DOI: 10.16538/j.cnki.jsufe.2025.05.008

市场导向下数字赋能绿色创新体系构建
——多政策文本扎根与机器学习聚类研究

谢吉青，周昱希，谢家平
（上海财经大学 商学院，上海 200433）

摘　要：在我国经济由高速增长转向高质量发展的关键阶段，创新驱动已成为绿色循环发展和生态文明建设的重要动力，而数字时代重塑了科研范式与创新范式的底层逻辑，面对“人均资源高度紧张”的禀赋约束，亟待构建高效的绿色创新体系。文章立足我国区域实践，运用扎根理论进行多期政策文件的手工编码分析，并创新性地将LDA主题建模与层次聚类深度嵌入扎根理论范式，对政策实施细则的大数据文本进行主题提取，进而高效揭示主题间的层级结构与逻辑关联。这超越了传统主题列表，揭示了绿色创新体系要素间的复杂网络关系。首创“人工诠释-机器验证”的迭代闭环验证机制，将编码结果反馈至机器模型进行反向验证，并据此不断修正迭代优化，实现数据与理论的深度耦合。据此，首次建构了绿色创新体系的钻石模型，以期推动科创成果高效转化与引领性产业实践，为我国自主创新驱动发展战略提供系统性解决方案。

关键词：数字技术；绿色技术；创新体系；市场导向
中图分类号：F124.3　　**文献标识码：**A　　**文章编号：**1009-0150(2025)05-0108-15

一、引　言

当前，全球正处于数字化、智能化、绿色化驱动的新一轮科技革命与产业变革关键阶段，生产力体系正加速向以新技术、新模式、新业态为特征的“新质生产力”形态跃迁。科技创新引领产业系统性变革，强调技术与产业深度耦合、生态环境高质量发展，其中绿色创新更是实现高质量发展的核心动能。在我国经济由高速增长转向高质量发展的重要时期，创新驱动发展理念已成为实现新时期的目标导向、动力来源和主要抓手。党的二十大明确提出“以科技创新引领新质生产力发展，建设现代化产业体系”的战略部署，因此，构建并优化以技术创新为核心、产业链协同为载体、生态绿色为导向的新型创新体系，成为新时代产业创新引领的关键命题。

我国技术创新研发投入近十年持续位居世界首位，但“世界级”的投入并未有效转化为“世界级”的成果产出，经济高质量发展的绿色技术创新驱动效力未能凸显。加之面临着“人均资源高度紧张”禀赋约束，与建设世界科技强国的要求相较，我国创新体制机制和能力还存在诸多

收稿日期：2025-06-16
基金项目：国家社会科学基金重大项目“我国市场导向的绿色技术创新体系构建研究”(20&ZD060)；国家社科基金后资助项目“科创平台的创新驱动效率：理论与实践”(24FGLB052)；上海财经大学研究生创新基金项目(CXJJ-2024-422)。
作者简介：谢吉青(1992—)，女，四川安岳人，上海财经大学商学院博士研究生；
周昱希(1999—)，女，浙江台州人，上海财经大学商学院博士研究生；
谢家平(1963—)，男，四川安岳人，上海财经大学商学院二级教授、博士生导师(通信作者)。

不足和亟待完善之处。目前，我国创新体系建设处于初级阶段，政府高度重视产业联盟和集群网络的建设，致力于推进以政府主导、园区集聚为特征的中国特色工业化，强调建设科创联盟和集群的世界一流创新体系。但如何发挥市场对创新资源的调配效能，引导园区集群和企业“依托科创中心创造专利的技术优势，衍生科创平台创造转化效率的服务优势，借助科创网络创造服务产业化的成本优势”，实现创新链价值位势耦合跃升，亟需政府在体制机制上为创新发展提供保障。党的十八届三中全会强调了“有效市场”与“有为政府”的有序职能分工，党的二十大报告、二十届三中全会《决定》，以及2024年中央经济工作会议上习近平总书记重要讲话，进一步明确“充分发挥市场在资源配置中的决定性作用，更好发挥政府作用”。鉴于此，文章将在市场主导、政府引导的“双重主次导向”机制下，对政府绿色技术创新体系构建的多期文件进行扎根编码与机器学习聚类识别研究，以期为实践运用和政策制定提供有价值的借鉴。

技术创新理论源于熊彼特的开创性工作，Freeman（1987）首次提出国家创新体系，日韩和中国等经济体逐渐从技术模仿者向技术创造者转变，区域创新体系渐受学界关注。技术创新体制是实现后发赶超的重要保障，其过程依赖于链式协同创新，最终形成链式网络组织与制度构成的技术创新体系。技术创新涵盖产品创新、工艺创新、市场创新、组织创新和制度创新等多维度（Freeman, 1987）。绿色技术概念由Brawn和Wield率先提出，强调减少环境污染、降低能源及原材料消耗，也称生态技术创新，系指降低消耗、减少污染、改善生态，促进生态文明建设、实现人与自然和谐共生的新兴技术。绿色技术创新涉及产品全生命周期中的根本性或渐进性创新，贯穿节能环保、清洁生产、清洁能源、生态保护与修复、城乡绿色基础设施、生态农业等领域，涵盖产品设计、生产、消费、回收利用等技术环节（发改环资〔2019〕689号）。唯有加大对创新资源的投入，方能提升绿色创新能力。创新资源系指为实现技术创新所需的、内嵌于创新网络之中有价值的关键资源，如数字信息、知识技术、设施设备、科技人才及创新资金等，且其资源配置又受制度因素显著影响（Malerba, 2002）。创新体系即围绕技术创新设定的一系列制度，关键在于机制改进及组织优化，二者构成复杂的技术-制度协同演化系统（Freeman, 1991）。制度配置与技术存在路径依赖，数据挖掘与人工智能等技术推动知识产权、科技评价、产学研合作等制度创新；反之，制度规制又引领技术发展方向，如重大专项与产业政策等制度具有导向作用，风险投资、科技金融等制度创新具有支撑作用等。而且，知识创新具有公共属性、溢出效应、长期性与风险性等特征，市场失灵理论认为市场机制具有局限性，需要政府之手干预引导（孔令丞等, 2019）。尤其是对于基础性、前瞻性研究投入，重大科创设施、公共技术平台等建设，以及健全知识产权保护、科技成果转化机制、营商环境优化等都需要政府有为干预（谢家平等, 2022）。因此市场机制与政府作用均为不可或缺力量，需要有效市场与有为政府共同发力（董盈厚等, 2021）。更重要的是，社会主义市场经济不断完善和政府职能的转变，政府作用逐步从直接行政干预转向提供制度供给引导。因此，创新体系建设需统筹推进制度建设、资源配置、协同组织、创新生态等工作。

