

基于 LDA 主题模型的省级数据政策文本挖掘与聚类分析

卢妍

西北师范大学, 甘肃 兰州 730070

摘要: 本文以 30 个省份 76 份省级数据政策文件为研究样本, 运用 LDA 主题模型提取政策文本的潜在主题结构, 识别出数字经济发展与数据流通、公共数据开放与获取规范、公共数据资源登记管理、政务数据共享与开放、授权运营管理五个核心政策主题, 并基于主题概率分布矩阵采用 K-means 聚类方法将省份划分为五类政策取向。研究发现, 省级数据政策在关注维度上呈现从宏观顶层设计到微观操作规范的层次递进; 东部省份在开放共享和授权运营上得分更高, 西部省份在大数据条例中较多涉及基础设施建设议题; 登记管理类政策覆盖面最广且地域差异最小。研究为理解省级数据政策的差异化格局提供了量化依据, 为政策协同优化提供了参考。

关键词: LDA 主题模型; K-means 聚类; 政策文本量化; 省级数据政策; 数据治理

0 引言

数据已成为数字时代的基础性战略资源。党的十九届四中全会首次将数据纳入生产要素体系, 此后《关于构建更加完善的要素市场化配置体制机制的意见》等文件相继出台, 对数据的开放共享、授权运营、登记管理等提出了系统性要求。在国家战略推动下, 各省份密集出台地方性法规和规范性文件, 截至 2025 年已初步形成覆盖数据治理关键环节的省级政策体系。但各省政策供给差异明显, 数据政策到底关注什么、差异有多大, 是值得研究的问题。

在已有研究中, 周毅等分析了公共数据治理政策的发展趋势, 发现政策主题范围持续扩大、数据开发利用日益受到重视^[1]; 周鑫构建了三维分析框架对国家层面数据治理政策进行文本分析, 发现宏观层面政策较为薄弱^[2]; 于兴尚等对 31 个省份公共数据政策进行 Nvivo 编码分析, 揭示了政策工具在结构配置上的失衡^[3]。在方法层面, LDA 主题模型等无监督机器学习方法逐渐被用于政策文本量化分析。孙远太等将 PMC 与 LDA 结合, 对 763 项城市人才政策进行了主题识别与类型分析^[4]。

基于此, 本文利用 LDA 主题模型提取潜在主题结构, 进而采用 K-means 聚类对省份进行分类, 揭示不同类型省份的政策取向差异, 以期为理解省级数据政策的差异化格局提供量化依据。

1 研究设计

1.1 样本与数据来源

本文以我国省级层面出台的公共数据相关政策文本为研究对象。政策文本检索来源自各省人民政府官方网站, 检索关键词涵盖“公共

数据”“数据条例”“政务数据”“数据管理”“授权运营”“登记管理”“数据开放”等, 并对其按以下标准进行筛选: 第一, 发文主体为省级人大、省政府或省级数据主管部门, 确保政策效力层级为省级层面; 第二, 政策内容直接涉及公共数据或政务数据的治理、管理、开放、共享、授权运营或登记等核心议题; 第三, 政策类型涵盖地方性法规、政府规章和规范性文件, 排除仅为政策解读、新闻通稿等非实质性文本。经筛选, 最终保留 30 个省份的 76 份政策文件作为分析样本, 文件类型包括数据条例、公共数据管理办法、授权运营管理办法、登记管理细则、政务数据共享管理办法等, 时间跨度为 2021 年至 2025 年。

1.2 研究方法

本文的核心研究问题是“各省数据政策在关注什么、关注差异有多大”, 这需要同时解决两个子问题: 一是从 76 份政策文本中提取出可解释的政策主题, 二是基于这些主题对 30 个省份进行分类。针对第一个问题, LDA 主题模型能够在无需人工标注的情况下从大规模文本中自动识别潜在语义结构, 近年来已被广泛应用于政策文本量化分析。梁益琳等便运用 LDA 对国家层面人工智能政策文本进行主题挖掘^[5]。针对第二个问题, K-means 聚类能够基于文档—主题概率分布矩阵将样本划分为若干类别。相较于单纯的政策主题提取, 聚类分析能够进一步揭示省份间政策关注的组合差异。

