全国科普日

科普中国
CHINA SCIENCE COMMUNICATION

9月15日 - 9月25日

2024年全国科普日活动

提升全民科学素质
协力建设科技强国

主办单位

中国科协、中央宣传部、中央网信办、教育部、科技部、国家原子能机构、自然资源部、生态环境部、水利部、农业农村部、国家卫生健康委、应急管理部、国务院国资委、中国科学院、中国工程院、国家林草局、全国总工会、共青团中央、全国妇联、中国作协、全国工商联

目次 Contents

情报工程
QINGBAO GONGCHENG

2024 年 第 10 卷 第 5 期

编委会 • EDITORIAL BOARD

doi:10.3772/j.issn.2095-915x.2024.05.001

国内卡脖子技术内涵特征和识别方法研究进展

刘盼盼[1,2,3] 钟永恒[1,2,3] 刘佳[1,2,3] 宋姗姗[1,2,3] 张萌萌[1,2,3]

1. 中国科学院武汉文献情报中心 武汉 430071；
2. 中国科学院大学经济与管理学院信息资源管理系 北京 100190；
3. 科技大数据湖北省重点实验室 武汉 430071

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 在全球科技竞争日益加剧的背景下，卡脖子技术是我国科技发展面临的重要问题。梳理国内卡脖子技术识别研究现状，以期为我国未来的卡脖子技术研究提供参考。[方法 / 过程] 首先对卡脖子技术及其相关概念进行辨析；其次从卡脖子技术识别方法和测度指标方面进行梳理；最后，总结现有研究不足，为未来卡脖子技术研究提出建议。[结果 / 结论] 当前卡脖子技术研究尚处于起步阶段，对卡脖子技术的概念内涵缺乏系统性研究；重点关注关键核心、技术差距等特征，对国家战略、国际关系特征的研究不足；通过文本挖掘、社会网络分析等技术手段识别卡脖子技术的研究方法体系仍不成熟；数据源较为单一；识别粒度较粗。未来，学者需加强对卡脖子技术的技术属性、内涵特征、关键主体、形成原因、内在作用机制等问题的研究，丰富卡脖子技术识别数据源，构建全方位多层次的卡脖子技术评价指标体系，积极探索采用机器学习、文本挖掘方法识别卡脖子技术，动态监测并预测卡脖子技术，更加科学有效地指导我国突破卡脖子技术问题。

关键词：卡脖子技术；卡脖子技术概念；卡脖子技术识别；卡脖子技术测度指标

中图分类号：G35；G250.2

Research Progress on Connotation Characteristics and Identification Methods of Strangle Technology in China

LIU Panpan[1,2,3] ZHONG Yongheng[1,2,3] LIU Jia[1,2,3] SONG Shanshan[1,2,3] ZHANG Mengmeng[1,2,3]

1. National Science Library(Wuhan), Chinese Academy of Sciences, Wuhan 430071, China;
2. Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China;
3. Hubei Key Laboratory of Big Data in Science and Technology, Wuhan 430071, China

基金项目 湖北省技术创新专项软科学研究类重点项目“湖北省重大科技创新平台建设若干重点问题研究”(2022EDA010)；湖北省软科学研究类项目“湖北省根技术识别及其培育发展研究”（2022EDA043）。

作者简介 刘盼盼（1998-），博士研究生，主要研究方向为产业大数据情报研究；钟永恒（1965-），博士，研究员，主要研究方向为产业大数据情报研究，E-mail: zyh@mail.whlib.ac.cn；刘佳（1986-），博士，副研究馆员，主要研究方向为产业大数据情报研究、科技评价；宋姗姗（1996-），博士，主要研究方向为产业智库研究；张萌萌（1997-），博士研究生，主要研究方向为产业情报研究。

引用格式 刘盼盼 , 钟永恒 , 刘佳 , 等 . 国内卡脖子技术内涵特征和识别方法研究进展 [J]. 情报工程 , 2024, 10(5): 3-17.

doi:10.3772/j.issn.2095-915x.2024.05.001

Abstract: [Objective/Significance] Against the backdrop of increasing global technological competition, strangle technology pose significant challenges to China's scientific and technological development. This paper reviews of the current research on strangle technology in China, aiming to provide references for future research in this field. [Methods/Processes] Firstly, this study distinguishes and analyzes the concept of strangle technology and its related concepts. Secondly, it clarifies the measurement indicators and identification methods for strangle technology. Finally, this study summarizes the shortcomings of existing research and proposes recommendations for future studies on strangle technology. [Results/Conclusions] The research finds that the current study on strangle technology is still in its early stage, with a lack of systematic research on the concept and characteristics of strangle technology. The current research focuses on key core and technological gaps, with insufficient research on national strategy and international relations characteristics. The research methods for identifying bottlenecks through text mining, social network analysis, and other methods are still immature. The data source is relatively single, and the recognition granularity is relatively coarse. In the future, scholars need to strengthen research on the technical attributes, characteristics, key actors, formation reasons, and intrinsic mechanisms of strangle technology, enrich the data sources for identifying strangle technology, construct a comprehensive and multi-level evaluation index system for strangle technology, explore the use of machine learning and text mining methods to identify strangle technology, dynamically monitor and predict strangle technology, and provide more scientific and effective guidance for China to overcome strangle technology issues.

Keywords: Strangle Technology; Concept of Strangle Technology; Identification of Strangle Technology; Measurement Indicators of Strangle Technology

引言

中美经贸摩擦爆发以来，我国面临更加复杂严峻的政治经济环境，卡脖子技术成为影响我国经济高质量发展和国家安全的重大隐患。因此，习近平总书记高度关注卡脖子问题，并作出一系列重要论述；2020 年 9 月在科学家座谈会上的讲话指出，“工业方面，一些关键核心技术受制于人，部分关键元器件、零部件、原材料依赖进口”[1]；2021 年 5 月 28 日在两院院士大会上再次强调“弄通卡脖子技术的基础理论和技术原理”[2]。习近平总书记在近年考察调研企业时多次强调自主创新，勉励企业掌握更多关键核心技术，只争朝夕突破卡脖子问题。解决国家关键核心技术卡脖子问题，是我国建设世界科技强国的必由之路，也是在国际产业竞争中实现弯道超越的必然选择。

鉴于卡脖子技术的现实意义，卡脖子技术已成为近年来国内的研究热点，国内学者从不同角度出发对卡脖子技术识别展开研究。曹宇轩等[3]从卡脖子技术研究的提出背景、科学内涵、现实挑战和战略路径进行阐述。然而，现有研究仍缺乏对卡脖子技术研究现状的深入剖析，尤其是对卡脖子技术的概念内涵、测度指标和具体识别方法的梳理总结还不够全面。因此，本文首先梳理卡脖子技术的概念、特征和分类，并对相关概念进行辨析；其次从专家智慧、文献计量、指标评价、文本分析和多方法综合角度归纳卡脖子技术识别方法，对比不同方法间的优劣，并从技术差距、关键核心、国家战略和国际关系角度归纳卡脖子技术识别指标；最后，总结卡脖子技术识别领域的现有研究的不足，并指出未来研究的方向，以期为后续研究提供参考。

本文的数据来源于 CNKI 数据库的北大核

心和 CSSCI 的期刊，检索式为 (SU= 卡脖子) OR (SU=(‘技术制裁’+‘技术管制’+‘技术限制’+‘技术遏制’+‘技术垄断’+‘技术封锁’+‘技术锁定’+‘技术脱钩’+‘技术孤立’+‘技术打压’))，检索时间为 2024 年 1 月 1 日，获取 210 篇文献。经阅读剔除不相关的文献，并检索关键文献的前后向引文，补充遗漏的相关文献，最终筛选获得 50 篇文献。

1 卡脖子技术概念研究

1.1 卡脖子技术的概念

《汉语大词典》中卡脖子有两种解释，分别是关键时刻发生致命性的事情、因某一部位受限制或出问题而影响整体工作 [4]。在技术创新领域，2009 年 4 月 20 日李克强同志视察柳工时指出，技术创新要抓被“卡脖子”的工序，实现关键部位的创新突破，从而提高效益、利润和质量 [5]。2018 年，科技日报梳理列举我国亟待攻克的 35 项“卡脖子”技术。2020 年，时任中国科学院院长白春礼提出要把美国卡脖子清单变成科研清单。

在学术界，国内学者从不同角度对卡脖子技术进行概念界定，如表 1 所示。由于卡脖子问题受政治、经济、技术等多因素耦合影响 [6]，卡脖子技术的概念内涵涉及技术本身、产业发展、国家战略、国际关系等多重维度。综合上述维度，认为卡脖子技术是具有重要战略地位、处于产业链关键核心位置，但与他国存在技术差距的技术，因技术来源较少且缺少替代，技术供给方的垄断使得技术需求方的产业链安全性和稳定性乃至国家安全存在风险。

表 1 卡脖子技术概念

维度	内涵	学者
技术本身	与其他国家存在技术差距，且短期内难以缩小； 因我国自主创新能力薄弱导致某些科技领域内的核心技术主要依赖技术引进，从而引发外部技术封锁风险。	陈劲等 [7]，任文华 [8]，邵颖红 [9]
产业发展	具有产业支撑带动作用，推动技术和产业发展； 处于产业链、供应链等的关键瓶颈位置； 核心环节受制于人影响产业链安全。	赵君彦等 [10]，赵雪峰等 [11]
国家战略	满足国家（地区）战略目标； 具有高技术价值，关键核心技术或高技术产品。	赵君彦等 [10]，李昱璇等 [12]
国际关系	因来源较少且缺少替代而形成技术垄断的技术； 技术供给方的垄断导致跨国、跨链、跨企合作受限，难以实现技术转移； 通过专利壁垒、技术封锁、投资限制、产品进出口限制和市场准入限制等手段引发卡脖子风险。	张婷等 [13]，何婉等 [14]，王敏等 [15]，宋立丰等 [16]，韩震等 [17]

1.2 卡脖子技术的特征分析

表 2 梳理了不同研究中卡脖子技术的特征。从表格中可以清晰地看到卡脖子技术的内涵特征丰富，多个学者强调卡脖子技术的技术差距、关键核心、国家战略、技术垄断等特征。

关于技术差距特征，国内与国外存在技术差距是学者们的共识，但对技术差距的程度意见不一。陈劲等 [7] 认为长期与其他国家存在较大技术差距，且短期内难以缩小是卡脖子技术的特征之一，杨玉良 [18] 认为卡脖子技术是尚未

doi:10.3772/j.issn.2095-915x.2024.05.001

完全掌握的技术。关键核心特征是指卡脖子技术处于产业链、创新链的关键核心位置，技术垄断导致一国在全球开放式创新环境下的产业链与创新链遭受断链，对产业链的安全稳定发展造成威胁。基于国家战略视角，多个学者认为卡脖子技术是满足国家或地区战略目标、国家竞合博弈的重点领域，对国家经济发展和科技安全具有关键核心作用。基于国际关系视角，由于技术来源较少且缺少替代，国际关系恶化导致技术贸易受阻，打乱正常的国际分工体系，形成技术垄断的局面，对外依存度较高的技术便很有可能形成卡脖子技术。

表 2　脖子技术特征

	国家战略	关键核心	技术差距	技术突破	技术垄断	技术依赖	产业需求	技术安全	产业安全	国家安全	技术对接	技术威胁	商品性
张婷等 [13]	√	√	√						√				
赵君彦等 [10]	√		√				√	√			√		
周海球等 [19]		√		√	√		√						
赵雪峰等 [11]		√								√			
汤志伟等 [20]	√	√		√	√	√				√		√	
陈劲等 [7]		√	√		√			√	√			√	
江瑶等 [21]		√	√	√									
俞荣建等 [22]	√	√	√		√					√			
江瑶等 [23]		√		√	√	√							
陈旭等 [24]		√	√	√									
李昱璇等 [25]	√	√	√	√	√					√			
徐霞等 [26]		√	√		√				√				
杨斌 [27]	√				√							√	√
董坤等 [28]			√										

1.3　相关概念辨析

卡脖子技术与关键核心技术、杀手锏技术、前沿技术、新兴技术以及颠覆性技术等不同类型的技术具有不同的产生背景和内涵特征。

卡脖子技术的形成涉及技术先发者和技术后发者两个参与主体。从技术先发者来说，技术先发者常常具有技术优势，随着关键核心技术竞争日益激烈，技术先发者试图通过专利壁垒、出口贸易限制等手段保持或进一步扩大技术差距，维护其技术垄断地位。从技术后发者来说，后发者在某些技术领域不能完全实现自给自足，对先发者产生技术依赖，在技术追赶的过程中先发者的技术锁定也会影响技术扩散和模仿，阻碍后发者实现技术创新。基础原材料、核心零部件、生产装备、核心工艺、技术路线、设计工具、系统软件等环节都有可能出现卡脖子问题 [7]，同时卡脖子技术也并非单纯的技术领域问题，而是涉及基础科学、技术科学和工程科学等多方面问题 [29]，如图 1 所示。

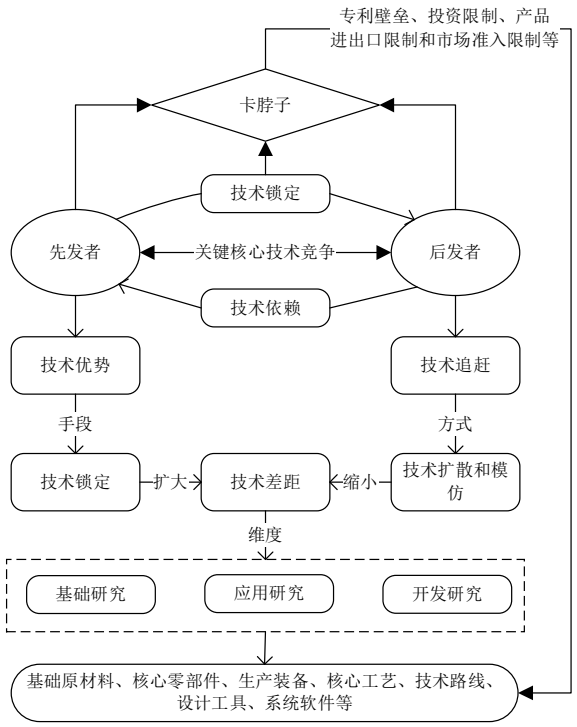


图 1　卡脖子技术形成

前沿技术是一定时间内技术关注度快速提升，处于技术发展的尖端，符合产业发展需求的技术[30]。新兴技术是开始引起关注并具有潜力创造一个新行业或改变某个老行业，但仍处于早期阶段的技术，其发展可能需要进一步的研究和验证[31]。颠覆性技术是对已有的技术进行革命性变革，能够改变市场规则从而引起竞争范式转变[32]。以上三种技术侧重于当下和未来，抢占科技制高点，从而实现先发优势，引领未来发展。卡脖子技术往往是前沿技术发展的关键瓶颈，没有突破这些卡脖子技术，前沿技术可能难以实现质的飞跃；前沿技术的发展又可能带动卡脖子技术的突破，为解决这些瓶颈问题提供新思路和新方法。新兴技术的出现可能会对现有的卡脖子技术构成挑战，推动原有技术体系的更新换代；卡脖子技术的突破往往能够为新兴技术提供更加坚实的基础和发展平台。颠覆性技术具有改变游戏规则的潜力，它们可能会绕过现有的卡脖子技术，通过全新的技术路径实现目标；卡脖子技术的存在也可能激发对颠覆性技术的追求，以摆脱对现有技术路径的依赖。卡脖子技术与前沿技术、新兴技术、颠覆性技术之间的关系是动态的、互相影响的，需要综合考量评估，灵活调整战略规划，以实现科技和产业的持续健康发展。

现有研究大多认同卡脖子技术本身是关键核心技术，符合关键核心技术的一般性特征[9, 20]。卡脖子技术和关键核心技术的区别有以下两方面：从技术角度来看，较关键核心技术而言，卡脖子技术缺少替代方案，研发与应用周期更长，并且短期内难以突破；从产业角度来看，卡脖子技术更加关注产业链、供应链的瓶颈环节，体现产业安全性特征[7]。因此，卡脖子技术较容易被技术供给方压制，更具技术威胁性、紧迫性和垄断性。

卡脖子技术与杀手锏技术两者之间既存在关联也有区别，如图 2 所示。二者都具有战略性、全局性和基础性[34]，核心区别在于我国是否具有技术优势。卡脖子技术是我国的技术短板，杀手锏技术则侧重拉长长板，巩固提升优势产业的国际领先地位，持续增强我国全产业链优势，提升产业质量，拉紧国际产业链对我国的依存关系，形成对国外技术垄断的强有力反制和威慑[35]。

外文文献中的“Bottleneck Technology”与卡脖子技术的概念也不同。从工程角度，“Bottleneck Technology”是指为限制一个技术系统的至少一项功能的技术；从市场术语则是指使系统过于昂贵而无法在特定细分市场销售

doi:10.3772/j.issn.2095-915x.2024.05.001

的技术[36]。“Bottleneck Technology”关注某一工程或市场中的技术瓶颈，而卡脖子技术则涵盖国家之间的科技竞争。

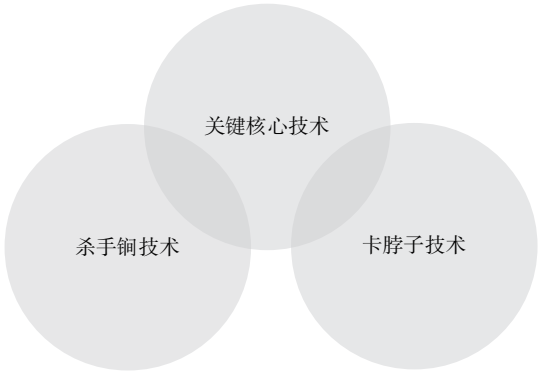


图 2 卡脖子技术与关键核心技术、杀手锏技术之间的关系

1.4 卡脖子技术分类

关于卡脖子技术分类，学者们主要从紧迫程度、依赖程度和产业类型等方面进行讨论，如图 3 所示。从紧迫程度角度，中国科学院原院长白春礼院士根据技术封锁的实际情况，将卡脖子技术划分为当前已受限亟待短期内攻克以及关系未来长远布局的两类技术。因此，解决卡脖子技术问题不仅要补短板，优先解决最紧急、最紧迫的问题，同时也要筑长板，瞄准未来科技和产业发展的制高点，甚至在“无人区”领域开展前瞻性的规划和布局，为未来中国的持续发展提供可靠的科技支撑[37]。

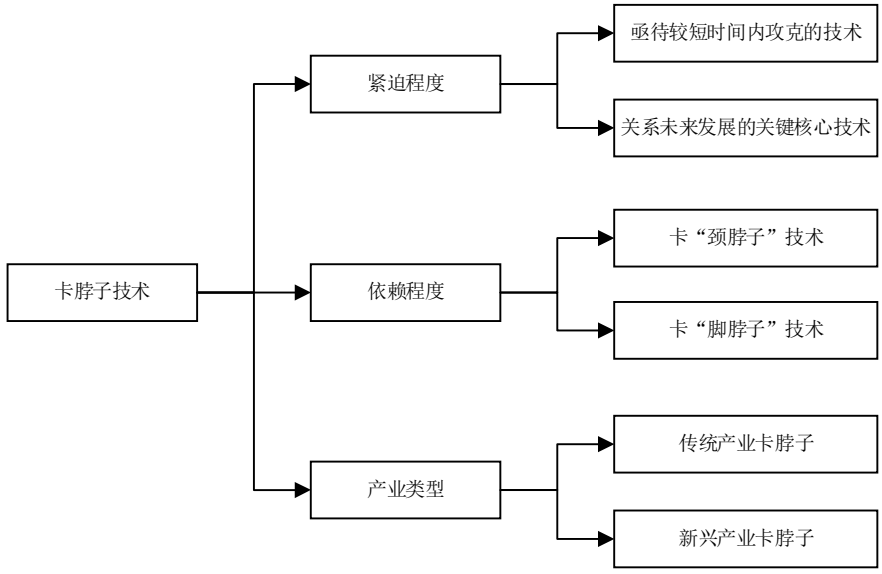


图 3 卡脖子技术分类

从依赖程度角度，卡脖子技术大致可分为卡“颈脖子”和“脚脖子”两类[38]。卡“颈脖子”技术是指必须使用国外技术，没有替代，且短时间内难以研发攻克，技术断供会导致依赖这类技术的产业链上下游企业停产。卡“脚脖子”技术是指国外技术可以提升产品性能，没有国外技术也能找到性能较差的技术替代，此类技术相比卡“颈脖子”技术来说，技术断供带来的损失和影响相对较小。因此，亟需动态跟踪和梳理当前对我国产业发展具有较大影响的卡脖子技术，科学界定攻克卡脖子技术的优先级，从而合理配置科技资源，分类探索和支持卡脖子技术攻关。

从产业类型角度，卡脖子技术分为传统产

业卡脖子和新兴产业卡脖子。在制造业等传统产业领域，全球技术链和产业链的分工基本稳定，发达国家与后发国家在知识积累基础、专业人才和产业转化等创新体系存在巨大差距。而在信息技术、生物医药、新能源等新兴产业领域，全球技术链和产业链仍在形成过程中，中国在部分技术领域取得了重大进展，但基础研究与领先国家仍有一定的差距。由于各产业领域的产业技术链发展程度不同，其技术创新突破的“机会窗口”和“产业化瓶颈”也存在显著差异[15]。此外，美国 2018 年由特朗普签署的《出口管制改革法》[39] 新增 14 项“新兴和基础技术”，进一步扩大技术出口管制范围。因此，需抓住新一轮科技革命和产业变革的机遇，加速推进我国基础技术和新兴技术发展，并根据不同产业领域的创新体系，选择合适的突破窗口，从而带动战略性新兴产业的发展。

2 卡脖子技术识别方法研究

在理解卡脖子技术的概念内涵之后，如何识别卡脖子技术成为近年的研究热点。通过梳理相关文献发现，现有研究主要从技术差距、关键核心、国家战略、国际关系等方面构建卡脖子技术识别指标体系，采用基于专家智慧、文献计量、指标评价、文本分析以及多种方法相结合的方法识别卡脖子技术。

2.1 基于专家智慧的识别方法

基于专家智慧的方法主要依赖专家的知识、经验和观点进行卡脖子技术识别，主要包括问卷调查、案例调研、德尔菲法、层次分析法等。

汤志伟等 [20] 采用问卷调查方式收集电子信息产业的关键核心技术，并通过垄断程度、攻克难度、技术价值筛选卡脖子技术。在此基础上，李昱璇等 [25] 采用“垄断—难度—价值—相对优势”四步走的筛选方式对某省电子信息产业卡脖子技术问题展开具体分析。杨斌等 [27] 从国家战略、产业安全、关键核心技术和技术差距角度归纳卡脖子技术的评估指标体系，并采用案例调研方法分析先进电子材料领域的卡脖子技术典型案例。郑国雄等 [40] 构建技术重要性、先进性、垄断性、可获得性和社会经济价值 5 个评价指标，结合德尔菲法和层次分析法筛选生物医药领域的卡脖子技术。

基于专家智慧的卡脖子技术识别方法，通过问卷调查、德尔菲法、层次分析法等将专家意见融入卡脖子技术识别过程中，能够充分利用和融合专家知识，结果较为可靠，但存在专家判断主观性较强的局限性，且耗费大量人力，难以实现大规模自动化识别卡脖子技术。

2.2 基于文献计量的识别方法

论文和专利文献是科学技术成果的重要载体，一定程度上反映技术创新活动的发展动向。因此，部分学者基于专利或论文数据，采用文献计量方法，通过统计专利数量或论文数量测度技术差距或技术优势，从而识别卡脖子技术。

唐恒等 [41] 通过对比分析中国和其他技术垄断国家在 IPC 小类集聚程度上的差异，挖掘我国卡脖子技术领域的短板，再聚焦这些技术短板领域中的国内外“领军机构”，构建技术—功效—机构三维分析模型，分析识别卡脖子技术卡点。董坤等 [28] 结合专利数量和专利质量计

doi:10.3772/j.issn.2095-915x.2024.05.001

算技术优势指数，识别省域视角下的产业潜在卡脖子技术，并解析潜在卡脖子的技术分支与关键技术点，明确其核心竞争主体与潜在竞争主体，预判其技术开发难度与可行性。程郁等[42]通过统计对比中国和其他国家的论文发表数量、专利申请数量，高被引论文数量、高价值核心专利数量、有效专利数量，发现在生物育种、实用功能基因、智慧育种等方面仍存在技术差距。

基于文献计量的卡脖子技术识别方法，常通过统计专利或论文的数量等外部特征测度技术差距或技术优势。该方法抓住卡脖子技术的关键特征，计算相对简单，具有较强的可操作性，弥补专家判断主观性较强的缺陷，但仅从专利数量信息难以衡量卡脖子技术，对卡脖子技术的关键核心、战略性、产业安全性等特征考虑不足。

2.3 基于指标评价的识别方法

部分研究采用多指标评价方法，构建多阶段甄选模型，在识别关键核心技术的基础上，进一步延伸识别卡脖子技术。目前卡脖子技术的定义尚未统一，学者们从不同定义下构建的指标也各不相同。本文根据卡脖子技术的概念内涵，从技术差距、产业发展、国家战略、国际关系等重要特征入手，归纳总结卡脖子技术测度相关指标。

2.3.1 技术差距维度

技术差距特征是卡脖子技术的重要特征之一。现有研究主要利用专利数据，通过计算专利数量或质量测度技术差距。唐恒等[41]通过对比分析中国和其他技术垄断国家在 IPC 小类上的专利数量差异，挖掘我国卡脖子技术领域的短板，但该研究仅从专利总量方面对比分析计算差距。除数量之外，董坤等[28]提出结合专利数量和专利质量构建技术优势指数，进而计算省域之间在特定技术领域的技术差距。后来，陈旭等[24, 43]、江瑶等[21]借鉴董坤的做法构建技术差距指数模型。徐霞等[26]也摒弃简单直接计算专利总量的方式，从市场占有率角度测度专利数量的国际水平，采用 IncoPat 专利分析模块中的合享价值度作为专利质量指标。范书琴等[44]利用关键核心特征指标计算结果，构建关键核心专利技术标准化质量指数作为技术差距特征指标。

除专利数据之外，虽然有学者还考虑了利用论文、实验室、商业等数据，但指标统计难度较大，停留在指标设计或定性评价阶段。例如，杨斌等[27]从基础研究差距、应用研究差距、实验开发差距、商业化差距设计 4 个一级指标。赵君彦等[10]从技术研究、技术转化、技术应用三方面衡量奶业卡脖子技术差距，邀请专家对各项指标进行评价。

总体来说，技术差距特征的指标体系围绕专利数量和质量展开，通过专利总量、计算均值等不同的计算方式构建不同的技术差距指数模型。但也存在一些不足之处：专利并不是唯一的评估技术差距的标准，技术差距还涉及研发投入、市场占有率、人才储备等其他因素的影响；有些公司或组织可能会通过大量申请专利来掩盖实际的技术差距，因此仅仅依靠专利数量和质量作为评估标准可能会产生误导；专利评估通常需要考虑多个因素，如技术领域、专利类型、法律和政策环境等，这些因素对专利评估结果的影响难以完全量化；专利申请和授权往往需要较长的时间，因此专利数量和质量评估的结果可能无法反映当前的技术差距。

2.3.2 关键核心维度

围绕产业发展特征，部分研究从理论层面构建卡脖子技术识别指标体系，为后续的卡脖子技术识别研究提供了研究基础，但未对指标进行量化。例如，杨斌等[27]从专利覆盖程度、技术可突破性、核心专利保护范围、生产装备的自主率、技术复杂度设计5个一级指标。

部分学者基于专利数据对卡脖子技术的关键核心特征进行量化测度，如表3所示。江瑶等[21, 23]从前沿影响、复杂创新、市场应用和战略辐射等维度设计关键核心特征指标。类似地，陈旭等[24, 43]从前沿技术、经济价值、战略安全构建衡量关键核心特征的指标体系。范书琴等[44]认为关键核心专利具有显著的技术价值、法律价值和商业价值，因此从技术属性、法律属性和商业属性构建关键核心特征识别指标体系。

表3 卡脖子技术的关键核心特征指标

维度	指标	学者
专利技术属性	专利影响力、先验知识量、科学关联度、技术宽度	范书琴等[44]，江瑶等[21,23]，陈旭等[24,43]
专利法律属性	保护范围、核心保护范围、保护力度、技术信息披露度、权利稳定性	范书琴等[44]，江瑶等[21,23]，陈旭等[24,43]
专利商业属性	剩余有效期、收益潜力、商业范围	范书琴等[44]

虽然卡脖子技术的关键核心特征指标体系愈加丰富，但其本质还是主要从专利的外部特征测度关键核心特征。但卡脖子技术不同于关键核心技术，卡脖子技术注重在产业发展中的作用和安全，其关键核心特征更加强调在产业链的重要地位，目前少有学者从产业链的角度衡量卡脖子技术的关键核心特征。

2.3.3 国家战略维度

关于国家战略维度，杨斌等[27]试图通过技术嵌入价值链的程度衡量其在全球价值链地位。张婷等[13]结合生物医药领域的特点，从是否为满足人群健康需求的技术、是否为国家社会经济发展的关键技术两个维度设置国家战略性指标，并细分为技术主要适应证的患病率、诊疗指南中的推荐程度、技术在国家级规划中被提及次数、是否符合国家政策演变趋势、国家科研项目资助数量或金额5个二级指标。赵君彦等[10]通过关键领域战略目标要求、国家技术发展战略、社会经济需求三个方面设置战略布局指标，邀请专家对各项指标进行评价。

目前从国家战略角度对卡脖子技术进行量化测度的研究较少，原因可能是国家战略通常涉及复杂的政治、经济、军事、社会等多个领域，而这些领域的指标往往难以直接量化和比较。此外，国家战略往往具有长远的目标和战略性的考量，对其影响的评估和量化更加困难。此外，部分学者在关键核心特征指标体系中也体现出国家战略特征，导致两者之间的边界不够清晰，这也反映了卡脖子技术的复杂性和多维度性。

2.3.4 国际关系维度

关于国际关系维度，杨斌等[27]设计核心零部件、基础材料和网络技术的对外依赖程度，以及技术标准参与水平和跨国技术转移难度5个指标测度技术自主可控程度，从而衡量产业安全性。周海球等[19]、汤志伟等[20]、郑国雄等[40]、通过问卷调查法或德尔菲法测度技术垄断性。范书琴等[44]采用国别集中度计算在特定技术领域中某一国家的技术垄断程度，国别集中度采用赫芬达尔指数HHI表示。江瑶等[23]以竞争

doi:10.3772/j.issn.2095-915x.2024.05.001

主体之间的专利数量差距来反映垄断依赖性。

目前从国际关系维度构建卡脖子技术识别指标体系的研究也较少，主要通过专家咨询或采用专利数量、质量差距来代替技术垄断性。此外，少有研究在国家关系指标方面区分国家之间的对抗竞争关系，对技术是否受国际环境影响显著的考虑不足，例如贸易政策、经济制裁、政治风险、知识产权保护、技术转移等方面。

