

情报工程

中国科技核心期刊
中国科技论文统计源期刊

情报工程

TECHNOLOGY INTELLIGENCE ENGINEERING



中国科学技术情报学会 | 主办
中国科学技术信息研究所

2025年 第11卷 第1期

二〇二五年 第十一卷 第一期

中国科学技术情报学会 主办
中国科学技术信息研究所

情报工程

QINGBAO GONGCHENG

2025年 第11卷 第1期
双月刊



主管：中国科学技术协会
主办：中国科学技术情报学会
中国科学技术信息研究所

编辑部

主编：刘琦岩
执行主编：姚长青
副主编：王莉军 刘志辉 桂 婕 董 诚
学术编辑：张均胜
责任编辑：孙 瑶 杨 岩 李孟秋 韩颖霄
美术编辑：翟芒芒

本刊为中国科技核心期刊（中国科技论文统计源期刊），数字化期刊全文数据库（万方）、中国学术期刊全文数据库（CNKI）、中文科技期刊数据库（维普）、《中国人文社会科学期刊评价报告（AMI）》引文数据库、国家哲学社会科学学术期刊数据库、超星学术期刊“域出版”等收录。

国内统一连续出版物号：CN 10-1263/G3
国际标准连续出版物号：ISSN 2095-915X
地 址：北京市海淀区复兴路 15 号（100038）
电 话：（010）58882458
E-MAIL：qbgc@istic.ac.cn
网 址：http://tie.istic.ac.cn/
印刷单位：北京科信印刷有限公司
发行范围：公开发行
编辑出版：科学技术文献出版社有限公司
广告发布登记：京海工商广登字 20170148 号

如发现印刷、装订等质量问题，请与本刊发行部联系调换。

ISSN 2095-915X



定 价：55.00 元

CN 10-1263/G3
ISSN 2095-915X



情报技术 | INFORMATION TECHNOLOGY

003

大模型赋能下学术文献服务中的智能化应用研究

Research on the Intelligent Application of Large Language Models in Academic Literature Retrieval Services

孟旭阳 白海燕 吕世灵 宋梦鹏

018

融合 BERTopic 和大语言模型的研究前沿识别 —— 以美国 NSF 人工智能领域资助为例

The Research Frontier Identification by Integrating BERTopic and LLMs: Taking the U.S. NSF Funding in the Field of Artificial Intelligence as an Example

范旭辉 穆智蕊

029

基于大模型和知识图谱的电网检修计划智能编排

Intelligent Orchestration of Grid Maintenance Plan Based on Large Model and Knowledge Graph

吴子辰 顾彬 张云翔 王莘然 高华

042

融合专利属性和用户行为分析的个性化专利推荐研究

Research on Personalized Patent Recommendation by Integrating Patent Attributes and User Behavior Analysis

赵学铭 王刚

054

大模型时代我国知识服务领域面临的挑战

Challenges of Large Models for Knowledge Service in China

彭鹏

信息分析 | INFORMATION ANALYSIS

063

基于扎根理论的政企数据协同影响因素模型构建

Construction of a Model for Factors Influencing Government-Enterprise Data Collaboratives Based on Grounded Theory

刘彬芳

072

教育—科技—人才一体化耦合协调发展研究 —— 以山东省为例

Research on the Integrated Coupling and Coordinated Development of Higher Education, Technological Innovation, and Talent Coupling: A Case Study of Shandong Province

姜明月 王舒扬 刘颖莹 赵林 张舒 刘进

085

老年慢性病患者在线健康替代搜寻影响因素研究

Research on the Influencing Factors of Online Health Information Surrogate Seek in the Elderly Patients with Chronic Diseases

王超 卢智增 徐文博

095

“健康中国”研究现状：格局、轨迹与热点

Current Status of Research on “Healthy China”: Patterns, Trajectories and Hotspots

刘樱 王语嫣

109

吉林省高校科研合作网络研究 —— 基于 2013—2023 年科研论文的分析

Research Collaboration Network of Universities in Jilin Province: Analysis Based on Research Papers from 2013 to 2023

孙毅

119

石油类高校高被引论文特征挖掘及影响因素研究 —— 以中国石油大学（北京）为例

Research on the Characteristics and Influencing Factors of Highly Cited Papers in Petroleum Universities: Taking China University of Petroleum (Beijing) as an Example

刘天琳 陆桃妹 张芹



开放科学
(资源服务)
标识码
(OSID)

大模型赋能下学术文献服务中的智能化应用研究

孟旭阳 白海燕 吕世灵 宋梦鹏

中国科学技术信息研究所 北京 100038

摘要: [目的/意义] 大模型等人工智能新技术正影响着人类信息获取和信息创造方式,加速催生学术文献信息服务的新产品、新形态。对学术文献服务中的智能化应用现状进行深入分析,有助于把握现阶段学术文献服务领域智能化应用的成效、不足和未来发展趋势,为进一步的转型升级和智能化创新服务提供参考和启示。[方法/过程] 在网络调研、文献调研和对各系统实际操作使用的基础上,对国内外典型文献检索发现 AI 应用工具的智能应用场景、功能和特点进行总结分析,并与传统学术文献服务进行对比分析,面向科研过程各阶段的科研需求进行支撑分析,对未来学术文献服务的智能化应用发展进行趋势分析和展望。[结果/结论] 普遍形成了自然语言检索及智能问答的创新服务模式、更细粒度的知识服务内容,在用户输入、信息匹配、结果呈现、交互体验、个性化服务等多个方面初具成效。目前仍处于大模型落地应用初期,结合科研全流程的需求和痛点,提出未来学术文献智能知识服务的展望,一是提升多轮对话下的精准服务能力,向“人—机—智”深度融合的高效交互发展,二是打造文本+表格+图像+多维度知识库的细粒度多模态统一智能知识服务,三是个性化、高价值知识创新服务场景与应用的拓展深化。

关键词: 大模型; 学术文献; 智能化应用; 科研过程

中图分类号: G35

Research on the Intelligent Application of Large Language Models in Academic Literature Retrieval Services

MENG Xuyang BAI Haiyan LV Shijiong SONG Mengpeng

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

基金项目 国家社会科学基金青年项目“大模型赋能下的学术文献智慧服务新模式与应用研究”(24CTQ008)。

作者简介 孟旭阳(1992-), 硕士, 助理研究员, 主要研究方向为自然语言处理、知识服务, E-mail: mengxy@istic.ac.cn; 白海燕(1973-), 硕士, 研究馆员, 主要研究方向为信息组织、数字图书馆、关联数据、知识组织系统; 吕世灵(1976-), 硕士, 工程师, 主要研究方向为软件工程、数字图书馆系统建设与服务; 宋梦鹏(1999-), 硕士研究生, 主要研究方向为自然语言处理与人工智能。

引用格式 孟旭阳, 白海燕, 吕世灵, 等. 大模型赋能下学术文献服务中的智能化应用研究[J]. 情报工程, 2025, 11(1): 3-17.

Abstract: [Objective/Significance] Large Language Model (LLM) technologies are reshaping the manner in which humans acquire and create information, expediting the emergence of novel products and formats within academic literature information services. A comprehensive analysis of the current state of intelligent applications in academic literature retrieval services can facilitate an understanding of their effectiveness, limitations, and future development trends in the field of academic literature services at this juncture. Furthermore, such an analysis offers valuable references and inspiration for further transformation, upgrading, and innovative intelligent services. [Methods/Processes] Based on network research, literature review, and practical operation and utilization of various systems, this study summarizes and analyzes the intelligent application scenarios, functions, and characteristics of typical literature retrieval AI application tools, both domestically and internationally. It also conducts a comparative analysis with traditional academic literature retrieval services, supporting the analysis of research needs at all stages of the scientific research process. Furthermore, it undertakes trend analysis and forecasts the future development of intelligent applications in academic literature services. [Results/Conclusions] The innovative service model for natural language retrieval and intelligent question answering, along with more fine-grained knowledge service content, has generally taken shape and has begun to show initial effectiveness in various aspects, including user input, information matching, result presentation, interactive experience, and personalized services. Currently, we are still in the early stages of implementing large-scale models into practical applications. Considering the needs and pain points throughout the entire research process, we propose the following prospects for future intelligent knowledge services in academic literature: Enhancing precision service capabilities in multiturn dialogues and fostering efficient interactions towards the deep integration of “human-machine-intelligence”, establishing a fine-grained, multi-modal unified intelligent knowledge service that integrates text, tables, images, and multi-dimensional knowledge bases, expanding and deepening the scenarios and applications of personalized, high-value knowledge innovation services.

Keywords: Large Language Model; Academic Literature; Intelligent Application; Research Process

引言

自 2022 年 11 月 OpenAI 发布 ChatGPT^[1] 以来, 以 ChatGPT 为代表的人工智能技术迅猛发展, 人工智能生成内容 (AI-Generated Content, AIGC) 在颠覆内容生产行业的同时, 也驱动着各行各业的变革与发展^[2]。2023 年 2 月, 微软发布 AI 驱动的新版必应 (Bing) 搜索引擎^[3], 将传统搜索引擎与 AIGC 技术整合, 推出了智能化的对话式搜索引擎; 2023 年 3 月微软推出了桌面办公软件 Microsoft

365 copilot, 将 GPT4 整合进 “office 全家桶”, 使办公软件与 AI 协同; 2023 年 4 月, Synthesis AI 发布生成式虚拟数字人方案, 可通过文本提示创建逼真虚拟数字人, 实现文本到 3D 的功能。AIGC 迅速扩展到多行业, 并带来崭新面貌。

科技文献信息服务领域也不例外, 国内外一些知名的出版机构、数据库商、图书馆等都在尝试借助大语言模型等人工智能技术, 积极推出智能检索、数据挖掘与数据分析等智能服务^[4], 以提升科研过程的效率, 为科研工作者

减轻负担,加速科研进程。并且,已经有不少的文献检索 AI 应用工具初现锋芒,科技文献信息与人工智能的结合,将对传统的科技文献信息服务工作模式和科研活动本身产生深远的影响。目前科技文献信息服务中的智能化应用研究现状如何?有哪些具体落地的应用场景?致力于在哪些方面或科研过程的哪些环节对科研工作者提供更智能化的帮助?还有哪些不足,未来如何进一步发展?对科技文献信息服务中的智能化应用现状进行调研、梳理和总结,有利于把握科技信息服务智能化发展的趋势,并对于科技文献的智能化服务创新发展研究具有一定的现实和应用意义。

为此,本文对国内外科技文献领域目前典型的智能化应用产品开展调研,梳理总结大语言模型等人工智能技术在学术发现方面的智能化应用场景、功能、特点,提出新阶段科技文献领域智能化服务的展望,以期对业界研究和科技文献服务工作的智能化知识化应用实践提供参考。

1 学术文献领域智能化服务研究现状

图书馆在确保获取使用、激活科学知识和创造新知识的进程中发挥关键作用,一直是学者们研究的重点。在数字图书馆文献的智能化服务研究中,呈现以下趋势:一是服务的再创新,力图更好地为用户提供高质量服务,主动贴近用户、方便用户;二是技术的再升级,力图借助智能化新技术不断强化服务,以满足用户多样化、复杂化和深层次的需求。

在理论研究方面,学者们不断重塑和丰富智能及智慧化服务理念与服务模式,在智慧服务的概念、特点^[5]、服务理念、实现路径、服务体系构建等方面开展数字图书馆智慧化服务的理论研究。初景利等^[6]将智慧图书馆定义为以人机交互的耦合方式、致力于实现知识服务的高级图书馆发展型态,应紧紧围绕“需求牵引,技术驱动,服务主导,能力检验”的基本原则加快做好图书馆向智慧服务转型升级。储节旺等^[7]从理论角度分析了数智赋能的创新生态系统构成及运行机制。

学者们利用大数据、人工智能等技术赋能科技文献服务,在知识管理、数据处理、文本挖掘、智能检索、分析评价、知识发现与创新服务等方面都开展了相关研究^[8-10]。以大模型为代表的新兴技术快速发展,为数字图书馆走向智能化提供了外部环境与技术动能。在科技文献领域,国内外相关学者、机构也积极开展探索研究,一些文献检索 AI 应用工具初现锋芒,如国外的 Elicit、Semmatic Scholar、Scispace、Scopus AI、Scite、Consensus 等^[11],国内的 Aminer、星火科研助手、NSTL 创新服务助手等,这些系统利用大模型新技术,一方面对传统已有的知识抽取、数据挖掘分析、学术评价、智能推荐等应用服务开展技术升级^[12-13],另一方面推出了问答检索、辅助阅读、智能综述、辅助写作及润色等 AI 应用实践,探索多形态新型智能知识服务,为读者提供更加高价值、人性化的知识服务。随着人工智能技术的不断发展,大模型技术在科技文献的智能知识服务建设中将发挥越来越重要的作用^[5]。

2 文献服务系统的智能化应用服务案例

2.1 Elicit

Elicit 于 2022 年发布, Elicit 团队最早从非营利组织 Ought 中孵化, 现在作为一家独立的公益公司运营。Elicit 作为面向科研人员推出 AI 赋能的“研究助手”, 旨在利用语言模型寻找相关论文、提取关键信息、整理概念逻辑, 让科学研究自动化。

相较于传统学术平台, Elicit 主要有智能搜索、智能辅助阅读、概念清单 (List of concepts) 等 AI 功能。

(1) 智能搜索

Elicit 支持以自然语言方式提出一个研究问题开展检索, 检索结果主要由两部分组成:

①自动文献综述。Elicit 依据 4~8 篇相关性最高的论文生成一份简短的文献综述, 并附有引用标记, 帮助用户快速了解查询主题的研究概况。

②文献结果集。不同于传统的检索结果列表页, Elicit 将论文中的详细信息提取到有组织的表格中, 信息呈现方式一目了然。Elicit 提供了 33 个可供选择的信息列, 包括主要发现、研究目标、研究设计、研究方法、理论框架、局限性、结论等, 同时也支持用户创建其他信息项。各信息项均利用智能技术从论文中提取与总结, 使用户方便快捷、结构化、个性化地了解论文信息。

(2) 从 PDF 中抽取数据

Elicit 支持用户上传和管理 PDF 文档, 利用智能技术从 PDF 文件中快速提取数据。用户可选择上传一篇或若干篇 PDF 文档, Elicit 读

取后以表格的形式展示抽取到的信息, 用户可以创建感兴趣的信息项或选择提供的信息项, 进行个性化定制的数据抽取和展示。解锁会员账户后, 用户还可以与论文全文对话。

(3) 概念清单

Elicit 提供在大量论文中快速找到相关主题和概念的功能。用户输入想了解的问题或主题, 平台会搜索主题相关的论文, 从这些论文中找到相关概念, 并过滤掉重复的概念内容, 以表格形式展现相关概念清单结果, 并提供了概念结果的来源片段和文献出处, 用户可点击查看原文相关内容。部分概念来自大模型生成, 尚未有论文明确指出, Elicit 将其来源标明为大语言模型, 仅供用户参考。

2.2 Semantic Scholar

Semantic Scholar 于 2015 年由美国艾伦人工智能研究所 (Allen Institute for Artificial Intelligence, AI2) 发布, 其使命是利用人工智能帮助学者找到和理解正确的研究, 建立学术联系, 克服信息超载, 从而加速科学突破^[14-15]。

除了传统文献检索筛选等基本功能外, Semantic Scholar 利用人工智能技术提供了智能摘要、AI 辅助阅读、AI 生成主题等特色功能。此外, 还开发了一系列的人工智能驱动的工具和服务, 如 Semantic Reader 增强型阅读器, 能够在阅读论文时提供实时的解释、注释、链接等, 让阅读更加易懂和有趣。

(1) TLDR: 智能摘要

TLDR^[16], 即“Too Long; Didn't Read”, 在 Semantic Scholar 上检索文献时, 检索结果列表页每篇文献带有 TLDR 标志后的一句话就是

利用 AI 技术自动为文章生成的一句话摘要总结。TLDR 实现了对论文内容的理解，以精炼的一句话总结论文的核心研究内容，帮助读者更快浏览和了解论文，减少筛选论文的时间。

（2）AI 辅助阅读

Semantic Scholar 利用人工智能技术自动识别和提取论文中的重要信息，如图表、主题、引用等。此外，在文献标题右侧提供 AI 问答功能，用户可以向 AI 提出有关论文的问题，系统利用大模型生成答案，并提供原文中的支持性论述，辅助用户以交互式问答的方式快速阅读和理解论文，但目前仅支持英文论文且仅在有限的文章中可用。

Semantic Scholar 还开发了一款增强型阅读器 Semantic Reader，在文献下方点击 “[PDF] Semantic Reader” 按钮，可以用该阅读器阅读文献。该工具提供丰富、实时解释，注释，链接，导航等，让阅读更高效、更智能；主要包括引文卡片、个性化引用增强、关键要点速读等智能化功能，即用户点击文章引用部分，系统以卡片形式智能提供引文详细信息；提供的目标、方法、结果 3 类文章关键要点抽取，并在页面上通过标签高亮显示，辅助用户高效阅读。Semantic Reader 目前仅支持 ArXiv 上发表的论文，研发团队也正在逐步优化功能，并将拓展覆盖到更多的论文来源。

（3）AI 生成主题

Semantic Scholar 提供主题页面功能，旨在通过收集 AI 生成的定义、该主题经常被引用的论文、最近的论文和相关主题等来帮助用户探索感兴趣的课题。

Semantic Scholar 在每篇文献详情页的 Top-

ics 功能部分，可以找到论文相关的主题并导航至主题页面，用户也可以在搜索栏中搜索主题。目前仅提供计算机科学领域的主题。

2.3 Scispace

Scispace 被誉为基于 ChatGPT 的论文阅读神器，其创始人在 2015 年创建 Typeset，为了打破科学研究的孤岛状态，让科学研究更具有协作性和可访问性，在 2022 年采用新的产品名称——Scispace，帮助用户理解和分析科学研究论文。

相较于传统学术检索，Scispace 提供了智能检索、智能辅助阅读、引文生成器、学术解释等特色功能，下面主要对前两者展开详细介绍。

（1）智能检索

Scispace 支持自然语言提问检索，检索结果页面与 Elicit 类似，主要由两部分组成：①自动文献综述，Scispace 依据最相关的 5~10 篇论文生成文献综述，并附有可循证引用标记，帮助用户快速了解查询主题的研究概况；②文献结果集，Scispace 将论文信息以表格形式展示，提供摘要、方法、结果、局限等 11 个可供选择的信息列，同时支持用户创建其他信息项。

（2）智能辅助阅读

Copilot-Read with AI 是 Scispace 的核心功能，可以通过对话的形式帮助理解论文，也支持用户上传 PDF 进行智能辅助阅读，提供有价值的建议。

用户提出问题后，Copilot 生成相应的答案，并可根据生成的答案定位文章中的位置。值得一提的是，Copilot 支持理解和解释文章中的术

语、图片、表格、公式等信息，用户只需要选择突出显示或框选区域即可，Copilot 可以解释并总结相应的内容，辅助用户在对话中轻松灵活阅读论文。此外，Scispace 支持 75 种语言与 Copilot 互动，用户可以通过选择语言设置，使用自己熟悉的语言进行交互和理解。

2.4 Scopus AI

2024 年 1 月 16 日，全球领先的科技出版与信息分析公司爱思唯尔（Elsevier）正式发布 Scopus AI，Scopus AI 是集成于 Scopus 网站的一款生成式人工智能科研工具。

该工具支持科研人员用自然语言形式进行提问，具备生成可溯源的智能概要、思维导图（Concept Map）、定位核心文献与领域专家等功能。

（1）智能概要

Scopus AI 支持科研人员对任意科学问题进行提问，几秒钟内基于相关科研论文生成一段概要，并提供主要引用的文献（5 篇左右），在页面右侧展示相关参考文献。此外，Scopus AI 还提供“扩展概要”（Expand Summary），进行渐进式研究和扩展，得到有逻辑的基本综述框架，并提供主要引用的科研文献（10 篇左右），在回答结束后为用户推荐 3 个提问，辅助开展进一步的研究。

（2）思维导图

Scopus AI 提供“思维导图”（Concept Map）功能，利用概述中的关键词为每个查询智能生成思维导图，帮助用户快速形成知识脉络，更全面地查看和了解研究主题及其与其他研究领域的关系，更直观地发现和了解研究现

状与趋势，从而促进学科交叉与发展。

（3）定位核心文献与领域专家

Scopus AI 通过优化的 AI 算法智能提供有关查询问题具有开创性意义或奠定基础作用的关键性论文，以及相关的领域专家，并为每位专家生成一份摘要，介绍专家与用户提问相关的工作和贡献。这使得用户无需阅读大量文献即可快速获得宏观全面的研究内容介绍，帮助用户有效识别各领域的专家。

2.5 Scite

Scite 被称为“智能引文”（Smart Citations）工具，由 2018 年成立的初创公司 Scite 研究发布，它通过智能引文帮助研究人员更好地发现和理解文章——显示引文上下文并描述文章是否提供支持或对比证据，能够更好地发现和评估科学文献，从而使其更加可靠。

Scite 目前主要的 AI 功能包括 Scite Assistant 研究助手、智能引文分析助力学术发现增强。

（1）Scite Assistant 研究助手

Assistant 是 Scite 开发的一款对话工具，提供带有参考资料支持的答案。Scite Assistant 以独特的智能引文数据库为后盾，最大程度地降低出现幻觉的风险，提高信息和真实参考资料的质量。

此外，用户还可以上传或搜索某篇文献、某份引文网络，指定 Scite Assistant 依据文献内容进行回答。Scite Assistant 支持自定义设置，如是否使用参考引用；指定证据来源；答案展示风格；搜索论文的时间范围、主题、出版物类型或期刊名称等指标作为过滤器等，引导研究助手在任务上沿着用户期望方向生成回答。

（2）智能引文分析

Scite 基于全文文献，通过人工智能算法抽取引文并判别引文立场，建立引文索引库，同时引入论文中引文位置的识别，基于语义引文增强文献发现与评价。除了文献的常规检索外，一是将引文陈述语句作为独特的检索字段，检索返回引文片段；二是支持按引文所处章节（Introduction、Methods、Results、Discussion 等）、引用立场类型（支持、中立、争议）进行文献筛选；三是同时标注文献的总被引篇数和不同立场的被引次数，支持用户按相应数量进行排序和筛选。

简而言之，Scite 让用户更容易地了解学术论文之间的相互关系和评价，达到学术发现增强的目的，同时提供了引文可视化功能，可以多角度查看文献被引用全局信息，如文献正面、负面评价，点击施引文献可查看详细引用表述，帮助用户更好地理解论文间的联系。

2.6 Consensus

Consensus 是 2022 年 9 月上线的的一个由人工智能驱动的学术搜索引擎，它使用人工智能来呈现和综合科学研究的答案，旨在让所有人都能更容易获取到专业知识。2023 年 10 月推出升级版本 Consensus 2.0，彻底改造了搜索的工作方式，并不断更新和优化服务。

Consensus 目前主要的 AI 功能包括智能检索、答案或见解的智能计量 Consensus Meter、研究助手 Consensus Copilot。

（1）智能检索

Consensus 支持用户以自然语言提问方式查找相关研究，突破传统关键字搜索方法的限制。

检索结果页面主要由四部分组成：①智能摘要 Summary，在页面左上方，Consensus 依据前 10 篇检索结果利用大模型生成一句简单的摘要回答，以使用户对搜索结果有直观快速的认知；②智能计量 Consensus Meter，如果用户检索为“是/否”问题，在页面的右上方，Consensus 依据前 20 篇检索结果，实时将与问题相关的答案结果智能分类为“是”“否”或“可能”，并进行统计和直观的展示；③文献结果集，Consensus 在文献筛选上，提供了更智能和细粒度的筛选项，如研究类型、研究细节（对照研究、人群类型、样本大小等详细参数）、期刊影响力等来自定义筛选结果；在文献结果呈现上，对每篇文献生成一句基于问题的答案、研究快照（方法、结果、人群、样本量）等文章要点信息，辅助用户快速评价和筛选；④相关搜索，在页面的底部，Consensus 提供相关研究查询的建议（推荐 3 个问题），以激发用户进一步研究探索。

（2）研究助手 Consensus Copilot

Consensus 于 2024 年 1 月推出研究助手 Consensus Copilot，为搜索带来了新的灵活性，可响应用户的指令，完成各种不同的研究任务并以基于科学证据的各种格式提供答案。

Consensus Copilot 基于 OpenAI 一流的大语言模型，结合和搜索超过 2 亿篇研究论文的数据库，提供有用的摘要、答案和主题概述。用户输入研究指令，如具体的研究任务、撰写特定的内容、解释术语或概念、寻找研究方向等，可以自定义输出格式并要求以准确引用来回答问题等。另外，Consensus 目前支持 100 多种语言的搜索^[17]，可以设置使用自己熟悉的语言开

展检索和研究。

2.7 Aminer

Aminer 是由清华大学计算机科学与技术系教授唐杰率领团队建立的,是具有完全自主知识产权的新一代科技情报分析与挖掘平台,旨在为科研人员提供全面的搜索和挖掘服务。

目前 Aminer 在利用大模型技术赋能上,主要包括的智能学术文献服务有:AI 检索助手、AI 对话、AI 辅写、趋势分析、智能推荐等,下面主要对前三项核心功能进行介绍。

(1) AI 检索助手

Aminer 支持自然语言的提问检索,在检索结果页面右侧提供了 AI 检索助手,可查看当前检索结果的综述回答,并支持继续提出问题进行交流,AI 检索助手根据当前页面检索内容进行回答。

在检索结果文献处,Aminer 对每篇文献提供了“ChatPaper”功能,利用大模型技术从文献中智能抽取总结文章核心内容,如要点、创新点、方法、实验等信息,在不阅读全文情况下能对论文内容有大致了解。

(2) AI 对话

Aminer 提供 AI 对话功能,依据知识库给出回答和依据。在对话范围上,可选择 Aminer 库,也可指定私有文献库/文件夹进行对话。在回答中,Aminer 依据知识库中相关的多篇论文进行回答,并附有参考引用标识和参考文献,支持内容溯源。回答结束后,为用户推荐问题,辅助用户更进一步地探索交流。

此外,Aminer 还提供单篇论文的 AI 对话功能,用户可点击文献右下方的“对话这篇论文”

或上传 PDF 文档进行对话,问答助手提供论文的总结概述,并且推荐论文相关问题,用户可点击提问或自己输入问题对文章进行提问。此外,还提供了论文评审功能、推荐相似论文功能。

(3) AI 辅写

Aminer 提供 AI 辅助写作功能,一是用户可以导入文件使用 AI 写作工具优化论文或借助 AI 进行创作,AI 提供翻译、润色、解释、精简、扩写和纠错等功能;二是提供文献综述功能,用户描述需求和想法,系统将从知识库中检索相关论文生成综述。在生成内容的风格上,支持用户选择简述或深入,并提供参考文献。

2.8 星火科研助手

星火科研助手是中国科学院文献情报中心与科大讯飞股份有限公司合作,共同研发的科技文献智能服务平台^[18]。2023 年 10 月 24 日,在第六届世界声博会暨 2023 科大讯飞全球 1024 开发者节的主论坛上,星火科研助手正式发布。

星火科研助手提供成果调研、论文研读、学术写作三大核心功能,旨在为科研工作者打造高效精准的科技文献服务,助力科研工作。

(1) 成果调研

成果调研功能旨在帮助用户快速了解某一研究方向的论文、学者和研究机构,为用户提供了综述生成、检索结果智能分析和智能问答功能。

用户检索之后,页面左侧是相关文献,右侧是调研助手,对检索结果生成概要总结,包括这些论文数量和时间分布、主题、未来研究趋势等,也支持在对话框中提问进行调研。此外,提供多篇文献综述功能,目前最多支持对 30 篇

选中的论文生成综述。

(2) 论文研读

星火科研助手支持用户上传论文进行交互式问答。完成论文上传后，系统会快速总结文章主要内容，包括摘要、方法、结论等；在阅读过程中支持用户对论文选中片段进行翻译、摘要或询问；用户也可以在对话框中询问论文外的知识内容。

此外，星火科研助手还提供了多文档对比功能，用户可以上传多篇文档（目前支持 2~5 篇论文），系统会给出每篇论文的摘要和贡献，并通过智能算法生成对比分析表，比较相同点和不同点，在对比分析表中展示每篇论文的标题、方法和优点，使用户一目了然。

(3) 学术写作

星火科研助手的学术写作功能提供学术翻译和英文润色，辅助用户更轻松地进行学术论文写作任务。学术翻译功能支持 docx 格式文档的上传和输出结果的下载。在语种上，目前支持中文和英文互译。

2.9 NSTL 创新服务助手

国家科技图书文献中心（National Science and Technology Library, NSTL）以构建数字时代的国家科技文献资源战略保障服务体系为宗旨，面向全国提供公益性、普惠性文献信息服务。

NSTL 持续完善网络服务系统建设，不断提升科技文献智能化服务水平，2024 年 5 月 22 日，NSTL 上线了 AI 智能应用“创新服务助手” V1.0 版本，重点围绕智能学术文献服务典型应用场景提供服务，包括 AI 检索、AI 问答和 AI 辅助阅读功能。

(1) AI 检索

NSTL 在传统检索的基础上，新增了 AI 检索功能，主要解决海量文献资源下，查找难、筛选难、进展跟踪分析难等问题。主要提供了自然语言检索、一句话摘要、文章脉络、智算综析等功能。

AI 检索支持用户通过使用关键词或自然语言提问的方式开展检索，检索结果页面左侧是相关文献，并利用 AI 为每篇文献生成一句简洁的中文摘要描述，同时还提供了“文章脉络”功能，以思维导图形式展示文章内容要点，如研究背景、目的、方法、结果、结论、下一步方向等，帮助用户快速了解、筛选和利用。页面右侧提供了“智算综析”功能，结合用户的问题，利用 AI 对左侧的相关文献进行智能总结与分析，辅助用户更高效地理解和吸收批量文献内容。

(2) AI 对话

NSTL 提供 AI 对话功能，旨在依据 NSTL 高质量文献资源，在一问一答的自然交互中，为用户提供来源可追溯可循证的知识答案，辅助用户研究探索。

在用户提问之后，AI 依据 NSTL 文献知识库的内容实现智能总结回答，并在页面右侧展示来源文献，点击回答下方的“文献溯源”可查看来源文献的核心内容要点。最后，在回答下方利用 AI 推荐相关问题，辅助和激发用户新的思考和更深入的探索。

(3) AI 辅助阅读

NSTL 的资源以外文文献为主，AI 辅助阅读旨在帮助用户更轻松、快速阅读和理解文献，同时也支持用户上传本地 PDF 文件进行辅助阅

读,主要包括智能综述、文章脉络、文章问答、全文翻译等功能。

阅读文献时,NSTL利用AI技术分析全文,总结文章核心内容和主要见解;梳理文章脉络图,即研究目的、方法、结果、结论、下一步研究方向等;支持与文献对话,用户提问后,系统依据当前文献内容回答,并提供文章内容循证;提供实时的全文翻译、划线翻译和解释等功能。此外,系统能够支持多个语种的外文文献辅助阅读,极大提高了文献阅读效率。

3 学术文献服务平台智能化应用分析

本章节主要从国内外的主要学术文献服务平台的智能化应用的整体情况、智能化应用分析以及面向科研过程各阶段的科研需求支撑分析三个方面进行梳理总结和分析。

3.1 智能化应用整体情况分析

目前国内外的主要学术文献服务平台的智能化应用归纳总结后的整体情况如表1所示(调研数据截至2024年7月2日)。

表1 国内外主要学术文献服务平台智能化应用情况

系统名称	链接地址	机构	智能应用
Elicit	https://elicit.com/	美国—Elicit公司	自然语言检索、简短综述、自定义文献要点信息抽取、PDF上传抽取数据、相关概念结果及出处……
Semantic Scholar	https://www.semanticscholar.org/	美国—艾伦人工智能研究所	智能摘要、文献图表信息抽取、单篇文献问答、引文卡片、关键要点速读、AI生成主题页面……
Scispace	https://typeset.io/	美国—Typeset公司	自然语言检索、简短综述、自定义文献要点信息抽取、单篇文献对话、答案溯源、PDF上传解读、图/表/公式的理解解释、多种语言个性化设置
国外 Scopus AI	https://www.scopus.com/search/form.uri?display=basic#scopus-ai	荷兰—Elsevier公司	智能摘要、知识溯源、推荐问题、知识脉络思维导图、定位核心文献与领域专家
Scite	https://scite.ai	美国—Scite公司	全库对话、知识溯源、引文卡片、上传文献对话、自定义设置回答风格/论文来源范围等、引文检索、语义引文分类筛选、引文可视化……
Consensus	https://consensus.app/search/	美国—Consensus公司	自然语言检索、智能概述、是否问题的智能计量、一句话摘要、研究快照、问题推荐、指令问答、知识溯源、多种语言个性化设置
Aminer	https://www.aminer.cn/	清华大学—唐杰团队	自然语言检索、文章要点提取、AI对话、上传PDF、论文评审、相似论文推荐、辅助写作、文献综述……
国内 星火科研助手	https://paper.iflytek.com/	科大讯飞公司	自然语言检索、检索结果智能分析、智能问答、文献综述、论文研读、文章要点提取、多文档对比、学术翻译、英文润色……
NSTL	https://www.nstl.gov.cn/	NSTL	自然语言检索、综述分析、一句话摘要、文章脉络、文章要点提取、AI对话、问题推荐、AI辅助阅读(支持用户上传PDF)……

由表1可看出,在学术文献服务平台中,目前利用大模型技术赋能较为普遍的应用主要包括:自然语言检索、文章要点提取、智能摘要/总结、文献综述、智能对话问答、知识溯源等。也有一些具有特色的应用,如Scopus AI的知识脉络思维导图、Scite的语义引文、引文卡片与引文可视化、Consensus对是/否问题的智能计量、星火科研助手的多文档对比、NSTL的文章脉络、全文翻译等。

从服务模式和场景来看,现有应用一方面利用大模型新技术提升原有知识服务能力,如智能文章要点提取、寻找主题概念、学术翻译、智能语义引文、智能推荐等;另一方面在AI赋能下涌现出一些新形式和新应用,创新文献及知识服务模式,如自然语言检索、智能对话问答、知识循证、智能辅助阅读、智能摘要/总结、综述生成、智能辅助写作等。

从知识服务的粒度来看,在AI技术赋能下极大增强了细粒度知识的抽取能力,使之能够提供更细粒度的知识服务,如文章要点、图片、表格等信息的智能抽取,知识到文章片段的循证溯源,支持研究类型、研究细节、章节、引用立场等更细粒度的文献筛选等。

从文献服务的友好性来看,大模型为科技文献知识服务带来了智能便捷的新体验,更通过个性化的服务支持满足了科研用户多元化的需求。如自然语言的便捷检索,文献详细信息的个性化定制抽取,支持用户上传PDF的解读和对话,支持不同语言的交互,自定义语言设置、生成内容风格等。

3.2 大模型赋能下的智能化应用特点与分析

在学术文献服务平台中,大模型技术的应用极大地提升了用户获取、理解和利用学术资源的效率,以下聚焦典型应用进行分析。

自然语言检索与智能对话问答,作为大模型技术深度融入文献服务领域所催生的新型服务模式,显著区别于传统学术文献检索中基于关键词的查询方式。这主要依托于大模型卓越的自然语言理解力、跨语言处理能力以及深度语义分析能力,这使得用户可以用更自然、接近日常表达的方式提出问题进行搜索,让用户的检索输入更简单。同时,大模型能够更好地理解用户查询的完整语义与上下文,更有效地匹配用户需求与文献内容,从而使用户能够以更快的速度定位并获取到所需的学术资源,提升了文献检索的效率与结果的精准性。

文章要点提取、智能摘要、文献综述与智能辅助阅读等功能,通过高效利用大模型对文本信息以及图片表格等多模态信息的深度解析与快速提炼优势能力,能够迅速捕捉并吸收文献中的关键要点信息,进而实现信息的精准提炼与高度概括。在文献资源的开发利用与快速信息抽取/提炼场景中,这些功能不仅显著提升了信息处理和知识吸收的效率,还能够为研究人员提供强有力的辅助,促进了学术研究 with 知识创新的进程。

知识溯源与循证等功能,主要针对大模型在生成内容方面的幻觉现象及可解释性缺陷,通过知识增强和可信度验证等技术,确保生成内容所引用知识的来源可靠、可信、路径清晰,

以适配科研领域对严谨性与专业性的要求。

智能写作、学术润色等功能，主要利用大模型生成能力与创作能力，结合知识库学习与人机协同机制等，辅助科研人员高效生成和优化文本内容。这些是人工智能技术发展下新兴的创新工具，拓展了传统学术文献服务的边界

与范畴。

通过智能化应用现状分析发现，相较于传统学术文献检索与服务，大模型技术赋能下的智能化应用在用户输入、信息匹配、结果呈现、交互体验、智能推荐、个性化服务等多个方面具有优势，如表 2 所示。

表 2 学术文献服务智能化应用与传统服务对比分析

对比维度	大模型赋能下的学术文献智能化应用服务	传统学术文献服务
用户输入	关键词或自然语言表达的提问	关键词或条件组合表达式
信息匹配	大模型技术赋能下，能够深入理解用户查询意图和文献内容语义，实现更精准的语义匹配	基于输入的关键词或查询条件构建查询语句
结果呈现内容	基于细粒度知识的智能抽取，提供文献智能摘要、文章要点（目的、方法等）、多模态详细信息（表格、图片、概念等），以及智能综述、对比分析等	文献元数据的基本字段
结果呈现形式	支持个性化定制的有组织表格形式，以及知识脉络、思维导图等可视化的呈现形式，直观清晰，一目了然	文献列表、文本信息
交互体验	对话式自然语言交互，提供更加流畅、自然的交互体验	多次修改关键词进行检索的被动式交互
智能推荐	在用户历史行为及兴趣偏好基础上，根据用户数据实时动态调整，实现个性化智能推荐；除文献外可推荐问题、专家等内容	根据用户查询及行为进行文献推荐
辅助阅读	实时生成文章概述、文章要点、文章翻译等信息，并提供文章问答、术语解释、引文卡片等智能服务	无
辅助写作	提供翻译、润色、解释、精简、扩写、纠错、综述生成等功能	无
个性化服务	用户上传 PDF 解读、结果呈现定制、语种及回答风格设置等	无

3.3 面向科研全流程的科研需求支撑分析

AI 赋能的目的是融入科研、嵌入科研、服务科研，为了分析目前智能应用对科研全流程的覆盖和支撑程度，进一步对目前各类 AI 功能按照科研阶段进行归纳分析，为 AI 在学术发现领域提供真正服务科研的高质量知识服务提供分析参考。

科研过程，即选题 / 创意、立项 / 规划、开始 / 计划、项目实施、成果产出、成果评价等阶段^[19]。将上述智能应用归纳为自然语言检索、智能对话问答、智能辅助阅读、智能综述、知识脉络、智能语义引文、智能

推荐、智能辅助写作共 8 类，分析情况如图 1 所示。

由图 1 可看出，目前的智能应用相对集中在科研过程的前几个阶段的需求支撑。如自然语言检索、智能语义引文在一定程度上解决了传统学术发现过程对海量科技文献查找难、筛选难的问题；智能对话问答、智能综述知识脉络等在一定程度上辅助用户解决研究进展跟踪分析难、总结归纳不易、认知负担重等问题；智能辅助阅读解决了用户在文献阅读特别是外文文献阅读过程中的吸收效率低、重点要点梳理费力等问题。相对薄

弱部分在于研究实施过程中除文献资源外的其他科研数据、工具、平台的关联分析获取，

论文投稿，文献引用，成果评价等内容的智能服务支撑。

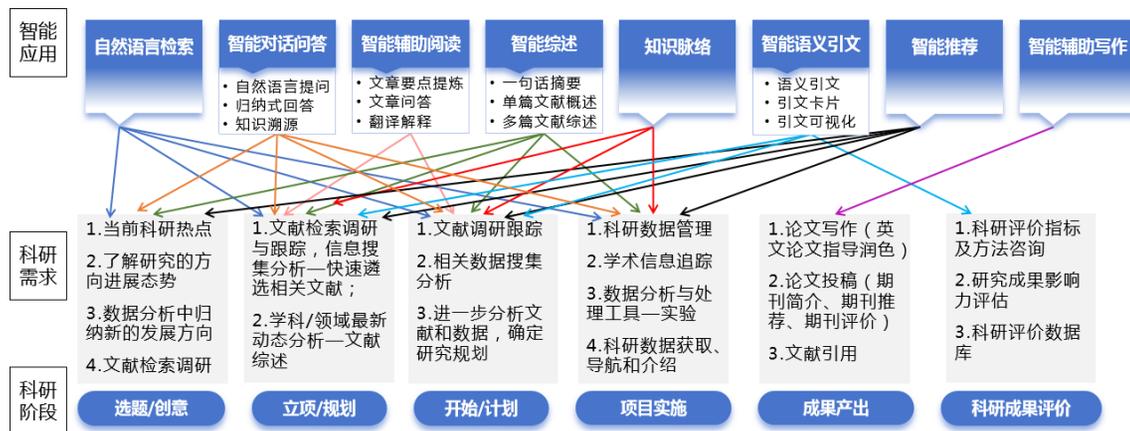


图1 面向科研过程各阶段的智能应用分析

3.4 发展趋势与展望

大模型技术在科技文献的检索发现和知识服务中已经取得了一定的应用落地成果，并且还有巨大的融合空间，如“人—机—智”深度融合的高效交互、精细精准以及多模态的知识服务、个性化高价值知识创新服务构建等是各方面需要进一步发展和强化的重要内容和趋势，大模型也将在其中发挥越来越重要的作用。

(1) 大模型技术为文献检索和知识服务带来全新的信息组织和检索利用方式，驱动文献信息服务向“人—机—智”深度融合的高效交互模式迈进。

大模型作为生成式语言模型，具有强大的语言生成和语义理解能力，①用户可以使用自然语言表达需求，模型能够更好地理解用户复杂的检索需求和意图，提供具有高度相关性的检索结果；②大模型的多语言理解能力打破了传统信息检索不同语种之间的壁垒，很好地实现跨语言检索；③对话式的问答搜索和知识获取成为大多数智能应用落地场景，提高了知识

获取的效率，以及自定义个性化的问答模式革新了信息检索的用户体验。自然语言的检索与智能问答是大模型与文献检索服务相结合的新型智能服务模式，更像是一个具有丰富知识的“人”，能够“读懂”用户需求，“了解”海量文献资源，及时“回答”相关信息和知识，已成为智能化文献知识服务的必然趋势。

从目前的文献发现等智能应用看，未来的趋势将围绕“人—机—智”深度融合的高效交互这一核心，不断深化技术融合，优化用户体验，特别是强化多轮对话机制、逐步追问下精准服务能力以及细粒度精准知识循证保障。具体来说，一方面，可以针对科技文献用户的复杂需求和交互信息反馈，研究大模型驱动的需求智能感知、意图精准识别、人机需求对齐等技术，有效处理用户需求，使得服务更加符合用户的期望；另一方面，攻关幻觉消除和可信可靠增强技术，使得生成内容具备可靠推理和可靠循证的特征，确保知识服务的严谨、可靠、可信、可追踪，提升知识服务的质量。

(2) 大模型赋能下,从科技文献中提取细粒度知识的能力大幅提升,推动知识服务向精细化、精准化、多模态方向发展。

大模型技术的应用极大增强了从文献中自动识别、解析并抽取深层次、细粒度知识单元的能力,促进了文献内容的结构化、精细化知识组织和深度揭示,更充分地挖掘和释放文献资源的宝藏价值,如文中图片、表格、公式等结构化内容,以及研究目的、方法、框架、结果、结论等关键点。相较于传统文献服务,实现了文献资源的“内容增厚”,更成为精细化知识服务模式转型升级的关键驱动力,支撑科研工作者更加便捷地获取到直接可用的内容或答案。

从目前大模型在文献知识提取领域的应用现状来看,未来的发展趋势将继续深化文献知识智能抽取的粒度、准确度以及不同模态信息的深度挖掘和关联,向文本+表格+图像+多维度知识库的多模态统一智能服务框架发展。具体来说,一方面,攻关多模态信息抽取的关键技术研发,充分挖掘文本中不同模态的隐性知识;另一方面,在应用场景的构建上,为用户提供多模态的知识呈现方式,丰富知识展示的维度和深度,如化学分子式的精准识别和理解,以及与其特性描述(文本、数据)、结构图(图片)等多元信息的关联整合,形成多元的综合性的知识网络,满足用户在复杂信息环境下的多元化知识需求及对“立体”知识的高效理解吸收。

(3) 大模型技术成为推动学术文献智能化知识服务转型发展的新引擎,催生更多个性化、高价值知识服务的创新应用。

基于大模型的人工智能技术在语言理解、

信息匹配、文本生成、上下文学习、指令跟随、逻辑推理和知识承载等多方面的能力优势,为科研创新服务带来了新的机遇,加速催生了一系列具有创新性的知识服务产品与应用场景,如自然语言检索、智能对话问答、智能辅助阅读、智能摘要、智能综述生成、辅助写作等,帮助用户更高效地获取、阅读和利用文献知识。

从当前大模型在学术文献智能服务场景的应用实践来看,一是对科研全流程的需求支撑还不够全面,一定程度上已经支撑的能力还需进一步精细化提升;二是创新服务大多独立应用,与传统学术发现过程的融合嵌入不够。未来的发展可包括:①围绕科研全流程的用户需求和痛点,进一步加强和拓展创新应用场景的构建与优化;②将大模型新技术深度嵌入传统学术文献发现过程,在资源的深度挖掘增值、高效精准的文献服务以及用户体验等多个方面赋能支撑;③加强跨学科知识融合服务,帮助研究者快速发现不同学科之间的联系,促进跨学科合作和创新;④强化“千人千面”的个性化智能服务。

4 结语

当下大模型等人工智能新技术的快速发展,驱动各行各业的变革与发展,学术文献信息服务领域也有了不少的基于大模型的智能化应用落地实践。本文对国内外学术文献服务系统的智能化应用服务现状进行梳理,总结了国外的 Elicit、Semantic Scholar、Scispace、Scopus AI、Scite、Consensus,国内的 Aminer、星火科研助手、NSTL 各系统基于大语言模型

等人工智能技术开展的智能应用场景、功能和特点。

通过对智能化应用现状的梳理总结、与传统服务的对比分析以及面向科研过程各阶段的科研需求的支撑分析,把握学术文献信息服务智能化发展的现状与趋势,对下一阶段学术文献领域智能化服务进行展望:①不断推动深度技术融合,提升多轮对话下的精准服务能力,向“人—机—智”深度融合的高效交互发展;②文献细粒度、多维度、多模态知识内容的挖掘,打造文本+表格+图像+多维度知识库的细粒度多模态统一智能知识服务;③个性化、高价值知识创新服务场景与应用的拓展深化。

参考文献

- [1] OpenAI. ChatGPT: optimizing language models for dialogue[EB/OL]. (2022-11-30)[2023-07-21]. <https://openai.com/blog/chatgpt>. OpenAI.
- [2] 储节旺,杜秀秀,李佳轩.人工智能生成内容对智慧图书馆服务的冲击及应用展望[J].情报理论与实践,2023,46(5):6-13.
- [3] Official Microsoft Blog. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web[EB/OL]. (2023-02-07) [2024-03-05]. <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>.
- [4] 赵瑞雪,黄永文,马玮璐,等. ChatGPT对图书馆智能知识服务的启示与思考[J].农业图书情报学报,2023,35(1):29-38.
- [5] 刘炜.智慧图书馆十问[J].图书馆理论与实践,2022(3):1-6.
- [6] 初景利,任娇菡,王译晗.从数字图书馆到智慧图书馆[J].大学图书馆学报,2022,40(2):52-58.
- [7] 储节旺,吴蓉,李振延.数智赋能的创新生态系统构成及运行机制研究[J].情报理论与实践,2023,46(3):1-8.
- [8] 王晓光,夏凌颖,段青玉.学术阅读智慧化:学术论文在线阅读系统优化研究[J].出版广角,2021(13):16-20.
- [9] BUCHKREMER R, DEMUND A, EBENER S, et al. The application of artificial intelligence technologies as a substitute for reading and to support and enhance the authoring of scientific review articles[J]. IEEE Access, 2019(7): 65263-65276.
- [10] FAN Y. Interactive AI Virtual Teaching Resource Intelligent Recommendation Algorithm Based on Similarity Measurement on the Internet of Things Platform[J]. Journal of Testing and Evaluation, 2024, 52(3): 1650-1662.
- [11] GLICKMAN M, ZHANG Y. AI and generative AI for research discovery and summarization[J]. arXiv preprint arXiv:2401.06795, 2024.
- [12] SEMANTIC S. Semantic scholar - A free, AI-powered research tool for scientific literature [EB/OL]. (2015-11-02)[2023-02-10]. <https://www.semanticscholar.org/>.
- [13] FIORINI N, LEAMAN R, LIPMAN D, et al. How user intelligence is improving PubMed[J]. Nature biotechnology, 2018, 36(10): 937-945.
- [14] 谢智敏,郭倩玲.基于深度学习的学术搜索引擎——Semantic Scholar[J].情报杂志,2017,36(8):175-182.
- [15] 姚翔宇,李洁.语义引文的实践进展研究——以 Semantic Scholar 和 Scite 为例[J].图书情报导刊,2022,7(1):48-57.
- [16] CACHOLA I, LO K, COHAN A, et al. TLDR: Extreme summarization of scientific documents[J]. arxiv preprint arxiv: 2004.15011, 2020.
- [17] Consensus. Consensus Product & Feature Updates [EB/OL]. (2024-06-08)[2024-07-02]. <https://consensus.app/home/blog/consensus-product-feature-updates/>.
- [18] 张智雄.在开放科学和 AI 时代塑造新型学术交流模式[J].中国科技期刊研究,2024,35(5):561-567.
- [19] 肖珑,张春红.高校图书馆研究支持服务体系:理论与构建——兼述北京大学图书馆的相关实践[J].大学图书馆学报,2016,34(6):35-42.