1.3 文章分析过程分四个步骤:

一是预处理, 对 76 份政策原文进行清洗, 去除特殊字符和数字。使用 Jieba 分词工具进行分词, 加载包含“公共数据”“授权运营”“数据要素”等 45 个政策术语的自定义词典以确

保切分准确。分词后建立词袋模型，过滤出现在低于 3 份或超过 80% 文本中的词汇，最终得到包含 2050 个特征词的字典。第二步，LDA 主题建模。基于预处理后的语料库，以连贯性得分作为指标，在 3 到 10 之间逐一遍历主题数，观察模型表现。主题数取 5 时，连贯性最高到达 0.5209，困惑度曲线在 K=5 附近也趋于平缓，两项指标一并指向最优主题数为 5。随后，设定迭代次数为 1000 次，训练最终模型，输出每份文档的主题概率分布和各主题的词汇概率分布，并按照高权重词汇的语义关联为每个主题命名。第三步，用 K-means 算法，把 LDA 输出的文档 - 主题概率分布矩阵作为输入，以轮廓系数

（取值在 [-1,1] 之间）作为评估不同聚类数下的分类效果。结合研究需要敲定聚类数 K 之后，本文对 30 个省份做政策取向分类，归纳每一类省份的主导主题与代表性政策文件。最后，把聚类结果与省份的地域分布结合到一起，剖析不同类型省份在数据政策关注维度上的差异。

2 研究结果与分析

2.1 LDA 主题提取结果

依据最优主题数 K=5 训练 LDA 模型，得到了五个政策主题及其核心词汇分布，如表 1 所示。

表 1 LDA主题词汇说明

主题	核心词汇（TOP5）	主题说明
数字经济发展与数据流通	数字、创新、基础设施、数据交易、协同	涵盖数据要素市场化、产业数字化转型、算力基础设施等宏观经济议题
公共数据开放与获取规范	开放、自然人、数据共享、校核协议、保密	强调公共数据向社会开放的法律规范、获取流程与利用协议
公共数据资源登记管理	受理、变更、注销、更正、有效期	聚焦公共数据资源的登记程序、变更规则与凭证管理
政务数据共享与开放	政务数据、数据共享、政府部门、采集、无条件	关注政府部门间的数据共享机制、目录编制与分类分级
授权运营管理	授权运营、专区、整体经营、评价、备案	聚焦授权运营的机构选择、收益分配、安全监管与成效评估

五个政策主题从宏观经济发展一直延伸到微观操作规范：数字经济发展与数据流通以及政务数据共享与开放，这两个主题的文本侧重制度与流通层面的设计；公共数据资源登记管理与授权运营管理的文本落脚在具体操作程序上；公共数据开放与获取规范的文本则衔接开放程序、条件与个人权益、数据协议。拿主题占比来说，各主题各有侧重并且相互关联，共

同勾勒出省级数据政策的内容谱系。

2.2 K-means 聚类结果

基于文档 - 主题概率分布矩阵，用 K-means 算法对 30 个省份进行聚类。轮廓系数在 K=5 时达到 0.785；继续增大 K 值，系数虽微有上升，但新增的类别多表现为单一省份独立成类，兼顾轮廓系数与分类可解释性，本文选定 K=5 作为最终聚类数，结果如表 2 所示。

表 2 省级数据政策K-means聚类结果

政策标签	省份数	主导主题	代表省份	政策特征概述
公共数据开放共享型	14	政务数据共享与开放（均值 0.878）	上海、山东、广东、河北、湖北等	以公共数据管理办法和开放细则为核心政策工具，强调向社会开放公共数据、建立开放平台与利用机制
数字经济发展引领型	20	数字经济发展与数据流通（均值 0.866）	吉林、安徽、贵州、陕西、甘肃、辽宁等	通过数据条例或大数据条例构建顶层法律框架，关注数据交易、基础设施与产业融合
授权运营先行型	18	授权运营管理（均值 0.926）	上海、北京、重庆、四川、贵州等	积极制定授权运营管理办法，探索数据产品的市场化开发与运营模式
登记管理制度完善型	20	公共数据资源登记管理（均值 0.950）	内蒙古、北京、天津、湖北、海南等	集中出台登记管理细则，规范公共数据资源的登记程序与凭证管理
开放获取规范型	4	公共数据开放与获取规范（均值 0.785）	云南、江西、浙江、西藏	侧重公共数据开放的法律规范与获取流程，强调数据共享协议与个人权益保护