综上，基于指标评价的卡脖子技术识别方法是目前常见的方法，根据卡脖子技术的概念内涵设计多维指标，设计指标判断标准，逐步收缩甄选卡脖子技术，最终将技术领域划分为非卡脖子技术、轻微卡脖子技术、一般卡脖子技术、严重卡脖子技术。该方法综合关键核心、战略性、垄断性、价值等特征识别卡脖子技术，但不同视角和定义下构建的指标各不相同，尚未形成通用的卡脖子技术识别指标体系。现有卡脖子技术识别指标重点关注关键核心和技术差距等方面，对国家战略、国际关系特征的研究不足。未来有待进一步从技术创新水平、产业竞争力的角度丰富卡脖子技术的技术差距特征，深入理解技术创新、市场需求等因素对技术差距的影响。同时，需要从产业链角度探索卡脖子技术的关键核心特征，了解在整个产业链中哪些环节具有关键性，从而更好地把握技术发展趋势和瓶颈。此外，将国家战略、国际关系等因素纳入评估指标体系是非常有必要的。卡脖子技术往往与国家战略、国际关系等紧密相关，这些因素对卡脖子技术的发展和应用产生重要影响。因此，未来需要综合考虑各种因素，建立全面的量化评估体系，以更好地评估卡脖子技术。

2.4 基于文本分析的识别方法

部分学者基于美国出口管制文本和专利数据，采用自然语言处理、机器学习、共词分析等方法发现卡脖子技术。

周磊等 [45] 首先计算实体清单企业的美国专利族与商品管制清单的文本相似性筛选卡脖子技术专利，再利用机器学习算法挖掘卡脖子技术属性，最后采用 LDA 建模提炼卡脖子技术主题及其演化规律。吕璐成等 [46] 提出基于 TF-IDF 和 Word2Vec 的商品管制清单数据与专利数据自动映射方法和效果评价指标，并进一步分析中美技术差距。陆天驰等 [47] 通过共词分析等方法对美国商品管制清单进行分析，探索人工智能领域中美国的管制重点。赵雪峰等 [11] 构建多分类轮询的高质量卡脖子专利识别模型 LSTM-Seq-BERT，首先以 LSTM 为基础识别核心技术专利，其后组合 Word2Vec、BERT 及全连接神经网络提取申请文件的技术特征，最后利用 Softmax 函数激活技术特征，识别出高质量卡脖子技术专利。

基于文本分析的卡脖子技术识别方法，通过分析美国政府文件例如实体清单、商品管制清单或与专利结合发现卡脖子技术。该方法深入政策文本和专利数据，有利于把握美国政府对华管制重点和倾向，但除了技术因素，出口管制清单还受政治、经济等其他因素影响。

2.5 多方法综合的识别方法

除了上述四种方法之外，学者们进一步拓展研究思路，采用机器学习、社会网络分析、指标评价、专家评价等多种方法结合的方式识别卡脖子技术，如表 4 所示。

表 4　基于多种方法综合的卡脖子技术识别方法

作者	研究视角	数据	方法	指标体系	领域
陈旭 [24]	技术本身、国家战略	专利	指标评价、专家研讨	关键核心技术、潜在卡脖子问题、突破机会	AI 芯片
周海球等 [19]	技术本身、国家战略、产业发展	技术方向	专家评价、聚类分析、层次分析法	技术战略价值高、对方垄断性强、攻克难度大、产业需求旺盛	不限
徐霞等 [26]	技术本身	专利	社会网络分析、指标分析	关键核心技术（共现网络 K- 核与中心性分析）、专利数量（市场占有率）、专利质量（合享价值度）	信息网络技术
范书琴等 [44]	技术本身、国际关系	专利	模糊层次分析法、熵权法、AHPSortII 分类方法、机器学习	关键核心专利（专利技术属性、法律属性、商业属性）、国别集中度（赫芬达尔指数 HHI）、差距程度（关键核心专利技术标准化质量指数）	锂电池隔膜技术
张治河等 [48]	技术本身、国家战略、国际关系	/	以德尔菲调查法为基础、融合多种定量与定性方法	技术重要性、国际竞争态势、发展基础、制约因素、预期时间、政策手段等	/

周海球等 [19] 在采用专家评价法对卡脖子技术进行评分的基础上，结合马氏距离聚类分析、层次分析法对备选技术进行分类判别。徐霞等 [26] 基于专利数据，综合采用社会网络分析方法和指标评价方法筛选卡脖子技术和杀手锏技术。范书琴等 [44] 首先结合模糊层次分析法和熵权法确定专利评价指标权重，然后采用 AHPSortII 和机器学习方法识别关键核心专利，最后通过国别集中度和差距程度分别确定国外技术的垄断程度和锁定风险等级。张治河等 [48] 设计定量与定性相结合的卡脖子技术方法，融合德尔菲法、文献计量分析、自然语言处理、层次分析等多种方法，甄选政府层面和企业层面的卡脖子技术。

近年来，越来越多的学者采用综合方法识别卡脖子技术，不断改进卡脖子技术识别效果，呈现出从以专家为主的定性研究方法转向数据驱动的定量研究方法和定性定量方法相结合的研究趋势。但同时计算方法复杂，且缺乏统一的理论根基支撑。

2.6　卡脖子技术识别方法述评

近年来，国内关于卡脖子技术的研究热度不断上升，涌现一批学术成果。通过表 5 对上文介绍的卡脖子技术识别方法进行汇总展示，以综合分析其研究思路、具体方法、研究视角和数据等。

上述每种卡脖子技术识别方法都有其独特的优势和局限性，并适用于不同的研究范畴，如表 6 所示。研究者需要根据具体的研究目标和技术领域的特点，选择最合适的识别方法或综合使用多种方法，以获得更全面和准确的识别结果。

通过对卡脖子技术的概念、特征以及主要识别方法进行归纳总结，现有研究仍存在不足之处，具体如下：

（1）在概念内涵方面，现有研究主要从技术本身、产业安全、战略地位、国际关系等角度定义卡脖子技术，尚未形成统一的概念。同时，对卡脖子技术的技术属性、内涵特征、关键主体、形成原因、内在作用机制等缺乏系统性研究，

doi:10.3772/j.issn.2095-915x.2024.05.001

表 5　不同卡脖子技术识别方法对比

识别方法	研究思路	具体方法	研究视角	数据
专家智慧	依赖专家的知识、经验和观点进行卡脖子技术识别	问卷调查、案例调研、德尔菲法、层次分析	技术本身、产业发展、国家战略	专家意见
文献计量	通过统计专利或论文数量测度技术差距或技术优势，从而识别卡脖子技术	文献计量、对比分析	技术本身	专利或论文
指标评价	构建多阶段甄选模型，在识别关键核心技术的基础上，进一步延伸识别卡脖子技术	指标评价	技术本身、国家战略	专利
文本分析	通过美国政府发布的实体清单、商品管制清单等文件发现卡脖子技术	自然语言处理、机器学习、共词分析	技术本身、国际关系	美国出口管制文本、专利
综合	采用多种方法结合的方式识别卡脖子技术	机器学习、社会网络分析、指标评价、专家评价	以技术本身、国家战略为主	专利

表 6　卡脖子技术识别方法优劣势

识别方法	优势	劣势	适用范畴
专家智慧	能够利用专家丰富的经验和深入的理解，对技术的重要性和潜在的瓶颈进行评估	依赖于专家的主观判断，可能受到个人经验和偏见的影响	适用于技术前景不明确，需要专业判断的领域
文献计量	计算相对简单，具有较强的可操作性，弥补专家判断主观性较强的缺陷	文献计量指标可能受到多种因素影响	适用于已有大量研究基础，且研究活动较为活跃的技术领域
指标评价	通过设定一系列评价指标，可以系统地评估技术的成熟度、影响力和潜在风险	指标的选取可能存在主观性，且不同技术领域适用的指标可能不同	适用于需要综合多方面因素进行评价的领域
文本分析	深入政策文本和专利数据，利用自然语言处理技术可以分析大量文本数据	出口管制清单受政治、经济等其他因素影响；文本分析的准确性受到算法和数据质量的限制	适用于文本数据丰富，需要快速筛选和识别信息的场景

理论基础薄弱。由于卡脖子技术尚未形成统一的定义，缺乏理论根基支撑，学者对卡脖子技术的内涵和特征的理解也各有不同，因此卡脖子技术识别的方法流程、指标设计、验证标准均存在差异。

（2）在识别指标方面，随着卡脖子技术研究的不断深入，指标体系也更加丰富。但存在以下问题：一是重点关注关键核心、技术差距等特征，对国家战略、国际关系特征的研究不足；二是卡脖子特征与测量指标之间缺乏联系，指标选取的科学性和合理性存在质疑，例如现有研究多选取专利数量和质量指标测度技术差距或垄断，但单一的专利外部特征指标具有片面性，仅仅依据该指标不足以衡量判断某项技术是否是真正的卡脖子技术；三是指标权重确定与阈值选取等没有统一的标准；四是部分指标难以量化等。

（3）在识别方法方面，现有研究常采用专家智慧、文献计量、指标评价、文本分析、社会网络分析等其中一种或综合多种方法识别卡脖子技术。卡脖子技术识别方法从基于专家智慧的定性研究方法，到基于数据驱动的定量研究方法以及两者相结合的方法。但仍存在以下问题：一是多从技术视角识别卡脖子技术，缺乏从产业发展、国家战略以及国际关系视角进行研究，无法全面评估卡脖子技术；二是通过自然语言处理、机器学习、文本挖掘、社会网络分析等新技术手段识别卡脖子技术的研究方

法体系仍不成熟；三是卡脖子技术识别研究所采用的数据源较为单一，仍以专家意见和专利数据为主；四是部分研究停留在得到卡脖子技术清单阶段，存在识别粒度较粗的问题，仍然需要进一步细化卡脖子技术的具体卡点、技术差距、技术分类等细节问题。

3 总结与展望

卡脖子技术作为推动国家科技进步、保障国家安全、促进产业升级的关键因素，其准确而有效的识别对于中国政府、管理机构以及研发人员而言，具有至关重要的意义。本文首先对卡脖子技术的概念、特征进行归纳辨析；其次系统梳理现有的卡脖子技术识别方法，将其总结为五类：基于专家智慧的方法、基于文献计量的方法、基于指标评价的方法、基于文本分析的方法以及多种方法结合的方法；最后对比各种方法的研究思路、优劣势、适用范畴等，提出现有研究存在的主要问题。

针对目前卡脖子技术研究领域的不足，未来可以在以下方面加强研究：

（1）加强卡脖子技术的概念和理论研究。目前学术界对卡脖子技术的概念、特征以及破解路径等理论问题进行了积极探索，但还未达成共识。未来应继续加强对卡脖子技术的技术属性、内涵特征、关键主体、形成原因、内在作用机制等问题的研究，为卡脖子技术识别和突破研究打下坚实的理论基础。

（2）构建全方位多层次的卡脖子技术评价指标体系。未来有待进一步从技术创新水平、产业链角度探索卡脖子技术的技术差距特征和关键核心特征的测度，深入理解技术创新、市场需求等因素对卡脖子技术的影响。此外，将国家战略、国际关系等因素纳入评估指标体系，进一步完善卡脖子技术的评估指标体系。

（3）构建动态精准的卡脖子技术识别方法。从技术本身、产业发展、国家战略以及国际关系多个视角全面评估卡脖子技术；积极探索机器学习、文本挖掘、社会网络分析等方法在卡脖子技术识别研究中的应用；拓展卡脖子技术识别的数据源；细粒度识别卡脖子技术，并分析技术卡点、技术差距和技术风险等方面；从紧迫程度、产业类型等角度设计更为精细的卡脖子技术识别方法；持续监测卡脖子技术领域相关动态，不仅识别当前的卡脖子技术，更要预测未来可能的卡脖子技术。

参 考 文 献

[1] 人民网 . 习近平主持召开科学家座谈会并发表重要讲话 [EB/OL]. (2020-09-11)[2022-12-28]. http://politics.people.com.cn/n1/2020/0911/c1024-31858723.html.

[2] 中国政府网 . 习近平：在中国科学院第二十次院士大会、中国工程院第十五次院士大会、中国科协第十次全国代表大会上的讲话 [EB/OL]. (2021-05-28)[2022-12-28]. http://www.gov.cn/xinwen/2021-05/28/content_5613746.htm.

[3] 曹宇轩 , 粟锋 . 中国“卡脖子”技术研究述评 [J]. 科技和产业 , 2022, 22(11): 38-44.

[4] 汉语大词典 & 康熙字典 .“卡脖子”词条 [EB/OL]. [2023-01-04]. https://hd.cnki.net/kxhd/Search/Result.

[5] 张治河 , 高中一 , 檀润华 , 等 . 突破“卡脖子”技术的思维模式 —— 基于 TRIZ 的设计 [J]. 科研管理 , 2022, 43(12): 54-68.

[6] 曹琨 , 吴新年 , 白光祖 , 等 . 基于专利文献的“卡脖子”技术识别研究 —— 以数控机床领域为例 [J]. 图书情报工作 , 2023, 67(19): 80-91.

doi:10.3772/j.issn.2095-915x.2024.05.001

[7] 陈劲 , 阳镇 , 朱子钦 . “ 十四五 ” 时期 “ 卡脖子 ” 技术的破解：识别框架、战略转向与突破路径 [J]. 改革 , 2020(12): 5-15.

[8] 任文华 . 钱学森技术科学观视域下关键核心技术 “ 卡脖子 ” 问题研究 [J]. 科学管理研究 , 2021, 39(3): 33-38.

[9] 邵颖红 , 周恺伦 , 程与豪 . 政府补助激励 “ 卡脖子 ” 技术企业创新中内循环能否提供助力 —— 以半导体及芯片行业为例 [J/OL]. 科技进步与对 策 : 1-9[2024-01-16].http://kns.cnki.net/kcms/detail/42.1224.G3.20230117.1553.003.html.

[10] 赵君彦 , 赵婧姝 . 奶业 “ 卡脖子 ” 关键技术甄选机制及培育路径 —— 基于产业链与创新链协同视角 [J]. 中国畜牧杂志 , 2023, 59(4): 323-328.

[11] 赵雪峰 , 吴德林 , 吴伟伟 , 等 . 基于深度学习与多分类轮询机制的高质量 “ 卡脖子 ” 技术专利识别模型 —— 以专利申请文件为研究主体 [J]. 数据分析与知识发现 , 2023, 7(8): 30-45.

[12] 李昱璇 , 方卫华 . “ 卡脖子 ” 技术概念辨析 —— 内生性矛盾、国家主体与外部限制的共同建构 [J]. 科学学研究 , 2024, 42(1): 31-37, 135.

[13] 张婷 , 陈娟 , 徐东紫 , 等 . 生物医药领域 “ 卡脖子 ” 技术的概念内涵辨析及评判原则思考 [J]. 中国新药杂志 , 2023, 32(16): 1615-1621.

[14] 何婉 , 樊斌 . 中国奶牛养殖业种源 “ 卡脖子 ” 问题识别、成因与破解路径 [J]. 黑龙江畜牧兽医 , 2023(12): 7-13.

[15] 王敏 , 银路 . 突破关键核心技术 “ 卡脖子 ” 困境的路径研究 [J]. 清华管理评论 , 2022(5): 45-50.

[16] 宋立丰 , 区钰贤 , 王静 , 等 . 基于重大科技工程的 “ 卡脖子 ” 技术突破机制研究 [J]. 科学学研究 , 2022, 40(11): 1991-2000.

[17] 韩震 , 赵宇恒 , 赵莉 , 等 . “ 卡脖子 ” 技术形成路径与破解策略选择 —— 基于芯片产业的案例研究 [J]. 科技进步与对策 , 2023, 40(24): 82-91.

[18] 杨玉良 . “ 卡脖子 ” 问题刍议 [J]. 科学与社会 , 2020, 10(4): 1-4.

[19] 周海球 , 黄晓林 , 李维思 , 等 . 基于 MD-AHP 的关键核心技术攻关任务甄选方法研究 [J]. 情报杂志 , 2023, 42(9): 149-154.

[20] 汤志伟 , 李昱璇 , 张龙鹏 . 中美贸易摩擦背景下 “ 卡脖子 ” 技术识别方法与突破路径 —— 以电子信息产业为例 [J]. 科技进步与对策 , 2021, 38(1): 1-9.

[21] 江瑶 , 陈旭 , 张凌恺 . 专利视域下 “ 卡脖子 ” 技术三阶段识别研究 —— 以芯片材料为例 [J]. 情报杂志 , 2023, 42(10): 132-139, 55.

[22] 俞荣建 , 王雅萍 , 赵一智 , 等 . 破解 “ 卡脖子 ” 技术难题 : “ 情境 —— 策略 ” 非对称匹配视角 [J]. 中国科学院院刊 , 2023, 38(4): 580-592.

[23] 江瑶 , 陈旭 , 胡斌 . “ 卡脖子 ” 关键核心技术两阶段漏斗式甄选模型构建及应用研究 [J]. 情报杂志 , 2023, 42(3): 94-101.

[24] 陈旭 , 江瑶 , 熊焰 , 等 . 关键核心技术 “ 卡脖子 ” 问题的识别及应用：以 AI 芯片为例 [J]. 中国科技论坛 , 2023(9): 17-27.

[25] 李昱璇 , 汤志伟 . “ 卡脖子 ” 技术问题的综合解决方案 —— 以 S 省为例 [J]. 中国科技论坛 , 2022(1): 7-13, 21.

[26] 徐霞 , 吴福象 , 王兵 . 基于国际专利分类的关键核心技术识别研究 [J]. 情报杂志 , 2022, 41(10): 74-81.

[27] 杨斌 . 先进电子材料领域 “ 卡脖子 ” 技术的研判与对策分析 [J]. 科技管理研究 , 2021, 41(23): 115-123.

[28] 董坤 , 白如江 , 许海云 . 省域视角下产业潜在 “ 卡脖子 ” 技术识别与分析研究 —— 以山东省区块链产业为例 [J]. 情报理论与实践 , 2021, 44(11): 197-203.

[29] 夏清华 , 乐毅 . “ 卡脖子 ” 技术究竟属于基础研究还是应用研究 ?[J]. 科技中国 , 2020(10): 15-19.

[30] 武川 , 王宏起 , 王珊珊 . 前沿技术识别与预测方法研究 —— 基于专利主题相似网络与技术进化法则 [J]. 中国科技论坛 , 2023(4): 34-42.

[31] 高楠 , 周庆山 . 新兴技术概念辨析与识别方法研究进展 [J]. 现代情报 , 2023, 43(4): 150-164.

[32] 周波 , 冷伏海 . 演绎逻辑与归纳逻辑视角下的颠覆性技术识别方法研究述评 [J]. 情报学报 , 2022, 41(9): 980-990.

[33] 王康 , 陈悦 , 宋超 , 等 . 颠覆性技术：概念辨析与特征分析 [J]. 科学学研究 , 2022, 40(11): 1937-1946.

[34] 陈光 . 从 P-TMO 模型看我国重大创新突破的关键因素 [J]. 国家治理 , 2020(45): 15-20.

[35] 中国政府网 . 习近平：国家中长期经济社会发展

战略若干重大问题 [EB/OL]. (2020-10-31)[2023-07-16]. https://www.gov.cn/xinwen/2020-10/31/content_5556349.htm.

[36] BOUTELLIER R, LOFFLER K. Bottleneck Technologies: Applying the Constraints Approach to Technology Management Evidence from Case Studies[C]//2006 Technology Management for the Global Future - PICMET 2006 Conference, 2006(1): 38-45.

[37] 中国政府网 . 如何加快解决 “ 卡脖子 ” 难题 [EB/OL]. (2021-08-12)[2023-02-10]. http://www.gov.cn/xinwen/2021-08/12/content_5630865.htm.

[38] 任继球 . 澄清认识 加快构建 “ 卡脖子 ” 技术攻关长效机制 [J]. 宏观经济管理 , 2021(4): 19-25, 33.

[39] INDUSTRY AND SECURITY BUREAU. Federal Register: Review of Controls for Certain Emerging Technologies[EB/OL]. (2018-11-19)[2022-12-28]. https://www.federalregister.gov/documents/2018/11/19/2018-25221/review-of-controls-for-certain-emerging-technologies.

[40] 郑国雄 , 李伟 , 刘溉 , 等 . 基于德尔菲法和层次分析法的 “ 卡脖子 ” 关键技术甄选研究 —— 以生物医药领域为例 [J]. 世界科技研究与发展 , 2021, 43(3): 331-343.

[41] 唐恒 , 邵泽宇 , 蔡兴兵 , 等 . 专利视角下 “ 卡脖子 ” 技术短板甄选研究 [J]. 中国发明与专利 , 2021, 18(1): 54-59.

[42] 程郁 , 叶兴庆 , 宁夏 , 等 . 中国实现种业科技自立自强面临的主要 “ 卡点 ” 与政策思路 [J]. 中国农村经济 , 2022(8): 35-51.

[43] 陈旭 , 江瑶 , 熊焰 , 等 . 基于专利维度的关键核心技术 “ 卡脖子 ” 问题识别与分析 —— 以集成电路产业为例 [J]. 情报杂志 , 2023, 42(8): 83-89, 19.

[44] 范书琴 , 刘国新 . 专利质量视角下国外技术锁定的模糊识别研究 —— 以锂电池隔膜技术为例 [J]. 科学学与科学技术管理 , 2022, 43(12): 132-152.

[45] 周磊 , 吕璐成 , 穆克亮 . 中美科技博弈背景下的卡脖子技术识别方法研究 [J]. 情报杂志 , 2023, 42(8): 69-76.

[46] 吕璐成 , 韩涛 , 陈芳 , 等 . 美国商业管制清单与专利自动映射方法及实证研究 [J]. 情报学报 , 2022, 41(1): 50-61.

[47] 陆天驰 , 闵超 , 高伊林 , 等 . 竞争情报视角下的中美人工智能技术领域差距分析 —— 以美国商品管制清单为例 [J]. 情报杂志 , 2019, 38(11): 25-33.

[48] 张治河 , 苗欣苑 . “ 卡脖子 ” 关键核心技术的甄选机制研究 [J]. 陕西师范大学学报 (哲学社会科学版), 2020, 49(6): 5-15.

doi:10.3772/j.issn.2095-915x.2024.05.002

基于文献关联的生成式人工智能技术演化分析

赛秋玥　徐峰　雷孝平

中国科学技术信息研究所 北京　100038

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 生成式人工智能（AIGC）作为一种前沿技术带来了新的工业革命，全球竞相推进 AIGC 技术的发展。深入理解 AIGC 的内涵是把握其发展趋势的关键。[方法 / 过程] 从多维视角进行 AIGC 技术分析，针对前沿技术的识别和路径演化设计了 YAKE-Apriori 算法展示领域内技术发展规律和路径。首先，基于 YAKE 算法对 AIGC 相关技术进行识别和可视化；其次，基于 Apriori 算法分析 AIGC 关键技术的关联技术，揭示技术发展规律；最后，基于关键技术—技术基础分析刻画技术演化路径，识别技术发展层级关系。[结论 / 结果] 以 AIGC 为对象，成功识别出 GAN、Transformer 和 Diffusion 等 AIGC 相关技术和 4 个主要发展阶段，研究有助于厘清 AIGC 的演化路径，同时验证了 YAKE-Apriori 算法对前沿技术的识别和路径演化分析的有效性。

关键词：生成式人工智能；技术特征；发展脉络；路径演化；热点分析

中图分类号：G35；G203

Research on the Evolution of Generative Artificial Intelligence Technology Based on Literature Association

SAI Qiuyue　XU Feng　LEI Xiaoping

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

Abstract: [Objective/Significance] The advent of Generative Artificial Intelligence (AIGC) marks a new industrial revolution, with nations around the world engaging in the competitive race to advance AIGC technology. Understanding the essence and characteristics of AIGC technology is vital for responding to its growth. [Methods/Processes] This study adopts a multi-dimensional perspective to identify the evolutionary pathways of AIGC technology. The YAKE-Apriori algorithm is designed to identify cutting-edge technologies and path evolution, illuminating the directions of technological development. First, the YAKE method is applied to extract key technologies and visualize literature. Then, the Apriori algorithm is utilized to analyze the

基金项目　科技创新 2030—“新一代人工智能”重大项目、“新一代人工智能风险防范与治理手段研究”课题、“新一代人工智能伦理风险评估与应对策略研究”（2023ZD0121701）。

作者简介　赛秋玥（1993-），博士，助理研究员，主要研究方向为人工智能政策和产业研究；徐峰（1976-），博士，研究员，主要研究方向为科技政策与管理和科技情报研究，E-mail: xufeng@istic.ac.cn；雷孝平（1979-），博士，研究员，主要研究方向为科技情报分析。

引用格式　赛秋玥 , 徐峰 , 雷孝平 . 基于文献关联的生成式人工智能技术演化分析 [J]. 情报工程 , 2024, 10(5): 18-28.

associations between key technologies, thereby revealing the underlying patterns of technological evolution. Finally, the study delineates the trajectory of technology evolution through an analysis of key technologies and their foundational bases, pinpointing the hierarchical relationships within technological development. [Results/Conclusions] Focusing on AIGC, this study successfully identifies relevant technologies such as GAN, Transformer, and Diffusion, along with four primary developmental stages. These insights are crucial for grasping AIGC's trajectory and affirm the effectiveness of YAKE-Apriori method in tracking the progress of emerging technologies.

Keywords: AIGC; Technical Characteristics; Development Trajectory; Evolutionary Pathway; Hotspots Analysis

引言

随着信息时代的快速进展，人工智能已经成为推动社会发展的核心力量，它突破了传统界限，提供了新的智能工具和服务[1]。在这些技术中，生成式人工智能（AIGC）凭借其深度学习和大数据分析能力，在文本、图像和音频内容生成方面展现了与人类智能相匹敌的能力[2]。AIGC 技术是当前世界科技发展的前沿领域之一[3]，并且将在未来推动社会智能化和自动化方向发挥更加重要的作用。然而，由于 AIGC 技术进步迅速，作为一种前沿技术，学术界对于 AIGC 的内涵尚不明确，缺乏一致的理论框架和详尽的技术演化路径分析，这限制了领域内的系统研究和理论深化，同时也制约了技术创新和应用领域的扩展。鉴于此，本研究针对前沿技术的路径演化分析需求，设计了 YAKE-Apriori 算法。研究以 AIGC 技术为例进行深入挖掘，以期揭示其技术特征和演化路径。通过对技术演化的可视化分析，本文将勾勒出 AIGC 技术的发展历程。此外，基于 Apriori 关联算法对 AIGC 技术的研究基础进行层级分析，探讨其技术基础和技术扩展。研究旨在针对 AIGC 技术构建一个全面的技术基础框架，为未来研究和应用提供技术指引。

1 文献综述

本研究对 AIGC 内涵、AIGC 技术特征和相关研究方法进行综述，旨在清晰展示现有 AIGC 研究现状和研究不足，为本研究的开展奠定研究基础。

1.1 AIGC内涵

AIGC 的研究尚无统一的定义标准。通过综合现有文献，发现当前 AIGC 的内涵定义一般涉及内容生成和技术创新两个维度。从内容生成分析，肖峰[4]认为，相比于感知决策阶段，AIGC 代表了人工智能的一个新阶段，不仅能识别和分类现有内容，还能创造全新的文本，因而此阶段被称为内容生成阶段。柴宝勇等[5]将 AIGC 与专业生成内容（PGC）及用户生成内容（UGC）进行比较，强调 AIGC 不仅仅是内容生产者视角下的一种内容类别，更是一种创新的内容生产方式，代表了自动化内容生成的技术集合。祝智庭等[2]认为 AIGC 是在人或机构主导的 PGC、UGC、人工智能辅助内容生成（AGC）之后的发展，利用深度学习等技术框

doi:10.3772/j.issn.2095-915x.2024.05.002

架实现了内容制作者由人变为人工智能技术。

从技术层面分析，李白杨等[6]将人工智能技术的发展划分为基础机器学习和深度学习两大阶段。深度学习技术通过构建复杂网络架构处理高维数据集，分为判别模型和生成模型。判别模型专注于条件概率模型的构建，以区分不同类别；而生成模型则建立特定类别的联合概率分布。生成对抗网络（GAN）的问世，标志着生成式机器学习模型在内容合成领域的应用起始，随着 GAN、Transformer、Diffusion 等模型的进展，2022 年被认为是 AIGC 元年[7]。Bandi 等[8]通过文献综述法梳理了 AIGC 模型，介绍了其关键技术和应用领域。唐宇希[9]根据技术专利对 AIGC 进行简单分析，指出 AIGC 融合了多项人工智能技术，包括 GAN、扩散模型，以及基于 Transformer 架构的大规模预训练语言模型等。但是当前文献很少完整综述生成式人工智能的主流发展技术脉络，不了解其技术演进就不能更好地促进其技术向纵深发展。

1.2 AIGC技术研究综述

AIGC 的技术研究主要围绕技术比较和技术演化路径进行探讨。在技术对比方面，研究多集中于 ChatGPT 和 GAN 的分析。王建磊等[10]探讨了基于 ChatGPT 的技术和算法所展现的传播特性及其对未来发展趋势，强调了 ChatGPT 的技术特点如单一模态和数据依赖性。郑世林等[11]指出 ChatGPT 在推动 AI 发展、劳动市场变革和教育体系重构方面的作用，同时也提出了伴随而来的伦理、价值和信息安全问题。王飞跃等[12]对 AlphaFold 和 ChatGPT 的基本原理及关键技术进行了分析，并概述了大模型时代 AI 技术的发展趋势。卢经纬等[13]从平行智能的角度，细致剖析了 ChatGPT 技术。Jin 等[14]对 GAN 模型的演化模型和重点模型进行了对比分析。Aldausari 等[15]对 GANs 模型用于视频领域的模型研究和应用进行了文献综述。然而，这些研究多采用文献综述法，缺乏定量分析，未能全面反映研究现状。

技术演化研究是分析技术发展历程、趋势和规律的方法，对技术的发展具有重要价值。在技术演化研究方面，根据研究的文本类型分类可以分为文献分析法[16-18]和专利分析法[19-20]。针对 AIGC 的技术演化研究尚处于起步阶段，李白杨等[6]基于文献资料分析了 AIGC 的技术特征与形态演化，从数据、模型和空间三个维度探讨了 AIGC 的技术特征。孙丽文等[21]利用专利数据在能量转换视角下解析了人工智能核心技术产业化的路径。祝智庭等[2]从时间维度勾勒了 AIGC 的发展历程。这些研究虽然从内容生成的角度定义了 AIGC，但未深入分析技术核心及其基础支撑技术。

