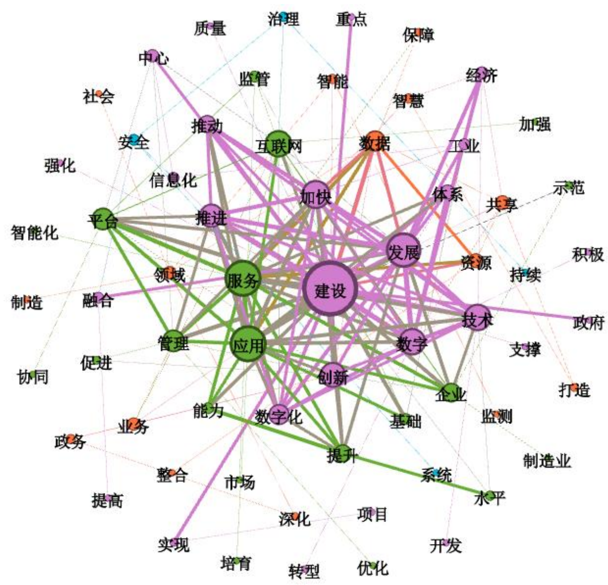

图 5.2　数字经济政策共现网络图

5. 3. 高频词聚类分析

为进一步揭示高频词语之间的关联性和特点，对高频词进行聚类分析。文本聚类分析是将文本中的高频词，按照共现模式或语义相似性分组的方法，以便理解文本数据的内在组织和主体结构，揭示政策文本中的潜在规律、关

联和内涵。常用的文本聚类方法是层次聚类分析法，又称系统聚类分析法，其灵感源自于“物以类聚”的思想，即通过挖掘数据的内在结构和共现模式，将相似的数据点聚合成具有相似特征的类别。本研究将上文的高频词共现矩阵采用 Ochiia 系数转化后，导入 SPSS 分析软件，选择平方欧氏距离进行层次聚类分析，通过识别高频词之间的共现模式，将分成具有内在联系的簇。聚类分析树状图如图 5.3 所示。

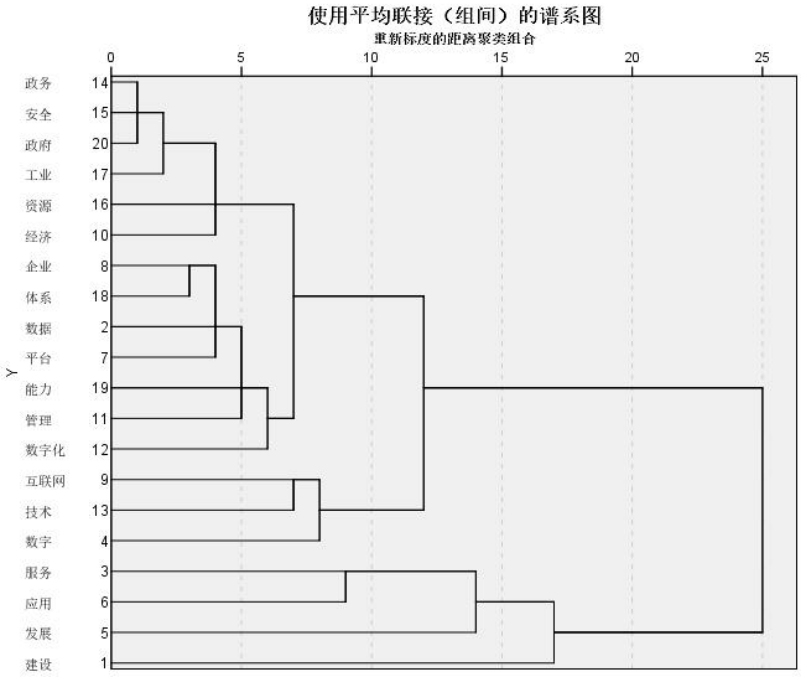

图 5.3　聚类分析树状图

聚类结果显示，数字经济政策高频词聚类可分为 3 个类簇，分别是数字化体系建设类（包括政务、安全、政府等）、社会服务建设类（包括服务、应用和发展等）、基础设施建设类（包括高频词建设）。三个类簇共同构建数字经济发展的整体框架。数字网络、数据中心等基础设施建设为数字经济的持续发展提供坚实基础和底座，特别是网络基础设施、5G 和数字基础设施的建设和升级，是数字化转型的关键驱动力。社会服务建设将数字化技术应用于各个领域，提高社会服务的效率和质量，促进了数字经济的良性发展。而数据安全、数字化平台、数字管理等数字化体系建设为数字经济的发展提供技术支持和保障。其中，政府、企业和数字技术是数字化体系建设的重要组成部分。首先，在政府层面，政务、安全和资源等高频词都是关注的重点领域，政府主要负责制定相关政策、法规和标准，为数字化体系建设提供指导框架和支持体系，以确保数字化体系建设的有序推进。其次，在企业层面，数据、平台和管理等高频词组成核心议题，作为数字化转型的主体和推动者，企业通过数字化转型，实现生产效率的提升、成本的降低以及市场的拓展，