登记管理制度完善型覆盖 20 个省份，集中对登记程序、变更规则和凭证管理做了细致规范。开放获取规范型更侧重开放的法律规范和个人权益保护。值得一提的是，同一省份跨多类的现象相当普遍——上海既在公共数据开放共享型中，也出现在授权运营先行型里；贵州的数据条例落在数字经济发展引领型，授权运营办法则分属其他类别，这说明各省政策体系本身具有多元组合的特征。五类政策取向呈现从宏观到操作的递进。数字经济发展引领型涵盖 20 个省份，这些省份倾向于适用数据条例搭建顶层法律框架，政策文本反复出现数据交易、基础设施与产业融合等词。公共数据开放共享型（14 省）与授权运营先行型（18 省）则各有重心：前者重点推动数据向社会开放并建设利用机制，后者较早探索数据产品的市场化开发。登记管理制度完善型覆盖 20 个省份，集中对登记程序、变更规则和凭证管理做了细致规范。开放获取规范型更侧重开放的法律规范和个人权益保护。

从地域分布看，五类取向表现出一定的扎堆现象。数字经济发展引领型集中了东北和西部省份以及部分中部省份，这些地方偏重以综合性条例搭建政策框架，基础设施和产业转型着墨较多，跟“东数西算”赋予西部的枢纽定位基本对应。公共数据开放共享型和授权运营先行型则多落在东部沿海，数字经济底子厚，

往往开放办法和授权运营办法双线推进，政策组合更完整。登记管理制度完善型覆盖面最宽，从内蒙古、北京到湖北、海南横跨东中西部，这块作为全国统一部署的基础制度，各省跟进节奏差不多，地域差异在这块表现得不怎么突出，这种分化既体现了“因地制宜”的政策设计逻辑，也反映出国家战略对地方政策取向的引导作用。

3 结语

LDA 模型从 76 份政策文本中提取出五个核心主题，呈现出从宏观议题（数据要素市场与产业数字化）到操作流程（登记管理与授权运营）的层次递进，表明省级数据政策体系已形成“顶层设计—制度安排—操作规范”的基本链条。K-means 聚类将 30 个省份划分为五类政策取向，发现了省级政策的关注重点与各省的经济基础、产业结构和国家战略定位密切相关。东部省份凭借较为成熟的数字经济生态，能够在基础制度之上进行更多创新探索；中西部省份则以落实国家统一部署为主，政策文本表现出更强的规范性。这种格局说明，国家层面的制度框架对省级政策起到了强有力的引导作用，但同时也提示，部分省份在授权运营和开放共享等需要更多制度创新的领域仍有提升空间。

参考文献：

- [1] 周毅, 陈必坤, 马江华, 等. 基于文本量化分析的我国公共数据治理政策发展研究 [J]. 情报学报, 2023, 42(4): 436-452.
- [2] 周鑫. 基于三维分析视角的我国政府数据治理政策文本分析 [D]; 长春工业大学, 2025.
- [3] 于兴尚, 谭洪, 罗碧, 等. 政策工具视角下公共数据治理研究——基于省级政策文本的分析 [J]. 图书馆, 2024, (2): 28-35.
- [4] 孙远太, 李润东. 中国大中城市人才政策评价及协同类型分析——基于 PMC—LDA 模型的实证研究 [J]. 科学学研究: 1-22.
- [5] 梁益琳. 基于 LDA 模型的我国人工智能政策量化分析 [J]. 价值工程, 2026, 45(17): 75-7.

作者简介: 卢妍 (2004.12—), 女, 汉族, 甘肃宁县, 本科在读, 研究方向: 大数据管理与应用。