科创平台在组织体系中发挥着关键的桥梁作用，通过整合科技资源对接双边用户，为科创活动提供资源共享、研发协作、专利转化等专业服务，且科创平台生态圈的多主体合作协同可导致价值共创的最终实现，开放创新与竞争合作共同塑造了技术创新的发生机制，有别于传统企业的自主创新；平台链接匹配并提供共享服务，引领多主体直接交互与协同创新（谢家平等, 2017）。平台组织与供应链结合形成平台供应链组织，呈现纵横竞合的网链关系，弯曲甚至打碎传统单向流动的价值链，端到端直接连接上下游用户群体，相比管道型供应链更具竞争力优势（Zhu等, 2023）。动力体系的数智技术发挥底层驱动作用，赋能多主体创新网链的合作互

通，提高了信息共享的准确性与及时性，消弭信任危机。总之，现有研究对技术创新体系的探讨囿于传统创新理论框架，缺乏对数字经济时代系统性、协同性的深度探索与扎根聚类研究。

扎根理论（Glaser等，1968）作为经典的质性研究方法，旨在解析管理现象与事件的内涵特征，揭示因素关系的内在机理逻辑，在学术界获得广泛认可。遵循自下而上的观察路径，对原始数据进行编码与聚类分析，其核心在于将观察到的实际资料进行数据定性构建理论，而非从理论假设出发进行演绎验证，这使之与传统文献研究法形成鲜明的方法论分野。然而，传统扎根理论在应对海量非结构化文本数据时，面临人工编码效率低、主观性强、复杂结构识别能力不足等显著局限。随着文本数据潜力日益显现，学者们开始探索“计算机技术+文本分析”双轨提升传统内容分析方法的科学性与效率。例如，政治学和组织学领域的学者通过完善DICTION工具（Alexa和Zuell，2000），实现了对政治语言多维度特征的量化测量。社会学家近年也积极将大数据聚类方法引入社会学研究实践中（Hanna，2013），推动了社会科学定性研究的技术革新，然尚未建立专用工具与公认最佳实践标准。随着计算机科学迅猛发展，机器学习技术逐渐为突破传统方法的诸多瓶颈、提升复杂结构识别和自动化处理能力，提供了新的理论与方法路径。主题模型（Blei等，2003）通过高效自动识别大规模文本中的潜在主题，为扎根理论提供了强有力的聚类分析工具。然而，虽已有学者尝试将主题模型应用于扎根分析（Nelson，2020），但其应用多停留于“主题发现”或“文本分类”层面，尚未深入挖掘主题间的层级关联与演化关系，导致机器学习解析复杂系统结构的潜力未获充分释放。另有文献虽联合采用主题模型与层次聚类等多种机器学习技术（Bybee等，2024），却未能将其深度融入扎根研究流程，致使机器学习“发现”与人工“诠释”相互割裂，缺乏闭环校验机制，削弱了理论建构的严谨性。

为破解上述局限，针对市场导向下绿色创新制度体系构建的研究近乎空白，本文聚焦目标体系与市场机制缺失、多主体协同创新不足、数字资源未发挥有效作用、政策保障体系缺位等关键科学问题，立足我国区域实践，运用扎根理论进行中央部委多期政策文件的手工编码扎根分析，并以地方政府实施细则文件为大样本，创新性地将机器学习的LDA主题建模与层次聚类深度嵌入扎根理论范式，实现相互验证。尤其是对大数据政策细则文本进行主题提取，进而高效揭示主题间的层级结构与逻辑关联，超越传统主题列表，从而揭示绿色创新体系要素间的复杂网络关系。据此，首创绿色技术创新体系的机制逻辑和“钻石模型”。同时，首创“人工诠释-机器验证”的迭代闭环验证机制——由研究者基于扎根理论原则的编码进行理论诠释与关系构建，机器学习生成主题与聚类结果，再将编码结果反馈至机器模型进行反向验证，并据此不断修正迭代优化，实现数据与理论的深度耦合。显著提升了大规模文本扎根研究的效率、客观性与结构洞察力。研究发现，机器学习聚类分析精准复现了扎根编码识别的6个生成维度，也验证了主范畴、副范畴与概念词组的潜在结构；同时，通过对层次结构进行多粒度命名映射，实现了质性结构与量化特征的跨模态互证，为科学解决政府文本隐性知识的结构化难题提供了数据驱动的新范式。通过识别数字驱动绿色技术创新运行的关键要素与作用机理，不仅丰富了创新理论，更结合中国政策实践，为绿色创新体系理论建构提供了具有现实意义的方法论参考。

二、多期文件的绿色创新体系文本扎根编码

扎根研究的本质是围绕管理问题收集整理相关资料，通过持续归纳、反思和比较实施系统性编码，对文本资料进行逐句解构与概念化，最终以崭新的因素逻辑重组概念要素（Glaser和Strauss，2017），揭示因素间的理论逻辑关系。

政策文件是重要的扎根理论素材来源，解读文本反映着政府组织部门或执行机构对上级

政策的理解与落实行动导向。国家发展和改革委员会与科技部联合下发“发改环资〔2019〕689号”和“发改环资〔2022〕1885号”政策文件，具有如何建构中国绿色创新体系的总领性作用，据此开展手工编码扎根分析，并将这两份指导性文件的各地方实施细则为大样本，再进行机器学习扎根聚类校验与补充完善。

（一）扎根编码过程

经典扎根要求研究者中立搜集相关资料，通过三轮编码与归类递进，实现理论要素的自然涌现。整个扎根编码环节包括开放性编码、主轴编码与选择性编码，严格遵循扎根理论范畴归类和模型构建步骤，对所收集的素材资料开展概念化和范畴化工作（谢家平等，2019）。参见图1。

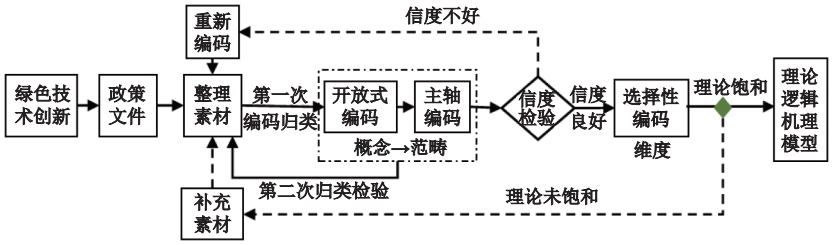

图 1　扎根编码程序

第一轮首先对“发改环资〔2019〕689号”文件素材实施初始开放式编码的贴标签、概念化和范畴化任务，初步建立主范畴类别。由1名绿色生态专家、1名科技创新领域教授以及3名绿色技术创新方向博士生组成编码小组，通过集体研讨完成标签贴注与概念化抽象。

基于此进行第二轮归类工作，对资料素材再次实施细粒度归类，验证研究问题主范畴的完整性与适配度。第二轮新增邀请2名绿色技术创新学者和1位技术创新管理的政府专家，三人独立分组对素材标签进行概念归类，归类结果存在三种情形：（1）完全相同；（2）两组相同；（3）完全不同。对第二轮归类结果实施信度检验。检验信度未达标则返回重启编码；反之，检验信度合规，则开展选择性编码，生成因素维度。