综上所述，一方面，现有研究对 AIGC 领域的内容研究相对匮乏，多数研究仅从时间维度介绍 AIGC，未能深入剖析其关键技术的逻辑关系，不利于从根本上促进 AIGC 的全面发展。另一方面，AIGC 的研究主要依赖于文献分析，较少运用数据挖掘方法来识别技术基础[22]，因此对技术演化过程的识别方法仍有待进一步优化。

2 研究设计

2.1 研究思路

本研究旨在通过文献关联分析方法构建

AIGC 演化的理论基础和逻辑框架。研究首先收集基于 AIGC 关键词检索得到的文献数据，并综合分析这些文献的引用和被引情况，以揭示 AIGC 的基础技术和其演化路径。“论文引用文献集”是指列在该论文参考文献部分的文献集，涵盖了与 AIGC 密切相关的基础技术；“论文被引文献集”是指引用该论文的后续论文的文献集，该论文一般出现在后续论文参考文献部分。被引论文内容包括了以此内容为基础的技术扩展和改进研究。深入研究这些文献可以帮助我们全面了解 AIGC 的发展脉络，从而为技术演化分析提供更为丰富的理论框架和实证数据。

进一步地，本研究设计 YAKE-Apriori 算法研究 AIGC 的演化路径。采用 YAKE 算法提取 AIGC 领域的关键技术词。之后，应用 Apriori 算法对 AIGC 技术的引用文献进行关联分析，旨在挖掘 AIGC 技术的研究基础。最终，结合关键技术与技术基础，构建了 AIGC 演化脉络。本研究技术路线图如图 1 所示。

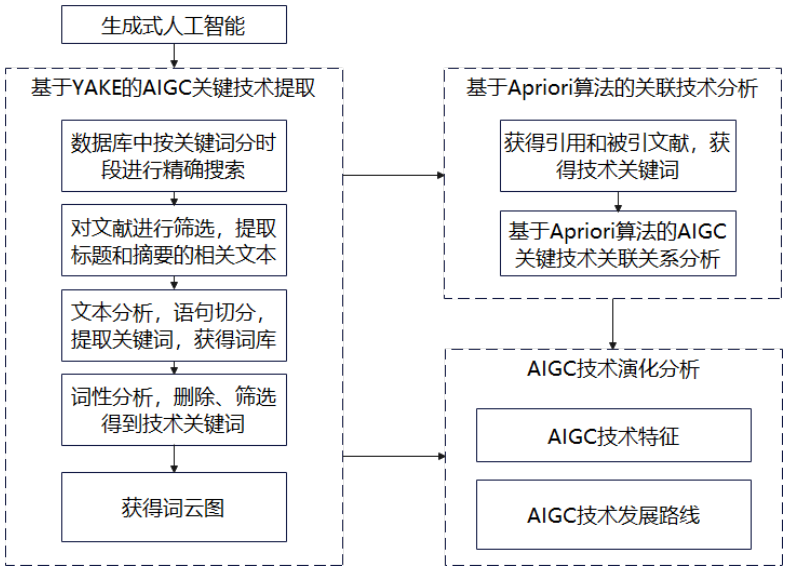


图 1　技术路线图

2.2　基于YAKE的AIGC关键技术提取

本研究采用 YAKE 算法自动提取论文标题和摘要中的关键词，以便快速概括文章的核心内容和方法。选择标题和摘要的原因在于：（1）标题和摘要集中体现了论文的要点；（2）标题和摘要部分的信息与全文的相关性更高；（3）虽然一些论文自带关键词，但这些关键词往往不能详尽反映文章的具体内容。YAKE 算法通过分析词的特征如大写特征、词的位置、词的频率、词与文章相关性和词的频率差 [23]，有效地从大量文献中提取出关键信息。具体特征量化方法详见 Campos[24] 的研究。联合计算指标 $S(w)$ 可以代表每个单词的重要性程度，该值越小，则词表现得越重要，如式（1）。

$$S(w)=\frac{W_{Re}*W_{Position}}{W_{Case}+\frac{W_{Freq}}{W_{Re}}+\frac{W_{DifSentence}}{W_{Re}}} \quad (1)$$

其中 W_{Re} 是词与文章相关性；W_{Case} 是大写特征；$W_{Position}$ 是候选词的位置；W_{Freq} 是词的频率；

doi:10.3772/j.issn.2095-915x.2024.05.002

$W_{DifSentence}$ 是候选词的句中频率。

候选关键词可以是单词或词组，因此本研究设置了滑动窗口大小为3来设定关键词序列。不考虑以停用词开头和结尾的连续序列。每个候选关键词分配一个 $S(kw)$，这样分数越小，关键词就越有意义。

$$S(kw)=\frac{\Pi_{w\in kw}s(w)}{TF(kw)*(1+\sum_{w\in kw}s(w))} \quad (2)$$

$S(kw)$ 是最长词数为 3 的候选词的得分。

2.3 基于Apriori算法的AIGC技术关联分析

Apriori 算法通常用于找出数据中的频繁项集，进而挖掘关联规则。在文献研究中，通过分析论文的引用和被引用关系，可以发现技术之间的基础关系。利用 Apriori 算法分析 AIGC 论文中文献与其引用和被引文献的关联关系[25]，可以有效地揭示该领域的技术联系，是分析 AIGC 技术发展脉络的一个重要方法。

（1）项与项集：$itemset=\{item_1, item_2, \ldots, item_m\}$ 是所有论文的集合，包含与生成式人工智能技术有关的引用论文和被引论文的集合。每一项是每篇论文和其关键词。

（2）关联规则：关联规则式形如 $C(A)\Rightarrow C(B)$，其中 $C(A)$、$C(B)$ 均为包含关键词 A 和 B 的且属于 $intemset$ 的子集的非空集合，且 $C(B)$ 为 $C(A)$ 对应论文引文的关键词集合。

（3）支持度：关联规则的支持度定义如式（3）。本研究中支持度指同时在论文中出现关键技术 A 和其引文出现关键技术 B 的次数，即出现关键词对 A 和 B，且 B 所在论文是 A 所在论文的引文。

$$support(A\Rightarrow B)=P(A\cup B) \quad (3)$$

（4）置信度：关联规则的置信度定义如式（4）。本研究中置信度指论文中出现关键技术 A 和其引文出现关键技术 B 的次数与出现关键词 A 的次数的比例。

$$confidence(A\Rightarrow B)=P(B|A)=\frac{support(A\cup B)}{support(A)} \quad (4)$$

（5）频繁项集：如果项集的相对支持度满足预先定义好的最小支持阈值，则为频繁项集[26]。本研究中，论文与其引用论文出现相同关键词对的频率越高，则关联程度越大。

Apriori 算法的两个输入参数分别是最小支持度和数据集。首先会查找所有生成式人工智能论文的项集列表。搜索每个 AIGC 中关键技术的支持度较大的前两对为与本技术关联性最大的技术。该过程重复进行直到所有关键技术均被搜索完全。

3 研究过程和结果分析

3.1 数据集介绍

研究从 Web of Science 论文库中获得数据，数据包括了论文题目、出版年月、摘要、论文引文、论文被引文。Web of Science 是相对权威的论文数据库，能够提供高质量的全面的论文研究信息，因此可以用于 AIGC 技术现状的研究。

在搜索的论文中，2006 年文献开始出现，因此检索时间段确定为 2006 年 1 月 1 日至 2023 年 10 月 1 日。检索时间为 2023 年 11 月 1 日。本研究主要对技术进展进行研究，因此检索研究领域为计算机科学、工程和数学。由于论文研究数量庞大，包括很多干扰性论文和关注度较小的论文。本研究选择引用次数大于 2

的论文进行导出分析。最终的检索式为 Topic=(“Generative AI” or “GenAI” or “AIGC” or “Generative Artificial Intelligence”) AND Time=(2006.01.01: 2023.10.01) AND Research Areas=(“Computer Science” or “Engineering” or “Mathematics”) AND Citations=(“>2”)。

为了防止检索式不能包含所有 AIGC 相关论文，根据论文的引文和被引也涉及 AIGC 内容的原则，研究再根据引用和被引文献条目，获得所有引用和被引的论文共同作为分析的数据。最终通过直接搜索 AIGC 关键词后获得 4168 篇论文，被引文献共 52151 篇论文。被引文献数量可以表示该领域的受关注度。关于 AIGC 论文的发表数和被引用量如图 2 所示。本研究引入了年增长率来划分研究时段，并新增了论文发表量年增长率和被引用量年增长率，如图 3 所示，从 2014 年开始，论文增长率一直保持正数，论文发表量持续增加，但在 2020 年后增长率明显下降。基于这一观察，我们将 AIGC 的发展历程划分为三个阶段。在 2014 年之前是 AIGC 的起步阶段，相关论文的发表量和被引用量都相对较少且稳定。2014 年被视为 AIGC 发展的关键之年，多个 AIGC 模型在这一年被提出。2014—2020 年为初步发展阶段，论文的发表量和被引用量持续增加，表明这一时期的论文对后续发展产生了重要影响。2020 年之后进入了快速发展时期，GPT-3 的诞生打破了神经网络创新纪录，引起了高度关注，论文的发表量和被引用量迅速增长。根据近几年发展态势，未来几年 AIGC 技术研究还将是研究的热点领域。

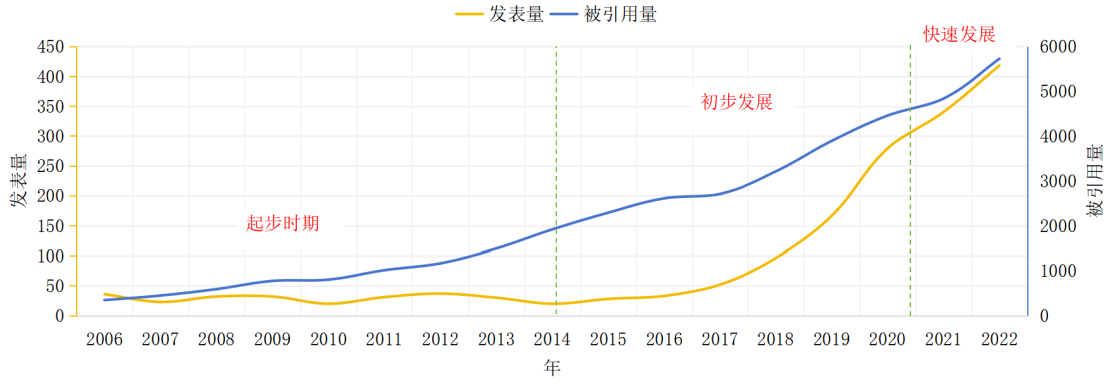


图 2　论文发表量和被引用量

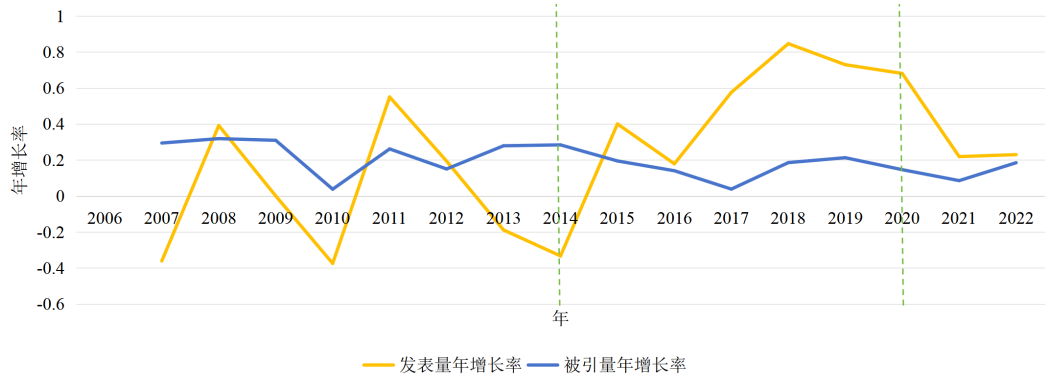


图 3　论文发表量年增长率和被引用量年增长率

doi:10.3772/j.issn.2095-915x.2024.05.002

3.2 基于YAKE提取AIGC技术关联词

研究根据技术关键词进行提取，我们获得了关键技术的列表并提取出技术关键词。对关键词进行分类，可以识别出基础模型系列，包括 DBN、RNN、LSTM、CNN、GAN 模型系列、Transformer 模型系列、Diffusion 模型系列以及流模型、CLIP、辐射场模型、变分自编码器模型等其他 AIGC 相关关键技术。为了获得 AIGC 的研究热点。研究按时间划分为 2006—2013 年，2014—2020 年，2021—2022 年。图 4 介绍了文献的关键词及被引文献的分析结果，其中，第一行是直接搜索得到的文献关键词，第二行是被引文献的关键词，第三行是此时间段发表的论文在之后被引用数量的时序图。

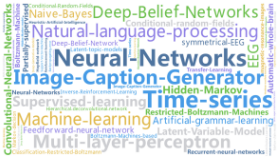
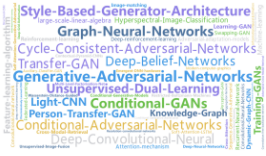
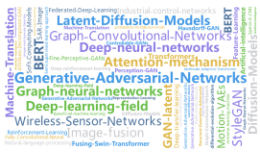
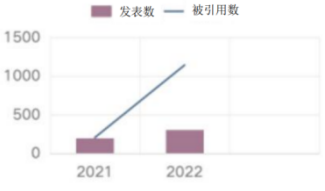
时间	2006—2013 年	2014—2020 年	2021—2023 年
发表论文			
	Deep Belief Nets（DBN）, convolutional neural network（CNN）	Generative Adversarial Networks (GAN), CNN	ChatGPT, Variational Auto-Encoder（VAE）, Transformer, GAN
被引论文			
	Deep Belief Nets（DBN）, convolutional neural network（CNN）, Feedforward neural network (FNN), EGG	StyleGAN, Conditional GAN	StyleGAN, BERT, Diffusion
时序图			

图 4　发展过程及研究基础分析

根据图 4 的关键词分析，对不同时间 AIGC 研究热点过程分析，可以获得 AIGC 领域的研究发展过程。AIGC 技术从最初的 DBN，CNN，逐渐演化到 GAN 等无监督生成网络。当前，研究的焦点集中在 Transformer 架构、ChatGPT、变分自编码器（VAE）以及 GAN 的进化模型。

对被引用论文研究发现，人工智能技术发展包括前馈神经网络（FNN）、脑电技术（EGG）、Transformer、BERT、GAN 和扩散模型（Diffusion model）的进化模型，AIGC 领域涵盖了广泛的机器学习和深度学习技术。这些技术的应用和发展不断推动 AIGC 研究的深入。根据被引论文的引用量趋势，我们预见 2021—2023 年间发表的文献将成为后续研究中的引用热点，尤其是围绕 Transformer、ChatGPT 和 GAN 模型的研究。这些研究领域不仅在学术界受到高度重视，也预示着它们在工业界的应用前景。

为了分析技术的研究热度，图 5 从时间维度呈现了各技术论文引用量的关系，论文引用量越大表示当前研究热度越大。从时间上可以看出 RNN、CNN 和 DBN 的研发时间较久远，是当前人工智能技术的研究基础。LSTM 模型出现时间也较久远，且其属于 RNN 的改进模型，因此我们认为其在 AIGC 技术中属于基础模型。近年来新产生且研究热度持续上升的模型主要是 Transformer 和 GAN 模型[27]。ChatGPT 和 GPT-3 是近两年研发的模型，属于前沿模型。

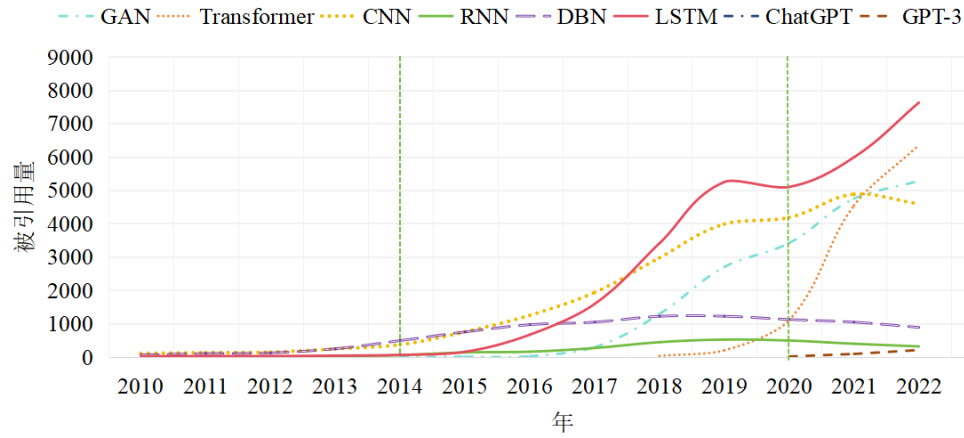


图 5　关键技术论文被引用量时序图

3.3　基于Apriori算法的AIGC技术关联分析结果

根据 3.2 内容，我们识别了 AIGC 的关键技术。通过分析关键技术相关文献的引用与被引用情况，我们构建了几个主流技术的研究基础，见表 1。通过提取每项技术文献中被引用和引用的关键词，并筛选出频率最高的两个，我们确定了每项技术最关键的技术基础和技术拓展方向。

表 1　技术关联分析结果

关键技术	高引用	高被引	重要能力变化
GAN	Unsupervised model (8%) Neural Network (5%)	Transformer（10%） Diffusion model（6%）	显著改善生成图像
Transformer	RNN（10%） LSTM（7.5%）	BERT（7%） ChatGPT（3%）	解决序列转换的挑战
Diffusion model	概率估计（14%） 贝叶斯模型（9%）	Improved DDPM（6%） Diffusion Probabilistic Model（2%）	生成图像

根据表 1 的数据，可以看出 AIGC 技术的发展建立在基础技术之上，这些技术通过引文关系相互关联。具体来说，GAN 模型的创新借鉴了无监督模型的理论基础，这个结论与 Kaswan 等[28]观点一致；而 Transformer 模型与 RNN 和 LSTM（长短期记忆网络）都是一种特征提取器并且有替代 RNN 和 LSTM 模型成为当前的主流模型的趋势。Hao 等[29]指出，当前大多数生成式人工智能的大语言模式均使用 Transformer 架构。

doi:10.3772/j.issn.2095-915x.2024.05.002

Transformer 模型奠定了 BERT 和 ChatGPT 发展的基础。Diffusion 模型以似然估计和贝叶斯模型为基础，并逐步扩展为改进的 Diffusion 模型。

3.4 AIGC技术演化分析

技术演化路径分析的依据包括技术发明的时间顺序和技术关联关系。首先，技术演化路径一般与时间有关，后续出现的技术一般在先前技术的基础上发明，根据 3.2 的内容，技术的出现和研究热点的时间顺序是技术演化的依据之一。此外，技术演化不仅关注于时间顺序，还要依据各技术之间的关联关系。3.3 介绍了与 AIGC 有关的技术基础关系。因此，综合时间关系和技术关联关系，研究总结出 AIGC 的发展层级，如图 6 所示。

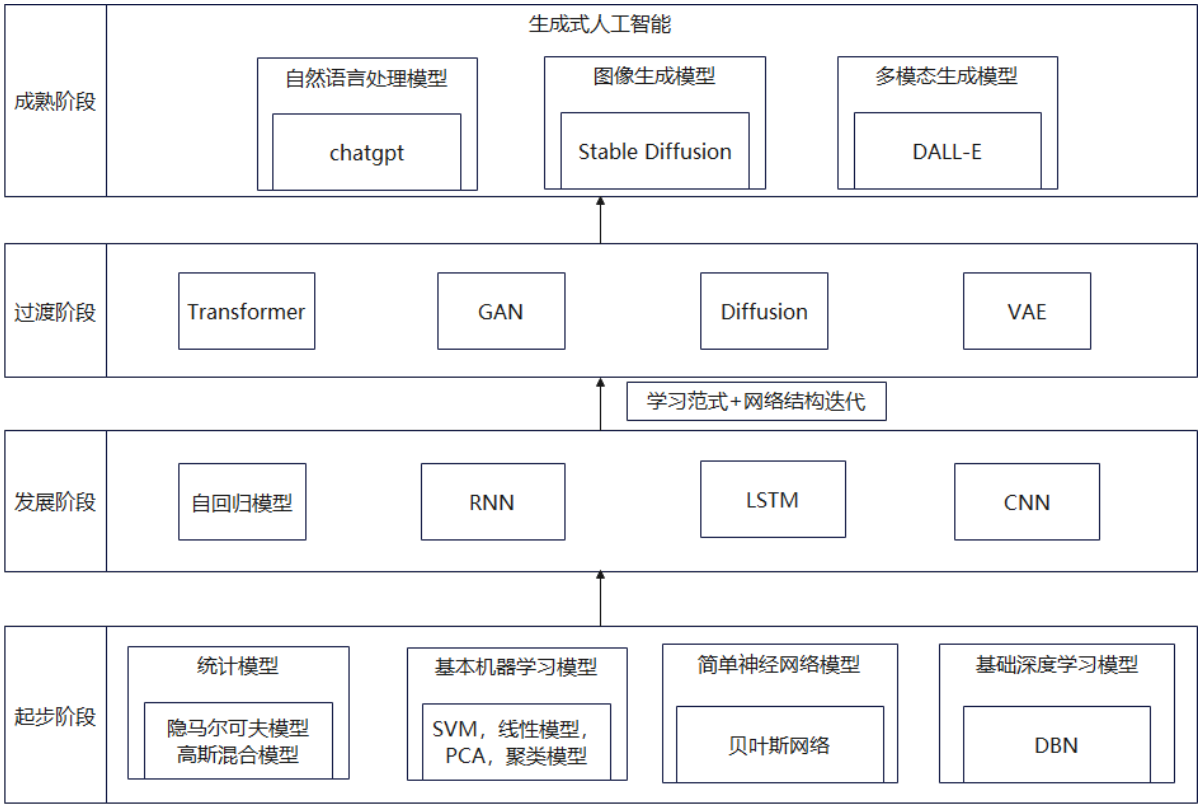


图 6 AIGC 技术演化层级分析

当前研究对 AIGC 发展进行了阶段划分，一些研究将 AIGC 发展时段划分为早期萌芽阶段，沉淀积累阶段和快速发展阶段。但是其没有区分技术过渡和技术成熟两个阶段，同时没有对每个阶段包含的基础技术进行介绍。由于这两个时期具有明显差异，因此本研究对其进行了区别和划分。本研究根据 AIGC 技术的时间发展和技术基础，同时根据历史经验知识，确定技术发展的层级结构。AIGC 发展阶段根据其生成内容的完善程度可以分为 4 个阶段，分别是起步阶段、发展阶段、过渡阶段和成熟阶段。起步阶段的 AIGC 模型主要是基础模型，在内容生成方面能力很有限，主要完成决策和预测的内容生成。发展阶段可以完成简单的数

据生成，在自然语言和图像生成领域能力有限。过渡阶段的模型发展逐渐成型，可以完成复杂的自然语言处理和图像生成，此阶段还处于科学研究阶段，到成熟阶段 AIGC 技术已经可以用于工业界。

AIGC 技术的发展主要从学习范式和网络结构两个方面进化迭代。一方面，模型结构的进步和多样化，从简单的统计模型，基本机器学习模型，简单神经网络模型和深度学习模型发展为 GAN、Transformer 和扩散模型等。另一方面出现多种生成学习范式，从基于规则的机器学习范式和有监督学习范式发展为预训练和精调结合的学习范式，构成了 AIGC 的核心框架。AIGC 技术的定义和实践正在随着这些模型的不断进化而扩展，展现出强大的创造力和应用潜力。未来的 AIGC 研究将进一步提升这些模型的生成效率、质量以及多样性，同时也将探索新的模型结构和训练方法，以更好地服务于各种生成任务。

4 总结和展望

本研究基于文献数据，对 AIGC 技术的核心要素及其技术演化路径进行了深入探讨。研究创新性地提出通过关键词提取和 Apriori 算法识别 AIGC 的关键技术方法。利用文献的引用和被引关系，以及关键技术之间的演化趋势，揭示 AIGC 技术演化规律。研究方法也适用于其他技术演化路径分析。本研究的重要贡献为研究首先以 AIGC 为研究对象，成功识别出 GAN、Transformer 和 Diffusion 等 AIGC 相关关键技术，以及 AIGC 技术的 4 个主要发展阶段，旨在为 AIGC 技术的未来发展提供理论支持和指导方向。当前看，这些技术在未来仍拥有较大的研究空间，学术研究不断增多且技术性能不断增强。尤其是 Transformer 和 Diffusion 是 AIGC 后续发展中的重要探索方向。其次，研究验证了 YAKE-Apriori 算法用于技术演化路径的可行性，为技术演化路径的研究方法提供了技术基础，为 AIGC 的演化路径提供清晰的历史参考。

未来研究将深入分析 AIGC 主流技术的技术内涵和发展规律，预测技术发展态势，旨在为 AIGC 技术的发展提供广阔视角，并为政策制定者、研究者和行业实践者提供有用的建议和指导。

参 考 文 献

[1] ROUMELIOTIS K I, TSELIKAS N D. ChatGPT and Open-AI Models: A Preliminary Review[J]. Future Internet, 2023, 15(6): 192.

[2] 祝智庭 , 戴岭 , 胡姣 . 高意识生成式学习 : AIGC 技术赋能的学习范式创新 [J]. 电化教育研究 , 2023, 44(6): 5-14.

[3] DE SCHRYVER G M. Generative AI and Lexicography: The Current State of the Art Using ChatGPT[J]. International Journal of Lexicography, 2023, 36(4): 355-387.

[4] 肖峰 . 生成式人工智能与知识生产新形态 [J]. 学术研究 , 2023(10): 50-57.

[5] 柴宝勇 , 陈若凡 , 陈浩龙 . 中国网络内容治理政策 : 变迁脉络与工具选择——基于政策文本的内容分析 [J]. 中南大学学报（社会科学版）, 2023, 29(5): 148-161.

[6] 李白杨 , 白云 , 詹希旎 , 等 . 人工智能生成内容 (AIGC) 的技术特征与形态演化 [J]. 图书情报知识 , 2023, 40(1): 66-74.

[7] 孙冰 , 潘建东 . 中信建投证券生成式 AI 技术研究与探索 [J]. 中国金融电脑 , 2023(10): 73-76.

doi:10.3772/j.issn.2095-915x.2024.05.002

[8] BANDI A, ADAPA P V, KUCHI Y E. The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges[J]. Future Internet, 2023, 15(8): 260.

[9] 唐宇希 . AIGC 技术专利分析 [J]. 通讯世界 , 2023, 30(7): 190-192.

[10] 王建磊 , 曹卉萌 . ChatGPT 的传播特质、逻辑、范式 [J]. 深圳大学学报（人文社会科学版）, 2023, 40(2): 144-152.

[11] 郑世林 , 姚守宇 , 王春峰 . ChatGPT 新一代人工智能技术发展的经济和社会影响 [J]. 产业经济评论 , 2023(3): 5-21.

[12] 王飞跃 , 缪青海 . 人工智能驱动的科学研究新范式 : 从 AI4S 到智能科学 [J]. 中国科学院院刊 , 2023, 38(4): 536-540.

[13] 卢经纬 , 郭超 , 戴星原 , 等 . 问答 ChatGPT 之后 : 超大预训练模型的机遇和挑战 [J]. 自动化学报 , 2023, 49(4): 705-717.

[14] JIN L, TAN F, JIANG S. Generative Adversarial Network Technologies and Applications in Computer Vision[J]. Computational Intelligence and Neuroscience, 2020, 2020(1): 1459107.

[15] ALDAUSARI N, SOWMYA A, MARCUS N, et al. Video Generative Adversarial Networks: A Review[J]. ACM Computing Surveys, 2023, 55(2): 30.

[16] 李修全 . 当前人工智能技术创新特征和演化趋势 [J]. 智能系统学报 , 2020, 15(2): 409-412.

[17] 邓莎莎 , 李镇宇 , 潘煜 . ChatGPT 和 AI 生成内容 : 科学研究应该采用还是抵制 [J]. 上海管理科学 , 2023, 45(2): 15-20.

[18] 王秀杰 , 陈云静 . 国内外智能网联汽车研究热点可视化对比分析——基于文献计量视角 [J]. 科技和产业 , 2023, 23(5): 98-107.

[19] 徐峰 , 冷伏海 . 专利技术形态分析方法研究进展 [J]. 图书情报工作 , 2010, 54(10): 54-57.

[20] 冯立杰 , 周炜 , 刘鹏 , 等 . 基于引文网络和语义分析的技术演化路径识别及拓展研究 [J]. 情报理论与实践 , 2023, 46(1): 90-99, 131.

[21] 孙丽文 , 李少帅 , 孙洋 . 能量转换视角下人工智能关键核心技术产业化路径解析 [J]. 科技进步与对策 , 2022, 39(14): 73-82.

[22] 雷孝平 , 张海超 , 桂婕 , 等 . 基于论文和专利的区块链技术研发状况分析 [J]. 情报工程 , 2017, 3(2): 20-32.

[23] 席耀一 , 高鑫 , 王小明 , 等 . 基于 ETM 模型的中亚国家” 一带一路” 网络舆情热点检测 [J]. 情报杂志 , 2020, 39(11): 82-89.

[24] CAMPOS R, MANGARAVITE V, PASQUALI A, et al. Yake! keyword ex-traction from single documents using multiple local features[J]. Information Sciences, 2020, 509: 257-289.

[25] 万小萍 , 刘向 , 闫肖婷 , 等 . 基于关联分析的技术演化路径发现 [J]. 情报学报 , 2018, 37(11): 1087-1094.

[26] ZHENG J G, ZHANG J M. The Research about Data Mining of Network Intrusion Based on Apriori Algorithm[C]//Proceedding of the 7th International Conference on Education, Management, Computer and Medicine (EMCM), 2017, 59: 661-664.

[27] KARRAS T, LAINE S, AILA T. A Style-Based Generator Architecture for Generative Adversarial Networks[C]//2021 IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4217-4228.

[28] KASWAN K, DHATTERWAL J, MALIK K. Generative AI: A review on Models and Applications[C]//2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), 2023: 669-704.