从而增强自身的竞争力。最后，在数字技术层面，人工智能、云计算等新型数字技术，为数字化体系建设提供技术支持，推动数字经济的快速发展。

5.4. 主题词分析

主题词分析是一种文本挖掘重要技术，通过分析文档中的词语分布和共现关系，揭示隐藏在海量文本数据中的主题和内容。主题词分析依赖于数学模型和算法的支持，最常用的模型是 Latent Dirichlet Allocation（LDA，潜在狄利克雷分配）模型。为实现主题词分析，本研究首先基于前文的分词文档，调用 corpora.Dictionary 构建文档-词频矩阵的词典，调用 doc2bow 将文档转换成词袋模型，通过创建 LDA 模型对象，计算每个主题的困惑度，以确定最优主题个数。其次，调用 LDAModel 构建 LDA 模型对象，传入文档-词频矩阵、词典和主题数等参数进行训练，并调用 pyLDAvis 进行可视化。

为进一步观察最优主题数的聚类情况，调用 gensim 库实现 LDA 模型，并采用 pyLDAvis 进行可视化，如图所示。

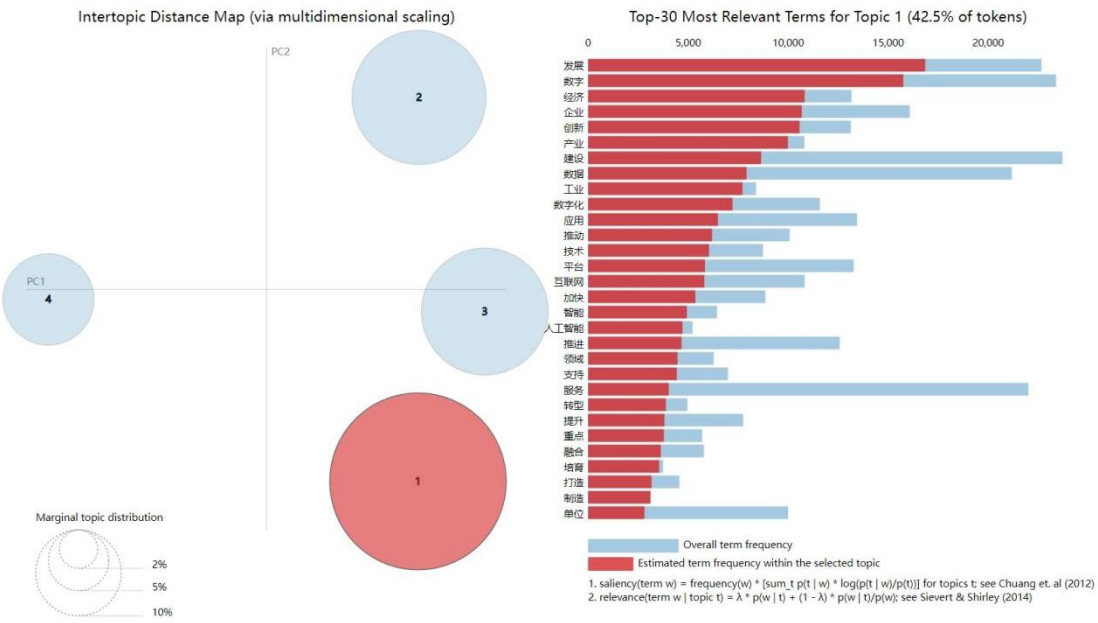

图 5.4　pyLDAvis 可视化图

图 5.4 展示了基于主题模型分析得到的主题聚类与关键词分布结果。图中左侧区域以气泡形式呈现了四个主题聚类，其分布疏密有致，聚类边界清晰，表明当前选择的主题数量具有较好的区分度，能够有效捕捉政策文本中的核心议题结构。气泡的大小直观反映了各主题在整体政策文本中的强度，

即主题的普遍性与影响力。其中，主题 1 的气泡尺寸最大，表明其在政策体系中占据主导地位；主题 2 与主题 3 次之，而主题 4 的气泡相对最小，说明其覆盖的政策内容较为聚焦或影响力相对有限。

气泡之间的空间距离体现了主题之间的语义关联程度：主题 1 与主题 3 位置较为接近，提示两者在政策内容上具有一定的相关性或重叠议题；而主题 3 与主题 4 之间距离较远，表明它们在议题取向上差异较大，关联性较弱。