基于贴标签、归类、验证等工作基础，再次依据“发改环资〔2022〕1885号”文件，重复上述编码全流程，验证范畴与维度，直至无新概念标签和新范畴涌现（理论饱和）为止。

（二）开放式编码

先对文件资料进行抽象归纳，并转化为相对简化的概念和范畴，完成贴标签、定义现象（概念化）、挖掘范畴（归类）并最终为范畴命名。若4人编码小组存在分歧，则通过充分讨论论证力求达成标签共识，以规避因编码者主观意见导致的编码结果偏差，确保过程的客观性。至此，完成了资料的分解、检验、比较、概念化和范畴化，结果如表1所示。

进而针对“发改环资〔2022〕1885号”文件，再次贴标签，编码验证概念与范畴，对“海量数据挖掘、关键场景团簇、众创空间、技术技能人才、供需匹配遴选”等概念标签，以及“团簇数实协同、孵化场景、成果转化”等副范畴进行了校正，直至无新概念标签为止。

经过以上操作，共得到470个概念标签，总结出68个概念，进一步比较和归类，将具有逻辑关系的概念进行整合，完成范畴化工作，最终形成32个副范畴、16个主范畴。

（三）主轴编码

基于上述开放性译码所得的具备操作性定义的数组概念与范畴，本研究进一步剖析这些范畴之间的逻辑关系与脉络，旨在建立有意义的内在联系。此时采用扎根理论中经典的译码典范，即借助所分析现象的因果条件、现象、脉络、行动/互动的策略及结果，对范畴进行逻辑关

表 1 开放式编码及基本范畴

概念化	副范畴	主范畴	概念化	副范畴	主范畴
技术交易 环境交易	市场主导	高质高效	知识共享 资源整合	集群效应	知识创造链
科创金融 创新补贴	政府引导		模仿学习 竞合行为	同群效应	
绿色技术特征 绿色创新战略	绿色技术突破	绿色生态	专利数量 专利质量	创新能力	技术创新链
绿色产业投资 绿色生态城市	绿色生态建设		孵化器 众创空间	孵化场景	专利孵化链
原创技术 颠覆式创新	自主创新	创新效率	首席专家 技术技能人才	专业团队	
引进吸收 模仿创新	集成创新		供需匹配遴选 成果转化率	成果转化	产业转化链
R&D人员 技术员	科创人才链	资源配置	新产品占有率 新产品营收	产品生产	
信贷资金 风投基金	科创资金链		通用标准体系 产品认证制度	技术认证	产业量产认证
科创园区 科创设施	科创数智链		企业商标认定 地理标识认证	标识认证	
网络设施 通讯设施 算法算力	数字基座	数字赋能	技术评价规范 考核技术成果	创新绩效评价	产业实践评估
云上算力 聚合挖掘	数据要素		打击侵权行为 健全保护制度	专利保护	约束政策
海量数据挖掘 关键场景团簇	团簇数实协同	产学研科创中心	环境容量 碳排放定额	环境规制	
项目合同式 竞合联盟 区域集群	产学研一体化		分类回收 循环再用	回收规制	
资源共享平台 公共服务共享 中试转化平台	科创平台	科创平台供应链	责任明确 政策衔接	统筹协调	激励政策
企业主体地位 多主体合作 网络引流协同	平台供应链		股权激励 减税激励 采购激励	激励机制	
			市场服务机制 公平监管机制	营商环境	服务政策
			国际交流合作	开放合作	

联，实现资料的结构化充足。各主范畴或核心体系维度的来源以及内涵释义见表2。

在主轴编码过程中发现，开放式编码获得的16个主范畴存在内在逻辑联系，因此，根据不同范畴之间的联系，按故事逻辑线归纳出6个核心维度（6大体系）。

（四）选择性编码

选择性编码旨在阐明故事线，凝练诠释创新体系的核心，形成核心范畴的故事线。通过多轮分析思辨和比较权衡16个主范畴与之相应的32个副范畴归属及其相关关系，并多次回访相关专家，最终厘清各范畴之间的逻辑关系。因此，绿色创新体系的作用机理逻辑如图2所示。

表 2　主轴编码及核心维度形成结果

体系维度	主范畴	范畴关系内涵
目标导向体系	高质高效	高科技数智技术和高效能生产能力协同加速技术变革与产业转型升级高质量发展
	绿色生态	以技术突破赋能产品升级为起点,经规模效应推动产业升级,构建可持续生态环境
	创新效率	以提升自主创新水平为目标,提高绿色技术创新全流程效率,缩短创新周期降本增效
动力驱动体系	资源配置	优化资金、人才、技术等资源要素配置,强化绿色技术创新内生动力
	数字赋能	依托大数据、AI等数字技术支持,系统发挥数据要素的赋能作用,驱动绿色技术创新
组织协同体系	产学研科创中心	数字能力与场景关系的数实融合联动,实践项目式、竞合联盟、网络集群等一体化
	科创平台供应链	重构专业化科创平台供应链,健全资源集聚、引流服务、网络协同等功能组织体系
过程运行体系	知识创造链	知识生产与传播形成集群效应与同群效应,筑牢0-1的科学研究知识基础
	技术创新链	1-10的科创过程并行裂变与交叠纠偏,提升创新主体的技术研发-专利申报创新活力
	专利孵化链	技术专利10-100的中试孵化和试生产关键进程,市场分析反馈校准转化方向
	产业转化链	推动创新成果100-1 000的产业化创新应用,培育1 000-10 000的绿色产业生态集群
制度保障体系	约束政策	健全环境规制和回收规制政策,降低碳排放和能耗强度,强化废弃物的循环利用
	激励政策	打造并完善财政补贴、税收优惠、金融支持等激励政策矩阵,激发多元主体创新活力
	服务政策	优化营商环境,深化国际交流合作,构筑绿色技术创新服务保障体系
实践评估体系	产业量产认证	建立产业量产认证制度,规范绿色技术产品的认定和标准体系,加速产业化进程
	产业实践评估	构建产业实践评估制度,量化创新成果的产业化效果,提供创新的反馈调试依据