[29] HAO J, YANEVA V. Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI[J]. Educational Measurement-Issues and Practice, 2024, 43(2): 16-29.

doi:10.3772/j.issn.2095-915x.2024.05.003

老年人健康信息获取行为影响因素研究

李芊晨　王丙炎

湘潭大学公共管理学院 湘潭 411105

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 我国人口老龄化、老年人的养老和健康问题越来越为政府和社会关注。老年人健康信息获取，可提升其健康素养及生活质量，促进健康养老。[方法 / 过程] 基于扎根理论方法，通过半结构化访谈方式收集了 30 位老年人的健康信息获取经历，以开放式编码、主轴编码及选择性编码对影响老年人健康信息获取行为的因素进行归纳与分析。[结果 / 结论] 老年人健康信息获取行为受环境因素、心理因素、能力因素、媒介因素、感知因素和信息因素的共同影响。其中，感知因素直接产生作用，其他因素均间接作用。在此基础上，可以进一步优化老年人健康信息获取策略和服务方式。

关键词：健康信息获取行为；影响因素；老年人；扎根理论

中图分类号：G35；G252

Research on the Influencing Factors of Health Information Acquisition of the Elderly from the Perspective of Information Encounter

LI Qianchen　WANG Bingyan

School of Public Administration, Xiangtan University, Xiangtan 411105, China

Abstract: [Objective/Significance] The aging of the population in China, the pension and health problems of the elderly are more and more concerned by the government and society. Access to health information for the elderly can improve their health literacy and quality of life, promote healthy elderly care. [Methods/Processes] Based on the grounded theory method, the health information acquisition experience of 30 elderly people was collected through semi-structured interviews, and the factors affecting the health information acquisition behavior of the elderly were summarized and analyzed by open coding, spindle coding and selective coding. [Results/Conclusions] The results showed that the health information acquisition behavior of the elderly was affected by environmental factors, psychological factors, ability factors, media factors, perception factors and information factors. Among them, the perceptual factors play a direct role, and the other factors play an indirect role. On this basis, the health information acquisition strategy and service mode of the elderly can be further optimized.

Keywords: Health Information Acquisition Behavior; Influencing Factors; Senior Citizen; Grounded Theory

作者简介　李芊晨（2000-），硕士研究生，主要研究方向为公共信息资源管理，E-mail：1585788379@qq.com；王丙炎（1971-），硕士，副教授，主要研究方向为知识管理与数字图书馆技术。

引用格式　李芊晨，王丙炎 . 老年人健康信息获取行为影响因素研究 [J]. 情报工程 , 2024, 10(5): 29-37.

doi:10.3772/j.issn.2095-915x.2024.05.003

引言

我国正面临着严峻的人口老龄化问题，保证老年健康已经成为社会关注的重点[1]。老年人对养生保健、医疗保险、疾病防范等健康信息的需求更加迫切[2]，越来越多的老年人通过获取健康信息实现健康决策支持[3]，通过社交互动、平台推送等渠道进行的随机性信息获取也成为老年人获取健康信息的重要途径[4]。目前老年人健康信息获取行为研究中多为搜寻和查找等主动信息获取研究，很少涉及其他形式的健康信息获取行为研究。因此，本研究聚焦具有随机性的老年人健康信息获取行为，主要探讨以下两个问题：（1）老年人实施健康信息获取行为时有哪些影响因素？（2）影响因素之间有怎样的作用关系？

1 老年人健康信息获取行为影响因素研究概述

健康信息获取行为包括信息需求、信息搜寻、信息检索、信息寻求等，在信息获取过程中涉及信息源、信息获取途径、信息获取工具和策略等具体内容[5]。针对老年人健康信息获取行为，国内学者们开始从健康信息搜寻行为的影响因素展开研究。总体而言，研究表明老年人健康信息搜寻行为的影响因素可分为四类：①个体因素，涉及个人心理因素、个人实施成本、健康素养等；例如朱姝蓓等[6]发现健康意识作为主导因素影响老年人健康信息搜寻行为；②信息因素，涉及信息特征、信息质量、信息可信度等；李锴[7]通过分析老年群体健康信息搜寻困境，指出健康信息质量和隐形推销内容是搜寻行为主要影响因素，而实行新媒体适老化改造、家庭数字反哺，则有助于提升老年群体健康信息搜寻能力；③媒介因素，涉及信息渠道、使用意愿、人脉资源等；例如戴艳清[8]发现信息来源是影响老年人健康信息搜寻行为的重要因素之一，指出公共图书馆通过举办健康知识讲座、提供中介检索服务能帮助老年人有效提升自身健康信息搜寻能力；④环境因素，涉及社会支持、生活世界等；如王琳等[1]的研究发现我国老年女性的健康信息搜寻行为受所处社群的社会规范、社会类型、世界观等的影响极大。

总体而言，目前我国学者对老年人健康信息获取行为已有的研究主要集中于信息搜寻行为及其影响因素，但是对于各要素间的关联与作用路径关注较少，同时对信息获取全过程关注较少，对老年人健康信息获取行为中信息获取全过程的相关研究尚不充分。基于此，本研究利用半结构化访谈和扎根理论，调查老年人健康信息获取情况，分析其影响因素和各因素间作用关系，构建老年人健康信息获取行为影响因素框架，为老年人健康信息获取行为相关研究提供理论指引。

2 研究设计

2.1 研究方法

扎根理论是由美国学者 Glaser 和 Strauss 提出的一种借助已有文献，根据研究者自身经验和资料开展编码分析，形成理论模型的质性研究方法，且这种方法能在系统性收集资料的基础上，揭示事物现象的核心概念，厘清现象内部发展逻辑关系，尤其适用于主题研究尚不完

善时进行理论构建。因此，本文选择扎根理论研究方法，在收集整理老年人日常生活经历的基础上对老年人健康信息获取行为进行探究。

2.2 数据收集与整理

本研究使用半结构化访谈法收集数据，主要使用面对面形式对共 30 位 60 岁及以上的老年人进行一对一访谈，平均访谈时间为 35 分钟。由于老年人互联网使用率较低，招募志愿者难度相对较大，考虑到方言口音等客观交流问题，为减少受访者对个人隐私泄露的担忧，笔者基于个人交际网络关系，选择弱链接关系人群作为受访人群进行质性研究数据收集。样本对象基本情况如表 1 所示。

表 1 样本对象基本情况

编号	性别	职业	受教育程度	年龄	编号	性别	职业	受教育程度	年龄
01	男	会计	高中	79	16	男	公务员	专科	67
02	男	教师	专科	76	17	女	公务员	高中	65
03	女	工人	初中	75	18	男	工人	初中	72
04	男	工程师	本科	80	19	男	警察	本科	61
05	女	工人	高中	75	20	女	个体	专科	60
06	女	教师	专科	62	21	男	国企员工	初中	71
07	男	国企员工	高中	65	22	男	医生	本科	77
08	女	个体	高中	60	23	女	教师	专科	67
09	男	个体	初中	63	24	男	工人	初中	74
10	女	公务员	高中	61	25	女	个体	高中	65
11	女	医生	本科	64	26	男	个体	高中	68
12	男	警察	高中	65	27	女	工人	初中	75
13	女	教师	专科	68	28	男	教师	高中	74
14	男	工人	初中	61	29	女	教师	高中	73
15	女	务农	初中	76	30	女	国企员工	专科	77

在正式访谈之前，笔者根据研究问题和研究目的，结合先前收集资料和相关文献设计了半结构化访谈提纲，并对 2 位老年人进行预访谈，以修正访谈提纲。半结构化访谈提纲为：（1）您关注健康信息的原因是什么？（2）您平时怎样获取健康信息？是否通过网络获取？（3）您平时获取健康信息时，比较关注哪些方面的信息？ 什么样的情境下关注这些信息？（4）您获取健康信息时遇到什么困难？怎样解决这些困难？（5）您相信您浏览到的健康信息吗？您比较相信哪种渠道获取的健康信息？

笔者在 2023 年 6—7 月间对以上 30 名受访者分别进行深度访谈，为确保研究严谨性和数据精准性，全程由笔者记录、录音和整理，形成 30 份文字资料，并从中选择 25 份资料进行编码分析，剩余 5 份资料作为饱和度检验。

doi:10.3772/j.issn.2095-915x.2024.05.003

3 数据编码与分析

3.1 开放式编码

在开放式编码阶段，笔者将 25 份原始资料打散，进行概念提炼，通过对比、合并后，删除了相互矛盾或重复的概念，最终获得 49 个基本范畴（分别以 B1、B2、B3……Bn 表示），对所有基础范畴进行归纳整合后，共形成 14 个初始概念（分别以 A1、A2、A3……An 表示），如表 2 所示。

表 2 开放式编码情况

范畴	初始概念	原始语句
A1 人际环境	B1 后辈影响	儿子媳妇买的手机，还经常转发给我。（23）
	B2 朋友影响	一起练瑜伽的一个退休老师告诉我这些健康信息的。（18）
	B3 同事影响	好些养生知识都是之前老陈发我的。（04）
	B4 街坊影响	问别人又不知道，都是问邻居，平时也聊聊保健信息。（10）
	B5 从众行为	上街人家都谈这些（健康信息）呀，不看这些没得聊，那就看看嘛。（19）
A2 需求动机	B6 年龄大	岁数大了，那不要多了解了解（健康信息）吗。（22）
	B7 自身患病	糖尿病快二十年了，平时看点健康知识学怎么保养。（13）
	B8 减轻子女负担	身体好就不给子女添负担。（01）
	B9 防微杜渐	就是平时注意健康，防微杜渐，未雨绸缪。（23）
A3 信息素养	B10 不会搜索	我不会搜欸。（11）
	B11 不了解平台使用方法	不晓得怎么用，只晓得划着看，也不会存下来也不会别的。（14）
	B12 不了解自身信息需求	也没有什么具体想法吧，我都是刷到什么（健康信息）就看什么。（09）
A4 生理机能	B13 视力下降	白内障看不清了，看东西费劲得很。（15）
	B14 记忆力下降	老啦，我孙子怎么教，我都记不住。（17）
A5 平台功能	B15 内容丰富	头条（健康知识）太多了，比报纸多多了，都看不过来。（21）
	B16 视听方便	抖音又能看又能听，方便呀。我去田里做事把抖音一放，一点不妨碍。（05）
	B17 满足碎片化阅读需要	坐地铁啦，等车啦，排队的时候啦，随便什么时候都可以看。（16）
	B18 成本低	抖音快手看这个不要花钱。（21）
A6 信息来源	B19 社交平台	关注了好多公众号和视频号。（09）
	B20 短视频平台	经常用抖音、快手看这些，别的不怎么用。（02）
	B21 报纸杂志	还是习惯看订的报纸。（12） 我老伴订了好几年的《家庭医生》，我跟着一起看。（03）
A7 感知有用性	B22 可读性	有些文章字体颜色太浅了，字又小，根本看不下去。（06）
	B23 可理解性	能看懂的才看，有些都不知道讲的什么。（24）
	B24 可掌握性	那我更关注能真的在生活里得到利用的（健康）信息，不能利用的看了也没用啊。（22）
A8 感知易用性	B25 便捷性	是欸，方便在家弄的，不方便的就划走了。（24）
	B26 个性化	对，经常看那种不笼统、讲得很详细的，尤其那种能听我们粉丝呼声，按照评论问题出的解答视频。（03）
	B27 匹配度	那当然更愿意看符合家里人情况的了，别的看了也没用。（20）

续表

范畴	初始概念	原始语句
A9 信息偏好	B28 健康养生	那我喜欢看养生的，尤其带练八段锦的那种，我还能跟着练。（25）
	B29 民间偏方	我还是蛮喜欢看那种小偏方的，方便效，平时就能用。（20）
	B30 食疗养生	刷那种药膳食补哦，广东的各种汤，我经常学。（01）
	B31 疾病治疗	就看看人家说这个高血压怎么治，这个颈椎病怎么保养哦。（13）
	B32 日常防范	有段时间看看那个新冠哦，怎么防，还有甲流啊。（06）
A10 信息效用	B33 节约金钱	看这个省钱啊，有些小问题就不用去医院挂号了，直接在药店开点药。（04）
	B34 节约时间	对，省时间。大医院离太远，学这些省的去医院了。（15）
	B35 缓解心情	那看这个能放松放松，改善一下心情。（25）
	B36 增长知识	万一以后能用上呢，多看看多了解一点哦。活到老，学到老嘛。（11）
	B37 有助健康	那些食补、锻炼的视频，跟着学学对身体好啊。（07）
A11 信息质量	B38 可靠性	这个公众号是我儿子告诉我的，我儿子说它可以相信，那我就信。（10）
	B39 及时性	这个医生的头条号发的都是紧跟时事的（知识），我就经常看。（08）
	B40 权威性	我就看什么央视新闻啊，人民网啊（的健康信息），别的不知道真假就不看。（16）
	B41 实用性	我关注了一个每天发不同食补的抖音号，能在家学着烧，别的用不上的不看。（02）
	B42 信息过载	网上（健康信息）太多了，看不过来欸。（17）
	B43 相关性	我只看养生的，别的不看。（03）
A12 信息获得	B44 算法推荐	它推什么我就刷着看。（08）
	B45 被动获取	我儿子给我发什么我就看什么，不搜（健康信息）的。（05）
A13 依赖程度	B46 完全依赖	基本只靠刷（手机）知道这些（健康信息），那又不会搜。（23）
	B47 部分依赖	我也问医生啊，不止刷手机（获取健康信息）的。（14）
A14 依赖时效	B48 持续依赖	我一直刷视频看（健康信息）的，看了不少年了。（19）
	B49 暂时依赖	偶尔看看，有需要才看。（12）

3.2 主轴编码

主轴编码指对开放式编码形成的范畴进行梳理、对比，以确定主范畴与子范畴之间的关联性。笔者通过 14 个基本范畴的辨析，共总结出环境因素、心理因素、能力因素、媒介因素、感知因素、信息因素与信息随机获取共 7 个主范畴，如表 3 所示。

3.3 选择性编码

选择性编码是指通过对现有范畴进行梳理和提炼，形成能覆盖各范畴的核心范畴，基于核心范畴组成能解释具体现象的主线。通过反复比较、关联现有主范畴，本研究得出以环境因素、心理因素、能力因素、媒介因素、感知因素、信息因素等 6 个主范畴和信息随机获取为核心范畴为主范畴的故事线，以此探究出各主范畴与核心范畴间的老年人健康信息获取行为关系构成，如表 4 所示。

3.4 饱和度检验

理论饱和度检验指在原始资料的基础上，不能获得新的范畴和关系。本研究使用预留的

doi:10.3772/j.issn.2095-915x.2024.05.003

表 3　主轴编码结果

主范畴	基本范畴	范畴内涵
C1 环境因素	人际环境	个体获得的来自人际环境的物质和行为支持
C2 心理因素	需求动机	个体自身接受健康信息的心理诉求
C3 能力因素	信息素养	个体具有检索、分析、评价和利用健康信息以解决自身健康信息需求的能力
	生理机能	个体获得健康信息时的生理问题
C4 媒介因素	平台特点	个体选择网络平台获取健康信息的因素
	信息来源	个体获得健康信息的渠道
C5 感知因素	感知有用性	个体感知到健康信息的使用价值
	感知易用性	个体感知到健康信息的使用难易程度
C6 信息因素	信息偏好	个体获取健康信息的态度
	信息质量	健康信息的特性
	信息效用	个体使用健康信息时的满足程度
C7 信息随机获取	信息获得	个体随机获取健康信息的方式
	依赖程度	个体通过随机方式获取健康信息的依赖程度
	依赖时效	个体通过随机方式获取健康信息的依赖时效

表 4　主范畴作用路径部分示例

作用路径	路径内涵	访谈文本示例
能力因素→感知因素	能力因素通过信息素养和生理机能两个基本范畴对感知因素产生直接影响。	眼睛不行了，很多视频看不清字，讲的东西也用不上。（15）
媒介因素→感知因素	媒介因素中平台特点对感知因素产生直接影响。	视频又可以看又可以听，平时一个手机走到哪里都可以刷，用着方便得很。（15）
能力因素→信息因素→感知因素	能力因素通过信息因素中信息偏好和信息质量两个基本范畴对感知因素产生间接影响。	平台有时候头条推送给我一些内容，比如说食补的，我平时也搜类似的食补视频，有些了解，觉得推送的内容我可以学着试试。（01）
心理因素→信息因素	心理因素对信息因素中信息偏好和信息效用产生直接影响。	本来就有糖尿病，更会刷些跟糖尿病相关的视频，看看能吃什么，不能吃什么，平时怎么保养。（13）
环境因素→能力因素	环境因素对能力因素中信息素养有直接作用。	我孙子教我怎么用手机，怎么看视频，就会了。（21）
感知因素→信息获取	感知因素通过感知有用性和感知易用性两个基本范畴对核心范畴信息随机获取产生直接影响。	觉得有的视频有用，平时能用上，（我）就停下来看看，下次推给我更多有用的，就看得更多。（19）

5 份原始资料进行理论饱和度检验。经讨论分析，未发现新的范畴和关系，由此得出该研究构建的理论已达到饱和。

4　研究发现

通过梳理上述关系，可得出以下故事线：感知因素作为重要因素，对信息随机获取行为直接产生作用，其他因素均在感知因素的间接作用下对信息随机获取产生影响。其中，媒介因素和信息因素直接通过感知因素的中介作用对信息随机获取产生影响。能力因素在环境因素前置变量的影响下，可对感知因素直接产生

影响，也可与心理因素在信息因素的中介作用下对感知因素产生影响。根据以上故事线，本研究构建出老年人健康信息获取行为影响因素框架，如图 1 所示。

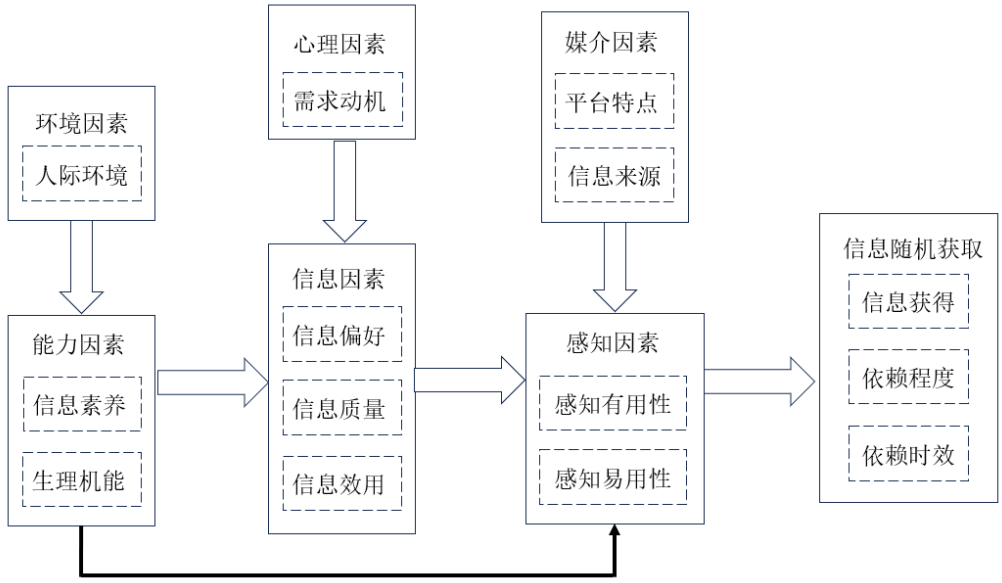


图 1　老年人健康信息获取行为影响因素框架

4.1　环境因素对老年人健康信息获取行为的影响

社会生态系统理论指出宏观环境可分为个人所处的特定微观环境、朋辈和职业所处的中观环境和外层的文化环境。本研究中，老年人群体所处的职业、朋辈环境作为老年群体所处的重要宏观环境，为其获取健康信息提供了大量物质和行为支持。本研究发现，环境因素对老年人健康信息获取行为的影响有个体差异。受人际环境影响较大的老年群体更了解健康信息的搜集、评估、采纳等全过程，操作更加熟练，获取健康信息的意愿也较高；受人际环境影响较小的老年群体则对获取健康信息的积极性较低，也更易放弃以随机方式进行健康信息的获取。

4.2　心理因素对老年人健康信息获取行为的影响

心理因素由需求动机这一基本范畴组成。动机作为决定行为的内在动力，为个体行动提供目标指向。本研究中，从老年群体自身情况看，随着社会经济的飞速发展，老年人不再仅停留于吃饱穿暖的基本生活需要上，也出于自身情况考量，更加关心自身健康，“现在比年轻时候条件好多了，多活一点多享受一点”“平时注意保养，想着以后八九十岁还能看看海逛逛街”。从家庭环境出发，当前老年群体普遍面临“421”倒金字塔家庭结构和“空巢老人”困境，老年人普遍出于“孩子还有家庭，（如果）病了拖累孩子”“平时注意点，小孩在外面少操心”考虑，为减少子女负担而获取健康

doi:10.3772/j.issn.2095-915x.2024.05.003

信息。

4.3 能力因素对老年人健康信息获取行为的影响

能力因素指老年人在获取健康信息时所需的信息素养和表现出的生理机能等。访谈结果显示，信息素养作为一种信息社会的必备适应能力，信息素养较高的老年群体能选择性获取健康信息，对健康信息进行多渠道采集、批判性评估、精确性使用。信息素养较低的老年群体则持续依赖于以随机方式进行健康信息的获取，难以判断信息的实用性和易用性。因此，信息素养较低的老年人更易陷入“信息茧房”中。随着年龄的增长，老年人不断下降的生理机能成为制约其获取健康信息的重要因素。视力和记忆力的下降致使老年人难以熟练使用电子产品，视频类健康信息成为这类老年人的首选。在国家大力实施“健康中国”战略的背景下，公共图书馆肩负着提高人们教育水平、提供健康服务的历史责任[9]。针对老年人对健康信息资源的迫切需求，公共图书馆可通过推荐相关健康书籍、电子资源，开设讲座手把手教授如何使用健康数据库，还可通过延伸服务范围，向老年群体的家人、陪护等周围人群提供健康信息咨询，以帮助老年群体脱离“数字鸿沟”的泥淖。

4.4 媒介因素对老年人健康信息获取行为的影响

媒介作为承载、传递知识的载体，承担着将信息从信息源传递到用户的责任。本研究中，媒介因素包括平台特点和信息来源等。访谈结果显示，不同于报纸杂志的时效性差和获取难度大，社交平台、短视频平台等在线平台的信息丰富性和便捷性使得老年人更加青睐在线健康信息获取。其中，在线平台内的专业医学自媒体和新闻官媒等官方权威发布最受老年群体关注。受限于互联网平台的算法推荐，老年群体无法自主决定想看的内容，这一定程度上影响了老年群体的健康信息获取体验。面对以上问题，公共图书馆与专业医疗机构、地方政府和社区合作，开展多角度、新形式的健康信息服务，向老年人科普健康信息，实行健康信息管理，以社区为单位制定健康信息服务策略[10-11]，帮助老年人了解自身健康信息需求，拓宽健康信息来源。研究发现，老年群体对来自亲友转发，日常聊天等周围环境的健康信息大多采取直接相信，不加辨析的态度，面对在线健康信息时则加以评估。

4.5 感知因素对老年人健康信息获取行为的影响

本研究中感知因素包含感知有用性和感知易用性。访谈结果显示，老年群体理解能力低，学习新事物的积极性下降，更偏向于浅显易懂的健康信息而非专业医学信息。生理机能的下降也导致老年人难以持续、准确记忆相关健康信息和查找旧有健康信息，完整、简短的健康信息更受到老年人的偏爱。信息运用难度也成为了老年群体利用健康信息的重要评判标准。此外，由于部分健康信息不够贴合老年群体实际需要，致使信息真实性分辨能力较低的老年人转而相信匹配度高的伪健康信息和医疗广告。

4.6 信息因素对老年人健康信息获取行为的影响

本研究中信息因素通过影响老年群体感知有用性和感知易用性对其健康信息获取行为产生作用。访谈结果显示，由于老年群体普遍怀有“不生大病不进医院”的思想，因此在信息偏好上，相比专业类医学消息，其更关注养生类健康信息。此外，老年人多年来的生活经验和节约金钱的思想也致使其关注民间偏方类信息。信息效用即老年人获取健康信息时取得的功效。老年群体为缓解年龄增长带来的紧张感，多借助获取健康信息缓解心情。部分老年人退休后仍从事某项工作，少有空闲去医院进行检查，因此获取健康信息以节约时间。研究发现，老年群体对信息的可靠性、权威性十分重视，并据此决定是否继续获取此类健康信息。但随着互联网信息井喷式增长，伪官方伪权威信息和各种反转新闻层出不穷，老年群体信息评估能力较弱，极易导致自身面临信息过载的问题，加剧自身信息焦虑，以至于个人数字囤积，最终降低健康信息获取的质量。

5 结语

本研究发现，信息随机获取已成为老年人健康信息获取的新方式。老年人健康信息获取行为受环境因素、心理因素、能力因素、媒介因素、感知因素和信息因素的共同影响。其中，感知因素对信息随机获取直接产生作用，其他因素均在感知因素的间接作用下对信息随机获取产生影响。能力因素在环境因素前置变量的影响下，可对感知因素直接产生影响，也可与心理因素在信息因素的中介作用下对感知因素产生影响。本研究阐释了老年人健康信息获取行为的影响因素及其之间关系，帮助优化老年人健康信息服务。本研究的样本对象选择可能有一定地域局限，研究结果可能更符合同样经济情况的地区，研究结果的普适性有待进一步观察。后续研究可进一步分地区、分类别地分析其在健康信息获取行为中的影响因素差异。

参 考 文 献

[1] 王琳 , 杨莹 , 朱华健 , 等 . 从“小世界”到“同心生活圈”：我国老年妇女健康信息搜寻行为 [J]. 图书馆论坛 , 2024, 44(3): 1-12.

[2] 郭明蓉 , 冯春 , 陈莉 . 乡镇老年人健康养老信息需求及获取途径调查分析 [J]. 智慧健康 , 2019, 5(16): 35-40.

[3] National Network of Libraries of Medicine. Healthy aging: Con-necting older adults to health information [EB/OL]. (2019-01-10) [2024-01-12]. https://nnlm.gov/class/ healthy-aging-jan19.

[4] 李熠煜 , 汪建冲 . 疫情防控常态化背景下老年人健康信息获取中的“堕距”与“融合”——基于CMIS模型的分析 [J]. 图书馆研究 , 2022, 52(3): 1-10.

[5] 乔欢 . 信息行为学 [M]. 北京 : 北京师范大学出版社 , 2010: 12.

[6] 朱姝蓓 , 邓小昭 . 老年人网络健康信息查寻行为影响因素研究 [J]. 图书情报工作 , 2015, 59(5): 60-67, 93.

[7] 李锴 . 新媒体环境下老年群体搜寻健康信息的路径与困境 [D]. 郑州 : 河南工业大学 , 2022.

[8] 戴艳清 . 社区图书馆为老年人提供健康信息服务初探 [J]. 图书馆论坛 , 2011, 31(4): 138-140, 146.

[9] 刘一鸣 , 朱萍萍 . 公共图书馆老年人健康信息服务策略研究 [J]. 国家图书馆学刊 , 2022, 31(1): 84-94.

[10] 彭丽徽 , 张芊慧 . 中老年人健康信息规避行为影响因素与关联路径研究——基于扎根理论的探索性分析 [J]. 图书馆学研究 , 2022(1): 77-86, 76.

[11] 刘一鸣 , 李露 . 澳大利亚公共图书馆老年人健康信息服务项目的研究与启示 [J]. 情报资料工作 , 2023, 44(4): 85-95.

doi:10.3772/j.issn.2095-915x.2024.05.004

粤苏沪科技数据要素市场化配置分析及启示

开放科学（资源服务）标识码（OSID）

乔振[1] 荀玥婷[2]

1. 山东省科学技术情报研究院 济南 250101；
2. 山东省科技发展战略研究所 济南 250121

摘要：[目的 / 意义] 探讨广东、江苏和上海科技数据要素市场化配置现状与经验，为山东开展工作提供借鉴。[方法 / 过程] 基于构建的科技数据要素市场化配置分析框架，从总体规划、科技数据要素、科技数据载体、数据主体等方面梳理三省市科技数据要素市场化配置现状和相关典型探索，总结三省市科技数据市场化探索的路径和模式。[结果 / 结论] 三省市进行了科技数据要素配置探索，形成了典型模式，示范性和可借鉴性强。与其相比，山东存在目标不明确、政策标准不完善等问题，应从明确建设目标、完善政策标准、推动数据集聚和应用等方面加强相关工作。

关键词：科技数据；数据要素；市场化配置；数据开放；数据共享

中图分类号：G35；F49

Analysis and Enlightenment on Market-oriented Allocation of Scientific and Technological Data Elements in Guangdong, Jiangsu, and Shanghai

QIAO Zhen[1] XUN Yueting[2]

1. Shandong Institute of Scientific & Technical Information, Jinan 250101, China;
2. Institute of Science and Technology for Development of Shandong, Jinan 250121, China

Abstract: [Objective/Significance] The current situation and experience of market-oriented allocation of scientific and technological data elements in advanced provinces and cities such as Guangdong, Jiangsu and Shanghai were explored, providing reference for Shandong's work. [Methods/Processes] Based on the constructed analysis framework for market-oriented allocation of scientific and technological data elements, the current situation and relevant typical explorations of market-oriented allocation of scientific and technological data in Guangdong, Jiangsu, and Shanghai was summarized from the aspects of overall planning, scientific and technological data elements, scientific and technological data carriers, data subjects,

基金项目 山东省重点研发计划（软科学）“数据要素市场化配置视角下粤苏沪鲁四省科技数据管理比较研究”（2023RKY04009）。

作者简介 乔振（1985-），硕士，副研究员，主要研究方向为信息资源管理，科技咨询与科技管理；荀玥婷（1986-），硕士，高级经济师，主要研究方向为科技管理与创新，E-mail：542962558@qq.com。

引用格式 乔振 , 荀玥婷 . 粤苏沪科技数据要素市场化配置分析及启示 [J]. 情报工程 , 2024, 10(5): 38-50.

and the characteristics and models of market-oriented exploration of scientific and technological data in the three provinces and cities were summarized. [Results/Conclusions] Guangdong, Jiangsu and Shanghai have carried out explorations on the allocation of scientific and technological data elements, forming a work path with strong demonstration and reference value. Compared with them, there are problems such as unclear goals and incomplete policy standards of Shandong. Therefore, efforts should be made to strengthen related work by clarifying construction goals, improving policy standards, and promoting data aggregation and application.