开放科学
(资源服务)
标识码
(OSID)

融合 BERTopic 和大语言模型的研究前沿识别 ——以美国 NSF 人工智能领域资助为例

范旭辉 穆智蕊

中国工程院战略咨询中心 北京 100088

摘要: [目的/意义] 为了解决研究前沿识别中主题建模缺少语义化的主题表达、基于关键词和人工判别的主题命名较为主观、未考虑主题相关文档内容等问题,引入大语言模型对生成的研究主题进行语义增强,以提高研究前沿识别的准确性和客观性。[方法/过程] 首先梳理了研究前沿的相关概念,以及主要识别理论、方法,然后以基金项目为数据源,通过 BERTopic 进行主题识别,使用大语言模型进行主题命名,识别出了隐含在基金项目中的研究主题,提出了研究前沿测度指标体系,并使用 Critic 客观赋权法确定了指标权重。[局限] 大模型生成的主题短语部分内容语义较为模糊,未能使用专业领域内的术语来表达,且研究所使用的数据仅限于基金项目数据。[结果/结论] 以人工智能领域 NSF 资助的科研项目为例,识别出了机器人、机器学习算法、智能教育、数据管理、模拟仿真等研究前沿,通过与美国人工智能规划及技术评估和预测报告的内容进行对比后发现,识别出的研究前沿具有一定的合理性和前瞻性。

关键词: 研究前沿; 前沿识别; BERTopic; 大语言模型; 主题建模; 科研基金

中图分类号: G353.1; TP391

The Research Frontier Identification by Integrating BERTopic and LLMs: Taking the U.S. NSF Funding in the Field of Artificial Intelligence as an Example

FAN Xuhui MU Zhirui

Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China

Abstract: [Objective/Significance] To address the issues in research frontier identification, such as the lack of semantic representation in topic modeling, the subjectivity in topic naming based on keywords and manual judgment, and the neglect

基金项目 中国工程科技知识中心项目“数据资源采集与加工”(CKCEST-2023-1-3), 中国工程院战略研究与咨询项目“2024年度全球工程前沿研究”(2024-XBZD-20)。

作者简介 范旭辉(1989-), 硕士, 助理研究员, 主要研究方向为数据分析与挖掘、智库信息服务; 穆智蕊(1983-), 硕士, 副研究员, 主要研究方向为前沿技术识别, E-mail: mzm@cae.cn。

引用格式 范旭辉, 穆智蕊. 融合 BERTopic 和大语言模型的研究前沿识别——以美国 NSF 人工智能领域资助为例[J]. 情报工程, 2025, 11(1): 18-28.

of relevant document content, this study enhances the generated research topics using large language models to improve the accuracy and objectivity of frontier identification. [Methods/Processes] This paper first reviews the relevant concepts of research frontiers, as well as the main identification theories and methods. Then, using funded projects as the data source, BERTopic is applied for topic identification, and a large language model is used for topic naming. Hidden research topics within the funded projects are identified, and a research frontier measurement index system is proposed. The Critic objective weighting method is used to determine the index weights. [Limitations] Some of the topic phrases generated by the large language models are semantically vague and fail to utilize specialized terminology in the relevant field. Additionally, the data used in this study is limited to funded project data. [Results/Conclusions] Using NSF-funded research projects in the field of artificial intelligence as a case study, this paper identifies research frontiers such as robotics, machine learning algorithms, intelligent education, data management, and simulation. By comparing these identified frontiers with the content of the U.S. artificial intelligence planning, technology assessments, and forecasting reports, it is found that the identified research frontiers demonstrate a certain degree of rationality and foresight.

Keywords: Research Frontiers; Frontier Identification; BERTopic; Large Language Models; Topic Modeling; Research Funding

引言

基础科学的研究水平决定着一国科技发展和进步的程度，世界上主要国家都设立了专门的科学基金资助本国的基础科学研究^[1]。政府基金资助在科学发展中发挥着重要作用，并引起了学术界的广泛关注^[2]。国家科学基金资助的研究内容和方向通常具有前瞻性、先导性和探索性，可以挖掘出大量的研究前沿信息^[3]。

准确识别学科领域的研究前沿，是洞察学科发展现状，预见未来趋势的基础，也是科技政策研究、制定和实施的依据和先决条件。识别研究前沿的关键在于确保其时效性、准确性和代表性。相对于论文、专利等科研产出数据，基于对基金项目申报环节的数据分析，可以更早识别出研究前沿。研究显示，在同一领域内，学术论文的研究主题相较于基金项目平均存在 2 年的滞后^[4]；相对于科技政策文件的分析，经

过同行评议的基金项目更能体现研究前沿，且避免了政策文件在落实阶段的不确定性。基金资助的方向通常用于解决关键问题和瓶颈问题，且往往与科学研究的前沿领域紧密契合^[5]。因此，研究世界科技强国科技领域布局情况，不仅可以精准识别研究前沿，把握学科发展趋势，而且对国家科学、前瞻部署科技方向、抢占创新制高点有重要的意义^[6]。

当前，随着人工智能技术的快速发展，世界各国高度重视并将其视为战略性新兴产业，在基础研究和应用研究领域都投入了大量资金。美国国家科学基金会（National Science Foundation, NSF）也将人工智能研究作为资助的重点，持续投入以推进并引领人工智能技术和应用的创新发展。本研究以 NSF 资助的基金项目作为研究前沿识别的数据基础，并提出了一种结合主题建模和大语言模型的文本挖掘方法，实现对研究前沿的识别和分析，具有一定的代表性和合理性。

1 相关研究

1.1 研究前沿的内涵界定

目前,对研究前沿内涵的界定尚未达成统一共识。结合国内外学者的相关研究,可以将其概括为两类:(1)基于引用关系,通常指一组被引频率极高的文献或者其施引文献所涉及的主题^[7-8];由于文献引用数据存在一定滞后性,此种方法识别出的前沿不一定能准确反映最新的科研成果和共识,特点是在快速发展的领域,存在一定的局限性。(2)基于研究前沿的特征,将其界定为突发、新兴或近期高度关注的主题。本文采用第2种定义,即研究前沿代表了近期某研究领域内最新涌现、引发学界广泛关注并促使学者们竞相探索的研究课题,具备创新性、活跃性和导向性等特点。所谓创新性,指的是这些研究前沿代表了该领域中最前沿、最尖端的研究议题。活跃性是指该研究主题能够引起学术界的高度关注。导向性体现在能够指引学术界和产业界加大投入,激发领域内的激烈竞争,推动整个研究方向的发展和创新。

1.2 研究前沿识别方法

研究前沿的识别方法大体上可分为专家研判法、文献计量法和主题建模法3种类型。

专家研判法首先邀请多位专家提名本领域研究前沿,然后对专家提名结果进行整理、归纳,再通过多轮讨论迭代收敛,获得最终结果。文献计量法又可以分为共被引分析方法和文献耦合分析方法,共被引分析是通过检测两篇以上文献被共同引用的频率来识别研究前沿,文

献耦合分析是基于文章发表时所引用的参考文献的相似性来识别研究前沿。这两种分析方法都是基于文献的引用关系,而文献从发表到被引用需要一定的时间积累,因此这类方法在识别研究前沿时会存在一定时滞,实际应用中往往无法及时发现新兴的研究前沿。

1.2.1 主题建模

研究前沿源于新的科学发现、技术突破或者显著的研究进展,因此必然会导致关键词、主题等与研究内容相关的文本内容发生显著变化。因此,通过基于文本内容的分析和挖掘来识别研究前沿具有一定的可行性。主题建模(Topic Modeling)是最为有效的手段之一,最早应用于自然语言处理、信息检索等领域,是一种无监督机器学习技术,主要用于发现大量文本数据中的隐藏主题结构^[9]。

主题建模方法在识别研究前沿时包括两个关键任务:主题抽取和主题命名。

主题抽取研究方面,常见的算法有 c-value、LDA、Top2vec、Doc2vec 和 BERTopic 等。逯万辉^[10]对 LDA、Top2vec、BERTopic 这3种主题建模方法在科学文献主题识别方面的性能进行了深入的研究和对比分析,研究显示基于文本预训练技术的 Top2vec 模型和 BERTopic 模型,在主题稳定性和离散性这两个关键指标上,表现优于传统的主题建模方法。Contreras 等^[11]使用 LDA 和 BERTopic 对文本进行分析,认为 BERTopic 算法在主题的连贯性、多样性和可解释性上表现更为突出。Egger 等^[12]认为,在各类主题建模算法中,BERTopic 在文档嵌入、主题自动合并和 c-TF-IDF 处理等方面具有明显优势。

主题命名研究方面，基于注意力机制的生成式语言模型展现出显著的技术优势。传统的主题抽取方法是将抽取出来的关键词进行组合，加上专家对内涵的归纳解读形成主题，耗时费力且存在一定的主观偏误性^[13-15]。大模型显著提升了文本理解和生成能力、从语料中获取和编码知识的能力以及复杂任务的推理能力^[13]。有研究表明，在进行文本摘要生成等任务时，与传统的抽取式方法相比，使用生成式模型得到的结果具有更好的连贯性和可读性^[16]。因此，大语言模型成为文本挖掘环境下处理复杂语义和生成式任务的更好选择，不仅能够实现主题的精准语义化表征，还可充分挖掘主题相关文档的深层语义信息。

1.2.2 研究前沿识别指标体系

在以基金项目为数据源进行研究前沿的识别时，许多学者根据研究前沿的概念内涵，设计了多种测量指标以提高准确性。徐路路等^[17]采用动态主题模型（DTM）对 NSF 资助的项目数据进行了主题抽取，并基于资助时长、主题强度、资助强度等指标，对主题随时间的变化趋势进行了详细的分析；王效岳等^[3]基于 PLDA 模型从基金项目文本中提取研究主题，并通过结合主题的资助时间、资助强度以及中心性等指标来识别出前沿主题。刘博文等^[14]结合基金数据和论文数据的特点，提出了一套包括主题强度、主题新颖度和主题相似度的指标体系，以识别和判定前沿主题；范丽鹏等^[18]在现有前沿主题探测方法的基础上，考虑到基金项目的特点，创新性地融入资助趋势、资助强度指标来综合判别前沿型项目。

2 研究设计

2.1 研究思路

主题模型法识别研究前沿一般是得到多组主题词的集合，再通过人工识别优化为主题短语来表达完整的语义，此种方法具有较强的主观性，很大程度上是基于主题词的含义。

本文创新性提出使用大语言模型增强的主题表示方法。以基金项目为数据源，运用 BERTopic 进行主题识别，运用大语言模型进行主题命名，识别隐含的研究主题，并提出了研究前沿测度指标体系。不同于以论文、专利数据为基础的研究前沿识别研究，科研项目数据具有以下特点：缺少引用等关系揭示项目间关系；没有关键词，需要进行主题抽取；具有资助内容、资助金额等文本信息，可以结合主题抽取对科研投入方向和强度进行深入分析。科研项目中揭示研究内容的关键字段同科技文献类似，包括项目名称、关键词、摘要以及科研计划、资助类型等。

本研究的具体框架如图 1 所示。首先，完成 NSF 资助的科研项目数据的采集与筛选工作；其次，利用数据集中的英文标题和摘要进行预处理，包括词性还原、去停用词、词干提取等；再次，借助 BERTopic 模型抽取研究主题，通过分析主题相关词以及评估主题之间的距离确定研究主题词集合；从次，利用大语言模型对上一步识别出来的研究主题词集合进行主题命名，具体步骤为选择各主题下最相关的代表性文档，连同该主题的关键词一起构建提示词并输入大模型，获取输出的短语作为主题名称，以代替上一步中生成的

一组关键词；最后，结合科研项目资助的特点，从资助强度、资助趋势等指标，构建了研究前沿识别模型，将满足特征的主题作为研究前沿。为

比较前沿识别的效果，将本研究成果与政府发布的科技战略和政策、智库发布的科技前沿与趋势分析报告中人工智能相关内容进行了对比。

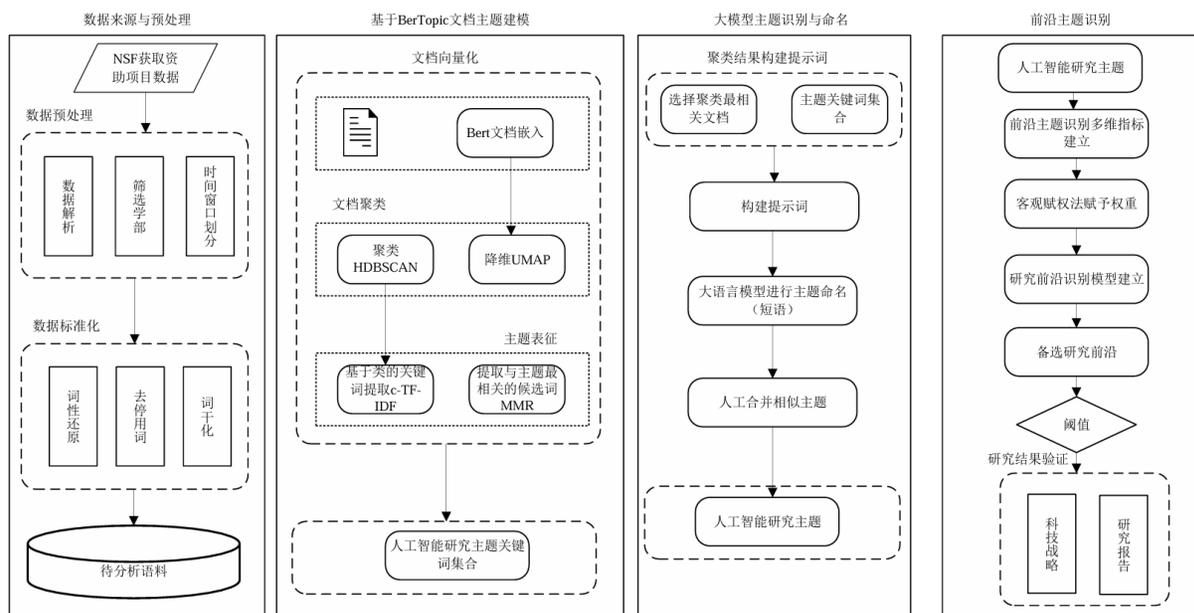


图1 研究框架

2.2 数据来源与处理

本文的研究数据为美国 NSF 在 2019—2023 年资助的人工智能项目。通过人工核查，剔除标题或者摘要为空的数据，对包含基金项目标题和摘要的文档进行去除停用词、同义词替换、词形转化等预处理，最终得到待研究数据集，共 2518 件，如表 1 所示。

表1 数据来源统计

年份	筛选前	筛选后	年份	筛选前	筛选后
2019	12975	683	2022	11457	473
2020	11301	501	2023	8423	348
2021	12651	513	合计	56807	2518

2.3 主题识别

由于科研项目数据没有正文，只有标题和

摘要包含与资助方向和内容相关的语义信息，其中摘要具有篇幅较短、内容复杂等特性，本文采用 BERTopic 对 NSF 资助的科研项目数据的标题和摘要进行主题抽取与分析，更好地挖掘相关项目的研究主题，发现其重点研究领域，把握其主题演变及发展趋势。主要流程包括：（1）文档向量化；（2）文档聚类，包括 UMAP 降维、HDBSCAN 聚类；（3）主题表征，包括基于类的关键词提取 c-TF-IDF、提取与主题最相关的候选词 MMR 算法。

具体流程及参数设置如下：（1）文档向量化：使用“all-MiniLM-L6-v2”的英文文本嵌入模型来将文档转换为向量表示，该模型能够有效地捕捉文本的语义信息。（2）使用 UMAP 降维：参考 Sánchez-Franco 等^[19]的研究，并通过反复

的实验验证，将相邻点数量（`n_neighbors`）设为 15，将投影后的维数（`n_components`）设为 2，使用余弦相似性衡量数据点之间的相似度，将点之间的最小距离（`min_dist`）调整到 0.01，以使嵌入更加紧密。（3）使用 HDBSCAN 聚类：将最小聚类规模（`min_cluster_size`）设为 10，将最小样本（`min_samples`）设为 5，即少于 5 个点的类簇会被视为噪声。（4）主题表征：将主题数（`nr_topics`）设为“auto”，以自动调整合并相似的主题。为了提高生成词组的质量，以发现结合较紧密的词组，调整 `n-gram` 窗口（`n_gram_range`）为 1~2。

2.4 主题表示

在主题识别阶段形成的每个主题都是以“编号 + 一组描述文档集合的关键词组”来表示。随着 NLP 技术的快速发展，尤其是大语言模型技术的突破，可以使用 GPT 等文本生成模型对主题进行微调，以得到更好的主题短语表示。

将 `c-TF-IDF` 生成的一组关键字视为候选主题，然后使用该主题的代表性文档进一步微调主题表示。通过实验确定每个主题代表性文档的数量，并使用候选关键字和文档构建提示词，传递到文本生成模型，并要求生成最适合该主题的输出。具体步骤如下：（1）选择文档，通过计算文档与主题表示的相似性（通过 `c-TF-IDF` 表示）来选择文档；根据反复实验，每个主题选择四个最佳匹配文档有较好效果。（2）构建提示词并优化，由于大语言模型有上下文窗口限制，将代表性文档内容按照长度截断后加入提示词中；本研究对比了“`en_core_web_sm`”“`KeyBERT`”“`GLM-4`”3 种主题表

示模型，最终采用大语言模型 `GLM-4` 进行主题命名，其中输入的单篇文档长度限定为 1000，温度设为 0.01，以确保生成的主题具有确定性。

（3）层次聚类与合并主题，本研究尝试使用 `topicGPT` 等方法进行分层主题建模，以便更好地组织和理解主题之间的关系，但效果不佳，再加上生成的主题数量不大，因此最终采用专家审定的方式，结合聚类结果合并相似的主题，生成总分结构的研究主题集合。

2.5 研究前沿识别

2.5.1 研究前沿识别指标体系

本文在识别出的研究主题基础上，按照研究前沿的定义，提出了主题新颖度、主题规模、主题趋势、主题强度等 4 个维度、5 个指标，用于表征研究主题的创新性、活跃性、动态性，构建了研究前沿识别模型。

（1）主题新颖度指标

新颖性是指一个研究领域最新涌现的研究主题。某个主题出现的年份越晚，说明该主题的新颖度越高。主题识别完成后，可以通过分析立项时间等信息，揭示每个主题的新颖性。计算方法为：

$$N_s = \sum_{i=1}^n y_i / n \quad (1)$$

式（1）中， N_s 表示主题 s 的新颖度； n 表示主题 s 内基金项目数量； y_i 表示第 i 个基金项目的批准立项年份与当前年的差值，该数值越小说明主题越新颖。

（2）主题规模指标

活跃性是指研究前沿应是能够激发学术界的广泛兴趣和高度关注的研究主题。利用近五

年主题内项目数量和金额衡量主题规模，直接统计即可。

(3) 主题趋势指标

研究前沿是正在兴起的主题，表现为近期主题的活跃度随时间而增加，具有一定的动态性、时效性。统计近五年主题内项目数量和金额的增长率的平均值，作为主题趋势的指标。

计算方法为：

$$Q_s = \frac{\sum_{i=2}^n (q_i - q_{i-1}) / q_{i-1}}{n-1} \quad (2)$$

$$M_s = \frac{\sum_{i=2}^n (m_i - m_{i-1}) / m_{i-1}}{n-1}$$

式(2)中， Q_s 、 M_s 分别表示主题 s 的数量和资助金额平均增长率； n 表示主题 s 内基金项目数量； q_i 表示 s 主题下基金项目在第 i 年份的数量， m_i 表示 s 主题下基金项目在第 i 年的金额。

(4) 主题强度指标

前沿主题在未来的持续发展会吸引更多投资，平均受资助金额可以反映研究主题的受资助力度。如果资助力度高，则说明在该研究主题下布局的单个项目的资助金额大，更能代表前沿。为消除资助时长和年份的影响，取近五年的年平均资助金额作为主题强度指标。计算方法为：

$$T_s = \frac{m}{n} \quad (3)$$

式(3)中， T_s 表示主题 s 的主题重要度； n 表示 s 主题内的基金项目的总量； m 表示 s 主题内的基金项目在近五年的受资助金额之和。

2.5.2 指标权重赋值

Critic法是一种客观赋权方法，它综合考虑

了指标内部的对比强度和指标之间的冲突性，同时考虑到指标内部变异性的尺寸，并兼顾了指标之间的关联性。该方法基于指标数据自身的统计特性进行评价，相较于熵权法和标准离差法，Critic法在处理指标权重赋值时更为精细和客观^[15]。

本文采用了Critic客观赋权法对指标体系进行权重赋值。具体实验步骤如下：(1)对原始决策矩阵进行无量纲化处理，以消除不同指标之间的量纲影响，得到标准化矩阵，本实验采用正向化处理，确保所有指标的趋势一致，便于比较和分析。(2)根据标准化矩阵，计算各指标的概率。(3)按照Critic法的客观赋权原理，计算各指标的权重，具体包括：a. 计算每个指标的平均值和标准差，以衡量指标的变异性；b. 计算指标之间的相关系数，以评估指标之间的冲突性；c. 根据指标变异性的尺寸和相关性的强弱，计算每个指标的权重。

3 实证研究

3.1 实验结果

3.1.1 研究主题识别结果

本次实验共识别出55个主题，经过人工审核，合并相似研究主题，最终形成20个研究主题，识别的研究主题如表2所示。可以看出，人工智能的研究前沿集中在数据管理等数据建设研究；深度学习、增强学习等算法研究；计算机视觉、语音、机器人、模拟仿真等应用研究；安全、隐私等技术与伦理研究；脑科学、医学等行业应用研究。

表 2 研究主题识别及人工合并结果

编号	一级主题	二级主题	主题短语 (LLM)	综合指数
1	机器人	自主机器人与人机交互	机器人与人机交互的进步 用于环境监测和群体机器人研究的先进无人机系统	100.00
		机器人部件	开发机器人踝足矫形器和假肢	
		水下机器人	使用机器人对有害藻类进行监测和应对	
2	计算机视觉	光学成像	计算机图形学和视觉中的微分绘制和动态光传输研究	41.92
		图神经网络	图神经网络的研究与开发	
		视频分析	从流媒体视频看计算机视觉和事件理解	
3	语音技术	语音合成	提高儿童学习成果的语音技术	32.51
		语音技术	人机交互和语音技术的研究	
4	数据管理	大数据管理	开放式数据保存和元数据管理系统 处理不确定性和大数据的高级数据管理和分析技术	56.06
		数据审核	内容审核研究与开发项目	
		数据可视化	数据可视化研究与开发	
		众包研究	主动学习与众包研究优化算法	
5	机器学习算法	深度学习	研究深度学习的可解释性和公平性 公民科学数据分析的智能支持系统	67.90
		大数据分析	数据分析和城市考古恢复的高级计算方法	
		数据挖掘	数据挖掘与知识发现	
		抽样估计	网络抽样与估计研究	
6	推荐系统	推荐系统	用于算法和界面实验的研究性新闻推荐基础设施的开发	10.64
7	信息检索	信息检索	高级搜索和信息检索技术	3.41
8	可解释 AI	因果推理	因果推理研究	18.00
9	虚拟现实	虚拟及增强现实	与人机交互和辅助技术相关的研究	28.89
		触觉技术	触觉技术与触觉研究	
10	模拟仿真	模拟仿真	使用并行计算和深度学习的大规模非线性可变形仿真技术的进展	52.09
		编程与软件开发	软件开发和编程接口增强研究	
11	大规模预训练模型	语言模型	语言技术与多语言处理	25.97
12	安全与伦理	技术伦理	人工智能伦理和多利益相关者决策 推进技术和数据科学伦理	44.46
		AI 公平	推进人工智能系统的公平	
		AI 安全	学习型自主系统的安全评估和风险管理	
13	AI 教育	AI 教育	机器学习研究和教育计划	57.77
14	脑科学	脑机接口	神经科学和网络优化研究	45.67
15	生物学	蛋白质功能预测	蛋白质功能预测研究	37.82

(续表)

编号	一级主题	二级主题	主题短语 (LLM)	综合指数
16	AI 新闻	新闻研究	通过人类技术合作推进新闻事业	42.53
17	医学与健康	流行病预测模型	传染病传播预测中的人工智能和数学建模	45.63
		医学数据分析	用于疾病理解和个性化医学的高级生物医学数据分析	
		疾病筛查	医学成像和深度学习在疾病检测和筛查中的应用	
		癌症与基因研究	癌症系统发育动力学和单细胞测序算法开发	
		非法药品供应	非法和假冒药品供应链中断研究	
		健康监测	睡眠医学中的睡眠表型感知和深度学习	
18	应急	灾害预测	人工智能和机器学习在环境科学和灾害预测中的应用	12.66
19	AI 制造	设计制造	用于定制和可访问产品开发的计算设计和优化进展	48.49
20	其他	AI 研究所	人工智能研发计划	23.32

3.1.2 多维研究前沿识别指标的计算结果

根据之前定义的研究前沿识别指标体系,在获得研究主题的基础上,先计算每个基金项目的资助时长、资助金额,然后计算 20 个研究主题的主题新颖度、项目数量、资助金额、项目数量增长率、资助金额增长率、主题强度等

6 个指标的值,通过 Critic 法确定的指标权重如表 3 所示,计算得到的综合指数见表 2。本研究识别出的研究前沿共有 20 个,其中前 10 个代表性前沿依次为:机器人、机器学习算法、AI 教育、数据管理、模拟仿真、AI 制造、脑科学、医学与健康、安全与伦理、AI 新闻。

表 3 Critic 法确定的评估指标权重

主题新颖度	项目数量	资助金额	项目数量增长率	资助金额增长率	主题强度
0.222874	0.16661	0.153197	0.135158	0.138261	0.183899

3.2 研究前沿识别结果与分析

根据研究前沿的识别结果,本研究通过文献调研与比对的方法,对比国家发布的人工智能领域战略规划和政策、智库机构发布的研究前沿识别报告,验证识别的研究前沿的准确性。

3.2.1 与科技战略及计划对比

2023 年 5 月,美国白宫公布了一系列围绕美国人工智能使用和发展的新举措并更新发布了《国家人工智能研发战略计划》。文件中提出要优先投资下一代人工智能以推动负责的

创新,包括推进基础人工智能能力(如:感知、表征、学习和推理),致力于开发更易使用和更可靠的人工智能以及评价和管理生成式人工智能相关的风险^[6]。

文件中与研究前沿相关的战略包括“战略 1:对基础和负责任的人工智能研究进行长期投资”和“战略 2:开发人—人工智能协作的有效方法”。其中战略 1 包含 10 个优先事项:推进以数据为中心的知识发现方法 (Topic_4 数据管理);促进联邦机器学习方法 (Topic_5 机器学习算法);了解人工智能的理论能力和局限性 (Topic_8 可解释性 AI、Topic_12 安全

与伦理)；开展可扩展的通用目的的人工智能系统研究 (Topic_11 大规模预训练模型)；在物理和虚拟环境中开发人工智能系统和仿真模拟 (Topic_10 模拟仿真)；增强人工智能系统的感知能力 (Topic_2 计算机视觉、Topic_3 语音技术)；开发功能更强大、更可靠的机器人 (Topic_1 机器人)；改进硬件以促进人工智能；创建改进硬件的人工智能系统；拥抱可持续的人工智能和计算系统。战略 2 涉及的研究领域包括成功的人智协作团队的属性和要求，评价人工智能协作应用的效率、有效性和性能的方法，以及降低人类滥用人工智能增强应用导致有害后果的风险 (Topic_12 安全与伦理)。可以看出，在识别出来的 20 个前沿方向中，有 10 个方向出现在战略文件中，表明识别的前沿具有较高的可信度。

3.2.2 与研究报告对比

3.2.2.1 斯坦福大学发布的《2024 年人工智能指数报告》

该报告对 2023 年 AI 技术的进展进行了全面总结，从高层次概述 AI 技术性能的整体演进开始，深入探讨了当前 AI 各领域的的能力，如语言处理、编程、计算机视觉、推理、音频处理、自动化智能体、机器人和强化学习等。详细介绍了过去一年中 AI 研究的重要突破，包括使用提示技术、优化和微调等方法来增强大语言模型，并最终探讨了 AI 系统对环境的影响。报告探讨了在负责任的 AI 领域内的几个关键趋势，包括隐私与数据管理、透明度与可解释性、安全性与保安以及公平性。

报告中涉及 Topic_1 机器人、Topic_5 机器学习算法、Topic_4 数据管理、Topic_12 安全与

伦理、Topic_2 计算机视觉、Topic_3 语音技术、Topic_11 大规模预训练模型、Topic_8 可解释性 AI 等 8 个识别出来的前沿方向，均包含在识别的研究前沿中。

3.2.2.2 欧洲研究委员会 (ERC) 发布的《预见：人工智能在科学过程中的使用和影响》报告

2023 年底 ERC 发布的该报告深入分析了其资助的 1048 个人工智能项目，主题包括：深度学习、机器学习、自然语言、计算机视觉、自适应和自治系统以及机器人。与报告相关的前沿主题包括 Topic_1 机器人、Topic_5 机器学习算法、Topic_2 计算机视觉、Topic_11 大规模预训练模型等 4 个前沿方向。

4 结语

围绕研究前沿识别这一研究主题，本文首先梳理了研究前沿的相关概念，以及主要识别理论、方法，然后以 NSF 基金项目为数据源，综合运用自然语言处理、主题识别、客观赋权、可视化分析等技术，识别出了隐含在基金项目中的研究主题，提出研究前沿测度指标体系，以人工智能领域为例，识别出了研究前沿主题，并与人工智能政策规划及智库研究报告的技术评估和预测内容进行了对比，显示了本文构建的前沿识别方法的可行性和有效性。

研究尚存在以下不足：

(1) BERTopic 主题模型获得的主题可解释性强，但其关键词分布倾向于高频词；使用大语言模型提取出来的主题短语，部分语义较为模糊，未能使用专业术语来表达，无法很好解释该主题内的主题词集合。

(2) 没有充分基于大语言模型来抽取和构建主题概念及其层次网络, 缺少主题抽取结果的对比和验证。

(3) 所利用的数据仅局限于 NSF 基金项目数据, 未能利用多元异构数据进行研究前沿的识别及测度。

在未来研究中, 将尝试使用大语言模型来进一步提取文本的关键词、摘要, 结合关键词抽取算法、专业领域术语来优化主题词的表达, 进而优化主题的语义表达。同时, 尝试使用监督学习的思路, 以代表性成果中的研究前沿作为目标构建训练集进行学习, 并考虑更好地评估前沿识别的效果。

参考文献

- [1] 孙海华, 张礼超. 美国国家科学基金会的重要资助举措及启示 [J]. 中国科学基金, 2021, 35(4): 663-671.
- [2] ZHOU P, CAI X, LYU X. An in-depth analysis of government funding and international collaboration in scientific research[J]. *Scientometrics*, 2020, 125(2): 1331-1347.
- [3] 王效岳, 刘自强, 白如江, 等. 基于基金项目数据的研究前沿主题探测方法 [J]. 图书情报工作, 2017, 61(13): 87-98.
- [4] 张雪, 张志强, 曹玲静, 等. 学科领域研究前沿识别方法研究进展 [J]. 图书情报工作, 2022, 66(12): 139-151.
- [5] 刘城, 廖林望, 丁倩. 基于美国国家基金数据的人工智能前沿分析 [J]. 中国高科技, 2021(4): 68-73, 88.
- [6] 张婧, 刘彦君, 张炜, 等. 基于科研项目数据的科技前沿识别有效路径实证探索 [J]. 科技管理研究, 2019, 39(16): 108-119.
- [7] PRICE D J. Networks of scientific papers[J]. *Science*, 1965, 149(3683): 510-515.
- [8] SMALL H. Co-citation in the scientific literature: A new measure of the relationship between two documents[J]. *Journal of the American Society for Information Science*, 1973, 24(4): 265-269.
- [9] FU X, HUANG K, SIDIROPOULOS N D, et al. Anchor-free correlated topic modeling[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(5): 1056-1071.
- [10] 逯万辉. 科学文献主题建模方法及其效果评估研究 [J]. 现代情报, 2024, 44(4): 22-31.
- [11] CONTRERAS K, VERBEL G, SANCHEZ J, et al. Using Topic Modelling for Analyzing Panamanian Parliamentary Proceedings with Neural and Statistical Methods[C]//2022 IEEE 40th Central America and Panama Convention (CONCAPAN). IEEE, 2022: 1-6.
- [12] EGGER R, YU J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts[J]. *Frontiers in Sociology*, 2022, 7(5): 1-16.
- [13] 张凯, 杨敏纳, 隗玲. 融合 Finetuned-BERTopic 和大模型的技术主题识别方法研究 [J]. 情报理论与实践, 2025, 48(3): 189-198.
- [14] 刘博文, 白如江, 周彦廷, 等. 基金项目数据和论文数据融合视角下科学研究前沿主题识别——以碳纳米管领域为例 [J]. 数据分析与知识发现, 2019, 3(8): 114-122.
- [15] 牛晓蓉. 基于政策文本和基金项目的研究前沿识别研究 [D]. 北京: 中国科学院大学, 2022.
- [16] 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展 [J]. 中国科学: 信息科学, 2023, 53(9): 1645-1687.
- [17] 徐路路, 王效岳, 白如江, 等. 基于 DTM 模型和文本特征分析的基金项目新兴趋势探测研究——以 NSF 石墨烯领域为例 [J]. 数据分析与知识发现, 2018(2): 87-97.
- [18] 范丽鹏, 王曰芬, 岑咏华, 等. 基金项目计划学部交叉及对前沿分布的影响研究——以美国 NSF 数据中 AI 领域为例 [J]. 情报学报, 2022, 41(9): 956-966.
- [19] SÁNCHEZ-FRANCO M J, CALVO-MORA A, PERIÁÑEZ-CRISTOBAL R. Clustering abstracts from the literature on Quality Management (1980–2020)[J]. *Total Quality Management & Business Excellence*, 2023, 34(7-8): 959-989.



开放科学
(资源服务)
标识码
(OSID)

基于大模型和知识图谱的电网检修计划智能编排

吴子辰 顾彬 张云翔 王莘然 高华

国网江苏省电力有限公司信息通信分公司 南京 210024

摘要: [目的/意义] 提出一种结合大语言模型和知识图谱技术的电网检修计划智能编排方法,以提高检修计划编排的效率和质量。[方法/过程] 利用大语言模型处理自然语言查询,并结合智能编排算法和图数据库技术,从电网公司诸多论文和专利中提取检修计划相关数据与总结编排规则,构建大规模问答数据集进行训练和验证。通过引入知识图谱技术,基于检修票据记录构建检修计划知识图谱,本方法能够更有效地表示和推理检修计划中的实体关系,从而提高实体识别和智能编排任务的性能。[结果/结论] 主要研究结论显示,相较于现有技术,所提方法在实体识别和智能编排任务中性能更优,准确率更高,且能显著减少人力成本投入,为电网检修计划智能化提供了有效解决方案。

关键词: 大语言模型; 知识图谱; 智能编排; 实体识别

中图分类号: G35; TP182

Intelligent Orchestration of Grid Maintenance Plan Based on Large Model and Knowledge Graph

WU Zicheng GU Bin ZHANG Yunxiang WANG Shenran GAO Hua

State Grid Jiangsu Electric Power Co., Ltd., Information & Telecommunication Branch, Nanjing 210024, China

Abstract: [Objective/Significance] This paper aims to propose an intelligent orchestration method for Grid Maintenance Plan that integrates large language models with knowledge graph technology to enhance the efficiency and quality of maintenance plan scheduling. [Methods/Processes] The research methodology involves using large language models to process natural language queries, combined with intelligent orchestration algorithms and graph database technology. This approach extracts data related to maintenance plans and summarizes orchestration rules from numerous patents of the power grid company to construct a large-scale question-answering dataset for training and validation. By introducing knowledge graph technology, this method constructs a maintenance plan knowledge graph based on historical maintenance ticket records, enabling more effective representation

基金项目 江苏省电力有限公司项目“基于数字孪生与自然语言处理的电力通信检修辅助决策关键技术研究”(J2023108)。

作者简介 吴子辰(1988-), 硕士, 高级工程师, 主要研究方向为电力通信运行与管理; 顾彬(1983-), 博士, 高级工程师, 主要研究方向为电力通信、人工智能; 张云翔(1978-), 硕士, 高级工程师, 主要研究方向为电力通信运行; 王莘然(1999-), 硕士, 主要研究方向为电力通信运行技术; 高华(1996-), 硕士, 主要研究方向为电力通信运行技术, E-mail: hua.gao@foxmail.com。

引用格式 吴子辰, 顾彬, 张云翔, 等. 基于大模型和知识图谱的电网检修计划智能编排[J]. 情报工程, 2025, 11(1): 29-41.

and reasoning of entity relationships within the maintenance plan, thereby improving the performance of entity recognition and intelligent orchestration tasks. [Results/Conclusions] The main research findings indicate that compared to existing technologies, the proposed method performs better in entity recognition and intelligent orchestration tasks, with higher accuracy, and significantly reduces the investment of human labor costs, providing an effective solution for the intelligent orchestration of grid maintenance plans.

Keywords: Large Language Model; Knowledge Graph; Intelligent Orchestration; Entity Recognition

引言

随着社会经济的快速发展和技术的不断进步, 智慧电网^[1-2]作为支撑现代社会运行的重要基础设施, 其稳定性和可靠性受到了前所未有的关注。电网检修计划的制定与执行, 直接关系到电网的运行安全和维护效率。传统的检修计划编排依赖于专业人员的经验和手动操作, 不仅效率低下, 而且容易受到人为因素的影响, 导致计划的不科学和资源的浪费。

在此情况下, 由于缺乏有效的冲突检测和资源分配机制, 检修计划之间常常出现冲突, 这导致电网运行中断和检修效率低下。此外, 随着电网规模的不断扩大和复杂性的增加, 传统的检修计划编排方法已难以满足现代电网的维护需求。因此, 如何实现电网检修计划的智能化、自动化编排, 已成为电力行业亟待解决的问题^[3-4]。

在图情领域, 信息的组织、管理和检索一直是研究的核心内容^[5-7]。近年来, 随着知识图谱和大数据分析技术的发展, 这些领域的研究成果为电网检修计划的智能化提供了新的视角和工具。知识图谱能够将电网检修的相关信息和检修实体间的关系进行结构化表示, 极大地提高了信息检索的准确性和效率。此外, 情报

学中的数据挖掘和分析方法也为检修计划的优化提供了科学依据^[8]。

与此同时, 人工智能技术的飞速发展, 特别是大模型技术在自然语言处理(NLP)领域的突破, 为电网检修计划的智能编排提供了新的思路 and 工具^[9-10]。利用大模型的强大语义理解能力, 结合图数据库中丰富的检修计划数据, 可以构建一个能够理解自然语言查询并自动生成最优检修计划的智能系统。这样的系统能够根据检修计划之间的潜在冲突和资源约束, 自动调整和优化计划安排, 从而提高检修工作的效率和质量^[11]。

本文提出了一种基于大模型和知识图谱的电网检修计划智能编排方法。该方法首先通过自然语言处理技术理解用户输入的问题, 然后利用检修计划冲突图数据库, 寻找最优的检修计划编排方案。通过这一过程, 系统能够有效地识别和解决检修计划间的冲突, 实现资源的合理分配和检修工作的高效执行。此外, 引入知识图谱技术, 基于过往大量电网检修票据记录构建检修计划知识图谱, 可以更加精确地表示检修计划中的实体和关系, 提供更加丰富的语义信息, 增强系统的推理能力, 从而在复杂场景下作出更加合理的决策。

在系统实现方面, 本文设计并实现了一个

综合考虑检修计划优先级和资源约束的智能编排模型。该模型以大模型为核心，通过深度学习技术对检修计划中的实体进行识别和链接，进而利用已有的冲突图，自动生成符合实际需求的检修计划安排。知识图谱的引入，使得系统能够更好地理解和处理检修计划中的复杂关系，提高智能编排的准确性和效率。

本文的工作有效缓解了电网检修计划存在冲突与资源分配不足的问题。该系统能够自动识别检修计划间的潜在冲突，并优化资源分配，减少电网运行中断，提升检修效率。随着电网规模的扩大，智能编排系统能够适应更复杂的电网结构，动态调整检修计划，确保电网维护的高效性和可靠性。

为了验证所提方法的有效性，本文调研了江苏省电力有限公司历史发表的论文以及申请的专利，搜集到与检修计划相关的5篇论文和82项专利，从中提取检修计划数据以及总结检修计划编排的相关规则，基于此构建了一个包含1000个涉及检修计划编排的问答对的数据集，并通过实验对比了大模型和其他方法在检修计划智能编排任务上的性能。实验结果表明，本文的方法在准确性和效率方面均优于现有技术，能够为电网检修计划的智能化提供强有力的支持。

1 相关工作

1.1 检修计划编排

电网检修计划是确保电力系统稳定运行的重要环节。随着人工智能技术的不断进步，电

网检修领域开始引入智能技术以提高检修工作的效率和智能化水平。研究者们探索了多种智能算法和模型，用于优化检修计划的制定和执行，以及提高故障诊断的准确性^[12]。

在这一背景下，知识图谱作为一种有效的数据组织和推理方法，被广泛应用于电网检修中。它通过整合和关联电网设备的各类信息，包括设备属性、检修历史和相互关系等，为复杂的检修决策提供支持^[13]。此外，概率模型和优化算法的应用，使得检修计划能够更加灵活地适应电网运行的变化，降低成本并提高可靠性^[14]。

1.2 大语言模型

大语言模型的应用，是近年来自然语言处理领域的重要突破。这些模型通过在海量文本数据上的预训练，学习到了丰富的语言表示，能够捕捉到语言的深层语义。大语言模型的引入，显著提升了机器对自然语言的理解和生成能力。著名的大语言模型，例如 ChatGPT^[15]，在多种自然语言处理任务上都展现出了强大的能力。然而也有研究指出，大模型存在难以自我纠正推理错误、对上下文的处理能力有限等一系列局限^[16-17]。

在电网检修计划智能编排中，大语言模型可以用于检修计划的自动生成、检修指令的语义理解以及用户复杂查询的解析等任务。本文在原有大模型的基础上，通过结合知识图谱技术来微调预训练模型，能够更有效地表示和推理检修计划中的实体关系，使大模型更好地适应电网检修领域的专业语境，从而提高实体识别和智能编排任务的性能。

1.3 知识库问答系统

知识库问答系统 (Knowledge Base Question Answering, KBQA) 是一种信息检索系统, 它能够直接回答用户基于知识库的自然语言查询。与传统的搜索引擎不同, KBQA 系统不仅检索信息, 还能理解用户的查询意图, 并提供精确的答案。前沿的问答系统, 如 RnG-KBQA^[18]、PanGu^[19]、Flexkbqa^[20] 等, 都在现有数据集上表现出了不错的效果。

知识库问答系统在电网检修计划智能编排中的应用主要体现在对检修计划数据的查询和分析。系统通过构建包含检修计划实体和关系的知识库, 能够准确理解用户的查询意图, 并从知识库中检索出准确的答案, 支持检修计划的决策制定。

1.4 知识库问答数据集

在知识库问答系统的构建中, 数据集的质量直接影响系统的性能。一个高质量的数据集应包括问题的自然语言表达形式、准确的答案以及清晰的推理路径。“一夜之间”(OVER-NIGHT) 方法^[21] 作为一种高效的数据集构建方法, 通过自动化和众包的结合, 实现了从知识库信息到自然语言问题的转换。

在构建过程中, 首先从知识库中提取信息并定义逻辑形式, 这些形式描述了如何从知识库中检索答案。基于这些逻辑形式, 生成一组规范的自然语言问句。众包技术被用于生成问句的多种变体, 不仅使数据集规模增加, 更提高了问答系统处理不同用户表达方式的能力。随着大语言模型的快速发展,

当前已可通过自动化方法替代众包技术, 生成问句变体, 以此降低人力成本, 并提高构建过程的效率^[22]。

对于电网检修计划的智能编排而言, 采用“一夜之间”方法构建的数据集可以提供多样化的用户查询表达和准确的检修计划信息。这使得系统能够理解各种复杂和非标准化的用户查询, 并提供及时准确的检修计划安排和冲突检测。通过这种方法构建的数据集, 可以训练出更加智能的问答系统, 从而提高电网检修计划编排的自动化和智能化水平。

2 智能编排数据集构建

本节详细介绍了智能编排数据集构建的过程, 该数据集包含 1000 个用于检修计划智能编排的〈自然语言提问, 查询语句〉问答对。

为了提取检修计划数据与总结检修计划编排规则, 本工作调研了江苏省电力有限公司历史发表的论文以及申请的专利, 最终保留下了与检修计划相关的 5 篇论文与 82 个专利, 从这些论文和专利中提取的原始数据涉及 567 个检修计划, 其中包含 3335 条检修光路, 6048 个检修业务。而与检修计划编排相关的规则, 则从对应论文和专利中总结了包含诸如主备光路冲突、OLP 冲突、环网冲突、业务冲突的冲突判断规则, 以及包含高等级计划优先、跨天计划优先、紧急检修最优的优先级判断规则等。这些规则将已提取的检修计划数据组织成检修计划冲突图数据库, 从而辅助构建检修计划智能编排数据集。数据集构建的流程如图 1 所示。

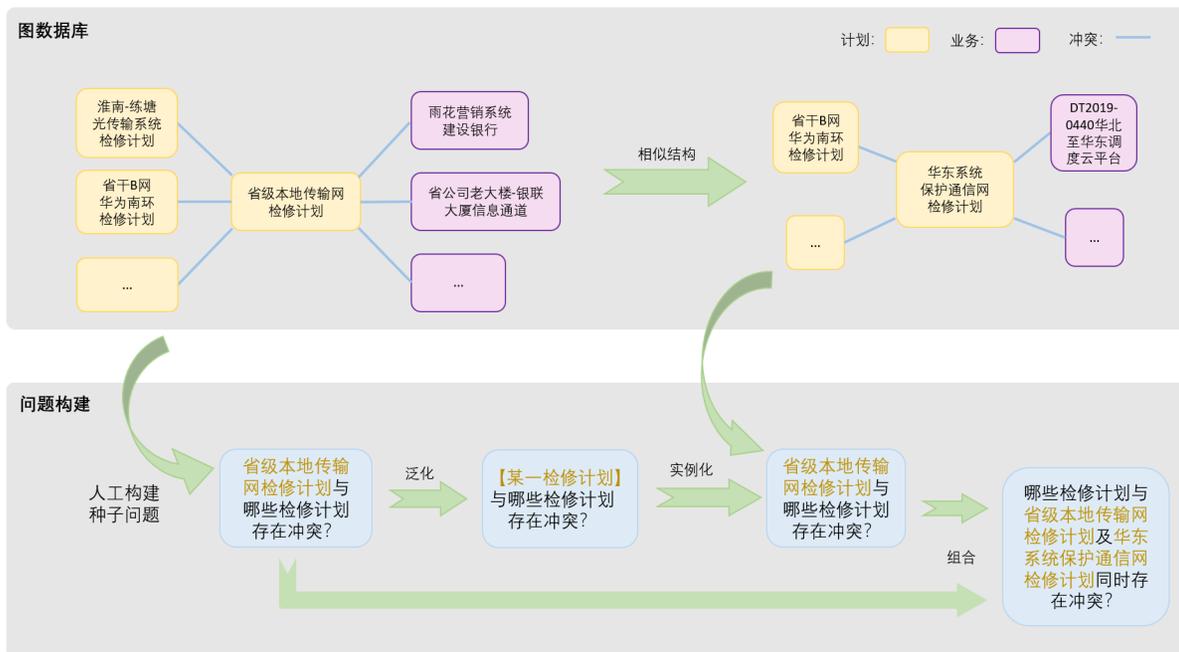


图 1 数据集构建流程

2.1 构造简单问题

为构建基础数据集，首先基于图数据库人工构造了 10 条简单问题。这些问题集中于识别单一检修计划与其他检修计划之间的直接冲突关系。

针对每个检修计划实体，设计了直接询问其冲突检修计划的问题。例如，针对“省级本地传输网检修计划”这一实体，提出了以下问题：“省级本地传输网检修计划与哪些检修计划存在冲突？”该问题的构造旨在确保仅涉及一跳的查询，即从选定的检修计划实体直接映射至与之冲突的其他检修计划实体。

为增强问题的多样性，采用了不同级别的检修计划实体进行问题构造，同时确保每个问题均基于图数据库中明确记录的冲突关系。通过这种方法，生成了一系列结构相似但实体不

同的问题，为数据集的初步构建提供了稳定的基础。

在问题构造完成后，进行了详尽的验证过程，以确保每个问题都能够得到准确的查询响应。验证过程包括对问题语义的审查以及对查询语句的测试，确保问题与数据库中的冲突记录相对应，且查询语句能够正确执行并返回预期结果。

通过这一过程，成功构建了一组简单、明确且可查询的问题，为数据集的进一步扩展和泛化奠定了坚实的基础。

2.2 问题泛化

继简单问题人工构造之后，本研究采取泛化策略以扩充数据集规模。泛化过程涉及将具体检修计划实体的描述抽象化，从而生成一系列新的问题。

泛化策略的核心在于将特定实体的描述泛化成更广泛的类别。例如，将“省级本地传输网检修计划”这一具体描述泛化为“某一特定检修计划”，以此生成更具普遍性的问题。通过这种方式，问题不再局限于特定的实体，而是能够涵盖更广泛的检修计划情况。