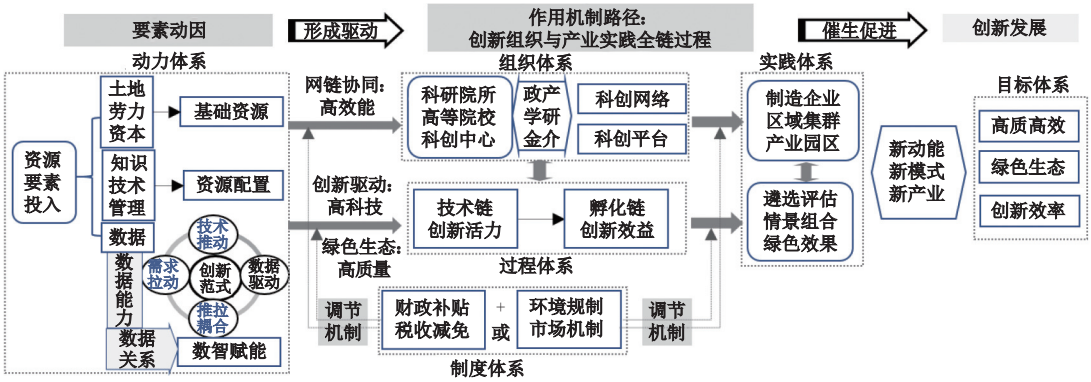

图 2　绿色技术创新体系作用机理逻辑

为提高准确性,编码小组针对归类完全不同的概念化标签展开充分讨论,必要时咨询外部专家意见,最终归入相应类别。经讨论,归类完全不同的5个概念标签中,1个归属“科创数智链”,1个归属“产学研一体化”,2个归属“激励机制”,1个归属“科创平台”。归类的一致性如下:

$$\text{信度}=\frac{n\times(\text{相互同意度})}{1+[(n-1)\times\text{相互同意度}]}$$

其中,n=编码者个数,“相互同意度 = 一致同意数/该类别拥有的总标签数”,文章编码和归类的信度结果均大于0.7,符合信度检验要求。

三、政府文件的机器学习扎根聚类研究

为验证扎根编码框架的科学性,下面引入机器学习技术,对1 382份政府文本进行大样本扎根分析。采用计算扎根框架,结合传统扎根编码方法与机器学习技术(Nelson, 2020),旨在通过算法约束主观编码偏差,提升研究可信度,并构建人工诠释与机器验证的相互检验机制。机器

学习聚类流程如图3所示。步骤如下：(1)通过Python、Excel等工具对原始文本进行清洗与预处理，构建标准化语料库；(2)应用LDA主题模型对语料库进行主题识别，提取核心主题；(3)基于LDA主题结果，进一步采用层次聚类方法解析各主题之间潜在的层次结构与内在关联，形成多层树状聚类体系；(4)以手工扎根编码为基础，对层次聚类结果逐层进行结构验证与标签确认，最终形成“机器学习分析—人工扎根贴标—机器反向验证”的闭环机制。

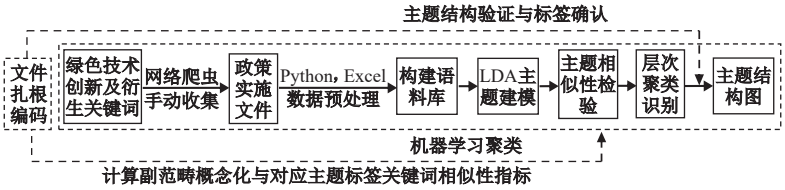

图 3 机器学习聚类流程

(一)文本数据来源

国家发改委于2019年4月印发“发改环资〔2019〕689号”文件，并于2022年12月印发“发改环资〔2022〕1885号”文件。基于此，设定2019年4月至2025年2月为地方实施方案文件样本收集时序区间。首先，以政府网在此时段发布的相关政府文本为数据来源，围绕“绿色技术创新体系”及其衍生关键词(即“绿色技术、绿色创新、绿色制造、绿色转型”)，采用网络爬虫技术采集“政策文本、政策解读及权威媒体报道”三类文本数据。其次，同样基于上述关键词，手动采集各省市政府门户网站同期发布的这三类文本数据。为保障数据库完备性与规范性，文件数据收集的范围涵盖除台湾地区外我国其余33个省级行政区划单位政府门户网站。

为确保数据样本可分析性与“绿色技术创新体系”主题的相关性，本研究采取以下措施：第一，剔除重复项与非文本格式。鉴于中国政府网与各省级政府门户网存在文件交叉引用，以及存在视频类、图片类报道，剔除重复数据样本以及非文本格式类数据样本。第二，排除机构职能说明文本与研究主题的关联性较弱文本。第三，筛选相关性主题文本。排除绿色食品、绿色旅游等无关主题，以及绿色农业中涉及生物多样性保护、生态农业模式创新等不涉及绿色技术的文本。最终获得初始数据集1 382份样本，共2,998,569字。

(二)文本数据预处理

在文本数据预处理阶段，文章遵循自然语言处理的标准流程，通过多阶段结构化处理将原始语料转化为符合LDA模型输入的规范化数据。具体流程如下：第一阶段实施基于正则表达式的文本清洗。通过构建多模式匹配规则，系统性清除HTML标签、URL链接及非中文字符(包括数字、字母及特殊符号)，仅保留具有语义承载功能的中文字符序列，有效消除噪声对特征提取的干扰。第二阶段采用领域自适应分词策略。基于jieba分词框架，整合自定义政策词典进行分词处理。针对政策文本的语体特征，设置词语长度阈值为3-4个汉字，从而在保证语义完整性的前提下避免过度切分问题。第三阶段执行多层次停用词过滤。综合集成四类权威中文停用词库(哈尔滨工业大学停用词表、百度停用词表、中国人民大学停用词表、四川大学机器智能实验室停用词表)实施基础过滤。同时，构建自定义停用词表，剔除如“国资委”“国家级”“事务局”等与政策部门相关但政策内容关联度较低的领域特定词汇，采用并集策略进行二次过滤。该过滤机制有效降低了特征空间维度，移除了冗余信息。第四阶段构建特征加权的文档表征。应用TF-IDF算法生成词项-文档矩阵，其中行向量对应政策文件，列向量表示唯一词项，矩阵元素值

由经平滑处理的TF-IDF权重填充。该数值化表征保留了词语在政策文件中的分布特征，并通过逆文档频率加权突出了领域关键术语，为LDA模型的概率分布估计提供了优化的输入数据结构。通过上述处理流程，原始政策文本被系统转化为符合主题模型分析要求的规范化数据，为后续的潜在主题挖掘奠定了可靠的数据基础。

（三）LDA主题建模分析

使用LDA主题模型对政府文本提取潜在主题，预先确定最优主题数量K作为超参数。

1.将各地政府实施出台的每个细则政策文本视为一个文档r。包含N个词语的政府文本记为$\mathbf{w}=(w_{m1},w_{m2},...,w_{mN})$，其中每个$w_{mn}$表示单个词语。语料库定义为$M$个政府文本的集合，记为$D=\{\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_M\}$。LDA的数据生成过程如下：（1）从狄利克雷分布$\alpha$中取样生成政府文本r的主题分布$\theta_m$，其中$\theta_m$表征政府文本中每个主题的概率。（2）从主题的多项式分布θ_m中取样生成政府文本r第n个词的主题z_{mn}。（3）从狄利克雷分布β中取样生成主题z_{mn}对应的词语分布ϕ_k，其中ϕ_k表示单词在主题z中的概率。（4）从词语的多项式分布ϕ_k中采样最终生成词语w_{mn}。

2.构建多个不同主题数量的LDA模型并计算模型困惑度和主题一致性指标。通过评估不同主题数模型的两项指标来选择最优主题数量K（Blei等，2003）。具体而言：

（1）困惑度（Perplexity）衡量模型对测试集的预测能力，反映概率模型对观测数据的泛化性能。困惑度越低，表明模型对数据的解释越高效。公式为：

$$Perplexity(D)=\exp\left\{\frac{-\sum_{m=1}^{M}\log[p(w_{mn})]}{\sum_{m=1}^{M}N_m}\right\}$$