Keywords: Scientific and Technological Data; Data Elements; Market-oriented Allocation; Data Opening; Data Sharing

引言

数据要素是各类活动中产生的信息、数据和知识的集合，其作为一种非物态的经济资源[1-2]，具有不可估量的资源配置基础能力、价值发掘与增值能力，对生产效率的提高也具有乘数作用[3-4]。科技数据是数据要素的重要组成部分，广东省、江苏省和上海市在科技数据管理和市场化配置方面做了大量工作，取得一定效益，具有一定的示范性。广东省科技政务数据治理已处于数据治理前期阶段[5]，在科技数据集聚，特别是在科学数据汇交方面做了大量探索，建设的广东科技基础平台通过向社会提供共享服务，实现超过 1500 亿元的经济效益[6]。江苏集聚了大量科技数据并开展市场化服务，2022 年挖掘企业技术需求 2570 项，服务企业 4700 余家，促成技术交易成交金额超 3.7 亿元，参与数据交易企业数量达 1056 家[7]。上海推动科技政策等科技数据的深度开发利用，建成的“政策北斗”累计访问量超过一百万人次[8]。本研究基于构建的数据要素市场化配置分析框架对粤苏沪现状进行了探讨，梳理了相关典型探索，总结相关启示，就山东省科技数据要素市场化配置提出针对性建议。

1 概念和分析框架

1.1 基本概念

1.1.1 科技数据

综合相关研究[9-14]，本研究将科技数据定义为省级科技行政部门及其所属单位等收集、管理的与科技活动相关的各类信息、数据和知识的集合，具体包括科技文献数据、科技政务数据、网络科技数据。同时，根据相关定义[5, 15-16]，本研究认为公共数据包含了政务数据，科技政务数据是政务数据的一种，科技数据包括科技政务数据和其他科技数据，因此科技数据和公共数据属于交叉关系，公共数据相关政策同样适用于科技数据。

1.1.2 科技数据要素市场化

根据相关定义[17-19]，科技数据要素市场化是在大量的科技数据集聚和供给前提下，通过数据授权运营、数据特许开发应用、数据资产凭证、数据共享和开放等方式推动科技数据要素在不同主体间流通，推动科技数据要素高水

doi:10.3772/j.issn.2095-915x.2024.05.004

平应用，充分发挥科技数据价值，赋能经济社会发展。

1.2 科技数据要素市场化配置分析框架

结合数据生命周期理论、市场化配置理论，本研究构建了基于数据生命周期的科技数据要素市场化配置分析框架，包括总体规划、科技数据要素、科技数据主体、科技数据载体、科技数据制度、科技数据监管。

科技数据主体是参与科技数据供给、流通和应用的主体。科技数据载体是承载科技数据配置的各类载体，如数据管理平台。科技数据监管内容可包括科技数据要素市场登记备案、流通安全与秩序、信用体系等内容[20]。

2 粤苏沪科技数据要素市场化配置现状

2.1 总体规划

围绕数据要素市场化配置，广东省提出了打造“理念先进、制度完备、模式创新、高质安全”的数据要素市场体系和市场化配置改革先行区的目标[21]。针对科技数据，明确要加强科技数据库 / 平台建设，构建具有国际影响力的科技信息高端交流平台[22]。

江苏省提出推动江苏成为数据要素高效配置先导区[23]，省科学技术行政部门开展科技资源的统筹集成工作[24]，建设省科技资源统筹服务中心，推进科技资源共享平台建设，支持建设科学数据中心[25-26]。

上海提出“建设具有国际影响力的数据要素配置枢纽节点和数据要素产业创新高地”，到 2025 年，基本建成数据要素市场体系[27]。科技数据方面，上海明确要“打造国际化的上海科技资源大数据中心，建设全球科技数据信息资源枢纽”[28]。

2.2 科技数据要素

广东、江苏和上海的科技数据要素均包括科技政务数据、科技文献数据和网络科技数据。科技政务数据中，一般包含了科技政策数据、科技计划项目数据、科技载体数据、科技成果、科技服务数据等，广东和江苏还将科学数据纳入其中，上海则有技术预见数据和科普数据。科技文献数据一般为购买的国内外期刊论文、专利、标准等。网络科技数据一般为媒体聚焦、互动交流数据等。

2.3 科技数据载体

三省市科技厅 / 科委均建有官方网站和微信公众号，提供科技政策和网络科技数据服务，均针对科技计划项目管理、科技报告建设相应平台。此外，广东还建设科技数据发布应用平台和科技资源共享服务平台，使科技计划项目数据和科学数据流通并进行供给。江苏省科技资源统筹服务中心平台、上海科技创新资源数据中心作为供给和流通载体，对科技文献、科技政务数据进行了一定集成，并提供政策分析等服务。此外，三省市还分别通过“开放广东”和广东政务服务网、江苏政务服务、上海市公共数据开放平台开展部分科技数据的流通。

2.4 科技数据主体

同一数据主体涉及不同类型的科技数据其

所发挥的作用不同；同一数据主体涉及同一类型的科技数据，在面向不同对象时，其发挥的作用也不同。如广东省科技厅作为数据供给、流通和应用主体提供科技政务数据服务，参与广东省各类科技活动的个人、法人单位是科技政务数据的供给方和数据应用方，具体见表 1。

表 1　广东省科技数据主体

序号	科技数据主体	数据内容	主体类型
1	广东省科技厅	科技政务数据	供给
		网络科技数据	供给
2	广东省科技创新监测中心	科技政务数据	供给
3	广东省科技基础条件平台中心	科技政务数据	供给；流通
4	广东省政务服务数据管理局	科技政务数据	流通
5	广东省科学技术情报研究所	科技文献数据等	供给；流通；应用
6	参与广东省各类科技活动的个人、法人单位	科技政务数据	供给；应用
		科技文献数据	应用
		网络科技数据	应用

江苏省科技厅是科技政策法规、科技计划管理数据及网络科技数据的主要采集方和提供方。江苏省科技情报研究所是科技成果、科技报告等科技政务数据的供给和应用主体，同时，面向社会提供科技查新、知识产权评议等服务，是科技文献数据的应用主体。江苏省科技资源统筹服务中心推进江苏全省科技资源数据的集成、展示与共享，面向社会提供科技政务数据的供给和科技文献数据的供给、流通和应用。江苏省政务服务管理办公室 / 江苏省大数据管理中心提供各类公共数据的汇聚、流通。个人、法人单位等则是各类科技政务数据的产生及应用主体，是科技文献和网络科技数据的主要应用主体。

上海市科委及其直属事业单位（如上海市研发公共服务平台管理中心）是各类科技政务数据的供给主体，上海市科技创业中心提供相关科技文献数据的供给、流通和应用，上海市大数据中心和上海数据集团公司主要提供公共数据的流通和应用，个人和法人单位是各类科技数据的应用主体。

2.5　制度标准建设

广东已出台一系列典型政策法规，详见表 2。截至 2024 年 3 月，除《广东省科技创新“十四五”规划》，尚未出台与科技数据直接相关的政策，但在政策覆盖范围、政策涉及内容方面都有了较大的突破。覆盖范围上，广东省，广州、深圳等市均出台了一系列相关政策；内容方面，除宏观管理内容，还涉及了数据确权（如政务数据资源所有权归政府所有）、数据资产登记、数据开放共享、数据流通交易、技术安全、数据监管及数据经纪人，可以说涵盖了数据生命全链条管理。

广东省围绕公共数据开放、平台建设、数据安全制定了 15 项地方标准和团体标准，见表 3，涉及了科技政务数据、科技大数据，起草单位部分为省科技厅所属单位，具有一定代表性。

doi:10.3772/j.issn.2095-915x.2024.05.004

表 2　广东省数据要素市场相关政策

序号	政策名称
1	《广东省国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》
2	《广东省数字经济促进条例》
3	《广东省公共数据管理办法》
4	《广东省公共数据开放暂行办法》
5	《广东省政务数据资源共享管理办法（试行）》
6	《广东省数字政府改革建设“十四五”规划》
7	《广东省数据要素市场化配置改革行动方案》
8	《广东省数字政府改革建设 2023 年工作要点》
9	《广东省科技创新“十四五”规划》
10	《广东省科学技术厅关于优化调整广东省科技报告管理工作的通知》
11	《广东省科学技术厅关于深入推进重大科研基础设施与大型科研仪器开放共享的若干措施》
12	《广东省数据流通交易管理办法（试行）》《广东省数据资产合规登记规则（试行）》《广东省数据流通交易监管规则（试行）》《广东省数据经纪人管理规则（试行）》《广东省数据流通交易技术安全规范（试行）》征求意见稿

表 3　广东省数据要素相关标准

序号	标准名称	标准类别
1	《电子政务数据资源开放数据管理规范》	地方标准
2	《电子政务数据资源开放数据技术规范》	地方标准
3	《政务信息资源标识编码规范》	地方标准
4	《政务信息共享平台接入规范》	地方标准
5	《科技平台建设规范》	地方标准
6	《公共数据安全要求》	地方标准
7	《大型科学仪器设施共享服务平台运行规范》	地方标准
8	《大型科学仪器设施共享服务平台数据交换规范》	地方标准
9	《广东省科技政务大数据应用平台数据采集管理规范》	团体标准
10	《广东省科技政务数据交换接口管理规范》	团体标准
11	《科技热点数据分析应用标准规范》	团体标准
12	《广东省科技政务大数据应用平台科技专家数据规范》	团体标准
13	《科技大数据平台业务规范》	团体标准
14	《科技大数据平台数据仓库开发指南》	团体标准
15	《科技大数据平台数据仓库建设规程》	团体标准

《江苏省公共数据管理办法》对公共数据供给、共享、开放、利用与安全等提出了明确要求。《江苏省数字经济促进条例》围绕公共数据资源开发利用模式、运营机制和应用等提出要求。新修订的《江苏省科学技术进步条例》对科技资源统筹机构、资源统筹集成、共享和

使用、收费等进行了规范。《江苏省“十四五”科技创新规划》对科技数据提出规划。《江苏省科技资源统筹服务管理办法（试行）》《江苏省科技资源统筹服务考核评价实施细则》等系列政策对江苏全省科技资源统筹集成与开放服务，科技资源统筹服务体系、考核等进行了规定，特色比较明显。标准规范方面，江苏虽出台了一些公共数据和政务数据地方标准，但这些标准为市级层面，如南京市的《政务数据安全管理指南》、泰州市的《政务数据安全分类分级指南》、无锡市的《公共数据分类分级实施指南》等。

《上海市数据条例》从法律的角度明确了上海数据法治管理体系，围绕公共数据共享和开发、授权运营等方面提出了创新举措。《上海市建设具有全球影响力的科技创新中心“十四五”规划》提出加快完善科技数据管理机制，将良好的数据管理纳入科技计划项目管理要求[28]。《上海市推进科技创新中心建设条例》明确市科学技术部门建立公共科技创新资源共享机制，推进重要科技信息、科学数据、科技报告等开放共享[29]，《上海市公共数据开放暂行办法》《上海市公共数据共享实施办法（试行）》对公共数据开放和共享进行了规范。标准方面，上海主要围绕公共数据制定了《公共数据资源目录》等6项地方标准，未见科技数据相关标准。

2.6 数据要素市场监管

《广东省公共数据管理办法》《广东省数据要素市场化配置改革行动方案》《广东省数据要素市场化配置改革重点任务分工表》《广东省数据流通交易管理办法（试行）》（征求意见稿）等对监管责任部门、职责与分工和监管内容等进行了规定，见表4。

表4 广东省科技数据市场监管

序号	数据监管主体	数据监管内容
1	广东省政务服务数据管理局（公共数据主管部门）	公共数据管理工作等；数据交易平台；数据经纪人监管；数据交易监管；数据交易流通安全监管
2	广东省科技厅（行业主管部门）	数据安全
3	广东省科技厅确定的公共管理和服务机构	公共数据日常监控
4	广东省商务厅	数据交易监管
5	广东省市场监管局	数据交易监管
6	网信、通信管理等部门	公共数据安全

《江苏省公共数据管理办法》《江苏省科技资源统筹服务管理办法（试行）》等规定了江苏省委网络安全和信息化委员会办公室负责网络安全协调，江苏省科技厅负责数据管理监管、安全监管、管理单位成本核算和服务收费标准监督，江苏省政务服务管理办公室统筹监管公共数据管理和日常监管，江苏省公安厅等负责数据安全监管，各科技数据提供部门负责数据安全监管和国有资产监管。

《上海市数据条例》《上海市公共数据和一网通办管理办法》《上海市公共数据开放暂行办法》等规定了各类监管主体、监管内容，具体见表5。

doi:10.3772/j.issn.2095-915x.2024.05.004

表 5　上海市科技数据市场监管

序号	数据监管主体	数据监管内容
1	上海市网信办	相关监管，数据开放安全
2	上海市公安、国家安全机关	数据安全
3	上海市政府办公厅	日常公共数据管理工作监督 公共数据开放
4	上海市大数据中心	公共数据质量监督 被授权运营主体实施日常监督 数据开放平台
5	上海市科委等数据开放主体	公共数据利用情况监管 行业监管

2.7　科技数据要素市场配置流程与模式

结合《广东省公共数据管理办法》《广东省公共数据开放暂行办法》，绘制了广东省科技数据市场配置流程图，如图 1。其中，公共管理和服务机构为广东省科技厅及相关执行单位。具体流程为科技厅等收集、汇聚参与科研活动的个人、法人等产生的数据，分别向广东政务服务网、“开放广东”、政务数据共享交换平台提供数据和目录，通过该部分平台供数据使用者使用。数据使用者利用公共数据开展科学研究等活动，相关活动产生的数据产品或者数据服务可通过数据交易平台开展数据交易。同时，科技厅也通过广东省科技数据发布应用平台等自建平台向数据使用者提供数据服务。针对科技文献数据、网络科技数据，广东一般通过自建平台开展共享和利用。

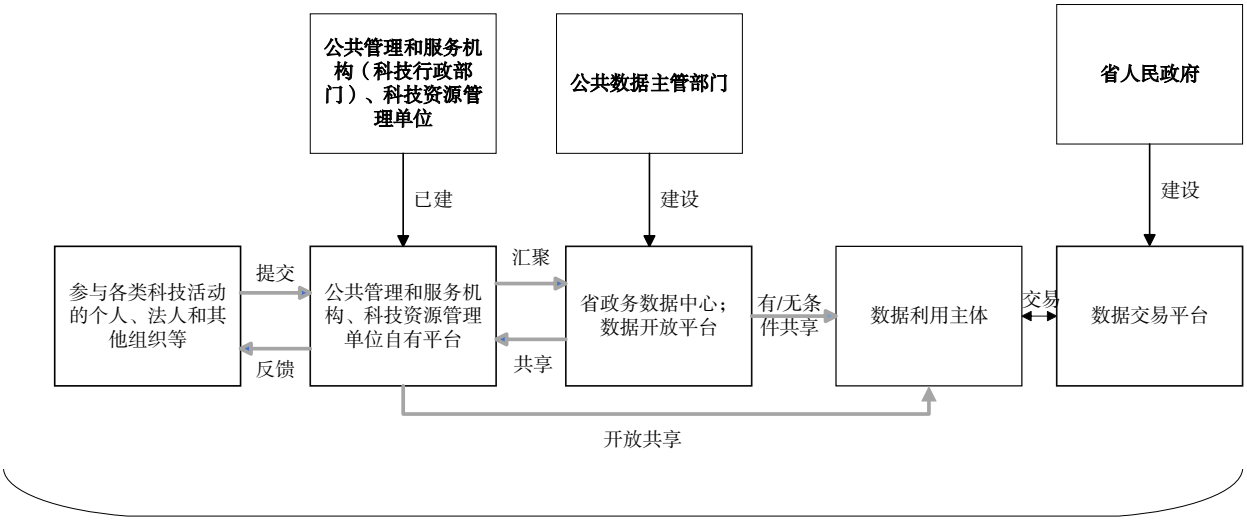


图 1　广东省科技数据市场配置流程与模式

根据《江苏省公共数据管理办法》《江苏省科技资源统筹服务管理办法（试行）》，绘制了江苏省科技数据要素市场配置流程与模式，如图 2。其中，公共管理和服务机构由江苏省政务服务管理办公室会同有关主管部门确定。配置流程上，各类用户通过省科技厅等公共管理和服务机构开发建设的平台办理业务并产生各类科技政务数据，省科技厅等进行加工、处理，部分结果数据通过自建平台反馈给个人、法人等，部分数据目录汇交至江苏政务服务网等公共数据平台、省统筹平台（江苏省科技资源统筹服务云平台），通过该部分平台面向其他公共管理和服务机构、数据利用主体实现共享和开放。

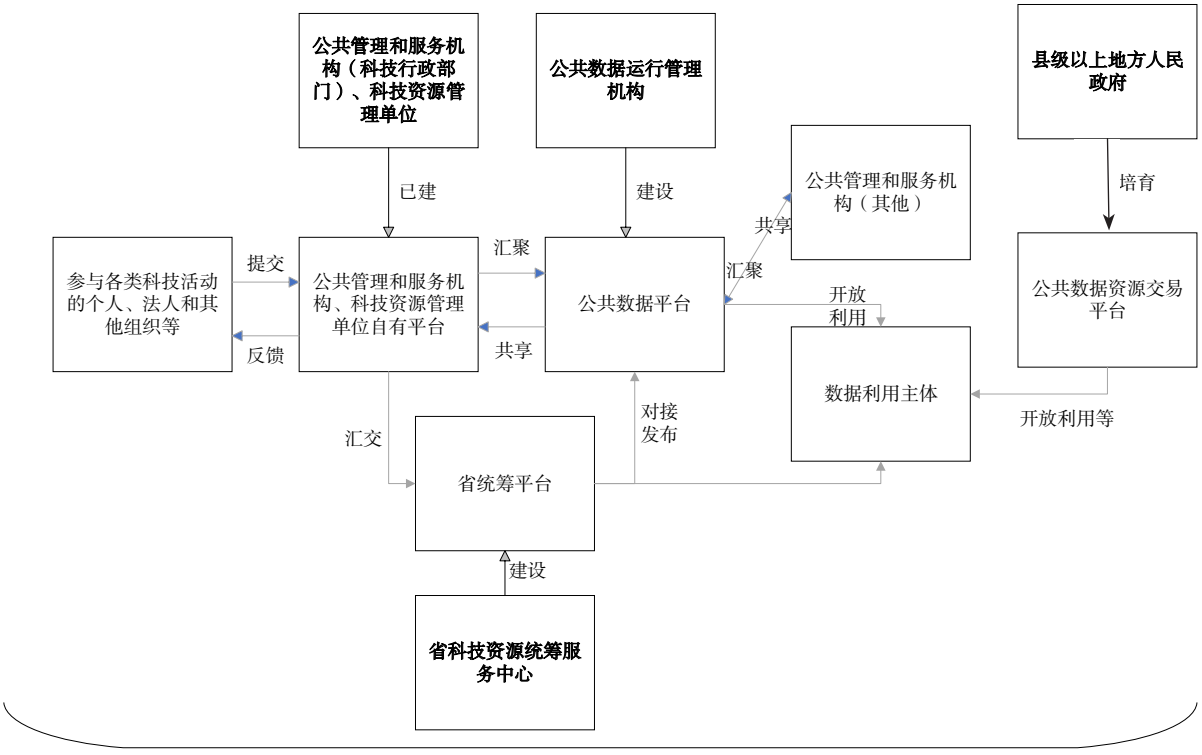


图 2　江苏省科技数据要素市场配置流程与模式

根据《上海市数据条例》《上海市公共数据开放暂行办法》《上海市公共数据和一网通办管理办法》等，绘制上海市科技数据配置流程与模式，如图 3。科技政务数据要素配置流程上，个人等通过参与科技活动产生并提交各类数据，公共管理和服务机构（科技）对科技政务数据进行归集、校核和整合，通过市大数据资源平台开展数据服务，其中，面向公共管理和服务机构，通过共享交换子平台开展数据共享服务；面向社会，通过开放子平台提供数据开放服务。公共管理和服务机构（科技）也通过自建平台面向社会提供科技政务数据浏览、科技政务数据收费增值服务（如政策可视化分析、各类科创数据服务）。科技文献数据主要由公共管理和服务机构（科技）采购后通过自建平台开展服务，网络科技数据通过自建平台开展共享。

doi:10.3772/j.issn.2095-915x.2024.05.004

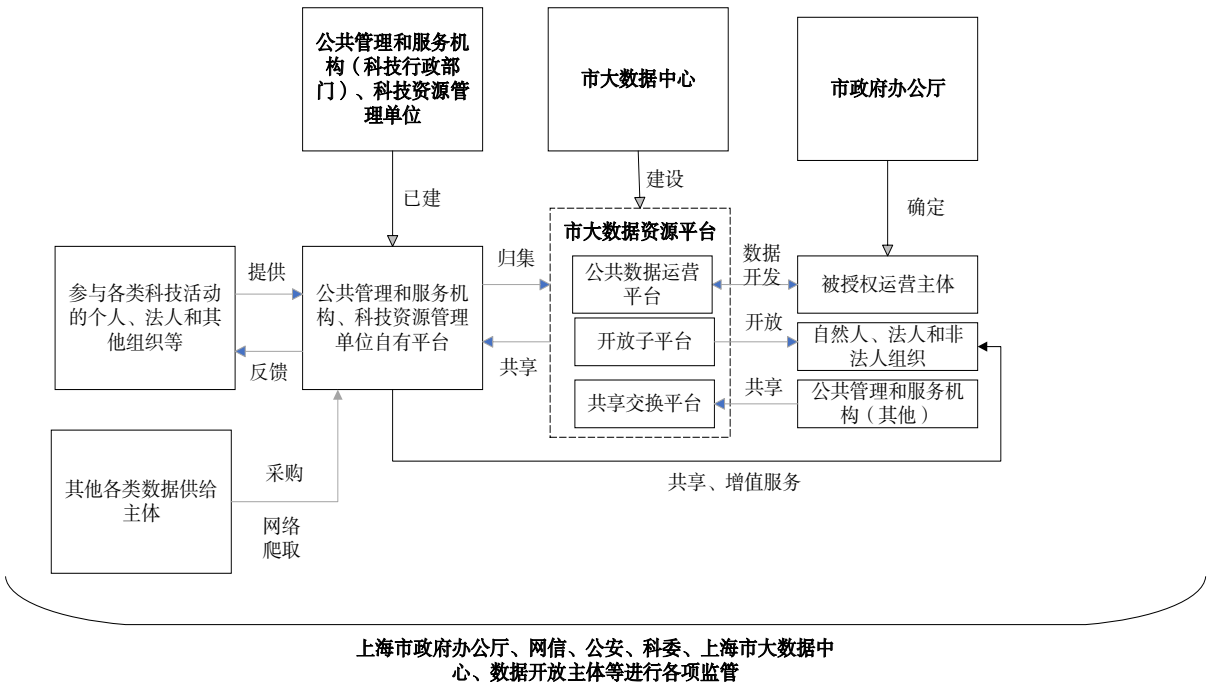


图 3　上海市科技数据要素配置流程与模式

配置方式上，三省市均采取了开放和共享的方式，其中，开放包括无条件开放、有条件开放和不予开放，共享包括无条件共享、有条件共享和不予共享。此外，广东率先通过公共数据资产凭证推动公共数据市场化配置，在高科技试点场景颁发了公共数据资产登记证书，这为开展科技数据资产登记提供了范例。江苏通过数据深加工等方式挖掘数据价值，采取市场化收费的形式向数据利用主体提供数据服务。上海还开展了授权运营探索，由上海市政府办公厅确定公共数据授权运营主体，被授权运营主体即上海数据集团有限公司依托统一规划的公共数据运营平台，实施数据开发利用，部分机构通过提供科技数据增值服务的方式进行市场化收费运营。

3　典型探索

3.1　广东省科学数据中心

广东省将科学数据中心定位为“国家科学数据中心体系的补充，与国家体系相辅相成”，探索采取“科学数据服务管理平台 + 科学数据中心”的建设模式，打造广东省科学数据服务管理平台和若干领域科学数据中心 [30]。

广东省科技厅通过广东省平台基地及科技基础条件建设项目支持科学数据中心建设，已建设服务管理中心 1 家、省科学数据中心 11 家，每家给予 300 万元立项经费支持。开展了国家科学数据中心在粤分中心建设，纳入广东省科学数据中心体系，并给予 100 万元立项经费支持。

建设有广东省科技资源共享服务平台（粤科汇），平台已汇集 10391 条科学数据目录，涵盖林业、基因组、地理、中医和农业等领域。

3.2 江苏省科技资源统筹服务中心

江苏省科技资源统筹服务中心是江苏省建设的省科学技术资源统筹服务专业机构，汇聚了科技载体资源、科技条件资源、科技成果资源，以及科技服务资源等科技数据，数据量超 5 亿条 [31]。

载体方面，除了统一建设的江苏省科技资源统筹服务云平台，还包括科技资源管理单位自行建设的科技资源在线服务平台，这些平台要与省资源统筹服务云平台互联互通。

江苏省科技资源统筹服务中心建设了数据平台并向社会提供统一检索、浏览和下载等服务。集合各类政策，研发“策立得”，提供政策匹配、企业体检等。同时，允许按照成本补偿和非盈利原则收取费用，服务收入纳入单位预算 [32]。

3.3 上海科技创新资源中心

作为科技数据类新型研发机构和上海新型研发与转化功能型平台 [33]，上海科技创新资源中心着力于科技资源数据的集聚融合、运行评估与服务 [34]，通过采购、合作协议等形式集聚了上海市科委等科技部门提供的科技政务数据、各类文献与服务资源及其他各类科技数据。

载体方面，上海科技创新资源数据中心建立了上海大型仪器设施信息服务数据库、研发基地资源数据库等平台。配置服务方面，通过注册登录、购买服务、服务外包、签署定制合同等，向各类用户提供科技数据检索服务、咨询服务、科技中介培育一站式科技资源服务 [33]。

4 粤苏沪鲁比较分析

山东围绕科技数据管理与市场化配置也开展了一些工作，但与粤苏沪相比，仍存在一定短板。

总体规划方面，山东虽提出了“打造形成全国数据要素市场化配置改革先行区”的目标，但对科技数据的总体规划和目标不明确，部分政策中仅提出加强科技计划数据管理，与三省市明确的目标和举措相比，仍需改进。

科技数据要素集聚方面，科学数据是科技数据的重要组成，与广东相比，山东尚未开展科学数据汇交，科学数据中心体系建设未有进展；同时，科技数据的开发利用也不足。

制度建设方面，广东和江苏对政务数据、公共数据等权属进行了界定，山东尚未涉及。标准方面，山东尽管出台了与公共数据、政务数据资源、政务服务平台和数据产品登记等相关的地方标准、工程标准和团体标准，但未见科技数据相关标准，与广东相比仍有短板。

科技数据主体和载体方面，江苏和上海均通过相对统一的数据主体建设了集中的数据平台，而山东科技数据分散于山东省科技云平台、科技报告服务系统、科技档案管理平台等多个数据载体，平台间互联互通不足。

市场化配置方面，山东虽通过公共数据开放网等开展了科技数据的开放和共享，但在数据授权运营、数据开发和市场化交易等方面探索不足。

doi:10.3772/j.issn.2095-915x.2024.05.004

5 启示与建议

5.1 启示

5.1.1 明确科技数据要素市场建设目标

广东、江苏和上海围绕数据要素市场、公共数据和科技数据进行了谋划，提出了数据要素市场建设目标，对公共数据提出了一定措施和任务，对科技数据进行了布局。尤其是上海，提出了建设国际化、全球性科技资源数据中心和枢纽的目标，对科技数据要素市场建设引领作用更强。

5.1.2 制定科技数据政策和标准

三省市出台了数字经济条例或数据条例，从法律高度确定了数据要素市场或公共数据相关内容。公共数据方面，江苏和广东出台了针对公共数据的管理办法，此外，广东还针对公共数据开放专门出台办法，上海更是增加了公共数据共享办法。科技数据方面，广东、上海将科技数据纳入科技创新十四五规划，江苏还将其纳入新修订的《江苏省科学技术进步条例》，专门出台了科技资源统筹服务办法和考核细则。

标准方面，上海主要围绕公共数据资源目录和公共数据交换工作规范制定了标准。广东除公共数据和政务数据地方标准外，还围绕科技政务大数据和科技大数据发布了团体标准，可为相关工作提供借鉴。

5.1.3 建设多元科技数据主体和载体

三省市科技数据载体和科技数据主体呈现多元化特点。公共数据开放平台是实现数据管理和流通功能的重要基础设施，三地均建立了公共数据开放平台，广东专门建立了广东省科技数据发布应用平台，江苏和上海对科技数据进行了一定的统筹，分别建设了江苏科技资源统筹服务中心和上海科技创新资源中心。

5.1.4 推动科技数据要素集聚和应用

三省市均聚集了一定的科技数据要素，广东在科学数据汇聚方面具有一定的特色，江苏和上海对科技数据进行了一定的统筹。其次，实现了科技数据不同程度的开发与应用。如江苏和上海就科技政策提供政策可视化分析、政策匹配等深度服务。上海等还建设了长三角科技资源共享服务平台，推动跨区域科技数据的共享和利用。

5.1.5 明确数据要素市场配置流程与模式

三省市对公共数据的市场配置流程和方式进行了明确，也具有一定的共性。如在配置方式上，面向不同主体实行共享和开放。广东对公共数据资产登记进行了试点。上海设置了公共数据授权运营即基于应用场景的授权，并将其纳入相关数据条例，同时，上海数据集团有限公司进行相关实践探索。针对科技文献数据和网络科技数据，三地一般实行免费共享。

5.1.6 加强数据要素市场监管

三省市均对数据要素市场监管进行了谋划，在监管主体、监管内容等方面进行了明确，但存在各项内容分散于多项政策，聚焦不足的问题。

5.2 建议

粤苏沪三省在科技数据要素市场化进行了探索，形成了工作路径。山东应充分借鉴粤苏沪三省市经验，围绕规划不明确、科技数据集聚不全、政策标准不完善、主体和载体不统一、

数据登记和市场化交易不足等短板，从以下几个方面开展工作：

（1）明确总体规划和建设目标。在“打造形成全国数据要素市场化配置改革先行区”目标的基础上，结合科技数据集聚和管理现状，提出具体的科技数据市场化目标，根据目标做好相关规划。

（2）推进科技数据要素集聚。加快山东省科学数据管理平台及省级科学数据中心体系等主体和载体建设，推动全省科学数据的汇交；加强对已有科技数据的挖掘和利用，发挥科技数据对科技创新的乘数效应。

（3）进一步完善政策标准。从法律、制度上做好数据确权[35]，从根本上解决数据产权问题。充分发挥山东数据要素创新创业共同体作用，探索出台科技数据相关标准，推动科技数据要素市场标准化建设。

（4）探索多元化配置方式。在数据开放和共享的基础上，加强与国家及各省市数据交易所的合作，同时，利用山东数据交易平台，探索开展数据知识产权登记、授权经营、市场化交易，积极培育数商，完善最大化释放数据价值的实现路径[36]。

参 考 文 献

[1] 陆岷峰 . 新发展格局下数据要素赋能实体经济高质量发展路径研究 [J]. 社会科学辑刊 , 2023(2): 143-151.