实施泛化时，针对每一类检修计划实体，定义了泛化模板。通过替换模板中的检修计划实体为更一般的描述，生成新的问题。例如，将“省级本地传输网检修计划”替换为“某一检修计划”，得到泛化后的问题：“某一检修计划与哪些检修计划存在冲突？”

泛化后的问题需经过验证，确保其能够映射至图数据库中的查询语句并返回正确答案。这一步骤是保证泛化问题准确性的关键。

通过泛化策略，原始的 10 条简单问题得以扩展成为 200 条更具普遍性的问题。这一过程不仅增加了数据集的规模，也提升了问题类型的多样性，为构建全面、高质量的数据集提供了坚实的基础。

2.3 问题组合

在问题泛化的基础上，本研究进一步通过问题组合策略来扩充数据集。该策略涉及将两个独立的泛化问题合并成一个复杂问题，以此探索更深层次的检修计划之间的冲突关系。

组合策略的核心在于识别具有逻辑关联性的泛化问题对，并通过添加适当的逻辑连接词，将两个问题合并成一个复杂问题。例如，将问题“省级本地传输网检修计划与哪些检修计划存在冲突？”与“华东系统保护通信网检修计划与哪些检修计划存在冲突？”进行组合，形成新

的问题：“哪些检修计划同时与省级本地传输网检修计划和华东系统保护通信网检修计划存在冲突？”

实施过程中，首先筛选出具有逻辑关联性的泛化问题对。然后，通过添加适当的逻辑连接词，将两个问题合并成一个新的复杂问题，合并后的问题需确保语义上的连贯性和查询上的可行性。

组合后的问题同样需要经过严格的验证过程。验证不仅包括问题语义的合理性，还包括查询语句的准确性和执行的有效性。通过这一步骤，确保组合问题能够正确反映图数据库中的复杂冲突关系，并且能够返回准确的答案。

通过问题组合策略，生成了 1000 条复杂问题。这一过程显著增加了数据集的深度和广度，为构建一个全面覆盖检修计划间复杂冲突关系的数据集提供了重要的支持。

2.4 数据集统计与分析

构建的数据集经过了详尽的统计与分析，以确保其质量和多样性。数据集最终包含 1000 条〈自然语言提问，查询语句〉问答对，覆盖了多种检修计划和业务场景。

在统计方面，数据集涵盖了 50 种不同的检修计划类型，每种类型对应 20 个问题，确保了数据集在检修计划方面的多样性。此外，数据集还涉及了 100 个不同的业务类型，每个业务类型关联 10 个问题，进一步增强了数据集的广泛性。

分析过程中，对问题的语言复杂度进行了评估，多跳占比达 70% 以上，这与数据集设计的初衷相符合。同时，对查询语句的复杂度也

进行了分析，结果显示查询语句主要涉及基础的图查询操作，如节点匹配和边查询，符合简单问题和复杂问题的设计要求。

为了确保数据集的质量，进行了人工审核。审核员对每个问题及其对应的查询语句进行了仔细检查，确保问题语义清晰、查询语句准确无误。审核结果显示，数据集中 98% 的问题和查询语句都是准确对应的，其余 2% 的问题经过修正后也达到了要求。

此外，还对数据集的可扩展性进行了分析。

结果表明，通过进一步的泛化和组合策略，数据集规模有望进一步扩大，以适应更多场景的需求。这表明，构建的数据集具有良好的多样性和准确性，为知识库问答系统的研究和应用提供了宝贵的资源。

3 智能编排系统

本节将阐述智能编排系统的构建过程，整体流程如图 2 所示。

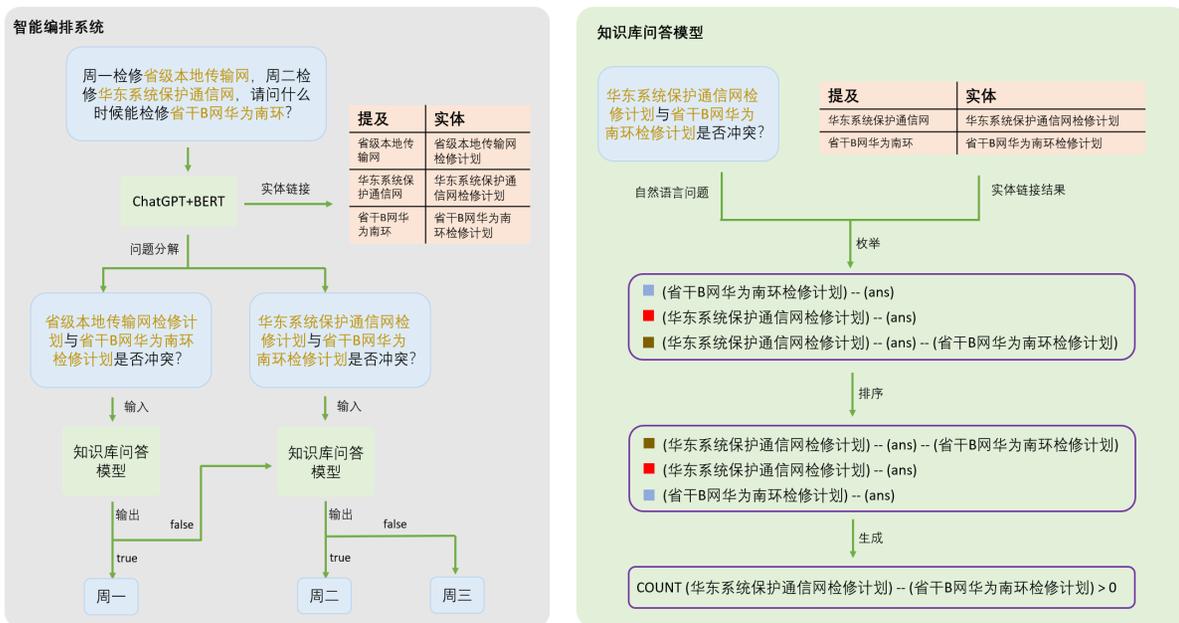


图 2 智能编排系统

3.1 问题分解与实体链接

在智能编排系统的初级阶段，问题分解与实体链接是至关重要的步骤。该步骤涉及将用户提出的自然语言问题用大模型进行分解，并同时利用小模型微调实体链接结果。这一流程的目的是将复杂问题拆解为可管理的子问题，并准确地识别出与问题相关的检修计划实体。

问题分解通过使用 ChatGPT 大模型，对用户输入的自然语言问题进行解析，旨在识别出问题中的关键信息并将其转化为明确的子问题。此过程对于后续的答案检索和精确的实体链接至关重要。大模型凭借其强大的语言处理能力，能够捕捉问题文本中的实体提及，并生成初步的上下文表示。

紧接着, BERT 模型^[23]作为小模型被用于微调初步的实体链接结果。BERT 模型经过预训练, 掌握了丰富的语言模式, 通过在特定数据集上的微调, 可以更准确地适应检修计划相关的实体链接任务^[24-25]。微调过程中, BERT 模型通过注意力机制聚焦于每个实体提及, 并捕捉到实体周围的上下文信息, 从而提高实体识别的准确性。

在实体消歧阶段, BERT 模型需要确定文本中提及的实体与知识库中的哪个实体是相同的。这一阶段的挑战在于可能存在大量同名或相似的实体, 模型需要根据上下文信息做出准确的判断。通过 BERT 模型的微调, 能够提高模型在特定数据集上的实体消歧能力, 从而选择最可能的实体作为链接结果^[26-27]。

通过问题分解与实体链接的流程, 智能编排系统能够准确地将用户的问题转化为具体的子问题, 并为知识库问答模型提供必要的上下文信息和精确的实体链接结果, 为生成最终答案奠定坚实的基础。这一流程不仅提高了系统的处理效率, 也确保了答案的准确性和可靠性。

3.2 知识图谱构建

知识图谱作为智能编排系统的核心组成部分, 其构建质量直接影响系统的推理能力和决策效果。本研究中的知识图谱构建基于丰富的电网检修计划历史数据, 包括近 600 张检修票据, 涵盖了超过 37000 个检修实体。这些实体不仅包括了检修计划的基本信息, 如计划编号、时间、地点和责任单位, 还详细记录了检修光路数等关键信息, 为检修计划之间的潜在冲突和资源约束提供了详尽的数据支持。

检修光路数作为电网检修计划中的一个重要参数, 涉及检修过程中需要中断的光路数量, 直接关系到检修对电网运行的影响程度。在知识图谱中, 检修光路数与其他检修计划实体的关系被精确表示, 使得系统能够全面评估检修计划的潜在影响, 并在智能编排过程中做出更加合理的决策。

为了构建高质量的知识图谱, 首先对历史数据进行了清洗和整合, 以确保数据的准确性和一致性。随后, 通过标准化处理, 将数据转换为图结构, 形成了包含节点(实体)和边(关系)的知识图谱。在这一过程中, 本文利用了自然语言处理技术提取实体和关系, 并采用了图数据库技术来存储和管理知识图谱, 以支持高效的图查询和推理操作^[28-30]。

知识图谱的构建不仅为智能编排系统提供了丰富的数据基础, 还通过实体链接和问题分解环节, 使得系统能够准确地理解和处理用户提出的复杂查询。此外, 知识图谱的引入还增强了系统的推理能力, 使得在复杂场景下能够作出更加合理的决策, 从而提高了智能编排的准确性和效率。

3.3 枚举图数据

在生成可能的查询候选项时, 对图数据库中的检修计划实体进行枚举是一种有效的策略。枚举过程旨在生成可能的查询候选项, 以供后续的排序和生成阶段使用。通过从每个检修计划实体出发, 查询其在知识库中的相邻实体, 并生成相关的查询语句, 可以覆盖多种可能的查询场景。这种枚举方法有助于提高问答系统在处理复杂查询时的鲁棒性。

一跳查询是查询图枚举的基础，它涉及从问题中的实体提及出发，查找与之直接相关的信息。例如，如果问题提到了“省级本地传输网检修计划”，一跳查询将直接检索该检修计划的属性以及与之相关联的实体，如检修计划的开始时间、结束时间或责任单位等。

随着查询复杂性的增加，可能需要进行两跳甚至多跳查询。两跳查询在一跳查询的基础上，增加了对间接关系的支持。例如，在了解“省级本地传输网检修计划”的基础上，可能需要查询其影响的业务类型。这时，查询不仅需要检索检修计划的直接信息，还需要进一步检索与检修计划相关的业务实体。具体到操作层面，可以通过连接检修计划顶点与业务顶点之间的边，来获取相关的业务信息。

在本研究中，查询图的枚举采用了一种迭代的方法。开始时，从问题中的实体提及出发，进行一跳查询，获取直接相关的信息。然后，根据问题的需要，逐步增加查询的跳数，直到获取到足够的信息来回答整个问题。在枚举过程中，为了提高效率，引入了剪枝策略，即在发现某些查询路径不可能得到正确答案时，及时停止这条路径的进一步查询。

此外，为了确保查询图的枚举能够覆盖所有可能的答案，本研究还采用了一种基于模式的扩展策略。通过定义一组可能的查询模式，系统能够自动地生成更多的查询候选项。这些模式基于先前类似问题的查询经验，能够指导系统更全面地探索图数据。

3.4 排序与生成

在问答系统中，排序和生成是两个核心步

骤。排序阶段根据问题的语义信息对所有可能的答案进行排序，选择最可能的答案作为候选。生成阶段则根据问题的语义信息和候选答案生成最终的查询语句。本研究中，为了提高问题回答的质量，采用了大模型 ChatGPT 进行初步排序，然后通过小模型 BERT 进行微调，以适应特定的数据集和查询需求。两者的协同工作显著提高了问答系统的性能。

排序模型的核心任务是从多个候选查询中识别出与自然语言问题最相关的查询。在本研究中，ChatGPT 模型被用作排序模型，因为它具有强大的语言理解能力和生成能力。模型接收由查询图枚举阶段生成的候选查询，并根据问题的内容和结构评估每个查询的相关性。使用注意力机制可以很好地聚焦于问题的关键部分。通过计算每个候选查询的置信度，ChatGPT 能够对查询进行排序，并将最有可能的候选排在最前面。

生成阶段的任务是将排序阶段筛选出的候选查询转化为可执行的查询语句。在本研究中，为了细化和优化 ChatGPT 生成的答案，采用了 BERT 模型进行微调。微调的目标是确保生成的查询语句在逻辑上与问题意图保持一致，并且能够在图数据库中执行以检索正确的答案。通过利用问题文本和候选查询中包含的模式信息，模型能够更具体地理解查询所需的上下文环境。

3.5 后处理

在智能编排系统的后处理阶段，主要任务是将子问题的解答整合，形成对原始查询的最终答案。该阶段侧重于依据逻辑顺序和既定规

则来确定最合适的答案，确保了处理过程的高效性。

具体来说，后处理步骤包括对先前由知识库问答模型生成的子问题答案进行评估，并基于这些答案以及预设的决策逻辑来形成最终的响应。例如，如果用户的原始问题涉及多个检修计划的冲突查询，系统将根据每个子问题的答案按照既定优先级和规则来决定最终的检修时间安排。

在整合答案的过程中，后处理机制确保了答案的准确性和一致性，同时提高了整个智能编排系统的响应效率。通过这种方式，系统能够直接为用户提供清晰、准确的最终答案，满足用户查询需求的同时，确保了检修计划的智能编排和优化。

综上，后处理阶段是智能编排系统中将子问题答案转化为最终答案的关键环节，通过简洁高效的处理流程，实现了对用户查询的快速响应和准确解答。

4 实验评测

4.1 实验环境

本文实验服务器配置了高性能的 NVIDIA RTX3090 显卡、Intel Xeon Gold 6326 CPU 以及 512GB 内存，确保了计算任务的高效执行。操作系统为 Ubuntu 18.04.3 LTS，提供了稳定的运行环境。编程语言采用 Python 3.6.3，深度学习框架则基于 PyTorch 1.10.2，这些工具的选用为模型的实现与训练提供了强大的支持。

4.2 消融实验

为了全面评估所提出的电网检修计划智能编排系统的性能，本文设计了一系列实验，分别测试了整个系统、仅使用大模型以及仅使用知识库问答模型 3 种情形下的表现，消融实验结果如表 1 所示。

表 1 消融实验结果

方法	精度	召回率	F1 值
仅使用知识库问答模型	0.527	0.467	0.495
仅使用大模型	0.613	0.546	0.578
完整系统问答	0.727	0.673	0.694

4.2.1 仅使用知识库问答模型

在这一测试中，选择直接将用户的问题输入到知识库问答模型中，不经过大模型的解析和分解。测试目的是评估知识库问答模型在没有大模型辅助情况下处理复杂查询的能力。结果显示，知识库问答模型在处理一些直接且明确的查询时表现良好，但在处理涉及多个检修计划和冲突检测的复杂查询时，性能受限，无法达到完整系统的处理水平。

4.2.2 仅使用大模型

在这一测试部分，选择仅利用大模型对用户的问题进行解析和分解，不涉及后续的知识库问答和后处理步骤。测试目的是评估大模型在独立处理复杂查询时的性能。结果表明，尽管大模型能够识别出问题中的关键实体并进行初步的分解，但在没有知识库问答模型支持的情况下，无法提供完整的答案。

4.2.3 完整系统问答

在这一测试中，选择利用完整的智能编排

系统处理用户查询。测试包括将用户的问题通过大模型进行解析和分解，然后使用知识库问答模型对分解后的子问题进行回答，最终通过后处理整合答案并给出最终响应。测试结果显示，该系统能够准确理解用户意图，并快速给出合理的检修计划安排，验证了系统的有效性和实用性。

4.3 错误分析

对电网检修计划智能编排系统在实验中回答错误的 50 个问题进行了详细分析，识别出四类主要错误及其原因占比，以供系统的进一步优化。

问题分解错误最为常见，占比 36%，这通常是由于大模型在解析复杂用户查询时产生的不精确性，导致后续处理基于错误的子问题进行。紧随其后的是实体链接错误，占比 28%，这类错误多发生在模型难以准确将问题中的实体与知识库中的实体匹配时，特别是面对专业术语和复杂的上下文环境。知识库问答模型错误占比 24%，这可能是由于知识库信息不全面或模型未能准确把握子问题意图。最后，后处理错误占比 12%，主要表现在答案整合阶段，系统有时会无法将子问题的答案恰当地合成针对原始问题的答复。

针对这些错误类型，未来可通过一些措施进行优化：增强大模型在问题分解阶段的准确性；扩充和更新知识库，以提供更全面的数据支持；改进实体链接算法，提高实体识别的准确度；以及优化后处理流程，确保答案的逻辑一致性和可读性。通过这些改进，期望能够显著提升系统的整体性能和用户满意度。

5 结束语

本文提出了一种基于大语言模型的电网检修计划智能编排方法，通过利用大模型的强大语义理解能力和先前提出的智能编排算法，实现了对电网检修计划的高效和智能化管理。通过构建大规模的电网检修计划问答数据集，并在此基础上进行实验验证，本文的方法在实体识别和智能编排任务中均展现出了优异的性能。

研究结果表明，本文提出的方法在处理电网检修计划的自然语言查询和编排任务时，不仅准确率更高，而且更加灵活、适应性更强。这主要得益于大语言模型在预训练阶段所获得的丰富语言知识和领域适应性，使其在处理专业领域文本时更具优势。此外，知识图谱的引入为系统提供了更加丰富的实体关系和语义信息，增强了系统的推理能力，使得在复杂场景下能够作出更加合理的决策。此外，本文的方法和理念也可以扩展到图情领域，为图书馆、档案馆和信息中心的信息组织、检索和服务提供智能化支持。例如，知识图谱可以用于构建图书和档案的语义网络，提高信息检索的准确性和效率；而大模型技术则可以用于理解和响应用户的复杂查询，提供更加个性化的服务。

通过自动化的编排流程，本文的方法显著提高了检修计划编排的效率和质量，降低了人力成本。这不仅为电网检修计划的智能化提供了一种有效的解决方案，也为电力行业的自动化和智能化进程做出了贡献。

未来将继续优化和完善本方法，探索更多的可能性。首先，将致力于进一步提升模型的性能，以适应动态变化的电网环境。这包括实

时更新模型以反映电网状态的变化,以及增强模型的泛化能力,使其能够处理不同规模和复杂度的电网系统。此外,计划引入更先进的机器学习算法和数据融合技术,以提高模型在处理大规模数据时的准确性和效率。

其次,考虑处理更复杂的检修场景,尤其是在多变量和不确定性较高的情况下,如何优化检修计划。这涉及开发新的算法来处理不确定性,例如通过引入概率模型和优化算法来评估和管理风险。还将研究如何利用历史数据和实时监测数据来预测潜在的电网故障,从而提前规划和调整检修计划。此外,计划探索多目标优化方法,以平衡检修成本、时间和质量等多方面因素。

最后,期望将本文的方法和理念扩展到图情领域,为相关研究者提供一定启发,将类似技术和方法应用于相关信息资源的管理和服务中,共同推动相关领域技术进步和创新发展。

本研究期望能够推动电网检修计划编排工作的自动化和智能化发展,为实现清洁、高效、可持续的能源供应贡献力量。随着技术的不断进步,电网检修计划的智能化将为电力系统的稳定运行和维护提供更加可靠和高效的解决方案。

参考文献

- [1] 高志远,姚建国,郭昆亚,等.智能电网对智慧城市的支撑作用研究[J].电力系统保护与控制,2015,43(11):148-153.
- [2] 李德仁,姚远,邵振峰.智慧城市中的大数据[J].武汉大学学报(信息科学版),2014,39(6):631-640.
- [3] 朱方,赵红光,刘增煌,等.大区电网互联对电力系统动态稳定性的影响[J].中国电机工程学报,2007,27(1):1-7.
- [4] 蔡帜,张加力,丁强,等.考虑两级协调和分区备用的多区域输电检修计划[J].智慧电力,2023,51(5):15-22.
- [5] 吴蝶,曹如中,熊鸿军,等.新文科建设背景下图书情报档案学科数字化转型发展研究[J].图书馆理论与实践,2023(2):12-16,26.
- [6] 张昱,吴振峰.面向科技情报服务的内容安全关键技术体系研究[J].数字图书馆论坛,2024,20(6):83-90.
- [7] 严驰.人工智能时代公共图书馆信息无障碍标准建设探索[J].数字图书馆论坛,2024,20(4):42-49.
- [8] 石栖,胡正银,王莉晓.基于LDA-WLC的图情领域学科交叉主题演化分析[J].数字图书馆论坛,2023,19(5):54-63.
- [9] 傅家伟,陈峥,杨佳圣,等.基于AI技术的电网业务智慧调度分配模型研究[J].粘接,2024,51(8):176-179.
- [10] 朱春宇.智能技术在电网智能检修中的应用[J].集成电路应用,2024,41(6):266-267.
- [11] 闻煜峰,赵世杰,赵玉成,等.电网设备异常处置智慧助手研究及应用[J].农村电气化,2024(8):53-57.
- [12] 凌亮,何强,熊正勇,等.基于多层迭代算法的电网检修计划优化方法[J].广东电力,2023,36(2):84-91.
- [13] 买买提·瓦热斯,王建军.基于检修价值评估的电容网检修优化决策方法[J].电工技术,2024(S1):378-380.
- [14] 郭子强,宋涛,郭杰,等.计及MCMC方法的电网概率检修计划指标优化模型[J].电子技术应用,2023,49(10):15-22.
- [15] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020(33):1877-1901.
- [16] HUANG J, CHEN X, MISHRA S, et al. Large language models cannot self-correct reasoning yet[J]. arXiv preprint arXiv:2310.01798, 2023.
- [17] LIU X, YU H, ZHANG H, et al. Agentbench: Evaluating llms as agents[J]. arXiv preprint arXiv:2308.03688, 2023.

- [18] YE X, YAVUZ S, HASHIMOTO K, et al. Rng-kbqa: Generation augmented iterative ranking for knowledge base question answering[J]. arXiv preprint arXiv:2109.08678, 2021.
- [19] GU Y, DENG X, SU Y. Don't Generate, Discriminate: A Proposal for Grounding Language Models to Real-World Environments[J]. arXiv preprint arXiv:2212.09736, 2022.
- [20] LI Z, FAN S, GU Y, et al. Flexkbqa: A flexible LLM-powered framework for few-shot knowledge base question answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(17): 18608-18616.
- [21] WANG Y, BERANT J, LIANG P. Building a semantic parser overnight[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1332-1342.
- [22] HUANG X, CHENG S, BAO Y, et al. MarkQA: A large scale KBQA dataset with numerical reasoning[J]. arXiv preprint arXiv:2310.15517, 2023.
- [23] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.
- [24] 田园, 原野, 刘海斌, 等. 基于 BERT 预训练语言模型的电网设备缺陷文本分类 [J]. 南京理工大学学报, 2020, 44(4): 446-453.
- [25] 孙玉芹, 肖静婷, 王海超. 基于多模型融合的电力运检命名实体识别 [J]. 科学技术与工程, 2024, 23(36): 15545-15552.
- [26] 潘正高. 基于规则和统计相结合的中文命名实体识别研究 [J]. 情报科学, 2012, 30(5): 708-712.
- [27] 何炎祥, 罗楚威, 胡彬尧. 基于 CRF 和规则相结合的地理命名实体识别方法 [J]. 计算机应用与软件, 2015, 32(1): 179-185.
- [28] 洪宝惜, 林泽斐. 基于社会资本理论的华侨华人知识图谱构建与应用 [J]. 数字图书馆论坛, 2024, 20(5): 35-47.
- [29] 胡汗林, 邓三鸿. 知识图谱在青铜器数字馆藏建设中的应用 [J]. 数字图书馆论坛, 2023, 19(4): 1-8.
- [30] 叶丁菱, 牟丽君, 许鑫. 基于知识图谱的中医古方交互式检索研究 [J]. 数字图书馆论坛, 2024, 20(2): 24-33.



开放科学
(资源服务)
标识码
(OSID)

融合专利属性和用户行为分析的 个性化专利推荐研究

赵学铭¹ 王刚²

1. 天津医科大学科学技术处 天津 300070;
2. 天津医科大学图书馆 天津 300070

摘要: [目的/意义] 提出一种融合专利属性信息和用户行为分析的个性化专利推荐方法, 以更好地满足用户的专利文献需求。[方法/过程] 首先量化用户对专利文献的检索与操作行为, 计算用户之间的相似性指数; 然后结合专利文献的属性信息(如 IPC 分类号、发明人等), 以实现个性化推荐。将该方法应用于自研的专利推荐系统中, 并与其他协同过滤推荐方法进行对比。[局限] 可能受到数据集的特定性限制, 如依赖于特定的用户行为和专利文献数据集。此外, 实验结果可能会因系统中的用户和专利文献选择不同而有所偏差。[结果/结论] 实验结果表明, 该个性化推荐方法优于传统的用户协同过滤推荐方法, 能够更好地理解用户的需求和偏好, 从而提高推荐的准确性与效率。

关键词: 用户行为分析; 协同过滤; 个性化推荐; 专利服务

中图分类号: G35; TP391

Research on Personalized Patent Recommendation by Integrating Patent Attributes and User Behavior Analysis

ZHAO Xueming¹ WANG Gang²

1. Science and Technology Office of Tianjin Medical University, Tianjin 300070, China;
2. Library of Tianjin Medical University, Tianjin 300070, China

Abstract: [Objective/Significance] This paper aims to propose a personalized patent recommendation method by integrating patent attribute information and user behavior analysis to better meet users' needs for patent literature. [Methods/Processes] The study first quantifies users' search and operation behaviors towards patent literature and calculates the similarity index between users. Then, it incorporates the attribute information of patent literature (e.g., IPC classification, inventors) to implement personalized recommendations. This method is applied to a self-developed patent recommendation system and compared with

基金项目 2023 年度天津市专利转化专项计划专项基金项目“供给侧专利培育与转化促进项目研究”。

作者简介 赵学铭(1983-), 硕士, 正高级工程师, 主要研究方向为知识产权成果转化; 王刚(1980-), 硕士, 副研究馆员, 主要研究方向为个性化信息服务, E-mail: wanggang80@tmu.edu.cn。

引用格式 赵学铭, 王刚. 融合专利属性和用户行为分析的个性化专利推荐研究[J]. 情报工程, 2025, 11(1): 42-53.

other collaborative filtering recommendation methods. [Limitations] The research may be limited by the specificity of the data, such as the reliance on specific user behaviors and patent literature datasets. Moreover, the experimental results may vary depending on the selection of users and patent literature in the system. [Results/Conclusions] The experimental results show that this personalized recommendation method outperforms the traditional collaborative filtering recommendation method. It better understands users' needs and preferences, thereby improving the accuracy and efficiency of recommendations.

Keywords: User Behavior Analysis; Collaborative Filtering; Personalized Recommendation; Patent Services

引言

专利文献以其特有的高实用性、专业性和前沿性在各行各业的发展中发挥着技术风向标的作用，是当今信息时代下最具研究意义的知识载体^[1]。伴随着科教兴国战略的实施，以及人们对于知识产权保护意识的不断增强，越来越多的科研成果和发明创造以专利申请的形式得到了保护，这也导致了专利数量的不断攀升。在专利数量庞大且不断增长的情况下，如何从中为科研人员和用户挖掘出满足其自身需求的专利文献是目前信息服务领域亟待解决的问题。为了应对这种“信息过载”的情况，实现向用户准确推荐有效的专利文献信息，需要根据用户的行为、偏好构建用户兴趣偏好模型，采用相应的推荐算法将匹配的专利文献信息推荐给目标用户。

1 相关研究

推荐系统在信息过滤、电子商务、社交网络等领域取得了广泛的应用。专利文献推荐作为其中的一个重要分支，主要关注如何为用户提供个性化的专利信息推荐服务。专利文献推荐不同于传统的推荐系统，它不仅需要考虑用

户行为和兴趣偏好，还需结合专利的复杂属性和技术内容进行推荐。随着用户行为数据的丰富，越来越多的研究者关注用户在专利文献搜索和使用中的行为数据。冉从敬等^[2]提出了一种结合用户行为和专利内容的混合推荐模型，该模型结合用户的操作行为和专利的文本相似度信息，显著提高了推荐的精准度。宋凯等^[3]进一步结合用户画像技术，将用户的长期操作行为和专利文献的技术属性关联起来，显著提高了专利推荐的准确性。然而，这些方法主要集中于用户的行为数据，对于专利文献的多维结构化信息（如国际专利分类号、发明人等）考虑不足，限制了推荐结果的多样性，降低了覆盖率。

与传统的商品推荐或电影推荐不同，专利文献不仅具有复杂的文本信息，还包含丰富的结构化属性信息，例如国际专利分类号（IPC）、发明人、申请人等。这些结构化信息对于提高推荐的精准度具有重要意义。近年来，研究者们开始探索如何将这些结构化数据与用户行为结合起来，进行更加智能化的专利推荐。黎楠等^[4]提出了一种基于LDA主题模型的专利推荐方法，通过将发明人与专利文献的主题分布相结合，实现了个性化推荐。该方法能够有效地挖掘用户在特定技术领域中的兴趣，并根据

主题相似性推荐相关的专利文献。然而，该方法仍然依赖于文本分析，未能充分利用专利文献的结构化信息。为了进一步提升推荐效果，一些学者将知识图谱等方法应用于专利推荐之中。王嘉瑜^[5]提出了基于知识图谱的专利推荐算法（KG-PR），将专利知识图谱向量化结果融入基于内容的推荐算法中，利用知识图谱中的专利标题实体计算专利之间的相似度，实现类似专利的推荐，从而完成专利所属技术领域的推荐。Deng等^[6]针对以往的专利推荐方法需要收集用户的专利检索的明确需求，无法有效捕捉专利转移背景下公司隐含的需求，提出了一种基于知识图谱的方法，通过构建专利知识图谱来捕捉专利领域关键词之间的语义信息，并将专利与公司分别建模为加权图。Xiao等^[7]提出了一种基于知识图谱（KG）表示的专利推荐方法，通过KG表示学习，建模产品与专利异构图中的语义关联和约束，从而提升推荐效率。Chen等^[8]将异构的专利信息组织为知识图谱，提取专利与公司配对的关联性和质量特征，并结合深度神经网络与相关传播技术设计了一个可解释的推荐模型。

本文提出了一种结合用户行为和专利文献多维属性的个性化推荐方法，不仅考虑了用户在专利文献中的操作行为，还通过引入专利的

结构化信息（如IPC分类号、发明人和申请人等），进一步提升了推荐的准确性和多样性。

2 专利文献信息数据模型

专利文献信息是集技术、经济、法律和战略等多方面于一体的综合信息，具有数量庞大、内容广泛、报道及时、结构统一等特点。广义上，专利信息指所有与专利活动相关的信息；狭义上，则是从专利局发布的文件中获取的技术、经济、法律及战略信息^[9]。专利文献包含摘要、技术背景和权利要求等文本信息，通过提取其中的主题词，能够发现潜在的知识或模式，便于用户检索。专利技术主题识别基于文献标题、摘要及技术要点中的技术特征词，经过文本预处理和专家筛选，建立主题词之间的关联关系，进而聚类形成技术主题^[10]。主题词不仅能深入揭示文献内容，还能将相同主题的不同作者或跨学科文献集中^[11]。因此，每项专利信息可以通过其所包含的主题词组合进行表示，采用相应的方法^[12-13]，诸如对专利的标题、摘要和全文进行分词，基于这些分词后的词汇进行分析，对每项专利文献进行主题词提取。通过主题词的提取工作，可以形成专利文献主题词集合，每项专利信息可以通过主题词集合中所包含的主题词进行表示，如表1所示。

表1 专利文献—主题词对应关系表

	主题词 ₁	主题词 ₂	...	主题词 _{n-1}	主题词 _n
专利 ₁	$w_{1,1}$	$w_{1,2}$...	$w_{1,n-1}$	$w_{1,n}$
专利 ₂	$w_{2,1}$	$w_{2,2}$...	$w_{2,n-1}$	$w_{2,n}$
...
专利 _{m-1}	$w_{m-1,1}$	$w_{m-1,2}$...	$w_{m-1,n-1}$	$w_{m-1,n}$
专利 _m	$w_{m,1}$	$w_{m,2}$...	$w_{m,n-1}$	$w_{m,n}$

每篇专利文献可以表示为一个由系统中所有主题词构成的向量，该向量的每个元素代表相应主题词在该专利文献中的权重，其值通过 TF-IDF 算法计算。当某个主题词出现在这个专利文献中，其权重为 $w_{i,j}$ ，反之为 0。主题词 t_i 在专利文献 p_j 权重的计算方法采用 TF-IDF 方法，计算方法如公式 (1) 所示。

$$w_{i,p_j} = \frac{N(t_i, p_j)}{\sum_{t_k \in T(p_j)} N(t_k, p_j)} \times \log \frac{|P|}{1 + |P(t_i)|} \quad (1)$$

其中， $N(t_i, p_j)$ 表示主题词 t_i 在专利文献 p_j 中出现的次数， $\sum N(t_k, p_j)$ ， $t_k \in T(p_j)$ 表示专利文献 p_j 所对应的主题词集合 $T(p_j)$ 中每个主题词在该文献中出现次数的总和。 $|P|$ 表示系统中专利文献总数， $|P(t_i)|$ 表示系统中包含主题词 t_i 的专利文献数量。

由于每项专利信息至少包含一个主题词，而且会和其他专利信息包含相同的主题词，因此专利信息与其所包含的主题词对应关系的数据模型如图 1 所示。

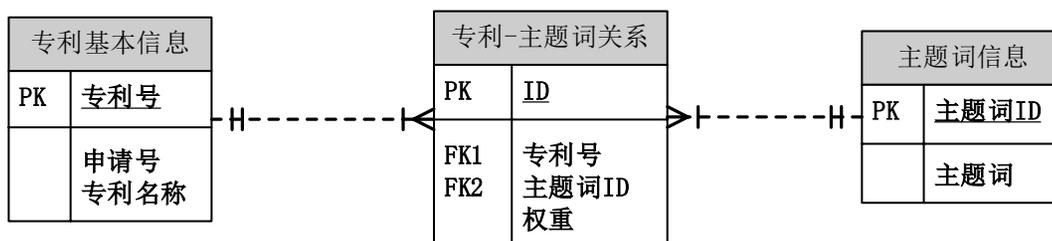


图 1 “专利—主题词对应关系”数据模型

每项专利以专利号作为唯一标识，专利信息中注明发明人和申请人。每项专利还包含一个或多个 IPC 分类号，用于专利的检索和分类。通过专利文

献中的发明人、申请人和 IPC 分类号，分别形成发明人集合、申请人集合和 IPC 分类号集合。专利文献与这些集合中的元素对应关系如表 2 所示。

表 2 专利文献与发明人、申请人以及 IPC 分类号对应关系表

专利文献 p	发明人 $_1$	发明人 $_2$...	发明人 $_{x-1}$	发明人 $_x$
	$a_{p,1}$	$a_{p,2}$...	$a_{p,x-1}$	$a_{p,x}$
	申请人 $_1$	申请人 $_2$...	申请人 $_{y-1}$	申请人 $_y$
	$b_{p,1}$	$b_{p,2}$...	$b_{p,y-1}$	$b_{p,y}$
	IPC 分类号 $_1$	IPC 分类号 $_2$...	IPC 分类号 $_{z-1}$	IPC 分类号 $_z$
	$c_{p,1}$	$c_{p,2}$...	$c_{p,z-1}$	$c_{p,z}$

表中 $a_{p,x}$ 、 $b_{p,y}$ 、 $c_{p,z}$ 分别表示发明人集合、申请人集合以及 IPC 分类号集合中的某个发明人、申请人以及 IPC 分类号是否出现在专利文献 p 中，如果出现，取值为 1，反之为 0。综上所述，专利文献及其所包含发明人、申请人以及 IPC 分类号属性的数据模型如图 2 所示。

发明人与专利间是多对多关系，一项专利可有多位发明人，发明人也可有多项专利，通过“专利—发明人关系”反映这种联系。类似地，申请人可申请多项专利，用“专利—申请人关系”表示。同样，“专利—IPC 关系”用于表示专利与其 IPC 分类号的对应关系。

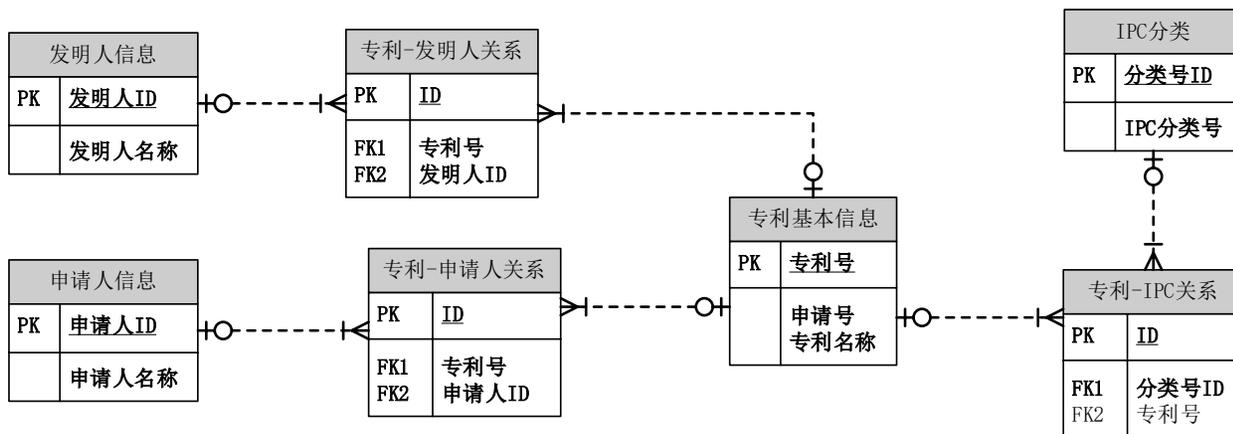


图2 专利文献数据模型

3 用户行为模型

3.1 用户检索行为模型

用户通过输入检索条件查找所需的专利文献。基于图1和图2中的专利文献信息数据模型，

本文考虑了用户按申请号、发明人、申请人和主题词四种类型进行专利检索。由于不同类型的检索条件反映了用户的兴趣偏好，系统会根据检索类型保存用户的搜索条件。用户的专利信息检索行为数据模型如图3所示。

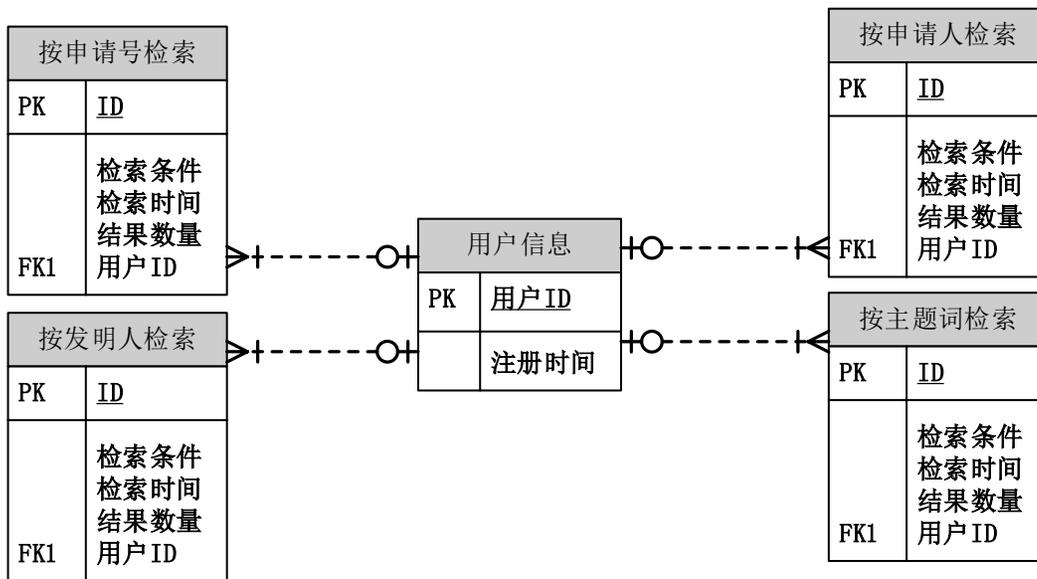


图3 用户检索行为数据模型

本文仅考虑用户输入检索条件后得到结果的行为，即“结果数量”不为空的情况。基于用户在四种检索类型下输入的条件，可将用户的专利检索兴趣偏好表示为向量 $S_u(t)$ ，其中 t 代

表四种检索类型（申请号、申请人、发明人和主题词）。每种检索类型对应一个向量，向量元素 $S_u(c_i(j))$ 表示用户 u 在该类型 t 下检索条件 $c_i(j)$ 的权重， j 表示该类型下保存的用户输入条

件集合。 $S_u(c_i(j))$ 的计算参考 TF-IDF 思想, $S_u(t)$ 及其元素 $S_u(c_i(j))$ 的计算方法如公式 (2) 所示。

$$S_u(t) = \{S_u(c_i(1)), S_u(c_i(2)), \dots, S_u(c_i(k-1)), S_u(c_i(k))\}$$

$$S_u(c_i(j)) = \frac{COUNT_u(c_i(j))}{\sum_{j=1}^k COUNT_u(c_i(j))} \times \log \frac{|User|}{1 + |User(c_i(j))|} \quad \begin{matrix} t = \{1, 2, 3, 4\} \\ j = \{1, 2, \dots, k\} \end{matrix} \quad (2)$$

其中, $COUNT_u(c_i(j))$ 表示用户 u 在某种搜索类型 t 条件下使用 $c_i(j)$ 进行搜索的次数, $\sum_j COUNT_u(c_i(j)), j = \{1, \dots, k\}$ 表示用户 u 在 t 类型搜索条件下进行搜索行为的总次数。 $|User|$ 表示用户总数量, $|User(c_i(j))|$ 表示在搜索类型 t 条件下使用 $c_i(j)$ 作为搜索条件进行搜索的用户数量。向量 $S_u(t)$ 用于后续根据搜索条件为目标用户查找兴趣偏好近似的最近邻用户集合做准备。

3.2 用户操作行为模型

用户对信息资源的操作行为 (如收藏、点击和浏览) 是其兴趣偏好的一种体现, 通过对这些行为的量化分析, 能够更准确地捕

捉用户在专利文献中的潜在兴趣。用户通常会收藏感兴趣的专利信息, 并对其进行多次点击或持续浏览超过一定时间。为了精确反映用户的行为偏好, 对用户的收藏、点击和浏览时长进行了量化。本文对系统后台数据库中保存的用户收藏、点击和浏览行为进行动态量化, 并为这三种操作行为赋予不同的权重, 记为 $e_j, j=1, 2, 3$ 。根据李聪等^[14] 通过研究电子商务领域中用户的行为发现用户对于商品收藏行为体现的兴趣偏好高于点击和浏览行为, 收藏行为对应的权重高于另两种行为, 因此 $e_1 \geq e_2 + e_3$ 且 $e_1 + e_2 + e_3 = 1$, 用户 u 对于专利文献信息 p 操作行为的动态量化记为 $B_u(p)$, 如公式 (3) 所示。

$$B_u(p) = \varepsilon_1 \times \exp\left(-\frac{t_{now} - t_{Fav}(u, p)}{t_{now} - t_{FirstFav}(u)}\right) \times Fav(u, p) + \varepsilon_2 \times \frac{\sum Brw(u, p)}{MAX_{x \in P_u} \{\sum Brw(u, x)\}} + \varepsilon_3 \times \frac{t_{LastClk}(u, p) - t_{FirstClk}(u, p)}{t_{Now} - t_{FirstClk}(u, p)} \times \frac{Clk(u, p)}{MAX_{x \in P_u} \{Clk(u, x)\}} \quad (3)$$

$Fav(u, p)$ 表示用户 u 对专利文献 p 是否进行收藏。 t_{now} 表示当前时间, $t_{Fav}(u, p)$ 表示用户 u 收藏专利文献 p 的时间, $t_{FirstFav}(u)$ 表示用户 u 首次在系统中收藏某个专利文献的时间。由于用户对某项专利文献信息会存在多次浏览的行为, $\sum Brw(u, p)$ 表示用户 u 对专利文献 p 浏览时间的总和, $MAX\{\sum Brw(u, x)\}, x \in P_u$ 表示用户 u 访问的专利信息集合 P_u 中各个专利信息浏览时间总和的最大值。 $Clk(u, p)$ 表示用户 u 对专利文献 p 的点击次数, $t_{FirstClk}(u, p)$ 表示用户 u 对项目 p

的首次点击时间, $t_{LastClk}(u, p)$ 表示用户 u 对专利文献 p 的最后一次点击时间。 $MAX\{Clk(u, x)\}, x \in P_u$ 表示用户 u 访问的资源集合 P_u 中各个专利信息点击次数最大值。

通过用户对专利文献信息操作行为的量化, 可以获取用户在操作行为方面对系统中专利文献的兴趣偏好, 这种兴趣偏好可以通过向量的形式进行表示, 假设系统中有 m 条专利文献信息, 则用户 u 在专利行为访问上的兴趣偏好记为 $Inst_u = \{B_u(1), B_u(2), \dots, B_u(p), \dots, B_u(m-1),$

$B_u(m)\}, p \in [1, m]$ 。

4 个性化专利推荐方法

4.1 用户相似性度量

4.1.1 用户搜索行为量化的相似性度量

根据前文用户搜索行为模型得到的每位用

$$SIM_{S_t}(u, v) = \frac{\sum_{j=1}^k [S_u(c_t(j)) \times S_v(c_t(j))]}{\sqrt{\sum_{j=1}^k [S_u(c_t(j))]^2} \times \sqrt{\sum_{j=1}^k [S_v(c_t(j))]^2}} \quad \begin{matrix} t = \{1, 2, 3, 4\} \\ j = \{1, 2, \dots, k\} \end{matrix} \quad (4)$$

在得到四种不同搜索类型所对应的用户间搜索行为兴趣相似性之后，对四个相似性结果

$$SIM_S(u, v) = \begin{cases} SIM_{S_t}(u, v) & |\{SIM_{S_t}(u, v) | SIM_{S_t}(u, v) \neq 0\}| = 1 \\ \sqrt[n]{\frac{n \times \prod_{t=1}^n SIM_{S_t}(u, v)}{\sum_{t=1}^n SIM_{S_t}(u, v)}} & n = |\{SIM_{S_t}(u, v) | SIM_{S_t}(u, v) \neq 0\}| \\ & 1 < n \leq 4 \\ 0 & |\{SIM_{S_t}(u, v) | SIM_{S_t}(u, v) \neq 0\}| = 0 \end{cases} \quad (5)$$

该公式中仅仅考虑四种搜索类型所对应的用户间搜索行为兴趣相似性不为 0 的结果，公式中 n 表示不为 0 相似性结果的数量。

4.1.2 用户操作行为量化的相似性度量

根据前文用户对专利文献信息操作行为模型所得到的每个用户对系统中各项专利文献访问行为量化形成的基于访问行为兴趣偏好向量 $Inst_u = \{B_u(1), B_u(2), \dots, B_u(p), \dots, B_u(m-1), B_u(m)\}$, $p \in [1, m]$ 。使用余弦相似性度量方法计算用户间对专利文献操作行为上的兴趣偏好相似性，计

户在四种不同搜索类型所对应的搜索行为兴趣偏好向量 $S_u(t) = \{S_u(c_t(j))\}$, 其中 $t = \{1, 2, 3, 4\}$, $j = \{1, \dots, k\}$ 。为了计算目标用户与系统中其他用户的行为兴趣偏好相似性，采用余弦相似性度量四种搜索类型下的用户行为兴趣向量之间的夹角，公式 (4) 给出了余弦相似性的计算方法，该方法能够量化用户行为偏好的一致性。

计算它们的调和平均数，得到用户搜索行为量化的相似性度量，计算方法如公式 (5) 所示。

算方法如公式 (6) 所示。

$$SIM_B(u, v) = \frac{\sum_{p \in P} [B_u(p) \times B_v(p)]}{\sqrt{\sum_{p \in P} [B_u(p)]^2} \times \sqrt{\sum_{p \in P} [B_v(p)]^2}} \quad (6)$$

4.1.3 综合相似性度量

在得到用户搜索行为偏好以及用户操作行为偏好的相似性度量结果之后，计算二者的调和平均数作为用户间兴趣偏好相似性的综合度量结果，计算方法如公式 (7) 所示。

$$SIM(u, v) = \begin{cases} \frac{2 \times SIM_S(u, v) \times SIM_B(u, v)}{SIM_S(u, v) + SIM_B(u, v)} & SIM_S(u, v) \neq 0 \ \& \ SIM_B(u, v) \neq 0 \\ SIM_S(u, v) & SIM_S(u, v) \neq 0 \ \& \ SIM_B(u, v) = 0 \\ SIM_B(u, v) & SIM_S(u, v) = 0 \ \& \ SIM_B(u, v) \neq 0 \\ 0 & SIM_S(u, v) = 0 \ \& \ SIM_B(u, v) = 0 \end{cases} \quad (7)$$

4.2 推荐方法的研究

4.2.1 目标用户最近邻的选择

传统的推荐方法中，根据与目标用户相似性度量的结果进行排序，可能会导致推荐结果的多样性不足。通过设定阈值 δ 可以扩大用户

$$\delta = \frac{1}{|U'|} \sum_{u \in U'} SIM(a, u) \quad U' = \{u | u \in U \& a \neq v\} \& SIM(a, u) \neq 0 \quad (8)$$

将系统中与目标用户 a 相似性综合度量结果高于阈值 δ 的用户组成目标用户兴趣偏好相似的最近邻用户集合，记为 U_N ， $U_N = \{v | v \in U \& SIM(a, v) > \delta\}$ 。