式中，M为文本数量，N_m为第m个文档的单词容量，w_m为第m个文档的词语序列。

（2）LDA模型的困惑度会随主题数量的增加而降低，但过多的主题数又会导致过拟合，仅将困惑度作为单一标准评价LDA模型并不足够，还需引入主题一致性（Coherence）指标。

$$C(z;S^z)=\sum_{n=2}^{N}\sum_{l=1}^{n-1}\log\frac{D_2\left(w_n^z,w_l^z\right)+1}{D_1\left(w_l^z\right)}$$

式中，$S^z=\{w_1^z,\ldots,w_N^z\}$为主题$z$的前$N$个高频词，$D_1\left(w_l^z\right)$为包含词$w_l^z$的文档数，$D_2\left(w_n^z,w_l^z\right)$是同时包含词语$w_n^z$和$w_l^z$的文档数。一致性越高，主题可解释性越强，语义上的连贯性越好。

结果呈现如图4和图5所示。基于此，采用多准则综合评估策略确定最优主题数。首先，困惑度在$K\geq16$时趋于稳定，确立基础约束条件。其次，系统识别一致性指标的局部峰值点（K=21、26、30、32、34等）作为备选方案（郭峰等，2024）。最后，通过敏感性分析发现：K<32时存在概念过度合并，如K=30时“绿色技术突破”与“绿色生态建设”无法有效区分；K>32时出现主题冗余。综合评估显示，K=32时主题一致性得分达到局部峰值0.467，相比K=31的0.451提升3.5%，且困惑度低于350，处于可接受范围。因此，结合困惑度与一致性指标，确定32为最优主题。

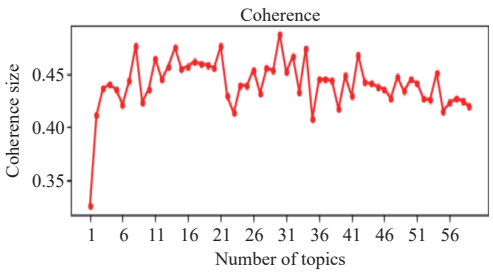

图 4　主题一致性

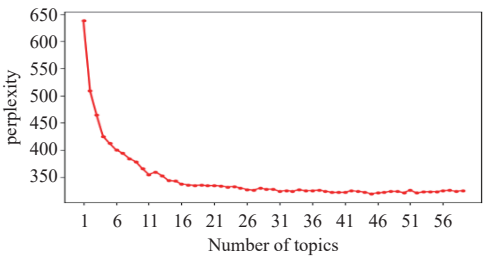

图 5　模型困惑度

各主题的关键词[①]，从词频排名前60的词语中遴选出最能代表主题内涵的10个核心词。需要强调的是，LDA主题模型作为一种无监督学习方法，完全基于数据自身特征进行主题识别，未引入任何人工标签或先验知识。因此，进一步依据文本扎根法识别出的范畴（见表1），归纳关键词列表，并为每个主题分配了最契合的标签。

（四）主题相似性检验

为验证LDA主题模型识别的主题与人工编码结果之间存在高度一致性，引入文本相似性指标*Sim_Simple*（Cohen等，2020），对各主题关键词与表1中各副范畴概念化文本进行逐一配对计算。*Sim_Simple*是微软Track Changes式相似性指标，通过计算文本对齐后的差异块比例来度量整体文本结构（包含词序与内容）的差异程度。该指标取值范围为[0,1]，数值接近1表示文本相似性高，接近0则表示相似性低。具体而言，为识别两份文档之间的内容差异，统计涉及添加（*additions*）、删除（*deletions*）和更改（*changes*）操作的单词数量之和。将此差异量c对两份文档D_1和D_2的总长度（$SizeD_1+SizeD_2$）进行归一化处理：

$$c=[additions+deletions+changes]/[SizeD_1+SizeD_2]$$

为获得取值在[0,1]区间内的相似性度量，通过特征缩放c归一化处理，计算如下值：

$$Sim\_Simple=[c_{\max}-c]/c_{\max}$$

表3展示了各副范畴概念化文本与其对应主题标签关键词之间的*Sim_Simple*相似性指标值。所有主题的相似性指标均接近1，这充分表明主题模型识别结果与人工编码具有高度相似性，从而验证了LDA主题模型结论的稳健性与客观性。

表 3　主题相似性指标

副范畴/主题标签	*Sim_Simple*	副范畴/主题标签	*Sim_Simple*	副范畴/主题标签	*Sim_Simple*	副范畴/主题标签	*Sim_Simple*
市场主导	0.979	科创数智链	0.978	同群效应	0.979	创新绩效评价	0.979
政府引导	0.978	数字基座	0.981	创新能力	0.980	专利保护	0.977
绿色技术突破	0.979	数据要素	0.980	孵化场景	0.979	环境规制	0.979
绿色生态建设	0.978	团簇数实协同	0.978	专业团队	0.979	回收规制	0.977
自主创新	0.980	产学研一体化	0.979	成果转化	0.979	统筹协调	0.978
集成创新	0.978	科创平台	0.981	产品生产	0.980	激励机制	0.979
科创人才链	0.978	平台供应链	0.976	技术认证	0.980	营商环境	0.979
科创资金链	0.977	集群效应	0.977	标识认证	0.979	开放合作	0.976

（五）主题词的层次聚类

为揭示各主题间的层次关系，对LDA所得32个主题的关键词进行层次聚类分析（Bybee等，2024）。结合前文表3所列主题关键词以及图6所示的层次聚类树状图，揭示主题潜在结构，实现对绿色技术创新体系相关政策文件的系统分类。

（1）目标导向体系分支：由主题04“政府引导”与13“市场主导”合并构成的“高质高效”；主题12“自主创新”与28“集成创新”构成的“创新效率”；主题08“绿色技术突破”与18“绿色生态建设”构成的“绿色生态”组成。

（2）动力驱动体系分支：由主题03“数字要素”与11“数字基座”构成“数字赋能”；主题00“科创人才链”、06“科创数智链”与15“科创资金链”构成“资源配置”组成。

（3）组织协同体系分支：由主题10“产学研一体化”与31“团簇数实协同”构成“产学研科创中心”；主题01“平台供应链”与25“科创平台”构成“科创平台供应链”组成。