[2] 王颂吉 , 李怡璇 , 高伊凡 . 数据要素的产权界定与收入分配机制 [J]. 福建论坛 (人文社会科学版), 2020(12): 138-145.

[3] 王谦 , 付晓东 . 数据要素赋能经济增长机制探究 [J]. 上海经济研究 , 2021(4): 55-66.

[4] 刘鹤 . 坚持和完善社会主义基本制度 [N]. 人民日报 , 2019-11-22(006).

[5] 邹倩瑜 , 郑宏松 , 胡意 . 基于中台架构的科技政务数据治理模式研究 —— 以广东为例 [J]. 科技管理研究 , 2021, 41(24): 169-176.

[6] 邵玉昆 . 科技数据资源的开放共享机制研究 [J]. 科技管理研究 , 2019, 39(13): 177-181.

[7] 庄雷 . 激活科技数据要素赋能新质生产力发展 [J]. 群众 , 2024(5): 60-61.

[8] 李宁 , 顾玲琍 , 杨耀武 . 上海科技政策智慧服务研究实践进展及建议 —— 基于上海科技“政策北斗”监测数据与调研 [J]. 科技中国 , 2022(5): 34-39.

[9] 曾文 , 车尧 . 科技大数据的情报分析技术研究 [J]. 情报科学 , 2019, 37(3): 93-96.

[10] 李辉 , 曾文 , 谭晓 , 等 . 科技大数据资源平台建设研究 [J]. 科技情报研究 , 2022, 4(1): 71-77.

[11] 符宁 . 科技政务大数据管理与挖掘平台设计 [J]. 软件导刊 , 2021, 20(5): 86-91.

[12] 钱力 , 谢靖 , 常志军 , 等 . 基于科技大数据的智能知识服务体系研究设计 [J]. 数据分析与知识发现 , 2019, 3(1): 4-14.

[13] 刘召栋 , 周亿城 . 科技大数据资源及分类分级研究 [J]. 科技与创新 , 2021(18): 123-126.

[14] 张勇 , 苏学 , 谢振峰 . 面向科技大数据的元数据仓储建设实践探索 [J]. 情报工程 , 2020, 6(6): 84-96.

[15] 邓峰 . “放管服”环境下科技政务大数据平台研究 [J]. 科技创业月刊 , 2020, 33(10): 75-77.

[16] 孟庆国 . 人民网权威解读：创新管理机制，推动数据资源体系开放共享 [EB/OL]. (2022-07-01) [2023-09-18]. http://finance.people.com.cn/n1/2022/0701/c1004-32463383.html.

[17] 孔艳芳 , 刘建旭 , 赵忠秀 . 数据要素市场化配置研究 : 内涵解构、运行机理与实践路径 [J]. 经济学家 , 2021(11): 24-32.

[18] 戚聿东 , 刘欢欢 . 数字经济下数据的生产要素属性及其市场化配置机制研究 [J]. 经济纵横 , 2020(11): 63-76, 2.

[19] 张会平 , 马太平 . 政府数据市场化配置：概念内涵、方式探索与创新进路 [J]. 电子科技大学学报 (社科版), 2022, 24(5): 1-8, 17.

[20] 王伟玲 . 中国数据要素市场体系总体框架和发展路径研究 [J]. 电子政务 , 2023(7): 2-11.

[21] 广东省人民政府 . 广东省人民政府关于印发广东

doi:10.3772/j.issn.2095-915x.2024.05.004

省数据要素市场化配置改革行动方案的通知 [J]. 广东省人民政府公报 , 2021(20): 10-15.

[22] 广东省人民政府 . 广东省人民政府关于印发广东省科技创新“十四五”规划的通知 [J]. 广东省人民政府公报 , 2021(29): 3.

[23] 江苏省工信厅 . 我省发布“十四五”大数据产业发展规划 [EB/OL]. (2021-09-03) [2023-09-11]. http://www.jiangsu.gov.cn/art/2021/9/3/art_60085_9998123.html.

[24] 江苏省人民代表大会常务委员会 . 江苏省科学技术进步条例 [N]. 新华日报 , 2023-02-06(006).

[25] 江苏省人民政府 . 江苏省人民政府关于印发江苏省国民经济和社会发展第十四个五年规划和二〇三五年远景目标纲要的通知 [J]. 江苏省人民政府公报 , 2021(5): 5-14.

[26] 江苏省人民政府办公厅 . 江苏省人民政府办公厅关于印发江苏省“十四五”科技创新规划的通知 [J]. 江苏省人民政府公报 , 2021(17): 67-87.

[27] 上海市人民政府办公厅 . 上海市人民政府办公厅关于印发《立足数字经济新赛道推动数据要素产业创新发展行动方案（2023-2025 年）》的通知 [EB/OL]. (2023-07-22) [2023-09-08]. https://www.shanghai.gov.cn/202316bgtwj/20230829/5472ef31541a49c9b84bac918e27b540.html.

[28] 上海市科学技术委员会 . 上海市建设具有全球影响力的科技创新中心“十四五”规划 [EB/OL]. (2023-06-27) [2023-09-08]. https://www.shanghai.gov.cn/hqkjcx1/20230627/cc3e2e0adb0548f6b535138e2bc8a353.html.

[29] 上海市人民代表大会 . 上海市推进科技创新中心建设条例 [EB/OL]. (2023-06-04)[2023-09-08]. https://www.shanghai.gov.cn/zdqyzjkxc/20230608/d557367e132e4f128ab1af6aeba29b9e.html.

[30] 广东省科学技术厅 . 广东省科学技术厅关于广东省十三届人大四次会议第 1164 号代表建议答复的函 [EB/OL]. (2023-11-16) [2023-11-16]. http://gdstc.gd.gov.cn/gkmlpt/content/3/3333/mpost_3333890.html#729.

[31] 戴力新 . 强化创新要素共享服务高水平科技自立自强 [J]. 群众 , 2023(8): 35-36.

[32] 江苏省科技厅 , 江苏省财政厅 , 江苏省教育厅 , 等 . 《江苏省科技资源统筹服务管理办法 (试行)》[J]. 江苏省人民政府公报 , 2020(15): 32-38.

[33] 王勰祎 , 孟溦 . 科技数据类新型研发机构运行特征与绩效评估研究 —— 以上海科技创新资源数据中心为例 [J]. 科技管理研究 , 2021, 41(5): 21-28.

[34] 马广瑞 . 上海科技创新资源数据中心 [J]. 张江科技评论 , 2022(2): 63.

[35] 张林轩 , 储节旺 , 蔡翔 , 等 . 我国地市级政府数据开放发展现状及对策探析 —— 以安徽省为例 [J]. 情报工程 , 2021, 7(4): 79-92.

[36] 刘枌 , 王瑛芝 . 基于 CiteSpace 的开放政府数据价值研究可视化分析 [J]. 情报工程 , 2023, 9(3): 43-57.

doi:10.3772/j.issn.2095-915x.2024.05.005

开放政府数据对全要素生产率的影响研究

开放科学（资源服务）标识码（OSID）

刘枏　薛世麟　王大海

重庆交通大学经济与管理学院 重庆 400074

摘要：[目的 / 意义] 现有研究在 OGD 和经济增长的关系上欠缺实证研究和定量分析，难以反映和诠释其推动作用。[方法 / 过程] 使用 DEA-Malmquist 指数法测算全国 19 个主要城市的全要素生产率，以及使用生产函数法测算了各个城市历年的资源配置效率，研究了 OGD 与全要素生产率之间的关系。[局限] 考虑到影响全要素生产率的作用机制有欠缺，数据可能未涵盖全面，同时部分变量的缺失和测算具有一定的局限性。[结果 / 结论] 政府数据的开放可以显著促进全要素生产率的提高，通过创新水平和优化要素配置效率的中介效应占据了开放政府数据促进全要素生产率总效应的 50% 以上。

关键词：开放政府数据；全要素生产率；双重差分模型

中图分类号：G35；F290

Research on the Impact of Open Government Data on Total Factor Productivity

LIU Nan XUE Shilin WANG Dahai

School of economics and management, Chongqing Jiaotong University, Chongqing 400074, China

Abstract: [Objective/Significance] Existing research lacks empirical research and quantitative analysis on the relationship between OGD and economic growth, making it difficult to reflect and interpret its driving role. [Methods/Processes] This article used the DEA Malmquist index method to measure the total factor productivity of 19 major cities in China, and used the production function method to measure the resource allocation efficiency of each city over the years. The relationship between OGD and total factor productivity was studied. [Limitations] Considering the lack of mechanisms that affect total factor productivity, the data may not be comprehensive, and the absence and measurement of some variables have certain limitations. [Results/Conclusions] The study found that the openness of government data can significantly promote the improvement of total factor productivity, and the mediating effect of innovation level and optimization of factor allocation efficiency accounts for more than 50% of the total effect of open government data on promoting total factor productivity.

Keywords: Open Government Data; Total Factor Productivity; Double Difference Model

基金项目 教育部人文社会科学研究一般项目“大数据资产的定价机理、方法和规范研究”（20XJAZH007）。

作者简介 刘枏（1966-），教授，博士，主要研究方向为大数据、数据分析、工程管理信息化等；薛世麟（2000-），硕士研究生，主要研究方向为数据分析，E-mail: 1046272035@qq.com；王大海（1997-），硕士研究生，主要研究方向为数据分析。

引用格式 刘枏 , 薛世麟 , 王大海 . 开放政府数据对全要素生产率的影响研究 [J]. 情报工程 , 2024, 10(5): 51-60.

doi:10.3772/j.issn.2095-915x.2024.05.005

引言

自从美、英等西方国家在 2009 年前后以开放数据的形式向社会提供政府数据以来，开放政府数据（以下简称 OGD）运动在世界各国迅速传播。2015 年，中国的《促进大数据发展行动纲要》明确提出数据的开放共享是国家大数据战略的核心和重要任务之一，强调积极推进数据资源的开放共享，引导社会力量共同开发利用，从而释放社会和经济效益。2021 年 3 月颁布的《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》提出激活数据要素潜能、优先推动高价值数据集向社会开放、鼓励第三方深化对公共数据的挖掘利用等目标。数据作为数字经济的基础，谁掌握数据，谁就拥有了先发优势，正如麦肯锡在其报告 *Big data: the next frontier for innovation, competition, and productivity* 中说数据已经成为除劳动力、资本以外的重要生产要素 [1]。

全要素生产率是考虑了技术水平、创新能力、管理效率和要素配置的综合生产率，其对经济增长的作用已经得到国内外学者的研究支持 [2]。但是针对其发展的关键与基础——数据要素在促进经济高质量发展及提高全要素生产率中扮演怎样的角色却鲜有研究，可供直接参考的成果很少。OGD 这一实践使得作为数据资源中质量最优、数量最多的政府数据得到了一定程度的开发与应用，为进一步理解上述问题提供了一个很好的切入点。但是当前关于 OGD 的研究多数集中于理论阶段，对经济增长和全要素生产率的影响实证研究偏少，缺乏对两者影响机制的分析，难以评估 OGD 对经济发展的实际贡献，这是本研究发起的初衷。

1 理论研究与研究假设

1.1 政府数据的开放及其作用

政府数据开放可理解为数字时代政府提供的新型公共服务 [3]，其中政府扮演数据供应商角色，社会主体扮演数据使用者角色。Attard 等 [4] 的研究指出，开放政府数据对于提升政府透明度、释放社会和商业价值、参与治理有着积极的意义。作为公共产品，政府数据兼具生产要素和治理要素二重性 [5]，不仅可以给企业带来巨大的价值潜力，还可以带动政府高效利用数据进行治理，更好地为公众提供服务 [6]。然而，为了充分利用 OGD 的经济价值和商业机会，需要合适的数据交易模式，所以有必要建立一个政府数据交易平台，通过维护企业用户社区、建立便捷交互系统以及安全可靠的数据源等平台核心体制 [7]，让政府在基础数据供给、保障数据共享、维护数据公平、发展包容性数据经济方面发挥积极作用 [8]，实现数据要素价值转化、释放数据价值，驱动数字经济高质量发展。

1.2 数据要素的价值创造与促进经济增长的机制

数据具有价值与使用价值，是数据要素开发利用的两个前提条件 [9]。原始数据由于其碎片化、虚拟性等特征，不能直接提供有价值的信息，需要经过收集、存储、清理、分析等环节，才能作为数据要素投入最终产品的生产 [10]，因此，数据的价值除了由使用价值贡献外，还来源于一系列人类劳动 [11]。按照蔡继明等 [12] 的

观点，数据要素的价值一部分来自属于公众的共有数据，另一部分来自生产数据时的劳动投入，体现了数据的原始价值和劳动价值。数据在经济活动中的价值体现在经济效益的实现和经济增长的促进上。数据要素促进经济增长的作用机制包括两条路径：驱动创新与技术进步、提高资源配置效率[13]。研究表明，数据要素可以优化企业的决策流程，提升企业的学习能力，促进企业的产品创新、管理创新，其蕴含的高价值信息被分析获取后能够降低企业运营时面临的不确定性，优化传统生产要素的配比[14]。也有研究指出，政府数据开放带来的数据要素的供给将直接推动经济增长，通过公开数据而带来的政府治理效率提升则是间接的推动力量[15]。为了衡量数据要素对于经济增长的作用效果，徐翔等[16]建立了包含数据资本、ICT 资本和传统生产要素的生产函数，研究了数据要素对经济增长的直接效应和溢出效应。蔡继明等[12]认为数据要素可通过单个企业数据的初始存量、前期收集处理数据所投入的劳动，以及当期在收集处理数据所投入的劳动等三种途径，提高企业的劳动生产力。因此，劳动生产率是一个适宜的反映数据作用效果的指标。

1.3 数字经济对全要素生产率增长的促进作用

全要素生产率反映了技术进步、管理提升、制度改善等非生产要素投入而提高的生产率水平，是研究经济增长源泉和评价增长质量的重要指标，一般使用狭义的全要素生产率来表达技术创新和技术引进[17]。近年来，全要素生产率常被研究者用来分析信息技术、数字经济等背景下的经济增长。他们的研究表明，数字经济的发展有效地促进了技术进步和效率改善，进而推动了全要素生产率稳步提升[18]。从微观层面来看，数字技术创新有助于改善企业的资源配置，提高创新效率，最终实现全要素生产率的提升[19]。从宏观层面来看，数字经济及数字技术具有显著的正外部性、高流动性以及低成本传播性，这使得数字经济不仅能够有效促进本地区全要素生产率提升，还可以通过空间溢出效应牵引邻近地区全要素生产率水平的提高[20]。

既有研究表明，数据要素可以推动宏观层面的资源配置优化和有效利用，从而助力数字经济的高质量发展。因此，本文考虑开放政府数据是否也能改善资源错配，以推优化动资源配置的方式促进经济发展？经济运行中各类经济主体的有限理性、信息不对称、缺乏市场竞争、产业结构不合理等因素的客观存在，都会使得要素配置难以达到帕累托最优。而开放政府数据可能在这几个方面有显著的改进作用。开放政府数据能够大幅地提高信息搜集的效率，有效改善要素供求双方的信息不对称，有利于减小供求缺口，降低交易成本，提高匹配效率，缓解要素的扭曲配置。

1.4 研究假设

OGD 可以被视为直接推动区域全要素生产率提高的技术冲击。在某种程度上，生产活动中使用的数据元素越多，可用于社会分析和挖掘的场景就越多，从而增加了数据要素的价值。因此，本文提出：

H_1：OGD 可以促进全要素生产率的提高。

doi:10.3772/j.issn.2095-915x.2024.05.005

本文认为，开放的政府数据可以通过两条主要途径提高全要素生产率：提高创新水平和优化要素配置效率。先前的研究表明，数据要素可以促进宏观层面的资源配置优化，有助于数字经济的高质量发展。因此，OGD 是否也能通过改善资源配置来增强资源匹配，支持经济发展，值得探索。

基于以上分析，本文提出：

H_2：OGD 可以通过提高创新水平来促进全要素生产率的提高。

H_3：开放的政府数据可以通过提高要素配置效率来提高全要素生产率。

2 研究设计

2.1 基于DEA-Malmquist指数法的全要素生产率测算

本文选取各个城市的年度实际国民生产总值作为产出变量，均以 2011 年为基期的不变价进行平减处理。

物质资本存量：永续盘存法是当前国内大量研究文献中常见的一种物质资本存量计算方法，本文参考张军等[21]的具体步骤对 2011—2020 年间各城市物质资本存量进行估计。永续盘存法的具体形式为：

$$K_{it} = K_{i(t-1)}(1-\delta) + I_{it} \quad (1)$$

其中，K_{it} 和 $K_{i(t-1)}$ 分别表示第 i 个城市在 t 年和 t–1 年的不变价物质资本存量；I_{it} 是第 i 个城市在 t 年的不变价物质资本投资额；δ 是折旧率，取 9.6%。

劳动力投入：本文采用各个城市的“全社会就业人数”为变量来衡量每年的劳动力投入水平。

以 t 和 t+1 时期决策单元的前沿技术为比较参考，我们得出：

$$TFP_{it} = M_{it}(x_t, y_t, x_{t+1}, y_{t+1}) = \sqrt{\frac{E_{it}(x_{t+1}, y_{t+1})}{E_{it}(x_t, y_t)} \frac{E_{i(t+1)}(x_{t+1}, y_{t+1})}{E_{i(t+1)}(x_t, y_t)}} \quad (2)$$

上述方程表示第 i 个城市在 t 时期的生产率。M 表示 Malmquist 指数的计算，E 表示 DEA 模型得出的效率结果。在本研究中，选择 CCR 模型，并将上一节的投入产出变量用作原始数据，使用 Dearun 软件计算全要素生产率。

2.2 双层差异分析（DID）模型的设计与解释

本文充分利用了 OGD 的实施所导致的在时间和城市两个层面的差异，并采用多时间点 DID 模型分析和检验了 OGD 对总要素生产率的影响。因此，基线回归模型设置如下：

$$TFP_{it} = \alpha_0 + \alpha_1 OGD_{it} + \alpha_2 X + \mu_i + \lambda_t + \varepsilon_{it} \quad (3)$$

其中 TFP_{it} 为解释变量，表示采用 DEA-Malmquist 指数法计算的第 i 个城市的全要素生产率；OGD_{it} 为 OGD 的虚拟变量，以开放政府数据平台启动后一年的值为 1，其他为 0；X 代表影响全要素生产率的其他一系列控制变量，有城市经济发展水平（人均实际 GDP 的自然对数）、城镇化水平（城镇化率）、政府干预程度（一般公共预算支出占地区 GDP 的比重）；μ 表示城市固定效应，消除城市层面不随时间变化的因素对实证结果的影响；λ 表示年份固定效应，以消除随时间变化的因素对全要素生产率的影响；ε 表示为误差项；α_0 为常数项，α_1 为被解

释变量的回归系数，α_2 为控制变量的回归系数。

在进行基本回归检验后，我们采用逐步法来检验创新水平和因素分配效率对 OGD 对总因素生产力的影响的中介作用。我们可以通过建立以下面板数据模型，进一步检验 OGD 对总因素生产力的中介效应：

$$\begin{aligned} \mathrm{Inn}_{it} &= \alpha_3 + \alpha_4 OGD_{it} + \alpha_5 \mathrm{X} + \mu_i + \lambda_t + \varepsilon_{it} \\ \mathrm{Con}_{it} &= \alpha_6 + \alpha_7 OGD_{it} + \alpha_8 \mathrm{X} + \mu_i + \lambda_t + \varepsilon_{it} \\ \mathrm{TFP}_{it} &= \alpha_9 + \alpha_{10} OGD_{it} + \alpha_{11}\ \mathrm{Inn}_{\ it} + \\ &\quad \alpha_{12}\ \mathrm{Con}_{it} + \alpha_{13} \mathrm{X} + \mu_i + \lambda_t + \varepsilon_{it} \end{aligned} \quad (4)$$

其中，Inn 代表创新水平，Con 代表因素分配效率。其他变量具有与基线回归模型中相同的含义。

3 数据分析与结果

3.1 平行趋势假设检验

为保证实证回归结果的有效性，本研究将使用基线回归模型来检验在 OGD 实施前，样本城市的总因子生产率的变化是否符合平行趋势假设。

考虑到开放政府数据实施前四年和后四年的数据较少，参考类似研究中的常见做法，将前四年的数据合并到模型的时点第 -4 期，将后四年的数据合并到时点第 4 期，以便更好地观测到不同时间的影响效果。图 1 所示的平行趋势检验结果表明，在开放政府数据实施前的各期，基准回归模型的估计系数均不显著。

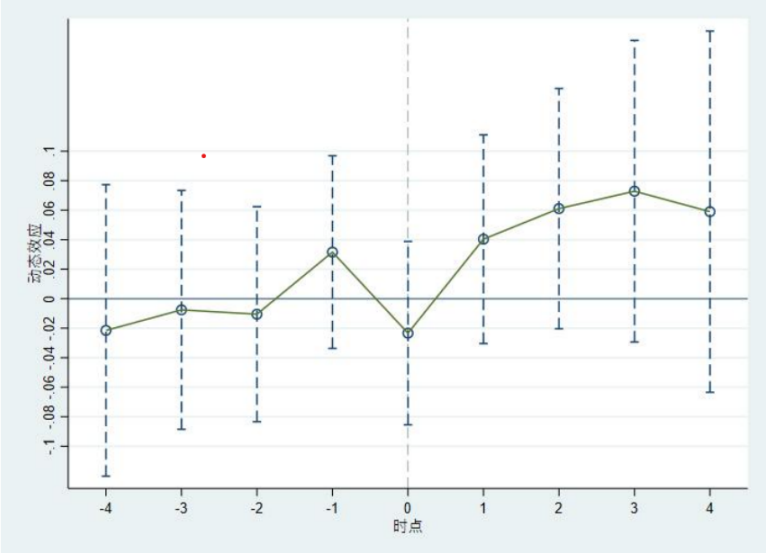


图 1　平行趋势试验结果

注：图中的小圆圈表示估计的系数，虚线表示估计的系数的 95% 置信区间

表 1 为多时间段 DID 模型的基线回归结果。基线回归模型的结果表明，当将城市固定效应控制在列（1）中时，OGD 的估计系数不显著。然而，在第（2）列添加控制变量时，OGD 的全要素生产率的回归系数显著为 0.0484。这表明，在控制了相关变量后，OGD 显著促进了全要素生产率的提高。在第（3）栏中，OGD 对全要素生产率的回归系数为 0.0537，在控制了城市固定效应和年份固定效应后，在 5% 水平上具有显著性。 回归系数的显著性和值有所提

doi:10.3772/j.issn.2095-915x.2024.05.005

高，根据客观统计回归，OGD 的系数变化呈现合理的趋势。从经济方面来看，根据第（3）栏的结果，与没有实施 OGD（0.0537/1.085）的时期相比，实施 OGD 一年后全要素生产率总体提高了约 4.95%。这表明，公开的政府数据确实对促进提高全要素生产率有显著作用。

表 1　基准回归结果

变量	（1） TFP	（2） TFP	（3） TFP
OGD	-0.029 6 （-0.020 2）	0.048 4* （-0.024 8）	0.053 7** （-0.025）
控制变量	否	是	是
城市固定效应	是	是	是
年份固定效应	否	否	是
常数项	1.094*** （-0.000 958）	1.710*** （-0.255）	1.372*** （-0.41）
R^2	0.045	0.260	0.409

注：标准误差在系数 *、* * 和 * * * 中分别表示 10%、5% 和 1% 的显著水平。下表中的内容相同

3.2　稳健性试验

（1）更改被解释变量

为了排除被解释变量的测算模型可能对研究结果产生的干扰，本文又分别使用了 DEA 方法中 CCR 模型全局参比、超效率 CCR 模型相邻参比、超效率 CCR 模型全局参比来测算全要素生产率并更换原本的被解释变量进行稳健性检验，以保证实证结果的可靠性。回归结果如表 2 所示，第（1）列中 CCR 全局参比的全要素生产率回归系数在 5% 水平上显著为正，第（2）列中超效率 CCR 相邻参比的全要素生产率回归系数在 5% 水平上显著为正，第（3）列中超效率 CCR 全局参比的全要素生产率回归系数在 5% 水平上显著为正。显然，更换了被解释变量后的回归结果依然保持了与基准回归中类似正向显著性。上述结果说明，本文的研究结论较为稳健。

表 2　替换所解释变量的回归结果

变量	（1） CCR 全局参比	（2） 超效率 CCR 相邻参比	（3） 超效率 CCR 全局参比
OGD	0.055 6** （0.024 9）	0.051 7** （0.025 5）	0.052 4** （0.025）
控制变量	（0.005 01）	（0.005 11）	（0.005 02）
城市固定效应	-0.123**	-0.113*	-0.130**
年份固定效应	（0.061 6）	（0.062 9）	（0.061 8）
常数项	1.242*** （0.409）	1.340*** （0.418）	1.286*** （0.411）
R^2	0.416	0.406	0.413

（2）改变实施时间

本文参照 Ferrara 等 [22] 的做法，通过改变开放政府数据的实施时间来进行稳健性检验。通过将实施时间向前推进 1~3 年，构建了 3 个“伪开放政府数据变量”，然后纳入基线回归模型进行实证分析。结果显示，在将开放政府数据虚拟变量提前 1~3 年后，全要素生产率的估计系数均不显著。上述结果进一步表明，本文的研究结论稳健。

表 3　改变实施时间的回归结果

变量	（1） TFP	（2） TFP	（3） TFP
	提前 1 年	提前 2 年	提前 3 年
OGD	-0.005 54 （0.025 1）	-0.024 9 （0.025 5）	0.007 69 （0.025）
控制变量	是	是	是
城市固定效应	是	是	是
年份固定效应	是	是	是
常数项	1.324*** （0.416）	1.307*** （0.415）	1.324*** （0.416）
R^2	0.392	0.396	0.392

（3）安慰剂检测

为了排除实证结果并非偶然性事件所致，参考魏志华等 [23] 的做法，本文使用多次重复随机生成开放政府数据开展的城市和开展的时间的方式进行安慰剂检验。研究结果如图 2 所示，在随机过程下，估计的系数集中在零附近，且大

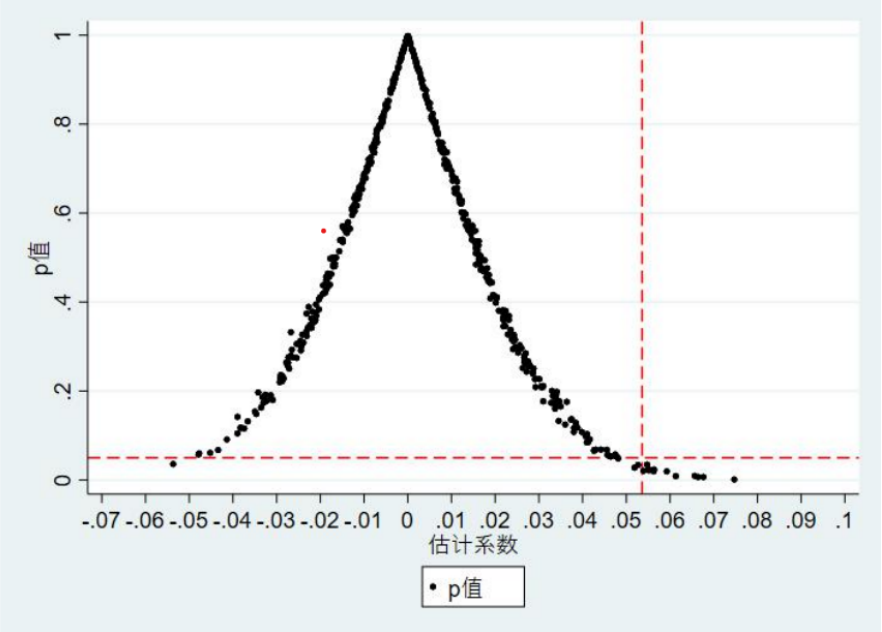


图 2　安慰剂试验的结果

doi:10.3772/j.issn.2095-915x.2024.05.005

部分回归结果在 5% 水平上不显著。此外，基线回归系数 0.0537 超出安慰剂检验的 98% 置信区间范围，表明基线回归对总因素生产力的改善影响是罕见事件，基线回归的结果得到安慰剂检验的支持。因此，OGD 对总要素生产力的促进作用并不是偶然的，本研究的研究结论是稳健的。

3.3 作用机制检验

本文使用中介效应模型式式（4）和逐步法，从“创新水平”和“要素配置”两个角度，检验假设二和假设三提出的开放政府数据影响全要素生产率的内在机制。

为明确 OGD 对总要素生产力的影响机制，进一步考察了 OGD 对区域创新水平和资源配置效率的影响。从表 4 可以看出，显然，第（1）列的创新水平（Inn）的回归系数在 5% 水平上显著为正，第（2）列的因子分配效率（Con）的回归系数在 5% 水平上也显著为正。这些结果表明，开放的政府数据显著提高了区域创新水平，优化了因子分配效率。以 TFP 为因变量，开放政府数据（OGD）、创新水平（Inn）、因子分配效率（Con）作为回归检验的自变量（3），开放政府数据（OGD）的回归系数在 5%（0.025 2）上仍然显著，说明创新水平（Inn）和因子分配效率（Con）的中介作用只是部分，这两种机制可以解释 OGD 对总要素生产力促进效应的 50% 以上。

表 4 中介效应回归结果

变量	（1） Inn	（2） Con	（3） TFP
OGD	0.049 0** （0.555 8）	0.028 8 ** （0.511）	0.0252** （0.025 2）
Inn			0.495*** （0.365）
Con			0.146*** （0.397）
控制变量	是	是	是
城市固定效应	是	是	是
年份固定效应	是	是	是
常数项	4.674*** （9.119）	-6.38*** （8.385）	1.545*** （0.445）
R^2	0.900	0.983	0.418

4 研究结论与启示

4.1 研究结论

我国开放政府数据的发展十分迅速，截止 2022 年已有超过 50% 的城市向社会开放了政府数据，以期能够对政府治理、社会建设、经济发展产生积极影响。学界也对此展开了充分的讨论，普遍认为这一举措或许能够突破当前面

临的一些发展难题，引领数字时代的发展方向。本文立足于开放政府数据对经济领域的深刻影响，梳理国内外相关文献，以国内 19 个主要城市为样本使用 DEA-Malmquist 指数法测算其全要素生产率，最后借助政府数据开放平台这一准自然实验，使用双重差异分析模型实证分析了开放政府数据对于全要素生产率的作用。主要结论可归纳为以下几点：

（1）OGD 能够显著促进全要素生产率的提高。以 DEA-Malmquist 指数法测算的全要素生产率为被解释变量的基准回归结果表明，开放政府数据确实能够显著促进全要素生产率的提高。这一研究结论在经过更换被解释变量、改变实施时间、安慰剂检验等稳健性检验后依然得到了支持。

（2）OGD 能够通过提升创新水平和优化要素配置效率的中介效应促进全要素生产率提高。通过中介效应模型检验发现创新和要素配置效率能够解释开放政府数据对全要素生产率一半以上的促进作用。

4.2 实践启示

在全力加速中国式现代化进程的宏伟蓝图中，各级政府应深刻认识到政府数据对于驱动数字经济发展与塑造数字政府新生态的不可替代价值。数据作为一种新兴生产要素，已并肩土地与资本，成为推动社会进步与经济发展的关键力量。无论是数字经济的蓬勃兴起，还是数字政府的高效构建，均深深植根于数据的支撑体系之中。充分挖掘并释放政府数据的生产要素与治理要素双重属性，不仅能够为数据要素市场的培育注入强劲动力，更能在提升政府决策的科学性、精准性以及管理服务效率方面发挥决定性作用，助力国家治理体系和治理能力现代化的跨越式发展。

在推动经济高质量发展战略中，主管部门应重点关注城市创新水平以及优化要素配置效率。通过开放政府数据，城市创新水平的提升会催生出一系列新兴产业和业态，如大数据、人工智能、云计算等。这些新兴产业不仅具有巨大的市场潜力，还能够带动相关产业链的发展，形成新的经济增长点。同时精准资源配置可以帮助企业更准确地了解市场需求和行业动态，从而进行更加精准的资源配置。企业可以根据政府提供的数据调整生产计划、研发方向和销售策略，减少资源浪费和盲目投资。还可以推动区域协同发展，有助于区域间形成比较优势，促进区域间的分工协作。通过数据引导创新要素流向效率更高的地区，可以实现区域经济的协调发展。

参 考 文 献

[1] MANYIKA J C M, BROWN B. Big Data: The Next Frontier for Innovation, Competition, and Productivity[R]. McKinsey Global Institute, 2011.