4.2.2 专利文献推荐方法

收集到最近邻用户集合中每位用户已访问的所有专利文献信息，找到这些专利文献信息中目标用户 a 访问过的专利，形成向目标用户推荐的专利文献信息推荐候选集合，记为 P_c ，即 $p_c = \{p_c | p_c \in P_a\}$ ，其中 P_a 表示目标用户 a 已访问的专利文献集合。

根据前文“专利信息数据模型”所述，通过相关数据模型保存了系统中所有专利文献所包含的主题词。表1“专利文献—主题词对应关系表”显示了每个专利文献与系统中所有主题词的对应关系情况，因此候选集合中每个专利文献与系统中所有主题词关系的向量表示为 $p_c = \{w(1), w(2), \dots, w(i), \dots, w(z-1), w(z)\}$ ，其中 $p_c \in P_c$ ， $i \in \{1, \dots, z\}$ ， $w(i)$ 为前文所述根据 TF-

集合的覆盖范围，从而提高推荐结果的多样性，满足用户的多元化需求。根据公式(7)，在求得目标用户与系统中其他用户相似性综合度量的结果之后，根据公式(8)所示求得均值，以此作为目标用户 a 的相似性的阈值。

IDF 方法计算的关键词 i 在专利文献 p_c 中的权重。 z 表示系统中所有专利文献所包含主题词的数量，当专利 p_c 不包含该关键词时， $w(i)$ 取值为 0。

此外，按照表2“专利文献与发明人、申请人以及 IPC 分类号对应关系表”的描述，每个专利文献可以通过三个向量表示其与这三类属性的对应关系，因此候选专利文献集合中的每项候选专利与这三类属性对应关系的向量表示为 $p_c = \{w_t(1), w_t(2), \dots, w_t(k), \dots, w_t(x-1), w_t(x)\}$ ，其中 $p_c \in P_c$ ， $t = \{1, 2, 3\}$ ， $k \in \{1, \dots, x\}$ ， x 表示系统中所有专利文献所包含的 t 类型属性的数量，当专利 p_c 包含该属性时 $w_t(k)$ 取值为 1，反之为 0。

在对专利及其属性向量化表示后，需要计算候选专利与目标用户已访问专利的相似性。相似性计算分为两部分：首先，通过余弦相似度衡量专利属性（如发明人、申请人和 IPC 分类号）的相似性，具体计算见公式(9)。

$$SIM_{2,t}(p_a, p_c) = \frac{\sum_{k=1}^x [w_{t,p_a}(i) \times w_{t,p_c}(i)]}{\sqrt{\sum_{k=1}^x [w_{t,p_a}(i)]^2} \times \sqrt{\sum_{k=1}^x [w_{t,p_c}(i)]^2}} \quad \begin{matrix} p_a \in P_a & p_c \in P_c \\ k = \{1, 2, \dots, x\} \\ t = \{1, 2, 3\} \end{matrix} \quad (9)$$

得到发明人、申请人以及 IPC 分类号等三个类型属性的相似性度量结果后，需要计算三

者的调和平均数作为基于属性的综合相似性结果，计算方法如公式(10)所示。

$$SIM_2(p_a, p_c) = \begin{cases} SIM_{2,t}(p_a, p_c) & |\{SIM_{2,t}(p_a, p_c) | SIM_{2,t}(p_a, p_c) \neq 0\}| = 1 \\ \sqrt{\frac{n \times \prod_{t=1}^n SIM_{2,t}(p_a, p_c)}{\sum_{t=1}^n SIM_{2,t}(p_a, p_c)}} & n = |\{SIM_{2,t}(p_a, p_c) | SIM_{2,t}(p_a, p_c) \neq 0\}| \\ & 1 < n \leq 3 \\ 0 & |\{SIM_{2,t}(p_a, p_c) | SIM_{2,t}(p_a, p_c) \neq 0\}| = 0 \end{cases} \quad (10)$$

公式中 n 表示三个类型属性的相似性度量结果中不为 0 的数量，如果三个结果均为 0，该综合相似性度量结果为 0。此外，根据专利

文献主题词向量形式表示的结果，可以计算基于主题权重的专利文献相似性，计算方法如公式 (11) 所示。

$$SIM_1(p_a, p_c) = \frac{\sum_{i=1}^z [w_{p_a}(i) \times w_{p_c}(i)]}{\sqrt{\sum_{i=1}^z [w_{p_a}(i)]^2} \times \sqrt{\sum_{i=1}^z [w_{p_c}(i)]^2}} \quad \begin{matrix} p_a \in P_a & p_c \in P_c \\ i = \{1, 2, \dots, z\} \end{matrix} \quad (11)$$

在得到基于专利属性和基于专利主题的相似性度量结果之后，再次计算这二者的调和平均数，

得到候选集中专利与目标用户已访问专利文献的相似性度量结果，计算方法如公式 (12) 所示。

$$SIM(p_a, p_c) = \begin{cases} \frac{2 \times SIM_1(p_a, p_c) \times SIM_2(p_a, p_c)}{SIM_1(p_a, p_c) + SIM_2(p_a, p_c)} & SIM_1(p_a, p_c) \neq 0 \ \& \ SIM_2(p_a, p_c) \neq 0 \\ SIM_1(p_a, p_c) & SIM_1(p_a, p_c) \neq 0 \ \& \ SIM_2(p_a, p_c) = 0 \\ SIM_2(p_a, p_c) & SIM_1(p_a, p_c) = 0 \ \& \ SIM_2(p_a, p_c) \neq 0 \\ 0 & SIM_1(p_a, p_c) = 0 \ \& \ SIM_2(p_a, p_c) = 0 \end{cases} \quad (12)$$

在公式 (9) — (12) 中， $\{p_a\}$ 表示目标用户 a 已访问的专利文献集合， $\{p_c\}$ 表示准备向目标用户 a 推荐的候选专利文献集合。将专利候选集中的每项专利与目标用户已访问专利的相似性度量结果的均值进行降序排序，最后将候选集中相似性均值最高的 N 条专利文献推荐给目标用户。

系统使用 HTML5 设计网站页面，系统主页面如图 4 所示。主页面为用户提供了四种类型的专利搜索形式，并显示出用户收藏数和点击数较高的热门专利文献。

用户登录系统后，系统按照本文设计的专利推荐方法，根据用户对专利的搜索和使用行为，向该用户推荐其感兴趣的专利文献列表，如图 5 所示。

5 系统实现与实验对比

5.1 系统实现

将本文设计的专利文献推荐方法用于实际应用中，开发了专利文献个性化推荐系统，本

5.2 实验对比

5.2.1 数据准备

本文实验对比使用的数据为所开发的专利文献个性化推荐系统中的数据，其中包括 362 名用户按四种搜索类型输入的搜索条件，以及



图4 系统主页面



感兴趣的专利文献推荐

用于提高肿瘤靶向CLDN18.2治疗疗效的药物
公开日期 (公开): 2024-03-15 | 申请人: 北京肿瘤医院 | 17

本发明涉及生物医药领域, 提供用于提高肿瘤靶向CLDN18.2治疗疗效的药物, 包括靶向抑制肿瘤细胞或纤维细胞及其关联分子的抗体、靶向抑制与T细胞耗竭相关的免疫检查点分子的抗体。上述治疗手段与靶向CLDN18.2的细胞治疗药物或抗体治疗药物联合, 能够显著提高肿瘤靶向CLDN18.2细胞治疗药物和抗体治疗药物的疗效, 使肿瘤患者获益。详细阅览

肿瘤靶向CLDN18.2治疗耐药标志物及其应用
公开日期 (公开): 2024-03-15 | 申请人: 北京肿瘤医院 | 20

本发明属于医学诊断技术领域, 公开了肿瘤靶向CLDN18.2治疗耐药标志物及其应用。本发明揭示了细胞外囊泡 (EV) 携带的CLDN18.2、PD1、PD-L2表达及其衍生评分或模型作为肿瘤靶向CLDN18.2治疗方案疗效和耐药标志物的用途, 能够划分有效人群、预测治疗效果、监测耐药发生, 且对细胞治疗药物和抗体治疗药物都具有明确的标志物潜力, 有望推动靶向CLDN18.2治疗方案的伴随诊断手段的开发, 为今后CLDN18.2靶向治疗药物的临床实践积累宝贵经验。详细阅览

粉防己碱作为线粒体SIRT5靶向抑制剂在制备抗肿瘤药物的应用
公开日期 (公开): 2024-03-08 | 申请人: 河北医科大学第三医院 | 7

本发明属于疾病治疗药物技术领域, 具体涉及中药成分粉防己碱作为线粒体SIRT5靶向抑制剂在制备抗肿瘤药物的应用。实验证明, 粉防己碱作为一种直接靶向SIRT5的天然化合物小分子, 当粉防己碱与细胞中SIRT5结合后可诱导细胞自噬的发生, 同时自噬发生后粉防己碱又可作为一种自噬体抑制剂, 切断了自噬体降解的途径, 诱导了自噬体发生自噬阻断, 从而抑制肿瘤的生长。因此受SIRT5所调控的肿瘤均可采用粉防己碱作为抑制剂进行治疗, 尤其对黑色素瘤、胃癌和食管癌等具有很好的抑制作用。详细阅览

FGF21作为肿瘤免疫治疗靶点在制备筛选抗肿瘤药物中的应用
公开日期 (公开): 2024-03-08 | 申请人: 上海交通大学医学院附属仁济医院 | 10

本发明提供了FGF21作为肿瘤免疫治疗靶点在制备筛选抗肿瘤药物中的应用, 本发明还提供了FGF21的中和抗体在制备治疗肿瘤的药物中的用途。本发明利用酶联免疫吸附试验 (ELISA) 筛选出FGF21在多种肿瘤细胞中分泌表达, 并与肿瘤的不良预后呈正相关。随后通过一系列体内体外实验证实肿瘤细胞来源的FGF21能显著抑制CD8⁺T细胞杀伤性功能, 如使用中和抗体阻断FGF21, 可显著恢复CD8⁺T细胞功能, 抑制肿瘤生长。联合使用anti-PD-1等免疫检查点抑制剂具有明显的协同效应。因此FGF21可作为一个新的肿瘤免疫治疗靶点用以制备抗肿瘤免疫应答, 具有很好的应用潜力和价值。详细阅览

LAYN分子在预测肿瘤内免疫细胞浸润、免疫检查点分子表达水平、T细胞耗竭中的应用
公开日期 (公开): 2024-01-19 | 申请人: 复旦大学附属华山医院 | 6

图5 用户界面

用户您好
您共计收藏36个专利文献, 点击此处修改个人信息

已收藏的专利文献

- 粉防己碱作为线粒体SIRT5靶向抑制剂在制备抗肿瘤药物的应用
收藏时间: 2024-1-17 点击次数: 36 | 17
- 基于基因组学机制的子宫颈鳞状上皮癌识别系统及方法
收藏时间: 2024-1-13 点击次数: 27 | 19
- 双靶双响应纳米药物的制备方法及其抗肿瘤用途
收藏时间: 2024-2-1 点击次数: 22 | 23
- 基于血清蛋白的乳腺癌筛查的标志物及应用
收藏时间: 2024-1-8 点击次数: 19 | 20
- CD47/SIRPα分子抑制剂及其应用抗肿瘤药物组合
收藏时间: 2024-2-2 点击次数: 13 | 11

1 2 3 Next » Last »

感兴趣的申请 (专利权) 人

- > 北京肿瘤医院
- > 复旦大学附属华山医院
- > 四川省肿瘤医院
- > 中国医学科学院肿瘤医院
- > 首都医科大学
- > 四川大学华西医院
- > 中山大学孙逸仙纪念医院

对系统中 6484 项专利文献的操作行为数据，操作行为数据中包括用户对专利文献信息的收藏，点击以及浏览时长数据。以 2023 年 12 月 31 日前的数据作为训练集数据，以 2024 年 1 月 1 日以后的数据作为测试集数据。

5.2.2 评价指标

为了验证本文所提出的方法在实验数据集上的推荐性能，将本方法（记为 MB-CF）与基于内容和协同过滤的融合推荐方法（记为 FUIM）进行对比。FUIM 方法^[15]改进了传统基于内容的方法，通过对提取文献特征词进行协同过滤获取用户兴趣实施推荐。

此外本文选择精确率（Precision）和召回率（Recall）作为实验的评价指标，精确率表示用户对系统推荐资源感兴趣的概率，召回率表明一个用户喜欢的项目被推荐的概率，精确率和召回率越高，表示推荐效果越好^[16]。

5.2.3 实验结果比较

将两种推荐方法在数据集上的实验结果进行比较，并将 Top-N 值设置为 {5,10,15,20,25,30}，实验对比结果如图 6 所示。

从图中可以发现随着 Top-N 取值的增长，两种推荐方法的准确率均有所下降，而召回率均有所提升。从对比实验结果可以发现本文提出的方法优于对比方法，其原因在于以下几个方面：首先，在用户行为与专利属性的多维结合方面，MB-CF 方法通过量化用户的搜索和操作行为，并结合专利的结构化信息（如发明人、申请人、IPC 分类号），提供更加精准的推荐。该方法充分利用专利文献的多维数据，不仅基于文本相似性，还结合专利背景，生成更具针对性的推荐结果。而 FUIM 方法主要依赖特征词提取和协同过

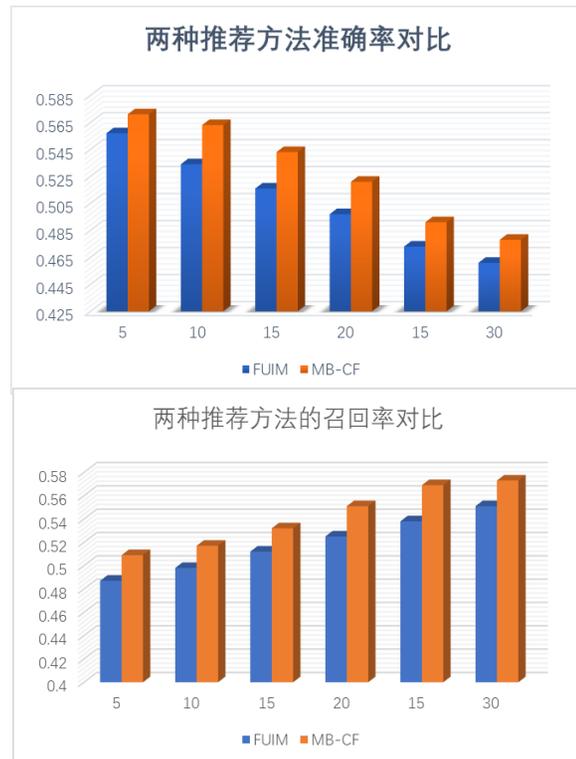


图 6 实验结果对比图

滤，对用户行为的利用不足，难以捕捉更深层次的需求和文献间的多样性。其次，在相似性度量的精确性方面，MB-CF 方法通过综合用户的搜索和操作行为，使用余弦相似度计算用户间的兴趣相似性，并量化浏览、点击、收藏等行为，准确找到与目标用户兴趣最相似的其他用户。相比之下，FUIM 方法过度依赖内容相似性，忽视用户操作行为的细节，导致推荐精度下降。最后，在推荐结果的多样性和覆盖率方面，MB-CF 方法结合用户行为和专利的多维属性，通过设定相似度阈值，扩大推荐范围，保证结果不仅最相似，还能满足用户多元化需求，提升推荐多样性。而 FUIM 方法由于过度依赖内容，导致推荐结果单一，难以覆盖用户的潜在需求。综上，MB-CF 方法通过用户行为、专利多维属性与精确相似性度量的结合，在推荐准确性和多样性上优于

FUIM, 更好满足了用户的复杂需求。

6 结束语

本文提出了一种融合专利属性与用户行为分析的个性化专利推荐方法,通过量化用户搜索和操作行为,并结合专利的多维结构化信息,实现了更为精准的推荐。与传统的协同过滤和基于内容的推荐方法相比,该方法能够更全面地捕捉用户偏好,提高了推荐系统的准确性与多样性。实验结果显示,本文方法在精确率和覆盖率上均优于传统方法,尤其在应对用户兴趣多样化方面具有显著优势。通过引入用户行为与专利属性的结合,系统能够有效提升推荐的多样性,并减少冗余推荐的出现。

虽然本文方法取得了良好的效果,研究仍有一定局限性,如实验数据集的规模较小、用户行为的复杂性可能影响结果的泛化性。未来工作可进一步扩展数据集,并结合深度学习或知识图谱等技术,以提升推荐系统的智能化水平与效率。总之,本文方法为个性化专利推荐提供了有效的解决方案,在技术领域应用中具有广泛的前景。未来研究将致力于提高方法的扩展性和鲁棒性,以期在实际应用中进一步提升推荐性能。

参考文献

- [1] 艾楚涵. 基于文本聚类的专利推荐研究 [D]. 昆明: 昆明理工大学, 2021.
- [2] 冉从敬, 宋凯. 基于混合方法的高校专利个性化推荐模型构建 [J]. 情报理论与实践, 2020, 43(10): 93-98.
- [3] 宋凯, 冉从敬. 基于企业画像的高校专利个性化推荐 [J]. 图书馆论坛, 2022, 42(9): 123-131.
- [4] 黎楠, 杜永萍, 何明. 基于主题发现的专利发明人推荐方法 [J]. 情报工程, 2015, 1(3): 90-97.
- [5] 王嘉瑜. 基于知识图谱的专利自动推荐技术研究 [D]. 哈尔滨: 哈尔滨工业大学, 2021.
- [6] DENG W, MA J. A knowledge graph approach for recommending patents to companies[J]. Electronic Commerce Research, 2022, 22(9): 1435-1466.
- [7] XIAO Y, LI C D, Thürer M. A patent recommendation method based on KG representation learning[J]. Engineering Applications of Artificial Intelligence, 2023(126): 106722.
- [8] CHEN H, DENG W. Interpretable patent recommendation with knowledge graph and deep learning[J]. Scientific Reports, 2023, 13(1): 2586.
- [9] 蒲佳炜, 文庭孝. 高价值专利信息挖掘及实证研究——以自动驾驶领域为例 [J]. 中国发明与专利, 2024, 21(2): 4-12.
- [10] 李姝影, 张鑫, 许轶, 等. 核心专利集筛选及专利技术主题识别影响 [J]. 情报学报, 2019, 38(1): 13-20.
- [11] 柳胜国. 主题词在计算机检索文献资料中的作用 [J]. 情报杂志, 2000(3): 32-33.
- [12] 公冶小燕, 林培光, 任威隆, 等. 基于改进的 TF-IDF 算法及共现词的主题词抽取算法 [J]. 南京大学学报 (自然科学), 2017, 53(6): 1072-1080.
- [13] 石晶, 李万龙. 基于 LDA 模型的主题词抽取方法 [J]. 计算机工程, 2010, 36(19): 81-83.
- [14] 李聪, 马丽, 梁昌勇. DFM-IA: 面向 B2C 电子商务的多源用户兴趣数据采集机制 [J]. 管理工程学报, 2017, 31(1): 58-70.
- [15] 杨武, 唐瑞, 卢玲. 基于内容的推荐与协同过滤融合的新闻推荐方法 [J]. 计算机应用, 2016, 36(2): 414-418.
- [16] SARWAR B, KARYPIS G, KONSTAN J, et al. Analysis of Recommendation Algorithms for E-Commerce[C]// Proceedings of the 2nd ACM conference on electronic commerce, 2000, 158-167.



开放科学
(资源服务)
标识码
(OSID)

大模型时代我国知识服务领域面临的挑战

彭鹏

中国科学技术信息研究所 北京 100038

摘要: [目的/意义] 大模型快速发展对各行业具有深远影响, 着重探讨大模型在知识服务领域的应用。[方法/过程] 大模型发展概况、主要国家或地区发布的相关政策、文献情报领域及知识服务行业针对大模型应用的研究进展进行分析, 探讨了国内外主要的知识服务类机构在大模型方面的相关布局 and 研发特点, 分析指出国内知识服务领域在运用大模型过程中面临的挑战, 并有针对性地提出应对策略。[结果/结论] 从国产知识服务大模型研发、评估体系构建以及人才队伍的组建和培养方面提出了针对性建议。

关键词: 大模型; 知识服务; ChatGPT; 生成式人工智能; 基础模型

中图分类号: G35; TP391

Challenges of Large Models for Knowledge Service in China

PENG Peng

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

Abstract: [Objective/Significance] Large Models have been developing rapidly and applied in multiple fields, which has had a profound impact on the multi-industry application. This paper focuses on exploring the application of large models in the field of knowledge service. [Methods/Processes] This paper analyzes the development overview of large models, relevant policies issued by major countries or regions, research progress in the field of literature and intelligence as well as in the knowledge service for the application of large models. It explores the layout and R&D characteristics of major knowledge service institutions in the large models. The analysis points out the challenges that may be faced by the domestic knowledge service field in the process of applying large models, and proposes targeted strategies to deal with them. [Results/Conclusions] Targeted suggestions are proposed in the research and development of domestic knowledge service large models, the construction of an evaluation system, and the formation and cultivation of talent teams.

Keywords: Large Models; Knowledge Service; ChatGPT; Generative Artificial Intelligence; Foundation Model

作者简介 彭鹏 (1967-), 硕士, 副研究员, 主要研究方向为企业管理、财务管理和信息资源管理, E-mail: pengp@istic.ac.cn。

引用格式 彭鹏. 大模型时代我国知识服务领域面临的挑战 [J]. 情报工程, 2025, 11(1): 54-62.

引言

自 OpenAI 公司发布 ChatGPT 以来,全球掀起了大模型研究热潮,迅速成为人工智能研究的热门方向。大模型也迅速应用到医疗、金融、法律、教育等领域,推动各行业技术改进与升级。本文主要探讨大模型在知识服务业的发展及应用,讨论当前国内知识服务领域面临的挑战和应对策略,以期对我国知识服务领域未来大模型的应用及发展提供借鉴。

1 研究现状

1.1 发展概况

大模型是可生成流畅的文字、代码、图像及音频和视频等的神经网络模型^[1-2]。相关产品包括大语言模型 GPT-4 Turbo、多模态大模型 GPT-4o、文生视频大模型 Sora、谷歌 Gemini、视觉语言模型 PaLM、Meta 的 Llama3-V、Anthropic 的 Claude 2.1、百度的文心 4.0、月之暗面的 kimi 和深度求索的 DeepSeek 等。其主要特点包括:

(1) 参数大;当前大模型参数已升至万亿级,表现出更强的理解、推理和学习等能力^[3],谷歌 PaLM 的参数为 5400 亿;

(2) 训练数据大;法国 BLOOM 训练数据达 3416 亿条,文本数据约 1.5T;

(3) 算力需求大;BLOOM 训练吞吐量约 150 TFLOP,耗时 3.5 个月,大计算导致训练成本高,谷歌 Gemini Ultra 约 1.91 亿美元^[4];

(4) 学习泛化能力、鲁棒性及认知互动能力越来越强;SignLLM 可支持美国手语、德国

手语等 8 种手语,能够通过文字描述生成手语视频^[5];

(5) 领域化、行业化趋势显著;大模型已快速应用于医药、金融、法律等多个行业^[6-8];从全球来看,美国领先中国、欧盟。2023 年全球共发布了 149 个大模型^[4],其中 61 个来自美国,远大于中国(20 个);国内百度、阿里、商汤、科大讯飞的大模型产品位于第一梯队,百川、智谱、昆仑万维位于第二梯队^[9]。

1.2 应用现状

相关研究已探讨了大模型对文献情报工作的影响^[10-12],包括文献情报数据组织、知识服务、情报分析以及可能引发的隐私安全、用户使用习惯和图书情报体系等多个方面,对整个文献情报体系产生了深远影响。多数研究和管理人员已使用 ChatGPT、Gemini 和 Bing Chat 等大模型辅助完成工作,并将其纳入日常工作管理流程^[13]。具体到知识服务方面,研究人员已着重探讨了大模型兴起对图书馆、出版行业等带来的机遇和挑战以及相关知识服务平台的构建^[14-16],当前需要做出大量技术和管理创新,解决如何基于人工智能技术整合多源异构数据、在已有数据集基础上关联和融合更具深度与广度的知识、更加精准理解用户需求并提供符合用户偏好的知识内容、评测所构建的大模型在情报研究方面的应用能力和效果等问题^[17-19],提升模型泛化能力、鲁棒性、安全性等性能。

科技信息服务企业正在积极训练生成式 AI 大模型,以爱思唯尔的 Scopus AI、美国国家医学图书馆的 GeneGPT 和华知大模型为代表。

Scopus AI 可通过大模型推荐文献或撰写综述, 包括了精确的引文获取和观点支持类文献搜索, 其回答基于大量真实文献并列引文来源。GeneGPT 通过提取 NCBI 基因数据库中的文本摘要, 并利用 GPT-3.5 生成每个基因的向量表示, 结合了基因别名和外部数据库进行映射, 并通过多源向量库融合强逻辑控制, 实现了思维链 + 知识框架的强逻辑内容可控生成, 有效抑制大模型幻觉。华知大模型面向知识服务与科研行业推出的专业知识增强大模型, 融合了语言、视觉、听觉等多模态理解和生成能力, 能满足知识问答、问题生成、文本生成和分析、信息抽取、总结聚合等大模型基本功能, 具有科研选题、文献理解、文章伴读、知识推荐、知识溯源、论文综述、图表解读、学术搜问、AIGC 检测等应用。

2 知识服务型机构的应用举措

随着科技发展, 知识服务内涵也在不断变化, 当前的知识服务是基于人工智能技术、自然语言处理技术、大数据技术等集内容创作、生产、传播, 甚至应用于一体的知识内容的全新服务模式^[20-21]。国内外主要知识服务相关机

构在大模型方面的进展不一, 主要通过研发垂直领域大模型、战略合作、收购等多举措并施, 不断推进大模型在知识服务领域的深度应用。

2.1 研发垂直领域大模型

知识服务机构通常以通用大模型为基础, 结合其专业数据库语料构建专用大模型。爱思唯尔、数字科学、汤森路透、美国国家医学图书馆 (NLM) 等都基于 OpenAI 通用大模型, 使用提示工程、检索增强生成 (RAG) 或微调技术研发垂直领域大模型。

爱思唯尔的 Scopus AI 具有观点总结、扩展概要、思维导图、主题专家等功能。科睿唯安的“学术 AI 平台”, 包括多种生成式人工智能助手。美国 NLM 的 GeneGPT 大模型^[22], 通过上下文学习使 Codex 利用 NCBI 的生物学数据库回答基因组问题, 提升研究人员获取信息的效率。智慧芽的生物医药大模型 (PharmGPT) 和专利大模型 (PatentGPT) 利用了研发情报库、专利数据库、新药数据库、生物序列数据库等产品^[23]。当前知识服务机构构建垂直大模型技术路线的对比总结如表 1 所示。

表 1 知识服务型机构垂直领域大模型的技术路径对比

技术路径	典型案例	优势	局限性	适用领域
微调 (Fine-tuning)	科睿唯安学术 AI 平台	针对性强, 输出专业度高	依赖大量标注数据, 成本高	高精度要求的学术研究场景
检索增强生成 (RAG)	美国 NLM GeneGPT	动态更新知识库, 缓解幻觉问题	依赖外部知识库质量, 响应速度受影响	需实时数据支持的医药领域
提示工程 (Prompt Engineering)	爱思唯尔 Scopus AI	轻量化部署, 快速适配新需求	可控性较低, 复杂任务处理能力有限	辅助性知识可视化场景

2.2 与大模型企业合作

知识服务机构通过与大模型企业战略合作研发人工智能相关工具，推动自身业务发展。2023年，科睿唯安与以色列 AI21 Labs 合作，用于研发生成式人工智能工具和应用程序，为学生、教师和研究人员提供快速访问详细的上下文信息和答案，以及更多个性化的服务和相关推荐。华知大模型基于华为盘古基础大模型以及知网的海量高质量知识数据。国家图书馆与百度战略合作，使 AI 技术与图书馆深度融合，使图书馆资源更广泛、更便捷地服务于全社会。总体来看，机构合作有技术互补型、数据融合型、市场拓展型等 3 种模式。

其中，技术互补型模式以华为和中国知网的合作为例，其结合了华为的底层算力与知网知识库，构建中华知识大模型，实现文献理解、技术分析等 8 大功能，知识溯源准确率提升 40%；数据融合型模式以百度和国家图书馆的合作为例，其利用百度自然语言处理（NLP）技术处理国家图书馆非结构化文献，将用户检索效率提升 60%，覆盖了长尾知识需求；市场拓展型模式以科睿唯安和 AI21 Labs 的合作为例，即双方通过合作快速切入教育市场，借助生成式工具使学生研究周期缩短了 30%。可以发现，技术互补型更利于构建竞争壁垒，而市

场拓展型能快速获得场景落地优势。

2.3 收购大模型机构

国外知识服务企业收购人工智能企业，加强对生成式人工智能技术的整合，快速推进大模型研发。2023年，科睿唯安已经收购了 Alethea、MotionHall 以及 Global QMS 等企业，相关收购能够在其产品路线图中快速引入 AI 解决方案，改进产品和工作流程，短期内将进一步增强 Cortellis 医药数据库的功能。汤森路透公司收购了 Casetext，推出了基于 GPT-4 的生成式 AI 法律研究助手 CoCounsel，而后汤森路透将核心产品 Westlaw 与 Casetext 进行技术融合，为法律研究工具 Westlaw Precision 发布了首个生成式人工智能助手，为客户群提供法律领域的生成式 AI 服务。施普林格·自然收购了 protocols.io 和 Slimmer AI，protocols.io 是一个开发和共享可重复方法的安全平台，Slimmer AI 推出了应对虚假科学论文的开源增强软件 PySciDetect，可在稿件提交阶段用于识别虚假研究和数据，加快和改善出版流程。数字科学收购的 Writefull，是基于 AI 在学术期刊上进行训练的工具，能够识别科学术语，提供符合学术写作的语法和风格建议。当前知识服务型机构进行大模型收购的策略对比如表 2 所示。

表 2 知识服务型机构收购策略对比

收购类型	代表案例	短期效益	长期价值	风险点
技术能力补全	汤森路透收购 Casetext	6 个月内推出 CoCounsel 法律助手	构建法律垂直领域 AI 生态	技术融合难度高
数据资产扩充	施普林格收购 protocols.io	新增 10 万 + 生物实验协议数据	强化开放科学平台的数据闭环能力	数据合规成本增加
流程优化驱动	科睿唯安收购 Global QMS	Cortellis 数据库功能迭代周期缩短 50%	建立医药研发全流程知识管理体系	原有工作流程重构阻力

3 国内知识服务业大模型应用挑战

大模型正处于快速发展中，国外知识服务企业开展了研发、战略合作和收购等多样化布局，国内知识服务机构通过研发和战略合作推进知识服务大模型应用，虽然这些大模型能够协助完成多种任务，但其发展既面临着通用大模型的固有缺陷，也伴随着专有大模型的特别问题，主要包括知识性不足、算力和算法、安全和隐私、科技企业竞争及地缘政治影响等多方面的挑战。

3.1 知识性不足

知识专业性不足。通用大模型依赖互联网公开语料训练，但垂直领域核心知识往往存在于 PDF 报告、手写病历、工程图纸等非结构化载体中。工程领域包含大量缩写、化学分子式、特殊符号等非标准化文本，非结构化知识的提取存在较大的困难。例如，医疗领域若未融合《临床诊疗指南》等知识库，诊断建议可能出现原则性错误。

知识更新滞后。知识服务领域的信息高速变化，与模型长训练周期形成矛盾。传统大模型需数月完成全量训练，难以实时响应政策变化。其次，新知识注入导致旧知识覆盖，可能产生灾难性遗忘问题。此外，新旧知识混杂还可能存在数据污染的问题。

3.2 算力算法

大模型研发应用离不开强大算力和先进算法的支持。算力方面，中国在芯片、软件工具等方面受到美国制裁。2023 年 10 月美国发布

的新禁令将从原来对单芯片算力（TTP）的关注向性能密度阈值（PD）转移，并首次提出了对小型高性能芯片的出口限制。考虑到未来垂直大模型将随着行业细分场景的渗透上升而加速增长，因此 AI 芯片国产化替代为大势所趋，未来一段时间内中国垂直大模型将存在一定程度上的算力短板。

算法方面，当基于 LLaMA-2 等开源大模型开展知识服务领域的垂直大模型研发时，算法缺乏必要的技术支持以及安全保障，存在版权纠纷、商业泄密等风险；当基于 ChatGPT 等闭源大模型时，核心算法难以做到自主可控，会影响知识服务大模型长远发展。

知识服务领域相关机构一般有自己的专业数据来源，前期尤其要考虑算力算法等相关基础设施对知识服务大模型的经济性、耗时和稳定性的影响。对于承担了国家公益职责的机构，其相关数据可能具有战略意义。事关国家安全，对大模型算力算法的要求更高。

3.3 多类型竞争

科技公司、科研机构等正在基于自身的通用大模型优势或者专业领域优势，利用免费或者经授权的行业数据库资源构建垂直大模型，将对知识服务机构构成竞争。OpenAI 推出的 AI 搜索引擎原型 SearchGPT，可提供对话式的搜索体验。OpenAI 与德国施普林格及美国《时代周刊》《大西洋月刊》、Vox Media、新闻集团等合作，不仅帮助用户发现优质内容，还为出版商带来更多流量，提供管理工具，让出版商可以控制其内容在 SearchGPT 中的展示方式。微软亚洲研究院发布的生物医学领域的大语言

模型 BioGPT^[25]，基于 GPT-2 大模型以及全球著名生物医学数据库 PubMed 中的 1500 万篇相关科学论文进行训练，能够理解专业术语、基因名、蛋白质序列等复杂概念，可快速准确地生成生物医学问题答案，完成文本挖掘、实验报告撰写、分子设计、文献综述撰写等任务。这类大模型一方面将对知识服务领域机构的大模型产生一定的影响，同时也将有利于促进知识服务领域相关机构不断提高自身大模型的应用水平，推动商业模式的发展。

3.4 地缘政治风险

中美竞争大格局下，美国以国家安全为名实施科技战。大模型自身的安全性以及所引发的地缘政治风险，对中国机构在通用大模型使用、公开的专业数据获取等方面带来挑战。OpenAI 于 2024 年 6 月宣布将阻止非支持国家和地区 API 服务，将对我国开发人员构成重大挑战，不仅限制其访问 GPT 等高端大模型，还可能对正在进行的项目和研究产生重大影响。在公开的专业数据获取方面，美国限制向中国、俄罗斯等 6 个国家相关的实体出售美国个人和政府的人类基因组数据等 6 类敏感数据，这可能会直接影响我国行业大模型研发过程中对美国知名数据库的使用，如 PubMed、“癌症基因组图谱”（TCGA）等。

3.5 可信性与可靠性

大模型的可解释性与可信度不足限制了其在知识服务领域的应用。首先，大模型的复杂性使得其决策过程难以被人类理解和解释。大模型通常包含海量参数，其内部复杂的神经网络

结构使得决策过程如同“黑箱”，难以用人类可理解的方式呈现其推理逻辑。在学术出版领域，内容的可追溯性和可验证性至关重要，模型生成的内容如果无法解释其来源或推理过程，可能难以被学术界接受。其次，幻觉现象、一致性问题。与聊天机器人等应用不同，在文献情报及知识服务领域，大模型生成与输入信息不符、缺乏事实依据的内容，可能会产生更加严重的后果，导致错误的文献解读或误导性的知识传播。编辑和审核人员需要花费更多时间和精力验证内容的准确性，增加了工作负担。最后，监管问题。大模型的复杂性和不透明性使监管部门难以进行全面、有效监管，这可能导致模型生成的内容存在学术不端等问题。

4 国内知识服务型行业的应对策略

垂直领域大模型有助于促进行业智能化转型升级，但对模型准确性和效果要求很高，国内知识服务机构在数据质量、算力与算法、安全与隐私以及多类型机构竞争等方面都存在一定挑战，需要从政策支持、高质量数据建设、大模型研发、标准化评估体系构建以及人才队伍等方面综合应对，以推动知识服务大模型的快速发展以及行业的进步。

4.1 加强相关政策支持和监管

随着各地网信部门陆续公布大模型备案信息，我国大模型治理已进入新阶段。由于大模型的固有缺陷，在知识服务领域应用存在诸多挑战。为推动大模型发展，需从研发支持、引导规范、加强监管和推动应用等方面系统发力。文献情报和知识服务领域具有专业性、知识密

集性、时效性和跨学科性，对信息的准确性、可信度和知识产权保护要求程度高。未来，在基础研究和技术创新政策方面需要提供政策支持，特别是加强可解释性和可信度的研究支持；并通过建设公共数据平台，丰富知识服务行业高质量、多样化的数据资源，支持领域大模型的训练和优化。同时在技术标准规范、数据安全、知识产权等方面加强政策支持和引导，建立数据安全保护体系，确保大模型应用符合标准规范，充分保护知识产权，推动大模型技术应用持续发展。

4.2 推进高质量数据资源建设

知识服务行业数据资源具有海量丰富、多样多样、动态复杂、高价值等特征，且多是同行评议的高质量资源及深度规范加工的高价值数据，对推进大模型研发应用具有重要意义。知识服务行业建设高质量数据资源关键在于构建高质量、安全、合规的数据集，并通过构建数据治理生态，推动数据驱动的创新应用，充分释放数据价值。高质量数据的特征主要包括：

- (1) 完整性，通过广泛的领域和多样情境以促进模型的全面理解，提高适应力；
- (2) 准确性，确保模型学习的正确模式，提升预测、分类和识别任务的精度；
- (3) 合规性，遵守法律法规和行业准则，确保数据处理的合法性和隐私保护；
- (4) 可操作性，数据应易于获取、处理和分析，确保高效应用于模型训练和业务决策；
- (5) 安全性，应采取保护措施防止数据泄露，确保数据安全；
- (6) 可审计性，保持透明可追溯，便于内

部审计和问题诊断，保障治理的严谨性。

4.3 研发知识服务垂直领域大模型

国家层面，对公益类机构的数据资源进行有效整合，推动建立知识服务领域可动态更新的高质量数据集，共同开发知识服务大模型，逐步推动算力、算法以及数据完全自主可控，促进行业数据安全、信息可信等。技术层面，知识服务行业的垂直领域大模型核心在于领域知识的深度嵌入与定制化开发，通过高质量科技文献、专利数据、技术报告等专业数据的训练，结合知识图谱和多模态处理技术，精准理解科技领域的专业术语、概念和逻辑关系，通过增量学习及时吸收新数据，保持对前沿技术的敏感性，提供全面准确的知识服务。市场层面，垂直领域大模型需根据不同用户的需求提供差异化服务，降低学习成本。构建开放协同的创新生态。通过与高校、科研机构、企业等建立产学研合作，共享资源，提供 API 工具吸引第三方开发应用，丰富应用场景，提升市场影响力。并积极参与国际合作与竞争，吸收全球科技信息资源的优势，提升全球竞争力。

4.4 建立大模型标准评估体系

当前大模型飞速发展，知识服务领域要适应并跟上技术进步的步伐，深入认识和理解大模型的能力及不足，预知并防范大模型的安全挑战和风险，结合实际需要针对知识服务大模型开展全方位评测，建立适应我国知识服务领域的标准化评估体系，以全面衡量知识服务大模型在实际应用中的表现，促进知识服务大模型的规范化发展和应用，提高大模型的应用水平。

技术性能方面,关注模型的准确性、覆盖率、响应速度和鲁棒性,确保满足多样化的知识服务需求。领域适配性方面,测试模型对专业知识理解能力、多模态处理能力和动态更新能力,确保适应复杂多变的实际场景。用户体验方面,关注模型的交互友好性、输出可解释性和个性化服务能力,提升用户满意度和使用效率。合规性方面,检查模型的数据隐私保护、知识产权合规性和内容安全性,确保符合相关法律法规和行业标准。社会影响方面,关注模型的公平性、可访问性和环境影响,推动向更加普惠和可持续的方向发展。

标准化评估体系的建立,不仅有助于提升知识服务大模型的技术水平和应用能力,还能够为行业监管和政策制定提供科学依据。通过全面、科学、可操作的评估,可以及时发现和解决大模型在实际应用中的问题,优化提升性能。评估体系还可以为用户提供参考,提高知识服务的效率和质量。

4.5 重视大模型人才队伍建设

随着大模型向垂直行业加快渗透,对于融合技术和业务能力的大模型人才的需求愈发凸显,当前急需结合我国知识服务领域发展需求,加快对我国知识服务大模型人才队伍的培育和储备,建立科研机构、企业等的人才培育联盟,加快培养与吸引大模型和知识服务领域的高素质人才,不断推动知识服务领域的技术融合与创新。

人才队伍建设需要重视构建多层次、多类型、高质量的人才培养和引进机制。除了需要培养和引进人工智能、自然语言处理、机器学

习等技术背景的人才,推动大模型在科技信息服务领域的前沿创新,还需注重领域知识交叉型人才的培养,既精通人工智能技术,又熟悉科技信息服务领域专业知识,将大模型技术与科技信息服务需求相结合,提升模型的领域适配性和应用价值。推动国际化人才引进与合作,吸引全球顶尖人才参与我国知识服务领域大模型的研发与应用,同时鼓励本土人才参与国际交流与合作,提升我国在全球人工智能领域的影响力和竞争力。在人才队伍建设模式上,注重多层次、多类型的人才培养体系构建,包括高等教育、职业培训、继续教育等。通过开设人工智能等相关专业课程,培养专业人才;针对大模型技术的实际需求,开展技能培训和认证,提升从业人员的专业水平。高质量的人才队伍将助力我国在全球科技竞争中占据领先地位,推动科技信息服务大模型在科研创新、技术转移、科技金融等领域的广泛应用,为我国经济社会高质量发展注入新的动力。

5 结语

在大模型技术迅猛发展的背景下,我国知识服务领域既迎来智能化升级的机遇,也面临多重挑战。一方面,大模型通过海量数据处理和智能生成能力显著提升了知识服务的效率与个性化水平,例如在医疗、金融、政务等领域的精准诊断、报告生成和资源配置优化中已取得初步成效。然而,技术门槛高、数据获取困难、成本效益平衡等问题成为制约其深度应用的核心瓶颈。国内知识服务机构在应用大模型时,既面临着通用大模型的固有缺陷,同时伴

随着专有大模型的特别问题,其中的关键挑战主要包括知识性不足、算力与数据安全双重制约、幻觉与可信度问题、多重竞争及地缘政治风险等。针对上述挑战,本文提出应加强相关政策支持和监管,同时推进高质量数据资源建设,加速国产知识服务大模型的研发,此外还要构建跨行业的评估体系,明确技术性能、数据合规性和经济效益的量化指标;最后,加强产学研合作,培养兼具技术能力与行业洞察的复合型团队,并通过开源生态降低技术应用门槛。唯有通过技术创新、制度完善与生态共建,我国知识服务领域才能在大模型时代实现从“追赶”到“引领”的跨越。

参考文献

- [1] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[J]. arXiv preprint arXiv:2108.07258, 2021.
- [2] FIROOZI R, TUCKER J, TIAN S, et al. Foundation models in robotics: Applications, challenges, and the future[J]. The International Journal of Robotics Research, 2023: 02783649241281508.
- [3] 廖俊伟. 深度学习大模型时代的自然语言生成技术研究[D]. 成都: 电子科技大学, 2023.
- [4] MASLEJ N, FATTORINI L, PERRAULT R, et al. Artificial Intelligence Index Report 2024[J]. arXiv preprint arXiv:2405.19522, 2024.
- [5] FANG S, WANG L, ZHENG C, et al. SignLLM: Sign languages production large language models[J]. arXiv preprint arXiv:2405.10718, 2024.
- [6] 罗锦钊, 孙玉龙, 钱增志, 等. 人工智能大模型综述及展望[J]. 无线电工程, 2023, 53(11): 2461-2472.
- [7] 钟新龙, 渠延增, 王聪聪, 等. 国内外人工智能大模型发展研究[J]. 软件和集成电路, 2024(1): 80-92.
- [8] 刘安平, 金昕, 胡国强. 人工智能大模型综述及金融应用展望[J]. 人工智能, 2023(2): 29-40.
- [9] IDC. 中国大模型市场主流产品评估 2024 [EB/OL]. (2024-06-12) [2024-06-18]. <https://www.idc.com/getdoc.jsp?containerId=CHC52324924>.
- [10] 刘细文, 孙蒙鸽, 王茜, 等. DIKIW 逻辑链下 GPT 大模型对文献情报工作的潜在影响分析[J]. 图书情报工作, 2023, 67(21): 3-12.
- [11] 张智雄, 于改红, 刘熠, 等. ChatGPT 对文献情报工作的影响[J]. 数据分析与知识发现, 2023, 7(3): 36-42.
- [12] 韩旭, 孙亚伟, 赵璐. 体系化人工智能与大语言模型在智能情报场景中的应用[J]. 北京邮电大学学报, 2024, 47(4): 11-19, 28.
- [13] CRAYON. 2024 State of Competitive Intelligence[R/OL] (2024-02) [2024-05-06]. <https://www.crayon.co/state-of-competitive-intelligence>.
- [14] 郭利敏, 付雅明. 以大语言模型构建智慧图书馆: 框架和未来[J]. 图书馆杂志, 2023, 42(11): 22-30, 133.
- [15] 袁虎声, 唐嘉乐, 赵洗尘, 等. ChatLib: 重构智慧图书馆知识服务平台[J]. 大学图书馆学报, 2024, 42(2): 72-80.
- [16] 李书宁, 刘一鸣. ChatGPT 类智能对话工具兴起对图书馆行业的机遇与挑战[J]. 图书馆论坛, 2023, 43(5): 104-110.
- [17] 洪贇, 叶鹰, 佟彤. 国内外大语言模型的图书情报应用探讨[J]. 图书馆理论与实践, 2024(2): 72-80.
- [18] 郭利敏, 付雅明. 融合 ReAct 模式的图书馆大语言模型知识服务系统构建[J]. 图书馆论坛, 2024, 44(6): 61-70.
- [19] 李晓松, 李增华, 赵柯然, 等. 科技情报研究领域的大语言模型测评工作思考[J]. 情报理论与实践, 2024, 47(11): 170-176, 200.
- [20] 张立, 吴素平, 周丹. 国内外知识服务相关概念追踪与辨析[J]. 科技与出版, 2020(2): 5-12.
- [21] 许洁, 袁小群, 朱瑞, 等. 基于大模型的轻量级智能出版知识服务: 理论基础与实现路径[J]. 中国数字出版, 2024, 2(1): 25-35.
- [22] JIN Q, YANG Y, CHEN Q, et al. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information[J]. Bioinformatics, 2024, 40(2): 1-8.
- [23] BAI Z, ZHANG R, CHEN L, et al. Patentgpt: A large language model for intellectual property[J]. arXiv preprint arXiv:2404.18255, 2024.
- [24] BI Z, ZHANG N, XUE Y, et al. OceanGPT: A large language model for ocean science tasks[J]. arXiv preprint arXiv:2310.02031, 2023.
- [25] LUO R, SUN L, XIA Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining[J]. Briefings in Bioinformatics, 2022, 23(6): bbac409.



开放科学
(资源服务)
标识码
(OSID)

基于扎根理论的政企数据协同 影响因素模型构建

刘彬芳

中央党校（国家行政学院）国家治理教研部 北京 100091

摘要: [目的/意义] 探究政企数据协同的影响因素及其关联关系, 有助于梳理政企进行数据协同的障碍和动力, 进一步挖掘多源数据公共价值, 提升政企数据协同赋能现代化治理的水平。[方法/过程] 通过扎根理论的程序化三级编码方法, 综合分析了国内外相关文献以及调研访谈数据, 在归纳演绎和逻辑推理的基础上, 根据政企数据协同的研究情境, 探索发现政企数据协同的影响因素, 进而构建政企数据协同影响因素的理论模型。[结果/结论] 研究得到包括 3 个主范畴和 13 个子范畴在内的 CPE 模型, 并探讨不同影响因素对政企不同主体的作用, 为后续政企数据协同相关研究、实践和政策制定提供重要的理论框架参考, 助力构建数据驱动的多元主体现代化治理体系。

关键词: 政企协同; 数据协同; 数据共享; 协同治理; 影响因素

中图分类号: G35; G203

Construction of a Model for Factors Influencing Government-Enterprise Data Collaboratives Based on Grounded Theory

LIU Binfang

National Governance Department, Party School of the CPC Central Committee (National Academy of Governance), Beijing 100091, China

Abstract: [Objective/Significance] Exploring the influencing factors of data collaboratives, analyzing the interrelationships between them, can help clarify the obstacles and motivations for government and enterprises data collaboratives, further explore the public value of multi-source data, and enhance the level of government enterprise data collaboratives empowering modern governance. [Methods/Processes] This paper uses the programmatic three-level coding method of grounded theory to comprehensively analyze relevant domestic and foreign literature, as well as interview data. Based on induction, deduction, and logical reasoning, according to the research scenario of data collaboratives between government and enterprises, the

基金项目 中共中央党校（国家行政学院）校级科研基金课题“政企数据协同赋能的公共服务与智慧决策研究”（2023QN066）。

作者简介 刘彬芳（1993-），博士，讲师，主要研究方向为数据治理、数字政府、数字治理，E-mail: liubf@ccps.gov.cn。

引用格式 刘彬芳. 基于扎根理论的政企数据协同影响因素模型构建 [J]. 情报工程, 2025, 11(1): 63-71.

influencing factors of government enterprise data collaboratives are discovered, and a theoretical model is constructed. [Results/Conclusions] This paper has developed a CPE model consisting of 3 main categories and 13 categories, and thoroughly explores the effects of different influencing factors on different entities of government and enterprises, which provides an important theoretical framework reference for subsequent research, practice, and policy formulation related to government enterprise data collaboratives, and helps to build a data-driven modern governance system with multiple entities.