①限于篇幅，主题关键词表备索。

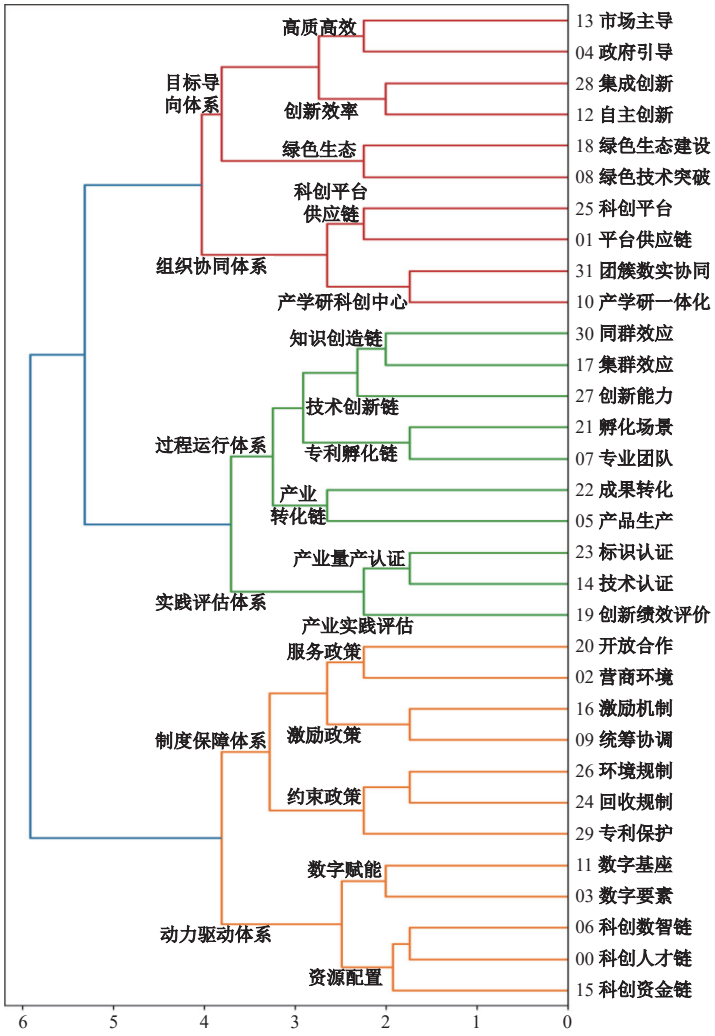

图 6　政府文本主题层次聚类图

（4）过程运行体系分支：由主题17“集群效应”与30“同群效应”构成“知识创造链”；主题27“创新能力”形成“技术创新链”；主题07“专业团队”与21“孵化场景”构成“专利孵化链”；主题05“产品生产”与22“成果转化”构成“产业转化链”组成。

（5）制度保障体系分支：由主题24“回收规制”、26“环境规制”与29“专利保护”构成“约束政策”；主题09“统筹协调”与16“激励机制”构成“激励政策”；主题02“营商环境”与20“开放合作”构成“服务政策”组成。

（6）实践评估体系分支：由主题14“技术认证”与23“标识认证”构成“产业量产认证”；主题19“创新绩效评价”形成的“产业实践评估”组成。

通过对LDA所得主题进行层次聚类，树状图划分为6类广义主题，具有明确的功能边界和内在逻辑。有效支撑了手工扎根编码6大维度的科学性。目标导向体系承担战略目标设定与方向引领功能，解决创新活动“做什么、朝哪个方向”的价值导向；动力驱动体系专注于要素配置与激励赋能，为创新活动提供资源支撑和数字化驱动力；组织协同体系侧重跨组织合作关系构建与平台搭建，解决“谁来做、如何协同”的主体组织；过程运行体系专注创新流程执行与运行

优化，解决“怎么做、如何运行”的过程机制；制度保障体系提供政策制度支撑与环境营造，构建创新活动的制度基础；实践评估体系承担效果评价与反馈改进功能，为体系优化提供数据支撑。同时，利用层次结构命名的多粒度映射实现了质性结构与量化特征的跨模态互证，树状图中各广义主题的分支与扎根理论提炼的主范畴、副范畴高度一致，深化了对主题结构的理解。

四、市场导向下绿色创新六大体系的钻石模型

再次思辨编码形成的故事线，识别出政府文本知识结构化的6个生成维度，并依据主范畴、副范畴与概念标签的潜在结构，首创提出如图7所示的钻石模型体系。

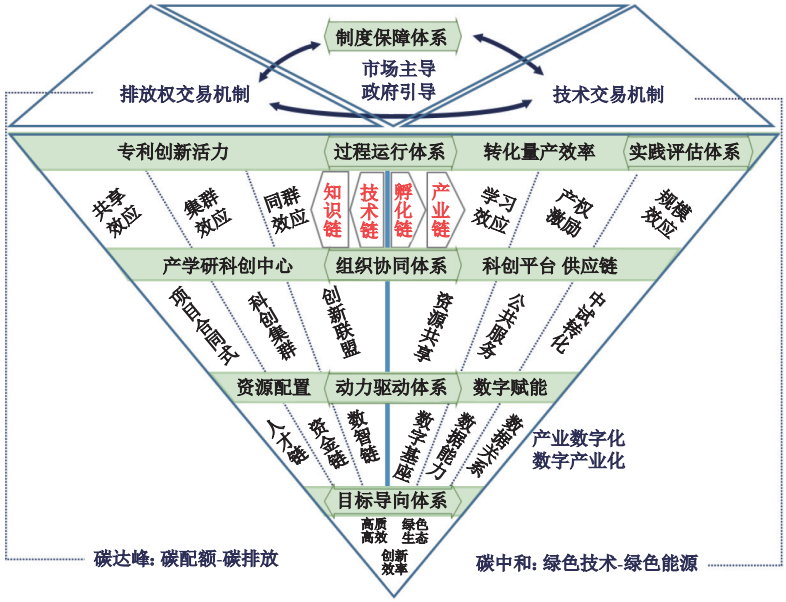

图 7　市场导向下绿色创新体系的钻石模型

市场导向的绿色创新体系须以企业为主体、市场为主导、政府为引导，加速“产学研金介”深度融合，尤其要发挥大型数字平台企业的聚合作用，依据全链创新过程，依次借助科创中心创造专利的技术优势→科创平台创造转化效率的服务优势→科创服务网络创造产业化集群的成本优势，重点建构与完善如图7中（遵从“自底向上”构建思路）的六大创新体系。

（一）建构数据洞察的绿色创新目标体系

兼顾政府与市场作用，在政府环境规制约束下提升创新效率，并运用大数据分析设定专利技术转化率与产业化率目标。一方面，充分发挥市场对资源配置的决定性作用，以市场机制激发创新主体积极性，弥补政府在方案选择、技术推广和研发资金投入等方面的局限，借助数字技术的算力算法能力优化全创新链资源配置。另一方面，利用科创平台汇聚的供需数据（Zhu等，2023）进行场景挖掘与评估，实现绿色技术创新即时检测与动态纠偏，强化自主创新技术路径升级。通过场景仿真识别绿色技术遴选与规划技术转化方案，赋能科创网络产业化服务，推动品牌创新路径升级（谢家平等，2022）。

（二）搭建数字赋能的绿色创新动力体系

一方面，在政府环境规制约束下完善创新资源市场配置体系，通过排污权等环境交易机制，引导企业/园区/区域采用节能降耗、减排增效等绿色技术，激活其绿色技术需求，引导绿色