[2] GIORDANO C, ZOLLINO F. Long-run factor accumulation and productivity trends in Italy[J]. Journal of Economic Surveys, 2021, 35(3): 741-803.

[3] 朱峥 . 政府数据开放的权利基础及其制度构建 [J]. 电子政务 , 2020(10): 117-128.

[4] ATTARD J, ORLANDI F, SCERRI S, et al. A systematic review of open government data initiatives[J]. Government Information Quarterly, 2015, 32(4): 399-418.

[5] 李刚 . 政府数据市场化配置的边界 : 政府数据的“生产要素”和“治理要素”二重性 [J]. 图书与情报 , 2020(3): 20-21.

[6] 付熙雯 , 郑磊 . 政府数据开放国内研究综述 [J]. 电

doi:10.3772/j.issn.2095-915x.2024.05.005

子政务 , 2013(6): 8-15.

[7] WYSEL M, BAKER D, BILLINGSLEY W. Data sharing platforms: How value is created from agricultural data[J]. Agricultural Systems, 2021, 193: 103241.

[8] 段盛华 , 于凤霞 , 关乐宁 . 数据时代的政府治理创新——基于数据开放共享的视角 [J]. 电子政务 , 2020(9): 74-83.

[9] 李松龄 . 数据要素属性及其市场化配置改革的深化认识 [J]. 学术界 , 2022(12): 38-46.

[10] 王颂吉 , 李怡璇 , 高伊凡 . 数据要素的产权界定与收入分配机制 [J]. 福建论坛 (人文社会科学版), 2020(12): 138-145.

[11] ZENG J, GLAISTER K W. Value creation from big data: Looking inside the black box[J]. Strategic Organization, 2017, 16(2): 105-140.

[12] 蔡继明 , 曹越洋 , 刘乐易 . 论数据要素按贡献参与分配的价值基础——基于广义价值论的视角 [J]. 数量经济技术经济研究 , 2023, 40(8): 5-24.

[13] 迟考勋 , 邵月婷 , 苏福 . 大数据价值来源、价值内容与价值创造机理——基于 2011—2021 年管理类和商业类 SSCI 期刊分析 [J]. 科技进步与对策 , 2022, 39(22): 151-160.

[14] 王谦 , 付晓东 . 数据要素赋能经济增长机制探究 [J]. 上海经济研究 , 2021(4): 55-66.

[15] 张晨 , 万相昱 , 姜智超 , 等 . 开放政府数据的经济增长效应研究 [J]. 中国软科学 , 2023(2): 1-11.

[16] 徐翔 , 赵墨非 . 数据资本与经济增长路径 [J]. 经济研究 , 2020, 55(10): 38-54.

[17] 严成樑 . 现代经济增长理论的发展脉络与未来展望——兼从中国经济增长看现代经济增长理论的缺陷 [J]. 经济研究 , 2020, 55(7): 191-208.

[18] 邱子迅 , 周亚虹 . 数字经济发展与地区全要素生产率——基于国家级大数据综合试验区的分析 [J]. 财经研究 , 2021, 47(7): 4-17.

[19] 罗佳 , 张蛟蛟 , 李科 . 数字技术创新如何驱动制造业企业全要素生产率？——来自上市公司专利数据的证据 [J]. 财经研究 , 2023, 49(2): 95-109, 124.

[20] 张微微 , 王曼青 , 王媛 , 等 . 区域数字经济发展如何影响全要素生产率 ?——基于创新效率的中介检验分析 [J]. 中国软科学 , 2023(1): 195-205.

[21] 张军 , 吴桂英 , 张吉鹏 . 中国省际物质资本存量估算 :1952—2000[J]. 经济研究 , 2004(10): 35-44.

[22] FERRARA E L, CHONG A, DURYEA S. Soap Operas and Fertility: Evidence from Brazil[J]. American Economic Journal: Applied Economcs, 2012, 4(4): 1-31.

[23] 魏志华 , 王孝华 , 蔡伟毅 . 税收征管数字化与企业内部薪酬差距 [J]. 中国工业经济 , 2022(3): 152-170.

doi:10.3772/j.issn.2095-915x.2024.05.006

开放科学（资源服务）标识码（OSID）

基于 SWOT 分析的高校图书馆与公共文化服务体系跨域合作路径研究

刘慧敏[1] 曾灵[2]

1. 广州商学院 广州 510700；
2. 广东工业大学 广州 511400

摘要：[目的 / 意义] 探索高校图书馆在公共文化服务体系中的作用和地位，寻找高校图书馆参与公共文化服务体系建设的合适发展路径，促进公共文化服务的提升，推动社会的发展。[方法 / 过程] 通过文献查阅法以及网络调查法，对我国高校图书馆发展现状进行了研究，分析当前我国高校图书馆在融入公共文化服务体系时所存在的优劣势、机遇和挑战，并结合实际情况给出相应的策略。[结果 / 结论] 高校图书馆在参与公共文化服务体系建设时要加强跨域合作与资源优化整合，制定发展规划，推动理念转变，加强创新能力，提高服务质量。政府应当从政策制度、资金投入、组织协调和宣传引导等方面来鼓励高校图书馆参与公共文化服务体系建设的进程，有效促进高校图书馆的良好发展。

关键词：SWOT 分析法；高校图书馆；公共文化服务

中图分类号：G35；G258.6；G252

Research on Cross Domain Cooperation Paths between University Libraries and Public Cultural Service Systems Based on SWOT Analysis

LIU Huimin[1] ZENG Ling[2]

1. Guangzhou College of Commerce, Guangzhou 510700, China;
2. Guangdong University of Technology, Guangzhou 511400, China

Abstract: [Purpose/Significance] Explore the role and position of university libraries in the public cultural service system, find suitable development paths for university libraries to participate in the construction of the public cultural service system, promote the improvement of public cultural services, as well as the social development. [Methods/Processes] This article used literature review and online survey methods to study the current development status of university libraries in China, and analyzed the advantages and disadvantages as well as opportunities and challenges of integrating university libraries into the public cultural

基金项目 广东省图书馆科研课题“基于 SWOT 分析的高校图书馆与地方公共文化服务体系的跨域合作探究”（GDTK22022）。

作者简介 刘慧敏（1997-），硕士研究生，主要研究方向为文本挖掘、公共图书馆；曾灵（1996-），硕士，助理馆员，主要研究方向为信息计量与科学评价，E-mail：1350790827@qq.com。

引用格式 刘慧敏 , 曾灵 . 基于 SWOT 分析的高校图书馆与公共文化服务体系跨域合作路径研究 [J]. 情报工程 , 2024, 10(5): 61-72.

doi:10.3772/j.issn.2095-915x.2024.05.006

service system, and provided corresponding strategies based on actual situations. [Results/Conclusions] When participating in the construction of the public cultural service system, university libraries should strengthen cross domain cooperation and resource optimization integration, formulate development plans, promote conceptual changes, enhance innovation capabilities, and improve service quality. The government should encourage university libraries to participate in the construction of the public cultural service system from the aspects of policy system, funding investment, organizational coordination, and publicity guidance, effectively promoting the good development of university libraries.
Keywords: SWOT Analysis; University Libraries; Public Cultural Services

引言

公共文化服务体系建设是国家“十四五”时期的重点建设内容，2015 年国务院印发《关于加快构建现代公共文化服务体系的意见》，要求加快构建现代公共文化服务体系 [1]。2021 年，文化和旅游部、国家发展改革委、财政部联合印发《关于推动公共文化服务高质量发展的意见》[2]，围绕进一步强化社会参与提出明确要求。“公共服务文化体系”是指面向大众的公益性的文化服务体系，主要包括先进文化理论研究、文艺创作、文化传授、传播、娱乐、传承和农村文化服务等七个方面 [3]。其载体包括实体和虚拟形态，如网络平台、数字资源，涵盖政府政策、制度和服务用户 [4]。高校图书馆参与地方公共文化服务体系的建设对提升人民精神生活质量、促进共同富裕和社会和谐发展具有重要意义 [5]。作为公共文化资源的一部分，高校图书馆不仅提供学术资源和服务，还承担着参与公共文化服务体系建设的责任。高校图书馆应积极参与地方公共文化建设，加强跨域合作，为促进文化服务水平的提升和社会和谐发展贡献力量。

1 相关研究

随着社会文化需求的升级和数字化技术的发展，高校图书馆参与公共文化服务体系的研究备受关注。高校图书馆承担着为师生提供学术信息资源和服务的重要职责。而公共文化服务体系则是为社会公众提供文化服务的重要平台。高校图书馆与公共文化服务体系之间存在着密切的关系，二者相互促进，共同推动着文化事业的发展。近年来，国内外关于高校图书馆参与公共文化服务体系建设的研究逐渐增多，主要包括以下几个方面的内容：

（1）高校图书馆在公共文化服务体系中的作用。高校图书馆作为公共文化服务体系的重要一员，需明确其在公共文化服务体系中的服务、角色、功能与发展定位 [6]，才能更好地承担公共文化服务体系建设的责任。高校图书馆被视为公共文化服务体系的主力军，应当充分发挥学科服务与技术、文献获取与咨询、与其他主体联合发展等作用 [7]。高校图书馆也被视为公共图书馆的补充形式，作为社会力量的重要一员，承担着传承文化、普及知识、促进学术研究、教育引导的重要任务，同时在服务高

校教育时向社会公众开放，为社会各界提供文化知识和信息服务 [8]。

（2）高校图书馆参与公共文化服务体系的现状。虽然有研究 [9] 在论述高校图书馆参与公共文化服务体系的优势时，强调了高校图书馆在信息、资源、技术、人才方面与公共图书馆相比具有的优势。但是高校图书馆参与公共文化服务仍然存在如服务有限、态度被动、形式单一、社会知晓度低和资源联动困难等问题 [10]。而且高校及其图书馆归因、政府机构归因和公众用户归因等都是高校图书馆参与公共文化服务体系的重要影响因素 [11]。

（3）高校图书馆与公共文化服务体系的合作路径。合作路径涉及多方面的融合与实践，通过研学旅行、馆藏资源、文创产业和红色文化等视角 [12]，高校图书馆积极参与文旅融合的实践。此外，在发展策略和实践路径方面，有研究 [7,13-15] 提出了一系列加强建议，包括资源共享、转变服务意识、强化宣传推广、促进合作发展以及推动公共文化服务均等化。在数字文化背景下，高校图书馆也被鼓励加强数字化服务并构建数字化服务平台 [16]。同时，图书馆应通过共享资源与社区、企业和政府建立合作关系 [7]，以发挥其公共文化服务的功能。例如，意大利费拉拉大学图书馆合作举办面向公众的文化遗产展览 [17]，以展示学生的创意。伦敦大学图书馆则在东伦敦社区建立研究与学习中心 [18]，为当地居民和国际访客提供学习设施和文化展览，促进社区参与与文化传承。

综上所述，高校图书馆在公共文化服务体系中扮演重要角色。研究者强调其定位，指出其应充分发挥学科服务、技术、文献获取、咨询等功能。但是，合作现状存在服务有限、态度被动等问题。部分学者提出加强资源共享、转变服务意识等策略，探讨数字文化背景下的服务模式，强调与社区、企业、政府合作，实现互利共赢。也有国外案例表明高校图书馆可通过展览、合作项目等方式与公众互动，促进文化传承。

然而，这些研究较少从高校图书馆的内外部环境进行全面的分析，缺乏在特定环境条件下的针对性合作路径探究。文本在知网以“SWOT”和“高校图书馆”为关键词进行检索，发现目前尚未发现有研究者将两者结合对跨域合作路径进行研究。因此本文将深入分析高校图书馆内外部环境，结合 SWOT 模型进行全面分析，为高校图书馆制定更具针对性和可操作性的发展策略，推动其在公共文化服务体系中扮演更加重要的角色，促进文化体系建设取得更大成就。

2 高校图书馆与公共文化服务体系合作的 SWOT 分析

SWOT 分析最早是由美国旧金山大学韦里克教授于 20 世纪 80 年代初提出的 [19]。SWOT 分析法，是指一种综合考虑企业内部条件和外部环境的各种因素，进行系统评价，从而选择最佳经营战略的方法 [20]。作为一种战略管理工具，SWOT 常用于评估组织内外部环境的优势、劣势、机会和威胁。高校图书馆作为高校组织的一部分，使用 SWOT 模型对其进行分析具有一定的合理性。为了提升高校图书馆的公共文化服务水平，本文将利用 SWOT 分析法对高校图书馆与公共文化服务体系跨域合作模式进行

doi:10.3772/j.issn.2095-915x.2024.05.006

研究，全面评估高校图书馆参与公共文化服务体系建设的优势、劣势、机遇和挑战，为高校图书馆探索出一条合适的发展路径。目前，高校图书馆拥有丰富的内部优势，如馆藏资源、专业人才和教育资源，而公共文化服务体系具有广泛的服务对象和特色资源，但存在人才缺失等问题。为充分利用双方优势并解决不足，需要整合资源并转变服务理念，以满足社会需求。随着经济发展和文化需求增加，高校图书馆应参与政府主导的公共文化服务建设，加强合作，应对资源分配不均、文化多样性和竞争压力等挑战，推动科技创新，提升服务水平。高校图书馆与公共文化服务体系合作的 SWOT 简要分析如图 1 所示。

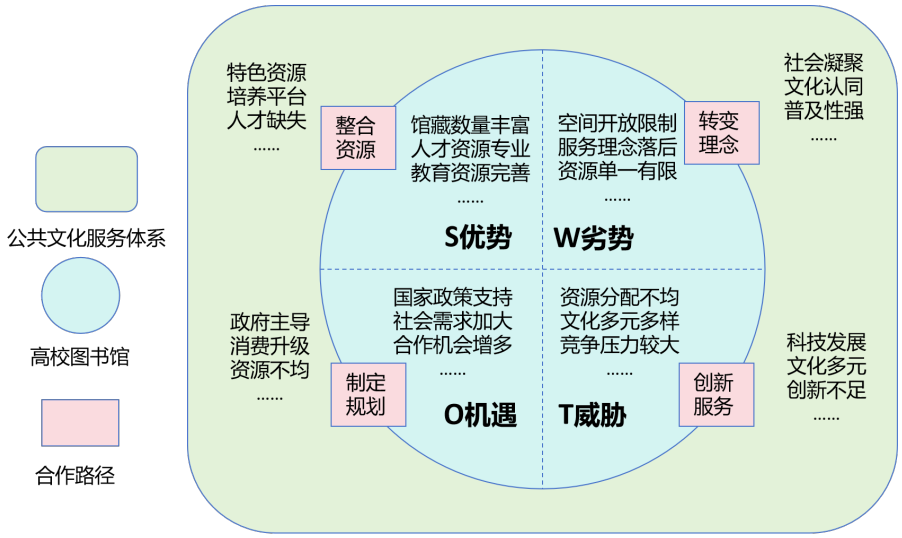


图 1　高校图书馆与公共文化服务体系合作的 SWOT 分析图

2.1　内部优势（Strengths）

（1）馆藏资源丰富。高校图书馆拥有丰富的学术资源和文献馆藏，包括图书、期刊、学位论文、学术数据库等。根据国家统计局 2022 年的统计年鉴，截至 2021 年，我国 3215 个公共图书馆总藏量 126178 万册 [21]，平均一个公共图书馆的总藏量仅为 39.25 万册。根据教育部高等学校图书情报工作指导委员会发布的《2021 年高校图书馆发展报告》[22]，共有 1290 所高校图书馆提交了 2021 年度馆藏纸质图书累积量的有效数据，纸质图书累积总量约为 16.97 亿册，馆均值为 131.5 万册。从馆藏资源上看，高校图书馆馆藏丰富，特别是综合类大学图书馆，拥有广泛的学科领域覆盖和深度的学术资源积累。因此高校图书馆作为公共图书馆服务公众的补充形式，与公共图书馆合作，建立图书馆资源共享联盟 [23]，为公众提供更广泛、更深入的学术资源和文献服务。

（2）拥有专业人才。根据 2022 年中国高校图书馆基本统计数据 [24]，截至 2022 年 10 月 7 日，共有 1373 所高校图书馆提交了 2021 年度在编工作人员的有效数据，总人数为 39696 人，馆均 28.9 人，在编博士学位工作人员总计 1532 人，每馆拥有博士学位工作人员 1.15 人；

在编硕士学位工作人员总计 13695 人，平均每馆聘用硕士学位工作人员 10.2 人。高校图书馆拥有专业图书馆人才，能根据需求制定公共文化服务策略，推行创新服务模式，建立合作关系推进文化服务体系建设，与其他文化机构开展项目合作，共享资源，提供全面高效的公共文化服务。

（3）学术氛围浓厚。高校图书馆拥有较好的学术环境，可以为读者提供参与学术文化活动和学术交流的平台。因其承载着学术文化的传承和发展的重要使命，图书馆可以利用自身的丰厚资源，与地方公共文化服务体系合作举办各种高质量的学术文化展览、学术文化讲座等活动，向公众展示学术文化的魅力和价值，推动学术文化的传播和交流。

（4）教育资源丰厚。高校图书馆作为高校组织的一部分，可以充分利用高校的教育资源和自身的馆藏资源，为公众提供学习辅导、培训等服务，例如开展学术写作、科研方法等方面的讲座，帮助公众提升学习能力和研究水平，也可以通过与社区学校合作，开展教育培训项目等。

2.2 内部劣势（Weaknesses）

（1）开放限制。一些高校图书馆对于校外读者的开放条件较为严格，只提供办理借阅证和馆内阅览等单一的服务形式。这种局限性与社会读者的期望相去甚远，限制了公共文化服务的范围和受众。但开放服务可能会导致一些负面影响，比如挤占高校图书馆读者的阅览座位、图书馆文献的流失以及外借文献不归还等问题 [25]。

（2）服务理念转变不足。高校图书馆作为高等教育机构的重要组成部分，其服务理念和社会意识的转变对于参与公共文化服务体系建设具有重要意义。目前高校图书馆的服务理念仍然以学术服务为主，对公共文化服务的意识和能力有待提升 [26]。传统上，高校图书馆主要以支持学术研究和教学为主要目标，服务对象主要是学校的师生群体。这种服务理念在一定程度上限制了高校图书馆参与公共文化服务体系建设的广度和深度。高校图书馆应该意识到，作为公共文化服务的一部分，其服务对象应该更加广泛，包括社会各个群体，并认识自身在社会中的角色和责任，将公共文化服务作为核心任务之一。

（3）资源有限。高校图书馆的主要任务是支持学术研究和教学活动 [27]，其内部资金分配主要用于购买学术图书、期刊和数据库，提供学术研究和教学支持服务。这就导致高校图书馆分配给公共文化服务方面的资源相对较少。其次，高校图书馆的设施主要用于学术研究和教学活动，如阅览室、电子资源区等，关于公共文化服务的设施有限，这也同样限制了公共文化服务的规模和质量。例如，高校图书馆的阅览室可能只对学校师生开放且开放时间有限；电子资源区可能只提供学术数据库和期刊，无法满足公众对其他类型的文化资源的需求。这些限制导致高校图书馆无法充分发挥其在公共文化服务中的潜力，满足公众对文化活动和交流的需求。

2.3 外部机会（Opportunities）

（1）国家政策支持。早在 2015 年，教育

doi:10.3772/j.issn.2095-915x.2024.05.006

部印发的《普通高等学校图书馆规程》就提倡高校图书馆应在保证校内服务和正常工作秩序的前提下，开展社会化服务[28]。2021年文旅部针对公共文化服务体系的建设提出了十四五建设规划，并明确指出新时期公共文化服务体系建设的主要任务之一，就是大力推动全国公共文化服务体系建设城乡一体化，缩小城乡公共文化服务资源的差距[29]。国家对公共文化服务的重视程度逐渐提高，为高校图书馆参与公共文化服务提供了政策支持和发展机遇。

（2）社会需求增加。随着社会的发展和进步，公众对文化服务的需求不断增加。人们对知识、艺术、历史等方面的了解和探索的需求日益旺盛。高校图书馆作为知识和文化资源的宝库，拥有丰富的图书、期刊、文献和电子资源，可以满足公众对学习和文化的迫切需求。

（3）合作机会增加。面对多元化的服务需求，高校图书馆拥有更多的合作机会。高校图书馆通过与公共图书馆、社会文化机构和企业等社会组织合作，拓展服务领域，提高公共文化服务的质量和覆盖范围，推动公共文化服务的整体提升。合作可以包括资源共享、人员培训、活动策划等，通过互相借鉴和学习，促进不同机构之间的交流和合作，实现资源的优化配置和互补发展，为公众提供更多元化、专业化的文化服务。

2.4 外部威胁（Threats）

（1）资源分配不均。主要体现在学校之间的经费差异上。部分高校由于缺少经费只能购买有限的图书和期刊资源，导致这些高校图书馆无法与头部高校进行竞争。同时，专业人才不足也会导致这些高校图书馆无法提供全面高质量的服务，在咨询服务、文化活动等方面的表现可能不尽如人意。另外，硬件设施条件也影响着公共文化服务水平。部分高校图书馆由于设施条件有限，如存在座位不足、设备老旧等问题，限制了公共文化活动的开展，导致公众的参与体验较差。

（2）文化多样性。不同地区和不同群体对公共文化服务的需求以及偏好均存在差异，高校图书馆需要根据高校所在地的实际情况进行差异化服务。在这个过程中，高校图书馆需要面对来自不同文化背景和学科领域的用户需求，提供多元化的文化资源和服务。这需要图书馆具备广泛的文化资源、举办多样的文化活动，并能够根据用户需求进行个性化的服务。

（3）竞争压力大。随着社会文化服务市场的竞争加剧，高校图书馆面临来自其他公共文化服务机构的竞争压力，它们也提供类似的文化服务，如图书借阅、学术资源获取等。高校图书馆需要化竞争为合作，通过合作提供更好的服务和资源，提升自身服务水平和竞争力，以吸引更多公众的关注和使用。

3 高校图书馆参与公共文化服务体系建设的路径

3.1 加强跨域合作，优化整合资源

（1）资源共建共享。高校图书馆与地方公共文化服务机构可以通过建立资源共享机制，实现资源的互补与共享。图书馆区域联盟是资源共建与共享的重要表现形式，承担了区域内

资源整合的责任[30]。自 2008 年“高校图书馆联盟”成为研究热点以来，图书馆联盟体系基本覆盖全国。例如，内蒙古自治区的乌兰察布市图书馆与当地的几所高校图书馆结成共建共享联盟，促进了资源整合和服务效能的提升[31]。此外，“政府机构—企业—学校”的散点合作也是资源共建共享的另一种形式。以文化和旅游部、北京大学与抖音集团为例，三方于 2023 年签署了《共建全国智慧图书馆体系框架协议》，旨在依据资源共享、优势互补和互惠互利的原则，充分发挥各自的优势[32]。此前北京大学已与国家图书馆和抖音集团展开深度合作，推出了“《永乐大典》高清影像数据库”等项目。

（2）合作举办活动。通过高校图书馆与地方公共文化服务机构合作，可以整合校内外资源，提供多样文化服务和活动。合作形式包括提供学术文化资源支持和场地宣传支持，共同打造多元文化活动。例如，鸦片战争博物馆与东莞理工学院合作举办展览《百年薪火——禁毒先驱林则徐家风传承展》，展开大思政课程、文创产品设计、学术交流等合作，拟建立战略合作伙伴关系，共同推动学术研究、人才培养和社会服务[33]。广东药科大学图书馆发起文化服务助力“双百行动”活动，赴罗定市怀集县开展科技文化服务，助力地方经济社会发展和教育帮扶，在学校和社会各方力量合作下，发挥了高校图书馆优秀传统文化传承、学科情报咨询服务和科普教育等优势，推动了“双百行动”取得扎实成效[34]。

（3）人才培养与交流。高校图书馆与地方公共文化服务机构开展人才培养和交流活动，相互合作共赢。高校图书馆派遣专业人员到文化机构进行培训，提高服务水平；文化机构邀请高校专家参与活动，促进学术交流。例如，南京图书馆与南京大学联合设立“图书馆大数据应用江苏省文化和旅游重点实验室”，开展基础研究，支持产业技术创新，培养科技人才，促进科技成果转化[35]。辽宁省图书馆与辽宁大学合作培养“古籍保护与修复”研究生，借助国家古籍保护人才培训基地，邀请专家进行实践课程教学。该合作发挥了资源优势，培养高层次专业人才，解决了古籍保护人才短缺的难题，促进了高校发挥社会服务功能，推动了学科的建设与发展[36]。

3.2 制定发展规划，推动理念转变

（1）制定发展规划。2020 年安徽省高校图工委举办数字图书馆“十四五”发展研讨会，安徽省教育厅副厅长储常连出席会议并指出，社会服务是高校图书馆要树立的“四个中心”建设目标之一[37]。高校图书馆应制定与公共文化服务合作相关的发展规划，明确目标和战略，包括文化活动、资源共享、人才培养等重点领域，制定工作方案和时间表。西安交通大学在“十四五规划”中，通过调研图书馆发展环境，分析机遇与挑战，明确了图书馆的定位、愿景和服务任务，针对人才培养、学科建设和科技创新港等服务，提出关键举措和具体行动计划，为规划实施指明路径和方向；同时，从资源、技术、空间和队伍等方面提出条件保障，推动高校图书馆与公共文化服务合作的发展[38]。

（2）推动开放理念转变。积极推动高校

doi:10.3772/j.issn.2095-915x.2024.05.006

图书馆的开放理念转变，逐步放开对公众的开放限制，提供更广泛的公共文化服务。可以考虑开放部分资源、设立公共文化活动场所等方式。北京工业大学耿丹学院图书馆是北京市第一家面向社会开放的高校图书馆[39]，耿丹学院图书馆秉承“开放办馆，服务社会”的理念，在为校内外读者提供高品质阅读空间的同时，持续开展丰富多彩的阅读活动，让更多师生和社会读者一起共享书香，助力全民阅读。

3.3 加强创新能力，提高服务质量

（1）创新合作模式。高校图书馆与公共文化服务机构可以通过探索和创新合作模式，形成更加紧密的合作关系。可建立共同的项目组或合作机构，共同制定合作计划，明确责任和任务分工。通过高校图书馆的学术资源和专业知识，与公共文化服务机构的社会服务经验和渠道优势相结合，实现资源共享和优势互补，提供更多样化、个性化的文化服务。例如，安徽农业大学图书馆与青禾书店“馆企团学”的合作模式下，高校图书馆为其提供物理空间和业务指导监管，出版企业提供人财物保障，校团委为活动开展和学生团队管理提供宏观指导[40]。

（2）推动文化创意与创新。高校图书馆与公共文化服务机构可以共同推动文化创意与创新，设立创新基金，支持创新项目和实践，鼓励馆员提出创新想法，鼓励公众参与创意产生和文化创新。例如，青禾书店与安徽农业大学图书馆展开深度合作，共同创办集文化休闲、产品设计展示销售、手工 DIY 体验等一体的青禾文创空间，同时出版企业派专人驻店指导，在合作中创新“书店 +”运营模式，探索“书店 + 文创”“书店 + 阅读推广”“书店 + 社会公益”等馆店融合新模式，挖掘书店和高校自身潜力，提升服务能力[39]。

（3）强化技术应用。高校图书馆可以加强技术应用，借助数字化、网络化的手段提供更便捷、高效的文化服务。通过建设和完善数字资源平台、开展在线阅读和学习服务，提供智能化的文献检索和推荐，满足公众对文化资源的需求。同时，也可以积极探索利用新技术，如人工智能、大数据等，提升服务的个性化和定制化水平。例如，西南大学图书馆引进三台智能机器人，它们不仅能实现前台接待、问路引领、场馆介绍等功能，还能依靠强大的数据库实现图书检索、智能导览、智能互动、知识库问答等功能。

4 完善高校图书馆参与公共文化服务建设的保障

4.1 政策制度保障

政府应在法律层面确立高校图书馆参与公共文化服务体系建设的权利和义务。法律保障需涉及知识产权保护、合作伙伴关系管理、服务质量评估等方面的内容，为高校图书馆的参与提供法律依据和保护。在国家“公共文化服务体系的十四五规划建设”等重大政策支持下，地方可以结合当地实际情况，制定支持高校图书馆参与公共文化服务体系建设的具体政策。政策可以制定经费支持、资源共享、人才培养、合作奖励等方面的具

体措施，为高校图书馆提供必要的政策支持和优惠政策。同时加强对高校图书馆参与公共文化服务体系建设的监管和评估，确保合作项目的顺利进行和效果的实现。通过定期的监管和评估，及时发现问题和不足，加强政策的实施及实施效果的监督，促进公共文化服务体系建设的不断完善。

4.2 资金投入保障

国家和地方政府需增加对高校图书馆与地方公共文化服务体系合作的经费投入，用于资源采购、数字化建设、设备更新等方面。鼓励社会组织或个人通过政府购买、捐赠与共享、志愿者服务等多种模式 [41] 支持高校参与公共文化建设。政府购买模式直接为高校图书馆提供经费购买文化资源和服务，通过项目制确保经费合理使用和服务效果；捐赠与共享模式鼓励社会力量和企业捐赠资源，促进资源共享和互补；志愿者服务模式鼓励志愿者参与文化服务工作，丰富服务内容，提升服务质量和效果。政府应综合利用多种支持模式，为高校图书馆提供必要支持，促进其参与公共文化服务体系建设。

4.3 组织协调保障

（1）建立专门机构。国家或地方政府可以建立专门的部门或机构负责协调和组织高校图书馆与地方公共文化服务体系的跨域合作。通过制定指导性的政策和方案，明确高校图书馆与地方公共文化服务机构的合作目标、内容和责任分工，以提供指导和支持，确保跨域合作得到有效的组织和推动。

（2）建立合作平台。国家或地方政府可以为高校图书馆与地方公共文化服务机构之间的合作提供便利性和规范性支持。通过建立合作协议、共享平台等方式，明确合作的具体安排和机制，促进资源共享和协同发展。

4.3 宣传引导保障

通过综合利用各种宣传渠道和手段，如新闻媒体、校园媒体等，提供支持和奖励，可以包括资金支持、荣誉表彰、知识产权保护等方面的支持措施，鼓励各大宣传媒体投身到高校图书馆与地方公共文化服务体系的合作中，并展示成功的合作案例和示范，宣传引导社会组织、企业和个人积极参与高校图书馆与地方公共文化服务体系的跨域合作。

5 结语

本文通过 SWOT 分析揭示了高校图书馆参与公共文化服务体系建设的优劣势。高校图书馆拥有丰富的馆藏资源、专业人才及浓厚的学术氛围，但也面临开放限制、服务理念转变不足和资源有限等问题。在外部环境中，国家政策的支持和日益增长的社会文化需求为高校图书馆与其他社会机构的合作提供了机会，但竞争压力也随之增加。其次，不同层次的高校图书馆在资源分配上存在不均，且各地区和群体对公共文化服务的需求和偏好存在差异，也增加了文化多样性所带来的挑战。

综上，本文认为高校参与公共文化服务体系建设的最优路径是不断加强跨域合作与资源优化整合，制定发展规划，推动理念转变，加强创新能力，提高服务质量。然而高校参与公

doi:10.3772/j.issn.2095-915x.2024.05.006

共文化服务体系建设并非一方唱独角戏，国家、社会也需要提供一定的支持，因此本文从政策制度、资金投入、组织协同、宣传引导等方面提出保障措施，以期能有效促进高校图书馆的良好发展。

参 考 文 献

[1] 中国政府网 . 中共中央办公厅、国务院办公厅印发《关于加快构建现代公共文化服务体系的意见》[EB/OL]. (2015-01-14) [2024-06-24]. https://www.gov.cn/xinwen/2015-01/14/content_2804250.htm.