Keywords: Collaboration between Government and Enterprises; Data Collaboratives; Data Sharing; Collaborative Governance; Influence Factors

引言

近期,《国家数据局等部门关于促进企业数据资源开发利用的意见》提出“推动建立新型政企合作机制”。《数字中国建设整体布局规划》也把数据要素价值有效释放作为目标之一,指出要畅通数据资源大循环,夯实数字中国建设基础。而早在2017年12月习近平总书记在中央政治局集体学习时就指出,要“统筹规划政务数据资源和社会数据资源”“推进同企业积累的社会数据进行平台对接”。随着社交媒体、物联网、人工智能等新技术的发展,与前互联网时代全社会数据资源主要集中于政府不同,谷歌、阿里、腾讯等大型互联网企业或私营主体逐渐积累了越来越多的数据,给政府共享和整合来自外部的企业数据以提升社会治理能力提出了全新的挑战。政府数据和企业数据作为主要的数据来源,必须进行跨区域的数据协同和深度共享,才能推动数据应用和价值发挥,提升国家公共治理能力和人民幸福感。在数字化时代,政企数据协同在政策要求和实践需求两方面都存在较大的变革,政府不再是唯一的社会行动者,需要动员整合社会资源为政府所用,发挥社会力量来共同解决社会公共事务,以有效改善公共服务和政府决策,提升

政企数据协同赋能现代化治理的水平。

1 研究现状

政企数据协同是一种新兴的政企合作机制,不同于以往的数据共享、数据开放和数据交易,来自公私不同背景的主体,通过建立机制或机构,在收集、共享、处理和利用数据方面进行合作,旨在利用未充分开发的数据(通常由企业实体持有)实现更大的公共利益,解决复杂的社会问题^[1-2]。目前国内有关政企数据协同的研究主要集中于:一是案例介绍,包括自动驾驶^[3]、电力^[4]等政企数据协同案例;二是现状描述,分析政企数据协同的问题和对策^[5-6];三是协同模式分析,包括平台化对接模式^[7]、企业数据向政府共享实践模式^[8]等;四是国外经验介绍,如梳理欧美企业数据向政府共享的政策体系和经验^[9-10]。孙道等^[11]基于案例探究了公众外部压力、电商销售规模、电商活跃度等因素对企业数据向政府共享意愿的影响,但该研究针对性较强,在影响因素研究的整体性、覆盖度和普适性等方面仍有待加强,影响因素和影响机制是未来需要进一步研究的内容^[12]。

在影响因素方面,现有研究主要集中在政府内部的数据和信息共享,涉及的影响因素主

要包括技术、组织和制度 3 个方面^[13]。而协同治理作为研究政企数据协同的另一理论视角，其影响因素大体上包括主体类^[14]、业务类^[15]、组织类^[16]3 个方面，缺少对数据和信息技术等方面的关注。总之，政企数据协同的影响因素和影响机制仍不明确。政企数据协同为数据共享活动带来了新的复杂性，私营部门将在治理中发挥更大的作用，通过形成新的结构、程序、流程和机制来组成新的工作方式，也会有新的影响因素和影响机制。因此，政企数据协同相关研究需要创新视角，在系统梳理影响因素及作用模型的基础上，拓展完善数据共享和协同治理等理论范畴和内容，构建对政企数据协同影响因素的整体认知，提升理论指导实践的参考价值。

2 研究方法与数据来源

本文使用扎根理论的系统化编码方法，从收集的文献、调研和访谈等数据中分析政企数据协同的影响因素理论模型。考虑到有关政企数据协同的研究仍在起步阶段，所以文献检索时更注重查全率，将“数据协同、数据共享、数据合作、政企信息化合作 / data sharing、

information sharing、data collaboration/collaboratives”“影响因素 / factors”等关键词在中国知网、Web of Science、Proquest 等数据库进行检索，共得到 2041 篇文献。对相关文献进行筛选：一是可靠性，剔除无参考文献等不规范文献，剔除篇幅短小的低质文献，剔除非同行评审论文；二是内容相关性，剔除和影响因素无关的文献，剔除与政府数据、企业数据或协同治理无关的文献；共留取强相关的高质量文献 24 篇。通过对引文进行滚雪球式的检索，补充获取文献 4 篇。进一步的，鉴于政企数据协同为新兴现象，在现有文献和理论研究不足，而实践和需求又较为丰富的情况下，在青岛、上海、浙江等地针对有关人员进行实地调研访谈，访谈内容主要包括政企数据协同项目的概况、对政企数据协同的整体评价、影响政企数据协同的因素、如何改进政企数据协同等。通过焦点小组访谈共获取 6 份访谈数据，形成约 1.4 万字访谈记录，从而对文献数据进行有效补充。在留取超过 20% 的数据做理论饱和度检验之外，对采集到的数据进行扎根理论的三级编码，即开放式编码、主轴编码、选择性编码，示例如表 1 所示，在进行理论饱和和检验后，构建形成政企数据协同影响因素理论模型。

表 1 纳入扎根理论三级编码的数据集示例

编码	标题	作者	年份
[1]	To share or not to share? Examining the factors influencing local agency electronic information sharing	Akbulut A Y, et al.	2009
[10]	政府基础信息资源跨部门共享的影响因素调查	龙健	2014
[12]	Exploring the determinants of cross-boundary information sharing in the public sector: An e-Government case study in Taiwan	Yang T, et al.	2014
[19]	基于 DEMATEL 模型的我国政府信息资源跨部门共享的关键影响因素研究	杨建梁等	2018
[26]	自愿性企业数据向政府共享的影响路径研究——基于模糊集定性比较分析	孙道等	2024
[27]	政府数据协同治理影响因素及其驱动—依赖机制研究	周霞等	2024

3 范畴提炼和模型建构

3.1 开放式编码

开放式编码指对具有相似主题的因素问题进行聚拢，发掘其属性并赋予相应概念，然后以新的方式重新组合起来。通过对收集的数据进行分解、比较、概念化和分类来确定相似性和差异，将每个从原始资料中产生的初始概念与从数据中提取的一个或多个文本段相关联，

为减少编码过程中个人偏见和认知的影响，尽量使用原文或原话作为标签形成初始概念。经过处理后，对 104 个初始概念进行重新分类组合范畴化，形成子范畴，开放式编码结果的示例如表 2 所示。通过开放式编码，对过去相关研究和访谈中涉及的影响因素进行提取和归纳，形成目标一致性、业务需求、数据基础、协同激励、协同成本、领导的支持、信任、沟通、反馈、组织机制、政策支持、法规限制和信息技术等 13 个影响因素的子范畴概念。

表 2 开放式编码结果示例

子范畴	初始概念	数据来源
B1 目标一致性	A1 共享总体规划、A2 共识、A3 价值观、A4 战略、A5 目标可及性、A6 跨部门统筹规划、A7 关于问题的一致认识、A8 共同理解、A9 共同动机、A10 群体绿色认同	[2][8][9][11][26]
B2 业务需求	A11 部门是否愿意向其他部门提供数据、A12 是否愿意接受和使用其他部门提供的数据、A13 对本部门的政策执行反馈需求、A14 对信息需求与管理实践的了解、A15 个人意识、A16 公共服务需求、A17 业务协作需求、A18 相互依赖性	[4][11][12][18][27]
B3 数据基础	A19 是否有能力提供数据、A20 信息质量、A21 信息误用、A22 信息的特征、A23 信息权属关系、A24 数据价值、A25 数据质量、A26 数据标准	[11][13][19][27]

3.2 主轴编码

主轴编码旨在发现和建立概念间的联系，形成一个具有关联关系和从属关系的层次结构。概念的主要类别和子类别可以由一组轴来表示，这组轴揭示了数据的模式和关系。通过主轴编码，概念上类似的代码将组合在一起。如果某些初始代码既不合适，又得不到足够数据的支持，便可被舍弃。通过主轴编码发现子范畴之间的潜在逻辑联系，进一步对开放式编码中产生的子范畴进行分类。本文根据不同范畴在概念层次上的相互关系和逻辑次序对 13 个子范畴进行归类，共归纳出协同基础、协同过程和协同环境 3 个主范畴，如表 3 所示。

表 3 主轴编码结果

主范畴	子范畴	范畴内涵
C1 协同基础	B1 目标一致性、B2 业务需求、B3 数据基础、B4 协同激励、B5 协同成本、B6 领导的支持	是促成政府和企业进行数据合作的必要准备，也是彼此协作的逻辑起点和基础条件因素。
C2 协同过程	B7 信任、B8 沟通、B9 反馈、B10 组织机制	是在进行数据协同治理的过程中，影响政企数据协同水平的影响因素。
C3 协同环境	B11 政策支持、B12 法规限制、B13 信息技术	是政府和企业进行数据协同所处环境的相关因素。

3.3 选择性编码

选择性编码通过不断分析识别研究中最重要、可以归纳总结其他代码的核心主题，将所有研究结果统一在该核心主题范围之内。

同时,选择性编码解释所识别概念之间的关系,分析核心范畴与主范畴及其他范畴的联结关系,发展出新的理论框架。本文通过系统处理概念类属之间的关系,基于上一步保留的概念,分析影响因素的联结,挖掘确定核心范畴即“政企数据协同的影响因素”,将绝大多数概念总括到一个具有涵盖性的理论框架中。本文主范畴的典型关系结构如表4所示,协同基础、协同过程和协同环境3个主范畴对政企数据协同水平存在显著影响;协同过程是直接影响因素,它直接决定政企数据协同的水平;协同基础影响政企数据协同的过程;协同环境则调节着协同基础、协同过程和协同水平之间的联结关系。以此关系结构为基础,本文建构和发展出一个全新的政企数据协同影响因素理论构架,即“基础—过程—环境”(Condition-Process-Environment, CPE)模型,如图1所示。

表4 主范畴的典型关系结构

典型关系结构	关系结构的内涵	参考来源
协同过程—协同水平	协同过程类影响因素对政企数据协同水平有显著影响。	[1][16][27]
协同基础—协同过程	协同基础条件对协同过程有显著影响。	[2][21]
协同环境—协同基础	协同环境类因素对协同基础有显著影响。	[19][27]
协同环境—协同过程	协同环境类因素对协同过程有显著影响。	[2][5]
协同环境—协同水平	协同环境类因素对协同水平有显著影响。	[6][10][22][24][27]

3.4 理论饱和度检验

当没有更多数据来形成新的类别,即概念类别已经足够并且完全覆盖了,同时现有类别之间的关系已由大量数据支持,则意味着理论

已经饱和。本文采用预留的20%数据进行理论饱和度检验。结果显示,对于政企数据协同的影响因素概念类别已经足够完全覆盖,没有新的类别形成,并且协同基础、协同过程和协同环境3个主范畴均没有发现形成新的重要范畴和关系,因而可以认为本文构建的“基础—过程—环境”理论模型通过了理论饱和度检验。

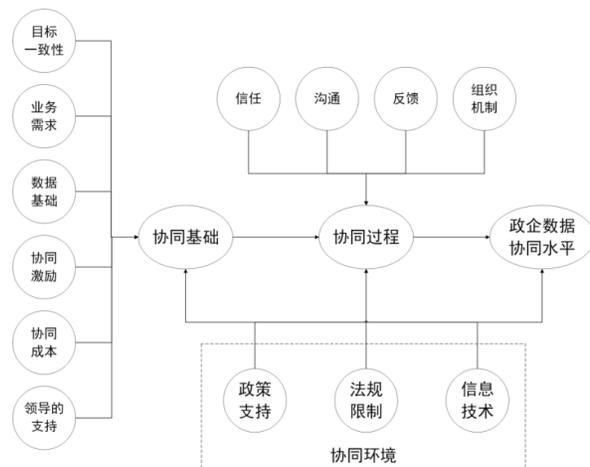


图1 政企数据协同影响因素的“基础—过程—环境”理论模型

4 政企数据协同影响因素分析

在政企数据协同的新情境中,政府等公共部门和企业等私营部门存在立场、价值观、专业水平等方面的不同,往往关注点和受影响的情况存在较大的差别,需要更加深入细致地探讨政企数据协同影响因素对不同主体的作用。下文分别从协同基础、协同过程和协同环境3个主范畴的维度出发,探讨各因素对政企数据协同的影响。

4.1 协同基础

合作开始时的条件可能会促进或阻碍利益相关者之间,尤其是公共部门和私营部门之间

的协同治理^[15]。协同基础是促成政府和企业进行数据合作的必要准备,也是彼此协作的逻辑起点,反映出相关主体参与数据协同活动的先决条件和意愿情况,主要包括目标一致性、业务需求、数据基础、激励、成本和领导支持等6个变量。

目标一致性是指政企数据协同中各主体具有相同或相似的目标和愿景。由于数据协同各主体之间立场、利益等方面的差异,对协同合作有不同的兴趣和动机,目标常常不完全一致,例如与企业相比,对于政府和很多公共组织来说,效率不是其主要目的^[17]。良好的数据协同需要达成目标一致性,而处于数据协同关系的政府和企业之间在进行数据合作时,目标和利益一致化程度越高,数据协同的效果就越好。这和 Soekijad 等^[18]研究有所呼应:相同的利益或目标能促进知识和信息资源的共享,并且目标结合程度会影响共享的深入程度。

业务需求是指政府和企业之间存在需要进行数据协同的业务。调查表明,政府在提供公共服务和解决公共问题方面存在权威性、资源支持等方面的优势,而企业在灵活性、信息技术等方面具有一定的优势。面对规模大、范围广、复杂性高的社会公共问题和服务项目,二者的优势互补性为业务合作创设了良好条件。已有研究表明内在和共同的业务需求是跨边界信息共享最基本、最持久的促进因素^[19]。政企之间要能进行数据协同并效果良好,必须基于业务产生数据合作意愿,不仅包括数据接收方的业务需求,还包括数据拥有方对数据的利用需求。

数据基础是政企数据协同的参与主体拥有和拟共享的数据资源情况。整个社会数据的急

剧增长成为促进政企数据协同的重要动力。研究证明数据质量和数据标准对管理机构共享数据的决定具有非常重要的影响^[20]。一些部门表示,因为不能保证数据的质量而无法共享数据。而保障信息资源质量等会促进政府信息资源跨部门共享^[21]。除了数据质量,数据资源的丰富度、信息知晓度和价值含量也是构成数据基础影响因素的重要内容。

协同的激励因素是激励政企数据协同的动力。没有持续的回报和奖励,数据协同的动力难以持久。激励不仅指物质激励,还包括非物质的激励。研究显示权力和资源的不平衡会影响群体参与协同过程的动机,当需要平衡合作所需的时间、精力和收益时,对协同结果的期望会影响动机^[22]。如果社会主体认为他们的投入仅仅是咨询性的或主要是仪式性的,激励效果就会减弱^[23]。同时,政企数据协同的声誉和公信力的提升也是重要的激励因素。

数据协同的实现需要参与方都付出一定的成本,不同参与主体对成本与收益有不同考量。公共部门考虑的成本因素包括工作量、权力资源、职员总数等。比如有受访者谈到:“成本收益比较是关键。协同谈判能否成功,就看成本收益。就是比谁花钱少,谁政治利益大。”企业在考虑自己的协同收益的同时,也会计算获得这些收益所需要付出的相应成本,不仅包括显现的外在化成本,还包括可能带来的风险和隐形成本。例如,可能导致核心数据的泄露、花费过多的时间以及丧失自身的竞争优势等,这些问题都会降低企业进行数据协同的意愿。

领导的支持与介入被许多受访者视作促进政企数据协同成功的最关键因素之一。在实践

中,不同层级都有领导支持数据协同工作,领导能力一直被视为驱动协同治理的关键因素^[16],高层领导关注后能带来截然不同的效果。对于企业来说,领导对数据协同也有着重要的影响。除了首席信息官(CIO)之外,其他领导的支持对政企数据协同项目的推动和实施也具有决定性作用。除了资金和人力,领导的支持还体现在对外沟通上,领导层面的协调沟通效果会远远好于基层员工之间的沟通协调。

4.2 协同过程

协同治理过程不能简单线性描述,协同过程应该是循环进行的,需要实现信任、沟通、结果等方面的良性循环。通过对协同过程的多个方面进行编码归类,本文将政企数据协同过程中广泛涉及的因素归纳为4个核心变量:建立信任、沟通、反馈和组织机制。

信任在政企数据协同中扮演重要角色。组织间的数据协同关系在很大程度上依赖于相关组织之间的信任建立,主要包括基于演算、身份和制度3种类型的信任^[24],可信度可由模型演算评估得出,可基于建立的长期个人关系,或者由制度、规范和法律法规保障。数据协同的实现需要建立信任机制进行保障,要建立一种能够保证各方利益的相互信任的机制^[25],既需要强制性的规范性约束,更离不开合作各方的伦理共识。

沟通是建立信任、相互尊重、共同理解和承诺的核心。政企数据协同过程中的沟通在某种程度上是一种跨文化沟通,是处于两种不同组织和业务背景下的政府和企业间进行的管理性质的信息沟通。Ansell等^[15]提出,直接对话

所允许的“密集沟通”对于利益相关者发现互利的机会是必要的。建立政企数据协同主要在于整合不同利益问题,通过沟通和协商对话,使各方在相互妥协和博弈的过程中达成共识,从而形成良好的利益共同体。

早期合作的反馈对政企数据协同会产生积极或消极的影响。由于政企数据协同过程是循环的,各影响因素的交互也拥有非线性特征。案例研究表明,如果协作能带来“小胜利”之类的反馈结果,那么合作就更有可能发生^[26]。过程成果反馈到政企数据协同过程中,能有助于建立成功协作的趋势,成为鼓励建立信任和承诺的良性循环。受访者曾说“他们看到了进行数据协同的实际作用,也就对我们的工作更加配合了。”协同过程中还应包括对协同问题和难点、工作进度等方面的反馈,进而及时调整协同过程,从而实现协同过程的良性循环。

组织机制指的是政企数据协同的基本协议和基本规则,保障数据协同的程序合法性。政企数据协同的治理结构需要考虑项目各参与者之间知识和信息的相互依赖性以及权责机制。王芳等^[27]调研发现,权责不清、问责机制不完善等问题影响着政府数据共享过程。不同的治理主体在不同的公共服务领域和层次中既有分工又有合作,并且各自的权责边界可能随着掌握资源与施行手段的不同不断发生变动,需要避免出现权力冲突或者治理真空^[28]。

4.3 协同环境

政策支持环境是指和政企数据协同实践相关的法律法规、政策制度和标准规范。其中法规政策对政企数据协同的目标、路径和边界进

行规定,是实践过程中程序正义的重要保障;标准是政企数据协同所参考的行动指南,是对实践过程中合规性的重要保障。研究表明,政府支持是自愿性企业数据向政府共享的核心条件^[1]。有关发挥数据的经济价值、推动信息公开和提供更优质公共服务方面的政策要求,也是推动政企数据协同的重要驱动力。

法规限制表现在推进政企数据协同面临缺少政策依据、缺少可操作的技术标准与合法性保证等问题。授权和立法作为制度合法性的支撑具有双重效果。尤其是政企数据协同过程中,公共部门组织可能只是众多参与者中的一个,对其他组织没有等级或立法控制^[29]。此外,相当一部分部门和企业的数据涉及国家安全、个人隐私与商业秘密。由于缺乏对不同数据进行分级数据共享和协作的明确法规标准,部分数据得不到充分利用,数据价值尚未充分发挥。

信息技术因素是指为实现数据协同所采取的信息技术措施,以及对某些具体信息技术的应用。政企数据协同可以被视为涉及信息系统建设、组织结构变革和业务流程再造的IT项目。Susha等^[29]研究表明,在政企数据协同中,由于异构的数据集和系统以及不同级别的准备工作,协作可能会更加复杂。但随着技术发展,数据协同的效率显著提高。开发标准、平台和应用程序互操作性、元数据以及算法的使用都在解决这一问题,而对安全的担忧成为影响政企数据协同的重要技术因素。

5 总结与讨论

政企数据协同是21世纪跨边界数据开发利

用和公私伙伴关系中的重要实践,通过政企数据交换共享利用能实现数据价值和公共利益,解决复杂的社会问题。本文系统研究梳理政企数据协同的影响因素及作用模型,创新研究视角,结合数据共享和协同治理两方面的理论视角,从更为细致的角度揭示不同因素的作用和影响,扩展相关研究边界范畴和内容,试图构建对政企数据协同影响因素的整体认知。未来还需要更多定量和实践研究来进一步丰富和完善这个模型。有关政企数据协同的研究也有待进一步深入,如:数据协同的主体间关系、参与动机、协同效益、协同风险、供需匹配机制、激励机制、协同模式、协同机制和相关技术等,从而为判断政企数据协同问题产生的原因和吸取政企数据协同成功经验提供理论支撑和指导。针对性地采取有效手段和措施来解决现有政企数据协同的影响因素问题,有助于提炼和推广政企数据协同优秀实践的经验,为相关政策制定提供参考,从而促进政企数据协同的发展和实践,进而构建现代化的多元主体社会治理体系。

参考文献

- [1] SUSH A I. Participation in open government[D]. Örebro: Örebro University, 2015.
- [2] KLIEVINK B, VAN DER VOORT H, VEENEMAN W. Creating value through data collaboratives[J]. Information Polity, 2018, 23(4): 379-397.
- [3] 李若兰. 企业数据向政府开放共享的价值逻辑和法治路径——以自动驾驶为例[J]. 行政管理改革, 2023(6): 65-72.
- [4] 王晓寅, 侯素颖, 徐帅, 等. 基于政企数据共享的电力创新服务模式探索[J]. 电力需求侧管理, 2018, 20(1): 59-60.
- [5] 郭明军, 童楠楠, 王建冬. 政务数据与社会数据共享利用中存在的问题及应对举措[J]. 中国经贸导

- 刊, 2019(8): 37-38.
- [6] 殷利梅, 赵令锐. 政企数据共享的现状与对策建议 [J]. 中国信息化, 2020(7): 111-112.
- [7] 于施洋, 郭明军, 王建冬, 等. 政务数据与社会数据平台化对接研究——模型构建与案例验证的视角 [J]. 情报理论与实践, 2020, 43(5): 74-79.
- [8] 孙道, 王盈盈. 企业数据向政府共享: 逻辑因由、实践模式与发展路径 [J]. 电子政务, 2024(3): 81-92.
- [9] 宋巍巍, 黄璜, 高银澜. 企业数据向政府共享: 美国实践及其对中国的借鉴 [J]. 中国行政管理, 2024(1): 100-111.
- [10] 翟云, 潘云龙. 政策设计视角下的 B2G 数据共享: 欧盟实践与政策启示 [J]. 中国行政管理, 2024, 40(9): 147-159.
- [11] 孙道, 王硕, 肖艳. 自愿性企业数据向政府共享的影响路径研究——基于模糊集定性比较分析 [J]. 情报科学, 2024, 42(1): 115-121, 142.
- [12] SUSHI I, M J A. Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development: 50th Hawaii International Conference on System Sciences (HICSS 2017), January 4-7, 2017 [Z]. Waikoloa, USA: 2017.
- [13] 李重照, 黄璜. 中国地方政府数据共享的影响因素研究 [J]. 中国行政管理, 2019(8): 47-54.
- [14] BRYSON J M, CROSBY B C, STONE M M. The Design and Implementation of Cross-Sector Collaborations: Propositions from the Literature[J]. Public Administration Review, 2006, 66(S1): 44-55.
- [15] ANSELL C, GASH A. Collaborative governance in theory and practice[J]. Journal of public administration research and theory, 2008, 18(4): 543-571.
- [16] EMERSON K, NABATCHI T, BALOGH S. An integrative framework for collaborative governance[J]. Journal of Public Administration Research and Theory, 2012, 22(1): 1-29.
- [17] 宋世明. 西方国家行政改革述评 [M]. 北京: 国家行政学院出版社, 1998.
- [18] SOEKIJAD M, ANDRIESEN E. Conditions for knowledge sharing in competitive alliances[J]. European Management Journal, 2003, 21(5): 578-587.
- [19] 郑磊. 跨边界信息共享中的领导力行为研究 [M]. 上海: 复旦大学出版社, 2012.
- [20] BIGDELI A Z, KAMAL M M, DE CESARE S. Electronic information sharing in local government authorities: Factors influencing the decision-making process[J]. International Journal of Information Management. 2013, 33(5): 816-830.
- [21] 杨建梁, 刘越男. 基于 DEMATEL 模型的我国政府信息资源跨部门共享的关键影响因素研究 [J]. 图书情报工作, 2018, 62(19): 64-76
- [22] GUNTON T I, DAY J C. The theory and practice of collaborative planning in resource and environmental management[J]. Environments, 2003, 31(2): 5-20.
- [23] FUTRELL R. Technical adversarialism and participatory collaboration in the US chemical weapons disposal program[J]. Science, Technology, & Human Values, 2003, 28(4): 451-482.
- [24] EMERSON K, NABATCHI T, BALOGH S. An Integrative Framework for Collaborative Governance[J]. Journal of Public Administration Research and Theory. 2012, 22(1): 1-29.
- [25] 兰旭凌, 沈霞. 政社合作危境规避管理研究 [M]. 成都: 四川大学出版社, 2017.
- [26] MCGUIRE M. Collaborative public management: Assessing what we know and how we know it[J]. Public Administration Review, 2006, 66(S1): 33-43.
- [27] 王芳, 储君, 张琪敏, 等. 跨部门政府数据共享: 问题、原因与对策 [J]. 图书与情报, 2017(5): 54-62.
- [28] 薛澜, 张帆. 治理理论与中国政府职能重构 [J]. 人民论坛·学术前沿, 2012(4): 6-15.
- [29] SUSHI I, PARDO T A, JANSSEN M, et al. A Research Roadmap to Advance Data Collaboratives Practice as a Novel Research Direction[J]. International Journal of Electronic Government Research, 2018, 14(3): 1-11.



开放科学
(资源服务)
标识码
(OSID)

教育—科技—人才一体化耦合协调发展研究 ——以山东省为例

姜明月 王舒扬 刘颖莹 赵林 张舒 刘进

山东省创新发展研究院 济南 252000

摘要: [目的/意义] 在中国式现代化背景下,教育、科技、人才耦合协调发展是建设现代化强省的重要前提。[方法/过程] 选取 2013—2022 年山东省 16 地市为研究对象,从创新投入、创新产出、教育规模、教育经费、人才储备水平等 9 个维度构建三系统评价指标体系,利用熵权法、耦合协调度模型、障碍度模型以及灰色预测模型,测算并预测 2023—2029 年山东省高等教育、科技创新、人才耦合协调度,全面揭示其时空格局演化及障碍因素。[结果/结论] 山东省高等教育、科技创新、人才发展水平整体上均呈现出增长趋势,科技创新涨幅>人才涨幅>高等教育涨幅,2018 年后涨幅明显;山东省 16 地市的三系统耦合协调度呈现出上升趋势,但整体耦合协调度不高,且不同地区的差异较为显著;山东省障碍因子主要集中在创新产出、人才发展规模、创新投入、人才资源和创新环境方面,其中创新产出和人才发展规模呈现出持续上涨趋势,成为影响耦合协调发展的关键;济南、青岛与其他地市的障碍因子具有较大差别;预测结果显示山东省 16 地市的耦合协调度整体上呈现出递增趋势,处于协调类型的地市数量持续增加。有助于推动构建高等教育、科技创新和人才协同联动机制,为山东省实施教育—科技—人才一体化战略提供参考依据和政策启示。

关键词: 教育—科技—人才; 耦合协调; 障碍因子; 预测分析

中图分类号: G35; F124; F224; G310

Research on the Integrated Coupling and Coordinated Development of Higher Education, Technological Innovation, and Talent Coupling: A Case Study of Shandong Province

JIANG Mingyue WANG Shuyang LIU Yingying ZHAO Lin ZHANG Shu LIU Jin

Shandong Academy of Innovation and Development Research, Jinan 252000, China

基金项目 山东省重点研发计划(软科学项目)“基于未来产业的概念验证中心建设初探”(2023RZB02022);山东省自然科学基金青年基金项目“山东省重大科技创新平台体系建设与新质生产力发展路径研究”(ZR2024QG217)。

作者简介 姜明月(1992-), 硕士, 助理研究员, 主要研究方向为科技战略、科技政策、科技统计分析;王舒扬(1992-), 博士, 副研究员, 主要研究方向为科技战略、科技政策分析;刘颖莹(1990-), 硕士, 助理研究员, 主要研究方向为区域科技创新、科技政策, E-mail: 781145934@qq.com;赵林(1980-), 学士, 副研究员, 主要研究方向为科技统计调查、科技人才政策;张舒(1999-), 硕士, 研究实习员, 主要研究方向为科技战略、科技政策分析;刘进(1991-), 博士, 助理研究员, 主要研究方向为科技创新、科技政策与战略管理。

引用格式 姜明月,王舒扬,刘颖莹,等.教育—科技—人才一体化耦合协调发展研究——以山东省为例[J].情报工程,2025,11(1):72-84.

Abstract: [Objective/Significance] In the context of the Chinese path to modernization, the coupling and coordinated development of education, technology and talents is an important prerequisite for building a modern province. [Methods/Processes] This study selected 16 cities in Shandong Province as research objects, with a time span from 2013 to 2022. An evaluation index system was constructed from 12 dimensions including innovation input, innovation output, education scale, education funding, and talent reserve level. The entropy weight method, coupling coordination degree model, obstacle degree model, and grey prediction model were used to calculate and predict the coupling coordination of higher education, scientific and technological innovation, and talent in Shandong Province from 2023 to 2029, comprehensively revealing their spatiotemporal pattern evolution and obstacle factors. [Results/Conclusions] Research has found that the overall level of higher education, technological innovation, and talent development in Shandong Province has shown an increasing trend, with technological innovation growth rate > higher education growth rate > talent growth rate, and a significant increase after 2018; The coupling coordination degree of the three systems in 16 cities in Shandong Province shows an upward trend, but the overall coupling coordination degree is not high, and the differences between different regions are significant. The main obstacles in Shandong Province are concentrated in innovation output, talent development scale, innovation investment, talent resources, and innovation environment. Among them, innovation output and talent development scale show a continuous upward trend, becoming key factors affecting coupled and coordinated development. There are significant differences in barrier factors between Jinan, Qingdao, and other cities. The prediction results show that the overall coupling coordination degree of 16 cities in Shandong Province shows an increasing trend, and the number of cities in the coordination type continues to increase. This study contributes to promoting the construction of a collaborative linkage mechanism between higher education, technological innovation, and talent, providing reference and policy inspiration for the implementation of the education-technology-talent integration strategy in Shandong Province.

Keywords: Education-Technology-Talent; Coupling Coordination; Obstacle Factors; Predictive Analysis

引言

习近平总书记在党的二十届三中全会提出，“教育、科技、人才是中国式现代化的基础性、战略性支撑”。推进教育、科技、人才三者统筹安排、一体部署，畅通教育、科技、人才的良性循环，是加快培育和发展新质生产力重要保障，也是实现中国式现代化的关键前提^[1-3]。然而，当前多数省份仍面临产教融合程度不深、科技人才结构性矛盾突出、区域发展不均衡以及资源配置不均衡等问题，使得教育、科技、人才之间的耦合协调仍存在不平衡、不协调的发展状况。因此有必要对科技、教育、人才三者之间的耦合协调情况进入深层次剖析，明确

当前发展情况和面临困境，深究问题根源，制定合理、针对性的对策建议，打通构建高等教育、科技创新和人才协同联动机制的难点堵点，实现三者高效耦合协调，为现代化强省建设提供强劲助力。

1 文献综述

学术界多是针对教育、科技、人才两两之间开展研究，对于三者的研究也多是从学理逻辑、耦合机理等方面开展。闫光才^[4]和何菊莲等^[5]对我国高等教育、人力资本进行测算分析。王海燕等^[6]研究了高校科技创新与区域创新能力耦合协调演变特征。高卉杰等^[7]基于全国视

角,运用耦合协调度模型分析人才聚集与科技创新耦合协调发展的时空演变规律。芮雪琴等^[8]运用协整检验和格兰杰因果检验等方法对人才聚集和科技创新的互动关系进行分析研究。刘在洲等^[9]和潘教峰等^[10]分析了教育、科技、人才“三位一体”战略发展的现实逻辑,剖析了三者的意义和作用。朱杰^[11]和段从宇等^[12]分析了教育、科技、人才的关系逻辑。李小球等^[13]和段从宇^[14]分析了教育、科技、人才三者之间理论逻辑。嵇慧敏等^[15]运用耦合协调模型和灰色关联模型,分析了三者的耦合协调度以及影响因素。黄永春等^[16]在全国视角下,运用耦合协调度模型和空间自相关分析等方法,分析了我国30个省区的教育、科技、人才的耦合协调度、影响因素以及演变趋势。蔡文伯等^[17]在全国视域下,构建了高等教育、人力资本和科技创新指标体系,并运用耦合协调度模型和GM(1,1)预测模型对三系统的耦合协调度进行测算和预测。

综上所述,现有研究已取得不少成果,但仍存在以下局限:一是研究多聚焦于教育、科技、人才之间的理论分析,缺乏对三者耦合协调的实证研究;二是部分系统评价体系指标设立不够全面,且运用主成分分析法或者赋值法确定权重,存在主观判断性,可能影响评价结果的有效性和准确性;三是当前三系统耦合协调研究缺乏全面性和深入性,多单纯对已有数据进行耦合性测算,较少同时进行耦合协调度测量、时空分析、障碍因子剖析以及未来预测的全方位分析;四是多是以国家视角开展研究,鲜少面向省域或者市域开展研究,缺少省域或市域层面的具体性、针对性对策建议。基于此,

本研究从创新投入、创新产出、教育规模等9个维度构建三系统评价体系,利用熵权法、耦合协调度模型、障碍度模型以及灰色预测模型,计算并预测2023—2029年山东省高等教育、科技创新、人才耦合协调度,全面揭示其时空格局演化及障碍因子,并立足山东实际提出对策建议。

2 研究设计

2.1 熵权法

(1) 标准化处理

$$Y_{ijt} = \begin{cases} \frac{x_{ijt} - \min x_j}{\max x_j - \min x_j} + L & \text{正向指标} \\ \frac{\max x_j - x_{ijt}}{\max x_j - \min x_j} + L & \text{负向指标} \end{cases} \quad (1)$$

式(1)中, x_{ijt} 为第 t 年 i 地市第 j 个指标值; Y_{ijt} 为 x_{ijt} 标准化后的数据; $\max x_j$ 、 $\min x_j$ 分别表示 j 项指标的最大值和最小值, $i=1,2,\dots,n$, $j=1,2,\dots,m$, $t=1,2,\dots,r$; L 取值为 10^{-5} 。

(2) 计算信息熵值

$$e_j = -k \sum_{t=1}^r \sum_{i=1}^n p_{ijt} \ln p_{ijt} \quad (2)$$

其中, $k = \frac{1}{\ln(rn)}$; $p_{ijt} = \frac{Y_{ijt}}{\sum_{t=1}^r \sum_{i=1}^n Y_{ijt}}$, n 为山东省地市总数, r 为年份总数。

(3) 计算权重系数

$$W_j = \frac{1 - e_j}{\sum_{j=1}^m (1 - e_j)} \quad (3)$$

(4) 计算综合得分

$$U_{it} = \sum_{j=1}^m W_j Y_{ijt} \quad (4)$$

2.2 耦合协调度模型

(1) 耦合度计算

$$C = \sqrt[3]{\frac{u_1 u_2 u_3}{[u_1 + u_2 + u_3 / 3]^3}} \quad (5)$$

其中, C 为三者之间的耦合度; u_1 、 u_2 、 u_3 分别为高等教育、科技创新、人才的发展水平。

(2) 耦合协调度计算

$$D = \sqrt{C(\alpha u_1 + \beta u_2 + \gamma u_3)} \quad (6)$$

其中, D 为耦合协调度, $D \in (0, 1]$, α 、 β 、 γ 为三系统的权重, $\alpha = \beta = \gamma = 1/3$ 。参考已有研究^[18-19], 提出三系统耦合协调度的划分标准, 如表 1 所示。

表 1 耦合协调度划分标准

耦合协调度	类型	区间
(0, 0.1]	极度失调	不可接受区间
(0.1, 0.2]	严重失调	不可接受区间
(0.2, 0.3]	中度失调	不可接受区间
(0.3, 0.4]	轻度失调	不可接受区间
(0.4, 0.5]	濒临失调	过渡区间
(0.5, 0.6]	勉强协调	过渡区间
(0.6, 0.7]	初级协调	可接受区间
(0.7, 0.8]	中级协调	可接受区间
(0.8, 0.9]	良好协调	可接受区间
(0.9, 1.0]	优质协调	可接受区间

2.3 障碍度模型

本研究引入障碍度模型深层次分析教育、科技、人才耦合协调关系的影响因子^[18]。

$$H_{ij} = \alpha_j w_{ij} (1 - Y_{ij}) / \sum_{j=1}^m (\alpha_j w_{ij} (1 - Y_{ij})) \quad (7)$$

其中, α_j 为第 j 个子系统权重; w_{ij} 为第 j 个子系统中第 i 个指标权重; Y_{ij} 为第 j 个子系统

中第 i 个指标的值; H_{ij} 为第 j 个子系统中第 i 个指标对三系统耦合协调发展的障碍度。

2.4 灰色预测模型

灰色 GM(1,1) 模型具有不需要大量样本且预测准确性高的优势^[19-20]。因此, 本文研究通过已有数据结果形成数列建立灰色模型, 以获取未来 7 年内三系统耦合协调发展的演变趋势。

(1) 初始序列设为:

$$x^{(0)} = \{x_{(1)}^0, x_{(2)}^0, \dots, x_{(k)}^0\} \quad (8)$$

(2) 灰色模型 GM(1,1) 为:

$$Y = B\mu \quad (9)$$

$$\text{式 (9) 中, } Y = \begin{bmatrix} x_{(2)}^{(0)} \\ x_{(3)}^{(0)} \\ \dots \\ x_{(k)}^{(0)} \end{bmatrix}, B = \begin{bmatrix} -z_{(2)}^{(1)} & 1 \\ -z_{(3)}^{(1)} & 1 \\ \dots & \dots \\ -z_{(k)}^{(1)} & 1 \end{bmatrix},$$

$\mu = \begin{bmatrix} a \\ b \end{bmatrix}$, a 称为发展系数; b 为灰色作用量; $z^{(1)}$ 为 $x^{(1)}$ 紧邻均值生成序列, $x^{(1)}$ 为初始序列进行一次累加, $z_{(k)}^{(1)} = 0.5x_{(k)}^{(1)} + 0.5x_{(k-1)}^{(1)}, k = 1, 2, 3, \dots, K$ 。

(3) 预测模型为:

$$\hat{x}_{k+1}^0 = \hat{x}_{k+1}^1 - \hat{x}_k^1, k = 1, 2, 3, \dots, K \quad (10)$$

3 指标体系构建及数据来源

3.1 三系统指标体系构建

根据科学性、系统性和全面性等原则, 本文在借鉴已有研究和充分考虑山东省实际情况的背景下构建了三系统综合评价指标体系, 包括教育规模、教育经费等 9 个二级指标, 又细化为 20 个三级指标。在高等教育中, 借鉴已有研究^[21-23], 并参考《山东省科技统计年鉴》高

等教育相关指标数据，从“教育规模—教育经费—教育产出”三个维度构建高等教育指标体系，包含人力、财力、物力以及成果产出等内容。科技创新指标体系已较为成熟，本研究从“创新投入—创新产出—创新环境”三个维度进行构建，运用 R&D 经费投入强度、科学技术支出占财政支出比重、R&D 人员全时当量表示创新投入^[24-25]，运用人均专利授权量、人均论文数量、技术市场成交合同额代表创新产出^[26-27]，而在创新环境中，参考程铭等^[28]研究，划分为职业学校师生比、规模以上工业企

业 R&D 人员数占就业人员比重细分指标。在人才体系建设方面，参照赵雪^[29]和刘颖等^[30]的研究，将科技人才总数作为人才储备水平，参考贺勇等^[31]和崔祥民等^[32]的研究，将具有本科学历以上的毕业生作为人才，并作为相应的人才结构指标。同时在参考黄永春等^[16]研究的基础上，考虑山东作为工业大省和国企体量大的特点，从规模以上企业 R&D 人数占总人口比重、规模以上企业 R&D 人员全时当量两个角度构建发展规模细节指标。具体情况如表 2 所示。

表 2 三系统指标体系

一级指标	二级指标	三级指标		
		名称	单位	指标属性
高等教育	教育规模	学校数量 (E1)	个	正向
		高校 R&D 人员数量 (E2)	人年	正向
	教育经费	高校 R&D 经费支出 (E3)	万元	正向
	教育产出	高校专利授权量 (E4)	件	正向
		高校发表科技论文 (E5)	篇	正向
		高校出版科技著作 (E6)	种	正向
		高校课题数量 (E7)	项	正向
科技创新	创新投入	R&D 经费投入强度 (I1)	%	正向
		科学技术支出占财政支出比重 (I2)	%	正向
		R&D 人员全时当量 (I3)	人年	正向
	创新产出	人均专利授权量 (I4)	件 / 万人	正向
		人均论文数量 (I5)	篇 / 万人	正向
		技术市场成交合同额 (I6)	亿元	正向
创新环境	职业学校师生比 (I7)	%	正向	
	规模以上工业企业 R&D 人员数占就业人员比重 (I8)	%	正向	
人才	储备水平	科技人才总数 (T1)	人	正向
	发展规模	规模以上企业 R&D 人数占总人口比重 (T2)	%	正向
		规模以上企业 R&D 人员全时当量 (T3)	人年	正向
	人才结构	本科 R&D 毕业生人数 (T4)	人	正向
		研究生 R&D 毕业生人数 (T5)	人	正向

3.2 数据来源

本研究数据主要来源于《统计年鉴》和《山东科技统计年鉴》，选取山东省16地市作为本次研究对象，时间跨度为2013—2022年。此外，考虑莱芜市于2019年正式并入济南市，故将2013—2019年莱芜市的数据也并入济南市。个别缺失数据采用线性插值法进行处理。

4 实证分析

4.1 三系统综合评价分析

2013—2022年，山东省高等教育、科技创新、

人才整体上均呈现出增长趋势，整体情况如图1所示，科技创新涨幅>人才涨幅>高等教育涨幅，2022年较2013年，分别增长了0.18、0.132和0.095。同时也可看出，科技创新和人才曲线2019年后增长速度明显快于2019年前，究其原因，2019年后山东省陆续实施科技强省战略和人才强省战略，相继印发《山东省人民政府关于加快推进新时代科技强省建设的实施意见》《山东省“十四五”科技创新规划》等政策文件，持续优化创新资源配置，搭建高水平创新平台，加大人才引育力度，不断壮大战略科学家、科技领军人才和青年科学家队伍。

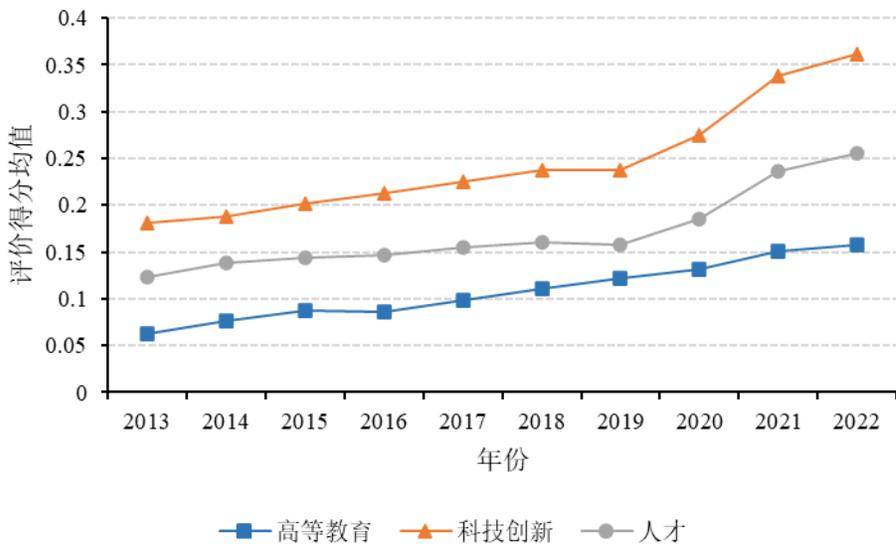


图1 2013—2022年山东省三系统综合评价结果

从个体来看，山东省高等教育、科技创新、人才发展水平存在明显的多级分化现象，其中济南和青岛始终位于发展头部，处于领先地位。在高等教育系统中，高等教育水平可划分为2个层级，第一层为济南、青岛，该地高等学府林立、教育实力雄厚，2022年得分分别为0.946和0.653；第二层级为烟台、潍坊、枣庄、菏泽

等地，高等教育资源匮乏，缺少高等学府，师资力量不足，教育水平滞后，评分分别为0.181、0.126、0.029以及0.019。在科技创新系统中，科技创新能力呈现出3个不同层级，其中第一层级为青岛、济南，两地深入实施创新驱动发展战略，科技创新实力雄厚，2022年综合得分为0.887和0.717；第二层级为烟台、潍坊、淄

博、威海等市，创新实力处于中等水平，得分分别为 0.476、0.452、0.397、0.366；第三层级为聊城、枣庄、菏泽等市，科技创新能力较低，得分分别为 0.234、0.158、0.158。在人才系统中，人才水平也划分为 3 个层级，第一层级为济南、青岛，人才水平遥遥领先，2022 年得分分别为 0.857 和 0.680；第二层级为烟台、潍坊、淄博等地，人才发展水平相对较弱，承受了济南和青岛两地产生的“荫蔽效应”，得分分别为 0.332、0.329、0.276；第三层级为聊城、枣庄、菏泽等地，人才发展水平较弱，经济较为落后，对科技人才吸引力低，得分分别为 0.126、0.081、0.076。可以看出，山东省教育、科技、人才水平较强的地区主要集中在鲁中和鲁东等经济发展较好

地区，鲁西和鲁南地区存在教育资源匮乏、教育水平不高、科技创新能力较弱、人才匮乏等多重困境。

4.2 三系统耦合协调度分析

4.2.1 耦合协调度时空演变分析

由表 3 可知，从整体来看，2013—2022 年，山东省教育、科技、人才的耦合协调度整体上呈现出上升趋势，2013 年耦合协调度均值为 0.273，2017 年为 0.348，2022 年为 0.439。但是，山东省整体耦合协调度不高，2022 年 75% 的地区耦合协调度低于 0.5。而且耦合协调度的总体跨度较大，最大值是最小值的 3.7 倍，说明不同地区的差异较为显著。

表 3 山东省 16 地市耦合协调度计算得分

地区	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	均值
济南	0.614	0.649	0.674	0.680	0.715	0.752	0.768	0.802	0.881	0.913	0.745
青岛	0.570	0.592	0.596	0.603	0.622	0.655	0.681	0.704	0.826	0.856	0.671
淄博	0.259	0.289	0.360	0.301	0.387	0.425	0.430	0.437	0.482	0.478	0.385
枣庄	0.156	0.182	0.168	0.180	0.196	0.204	0.210	0.229	0.257	0.268	0.205
东营	0.107	0.220	0.339	0.334	0.365	0.341	0.350	0.380	0.298	0.317	0.305
烟台	0.380	0.416	0.409	0.424	0.436	0.446	0.423	0.479	0.535	0.553	0.450
潍坊	0.337	0.367	0.394	0.383	0.386	0.400	0.408	0.454	0.510	0.515	0.415
济宁	0.298	0.334	0.335	0.331	0.352	0.358	0.337	0.363	0.412	0.408	0.353
泰安	0.366	0.378	0.392	0.382	0.375	0.368	0.395	0.381	0.381	0.396	0.382
威海	0.112	0.211	0.247	0.264	0.270	0.281	0.285	0.305	0.341	0.338	0.265
日照	0.095	0.171	0.188	0.194	0.222	0.258	0.254	0.271	0.276	0.277	0.221
临沂	0.235	0.246	0.265	0.291	0.284	0.285	0.271	0.296	0.355	0.368	0.290
德州	0.218	0.237	0.240	0.240	0.241	0.280	0.294	0.334	0.357	0.396	0.284
聊城	0.203	0.235	0.249	0.262	0.237	0.271	0.230	0.306	0.334	0.344	0.267
滨州	0.262	0.265	0.256	0.279	0.279	0.285	0.299	0.323	0.331	0.343	0.292
菏泽	0.158	0.177	0.184	0.189	0.198	0.186	0.175	0.191	0.229	0.246	0.193
均值	0.273	0.311	0.331	0.333	0.348	0.362	0.363	0.391	0.425	0.439	—

从个体来看，在高于全省耦合协调度平均值方面，2013 年，仅有济南、青岛等 6 个地市超过全省平均水平；2017 年，上涨为 8 个地市，

新增了淄博、东营两地；2022 年，济宁、泰安、东营跌出，仅有济南、青岛等 5 个地市高于均值。在耦合协调度地区分布方面，高耦合

协调度地区主要为济南、青岛、淄博等中东部地区，低耦合协调度地区则主要为枣庄、聊城、菏泽等鲁西南地区，且差距逐渐增大。在耦合协调度增幅方面，16个地市呈现出不同的增长态势，2013—2022年，威海（201.803%）、日照（190.622%）、东营（197.377%）等地增长快速，泰安（8.393%）、滨州（30.981%）和济宁（36.858%）等地增长缓慢。济南（48.663%）和青岛（50.175%）等耦合协调度领先的地区，在增长幅度方面表现不佳。