技术的可持续开发与开放式转化创新（Albats等，2023），以此带动科技人才、科创设施设备、创新资金、数字信息等资源的市场有效配置。另一方面，建成具备区域网络稳固、数据智能融合、信息集成共享、网络组织生态、平台创造活力、产业实践应用、安全可信可控等特征的数字化赋能信息网络，加速数字赋能驱动创新过程（谢家平等，2024）。

（三）重构数据关系的绿色创新组织体系

传统创新串链在数据时代难以有效驱动科技创新向产业创新转化。巨量数据的深度嵌入，赋予科技创新呈现数据密集型创新特征，催生数字经济与实体经济深度融合的创新机制（洪银兴和王坤沂，2024）。科技成果向产业转化的创新链发生了根本性变革，由数据驱动的科研范式和创新范式也正在深刻演进（江小涓和靳景，2022）。为此，需依据实际场景采用项目合同式合作、横向竞合联盟、网络纵横集群等组织模式，促进科创网络主体协同创新，形成高效的网络组织体系。这区别于传统的创新主体关系：数字时代的科创网络既生产数据，也聚合数据生产者。科创平台企业既是创新主体，也是众多制造企业开展科技创新和产业创新的枢纽（Zhu等，2023）。其规模效应加持使科创平台的创新链位势持续提升，其开放性特征不断吸引高异质性、强多样化的数据和资源，推动创新从传统的连续性向数字时代的裂变性演进。因此，数字赋能产学研一体化创新，亟需发挥科创平台的资源共享、公共创新、设备托管等集聚服务功能，重构数字科创网络产学研融合与合作模式，提升数智网络协同治理能力，确保数智网络效应有效发挥，健全数智科创平台的市场化运行体系（Xie等，2021）。

（四）优化数据能力的绿色创新过程运行体系

激发创新主体在绿色技术专利全生命周期管理的作用，健全科创网络协同治理机制，调动科创平台多主体间的资源聚集、共享共用和合作创新等协同效应，实现技术链中知识吸收创造和有效溢出、专利的权益保护与效率提升（孔令丞等，2019）。数字时代创新链各环节呈同步形态和异构特征。数据能力深刻变革了技术创新与产业创新方式，呈现出创新活动“正向过程”（并行、交叠、裂变）与“反向过程”（纠偏调整与认证确认）双螺旋演进特征。数据要素将串联创新链重组为并行交叠的创新团簇，有力推动科研、科创和科技的互动耦合，加速产业转化（江小涓和靳景，2022）。数据链重塑知识/专利创造的技术链活力，通过市场遴选与评估机制，不断优化创新技术工艺、强化中试环节、提升专利技术的孵化转化效率，最终释放产业化应用的价值链效益，进而优化创新链各环节的并行交叠结构、协调其作用、激发其活力。由此实现研发、孵化、产业创新的一体化贯通融合，形成具备团簇校准与即时纠偏的协同创新过程体系。

（五）健全数智化转型的绿色创新制度保障体系

数智化转型下，创新链通常被定义为从创意产生到市场推广的全过程，需政策制度保障（黄凯南和乔元波，2018），而数字时代的科技创新显著增加了其复杂性、动态性与环节交互性，纯市场竞争模式存在明显缺陷，主体间不协调问题难以被彻底解决。数据共享与交流催生知识激增，强化“数据激活、语料筑基”的认知理念，吸引更多主体加入创新行列。制度保障需涉及绿色技术创新市场引导、激励、开放合作、市场服务、市场规范等机制（孔令丞等，2022）。因此，政府导向（环境规制与政策激励）和市场机制的双重作用，满足多主体网络组织在绿色技术创新过程中的组织模式与运行机制。从政府引导、市场激励、开放合作、共享服务和公平规范等机制入手，健全市场导向的绿色技术创新制度体系，并洞察相关机制的影响作用关系；同时，借助物联网、区块链和云计算等数字技术，科学识别制度体系构建的制约因素，才能确保责任主体与创新主体间的交易公平、公正与信息透明（黄凯南和乔元波，2018），并依靠政府政策引导竞合联盟或网络集群开展合作创新共对风险，健全环境市场交易体系与绿色技术交易体系。

（六）完善数字网络生态的绿色创新实践评估体系

数字时代的数据能力和数据关系，既赋能了科研范式与创新范式的剧变（江小涓和靳景，2022），亦助力科研组织与创新过程快速迭代和验证，赋能科技实践量产与市场规模洞测。因此，需全面系统评估绿色技术创新政策与创新绩效，将创新主体、创新要素、创新成果及实践应用纳入以创新技术引领的绿色发展、高质量发展、生态文明建设评价体系。遴选优选区域/园区/企业绿色创新指数，促进创新成果转化与产业应用（孔令丞等，2022），并借助政策组合情景仿真，选择、评估实践效果、识别潜在机遇与挑战，加速其产业化发展与应用。

为此亟需建构数据赋能的科技与产业创新团簇体系：一方面，数字时代科创平台聚合产学研资源形成平台团簇网络，集结数智企业构建虚拟空间，由平台企业主导技术与产业创新生态，筑牢左轮科技创新基座，为前沿技术基础研究注入创新动能。另一方面，科技与产业创新相互交叠，驱动新技术孕育新产业（刘世锦，2022）；反之，也能促成跨团簇融合的“并链耦合”机制，实现科技研发到商业市场“供-产-销”全流程动态耦合匹配路径。

五、研究结论与展望

文章立足环境交易和绿色技术交易的双轮市场导向驱动，融合小样本的经典扎根手工编码和大样本的机器学习识别与聚类分析，实施了多期政策的文件编码扎根和实施细则文本爬虫的机器学习聚类识别。依据技术-制度理论框架，以环境规制为约束，以激励政策为引导，以网络生态企业为创新主体，以科创平台供应链为载体实现科创设施资源共享，由此首创全网链创新的六维钻石模型体系。研究发现，一方面，要侧重构建企业联合高等院校、科研院所、金融机构、科创中介等“产学研金介”一体化组织协同生态体系，强化技术创新与产业创新耦合机制；另一方面，要利用数字时代的数据能力和数据关系，打造“并行交叠与反馈纠偏”的数智平台运行模式，揭示技术迸发裂变与产品市场需求的赋能反哺机制。提升数字新业态下多方异构资源团簇协同创新能力，为构建科创网络团簇合作模式提供变革路径。

科创平台具有风险共担特征，有别于淘宝、京东等电商平台。未来研究拟基于长三角、珠三角、京津冀等区域科创平台的创新风险特征场景案例：（1）依托企业和平台调研数据，多案例实证检验绿色创新体系的作用机制和调节机制。（2）将典型科创平台案例访谈调研与统计数据相结合，综合运用数据拟合、算法模拟等方法，分析不同政策组合对科创平台供应链各主体的决策影响强度和趋势，揭示科创要素资源优化协同配置的政策阈值及实践路径。（3）为提升科创平台供应链配置共享资源的效率，基于双边市场理论，探究科创平台如何引流决策。

主要参考文献：

[1] 董盈厚，马亚民，董馨格，等. 金融资产配置与盈余价值相关性——“有效市场”抑或“功能锁定”[J]. 会计研究，2021，(9).