右侧条形图列出了各主题对应的前 30 个高频关键词及其词频统计。其中，浅色条形表示该词在所有政策文本中的总词频，深色条形则代表该词在当前主题内的词频。“发展”“数字”“建设”“服务”等词在多个主题中均具有较高的权重，反映出这些概念在整体政策话语体系中的核心地位，也进一步印证了数字化转型与服务能力建设在当前政策议程中的重要性。

6. 结论

本研究通过系统性的文本挖掘与计量分析，对我国 2015—2023 年间的数字经济政策进行了多维度剖析，得出以下主要结论：

第一，从政策演进趋势来看，我国数字经济政策体系呈现出明显的阶段性增强与体系化发展特征。2018 年后政策发布频率显著提升，尤其是“十四五”以来，政策密度与层级不断提高，反映出数字经济已从局部探索进入全面布局与深化实施的新阶段。政策类型以“意见”“通知”“方案”为主，体现了政策推进中兼顾指导性、操作性与阶段调整的灵活机制。

第二，政策发文主体结构凸显“央地协同、多部门联动”的治理格局。中央部委在战略引领与顶层设计方面发挥主导作用，地方政府则承担重要的政策落地与区域实践职能。东部地区在政策供给上明显领先，与区域数字经济发展水平形成呼应，也提示中西部地区在政策适配与执行效能上仍有提升空间。

第三，文本挖掘结果显示，“建设”“数据”“数字”“服务”“发展”等成为政策高频核心词汇，共同勾勒出以“基础设施—数据要素—应用服务—融合发展”为主线的政策话语体系。主题模型进一步识别出数字化体系建设、社会服务建设和基础设施建设三大主题簇，三者相互支撑、协同演进，共同构建了数字经济政策的立体框架。

第四，LDA 主题可视化分析表明，政策主题聚类清晰、区分度显著，说明当前政策体系在议题设置上具有较好的结构性与系统性。主题之间既有一定的关联性，又保持相对独立性，反映出政策设计在整体协同与重点突出之间的平衡。

参考文献

[1] 李文军,郭佳.我国战略性新兴产业发展：成效、挑战与应对[J].经济纵横,2022(08):65-75.

[2] 王灵桂,洪银兴,史丹,洪永淼,刘俏,周文.阐释党的十九届六中全会精神笔谈[J].中国工业经济,2021(12):5-30.

[3] 于洪君,戴长征,吴志成,朱锋,金灿荣,苏长和.深入学习贯彻党的二十大精神：以中国方案应对世界之变[J].国际论坛:1-31.

[4] 国家发展改革委评估督导司课题组,王青云,李东.关于构建中国特色“三个重大”评估体系的几点思考[J].管理世界,2022,38(12):76-83+91+84.

[5] TRUSOVA,N.Government socioeconomic policy under the digital economy in the foreign countries and Russia[J].Economic Annals,2019,180(11-12):88-96.

[6] 李志军.加快构建中国特色公共政策评估体系[J].管理世界,2022,38(12):84-92.

[7] Alim P. Hannoura,Gianna M. Cothren,Wael M. Khairy. The development of a sustainable development model framework[J]. Energy,2006,31(13).

[8] 马雨萌,黄金霞,王昉,汤怡洁,周子健.基于政策文本量化研究的科技政策分析服务平台建设[J].情报科学,2022,40(07):169-176+185.

[9] 谢青,田志龙.不同情境下新能源汽车商业模式创新的制度创业比较[J].中国科技论坛,2015(10):46-52.

[10] 李靖华,常晓然.我国流通产业创新政策协同研究[J].商业经济与管理,2014(09):5-16.

[11] 李晓园,陈颖.基于模糊综合评价法的易地扶贫搬迁绩效评价及政策建议——以修水县“进城入园”扶贫搬迁工程为例[J].江西师范大学学报(哲学社会科学版),2019,52(03):130-137.

[12] 时丹丹,嵇国平.基于 BP 人工神经网络的工艺创新科技政策评价[J].统计与决策,2011(16):64-66.

[13] 孙岩,胡茗,张备.政策工具视角下上海生活垃圾分类政策文本量化[J].资源科学,2021,43(11):2224-2235.

[14] 谭春辉,周一夫,冯扬文.政策工具视角下的我国信息服务政策文本量化研究[J].现代情报,2021,41(11):91-101.

[15] 栾丽霞,张梦圆.治理视域下我国体育公共服务政策工具选择逻辑研究——基于 29 份政策文本的量化分析[J].武汉体育学院学报,2022,56(07):5-12.