4.2.2 耦合协调度地区等级分析

由表4可知，2013年各地市三系统耦合协调度等级较低且分布差异显著，除济南、青岛处于勉强协调等级外，其他14个地市皆处于失调状态，其中，轻度失调地区占

21.429%，中度失调地区占42.857%，严重失调地区占28.571%，极度失调地区占7.143%。2017年，虽然多数地区耦合协调度都提升了等级，但仅有济南、青岛两地处于协调状态，其他地市皆为失调状态，其中东营跨度最为显著，由严重失调上升为轻度失调。2022年，16个地市耦合协调度等级又进一步提升，济南、青岛、烟台、潍坊处于协调状态，其他地市处于失调状态，其中濒临协调、轻度失调、中度失调的地区分别占比16.667%、58.333%和25%。综合可以看出，济南、青岛为耦合协调度的头部地区，凭借其经济实力、科技实力等发展优势，通过以点带面的辐射作用影响了其他地区，使得整体耦合协调度向更高等级发展。

表4 山东省16地市耦合协调度等级划分结果

等级	2013年	2017年	2022年	2029年
极度失调	日照			
严重失调	枣庄、东营、威海、菏泽	菏泽		
中度失调	淄博、济宁、临沂、德州、聊城、滨州	枣庄、威海、日照、临沂、德州、聊城、滨州	枣庄、日照、菏泽	菏泽
轻度失调	烟台、潍坊、泰安	淄博、东营、潍坊、济宁、泰安	东营、泰安、威海、临沂、德州、聊城、滨州	枣庄、东营、泰安
濒临失调		烟台	淄博、济宁	威海、济宁、日照、临沂、聊城、滨州
勉强协调	青岛		烟台、潍坊	
初级协调	济南	青岛		烟台、潍坊、德州
中级协调		济南		淄博
良好协调			青岛	
优质协调			济南	济南、青岛

4.3 障碍因子分析

根据障碍度模型分别计算出山东省整体障碍因子和各地市障碍因子演变趋势，分别如图2和表5所示。由图2可知，山东省排

名前5的障碍因子主要集中在创新产出、人才发展规模、创新投入、人才资源和创新环境方面，5类障碍因子在不同时间段内占比程度具有不同的变化趋势，其中，创新产出呈

现出明显上升趋势，人才投入呈现稳定下降趋势，人才发展规模呈现先下降后上升趋势，创新环境呈现出基本稳定趋势，2022年障碍度排序为创新产出（24.033%）>人才发展规模（17.643%）>创新环境（12.654%）>创新投入（11.776%）>人才结构（11.451%）。

这就说明，科技创新系统和人才系统是影响三系统耦合协调发展的关键因素，山东省要持续提升科技成果产出和成果转移转化水平，提升规模以上企业的研发能力，优化创新环境，增加创新投入，加大人才引进、培养力度，持续优化人才结构。

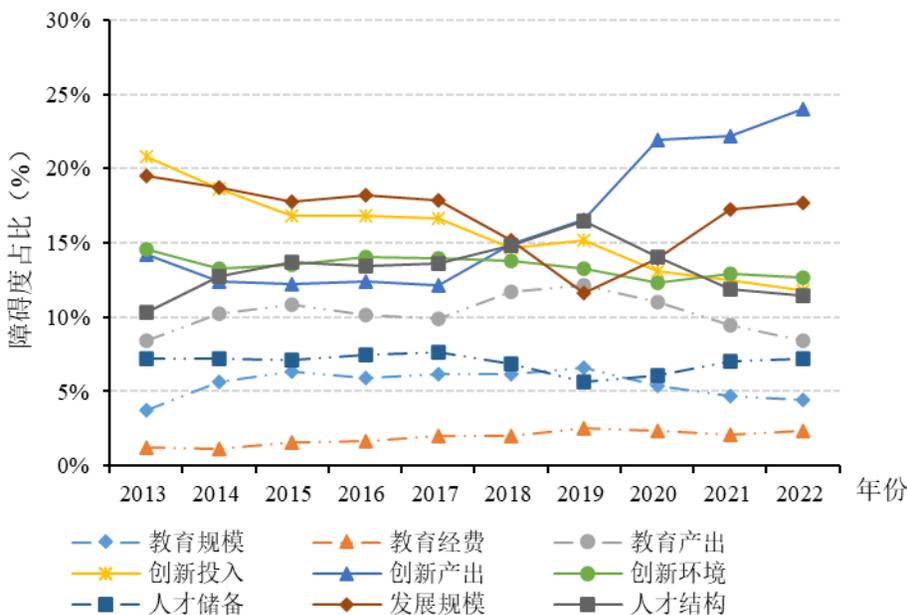


图2 山东省各障碍因子演变趋势

表5反映了2022年各市排名前5的障碍因子情况，从三系统障碍因子类型数量上来看，在80个主要障碍因子中，人才系统的障碍因子占比为56.25%，说明人才系统指标已经成为影响各地市三系统耦合协调发展的重要因素，各地市应该加大人才引育力度，持续优化人才结构，提高企业研发人才占比，进一步促进教育、科技、人才三系统耦合协调发展。同时济南、青岛与其他地市的障碍因子具有较大差别，济南、青岛两市障碍因子涉及三系统部分指标，主要为研究生R&D毕业生人数、科技人才总数、人均论文数量、技术市场成交合同额、高校R&D经费支出等，

因此两市应该加大教育经费支出强度，注重科技人才培养，增加高层次人才数量，提升科技成果产出和转移转化水平。而淄博、枣庄、东营等地市的障碍因子涉及科技创新和人才两个系统部分指标，障碍因子主要为技术市场成交合同额、规模以上企业R&D人数占总人口比重、规模以上企业R&D人员全时当量、研究生R&D毕业生人数、人均专利授权量等，说明该类地区应该着重完善成果转化服务体系，提升科技成果转移转化水平，加大规模以上企业引才力度，提升企业人才存量和企业研发核心竞争力，增加论文成果产出，提升高层次人才比重。

表5 山东省16地市主要障碍因子及障碍度

地区	1		2		3		4		5	
	障碍因子	障碍度 /%								
济南	T5	13.449	T4	9.738	I5	11.110	I6	7.633	E3	7.172
青岛	T5	8.646	T1	6.995	I5	16.005	E5	7.549	E4	6.567
淄博	T2	10.085	T3	7.448	T4	7.398	I6	13.497	I8	9.525
枣庄	T2	10.012	T4	7.096	I6	17.058	I4	11.961	I8	9.072
东营	T2	21.245	T3	7.204	I8	17.029	I4	13.531	I6	8.987
烟台	T3	9.026	T1	7.326	T2	7.220	I5	10.120	I6	10.043
潍坊	T3	12.737	T1	9.510	T4	8.551	I6	8.786	I3	7.385
济宁	T5	9.041	T3	8.423	T1	7.478	T4	7.233	I6	12.366
泰安	T5	8.277	T2	7.752	T3	7.077	T1	6.557	I5	8.659
威海	T2	17.845	I8	13.864	I4	12.819	I6	10.047	I7	7.599
日照	T2	11.551	I5	13.504	I6	10.912	I8	9.560	I4	9.554
临沂	T3	13.107	T1	11.815	T2	6.864	I6	12.982	I3	8.060
德州	T4	10.750	T2	10.329	T1	8.105	T3	7.903	I8	8.088
聊城	T5	9.230	T3	7.752	T2	7.460	I6	13.379	I4	8.320
滨州	T2	15.456	T3	9.048	T1	7.304	I8	12.898	I6	9.430
菏泽	T3	13.204	T1	10.860	T2	8.878	I6	15.970	I8	7.567

4.4 预测分析

基于2013—2022年已有数据，运用灰色模型预测出2023—2029年山东省16地市的预测值，预测结果及相应的耦合协调度等级划分结果分别如表6和表4所示。

从表6和表4可以看出，2023—2029年山东省16地市的耦合协调度整体上呈现出递增趋势，2023—2029年耦合协调度均值分别为0.452、0.473、0.493、0.511、0.529、0.547、0.566。从耦合协调度类型来看，截止到2029年，将有济南、青岛、淄博、烟台、潍坊、德州等6个地市处于协调类型，济南、青岛为优质协调；淄博为中级协调；烟台、潍坊、德州为初级协调；威海、济宁、日照等6个地市为濒临失调；东

营、泰安、枣庄为轻度失调；仅有菏泽为中度失调。从增长幅度来看，各地市增长幅度有所不同，2023—2029年，淄博、德州的增长范围为0.1~0.2，青岛、烟台等地为0.1~0.2，东营、泰安等地为0~0.1。由此可见，虽然山东省各地市高等教育水平稳步提高、科技创新水平不断上升、经济发展持续向好，但是由于人才培养能力、高层次人才引育、颠覆性技术创新、科技成果转化等仍存在制约，使得部分地市耦合协调度等级晋升跨度不明显，仍处于失调状态。因此，各地市仍需要针对自身情况，进一步提升科技人才引进和培养水平、完善科技创新机制、推进科技成果产出和转移转化，实现三者耦合协调发展的良好局面。

表 6 2023—2029 年山东省 16 地市耦合协调度预测

地区	2023	2024	2025	2026	2027	2028	2029	均值
济南	0.940	0.982	1.000	1.000	1.000	1.000	1.000	0.989
青岛	0.871	0.916	0.964	1.000	1.000	1.000	1.000	0.964
淄博	0.530	0.562	0.596	0.632	0.671	0.711	0.755	0.637
枣庄	0.279	0.296	0.314	0.333	0.353	0.375	0.398	0.336
东营	0.354	0.360	0.366	0.371	0.377	0.384	0.390	0.372
烟台	0.552	0.574	0.596	0.620	0.644	0.670	0.696	0.622
潍坊	0.527	0.551	0.576	0.603	0.631	0.660	0.690	0.605
济宁	0.410	0.422	0.434	0.446	0.458	0.471	0.484	0.446
泰安	0.388	0.389	0.390	0.391	0.392	0.393	0.394	0.391
威海	0.363	0.383	0.403	0.424	0.447	0.471	0.496	0.427
日照	0.313	0.333	0.353	0.375	0.398	0.423	0.449	0.378
临沂	0.366	0.382	0.399	0.417	0.436	0.455	0.476	0.419
德州	0.412	0.443	0.477	0.513	0.552	0.594	0.640	0.519
聊城	0.347	0.365	0.383	0.403	0.423	0.444	0.466	0.404
滨州	0.354	0.367	0.381	0.395	0.410	0.425	0.441	0.396
菏泽	0.233	0.242	0.250	0.259	0.268	0.278	0.287	0.260
均值	0.452	0.473	0.493	0.511	0.529	0.547	0.566	—

5 结论与建议

5.1 研究结论

(1) 山东省高等教育、科技创新、人才发展水平整体上均呈现出增长趋势，科技创新涨幅 > 人才涨幅 > 高等教育涨幅。山东省 16 地市教育、科技、人才发展水平较强的地区主要集中在鲁中和鲁东等经济发展较好的地区，鲁西和鲁南地区存在教育资源匮乏、教育水平不高、科技创新能力较弱、人才匮乏以及大量人才外流等多重困境。

(2) 山东省 16 地市的教育、科技创新以及人才的耦合协调度整体上呈现出上升趋势，但整体耦合协调度不高，耦合协调度的总体跨度较大，不同地区的差异较为显著。高耦合协调度地区主要为济南、青岛、淄博等中东部地区，

低耦合协调度地区则主要为德州、聊城、菏泽等鲁西南地区，且差距逐渐增大。

(3) 山东省障碍因子主要集中在创新产出、人才发展规模、创新投入、人才资源和创新环境方面，其中创新产出和人才发展规模呈现出持续上涨趋势，成为影响山东省地区耦合协调发展的关键。从各市主要障碍因子类型可以看出，济南、青岛与其他地市的障碍因子具有较大差别，人才系统指标已经成为影响各地市三系统耦合协调发展的重要因素，各地市应该加大人才引育力度，持续优化人才结构，提高企业研发人才占比。

(4) 预测结果显示山东省 16 地市的耦合协调度整体上呈现出递增趋势，处于协调类型的地市数量持续增加。截止到 2029 年，将有济南、青岛等 6 个地市处于协调类型，其中济南、

青岛为优质协调,淄博为中级协调,烟台、潍坊、德州为初级协调,威海、济宁、日照等6个地市为濒临失调,东营、泰安、枣庄等地为轻度失调;仅有菏泽为中度失调。

5.2 政策启示

(1) 提升高等教育水平,发挥高等教育支撑作用

一方面,经济实力雄厚、教育资源丰富的济南和青岛两市,应该着重推进基础研究和原始创新,加快构建高能级平台建设,加强双一流学校和学科建设,围绕山东强省战略需求,加快颠覆性技术和“卡脖子”技术突破,实现高水平科技自立自强。另一方面,其他地市可以探索引入大学机制,抢占稀缺的高校优质资源,高等教育深度对接本地产业发展需求,持续开展产学研全方位合作。

(2) 强化科技创新水平,发挥科技创新推动作用

一是强化核心技术攻关。山东应增加科研经费投入,加快推进基础创新平台、产业创新平台建设,以国家级科技示范工程和省级科技示范工程为抓手,加强企业创新主体地位,聚焦重点领域持续实现关键核心技术和重大原创性成果突破。二是加强区域创新协同发展。发挥济南和青岛作为创新核心区域的辐射带动作用,以项目合作、资源共享和人才交流等方式加强各地市之间的联系,构建优势互补、科技协同的发展局面。三是畅通科技成果转化渠道。完善科技成果转化机制,推动科技成果与市场紧密结合,优化“山东好成果”遴选机制,高水平建设“山东科技大市场”,实现联动全国、

辐射黄河、覆盖重点产业链的布局。

(3) 完善人才引进模式,发挥人才主体作用

一是以高能级载体集聚领军人才。发挥国家级、省级高能级平台的人才载体功能,实施“一事一议”直通车机制,深化提升泰山人才工程广纳贤士,“软硬兼施”引进顶尖人才团队,填补重点产业领域领军人才空缺。二是产教融合加强青年人才培养。鼓励企业与高校院所共建人才培养基地,面向人工智能、量子科技等重点产业,在有基础、有潜力的高校院所设立“基础研究特区”,打造一批具有全国特色的前沿交叉学科,持续培养年轻科学家,形成青年人才储备库。三是优化科技人才保障机制。尝试革新科技人才评价机制,鼓励省内重点科研单位积极探索,为科技人才发展晋升、持续发挥作用提供制度保障。

(4) 强化三系统整合思维,实现“三位一体”螺旋发力

要因地制宜,强化三系统整合思维,促进三者同轨同趋、交叉协调。一方面,济南、青岛作为耦合协调度的头部地区,应发挥其以点带面的辐射作用,带动其他地市耦合协调度向更高等级发展。另一方面,各地市仍需要针对自身情况和障碍因子分析,进一步优化高等教育结构、提升科技人才引进和培养水平、完善科技创新机制,实现三者耦合协调发展的良好局面。

参考文献

- [1] 阙明坤,沈阳.教育、科技、人才“三位一体”统筹推进的价值意蕴、内在机理和实践路径[J].现代教育管理,2024(9): 22-33.
- [2] 陈书洁.教育、科技、人才一体化赋能新质生产力:趋势、挑战与应对[J].人口与经济,2024(4): 9-14, 18.

- [3] 翁铁慧. 发挥教育、科技、人才在推进中国式现代化中的支撑作用[J]. 中国高校社会科学, 2024(4): 4-12, 156.
- [4] 阎光才. 学校教育与科技人才培养[J]. 中国高教研究, 2023(10): 17-24.
- [5] 何菊莲, 陈郡, 梅焯. 基于经济高质量发展理念的我国高等教育人力资本水平测评[J]. 教育与经济, 2021, 37(6): 44-52.
- [6] 王海燕, 苏博谦. 高校科技创新与区域创新能力耦合协调的时空演化研究[J]. 中国科技论坛, 2023(9): 144-154, 162.
- [7] 高卉杰, 李正风, 任莎莎, 等. 科技人才聚集与区域科技创新的耦合协调度研究[J]. 数学的实践与认识, 2018, 48(12): 109-118.
- [8] 芮雪琴, 李环耐, 牛冲槐, 等. 科技人才聚集与区域创新能力互动关系实证研究——基于2001—2010年省际面板数据[J]. 科技进步与对策, 2014, 31(6): 23-28.
- [9] 刘在洲, 汪发元. 教育、科技、人才一体推进的内在逻辑与实践方略[J]. 中南民族大学学报(人文社会科学版), 2024, 44(11): 187-196, 204.
- [10] 潘教峰, 左晓利. 教育科技人才一体推进: 内在逻辑、理论框架与实践路径[J]. 科教发展研究, 2023, 3(4): 20-32.
- [11] 朱杰. 推进教育、科技、人才“三位一体”融合发展[J]. 中国高等教育, 2023(9): 10-13.
- [12] 段从宇, 胡礼群, 张逸闲. 中国式现代化进程中教育、科技、人才三者关系的科学识辨与正确处理[J]. 教育科学, 2023, 39(2): 48-55.
- [13] 李小球, 宋杰. 教育、科技、人才“三位一体”发展的内涵特征及其圈层体系构建研究[J]. 当代教育论坛, 2024(3): 17-24.
- [14] 段从宇. 中国式现代化进程中教育、科技、人才一体推进的理论逻辑与实施路径[J]. 学术探索, 2023(3): 124-128.
- [15] 嵇慧敏, 张鹏. 教育、科技、人才耦合协调发展及其影响因素探究——以山东省为例[J]. 济宁学院学报, 2024, 45(1): 48-55.
- [16] 黄永春, 钱春琳, 钱昕怡, 等. 教育—科技—人才耦合协调度研究——基于30个省级区域的实证分析[J]. 中国科技论坛, 2024(10): 12-24.
- [17] 蔡文伯, 龚杏玲. 我国高等教育、人力资本与科技创新的耦合协调研究[J]. 高校教育管理, 2024, 18(4): 13-29, 59.
- [18] 周筱扬, 左国存. 我国中部地区科技创新与经济高质量发展耦合协调度的时空演化[J]. 科技管理研究, 2022, 42(22): 77-85.
- [19] 陈明华, 李倩, 王哲, 等. 中部地区城市经济高质量发展与生态可持续耦合研究[J]. 城市问题, 2022(4): 77-86.
- [20] 杨国华, 颜艳, 杨慧中. GM(1,1)灰色预测模型的改进与应用[J]. 南京理工大学学报, 2020, 44(5): 575-582.
- [21] 高金岭, 海颖. 我国高等教育、人力资本、科技创新与区域经济耦合协调的时序演变及空间集聚性分析[J]. 黑龙江高教研究, 2024, 42(9): 49-57.
- [22] 贺建清, 刘德鹏. 高等教育发展与共同富裕耦合协调水平测度及时空分异演变分析[J]. 高教探索, 2024(3): 48-57.
- [23] 赵志强. 高等教育高质量发展水平测度与时空演变[J]. 统计与决策, 2023, 39(16): 37-42.
- [24] 王守文, 赵敏, 章杰嘉, 等. 长三角城市群科技创新合作的空间关联网络特征及影响因素分析[J]. 统计与决策, 2023, 39(9): 91-96.
- [25] 金巍, 蒋薇. 长江经济带人口集聚对科技创新影响的空间异质性研究[J]. 长江流域资源与环境, 2025, 34(1): 14-27.
- [26] 刘青, 张晓南, 李曼. 山东省科技创新效率的时空演变、区域差异及集聚性评价研究[J]. 甘肃科学学报, 2024, 36(4): 116-125.
- [27] 汪树坤, 赵婷婷. 我国区域高等教育发展水平与科技创新效率的耦合协调关系及空间溢出效应[J]. 教育与经济, 2024, 40(6): 37-47.
- [28] 程铭, 袁文琦, 姜媛, 等. 科技创新与经济高质量发展耦合协调的时空演变分析及障碍因素诊断——以山东省为例[J]. 情报工程, 2024, 10(3): 14-27.
- [29] 赵雪. 科技人才聚集、科技产出与绿色GDP的互动关系研究——以京津冀区域为例[J]. 生态经济, 2024, 40(2): 67-74.
- [30] 刘颖, 王野, 曹琦. 科技人才聚集演化与区域协同治理——基于京津冀城市群的计量分析[J]. 中国行政管理, 2024, 40(7): 50-60.
- [31] 贺勇, 廖诺, 杨倩霞. 基于聚类分析和Cobb-Douglas函数的我国人才经济贡献率测算[J]. 数学的实践与认识, 2014, 44(19): 130-138.
- [32] 崔祥民, 张子煜, 裴颖慧. 江苏省“科技人才—科技创新—经济发展”复合系统协同发展评价体系构建及其协同发展水平评价分析[J]. 科技管理研究, 2023, 43(16): 52-62.



开放科学
(资源服务)
标识码
(OSID)

老年慢性病患者在线健康替代搜寻影响因素研究

王超¹ 卢智增² 徐文博³

1. 辽宁工业大学图书馆 锦州 121000;
2. 桂林理工大学公共管理与传媒学院 桂林 541000;
3. 锦州医科大学附属第一医院 锦州 121000

摘要: [目的/意义] 随着我国人口老龄化加剧,老年慢性病患者对健康信息的需求日益迫切。旨在从老年慢性病患者的视角出发,探讨其在线健康信息替代搜寻行为的影响因素,为优化老年人健康信息服务提供参考。[方法/过程] 采用质性研究方法,通过半结构化访谈收集23名老年慢性病患者的数据,利用NVivo12软件进行扎根理论分析,通过三级编码构建影响因素模型,并采用解释结构模型厘清各因素间的层级关系。[结果/结论] 老年慢性病患者的在线健康信息替代搜寻行为主要受自我效能、情感需求、信息价值判断、健康状态和电子健康素养等因素影响。其中,自我效能和情感需求为直接因素,信息价值判断为关键因素,健康状态和电子健康素养为根源因素。研究为信息服务机构优化老年人健康信息服务提供了启示,建议加强信息素养培训、优化信息平台设计、鼓励家庭参与及提供个性化服务。

关键词: 老年人; 替代搜寻; 扎根理论; 解释结构模型

中图分类号: G35

Research on the Influencing Factors of Online Health Information Surrogate Seek in the Elderly Patients with Chronic Diseases

WANG Chao¹ LU Zhizeng² XU Wenbo³

1. Library of Liaoning University of Technology, Jinzhou 121000, China;
2. School of Public Administration and Media, Guilin University of Technology, Guilin 541000, China;
3. The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121000, China

Abstract: [Objective/Significance] With the intensification of population aging in China, the demand for health information among elderly patients with chronic diseases is becoming increasingly urgent. This study aims to explore the influencing factors of online health information alternative search behavior from the perspective of elderly patients with chronic diseases, providing

基金项目 国家社会科学基金一般项目“农村人居环境与居民健康协同治理研究”(22BKS009)。

作者简介 王超(1983-), 硕士, 副研究馆员, 主要研究方向为信息素养教育; 卢智增(1973-), 博士, 教授, 主要研究方向为边疆地区健康治理, E-mail: hjllzz@126.com; 徐文博(1987-), 硕士, 副主任护师, 主要研究方向为健康教育。

引用格式 王超, 卢智增, 徐文博. 老年慢性病患者在线健康替代搜寻影响因素研究[J]. 情报工程, 2025, 11(1): 85-94.

references for optimizing health information services for the elderly. [Methods/Processes] A qualitative research method was adopted. Data from 23 elderly patients with chronic diseases were collected through semi-structured interviews. NVivo12 software was used for grounded theory analysis. A model of influencing factors was constructed through three-level coding, and the hierarchical relationship among the factors was clarified using the interpretive structural model. [Results/Conclusions] The online health information alternative search behavior of elderly patients with chronic diseases is mainly influenced by factors such as self-efficacy, emotional needs, information value judgment, health status, and e-health literacy. Among them, self-efficacy and emotional needs are direct factors, information value judgment is a key factor, and health status and e-health literacy are root factors. The study provides implications for information service institutions to optimize health information services for the elderly. It is suggested to strengthen information literacy training, optimize the design of information platforms, encourage family participation, and provide personalized services.

Keywords: Elderly; Surrogate Seek; Grounded Theory; Interpretative Structural Model

引言

我国人口老龄化不断加剧，老年慢性病患者数量也不断增加，老年人健康服务成为大众普遍关注的话题。有统计数据显示，我国有六成左右的老年人患有不同程度的慢性疾病^[1]，而这些老年人对在线询医、疾病预防、医疗保险等健康信息有较为迫切的需求^[2]。随着互联网的深度普及，“健康信息替代搜寻”理念在国外兴起，其核心理念是家属或社区医生作为替代者为功能障碍或知识技能较低的老年患者搜寻网络健康信息并为其提供简单护理服务。近年来“替代搜寻”在国内也逐渐普及，结合国内学者的研究^[3-5]和研究者的前期调查^[6]，研究者认为，替代搜寻行为应指非医学专业或者从事非医疗相关职业的人员源于某种动因或伦理关系主动替代老年人在互联网上搜寻健康信息的行为。在许多地区尤其是农村地区，老年人对健康信息的获取渠道有限，他们更依赖于子女或其他家庭成员来帮助他们搜索健康信息和获取健康建议。因此，对老年人线上健康信息替代搜寻的研究具有重要意义，可以为老年人

提供更好的健康信息获取和管理方式，并提高老年人的健康水平和生活质量。近年来学者们开始关注老年人线上健康信息替代搜寻并展开相关研究。

1 文献回顾

目前国内外学者关于健康信息“替代搜寻”的研究仍处于探索阶段，已有研究成果不算丰硕，而且尚未形成较为完整的理论体系。“替代搜寻”（Surrogate Seek）最早出自 McKenzie^[7]关于人们日常信息搜寻模式的研究。健康信息替代搜寻行为具有显著的情感纽带和社会关系特征。Cutrona 等^[8]将健康信息替代搜寻定义为非医学专业人员被要求或自愿替代他人搜寻健康信息的行为，这种行为通常发生在家庭成员、朋友或同事之间。McKenzie^[7]指出健康信息替代搜寻包括行为触发、信息源选择等七个阶段，涉及替代搜寻者和被替代搜寻者两类主体。替代搜寻者多为已婚、女性、健康状况良好且搜索经验丰富的个体，而被替代者则主要是具有健康问题的信息弱势群体，如老年人、

低收入者、低文化者及农村居民等^[6]。此外，替代搜寻可分为主动和被动两种形式，主动替代搜寻源于替代者的关心与利他主义，信息需求描述不明确；而被动替代搜寻则受被替代者明确的健康信息请求驱动^[9]。

替代搜寻行为受多种因素的影响，可归纳为替代搜寻者的自然特征、相关特征、内在动因、被替代者的特征以及外部环境因素。替代搜寻者的自然特征，包括年龄、性别、学历和婚姻状况，显著影响其搜寻行为；宋小康等^[3]发现已婚中年女性更倾向于为家人搜寻健康信息，这可能与她们在家庭中的角色定位及对健康问题的关注程度有关。相关特征，即替代搜寻者的健康状况、媒介技能和与被替代者的亲疏关系等；健康状况较好且具有较高媒介素养的个体更可能充当信息替代者，而亲情关系、同事情谊等情感纽带也在一定程度上强化了搜寻行为的动机^[11]。内在动因方面，同理心、利他主义和心理满足感是驱动替代搜寻的重要心理因素；Cutrona等^[12]认为这些内在动因不仅能增强替代者的搜寻动力，还可能影响其在搜寻过程中对信息来源的选择和信任程度。被替代者的特征也是一个不可忽视的影响因素；老年人、低收入人群或低文化水平者等健康信息弱势群体，在搜寻健康信息时面临诸多障碍，因此更倾向于依赖家庭成员或亲密关系中的他人完成信息搜寻^[13]。此外，外部环境因素，如信息资源的可用性、搜寻平台的便捷性和信息技术的发展，也对替代搜寻行为产生重要影响。虚假信息的泛滥可能降低替代者对网络信息的信任度，从而影响其搜寻效果。同时，社会支持网络的规模和质量也会影响替代搜寻的频率

和效果。

替代搜寻行为的理论支持主要来源于健康信念模型、社会支持理论以及信息搜寻相关模型和理论。健康信念模型认为，个体的健康行为受到其对健康问题的感知、对疾病严重性的认知等因素的影响。在替代搜寻行为中，替代搜寻者可能基于对被替代者健康问题的感知和认知，从而决定是否进行替代搜寻。社会支持理论则强调个体在社会网络中获得的支持对其行为和心理状态的影响。替代搜寻者往往从亲朋好友、医疗专家等社会支持网络中获取信息和建议，这些支持不仅增强了他们对健康信息的信任度，还促进了替代搜寻行为的实施。此外，ELIS模型和CMIS模型也对替代搜寻行为具有解释力。ELIS模型聚焦于日常生活信息搜寻行为，强调信息需求的情境性特征，有助于理解替代者如何在不同情境下调节搜寻策略。CMIS模型则强调信息搜寻过程中的动机、背景和能力因素，能够解释替代搜寻如何在满足被替代者需求的同时强化替代者的自我效能感^[14]。

基于已有研究，健康信息替代搜寻主动行为及其影响因素应为当前的研究重点之一。此外，已有研究主要是从替代者（近亲属）视角展开研究，仅陈明红等^[15]从被替代者（老年人）视角探讨替代搜寻行为，但其研究对象的选取具有随机性，也未明确区分替代搜寻行为的动因，即出于“关爱”的主动行为或源于“求助”的被动行为。本文为了弥补已有研究的不足，选择老年慢性病患者为研究对象，充分考虑健康信息搜寻情境、老年人群体的特殊性，探讨健康替代搜寻行为受到哪些因素影响（主要指与老年人相关的影响因素，如老年人的健康

状况、自我效能、亲情关系等),各因素之间的层次关系如何,以期丰富理论研究,并为优化老年慢性病护理和老年人健康服务提供借鉴,进而推动《健康中国行动(2019—2030)》^[16]。

2 对象与方法

2.1 研究对象

2022年4~6月,便利选取某医科大学附

属医院收治的慢性病患者为访谈对象。纳入标准:(1)患有1种以上慢性疾病(根据ICD-10标准);(2)年龄 ≥ 60 岁;(3)有自主搜寻或亲属替代搜寻网络健康信息的经历;(4)意识清楚、能够准确表达想法;(5)自愿参与访谈。最终确定23名老年人(67.87 ± 4.60 岁)为研究对象,符合质性研究(扎根理论)建议20~30的访谈样本量^[17]。研究对象详细情况如表1所示。

表1 受访者一般资料(n=23)

基本信息	性别		年龄(岁)				学历				访谈时长(分钟)		居住地	
	男	女	60~65	66~70	71~75	76~80	小学	初中	高中	大学	15~20	21~25	农村	城市
人数	11	12	8	9	4	2	5	9	5	4	14	9	13	10

2.2 研究方法

2.2.1 研究设计与访谈提纲拟定

目前国内关于健康信息替代搜寻影响因素的系统研究较少,需要利用质性研究方法对搜寻行为进行观察和访谈,对问题进行概念化描述,进而构建理论和模型^[18]。质性研究方法通过深入访谈被替代主体,可以更深入地了解他们的需求和行为,能够更好地发现影响因素和作用路径。通过回顾已有研究文献及结合研究者前期调查结论制定初步访谈提纲并咨询相关专家,根据专家建议对提纲进行调整和修订。提纲主要涉及研究对象的自然特征、信息素养与知识技能、行为影响因素等,以开放式问题为主,包括:(1)您使用手机或者电脑上网搜索、浏览信息频率如何?(2)您认为网络健康信息对您的健康是否有帮助?(3)能否自主分

辨网络信息真伪,是否担忧受到不良信息的困扰?(4)搜寻网络健康信息对您的康复帮助程度如何?(5)如果您自主搜寻健康信息有困难,家人会为您提供帮助或替代您搜寻吗?(6)在什么情况下,大多由谁帮您搜寻以及如何处理健康信息呢?(7)您主要关注哪方面的健康信息,通过哪些网站或平台获取?

2.2.2 资料收集与分析方法

采用半结构式访谈法收集资料并全程录音转化文本,平均时长20分钟。应用NVivo12软件,依据扎根理论,对访谈文本进行三级编码并进行理论饱和度检验,总结归纳老年慢性病患者视角下健康信息替代搜寻行为的主要影响因素,采用构建解释结构模型(ISM)的方法,分析各影响因素间的层级关系及其关联路径,并构建结构模型。

2.3 质量控制及伦理

访谈前向受访者介绍本研究的目的、意义，访谈过程中避免诱导性的提问，研究者始终保持中立态度。本研究已获医院伦理委员会批准（2022047），使用 Excel 2019 录入数据，应用 SPSSAU 21.0 在线分析软件计算理解矩阵、可达矩阵、层次分解等。

3 结果

3.1 开放编码

开放式编码是将研究资料逐步概念化和范畴化的过程^[19]。为提高编码的信度，由两名研

究者分别使用 NVivo12 软件对原始资料逐字逐句分析、编码并赋予其概念化标签，再相互比对、分析修正，共获 81 个初始概念，6 个范畴，如表 2、表 3 所示。

3.2 主轴编码、选择性编码

主轴编码是指将开放编码依据潜在的相互关系进一步统筹归类，并按照逻辑次序抽出主轴编码^[20-21]。在本研究中，开放编码（A₁ 健康感知、A₂ 健康关注度、A₃ 信息搜寻能力、A₄ 信息价值判断、A₅ 自我效能、A₆ 亲情关怀）可合并归类为 4 个主范畴，分别为健康状态（A₁、A₂），电子健康素养（A₃、A₄），自我效能（A₅），情感需求（A₆），如表 4 所示。

表 2 概念提取实例

部分初始概念	部分原始资料语句
a10 多种慢性疾病缠身	我身患好几种慢性疾病，比如高血压让我经常头晕，糖尿病又让我得控制饮食，这些病一直困扰着我，让生活变得很不方便。
a15 行动困难、就医不便	我的腿脚不太灵便，每次去医院都要费很大劲，上下楼梯特别吃力，而且医院人多，排队等候也让我感到很疲惫。
a28 通过社交媒体交流健康信息	我加入了几个健康养生的微信群，经常和群里的病友们交流治疗经验、分享养生食谱，还互相鼓励打气，感觉很有收获。
a38 自己可以搜寻但看不懂网络健康信息	我虽然能自己在网上搜到很多健康信息，但那些专业的医学术语和复杂的数据报告让我看不懂，还是得请教医生或懂行的朋友。
a51 网络预约问诊很方便、作用大	我发现现在网络预约问诊真的太方便了，只需要在手机上点几下就能预约到专家号，不用再去医院排队挂号了，省了很多时间和精力。
a52 大网站健康信息可靠	我一般都去那些大型知名的健康网站查询信息，比如丁香园、好大夫在线等，我觉得这些网站的信息比较可靠，有专业医生审核把关。
a61 老年人查信息麻烦	我这老年人查信息可真麻烦，眼睛花了看不清屏幕上的字，手指也不灵活点不动那些小小的按钮，每次都得麻烦年轻人帮忙。
a71 依赖家人买药和健康产品	我现在买药和健康产品都得靠家人帮忙，我自己不懂怎么挑选，也怕买到假货，所以都是让儿女或者老伴去买。
a72 告知子女健康产品使用感受	我每次用了新的健康产品都会告诉子女我的感受，比如这个保健品吃了有没有效果，那个医疗器械用起来方不方便，让他们也了解了解。
a73 儿女距离远见面少	我的儿女都在外地工作，离我很远，我们一年到头也见不了几次面，只能通过电话或者视频聊聊天，挺想念他们的。

表3 开放编码

范畴化	初始概念	范畴内涵
A ₁ 健康感知	a1 腰背酸痛 / a2 高血压 / a3 哮喘 / a4 肾炎 / a5 风湿病 / a6 前列腺炎 / a7 冠心病 / a8 胃溃疡 / a9 糖尿病 / a14 脑卒中 / a10 多种慢性疾病缠身 / a11 经常感冒 / a12 术后恢复不理想 / a13 频繁就医 / a15 行动困难、就医不便	老年人对身体状态的评估，包括生理疾病、身体疼痛、就医体验等方面。
A ₂ 健康关注度	a16 不经常锻炼 / a17 查询慢性疾病新药 / a18 要经常体检 / a19 饮食疗法 / a20 必须坚持锻炼 / a21 关注饮食疗法 / a22 不经常关注健康信息 / a23 和朋友、病友经常聊药物或耗材信息 / a24 偶尔锻炼但不经常 / a25 饮食疗法有效 / a26 定期积极体检 / a27 需频繁就医 / a28 通过社交媒体交流健康信息 / a29 购买上门推销的药品 / a30 向朋友询问药物疗效 / a31 很关注自身健康	老年人对健康的重视程度和采取的健康措施，包括生活习惯、健康信息获取、治疗选择等方面。
A ₃ 信息搜寻能力	a32 不会拼音 / a33 不知道搜索 / a34 打字很困难 / a35 获取所需信息困难 / a36 自己不行依靠女儿搜寻 / a37 自我感觉可以自己搜寻 / a38 自己可以搜寻但看不懂网络健康信息 / a39 信息检索能力差 / a40 基本不会上网 / a41 家里没网络，偶尔听广播看电视 / a42 没使用过网络 / a43 不会用智能手机 / a44 只会聊微信 / a45 不会用电脑上网 / a46 没接触网络	老年人获取健康信息的技能和能力，包括网络使用、信息检索等方面。
A ₄ 信息价值判断	a47 大量药品广告来袭感觉烦 / a48 网络信息质量真假难辨 / a49 烦电话药品推销 / a50 女儿或朋友提供的信息比较有效 / a51 网络预约问诊很方便、作用大 / a52 大网站健康信息可靠 / a53 信任医院官方信息 / a54 海量信息感觉疲惫 / a55 信任老中医 / a56 辩证评价网络健康信息价值 / a57 网络信息都是骗人的 / a58 网络购药对慢性病有帮助	老年人对线上健康信息的真实性、可靠性和有用性的判断能力。
A ₅ 自我效能	a59 感觉无助依赖子女 / a60 视疲劳、心烦 / a61 老年人查信息麻烦 / a62 年龄大很难学会上网 / a63 有尝试过但效果不理想 / a64 主动获取信息态度消极	老年人在面对健康信息搜寻时的自信心和积极态度，以及克服困难的决心和能力。
A ₆ 亲情关怀	a65 孩子们工作很忙 / a66 不希望他们担忧 / a67 家人主动关怀 / a68 和子女视频聊天 / a69 家庭和睦 / a70 充分信任亲人 / a71 依赖家人买药和健康产品 / a72 告知子女健康产品使用感受 / a73 儿女距离远见面少 / a74 儿女不在身边 / a75 与子女同住 / a76 家庭内部分享网络健康信息 / a77 很少谈论健康话题 / a78 主动告知家人健康状况 / a79 家人出于关心分享网络健康信息 / a81 听从家人或朋友的治疗建议	老年人在健康方面感受到的家人关怀和支持，以及他们对这种关怀的主观感受和满意度。

表4 主轴编码

核心范畴	主范畴	子范畴	逻辑关系及内涵	原始语句示例
老年慢性病患者健康信息替代搜寻行为的影响因素	健康状态	A ₁ 健康感知	健康状态影响老年人的健康信息需求	“我身患好几种慢性疾病，需要了解很多健康信息。”
		A ₂ 健康关注度	对健康的关注程度影响健康信息的搜寻动力	“我很关注自己的健康，所以经常上网查找相关信息。”
	电子健康素养	A ₃ 信息搜寻能力	信息搜寻能力决定老年人能否有效获取健康信息	“我不会上网，所以很难获取到有用的健康信息。”
		A ₄ 信息价值判断	信息价值判断能力影响老年人对健康信息的信任度和利用率	“我觉得大型知名网站的信息比较可靠，所以会优先查看。”
	自我效能	A ₅ 自我效能	自我效能影响老年人健康信息搜寻的自信心和积极性	“我觉得自己年纪大了，学不会上网，所以不愿意尝试。”
	情感需求	A ₆ 亲情关怀	亲情关怀影响老年人对健康信息的接受度和分享意愿	“我会和家人分享我找到的健康信息，让他们也了解。”

选择性编码也称核心编码，是指从本研究4个主范畴中提取核心范畴，建立核心范畴和其他范畴之间的联系，如图1所示，其“故事线”为：老年慢性病患者健康状态影响其情感需求，老年人电子健康素养影响自我效能，情感需求和自我效能共同引发替代搜寻行为。也就是说，健康信息替代搜寻行为是由自我效能和情感需求直接引发，而受到老年人健康状态和电子健康素养的影响。为检验理论饱和度，将访谈文本按3:1比例随机分两份，分别进行扎根理论编码。四分之三文本与四分之一文本的分析结论基本一致，且后者未提炼出新范畴，表明编码已达到理论饱和度。

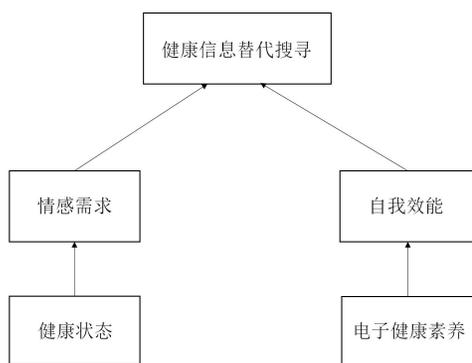


图1 健康信息替代搜寻主范畴的结构关系

3.3 解释结构模型构建

ISM适用于元素众多、结构关系复杂的问题，其通过将复杂系统进行分解，厘清各元素之间的关系，划分各元素之间的层级关系，将混乱无序的结构转化为具有明显结构层级关系的模型^[22]。本研究通过扎根理论分析总结，归纳影响健康信息替代搜寻行为主要有4个主范畴和6个副范畴。为了进一步厘清因素之间的层次关系及关联路径，研究者采用ISM方法分

析并构建结构模型。

3.3.1 构建邻接矩阵

通过德尔菲打分的方式建立邻接矩阵，课题组成员充分讨论并打分，相关结果再通过电子邮件函询相关专家，最终确定6个主要影响因素（ A_1 健康感知、 A_2 健康关注度、 A_3 信息搜寻能力、 A_4 信息价值评价、 A_5 自我效能、 A_6 亲情关怀）之间的相互关系。以 $A_1 \cdots A_6$ 代指各个因素构建 6×6 邻接矩阵A，如表5所示，矩阵中数字“1”表示因素之间相互影响，数字“0”表示因素之间无相互影响。

表5 邻接矩阵A

因素	A_1	A_2	A_3	A_4	A_5	A_6
A_1	0	1	0	0	0	1
A_2	1	0	0	1	0	0
A_3	0	0	0	1	1	0
A_4	0	0	0	0	1	0
A_5	0	0	0	0	0	0
A_6	0	0	0	0	0	0

3.3.2 计算可达矩阵及层级关系

把邻接矩阵A输入SPSSAU 21.0在线分析软件求解可达矩阵M以及可达矩阵M的可达集合R、先行集合Q及其交集，如表6、表7所示，依据结果优先的层次划分原则得三个层级： $L_1 = (A_5, A_6)$ ， $L_2 = (A_4)$ ， $L_3 = (A_1, A_2, A_3)$ 。

3.3.3 解释结构模型的构建

根据表7的层级划分结果，构建健康信息替代搜寻影响因素层级结构模型，如图2所示。该模型共包括3个层级，分别是表象层（直接因素）、中间层（关键因素）、根源层（根源因素）。

表 6 可达矩阵 M

因素	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆
A ₁	1	1	0	1	1	1
A ₂	1	1	0	1	1	1
A ₃	0	0	1	1	1	0
A ₄	0	0	0	1	1	0
A ₅	0	0	0	0	1	0
A ₆	0	0	0	0	0	1

表 7 可达矩阵 M 的可达集、先行集和交集

	可达集合 R	先行集合 Q	交集 R∩Q	层级
A ₁	A ₁ A ₂ A ₄ A ₅ A ₆	A ₁ A ₂	A ₁ A ₂	L ₃
A ₂	A ₁ A ₂ A ₄ A ₅ A ₆	A ₁ A ₂	A ₁ A ₂	L ₃
A ₃	A ₃ A ₄ A ₅	A ₃	A ₃	L ₃
A ₄	A ₄ A ₅	A ₁ A ₂ A ₃ A ₄	A ₄	L ₂
A ₅	A ₅	A ₁ A ₂ A ₃ A ₄ A ₅	A ₅	L ₁
A ₆	A ₆	A ₁ A ₂ A ₆	A ₆	L ₁

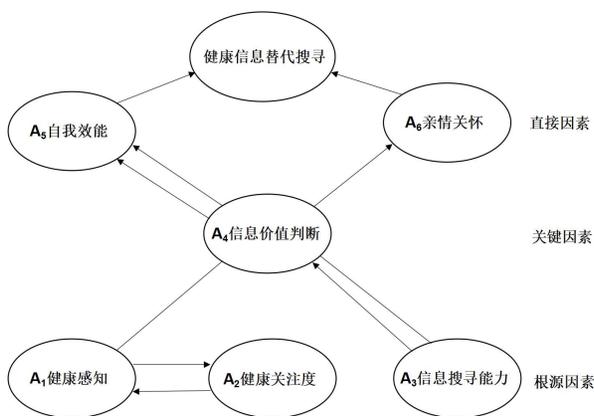


图 2 健康信息替代搜寻影响因素的层级结构模型

4 模型阐释与讨论

影响老年慢性病患者健康信息替代搜寻行为的因素包括：直接因素（A₅、A₆）、关键因素（A₄）、根源因素（A₁、A₂、A₃），如图 2 所示，替代搜寻行为源于老年人对健康状况的感知和关注，两者受到老年人健康信息价值判

断能力的影响，并由自我效能和情感需求直接引发。以下分别对各个因素进行分析。

4.1 健康信息替代搜寻行为是由老年人的自我效能和情感需求直接引发

访谈中发现，老年慢性病患者普遍知识技能有限、电子健康素养较低，对自主上网搜寻健康信息信心不足，在线健康信息的实用性、易用性和有效性又使得大多数老年人对其持积极态度，而由亲属或朋友替代搜寻成为最佳选择，这符合心理学上的最省力原则。直接引发替代搜寻的另一因素是亲情关怀，老年人的家庭亲疏关系成为引发替代搜寻行为的关键。在访谈中以及研究者的调查发现，老年人向家人倾诉自己的健康状况，希望得到家人（子女）的重视，通过亲情互动获得家人的关爱与支持，也有老年人收到家人代其购买的健康产品积极交流使用感受，以此增进亲情互动。老年人退休在家无人陪伴，随着年龄的增长，其活动与社交范围愈发局限，容易感到空虚与寂寞，又受到慢性病困扰难免产生孤独无助的焦虑情绪，健康信息替代搜寻在很大程度上既满足了老年人的健康需求，又增强了亲情互动。

4.2 健康信息价值判断是影响健康信息替代搜寻行为的关键因素

老年人对健康信息价值的判断既包括对线上健康信息价值的判断（如新药疗效评估），又包括甄别纷繁复杂的线上信息，排除不良信息的困扰。有替代搜寻经历的老年人，一般不会评价线上信息质量与效用，对自主搜寻缺乏信心，需要依靠家人（子女）替代。访谈中

解到,有替代搜寻经历的家庭大多亲情关系比较和睦,老年人对子女信赖度和满意度较高,而父辈信任是推动替代搜寻行为的重要因素,这与研究者此前的结论一致。也就是说,老年人对子女的信赖度和满意度高,则希望子女替代自己搜寻健康信息;而子女获得的父辈信任越多,越倾向于进行线上健康搜寻并在信息解读上会付出更多的努力,以求简单准确地传递给老年人。因此,信息价值判断与评价对老年人自我效能有直接影响,同时也会对亲情关怀产生影响。

4.3 老年人对健康状况的感知和关注是健康信息替代搜寻行为的根源

本次访谈对象都是患有有一种以上慢性疾病的老年人,访谈中他们表现出对自身的健康状况比较清晰的感知,他们对自身状态的变化都有明显的感知,对线上健康信息比较重视,在日常生活中渴望获得来自亲人的关怀。反之,健康状况较好的老年人对线上健康信息一般不关注,更不了解其价值,也不太需要亲属、子女的关心和问候。老年人对健康的关注度与健康感知相互影响,有替代搜寻经历的受访老年慢性病患者对自身健康需求(如康复器械、药品、耗材等)更为了解,同时了解自身健康需求的老年人对使用康复器械或服用药物带来的健康状态变化感知也越明显。此外,在人们的印象中老年人知识能力有限,大多不会使用互联网检索信息,更无法准确评估线上健康信息价值,对自主搜寻缺乏信心。而访谈中有位老年糖尿病患者多年来都是自主上网搜寻和购买糖尿病药物和耗材,随着视力下降才改由子女替代搜

寻。上述研究表明,并不是所有的老年人都不会使用互联网,部分老年人具有信息检索能力,是由于一些原因受到限制,这与已有研究结论^[5]一致。由此可见,老年人对自身健康状况的感知和关注是健康信息替代搜寻行为影响层次最深的因素,但绝大多数的老年人信息检索能力确实不足,把信息检索能力作为根源因素之一同样具有合理性。

5 结论

本文不仅通过深度访谈和质性分析方法厘清了老年慢性病患者在线健康信息替代搜寻行为的影响因素及其层次关系,为信息行为理论研究提供了坚实的基础,也为老年人在线健康服务的优化提供了重要的启示。首先,信息服务机构应加强对老年人信息素养的培训和教育,提升其信息搜索能力和信息价值判断能力,从而帮助他们更有效地获取和利用网络健康信息。其次,针对老年人的特殊需求,优化网络健康信息平台的设计和功能,使其更加易于老年人使用和理解,提高信息的可用性和易用性。再次,鼓励家庭成员积极参与老年人的健康信息管理,提供必要的支持和帮助,满足老年人的情感需求,同时促进家庭和谐与亲密关系。最后,根据老年人的健康状况、信息素养和自我认知等因素,提供个性化的健康信息服务,帮助他们更精准地获取所需的健康信息。

综上所述,本文不仅丰富了信息行为理论的研究内涵,也为老年人在线健康服务的实践提供了有益的参考和启示。研究局限性:一是访谈对象为老年慢性病患者,研究对象较为单

一；二是对访谈文本进行编码和分析可能存在主观性。未来研究可扩大研究对象覆盖范围，纳入不同健康状态（如急症、亚健康）及城乡差异显著的老年群体，以检验结论的普适性与差异性。此外，可结合技术接受模型（TAM）与社会支持理论，探究代际数字鸿沟、文化观念对替代搜寻行为的作用机制，评估其对老年人信息素养的干预效能。

参考文献

- [1] 闫伟, 何梦娇, 路云, 等. 基于 CLHLS 的我国老年人失能现状及其影响因素研究 [J]. 护理研究, 2021, 35(10): 1807-1811.
- [2] 章轶立, 黄馨懿, 齐保玉, 等. 老年人群共病问题现状挑战与应对策略 [J]. 中国全科医学, 2022, 25(35): 4363-4368.
- [3] 宋小康, 赵宇翔, 宋士杰, 等. 互联网环境下我国健康信息替代搜寻者特征及其行为模式研究 [J]. 图书情报工作, 2020, 64(22): 107-117.
- [4] 宋小康, 赵宇翔, 朱庆华. 在线健康信息替代搜寻对被替代者健康行为和健康水平的影响研究 [J]. 情报学报, 2022, 41(6): 625-636.
- [5] 张肖, 王文韬, 李晶, 等. 老年人在线健康信息替代搜寻: 内容分析与研究展望 [J]. 情报资料工作, 2021, 42(5): 84-93.
- [6] 王超, 卢智增. 农村老年人健康信息替代搜寻者特征及其行为影响因素探析 [J]. 图书情报导刊, 2022, 7(10): 37-44.
- [7] MCKENZIE P J. A model of information practices in accounts of everyday-life information seeking[J]. The Journal of Documentation, 2003, 59(1): 19-40.
- [8] CUTRONA S L, MAZOR K M, VIEUX S N, et al. Health Information-Seeking on Behalf of Others: Characteristics of "Surrogate Seekers"[J]. Journal of Cancer Education, 2015, 30(1): 12-19.
- [9] 谢兴政, 张大伟, 张潜, 等. 农村大学生线上健康信息替代搜寻意向形成机制研究 [J]. 图书馆学研究, 2021(5): 69-81.
- [10] SADASIVAM R S, KINNEY R L, LEMON S C, et al. Internet health information seeking is a team sport: Analysis of the Pew Internet Survey[J]. International Journal of Medical Informatics, 2013, 82(3): 193-200.
- [11] 万文智, 宋小康, 赵宇翔, 等. 在线健康信息替代搜寻行为的影响因素探究: 基于扎根理论的实证 [J]. 情报资料工程, 2020, 41(6): 88-94.
- [12] CUTRONA S L, MAZOR K M, AGUNWAMBA A A, et al. Health information brokers in the general population: an analysis of the health information national trends survey 2013-2014[J]. Journal of medical Internet research, 2016, 18(6): e123.
- [13] 张儒雅. 农村老年人在线健康信息替代搜寻行为及影响研究 [D]. 保定: 河北大学, 2024.
- [14] REIFEGERSTE D, BACHL M, BAUMANN E. Surrogate health information seeking in Europe: Influence of source type and social network variables[J]. International Journal of Medical Informatics, 2017(103): 7-14.
- [15] 陈明红, 李思婷, 曾贝欣. 老年人网络健康信息替代搜寻行为影响因素及关联路径研究——从被替代搜寻者视角 [J]. 信息资源管理学报, 2022, 12(5): 41-52.
- [16] 新华社. 中共中央、国务院印发《“健康中国2030”规划纲要》[EB/OL]. (2016-10-25) [2022-06-30]. http://www.gov.cn/zhengce/2016-10/25/content_5124174.htm.
- [17] CRESWEII J W. Qualitative Inquiry and Research Design: Choosing Among Five Traditions[M]. Thousand Oaks: Sage, 1998.
- [18] 孙晓娥. 深度访谈研究方法的实证论析 [J]. 西安交通大学学报(社会科学版), 2012, 32(3): 101-106.
- [19] 张敬伟, 马东俊. 扎根理论研究法与管理学研究 [J]. 现代管理科学, 2009(2): 115-117.
- [20] 李娟, 孙启, 傅利平. 基于扎根理论的唐山市医养结合机构养老服务影响因素质性研究 [J]. 医学与社会, 2022, 35(9): 37-42.
- [21] 何秋, 何瑛, 李贵妃, 等. 社区老年人对陪护机器人使用意愿的质性研究 [J]. 护理学报, 2022, 29(19): 6-10.
- [22] 李立清, 丁海峰. 我国城市老年人慢性病防控的多级阶梯结构及逻辑阐释 [J]. 医学与社会, 2022, 35(6): 29-34.