[2] 郭峰，曹友斌，吕斌，等. 机器学习与社会科学应用[M]. 上海：上海财经大学出版社，2024.

[3] 洪银兴，王坤沂. 新质生产力视角下产业链供应链韧性和安全性研究[J]. 经济研究，2024，(6).

[4] 黄凯南，乔元波. 产业技术与制度的共同演化分析——基于多主体的学习过程[J]. 经济研究，2018，(12).

[5] 江小涓，靳景. 数字技术提升经济效率：服务分工、产业协同和数实孪生[J]. 管理世界，2022，(12).

[6] 孔令丞，许建红，刘鲁浩，等. 科创网络推动区域创新的作用机理及实证分析——来自省级面板数据的证据[J]. 上海经济研究，2019，(4).

[7] 孔令丞，王悦，谢家平. 长三角区域一体化扩容、协调集聚与区域创新[J]. 财经研究，2022，(12).

[8] 刘世锦. 加快发展数字化、绿色化的实体经济[J]. 中国改革，2022，(1).

[9] 谢家平，孔訸炜，梁玲，等. 自主创新的科创平台治理因素机理：扎根理论质性研究[J]. 上海财经大学学报，2019，(6)：.

[10] 谢家平，孔令丞，梁玲. 数字时代市场导向下中国绿色技术创新体系构建研究[J]. 当代经济管理，2022，（12）.
[11] 谢家平，张广思，胡强，等. 科创平台服务供应链定价策略——平台“烧钱”还是引流联盟?[J]. 管理科学学报，2024，（8）.
[12] 谢家平，孔訸炜，张为四. 科创平台的网络特征、运行治理与发展策略——以中关村、张江园科技创新实践为例[J]. 经济管理，2017，（5）：.
[13] Albats E, Podmetina D, Vanhaverbeke W. Open innovation in SMEs: A process view towards business model innovation [J]. Journal of Small Business Management, 2023, 61(6): 2519–2560.
[14] Alexa M, Zuell C. Text analysis software: Commonalities, differences and limitations: The results of a review [J]. Quality and Quantity, 2000, 34(3): 299–321.
[15] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3: 993–1022.
[16] Bybee L, Kelly B, Manela A, et al. Business news and business cycles [J]. The Journal of Finance, 2024, 79(5): 3105–3147.
[17] Cohen L, Malloy C, Nguyen Q. Lazy prices [J]. The Journal of Finance, 2020, 75(3): 1371–1415.
[18] Freeman C. Technology policy and economic performance: Lessons from Japan[M]. London: Pinter Publishers, 1987.
[19] Freeman C. Networks of innovators: A synthesis of research issues [J]. Research Policy, 1991, 20(5): 499–514.
[20] Glaser B, Strauss A. Discovery of grounded theory: Strategies for qualitative research[M]. New York: Routledge, 2017.
[21] Glaser B G, Strauss A L, Strutzel E. The discovery of grounded theory:Strategies for qualitative research [J]. Nursing Research, 1968, 17(4): 364.
[22] Hanna A. Computer-aided content analysis of digitally enabled movements [J]. Mobilization: An International Quarterly, 2013, 18(4): 367–388.
[23] Malerba F. Sectoral systems of innovation and production [J]. Research Policy, 2002, 31(2): 247–264.
[24] Nelson L K. Computational grounded theory: A methodological framework [J]. Sociological Methods & Research, 2020, 49(1): 3–42.
[25] Xie J P, Wei L H, Zhu W J, et al. Platform supply chain pricing and financing: Who benefits from e-commerce consumer credit? [J]. International Journal of Production Economics, 2021, 242: 108283.
[26] Zhu W J, Xie J P, Xia Y, et al. Getting more third-party participants on board: Optimal pricing and investment decisions in competitive platform ecosystems [J]. European Journal of Operational Research, 2023, 307(1): 177–192.

Construction of Digital-empowered Green Innovation System Based on Market Orientation: A Multi-policy Text Rooting and Machine Learning Clustering Research

Xie Jiqing, Zhou Yuxi, Xie Jiaping
(*College of Business, Shanghai University of Finance and Economics, Shanghai* 200433, *China*)

Summary: The digital era has fundamentally transformed the underlying logic of scientific research and innovation. Given China's relative resource constraints on a per capita basis, it is

particularly essential to establish an efficient, collaborative, and market-oriented green innovation system. Rooted in China's regional context, this paper integrates the classical grounded theory with the machine learning method to analyze multi-phase government policy texts, aiming to uncover the intrinsic structure and operational mechanisms of the green innovation system.

This paper firstly performs manual coding and grounded analysis based on two overarching policy documents: the Guiding Opinions on Establishing a Market-Oriented Green Technology Innovation System and the Implementation Plan for Further Improving the Market-Oriented Green Technology Innovation System (2023-2025), jointly issued by the National Development and Reform Commission (NDRC) and the Ministry of Science and Technology (MOST). These documents provide foundational guidance for the construction of China's green innovation system. Through this analysis, it identifies six core systems: goal orientation, dynamic drive, organizational collaboration, process operation, institutional guarantee, and practical evaluation. Secondly, this paper analyzes a large sample of local implementation rules derived from these two guiding policies. It innovatively incorporates LDA topic modeling and hierarchical clustering algorithms into the grounded analysis process. A large-scale data analysis of 1,382 local implementation documents is conducted, extracting 32 topics and applying hierarchical clustering to validate and refine the grounded coding. The results show strong consistency between machine-generated clusters and manual coding, confirming the rationality and structural coherence of the six systems.Thirdly, this paper proposes a novel "diamond model of a market-oriented green innovation system". This model reinforces a development philosophy centered on enterprises, driven by the market and guided by the government. It employs a chain transmission mechanism of "Sci-Tech Hubs—Sci-Tech Platforms-Sci-Tech Networks" to propel the innovation process, thereby achieving progressive enhancements in technological autonomy, shared services, and the synergistic integration of technological and industrial innovation.Finally, this paper explores key strategies for digitally empowering the green innovation system. On the one hand, it utilizes big data technology to accurately identify supply-demand matching in the green technology market and facilitate collaborative innovation clusters. On the other hand, it constructs a sci-tech innovation network cluster system to leverage the aggregation and bridging functions of platform enterprises, enabling deeper integration of technological and industrial innovation. The study preliminarily reveals the parallel, overlapping, and dynamically adaptive characteristics of the green innovation system in the digital era, offering a systematic and structured analytical framework for the innovation theory. It also provides practical insights for policymakers in policy mixes and path design to establish a market-oriented green technology innovation system, thereby supporting the efficient translation of sci-tech achievements and enhancing industrial leadership.

Key words: digital technology; green technology; innovation system; market orientation

（责任编辑：倪建文）