开放科学
(资源服务)
标识码
(OSID)

“健康中国”研究现状：格局、轨迹与热点

刘樱 王语嫣

武汉大学人民医院胃肠外1科 武汉 430060

摘要: [目的/意义] “健康中国”战略关系到国家运行和民生福祉, 研究“健康中国”战略, 探讨哲学社会科学研究生态体系对其形成与发展的影响, 并提出建议与展望。[方法/过程] 采用描述性统计、相关性分析等方法与 Echarts 等工具, 从基金立项的宏观层面与成果产出的微观层面对“健康中国”的研究格局、轨迹与热点进行探究。[结果/结论] “健康中国”的哲学社会科学研究体系正逐步形成, 但研究主体分布态势的均衡性缺位显著, 细分学科力量布局差异性日益明显, 跨学科的交汇融合尚不充分。今后应统筹各学科的研究力量, 加强学科间的协作创新和立体化运作, 以促进“健康中国”战略实现。

关键词: 健康中国; 研究热点; 发展趋势

中图分类号: G35; R197.1; G252

Current Status of Research on “Healthy China”: Patterns, Trajectories and Hotspots

LIU Ying WANG Yuyang

No.1 Department of Gastrointestinal Surgery, Wuhan University People's Hospital, Wuhan 430060, China

Abstract: [Objective/Significance] The “Healthy China” strategy is related to the operation of the country and the well-being of the people. This article aims to research the “Healthy China” strategy, explore the impact of the philosophical and social science research ecosystem on its formation and development, and propose suggestions and prospects. [Methods/Processes] Using methods such as descriptive statistics and correlation analysis, along with tools like Echarts, this research examines the research pattern, trajectory, and hotspots of the “Healthy China” initiative from both the macro level of funding approvals and the micro level of research output. [Results/Conclusions] The philosophy and social sciences research system for the “Healthy China” is gradually taking shape. However, the balance in the distribution of research entities is lacking, the differences in the strength layout of sub-disciplines are increasingly, and interdisciplinary integration is insufficient. In the future, it is essential to coordinate the forces of various disciplines, enhance collaborative innovation and multidimensional operations across disciplines, to promote the realization of the “Healthy China” strategy.

Keywords: Healthy China; Research Hotspots; Development Trends

作者简介 刘樱 (1980-), 学士, 主管护师, 主要研究方向为临床医学研究; 王语嫣 (1994-), 学士, 主管护师, 主要研究方向为医学信息学研究, E-mail: 526212876@qq.com。

引用格式 刘樱, 王语嫣. “健康中国”研究现状: 格局、轨迹与热点 [J]. 情报工程, 2025, 11(1): 95-108.

引言

2016年10月,中共中央、国务院印发实施《“健康中国2030”规划纲要》,将人民健康提上了最高议程^[1];2019年7月,国务院印发《国务院关于实施健康中国行动的意见》,成立健康中国行动推进委员会;2020年,面对公共卫生事件,党始终将“人民生命健康重于泰山”作为行动指南^[2]。“健康中国”作为民族昌盛和国家富强的重要单元,需要自然科学“硬实力”给予坚毅支持。各学科学者已从多视角开展研究,涌现大量成果,但目前尚缺乏针对该领域周期性、阶段性的梳理与总结,难以清晰呈现出“健康中国”前沿热点。

综上,本文从国家社科基金项目 and CSSCI 来源期刊两个维度出发,综合运用网络调研、内容分析、统计分析等多种研究方法,对2017—2021年内的研究内容进行解析和梳理,并借助 Echarts、微思词云、CiteSpace 等多种可视化工具予以展示,明晰“健康中国”在哲学社会科学领域的研究格局、轨迹与热点,为各个学科对该领域的未来研究提供新视角,从而更好地服务于个体健康保障、社会“大健康”布局以及国家“健康中国2030”建设的整体进程。

1 数据来源与研究思路

1.1 数据来源

从全国哲学社会科学工作办公室官网获取2017—2021年的基金立项清单^[3],保留名称涵盖“老龄化”“大健康”等的项目,获取297项项目;同时,在中国知网上,构建检索式

为:主题=“健康中国”OR主题=“健康中国2030”OR主题=“健康中国战略”,以精确匹配检索2017—2021年,来自CSSCI期刊的文献,共检出983篇;删除简讯、会议纪要文献后获取601篇文献,其中117篇为上述社科项目研究成果,保留题名、摘要、关键词、作者、机构、来源刊物及发表时间在内的数据项,并以“Refworks”格式导出用于进一步分析。

1.2 研究思路

针对社科基金项目,利用EXCEL进行项目年度、学科等分布态势统计,借助Echarts进行可视化;以聚类分析视角,结合NLPIR等进行主题统计。针对CSSCI,进行年度、刊物、学科分布态势在内的描述性统计,利用SPSS22.0软件进行文献下载和被引情况分析;利用CiteSpace开展研究核心力量和热点研究方向及演进轨迹的可视化分析。研究架构如图1所示。

2 “健康中国”研究现状探索

2.1 基于社科基金项目的探索

2.1.1 描述性统计

(1) 立项年度分布

立项数量可以反映国家对特定主题的重视程度,统计结果如表1所示,发现5年间立项数依次为49、63、60、67、58项。总体而言,除2017年,其余4年立项数均超过55项,呈现出稳定发展态势。说明国家给予“健康中国”领域极高的重视,专家学者积极投身于该领域的探索,展现出较好的研究前景。进一步探析

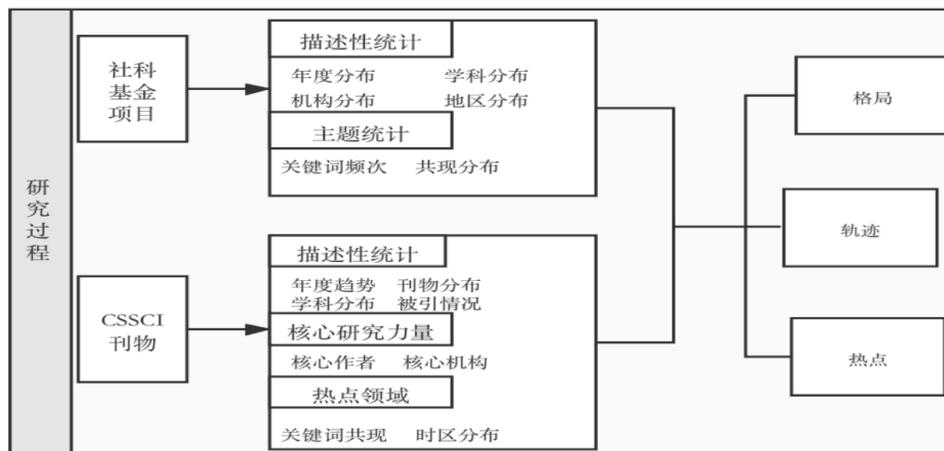


图1 研究过程整体框架图

表1 2017—2021年“健康中国”领域社科基金类别年度分布

时间	重大项目	重点项目	一般项目	青年项目	西部项目	总数
2017	5	2	28	12	2	49
2018	5	3	38	16	1	63
2019	2	4	44	10	0	60
2020	3	6	41	15	2	67
2021	2	4	40	12	0	58

不同类别基金立项情况，发现一般项目占主体地位，青年项目次之，体现出青年学者受到大力支持。重大和重点项目集中攻关高端项目，两者具有一定立项数，体现出该领域具有较大的影响力。西部项目立项数则相对较少。

（2）立项地区分布

按区域统计立项数，能对其进行整体把握。统计除港澳台外的31个省、直辖市和自治区立项情况。发现立项数同经济状况保持较高耦合：一是立项地区覆盖面广，31个省、直辖市、自治区均有立项，体现研究广泛性；二是地域分布不均衡，呈现出“东多西少，南多北少”的格局，立项数量前3位为北京（43项）、江苏（35项）、上海（28项），共占据总立项数的35.6%，反映出在社科基

金资源获取层面显著的区域差异性；三是教育大省和沿海经济强省立项优势显著，湖北（24项）、广东（23项）、陕西（19项）立项数量突出，区位优势突出的浙江（14项）、山东（11项）、福建（10项）在立项数量上表现较好。

（3）立项主持机构分布

主持社科基金项目的数量，能够体现机构科研实力，统计结果如表2所示。其中，中山大学（10项）名列首位，占主导地位。此外，立项数量前10的高校仅占立项高校总数（107所）9.3%，但其立项数为63项，占总立项数21.2%，既体现该领域拥有一批实力雄厚的核心高校开展研究工作，又一定程度上反映了立项主持机构的不均衡性。

表 2 2017—2021 年立项数量前 10 的高校

高校名称	立项数量	占比 /%	高校性质	所属地区
中山大学	10	3.4	综合	广东
中国人民大学	8	2.6	综合	北京
北京大学	8	2.6	综合	北京
温州医科大学	6	2.1	医药	浙江
上海交通大学	6	2.1	综合	上海
东南大学	6	2.1	综合	江苏
中南大学	5	1.6	综合	湖南
中南财经政法大学	5	1.6	财经	湖北
西安交通大学	5	1.6	综合	陕西
北京师范大学	5	1.6	师范	北京
合计	63	21.2	—	—

(4) 立项学科分布

对立项学科进行分析，以明晰不同学科的参与度，结果如图 2 所示，项目涉及的学科数量多、范围广，其中管理学、社会学与人口学占比较高，体育学也表现强劲。尽管立项的学科来源广泛，但总体上呈现出前述 4 类学科主

导的局面，而其他学科则尚未充分发挥出学科优势和价值。“健康中国”应是一项涉及个体、政府、社会在内的“全民事业”，其推进、发展及实现同样离不开民族学、新闻传播学、法学等专业的支持与发力。

2.1.2 立项主题分析

项目名称一般由研究视角、研究内容等元素构成，对其进行分析，能反映研究的核心要素。利用 NLPPIR 对 297 项项目名进行分词，同时剔除“健康”“战略”等高密度低价值的词条，保留频次大于 4 的语词，同时借助微思词云软件绘制主题词云，如图 3 所示。可以发现，医疗、养老、老人等词的频次较高。语词可以划分为 5 大类主题：多元主体健康服务（老人、青少年、少数民族、养老）、政府健康治理与保障（政策、治理、协调）、“大健康”融合与重构（心理、体育、体医融合）、健康平等与协调（贫困、平等、平衡）、基础医疗巩固与促进（医疗、卫生、供给）。

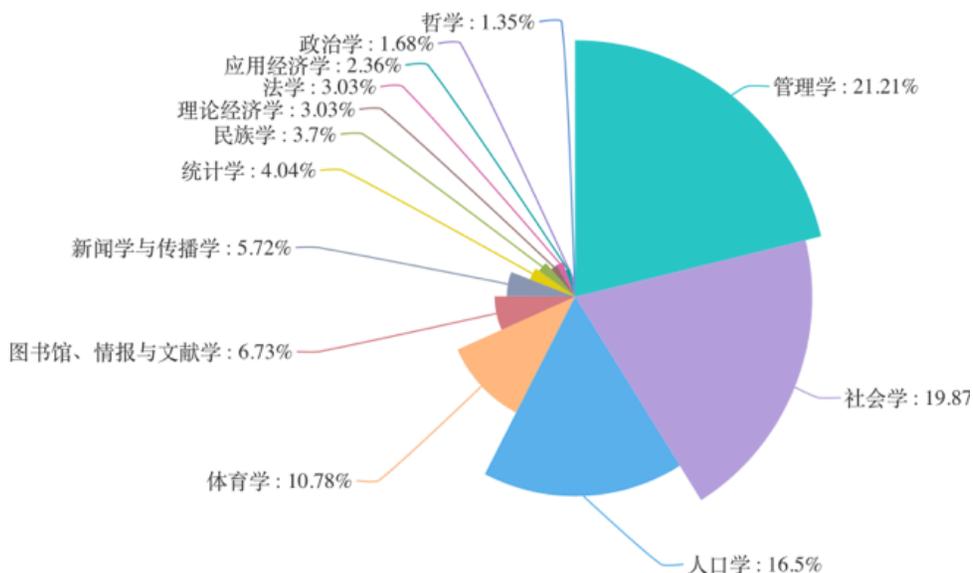


图 2 2017—2021 年立项学科分布

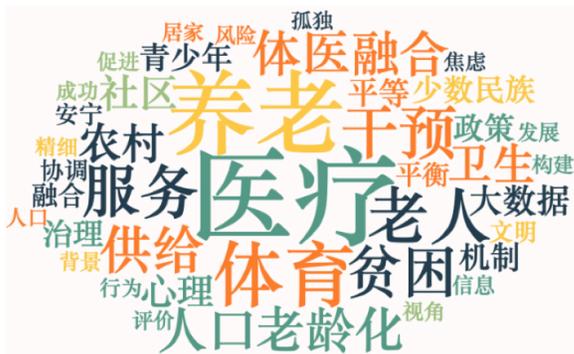


图3 2017—2021年社科基金立项主题词云

2.2 基于CSSCI来源刊物的探索

2.2.1 描述性统计

(1) 年度趋势

对年度趋势进行分析，能够对领域研究成果产出予以把握，结果如表3所示。从发文量来看，“健康中国”提出至今，CSSCI来源刊物发文量基本呈逐年攀升趋势，表明科研工作者们对其的重视程度在不断增加，始终紧跟国家宏观层面的战略引导，产出大量兼具创新性与启发性的科研成果。从社科基金支持的比率来看，2017—2021年每年的相关产出成果中由社科基金支持的文献占比逐年攀升，体现出社科基金的成果转化率呈现出稳中有进的良好

趋势。

表3 2017—2021年CSSCI来源刊物发文量及基金支持情况

年度	文献数量	社科基金支持数量	社科基金支持率/%
2017	67	9	13.43
2018	106	19	17.93
2019	141	26	18.44
2020	140	31	22.14
2021	147	34	23.13

(2) 刊物来源分布

对刊物来源进行分析利于了解“健康中国”研究的学科范畴。经统计，文献成果分布于231种刊物，发文量排名前20的刊物如表4所示。排名前4的期刊中有3个体育类期刊，说明其依托学科特色在成果传播中处于核心地位。靠前的期刊还包括人口学、管理学相关期刊，从不同维度对“健康中国”战略进行深层次的研究。同时，期刊影响因子也能够体现出成果的质量。影响因子最高的是《中国图书馆学报》(6.434)，影响因子大于2的期刊占比70%，体现出各学科均高度关注“健康中国”主题探索并持续产出高质量研究成果。

表4 2017—2021年发文量前20的期刊

期刊	发文量	影响因子	期刊	发文量	影响因子
体育文化导刊	27	1.876	中国行政管理	10	4.137
北京体育大学学报	26	2.702	华中师范大学学报(社科)	8	1.736
人口与发展	23	3.135	科技与出版	7	1.240
武汉体育学院学报	22	2.347	山东社会科学	6	1.760
西安体育学院学报	17	2.215	上海体育学院学报	5	3.372
体育科学	16	5.367	心理与行为研究	5	2.220
人民论坛	14	1.763	体育学研究	4	2.885
体育学刊	13	2.317	西北人口	3	1.990
沈阳体育学院学报	13	2.083	中国图书馆学报	3	6.434
中国人口科学	11	4.887	中国软科学	3	5.418

(3) 学科分布

从学科分布来看,如表5所示,“健康中国”相关成果除外国文学与考古学外,其余学科均有涉及,体现出成果来源广泛性;从发文量来看,尽管体育学学科在社科基金立项方面表现并不突出,但其研究成果产出却处于绝对的优势地位,其次是社会学、人口学。这表明自“健康中国”提出以来,体育学科的科研工作者,开展了大量研究,而人口学与社会学领域的专家学者们也从自身的学科视角探析“健康中国”视阈下社会发展的微观表征。

表5 CSSCI来源期刊发文学科分布

学科	发文量	学科	发文量
体育学	142	图书馆、情报学与文献学	17
社会学	79	马克思主义理论	13
人口学	59	人文经济地理	8
高校人文社科及综合类学报	60	法学	7
综合社科	52	政治学	6
经济学	37	宗教学	5
管理学	33	哲学	4
新闻学与传播学	26	民族学与文化学	3
心理学	24	历史学	3
教育学	21	其他	2

(4) 被引下载情况

引证行为是学术交流、知识共享的客观体现^[4],被引次数是衡量文献学术价值的重要指标之一。经统计,601篇文献共有429篇被同行引用,被引率达71.3%,体现出围绕“健康中国”的研究已逐渐形成良好的学术生态体系。其中,被引频次最高的文献是《全民健身国家战略:内涵与发展思路》,被引次数245次,主要论述了“全民健身”对于“健康中国”愿景实现的重大战略意义及其基本实现路径^[5]。

同时,被引超过50次的文献有21篇,主要来自体育学、人口学、社会学领域。

下载量是数字出版时代学术论文成果在数据库中被研究人员保存、采纳的直接记录,其反映了学术成果的价值性^[6]。下载量最高的文献仍为《全民健身国家战略:内涵与发展思路》,下载数为10385次,同时,下载量超过1000次的文献有183篇。探析该领域文献成果下载量和被引量的关系,调用SPSS22.0进行相关性分析,发现P值小于0.05,即被引和下载间具有显著相关线性关系成立,两者间呈现出较强的互助促进作用。

2.2.2 核心研究力量

(1) 核心作者

目前“健康中国”领域的作者分布结构尚未稳定,核心作者群尚未形成,但已有部分代表性的高产学者,发文前3位学者包括:北京大学的刘继同(7篇)和郑晓瑛(5篇),两位学者分别关注社会健康福利体系建构和多元人群的健康保障,中山大学的申曙光(5篇)则关注“健康中国”建设的框架、路径等基础理论。同时,以发文量3篇为标准进行统计,共有21位学者,其发文量占总发文量的10.48%。

(2) 核心研究机构

对发表研究成果第一署名机构进行统计,以分析核心研究机构。表6列举出了前10的机构,其中体育类机构发文量16篇,社会与人口类机构和管理类机构发文量均为22篇,可见其在我国“健康中国”研究中发挥着重要作用,而随着“全民健身”“体医融合”理念的不断推进,体育类机构在“健康中国”研究中也扮演着重要的角色。

表 6 成果产出数量前 10 的机构

研究机构	发文量	研究机构	发文量
北京大学人口研究所	10	华东师范大学体育与健康学院	5
上海体育学院体育教育训练学院	7	北京大学社会学系	5
中山大学岭南学院	7	中央财经大学政府管理学院	4
中国人民大学社会与人口学院	7	上海财经大学公共经济与管理学院	4
中南财经政法大学公共管理学院	7	南京体育学院	4

为更加直观地展现发文机构间的合作网络情

况，利用 CiteSpace 进行可视化分析，如图 4 所示。共现知识图谱选取了文献发表数量为 3 篇及以上的机构，节点数为 159，连接数为 61 条，网络密度等于 0.0049。整体来看，发文机构众多，但机构间的合作却较为松散，仅有 3 个较大的共现网络，分别以北京大学人口研究所、中山大学岭南学院、中国人民大学社会与人口学院为中心。总的来说，“健康中国”发文机构具有数量多、来源广泛、学科丰富的特点，但机构间合作不够紧密，未形成研究的核心机构集群。

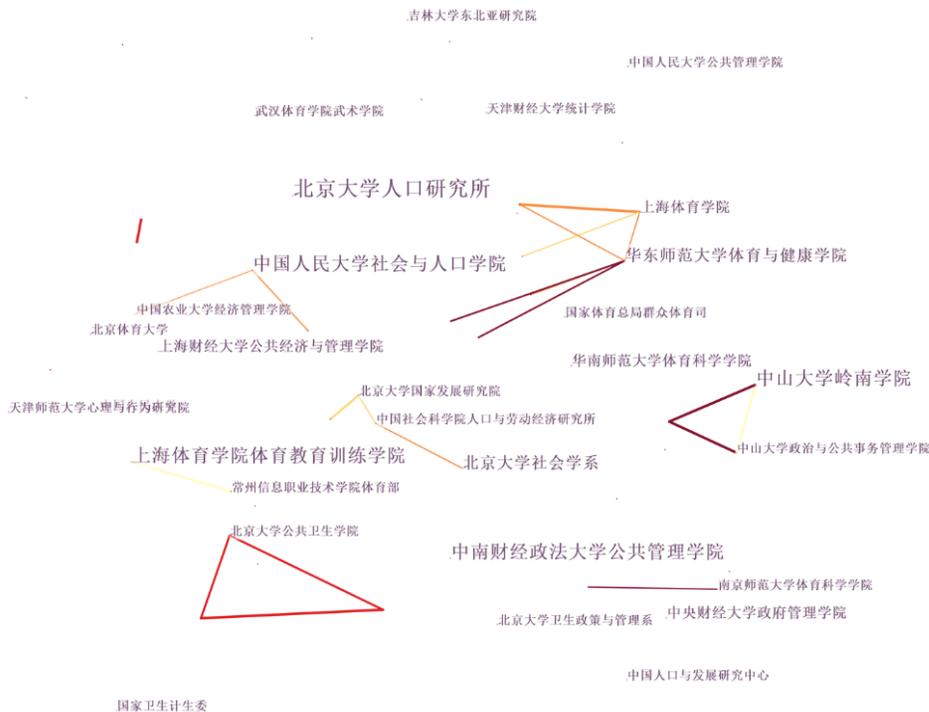


图 4 2017—2021 年“健康中国”机构共现图谱

2.2.3 研究热点与演进轨迹

(1) 关键词共现分析

借助 CiteSpace，将 Node Types 设置为 Key-word，以生成健康中国研究领域的关键词共现知识图谱，如图 5 所示。图谱中共有 248

个节点，486 条连接线，网络密度为 0.0159，反映出健康中国研究领域已初具规模。高频次关键词、高中介性关键词能对领域热点研究问题予以呈现，对频次最高和中心度最高的 10 个关键词进行统计，如表 7 所示。关键词“健康

要组成要素。针对多地区、多人群开展的心理状况、抑郁状态测评,以及援助路径探讨,是该领域研究人员关注的重点^[14],其在提升应对心理疾病素养和社会整体心理服务体系构建等重点突破口均在持续性的探索。

主题四:人口老龄化背景下的健康服务与治理。基于即将步入中度老龄化时代的基本国情,“健康老龄化”是当前健康社会建设中必须面对的难题^[15]。目前研究主要从老龄化健康服务的内涵、主要障碍、重难点等理论层面展开,同时也从各维度出发构建社会保障体系和多方联动的动力机制^[16],这些将是“十四五”时期走出“未富先老”困境的重要探索领域。

主题五:“体医融合”推进的人群健康保障。在国家、区域政策部署下,对涉及体育系统和医疗系统的有形资源和无形资源的协同和重构^[17],进行同国家现实经济社会发展状态的“体医融合”模式探索^[18],是领域研究的一个重要方向。目前的研究主要关注两者在宏观层面的系统机理探索和顶层战略性的政策工具的分析。

主题六:医养结合的养老模式构建。针对老年人规模大、健康问题突出等问题,十九大报告中明确强调“养老多元化、多样化发展,加速发展医养结合养老产业”。专家学者针对这一主题,从“全周期”和“全人群”等角度出发,探索出“养中设医”“医中设养”“医养合作”等模式^[19],同时在行政政策、社会经济模式、相关产业等方面进行探索^[20]。随着社会实践层面新障碍的出现,该研究主题将需要长期关注,开展广泛探析。

主题七:基本医疗卫生制度建设。随着《“健

康中国2030”规划纲要》的深入实施,健康中国建设过程中关于基本医疗卫生制度建设的政府责任和市场机制协同^[21],以及不同层级的公共资源的分配成为了研究重点,一方面要实现医疗卫生政策同社会发展相适应,另一方面又必须兼顾区域发展差异以及考量个体在健康资源获取上的相对平等^[22]。

主题八:全球健康治理中的“中国方案”。2017年1月,习近平总书记首次提出“健康丝绸之路”的战略构想^[23];2020年3月,提出构建“人类卫生健康共同体”。如何将健康中国建设的经验、成果予以提炼,并向全球健康治理发出中国声音,贡献中国智慧^[24],是专家学者一直探索的重要内容,其研究也应是今后着重关注方向。

主题九:国民健康自评现状与健康公平。关注国民健康自评状况和保障相对健康平等,是健康中国战略发展的内在要求。该领域研究多基于中国家庭追踪调查数据、全国老年人口健康状况调查截面数据等^[25],利用定量研究方法展开研究,相关成果能够从各维度掌握多元人群的健康状况^[26],不仅能够帮助推进健康平等化进程,同时能够较好服务于政府职能部门在社会健康建设进程中的政策制定和产业部署。

主题十:COVID-19背景下的公共卫生治理与反思。面对突发的公共卫生事件,在我国取得抗疫阶段性成果和全球抗疫态势仍不明朗的现实环境中,各学科专家学者充分利用学科优势,始终关注个体、社会和国家主体的现实需求,从应急管理、全球价值链、舆论引导等多个方面展开探索^[27],以期今后的公共卫生

时则是从微观层面分析了“健康中国”研究成果的产生及其使用情况,找出了主要研究机构及学者并挖掘其相互关联,同时从关键词、热点主题等方面分析了“健康中国”的研究重点。总体而言,基于社科基金数据可以对“健康中国”的研究现状进行整体把控,而基于 CSSCI 期刊数据则可以更好地分析“健康中国”研究的关键细节。

3 研究结论及展望

3.1 研究结论

(1) 面向“健康中国”的哲学社会科学研究生态体系正逐步形成,并呈现出良好的研究前景和发展空间。在宏观基金立项层面,该领域年度立项数稳中有进,立项来源的地区分布、机构分布、学科分布日益普遍化、成熟化,立项主题内涵始终同现实社会的发展需求保持较高的契合度;在微观研究成果层面,该领域研究的年度产出逐年攀升,成果的学科来源多元且质量较高,同行引证及知识分享的氛围活跃,并有一批高产作者、核心研究机构推动领域研究的外延扩展。

(2) 研究主体分布态势的均衡性缺位显著,细分学科力量布局差异性日益明显。当前“健康中国”研究领域无论是在国家立项支持的空间分布、院校实力格局方面,还是在成果产出的学科整体维度和高产作者层面,均有极大的差异性。部分学科、部分核心机构及其研究人员拥有明显的社会资源和成果积累优势,虽能够在一定程度上对领域研究的力量予以支撑,

但长此以往逐渐形成的“马太效应”可能导致非主导研究力量内生动力的不足,进而在边缘化状态长期过溢的情况下影响“健康中国”研究生态体系发展的可持续性。

(3) 研究视角多元且热点主题丰富,但跨学科的交汇融合尚不充分且研究的深度和广度有待提升。通过梳理可以发现,目前研究在立项层面聚焦国运民生的现实需求,紧握社会矛盾转化的发展方向,同时其研究成果中的十大热点主题,也充分展现出中国特色哲学社会科学领域研究人员在“健康中国 2030”进程中的身体力行和使命担当。但目前的研究多从各个学科自身优势出发以为社会健康服务提供支持,而跨学科领域的研究较少,还未充分发挥出学术共同体的合力,同时相对于其他已有数十年研究的主题而言,该领域刚完成“一五”周期的路途,其研究在深度和广度上均有继续拓展的空间。

3.2 研究展望

(1) 在保持现有研究格局的基础上,进一步明晰已有研究成果演进逻辑的倾向性,以拓展领域研究的外延和实现研究的可持续性发展。一方面在立项上继续立足社会实践的具体要求,向更精细化的维度予以立项支持,并在重视意识形态引领和理论体系构建的同时,能够向技术维度的研究予以倾斜;另一方面促进高产作者、核心机构在保持已有优势的基础上,开展更加多元化的探索。

(2) 统筹各学科的研究力量,避免超强化和边缘化的极端状况,加强学科间的协作创新

和立体化运作，形成领域研究的学术共同体。在继续鼓励管理学、社会学、体育学等主力学科开展研究的同时，其他学科的优势也不容忽视，注重立项层面的各学科的分配均衡性以及成果实践转化上的科学规律。

(3) 研究主题进一步深化，不仅关注宏观层面的“大健康”议题，同时也需紧跟国民“小健康”的现实需求，并重视以下内容：①信息技术助力研究内容转化，充分利用5G、大数据、云计算等新兴技术开展领域研究，完成宽泛领域向精准研究转化；②汲取传统中医药传承与创新的文化自信，现有研究鲜有针对中医药传统服务于社会健康进程的探索，今后研究应向这一方向拓展；③重视“以人为本”的群众健康教育与全民健康素养提升，健康建设归根结底是面向人的服务，今后的研究应逐渐过渡到健康自评基础上公民个体健康能力提升的维度上；④保持大健康研究的拓展，突出面向大健康产业的多元融合和人才培养的深入研究；⑤继续探索人类命运共同体视野下的中国使命与担当。

4 结语

“健康中国”是关系国运民生的重大战略，其实施与发展已走过第一个五年。本文立足中国特色哲学社会科学面向该领域的研究，从社科基金和CSSCI来源刊物双维度出发，借助多种研究方法和多类可视化工具，对“一五”周期内的研究现状展开全面系统的梳理和总结，把握其研究格局、轨迹和热点，并探析其存在的部分研究壁垒，同时对其未来的发展予以展

望，以期实现可持续的长久发展。

参考文献

- [1] 岳经纶, 黄博函. 健康中国战略与中国社会政策创新[J]. 中山大学学报(社会科学版), 2020, 60(1): 179-187.
- [2] 李姝婧, 康秀云. 在应对重大疫情中彰显党的初心使命[J]. 人民论坛, 2020(18): 96-97.
- [3] 全国哲学社会科学工作办公室. 国家社会科学基金年度项目立项结果[EB/OL]. (2021-09-24)[2023-05-05]. <http://www.nopss.gov.cn>.
- [4] 谢娟, 龚凯乐, 成颖, 等. 论文下载量与被引量相关关系的元分析[J]. 情报学报, 2017, 36(12): 1255-1269.
- [5] 胡鞍钢, 方旭东. 全民健身国家战略: 内涵与发展思路[J]. 体育科学, 2016, 36(3): 3-9.
- [6] NAUDE F. Comparing downloads, Mendeley readership and Google Scholar citations as indicators of article performance[J]. The Electronic Journal of Information Systems in Developing Countries, 2017, 78(1): 1-25.
- [7] 钟镇. 知识图谱分析方法的可靠性检验研究——以共词分析为例[J]. 科学学研究, 2015, 33(5): 647-653.
- [8] 余达淮, 王世泰. 习近平关于人民健康重要论述的内涵、实践价值与世界意义[J]. 南京社会科学, 2020(12): 1-8.
- [9] 鞠忠美. 习近平关于人民健康重要论述探析[J]. 山东社会科学, 2020(9): 88-93.
- [10] 杨抗抗. 习近平关于人民健康重要论述的深刻思想内涵[J]. 人民论坛·学术前沿, 2020(15): 140-143.
- [11] 潘慧. 广西面向2035年民族大健康产业发展愿景及其对科技需求研究[J]. 广西民族研究, 2020(1): 136-141.
- [12] 何秋洁, 杨晓维. 大健康产业与养老服务的耦合协调度分析[J]. 软科学, 2019, 33(10): 45-49.
- [13] 周晓英. 健康服务: 开启公共图书馆服务的新领域[J]. 中国图书馆学报, 2019, 45(4): 61-71.
- [14] 江光荣, 李丹阳, 任志洪, 等. 中国国民心理健康素养的现状与特点[J]. 心理学报, 2021, 53(2): 182-

- 201.
- [15] 夏翠翠, 李建新. 健康老龄化还是病痛老龄化——健康中国战略视角下老年人口的慢性病问题 [J]. 探索与争鸣, 2018(10): 115-121.
- [16] 陆杰华, 阮韵晨, 张莉. 健康老龄化的中国方案探讨: 内涵、主要障碍及其方略 [J]. 国家行政学院学报, 2017(5): 40-47.
- [17] 胡扬. 从体医分离到体医融合——对全民健身与全民健康深度融合的思考 [J]. 体育科学, 2018, 38(7): 10-11.
- [18] 冯振伟, 张瑞林, 韩磊磊. 体医融合协同治理: 美国经验及其启示 [J]. 武汉体育学院学报, 2018, 52(5): 16-22.
- [19] 于潇, 包世荣. 健康中国背景下医养结合养老模式研究 [J]. 社会科学战线, 2018(6): 271-275.
- [20] 包世荣. 国外医养结合养老模式及其对中国的启示 [J]. 哈尔滨工业大学学报 (社会科学版), 2018, 20(2): 58-63.
- [21] 姚力. 从卫生与健康事业发展看新中国 70 年的成就与经验 [J]. 毛泽东邓小平理论研究, 2019(11): 52-57.
- [22] 刘艳飞, 胡晓辉. 健康中国战略下的健康服务供给模式优化研究 [J]. 福建论坛 (人文社会科学版), 2019(3): 59-66.
- [23] 杨鸿柳, 杨守明. 构建人类卫生健康共同体的现实境遇与路径选择 [J]. 福建师范大学学报 (哲学社会科学版), 2020(4): 41-49.
- [24] 陶倩. 携手打造人类卫生健康共同体 [J]. 红旗文稿, 2020(11): 44-45.
- [25] 陆杰华, 汪斌. 居民互联网使用对其自评健康影响机制探究——基于 2016 年中国家庭追踪调查数据 [J]. 中山大学学报 (社会科学版), 2020, 60(3): 117-127.
- [26] 石郑. 流动人口健康自评状况及影响因素分析 [J]. 江汉学术, 2020, 39(2): 17-28.
- [27] 文军. 新型冠状病毒肺炎疫情的爆发及共同体防控——基于风险社会学视角的考察 [J]. 武汉大学学报 (哲学社会科学版), 2020, 73(3): 5-14.



开放科学
(资源服务)
标识码
(OSID)

吉林省高校科研合作网络研究 ——基于 2013—2023 年科研论文的分析

孙毅

吉林大学图书馆 长春 130012

摘要: [目的/意义] 区域内高校科研合作是其开展科学研究活动的主要方式之一,了解吉林省高校科研合作情况,有利于优化合作网络结构,促进区域协同发展。[方法/过程] 以 2013—2023 年吉林省 31 所高校的科研合作论文为研究样本,通过文献计量及社会网络分析的方法,构建科研合作网络,提取合作网络结构特征,分析其演化规律。[局限] 研究基于论文成果展开,未对科学研究中其他合作方式进行定量分析。[结果/结论] 吉林省高校广泛参与到了科研合作之中,合作网络具有小世界特性。合作网络节点中心度分布不均匀,马太效应明显。各高校处于不同的位置层次,合作网络具有明显的特色聚集性,地域聚集性不显著。演化过程中,合作频次逐年上升,合作深度不断提高,网络整体平衡性有所改善。

关键词: 吉林省高校; 科研合作; 社会网络; 网络结构; 网络演化

中图分类号: G35

Research Collaboration Network of Universities in Jilin Province: Analysis Based on Research Papers from 2013 to 2023

SUN Yi

Jilin University Library, Changchun 130012, China

Abstract: [Objective/Significance] Scientific research collaboration among universities within a region is one of the essential ways in which they carry out scientific research activities. Understanding the collaboration among universities in Jilin Province is beneficial for optimizing the structure of collaboration network and promoting the development of research in the region. [Methods/Processes] Using research papers of 31 universities in Jilin Province, from 2013 to 2023, as research samples, this study constructs a research collaboration network through bibliometric and social network analysis, extracts the structural feature of the collaboration network, and analyzes its evolution laws. [Limitations] The study focuses on the research papers and does not

基金项目 吉林省高教学会高教科研课题“特色高水平大学建设背景下吉林省高校科研合作态势研究”(JGJX24C004)。

作者简介 孙毅(1991-), 硕士, 馆员, 主要研究方向为社会网络分析、科技情报分析, E-mail: yisun@jlu.edu.cn。

引用格式 孙毅. 吉林省高校科研合作网络研究——基于 2013—2023 年科研论文的分析[J]. 情报工程, 2025, 11(1): 109-118.

quantitatively analyze other forms of collaboration in scientific research. [Results/Conclusions] Universities within Jilin Province have widely participated in scientific research collaboration, and the network of collaboration demonstrates the characteristics of small-world. The distribution of centrality in the collaboration network is uneven, with a pronounced Matthew effect. Universities being positioned at different levels, the collaboration network exhibits distinct characteristic clustering, while regional clustering is not significant. In evolution, the frequency of collaboration has increased, the collaboration has been enhanced, and the overall balance of the network has been improved.

Keywords: Universities in Jilin Province; Scientific Research Collaboration; Social Network; Network Structure; Network Evolution

引言

在 2024 年的全国科技大会、国家科学技术奖励大会、两院院士大会上，习近平总书记指出，“科技进步是世界性、时代性课题，唯有开放合作才是正道”^[1]。随着科学研究的不断深入与发展，科研合作已经成为研究活动的主要方式，同时也是研究机构之间协同创新的重要途径之一。高等院校作为我国科技创新体系的重要组成部分，既是培养高素质人才的基地，也是开展科研合作的核心参与者。高校之间往往通过开放共享实验资源，共同承担研究项目，组织召开学术会议，联合培养创新人才等方式建立科研合作网络。2022 年教育部在《关于加强高校有组织科研 推动高水平自立自强的若干意见》中指出，提升区域高校协同创新能力，服务区域高质量发展是加强高校有组织科研的重点举措^[2]。基于高校区域内科研合作的研究，能够明确区域高校合作发展现状，进而服务区域未来科技合作活动的开展。

论文合著是科研合作的重要表现形式，1963 年 Price^[3] 提出运用一系列计量指标，如合著指数、篇均作者数、合著率以及合作强度等来研究科研合作的方式和程度等内容。随着科

学研究交流的不断深入发展，科研合作逐渐呈现网络化的特点。2001 年 Newman^[4] 将社会网络分析法引入科研合作的研究中，合作网络的结构特征及其演化规律成为相关研究的重点之一。重点高校间的科研合作关系已获得了广泛关注，邱均平等^[5] 对我国“985 工程”高校间的科研合作关系网络进行了研究；郭崇慧等^[6] 进一步比较了“985”高校国际与国内期刊的科研合作网络。柴玥等^[7] 对“211 工程”大学间的合作论文定量分析，计算了各高校的中介中心度，并绘制了合作图谱。赵蓉英等^[8] 基于一流大学建设高校的合作网络的结构、关系、规律，分析了合作网络演化产生的主要原因；李俊良^[9] 将我国双一流高校与美国知名高校的合作网络进行了对比研究。与此同时，区域高校合作的研究也在逐步展开，苏一凡等^[10] 以广州地区 23 所本科院校在 2012 年的合著论文为样本，探讨了学科相似性对科研合作的影响。李非凡^[11] 通过对京津冀地区“211”及省部共建的 26 所高校合作网络的分析，发现地域聚集特性是影响科研合作关系的主要因素。宗晓华等^[12] 对长三角地区的“双一流”建设高校合作网络进行分析发现，上海地区高校有着突出的中介优势。吉林省高等教育资源、科技资源丰富，

在科研合作中具有显著的信息“输出型”特点，主导性较强^[13]，其省内高校合作特征与发展态势有一定的研究价值。本文以吉林省高等学校为研究对象，基于各高校论文合作情况，构建科研合作网络，提取其结构特征，分析其发展态势，描绘出吉林省高校科研合作图景。

1 数据来源与研究方法

1.1 数据来源

本文依照教育部公布的普通高等学校名单^[14]，根据高校具体情况及数据的可获取性确定 31 所吉林省高校为研究样本，具体名单如表 1 所示。

表 1 吉林省高校名单

序号	吉林省高校	序号	吉林省高校
1	白城师范学院	17	吉林工程技术师范学院
2	北华大学	18	吉林工商学院
3	长春大学	19	吉林化工学院
4	长春大学旅游学院	20	吉林建筑大学
5	长春工程学院	21	吉林警察学院
6	长春工业大学	22	吉林农业大学
7	长春建筑学院	23	吉林农业科技学院
8	长春理工大学	24	吉林师范大学
9	长春汽车职业技术大学	25	吉林体育学院
10	长春师范大学	26	吉林外国语大学
11	长春中医药大学	27	吉林医药学院
12	东北电力大学	28	吉林艺术学院
13	东北师范大学	29	通化师范学院
14	吉林财经大学	30	延边大学
15	吉林大学	31	中国人民解放军空军航空大学
16	吉林动画学院		

InCites 是科睿唯安公司基于 Web of Science 数据库建立的科研数据分析平台，利用 InCites

对吉林省高校产出的 SCI 论文进行统计分析。使用机构分析模块，限定机构类型为“Academic”，时间检索范围为 2013 年至 2023 年，文献类型为“Article”及“Review”两种，国家/地区为“JILIN, CHINA MAINLAND”，获取吉林省内高校文献计量数据及相关文献信息。进一步通过限定合作机构，逐一获取与吉林省各高校产生科研合作的高校集合，得到相应高校之间的合作数据。

1.2 研究方法

1.2.1 文献计量分析

文献计量是一种利用数学统计方法分析文献特征规律及变化的定量研究方法。合著论文是高校之间科研合作的重要表现形式，是研究高校科研合作的重要切入点^[15]。本文采用合作频次、合作率、Salton 指数等计量指标对吉林省高校科研合作情况进行分析。

1.2.2 社会网络分析

社会网络分析是一种将社会关系抽象成为数学表达的研究方法。社会网络分析可以兼顾定量与定性研究，与文献计量分析相结合，能够有效构建各机构之间的合作关系网络，并提取网络结构特征。绘制可视化图谱，可以直观地展现机构间的拓扑关系，进而多维度、动态地刻画高校合作情况。

本文利用 UCINET 软件计算节点数、边数、网络密度、平均距离、群集系数、网络效率等指标，对吉林省高校科研合作整体网络进行分析；计算中心度及核心值等指标，对个体网络进行分析；使用 CONCOR 程序，对合作网络进行块模型聚类分析。吉林省高校科研合作网络

图谱由 Gephi 软件绘制。

2 吉林省高校科研合作情况

2.1 整体合作情况

高等院校之间的合作频次可以通过共同发表的学术论文数量来表征，是评价高校间合作深度的重要指标。高校在样本集合中合作论文数量（不含高校内部合作论文）与高校论文总

量的比值记为高校间的合作率，能够有效反映高校科学研究的合作意愿。

2013—2023 年，吉林省高校间共发表区域内科研合作论文 22621 篇，占全省高校论文总量的 18.55%。具体科研合作态势如图 1 所示。吉林省高校之间的合作频次整体呈稳定上升趋势，随着科学研究的持续深入，科研合作的需求扩大，吉林省高校对科研合作的重视程度逐步提高，高校之间的合作正在不断地扩展与深化。

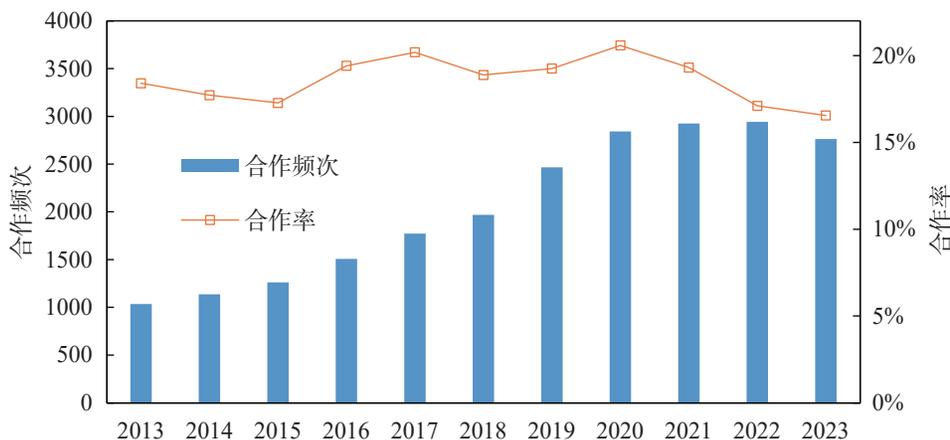


图 1 2013—2023 年吉林省高校科研合作趋势

基于 ESI 分类体系，对吉林省高校合作的研究领域进行分析。吉林省高校合作领域分布广泛，各领域均有成果产出。研究主要集中于化学与物理学等自然科学领域，材料科学与工程学等应用科学领域以及临床医学等生命健康领域，其中以化学领域发文最多，为 3362 篇。具体研究领域分布如图 2 所示。

2.2 合作关系情况

为反映个体之间的合作情况，本文利用各高校间的科研合作频次，构建了 31×31 的合作频次矩阵 F_{ij} 。以高校间是否发表合著论文为依据，确定其是否存在合作关系，对 F_{ij} 进行二值

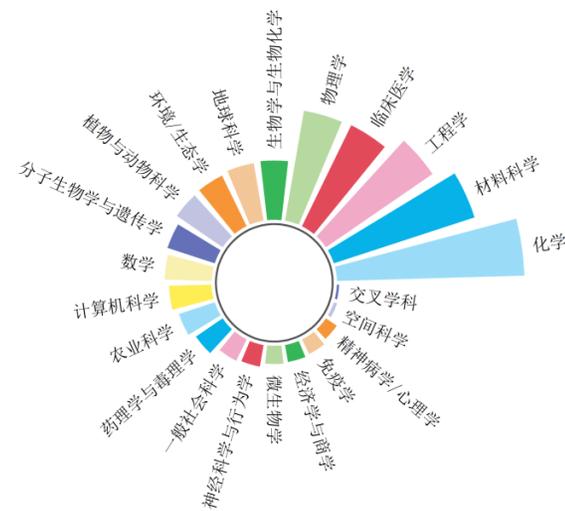


图 2 吉林省高校科研合作研究领域分布

化处理，构建相应合作关系矩阵（无权矩阵） R_{ij} 。由于高校科研产出的差异性较大，为消除

其影响,本文采用改进的 Salton 指数^[16]对高校间的科研合作强度进行测量,构建合作强度矩阵 S_{ij} ,其计算公式为:

$$S_{ij} = \frac{n_{ij}}{\sqrt{c_i c_j}} \quad (1)$$

其中, n_{ij} 表示高校 i 与高校 j 之间合作论文数量, c_i 与 c_j 分别表示样本内高校 i 和高校 j 各自合作论文总量。两所高校之间的 Salton 指数数值越大,高校之间的合作联系越紧密。

对 S_{ij} 进行统计分析,得到吉林省高校主要合作关系(合作强度前 20)的组合,如表 2 所示。

(1) 高校开展科研合作有中心化趋势,吉林大学与其他高校形成了 11 对组合。一方面,吉林大学

是吉林省唯一的“985”高校,自身科研实力较强,对其他高校具有一定的辐射作用;另一方面,吉林大学学科门类齐全,能够与不同类型的高校在各领域开展较为深入的合作。(2) 办学特色相近的高校具有较强的合作倾向,吉林农业大学和吉林农业科技学院、吉林师范大学和白城师范学院、东北师范大学和长春师范大学等组合的合作强度较高。相近的办学特色使各高校能够更好地进行资源共享与人才交流,从而形成更为紧密的合作关系。(3) 城市区域邻近性对吉林省高校科研合作影响并不显著,主要合作关系中有 9 组为跨城市地区间的高校合作,高校间频繁的交叉合作可能促进了省域内科研合作的一体化进程。

表 2 吉林省高校合作强度前 20 的组合

序号	组合	合作强度	序号	组合	合作强度
1	吉林大学; 东北师范大学	0.311	11	吉林大学; 吉林化工学院	0.163
2	吉林大学; 长春工业大学	0.276	12	吉林大学; 北华大学	0.160
3	吉林大学; 长春理工大学	0.257	13	东北师范大学; 吉林农业大学	0.159
4	吉林大学; 吉林农业大学	0.237	14	吉林大学; 长春大学	0.155
5	吉林大学; 长春中医药大学	0.232	15	吉林师范大学; 白城师范学院	0.131
6	吉林农业大学; 吉林农业科技学院	0.205	16	东北师范大学; 长春师范大学	0.129
7	吉林大学; 吉林师范大学	0.187	17	长春理工大学; 吉林师范大学	0.120
8	吉林大学; 吉林建筑大学	0.174	18	吉林大学; 延边大学	0.118
9	延边大学; 吉林医药学院	0.165	19	吉林大学; 东北电力大学	0.117
10	东北师范大学; 长春理工大学	0.163	20	吉林大学; 长春师范大学	0.115

3 吉林省高校科研合作网络

3.1 整体网络结构特征

将高校抽象为节点,合作关系抽象为节点之间的边,点边的结合构成了高校科研合作网络的基本形态,了解其基本特征,对高校科研合作模式的研究有着重要作用。利用 UCINET 软件计算吉林省科研合作关系网络的特征值如

表 3 所示;利用 Gephi 绘制吉林省科研合作强度网络,如图 3 所示。

表 3 吉林省高校科研合作网络特征

特征	特征值	特征	特征值
节点数	31	平均距离	1.484
边数	482	群集系数	0.849
关系网络密度	0.518	网络效率	0.515
强度网络密度	0.021		

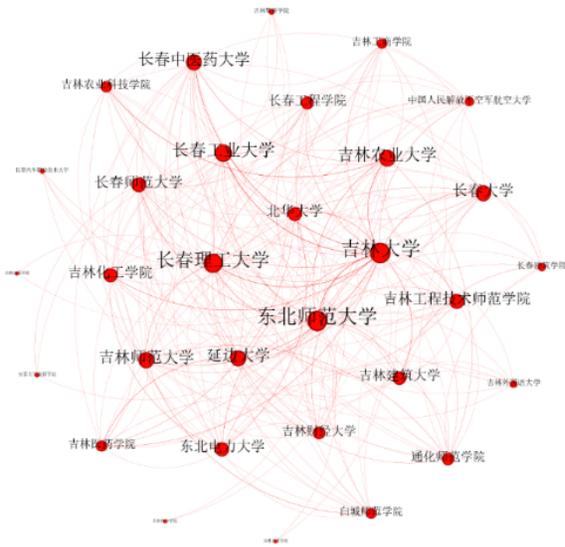


图3 吉林省高校科研合作网络

吉林省高校科研合作网络节点数为31，边数为482，网络中不存在孤立点，吉林省高校广泛参与到了科研合作网络之中。合作关系网络密度为0.518，已建立的合作关系是可能合作关系的一半左右，兼具结构的稳定性与发展的潜力。合作网络的平均距离为1.484，群集系数为0.849。高校间平均仅需通过2个节点即可建立合作关系，网络的连通性较高，聚集性较强，具有小世界网络的特性。

3.2 个体网络结构特征

计算合作强度网络中各高校的3种中心度及核心度，部分结果如表4所示。点度中心度能够衡量高校在网络中的合作强度，是确定合作网络中重要参与者的关键。接近中心度能够表征高校间合作通达性，其值越高寻求合作越为便捷，也易与其他高校建立更为有效的合作关系。中介中心度可以揭示高校在网络中的连接性，中介性较高的节点是网络中高校间开展合作的重要纽带与桥梁。核心边缘位置关系能够确定节点在网络中

所处的位置，吉林大学的核心度最高，处于网络的核心位置；东北师范大学、长春理工大学、吉林农业大学、长春工业大学、长春中医药大学、吉林师范大学、北华大学、长春大学、吉林建筑大学、吉林化工学院等高校处于核心边缘位置，其余高校处于边缘位置。可以发现，在吉林省高校科研合作网络中，节点特征值分布不均匀，吉林大学等高校多个特征值均位于前列，处于网络中最为重要且最为有利的位置，对其他高校影响力较大，整体网络呈现出一定的无标度网络倾向，马太效应较为显著。

表4 个体网络中心度及核心度(部分)

	点度中心度	接近中心度	中介中心度	核心度
吉林大学	36.531	96.774	13.717	0.739
东北师范大学	19.897	96.774	13.257	0.309
长春理工大学	15.255	90.909	7.395	0.239
吉林农业大学	12.806	81.081	1.388	0.211
吉林师范大学	9.784	78.947	1.494	0.164
长春工业大学	9.413	83.333	3.627	0.195
长春中医药大学	8.621	78.947	1.241	0.167
延边大学	8.059	78.947	1.199	0.116
北华大学	7.926	73.171	0.396	0.133
长春大学	6.841	76.923	1.183	0.129

3.3 合作网络层次特征

块模型方法可以构建网络相对独立的区域，以便更直观地观测网络中的层次结构。对高校合作强度网络进行块模型聚类，聚类结果如图4所示，块密度矩阵如表5所示。采用“ α -块密度指标”模型方法对像密度矩阵进行变换，利用像矩阵绘制像矩阵简化图，如图5所示。

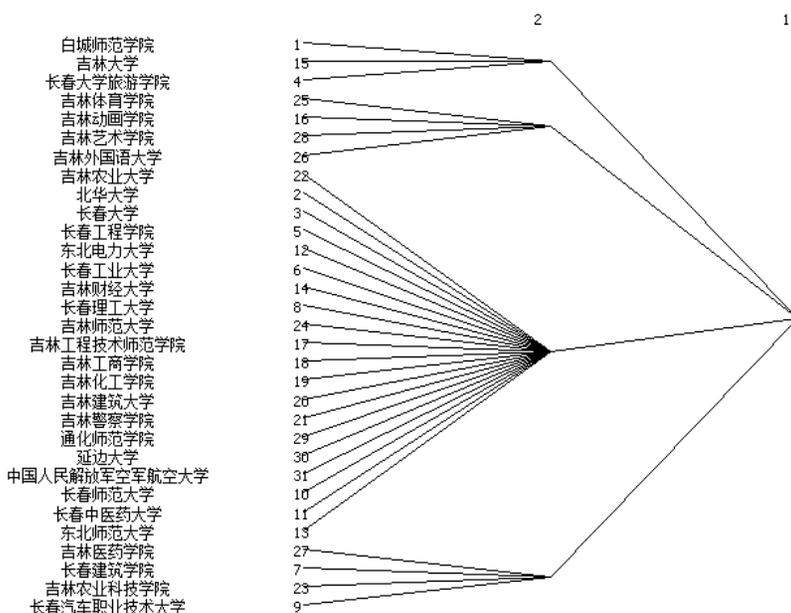


图 4 吉林省高校科研合作网络块模型聚类

表 5 吉林省高校科研合作网络块密度矩阵

	块 1	块 2	块 3	块 4
块 1	0.032	0.012	0.100	0.058
块 2	0.012	0.000	0.025	0.000
块 3	0.100	0.025	0.031	0.033
块 4	0.058	0.000	0.033	0.005

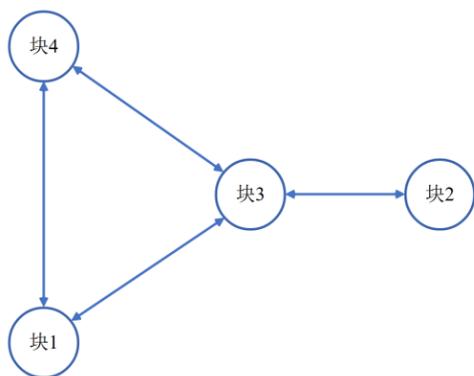


图 5 吉林省高校科研合作网络像矩阵简图

吉林省高校科研合作网络可以划分为 4 个块。块内密度能够衡量块内部合作紧密程度，往往由节点特征值与节点的分布决定。吉林大学作为核心节点的重要作用，促进了块 1 内部成员的交流；与此同时，块 3 内部成员较多，

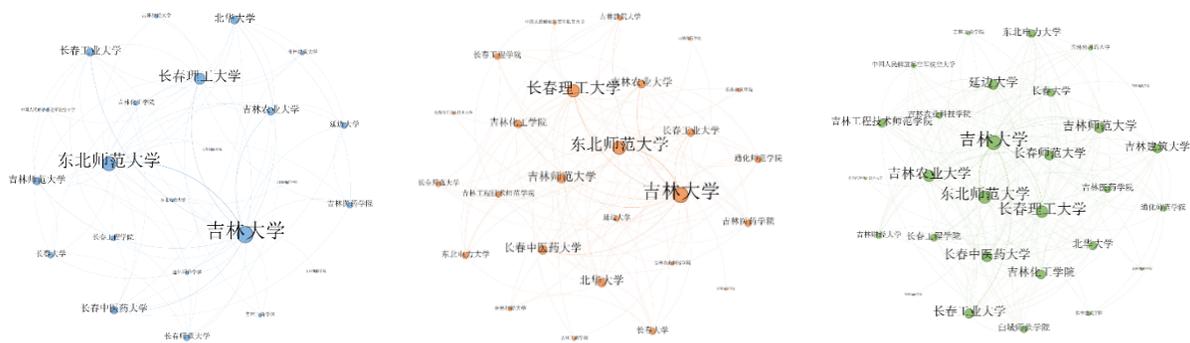
分布较广，这使得块 1、块 3 的块内密度显著高于块 4 与块 2。块 4 中包含吉林艺术学院、吉林外国语学院、吉林体育学院以及吉林动画学院等 4 所办学特色鲜明的高校，其在网络中具有相似的位置特性，但由于办学差异性较大，无法建立有效的合作关系，块密度为 0。像矩阵能够有效地揭示块与块之间的合作关系。块 3 与其余各块均有联系，在网络中处于枢纽位置，显示出较高的凝聚力；相较之下，块 2 仅与块 3 建立联系，处于网络的边缘位置，不利于该块内各高校发挥特色办学的优势与潜力。

3.4 合作网络演化特征

社会网络的拓扑结构、关系层级往往随时间变化，为研究吉林省高校科研合作网络的演化趋势，计算 2013—2023 年吉林省高校科研合作网络特征，如表 6 所示，分别绘制 2013、2018、2023 年的网络图谱，如图 6 所示。

表 6 2013—2023 年吉林省高校科研合作网络特征

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
节点数	23	25	23	25	26	26	28	27	27	29	29
边数	144	174	156	196	210	218	272	286	296	334	356
网络密度	0.285	0.290	0.308	0.327	0.323	0.335	0.360	0.407	0.422	0.411	0.438
群集系数	0.735	0.770	0.745	0.728	0.777	0.755	0.773	0.793	0.785	0.788	0.767
平均距离	1.755	1.747	1.704	1.737	1.677	1.680	1.661	1.604	1.621	1.613	1.576
网络效率	0.784	0.772	0.758	0.732	0.730	0.720	0.690	0.640	0.625	0.632	0.603



2013 年吉林省高校科研合作网络

2018 年吉林省高校科研合作网络

2023 年吉林省高校科研合作网络

图 6 吉林省高校科研合作网络演化

2013—2023 年，吉林省高校科研合作网络节点数、边数逐渐增加，孤立点逐渐减少，更多的高校参与到合作网络之中。网络密度一直保持上升趋势，各高校参与合作的意愿增强，合作频次不断增加。群集系数基本保持不变，网络平均距离逐渐缩短，各高校之间的连通性增强，网络整体的小世界属性增强。网络效率呈现逐年下降的趋势，一方面是由于网络规模的不断扩大，边缘节点对网络效率有一定影响；另一方面是高校间合作逐步优化，网络冗余度增加的结果。

图 6 更为直观地揭示了吉林省科研合作网络的演变趋势，吉林省高校科研合作网络逐步由稀疏网络演化出更为复杂的网络结构。吉林大学等高校最初即是合作网络的关键节点，随

着亚核心高校的不断发展和，网络外高校的不断加入，最终形成了以吉林大学、东北师范大学、长春理工大学等高校为核心节点，向边缘高校辐射的一体化网络格局。吉林省高校科研合作网络整体不平衡程度有所下降，网络稳定性日益增强。

4 结论与展望

4.1 研究总结

本文对吉林省 31 所高校 2013 年至 2023 年的科研合作情况进行分析，并构建合作网络，揭示网络结构及演化特征，得到以下结论：

(1) 吉林省高校科研合作网络是一个联系广泛的网络，没有孤立高校的存在。同时，高

校合作的强关联围绕吉林大学、东北师范大学、长春理工大学等高校展开。吉林省高校合作具有较强的连通性与聚集性，合作网络具有小世界特性，但未发现明显的城市地域聚集现象。高校科研合作网络的建立有助于科研活动的协同开展以及资源流动的高效进行。

(2) 吉林省高校科研合作网络节点中心度分布不均匀，具有一定无标度网络倾向，马太效应明显。吉林大学由于其较强的科研实力以及齐全的学科门类，处于合作网络的核心位置；东北师范大学、长春理工大学等高校处于亚核心位置；几所头部高校主导着科研合作关系的建立与开展。部分边缘高校与合作网络联系较弱，某些时期较易成为孤立点，影响到合作网络整体的连通性与稳定性。

(3) 吉林省高校在合作网络中处于不同的位置层次，合作网络具有明显的特色聚集性。吉林艺术学院、吉林外国语大学、吉林体育学院以及吉林动画学院等不同特色的高校具有相同的网络位置层次特性，没有与其他高校建立有效的合作关系。与此同时，特色相近的高校，如吉林农业大学和吉林农业科技学院、吉林师范大学和白城师范学院、东北师范大学和长春师范大学之间具有较强的合作关系。

(4) 吉林省高校合作频次逐年上升，高校间的合作深度不断提高。科研合作网络规模逐渐扩大，小世界特性不断增强。合作网络效率下降，网络冗余度增加，同时合作网络有从单一核心向多核心发展的演化趋势。吉林省高校间合作逐渐深入，整体不平衡程度有所下降，网络的稳定性日益增强。

4.2 发展建议

吉林省高校在多年的合作发展过程中已经形成了较为复杂的合作网络，伴随着科学研究不断深入，合作网络结构也日渐合理；与此同时，应有意识地加强合作交流，优化网络结构，进一步提高科研效率，实现省域高校的协同发展。

一方面，积极促进不同类型、不同层次的高校开展合作。吉林大学、东北师范大学等核心高校应发挥示范作用，承担一定的培训交流任务，服务于整体科研合作网络的结构优化。区域特色高校需明确定位，突出特色优势，结合吉林省内高校合作领域分布特点，加速学科领域的融合发展。边缘高校要切实融入科研合作网络，防止自身成为信息孤岛，保证整体网络连通性，提高网络均衡性，持续优化吉林省高校合作网络一体化格局。

另一方面，加强政策倾斜与项目支持，完善科研合作机制，推动有组织科研。充分把握科研合作的现实情况与发展脉络，围绕吉林省高质量发展战略需求，制定科技合作政策，指导高校间科研合作活动，赋能区域产业、经济发展。组织规划跨学科、跨领域的重大科研项目，打造创新合作平台，为科研合作提供全方位支持，借由科研合作实现解决重要问题、突破关键技术、促进科技进步、服务社会发展的长期目标。

参考文献

- [1] 习近平. 在全国科技大会、国家科学技术奖励大会、两院院士大会上的讲话 [EB/OL]. (2024-

- 06-24)[2024-10-10]. https://www.gov.cn/yaowen/liebiao/202406/content_6959120.htm.
- [2] 教育部. 教育部印发《关于加强高校有组织科研推动高水平自立自强的若干意见》[EB/OL]. (2022-08-29)[2024-10-15]. http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202208/t20220829_656091.html.
- [3] PRICE D. Little Science, Big Science[M]. West Sussex: Columbia University Press, 1963.
- [4] NEWMAN M E J. The structure of scientific collaboration networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(2): 404-409.
- [5] 邱均平, 温芳芳. 我国“985工程”高校科研合作网络研究[J]. 情报学报, 2011, 30(7): 746-755.
- [6] 郭崇慧, 王佳嘉. “985工程”高校校际科研合作网络研究[J]. 科研管理, 2013, 34(S1): 211-220.
- [7] 柴玥, 刘趁, 王贤文. 我国高校科研合作网络的构建与特征分析——基于“211”高校的数据[J]. 图书情报工作, 2015, 59(2): 82-88.
- [8] 赵蓉英, 王旭, 亓永康. 我国世界一流大学建设高校间科研合作网络及演化研究[J]. 现代情报, 2019, 39(3): 132-143.
- [9] 李俊良. 科研合作网络特征分析及比较——以我国“双一流”大学与美国知名大学为例[J]. 情报探索, 2020(4): 35-43.
- [10] 苏一凡, 朱少强, 谢卫红, 等. 广州地区高校间科技创新合作现状探讨——基于2012年合著论文的计量图谱分析[J]. 科技管理研究, 2014, 34(6): 90-95.
- [11] 李非凡. 高校科研合作网络及演化研究——以京津冀地区211及省部共建高校为例[J]. 农业图书情报, 2019, 31(8): 31-39.
- [12] 宗晓华, 王立成. 高校科研合作的网络结构及其演变[J]. 高教发展与评估, 2022, 38(6): 9-23, 119-120.
- [13] 李百华, 董志良, 王刚, 等. 中国省际科研合作网络信息流向与关键路径研究[J]. 中国高校科技, 2023(Z1): 14-19.
- [14] 教育部. 全国高等学校名单[EB/OL]. (2024-06-21)[2024-10-10]. http://www.moe.gov.cn/jyb_xxgk/s5743/s5744/A03/202406/t20240621_1136990.html.
- [15] 石慧. 面向区域和类型合作的我国大陆地区图书馆整体科研现状研究[J]. 图书馆工作与研究, 2022(6): 75-85.
- [16] LIANG L, ZHU L. Major factors affecting China's inter-regional research collaboration: Regional scientific productivity and geographical proximity[J]. Scientometrics, 2002, 55(2): 287-316.



开放科学
(资源服务)
标识码
(OSID)

石油类高校高被引论文特征挖掘及影响因素研究 ——以中国石油大学（北京）为例

刘天琳 陆桃妹 张芹

中国石油大学（北京）北京 102249

摘要: [目的/意义] 为提高石油类高校科研竞争力、国际影响力, 梳理石油类高校高被引论文的整体情况, 以供同行业高校在提升论文学术质量及实际应用价值方面参考。[方法/过程] 以中国石油大学（北京）2018—2023 年高被引论文为研究样本, 总结了论文本身、期刊、作者及参考文献类型 4 类 21 个评价因素, 综合运用随机森林方法, 挖掘高被引论文的特征及影响因素。[结果/结论] 在论文本身方面, 早期被引量、研究领域较重要; 在期刊方面, 期刊影响因子则是最为重要的因素; 在作者方面, 作者声誉反映了作者在学术界的地位和影响力, 作者合作规模在一定程度上反映了研究团队的实力和合作能力; 在参考文献方面, 参考文献平均影响因子、参考文献的累计被引量两个指标需要关注。根据数据特征挖掘, 提出了提升论文质量与创新性、优化论文发表策略等 4 方面的提升石油类高校高被引论文数量的策略。

关键词: 石油类高校; 高被引论文; 特征挖掘; 影响因素

中图分类号: G35; G255.2

Research on the Characteristics and Influencing Factors of Highly Cited Papers in Petroleum Universities: Taking China University of Petroleum (Beijing) as an Example

LIU Tianlin LU Taomei ZHANG Qin

China University of Petroleum (Beijing), Beijing 102249, China

Abstract: [Objective/Significance] In order to improve the scientific research competitiveness and international influence of petroleum universities, the overall situation of highly cited papers in petroleum universities is sorted out, so as to provide

基金项目 中国石油大学（北京）研究生教育质量与创新工程重点项目子课题“AIGC 时代学术诚信教育与 AI 学术伦理研究（yjs2024018）”。

作者简介 刘天琳（1992-），硕士，馆员，主要研究方向为石油情报及知识产权信息分析；陆桃妹（1992-），硕士，馆员，主要研究方向为图书情报、知识产权分析，Email: 1016195164@qq.com；张芹（1981-），博士，副研究馆员，主要研究方向为石油情报分析。

引用格式 刘天琳, 陆桃妹, 张芹. 石油类高校高被引论文特征挖掘及影响因素研究——以中国石油大学（北京）为例 [J]. 情报工程, 2025, 11(1): 119-127.

reference for the same industry universities in improving the content, academic quality and practical application value of papers. [Methods/Processes] The author took the highly cited papers of China University of Petroleum (Beijing) in the past six years from 2018 to 2023 as research samples, and summarized twenty-one evaluation factors in four categories: the papers themselves, journals, authors and reference types. The random forest method was comprehensively applied to explore the characteristics and influencing factors of highly cited papers. [Results/Conclusions] The results show that: In terms of the papers themselves, the early citation amount and research field are more important; In terms of journals, journal impact factor is the most important factor. As for the author, the author reputation reflects the status and influence of the author in the academic circle, and the scale of the author cooperation reflects the strength and cooperation ability of the research team to some extent. In terms of references, the average impact factor of references and the cumulative number of references need to be paid attention to. Based on data feature mining, the paper puts forward four strategies to improve the number of highly cited papers in petroleum universities, including improving the quality and innovation of papers and optimizing the publishing strategy.

Keywords: Petroleum Universities; Highly Cited Papers; Feature Mining; Influencing Factors

引言

在全球科研竞争日益激烈的背景下，高被引论文作为衡量科研影响力和学术质量的重要指标，受到学术界广泛关注。根据基本科学指标（ESI）数据库的界定，高被引论文指近十年间累计被引用次数进入各学科世界前1%的论文。高被引论文评价受多种因素共同影响，根据现有研究，论文本身、被引用论文的作者、摘要、期刊、领域和参考文献等是影响论文被引频次的主要因素^[1-2]；部分探究影响因素与被引频次的相互关系，如呈线性、非线性、U型关系等^[3-4]。另外，研究课题新颖程度、文章类型、不同学科、国际合作关系也有影响^[5-6]。

现有研究成果是从全局角度开展高被引论文影响因素评价及被引量预测，研究样本并没有针对具体科研机构、具体专业领域，更没有涉及石油领域高校的深入分析，特别是在能源战略性领域的研究尤为不足。作为全球最大的能源消费国和生产国，中国的石油能源安全对国家发展具有重大战略意义。石油领域高校在

保障国家能源安全、推动能源科技创新方面肩负着重要使命。在此背景下，深入研究石油领域高校的高被引论文特征及其影响因素，对于提升科研核心竞争力、增强国际学术影响力具有重要的现实意义。

基于此，本文聚焦石油行业重点高校——中国石油大学（北京），通过构建科学评价指标体系，深入分析高被引论文的特征及其影响因素。研究采用机器学习方法，系统挖掘影响高被引论文的关键因素，并在此基础上提出培育高被引论文的有效路径。本文不仅填补了石油领域高被引论文研究的空白，也为提升石油领域高校的科研水平和国际影响力提供了理论依据和实践指导。

1 中国石油大学（北京）ESI 高被引论文概况

1.1 数据来源

本文以中国石油大学（北京）2018—2023

年去除重复论文后的 223 条高被引论文作为分析样本，在 SCI、INCITES 网站中进行相关指标数据下载。

1.2 高被引论文特征分析

中国石油大学（北京）作为石油领域的重点高校，正在全面构建能源特色鲜明学科体系，争创世界一流学科，高被引论文呈现 5 个方面的特征：（1）高被引论文数量规模增长，2023 年该校产出 216 篇高影响力论文，总体数量比 2018 年增长 18%；（2）论文影响广泛，被中国、美国、印度、澳大利亚等 23 个国家或地区引用，国内、国外引用比为 7:3，自引、他引比约为 1:9；（3）优势学科明显，高影响力论文产出于本校优势学科，尤其以工程学、地球科学、化学三大学科产出最多；另外，在物理学、经济与商业、临床医学、生物与生化等学科中也有少量高影响力论文产出；（4）论文整体质量较高，高达 84% 的高影响力论文属于中科院分区 1 区；（5）国际合作广泛，在高影响力论文中，国际合作论文占比 52%，与 210 个国际机构有论文合作，且大多数是与美国等发达国家的国际合作。

通过对发文关键词进行聚类分析，可以发现高被引论文主要围绕多相热流体与热传递特性、非冷凝性气体与性能分析、分离与吸附、优化与模型恢复以及其他相关主题进行聚类，如图 1 所示。这些主题共同构成了石油领域的重要研究内容和技术挑战，主要研究方向包括：pore structure（孔隙结构）、photocatalysis（光催化）、shale gas（页岩气）、pyrolysis（热

解）、natural gas hydrate（天然气水合物）、CO₂ emissions（二氧化碳排放）、heat-transfer characteristics（传热特性）等。例如，图中红色圆圈展示的和页岩气（shale gas）研究相关的关键词网络，中心是 pore structure（孔隙结构），这表明它是主要的研究领域。与“pore structure”直接相连的关键词包括“permeability”（渗透率）、“porosity”（孔隙度）、“fractal dimension”（分形维度）和“sorption”（吸附作用）。这些都与岩石物理性质和气体在多孔介质中的行为有关。其他重要术语如“mississippian barnett shale”（密西西比巴尼特页岩）和“ordos basin”（鄂尔多斯盆地）则指出了研究区域。关键词以不同深浅的红色表示，代表这些研究方向的重要性及研究频率，线条表示关键词之间的关联性，线条越粗意味着两个关键词之间的关系越密切。总体来看，反映了关于页岩气研究中各个关键领域的相互联系，强调理解孔隙结构及其相关属性对于页岩气开采和应用的重要性。

为了从时间维度理清研究重点领域的演化过程，分析了关键词时区分布图，展示学校研究关注点的时间序列，如图 2 所示，该图能够从整体上反映研究路径的变化。图中圆圈大小代表该研究主题在时间段内的重要性和流行度。每个时间段展示的均为该时段内首次涌现的关键词，若这些新关键词与先前时间段中的关键词在同一篇文章中共现，则会通过线条相连，且前期关键词的计数增加 1，其对应的圆圈随之扩大。图 2 展示了 2010—2023 年关键主题之间的关系网络。

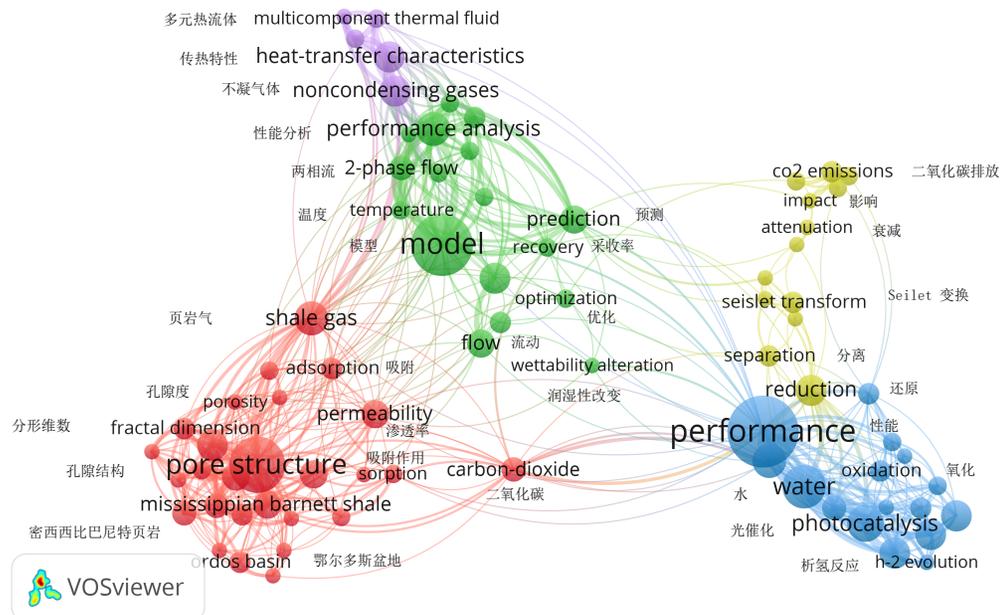


图1 关键词聚类

CiteSpace v. 5.2.R4 (64-bit) Advanced
 April 17, 2025 at 8:16:01 PM CST
 WGS: C:\Users\dmin\Desktop\paper\data
 Timespan: 2010-2023 (Slice Length=1)
 Selection Criteria: q=0.25, LRF=0.0, L/N=10, LBY=5, e=1.0
 Network: N=358, E=1305 (Density=0.0204)
 Largest CCs: 358 (100%)
 Nodes Labeled: 1.0%
 Pruning: None

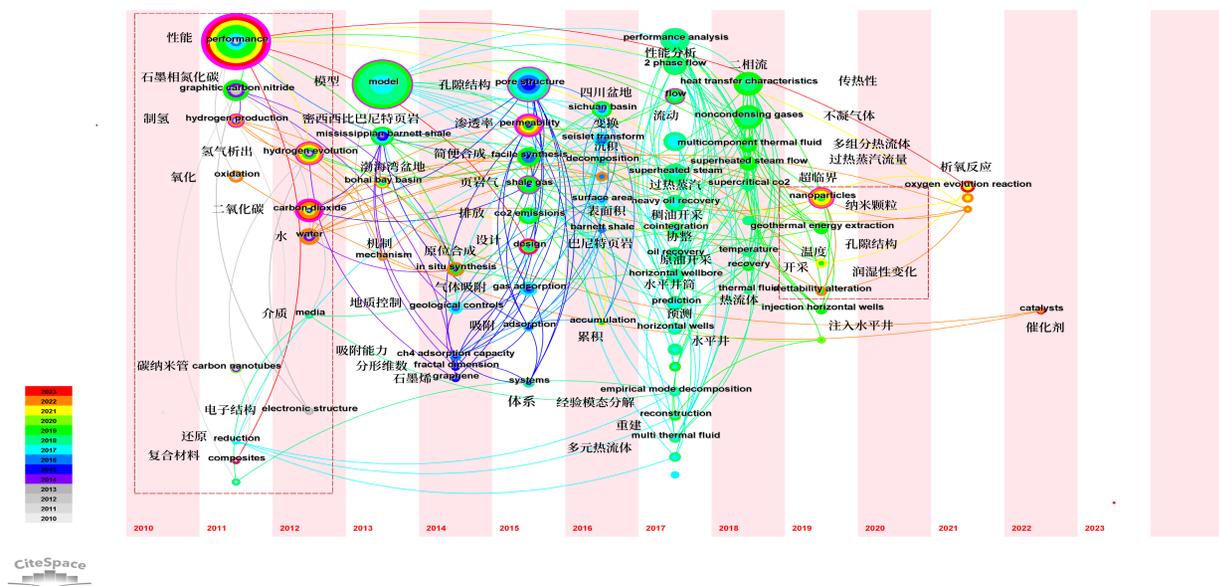


图2 关键词时区图

进一步通过聚类分析，评价主题之间的内在联系和相互作用，这些主题涵盖了从研究理论、实验技术到实际应用等多个方面，为石油能源领域的研究和发展提供了重要的参

考。例如图2中一个2011年出现的突出关键词“performance”（性能），该关键词对应的圆圈，由圆心向外逐步扩大，且由不同颜色组成。对照年份色标可以看出，这个关键词

在 2011 年至 2023 年间受关注程度的变化。可以看到，“performance”在 2019 年（绿色色标）和 2020 年（浅绿色标）期间尤为突出，表明这段时间对其研究的兴趣达到高峰。该词与同年出现的“composites”（复合材料）连线为红色，表明了研究的重点应用和扩展方向。

“performance”在后续年份的圆形面积虽有所波动，但仍然保持在一定的重要位置；另外一个关键词“nanoparticles”（纳米颗粒）在 2019 年（绿色色标）开始出现，并在 2021 年（黄色色标）变得更加突出。这表明纳米颗粒作为一个新兴的研究主题正在逐步获得更多的关注。与“nanoparticles”相关的关键词“wettability alteration”（润湿性变化），揭示了纳米颗粒改变润湿性的可能性。

2 数据来源与指标选择

2.1 指标选择与数据预处理

石油领域论文具有“侧重解决实际生产问题、跨学科融合并涉及新技术新方法新材料、关注国际合作和交流”等特点，针对可以量化的指标对 223 条高被引论文进行数据预处理，最终从论文本身、参考文献、作者、期刊 4 个方面，共选取 21 个定量指标。

2.2 基于随机森林的预测模型构建

传统单一模型往往难以应对复杂数据集中的噪声和多样性，而机器学习方法能够有效地解决这一问题，因此本文采用机器学习方法预测高被引论文的年均被引量。在众多的机器学习方法中，集成学习是一种有效提高模型性能

的方法，其中，随机森林算法（Random Forest）因其简单易用、效果显著而成为常用的方法之一。该方法通过构建多个决策树并将其结果进行综合，从而完成高效的分类或回归任务。由于其在处理高维数据、缺失值和非线性关系方面的优越性能，随机森林已经广泛应用于生物信息学、金融预测、图像识别等领域。

表 1 指标选择

指标	描述
论文	研究领域：wos 学科分类个数
	论文总页数
	发表论文时间
	摘要字数
	关键词数量
	论文早期被引量：论文发表后两年内被引总次数
参考文献	是否有基金资助
	类型：综述性论文、研究型论文
	论文所引用的参考文献数量
	参考文献质量：参考文献期刊平均影响因子
	参考文献的年龄：参考文献的出版年份均值
作者	参考文献的累积被引：截至计算日期参考文献的累积被引
	作者声誉：第一作者 h 指数
	作者合作规模：参与作者数量
	国家合作规模：国家数量
期刊	合作方式：国际合作、国内合作、组织内合作、组织外合作
	自引率：论文作者对论文的被引量占比
	期刊质量：期刊影响因子
	是否开放获取
	发文量：发文当年刊载的论文数
	期刊的语言类型

随机森林算法由 Breiman^[10]于 2001 年提出，其核心思想是通过构建多个决策树并结合其投票或平均结果来降低模型的方差。基于算法的

基本设计原则，通过 Python 语言编写预测模型，流程图如图 3 所示。选择 80%（178 篇）的数

据建立训练集，利用剩余 20%（45 篇）数据进行验证，评价模型准确性。

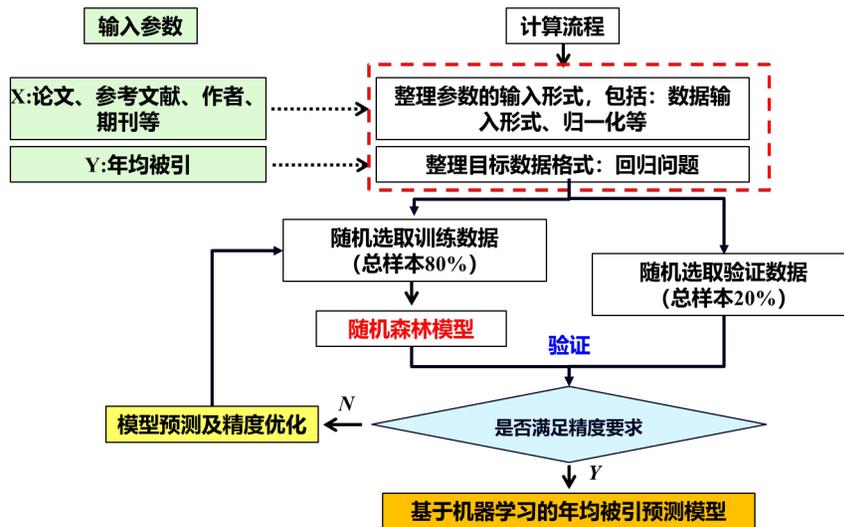


图 3 机器学习预测流程

预测值与实际值的比较如图 4 所示，图中虚线为相对误差 $\pm 15\%$ 的范围，实线为误差 0

的标准线，验证数据处于误差范围内，证明该模型能够准确预测论文的年均被引量。

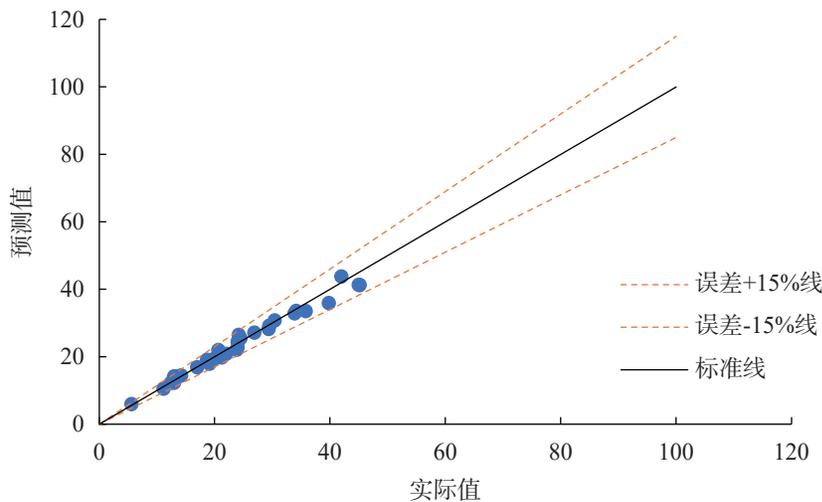


图 4 论文的年均被引量预测结果

3 影响因素研究

3.1 影响因素权重分析

基于随机森林预测结果，对 21 个指标进行

重要性排序，如图 5 所示。在论文方面，论文早期被引量、论文的研究领域比较重要。论文早期被引量反映了论文发表后的影响力和被认可程度。在石油领域，早期被引量反映了论文

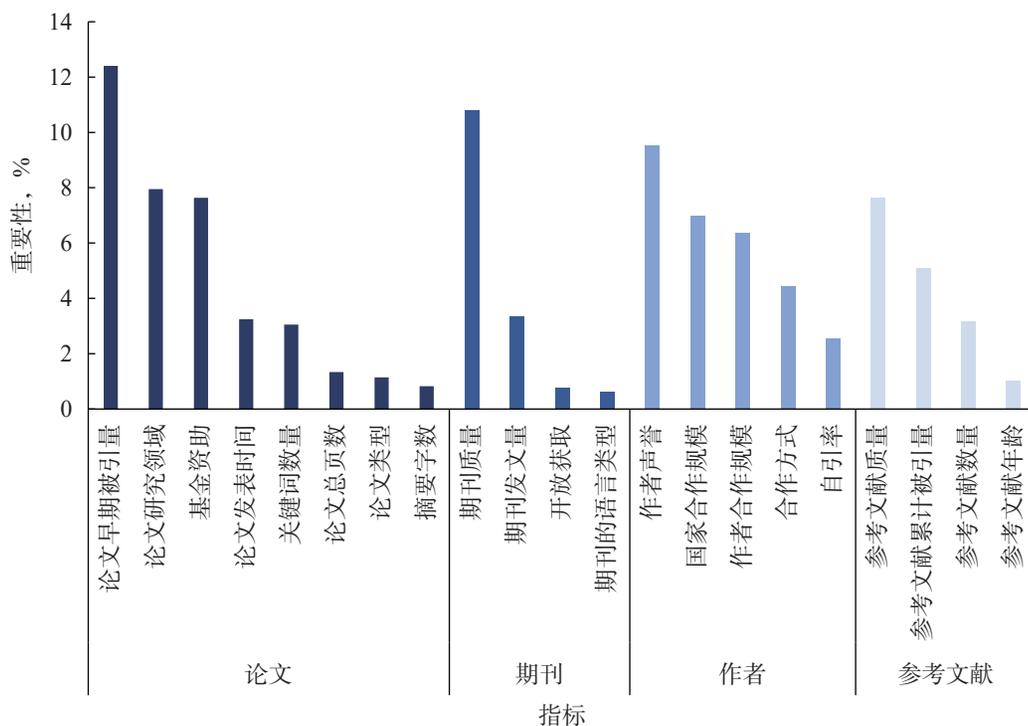


图5 指标重要性

在发表后迅速被学术界和工业界关注的程度。石油领域的研究往往具有较强的应用导向，因此早期被引量高的论文通常涉及前沿技术或解决行业关键问题；石油领域的研究方向，如油气勘探、油田开发、石油化工等，对论文的被引量有显著影响。具有创新性和突破性的研究，如页岩气开发、深海钻井技术更容易受到广泛关注和引用。

在期刊方面，期刊影响因子是最为重要的因素。石油领域的高质量期刊（如《Journal of Petroleum Science and Engineering》）往往能够吸引高水平的论文投稿，并提升论文的学术价值和实践价值；石油领域的开放获取期刊能够加速研究成果的传播，特别是在国际合作中，开放获取有助于提升论文的可见性和引用率。

在作者方面，作者声誉反映了作者在学术

界的地位和影响力，在石油领域，知名学者或研究团队的研究成果更容易被引用；作者合作规模在一定程度上反映了研究团队的实力和合作能力，石油领域的研究往往需要跨学科、跨机构的合作。大规模的合作团队通常能够整合更多资源，提升研究的深度和广度。

在参考文献方面，参考文献平均影响因子、参考文献的累计被引量两个指标需要关注。高质量的参考文献是论文研究深度和广度的体现，也是评估论文学术价值的重要依据。在石油领域，引用高被引文献能够增强论文的学术基础；累计被引量反映了参考文献的影响力和被认可程度，是评估论文研究基础扎实程度的重要指标。

与材料科学、计算机科学等领域相比，石油领域的研究更注重实际应用，因此论文早期被引量和研究领域的重要性更为突出。全球范

围内，高被引论文的影响因素可能更侧重于国际合作和跨学科研究，而石油领域的高被引论文则更依赖于行业内的创新性和实用性。

3.2 影响因素的作用机制

石油类高校高被引论文的影响因素及其作用机制是一个复杂而多维的问题，本文将对随机森林方法挖掘的影响高被引论文的因素进行分析。

(1) 论文的学术水平是被引次数最重要的影响因素，这与内容质量密切相关。一篇高质量的论文通常具有创新性、前沿性和实用性，能够解决该领域内的关键问题或提出新理论和方法。这样的论文更容易被同行认可和引用，从而成为高被引论文。石油领域的高被引论文往往与石油行业的实际需求密切相关。例如，针对特定地质条件下的油气勘探开发技术、提高采收率的方法等，这些研究能够直接解决石油行业面临的实际问题，因此更容易受到关注和引用。

(2) 期刊的学术影响力与论文的被引频次密切相关，具体而言，那些拥有较高影响因子的期刊，通常享有更高的学术地位，并能吸引更庞大的读者基础。在这样的期刊上发表论文，有助于论文迅速进入主流学术视野，提高被引率。例如，《Petroleum Exploration and Development》作为石油工程类期刊中科院分区1区、影响因子为7的高水平期刊，发表在该刊上的论文往往能够获得更高的被引次数。

(3) 作者的学术声誉、研究经验和合作网络等是影响论文质量和被引频次的重要因

素。知名学者或研究团队发表的论文往往更容易受到关注和引用，因为他们的研究成果属于石油领域的技术突破，推动石油行业的发展和进步。

(4) 高被引论文往往引用了大量权威且相关的文献。石油领域的研究往往涉及多个学科的交叉和融合，如地质学、地球物理学、化学、工程学等。跨学科的合作与融合有助于形成综合性的研究成果，提高研究的深度和广度，从而增加论文的被引频次。这些文献不仅为论文提供了坚实的理论基础，还增加了论文的可信度和说服力。并且引用的范围广泛，涵盖了多个研究方向和领域，体现了论文的跨学科影响力，促进了不同领域之间的知识交流和融合。同时，这些论文在引用时往往深入挖掘了相关文献的深层含义和价值，进一步提升了论文的学术价值。

(5) 其他因素。一些其他因素也会影响石油类高校高被引论文的数量，例如，研究主题的前沿性对论文被引次数具有重要影响；选择一个既符合学科发展趋势又填补现有研究空白的课题，能够显著提高论文的吸引力和被关注的可能性；论文的标题、摘要、关键词等元素也可以提高论文的可见度和被检索到的概率，从而增加被引次数。

4 提升石油类高校高被引论文数量的策略

为了增加高被引论文的数量，石油类高校需要注重提升论文内容质量、选择高声誉期刊发表、培养知名学者和研究团队、加强交流与

合作等方面的工作。

（1）提升论文质量与创新性

聚焦当前石油领域的热点和前沿问题进行研究，这些领域往往更容易获得关注和引用。强化开创性研究，提出新的理论、方法或技术，这样的研究成果更容易被同行认可和引用。注重论文的实用性和应用价值，具有实际应用价值的研究成果更容易被工业界和学术界所关注，从而增加被引用的机会。

（2）优化论文发表策略

选择高质量期刊，投稿时应选择在本领域内具有影响力的高质量期刊，这些期刊的读者群更广，引用机会更多。优化论文标题和摘要，一个简洁明了、引人入胜的标题以及清晰、准确的摘要能够吸引读者的注意力，提高论文的可见度和被引率。

（3）注重人才培养及热点跟踪

对研究团队提供经费及政策支持，鼓励学者参与科研项目。并及时跟踪研究领域的最新进展，根据最新进展调整研究方向和策略，保持研究成果的时效性和前沿性。对已发表的研究成果进行持续的关注和更新，可以进一步提高其影响力和被引率。

（4）加强学术交流与合作

积极参加学术会议，通过参加学术会议，可以展示研究成果，与同行进行交流，增加论文的曝光度和引用机会。与其他学者合作撰写论文，合作研究可以汇聚多方智慧和资源，提高研究质量，同时增加论文的引用机会。

石油资源的稀缺性和在全球能源结构中的重要地位，使得石油领域的研究具有更高的关

注度和重要性。结合行业的实际需求，提升石油类高校高被引论文数量，需要从论文质量、发表策略、学术交流与合作、学术声誉等多个方面入手。这些策略的实施将有助于石油类高校在学术界中提高声誉和影响力。为石油领域科研论文的发展提供指导。

参考文献

- [1] 方红玲, 张亚杰, 徐自超. 第一作者和合著者的生产力、影响力与论文被引频次的相关性对比研究[J]. 中国科技期刊研究, 2024, 35(2): 273-279.
- [2] 常宗强, 叶喜艳, 张静辉, 等. 基于被引频次的期刊论文被引质量评估指标构建[J]. 编辑学报, 2023, 35(S1): 238-240.
- [3] 许林玉. 高被引论文核心影响因素判别研究[J]. 信息资源管理学报, 2023, 13(5): 137-148.
- [4] 吴冰, 齐思贤. 集成传统学术评价和 Altmetrics 指标的论文高被引预测研究[J]. 数字图书馆论坛, 2023, 19(9): 30-37.
- [5] 毛璐, 许鑫, 邓璐芾. 基于研究数据评价的引证优化: 高被引数据集特征视角[J]. 情报科学, 2023, 41(2): 126-134, 142.
- [6] 刘红煦, 唐莉. 获评高被引学者会提升学术产出与影响力吗?——来自整体与个体层面的双重验证[J]. 科学学研究, 2021, 39(2): 212-221.
- [7] 马荣康, 李真真. 高被引还是零被引: 基于论文被引的最佳科研合作规模研究——来自 Financial Times TOP 45 商学院期刊的证据[J]. 情报学报, 2020, 39(11): 1182-1190.
- [8] 赵婉忻. 引文视角下国内高被引论文的 Altmetrics 指标相关性研究[J]. 情报理论与实践, 2020, 43(11): 47-53.
- [9] 涂静, 李永周, 张文萍. 国际合作网络结构与高被引论文产出的关系研究[J]. 图书馆杂志, 2019, 38(7): 69-75.
- [10] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.