[2] 中国政府网 . 文化和旅游部、发展改革委、财政部关于推动公共文化服务高质量发展的意见 [EB/OL]. (2021-03-08) [2024-06-24]. https://www.gov.cn/gongbao/content/2021/content_5602033.htm.

[3] 黄丽茵 . 南宁高校图书馆参与周边乡镇公共文化服务建设研究 [D]. 南宁 : 广西大学 , 2022.

[4] 钱兰岚 , 王建涛 . 服务 “ 可及性 ” 视角下的新时代公共文化服务体系建设路径研究 [J]. 图书馆杂志 , 2022, 41(3): 41-47.

[5] 黎梅 , 奉晓红 . 高校图书馆参与地方公共文化服务体系构建研究 [J]. 图书馆 , 2014(5): 107-109, 115.

[6] 王文莉 , 陈朋 . 高校图书馆参与社区公共文化服务模式探索 —— 以武汉市高校为例 [J]. 湖北经济学院学报 (人文社会科学版), 2021, 18(11): 102-104.

[7] 王笑梅 . 新时代背景下，高校图书馆公共文化服务功能定位与服务价值探究 [J]. 兰台内外 , 2022(34): 73-75.

[8] 罗彬香 . 高校图书馆参与地方公共文化建设服务策略研究 [J]. 河南图书馆学刊 , 2023, 43(8): 77-79.

[9] 常琛 . 高校图书馆参与公共文化服务体系构建探析 [J]. 图书馆工作与研究 , 2012(5): 97-99.

[10] 王金花 . 观我国高校图书馆参与公共文化服务 [J]. 文化产业 , 2023(31): 55-57.

[11] 曹国凤 , 贾晓彦 , 张丽 . 高校图书馆参与公共文化服务体系影响因素研究 [J]. 图书情报工作 , 2019, 63(6): 35-40.

[12] 王颖 . 文旅融合下高校图书馆参与公共文化服务的研究 [J]. 内蒙古科技与经济 , 2023(24): 135-137, 141.

[13] 叶彤 . 高校图书馆对外开放的现实审思 —— 以书香淮安建设为例 [J]. 传播与版权 , 2023(21): 77-80.

[14] 杨小玲 . 高校图书馆智慧建设中的公共文化服务新思路 [J]. 河南图书馆学刊 , 2023, 43(10): 61-64.

[15] 王现忠 . 共享理念下石家庄市属高校图书馆参与优质公共文化服务的对策研究 [J]. 文化学刊 , 2023(12): 143-146.

[16] 齐方圆 , 石小康 . 数字文化背景下高校图书馆参与公共数字文化服务研究 [J]. 工业技术与职业教育 , 2023, 21(6): 112-115.

[17] BERNABÈ A, CONTARINI M, MANFRA M, et al. Design for Cultural Heritage at the University of Ferrara[C]//6th International Conference on Higher Education Advances (HEAd’20). Editorial Universitat Politècnica de València, 2020 (30-05-2020): 455-463.

[18] MEUNIER B. Library Technology and Innovation as a Force for Public Good: A Case study from UCL Library Services[C]//Greater Noida, INDIA. 2018. New York: IEEE, 2018: 159-165.

[19] 倪义芳 , 吴晓波 . 论企业战略管理思想的演变 [J]. 经济管理 , 2001(6): 4-11.

[20] 张沁园 . SWOT 分析法在战略管理中的应用 [J].

企业改革与管理 , 2006(2): 62-63.

[21] 国家统计局 . 中国统计年鉴 [M]. 北京 : 中国统计出版社 , 2022: 23-24.

[22] 教育部高等学校图书情报工作指导委员会 . 2021 年高校图书馆发展报告 [DB/OL]. (2022-11-21) [2024-06-24]. http://www.scal.edu.cn/sites/default/files/attachment/tjpg/2021 年高校图书馆发展报告_0.pdf.

[23] 沈璐 , 仇恒喜 . 高校图书馆参与地方公共文化建设研究 —— 兼论参与模式 [J]. 嘉兴学院学报 , 2022, 34(3): 103-109.

[24] 吴汉华 , 王波 . 2022 年中国高校图书馆基本统计数据分析 [J]. 大学图书馆学报 , 2023, 41(6): 63-72.

[25] 于蓓莉 , 陈萍秀 , 刘艳磊 . 高校图书馆面向社会开放研究 [J]. 成都航空职业技术学院学报 , 2016, 32(2): 90-92.

[26] 李育嫦 . 公共文化服务视角下高校图书馆的社会合作研究 [J]. 情报理论与实践 , 2015, 38(12): 43-46.

[27] 李育嫦 . 公共文化服务体系中高校图书馆的定位与实现对策 [J]. 情报理论与实践 , 2015, 38(9): 51-54.

[28] 教育部 . 普通高等学校图书馆规程 [EB/OL]. (2015-12-31) [2024-06-24]. http://www.moe.gov.cn/srcsite/A08/moe_736/s3886/201601/t20160120_228487.html.

[29] 文化和旅游部 . 文化和旅游部关于印发《“十四五”公共文化服务体系建设规划》的通知 [EB/OL]. (2021-06-10) [2024-06-24]. https://www.gov.cn/zhengce/zhengceku/2021-06/23/content_5620456.htm.

[30] 王怡 , 邵波 . 从区域联合到服务联合 : 高校图书馆联盟智慧化转型研究 [J]. 图书馆学研究 , 2022(8): 69-75.

[31] 乌兰察布市文化旅游体育局 . 乌兰察布市图书馆与集宁师范学院图书馆、乌兰察布职业技术学院图书馆、乌兰察布医学高等专科学校图书馆结成共建共享联盟 , 促进资源整合 , 提升服务效能 [EB/OL]. (2022-09-02) [2024-06-24]. http://whlytyj.wulanchabu.gov.cn/gzdt/912809.html.

[32] 北京大学新闻网 . 文化和旅游部、北京大学、抖音集团共建全国智慧图书馆体系 [EB/OL]. (2023-06-13) [2024-06-24]. https://news.pku.edu.cn/xwzh/4dd8e23905624e84912261a3ab770849.htm.

[33] 中国纪念馆 . 博物馆 + 高校，破圈合作擦出新火 花 [EB/OL]. (2024-4-18) [2024-06-24]. https://mp.weixin.qq.com/s?__biz=MzI3MDU2MDg0OQ==&mid=2247686801&idx=5&sn=da73e9883cc462c5f2449f6b385aaa9a&chksm=eb88e524d17257e20fcc987426804c3b215aa08290cc7cf6dc3e535d4b09f6a2b5c38f982c90&scene=27.

[34] 广东药科大学图书馆 . 双百行动 | 广东药科大学与罗定市、怀集县携手启动文化服务助力活动 [EB/OL]. (2024-04-17)[2024-06-24]. https://mp.weixin.qq.com/s?__biz=MjM5MDMwMjA0MQ==&mid=2247503428&idx=1&sn=156c4a1c8d1e2416e2ac9de75d676235&chksm=a7af674cc97671bb89fcb301993f76d0d83aad0e07422aa13a8d44c524a5558f915d42a53baa&scene=27.

[35] 南京图书馆 . 南京图书馆联合南京大学成功申报“图书馆大数据应用江苏省文化和旅游重点实验室”[EB/OL]. (2020-12-29)[2024-06-24]. http://www.jslib.org.cn/zx/gqsb/202012/t20201229_198957.html.

[36] 中华人民共和国文化和旅游部 . 辽宁省图书馆与辽宁大学合作培养古籍保护人才 [EB/OL]. (2017-10-30) [2024-06-24]. https://www.mct.gov.cn/whzx/

doi:10.3772/j.issn.2095-915x.2024.05.006

qgwhxxlb/ln/201710/t20171030_780760.htm.

[37] 孙鹏 . 我国高校图书馆 “ 十四五 ” 规划的进展与热点聚焦 [J]. 大学图书情报学刊 , 2021, 39(6): 14-19, 65.

[38] 贾申利 , 邵晶 , 时莹 . 西安交通大学图书馆 “ 十四五 ” 发展规划编制框架 [J]. 大学图书馆学报 , 2021, 39(1): 21-23, 27.

[39] 光明网 . 北京市首家面向社会开放的高校图书馆开馆 [EB/OL]. (2019-11-18) [2024-06-24]. https://news.gmw.cn/xinxi/2019-11/18/content_33329029.htm.

[40] 朱建军 , 吴小冰 . 高校校园书店 “ 馆企团学 ” 合作共赢模式研究 —— 以安徽农业大学图书馆青禾书店为例 [J]. 大学图书情报学刊 , 2023, 41(1): 113-117.

[41] 郭利伟 , 冯永财 , 彭逊 . 共享理念下高校图书馆参与地方公共文化服务研究 [J]. 图书馆工作与研究 , 2020(12): 123-128.

doi:10.3772/j.issn.2095-915x.2024.05.007

开放科学（资源服务）标识码（OSID）

基于深度语义理解的“技术—知识”关联实体识别及演化分析研究

杨金庆[1,2] 李嘉琦[1] 杨儒汉[1] 罗星雨[1] 程秀峰[1]

1. 华中师范大学信息管理学院 武汉 430079；
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

摘要：[目的 / 意义] 从细粒度视角理解“技术—知识”关联实体，构建专利文献中技术要素与知识要素识别的实现方案。[方法 / 过程] 选取传统机器学习模型 *HMM*、*CRF*，深度学习模型 *BiLSTM-CRF*、*BERT-Softmax*、*BERT-CRF* 和 *BERT-BiLSTM-CRF* 进行任务训练学习，以便确定性能最优的细粒度技术与知识实体识别模型。[结果 / 结论] 为了验证所构建技术与知识实体识别的理论框架，以出版印刷领域文本作为实验验证场景，从专利文本中随机抽取 7853 条有效语料句，标注了 71626 个实体，通过训练学习确定 *BERT-BiLSTM-CRF* 为性能较好的实体识别模型，其对知识与技术实体识别综合性能 F1 值为 0.82。此外，运用训练出的最优模型从 66665 篇专利文本的第一权利要求、权利要求、独立权利要求和技术功效中识别出 4769296 对知识与技术实体关联组合体，并分析了技术演化路径和“技术—知识”关联网络结构的演化规律。

关键词：科技情报；学科技术关联；实体识别；BERT

中图分类号：G35；TP391

Research on Related Entity Recognition and Evolution Analysis of “Technology-Knowledge” Based on Deep Semantic Understanding

YANG Jinqing[1,2] LI Jiaqi[1] YANG Ruhan[1] LUO Xingyu[1] CHENG Xiufeng[1]

1. School of Information Management, Central China Normal University, Wuhan 430079, China;
2. Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038, China

基金项目 富媒体数字出版内容组织与知识服务重点实验室开放基金项目“基于深度语义理解的技术谱系自动构建研究”（ZD2023/11-03）；中央高校基本业务费项目“基于技术要素深度语义理解的创新路径识别及引导策略研究”（CCNU24ZZ140）；中央高校基本业务项目“基于知识网络科学知识角色转变研究”（CCNU23XJ013）。

作者简介 杨金庆（1991-），博士，副教授，主要研究方向为科技情报、学科知识演化，E-mail: jinq_yang@163.com；李嘉琦（2001-），硕士研究生，主要研究方向为信息检索、用户行为；杨儒汉（2000-），硕士研究生，主要研究方向为政务智能问答；罗星雨（2002-），硕士研究生，主要研究方向为科技情报、学科知识演化；程秀峰（1981-），博士，教授，主要研究方向为信息组织与检索、大数据分析及应用。

引用格式 杨金庆，李嘉琦，杨儒汉，等 . 基于深度语义理解的“技术—知识”关联实体识别及演化分析研究 [J]. 情报工程，2024, 10(5): 73-84.

doi:10.3772/j.issn.2095-915x.2024.05.007

Abstract: [Objective/Significance] Understand the "technology-knowledge" related entities from a fine-grained perspective, and construct a implementation scheme for identifying technology elements and knowledge elements in patent documents. [Methods/Processes] The traditional machine learning models *HMM* and *CRF*, and the deep learning models *BiLSTM-CRF*, *BERT-Softmax*, *BERT-CRF*, and *BERT-BiLSTM-CRF* are selected for the task training and learning in order to identify the fine-grained technology with optimal performance and the knowledge entity recognition model. [Results/Conclusions] In order to validate the theoretical framework of technology and knowledge entity recognition, this paper takes the text in the field of publishing and printing as the experimental validation scenario, randomly selects 7853 valid corpus sentences from patent texts, and annotates 71626 entities, and determines that the *BERT-BiLSTM-CRF* is the entity recognition model with better performance through training, and the F1 value of its comprehensive performance for knowledge and technology entity recognition is 0.82. In addition, this paper applies the trained optimal model to identify 4769296 pairs of knowledge-technology entity association combinations from the first claim, claims, independent claims and technical effects of 66665 patents, and analyzes the technology evolution paths and the evolution pattern of the "technology-knowledge" association network structure. We also analyzed the technology evolution path and the evolution law of "technology-knowledge" association network structure.

Keywords: Scientific and Technological Information; Subject Technology Correlation; Entity Recognition; *BERT*

引言

新一轮科技革命和产业变革加速演进，科学与技术的交互日益紧密，共同驱动科技创新发展。尽管科学与技术拥有独立的知识结构和演化过程，但是在科技创新中有多层次交融。技术的本质是对科学发现的利用，新技术源自新的科学发现实践应用，或是现有技术的改进、调整和组合[1]。同样，新科学的发现也是依赖于科学与技术之间的合作，例如青霉素、晶体管等产品的研发，表现出科学与技术之间不仅遵循自身的发展规律，而且还相互影响，既促进科学进步，又催生新技术[2]。目前，科学与技术之间的关系一般包括技术与科学的依赖关系、结合关系、互动关系以及知识传递关系[3]。对技术而言，科学为技术提供强有力的理论基础，是技术开发的基本依据。从构成要素来看，技术由经验型要素、知识型要素和实体应用型要素构成[4]，反映出新技术的形成离不开科学知识与现有技术的双重支撑。

目前，"技术—知识"关联主要由引用关系连接的期刊论文和专利文献表示[5]。可是，技术本质上是人类为满足自身的需要，在实践活动中根据实践经验或科学原理所创造或发明的各种手段和方式方法的总和。技术本身由多种要素构成，可理解为主体知识、经验、技能与客体（工具、机器设备等）要素的统一[6]。专利文献作为技术核心内容呈现的载体形式，不仅包含了表述工具、机器设备等的技术要素，还包含了表示主体知识及科学原理的知识要素。"知识要素"是指专利文本中表述自然科学原理以及理论知识等的知识实体，"技术要素"是指表示技术实践中的方法、技巧、手段和工具等的技术实体[7]。鉴于此，本文试图借助深度语义理解技术构建专利文献中技术与知识实体识别的实现方案，识别出专利文本中共现关

联的知识与技术实体对，并对细粒度“技术—知识”实体关联组合进行初步的演化分析。

1 相关研究

1.1 基于深度语义理解的科技文献实体识别研究

科技文献作为知识的重要载体，包括论文和专利两大类[8]，是科学技术演化进程和发展水平的直观反映。如何高效地对科技文献内容中的知识与技术实体进行识别与抽取，是生物医学领域、社交媒体领域、图书情报等领域的重要研究主题[9]。实体识别（*Named Entity Recognition, NER*）是自然语言处理（*NLP*）中的一个基本任务，目标是从文本中识别出具有特定意义的实体（如人名、地点、组织名），并将其分类为预定的类别[10]。科技文献是蕴含大量知识的非结构化文本，其中涉及多种类型实体、长实体链，呈现出更新快、专业术语密集、语义识别困难、重叠关系复杂[11]等特点，导致科技文献实体识别相对于通用领域实体识别难度大幅度提升。此外，不同领域科技文献在语言风格、结构和术语等方面的差异也要求开发适用于各自特点的实体识别方法和模型。

在早期科技文献实体识别研究中，基于统计的机器学习方法，如支持向量机（*SVM*）、隐马尔可夫模型（*HMM*）、条件随机场（*CRF*）等被广泛使用[12]。随着科技文献术语的复杂化和结构的多样化发展，基于统计的机器学习方法在语义理解方面逐渐显得不足，而深度学习的发展为实体识别带来了突破性进展。深度学习的方法在大规模数据集训练基础上学习丰富的语言规律和语义关系，能够捕捉和建模复杂的、非直观的语言现象和语义模式，进而提升实体识别性能。典型的深度学习方法包括递归神经网络（*RNN*）、卷积神经网络（*CNN*）、双向长短期记忆网络（*BiLSTM*）以及 *BERT* 和其他基于 *Transformer* 的模型[13]。目前较为流行的实体识别方法主要为基于 *BiLSTM* 和 *CRF* 框架的变体或以 *BERT* 作为预训练语言模型进行优化的方法。如，Gao 等[14]构建了一种适用于稀少标记样本数据集的技术实体识别的对抗性多任务神经网络，利用 *BERT* 捕捉文本深层次语义特征，再通过引入对抗性训练机制增强模型对实体边界和关系的识别能力。Ma 等[15]提出了 *BiLSTM +CRF* 模型与基于特征的命名实体知识库融合的实体抽取方法，提取文献中时间、地点及技术实体，以分析最具潜力的中国脆弱生态修复技术。于润羽等[16]基于关键词－字符 *LSTM* 和注意力机制提出了一种能够考虑文本中的上下文信息以及全局信息的命名实体识别算法，应用于科技学术会议论文数据的识别，实验结果显示该方法优于对比算法。Chen 等[17]提出一种基于外部医学知识和自注意力机制的 *MFA-BERT-BiLSTM-CRF* 模型，能够有效地捕捉丰富的医学语义特征和全局上下文信息，优化了复杂医学文献命名实体识别的性能。王宇晖等[18]提出基于 *Transformer* 与技术词信息的知识产权实体识别方法，包含技术词信息融合层、知识产权语义抽取层和实体标签优化层三个层次，有效解决了知识产权实体的多义性问题，以及 *Transformer* 模型的相对位置感知弱等问题。Zhao 等[19]提出了针对中国专利文本

doi:10.3772/j.issn.2095-915x.2024.05.007

的联合实体和关系提取的 *BERT-BiLSTM-DGAT* 模型，能够准确捕获中国专利文本中的复杂长实体和重叠关系，提高了实体和关系提取的准确性。

随着技术的不断发展，针对科技文献实体识别任务的研究越来越多，识别效果也不断提升，但仍面临一些挑战，如多类型实体的识别、领域内实体识别精度的差异，以及如何处理专业术语的多义性问题等。为解决这些问题，未来科技文献实体识别研究应向深层次语义、跨领域适应性、自动学习和调整的实体识别方向发展。

1.2 “科学—技术”关联视角下科学演化规律研究

科学是由人类对认识客体的知识体系、产生知识的活动、科学方法等按一定层次、一定方式所构成的一种动态知识系统，它不仅仅是一套固定的知识，更是一种持续发展和进步的过程 [20-21]。技术既是一种知识体系，也是实物工具的集合，包括技能、方法、物理设备和软件等 [22]。科学与技术关联本质是科学知识体系和技术体系内部知识要素之间的复杂关系，表现为技术专利与科学论文通过相互引用、知识耦合、学者—发明人等途径产生的关联 [23]。在科学与技术的关联关系上，学术界持有不同见解，Bundy[24] 认为科学与技术虽有各自独特的知识积累结构，但知识流动的路径可以是双向的。Bush[25] 认为基础研究是技术进步的先驱，科学为基础研究，而技术则将科学原理应用于实践。Guan 等 [26] 提出了双螺旋理论，认为科学与技术在信息交互反馈中呈现螺旋上升的发展趋势。总体而言，科学与技术是独特而相互作用的共同体，二者共同塑造了科技发展的脉络和走向。因此，测量科学与技术之间的互动转化，识别二者间潜在的推动或启示作用，是厘清科学演化规律和模式的关键，对预测科技趋势、指导科研政策和推动技术创新具有重要意义。

当前，科学技术关联研究中，学者们通常以文献和专利分别表征科学和技术，利用科学计量学工具和网络分析技术结合历史分析的方法进行探讨，主要包括专利的论文引文分析法，论文的专利引文分析法和专利发明人—论文作者关联关系分析法 [8]。如，Verbee 等 [27] 通过分析生物技术和信息技术领域论文与专利间的引用网络、合作模式和知识流动，探讨了领域科学知识的发展脉络以及科学技术之间的交互。韩芳 [28] 基于不同技术领域专利与文献的引用关系，揭示了科学理论与技术成果之间的相互转化过程和规律。研究发现，科技专利的增长与科学文献的增长密切相关，反映出共演化特征。Ke[29] 针对生物技术和药物专利领域，对不同时间段、不同国家及机构类型的专利引用进行了详细探讨，揭示了科学与技术之间的关联和演变，为科技政策制定和国际合作提供了依据。刘雨梦等 [30] 构建了基础科学（*Science*）、技术科学（*Engineering science*）及工程应用（*Engineering*）的 SESE 研究模型，应用在辽宁省重大装备协同创新中心和智能型新能源汽车协同创新中心中，以分析科学技术的互动行为和共演过程。吴宁等 [31] 利用多层网络方法，基于专利和论文数据，构建了“科学—技术”知识流动微观网络分析框架，揭示了科学与技术之间的知识流动网络结构特征，为理解科技联系与

挖掘创新点提供了新视角。

随着科学技术关联研究的深入，少数学者开始从微观层面进行研究，但整体而言，现有科技关联视角下的科学演化规律和模式研究多以引文为纽带，通过专利与论文间的引证映射，揭示科学与技术的知识关联和演化规律。这类研究侧重于文献层次的引用关系和主体层次的相似关系的计量分析，未能充分从语义角度揭示科学技术内在关联，难以透视文献中知识单元间的直接联系，限制了对科学演化规律和模式的全面剖析与理解。未来的研究应更注重从深度语义角度出发，以文本内容为基础，知识与技术实体为核心，探索科学技术之间错综复杂的联系，更有效地推动科技创新和应用。

2 数据来源及标注过程

为了更好地构建细粒度“技术—知识”关联实体识别模型，笔者选取了技术与知识交叉较大的出版印刷领域作为实验场景，并从商业专利库 *Patsnap* 中运用分类号搜索的方式获取 9880 份专利文本，随机抽取 7853 条有效语料句。本文采用了 *Label studio* 数据标注工具对有效语句标注技术实体（机器、工具等）和知识实体（理论、知识等），共标注了 71626 个实体，两种实体标注的示例如图 1 所示。

图 1　技术和知识实体在专利文本中的示例

3 实体识别结果分析

3.1 “技术—知识”关联实体识别模型构建

细粒度技术实体和知识实体的识别需要在对专利文本语义进行深入理解的基础上，对输入模型的语句中的每个词组进行处理，预测两类实体标签。本文拟通过参照当下命名实体识别任务中的常见语义理解框架，构建“技术—知识”关联实体识别模型。目前，命名实体识别的深度语义理解框架前后经历两个阶段：一是以卷积神经网络（*CNN*）和循环神经网络（*RNN*）为文本特征编码器，学习单词序列的隐层表示，再借助 *CRF* 完成实体标签分类，此阶段以 *LSTM-CRF* 实体识别模型为代表；二是借助以 *BERT* 预训练模型为底层的文本特征编码器，利用预训练模型强大的文本语义理解能力预测实体标签。

结合以往实体识别研究结论，本文以 *BERT* 为底层文本特征编码器，构建 *BERT-CRF* 或 *BERT-BiLSTM-CRF* “技术—知识”关联的两种类型实体的识别模型。模型训练数据构建环节，本文采用“*BIO*”数据标注模式，将技术实体（机器、工具等）和知识实体（理论、知识等）分别标注为 *B-tool/app\I-tool/app*, *B-knowledge\I-knowledge*，适应以 *BERT* 为底层的技术要素深度学习识别架构，如图 2 所示。

doi:10.3772/j.issn.2095-915x.2024.05.007

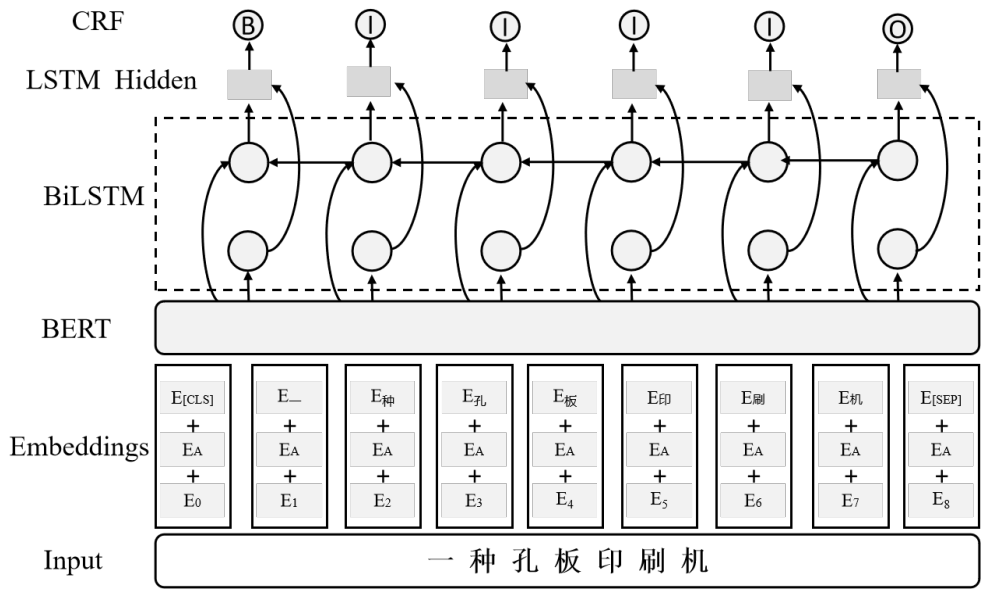


图 2 “技术—知识”关联实体识别深度学习框架

由图 2 可知，输入有效的专利文本序列后由嵌入层（*Embedding layer*）将序列嵌入转化为向量表示，随后借助 *BERT* 或 *BERT-BiLSTM* 文本特征编码能力对输入的文本向量表示进行建模，以获取文本深层语义信息，最后通过 *CRF* 分类算法对技术与知识两类实体标签进行分类预测。

3.2 基于深度语义理解的模型训练与评价

3.2.1 实验设置及评价指标

本文标注的 71626 个实体中包含 55444 个技术实体和 16182 个知识实体，按 1:1:3 划分为测试集、验证集和训练集，两者在测试集、验证集和训练集中存在数据不均衡现象，如表 1 所示。

表 1 两类实体在测试集、验证集和训练集中的分布情况

数据划分	技术实体	知识实体
训练集	25272	8799
验证集	15486	3507
测试集	14686	3876

在模型训练过程中，本文以出版印刷领域的数据集为基础，采用 *Adam* 优化器微调 *BERT* 预训练模型，对该领域专利文本中的语义进行深层次理解，并设定 *BERT* 和 *CRF* 的超参数，如表 2 所示。

表 2 ***BERT/CRF*** 超参数设定结果

超参数	值
优化器	Adam
BERT 学习率	5e-5
CRF 学习率	5e-3
隐层维度	768
Epoch	10
Batch_size_train	32
Batch_size_eval	32
Max_len	150

为了对模型识别性能进行评价，本文采取实体级的评价方式，精确度（*precision*）是指正确识别的技术与知识实体数与模型识别出的技

术与知识实体总数的比例，召回率（*recall*）是指正确识别的技术与知识实体数与数据集中实际存在的技术与知识实体总数的比例，F1 值是精确度和召回率的调和平均值，可衡量模型的综合性能，计算公式如下：

$$precision = \frac{\text{正确识别的技术与知识实体数量}}{\text{模型识别的技术与知识实体数量}} \quad (1)$$

$$recall = \frac{\text{正确识别的技术与知识实体数量}}{\text{数据集中实际的技术与知识实体数量}} \quad (2)$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

3.2.2 模型训练与对比评价

为了确定性能最好的“技术—知识”关联实体识别模型，本文先后对 *HMM*、*CRF*、*BiLSTM-CRF*、*BERT-Softmax*、*BERT-CRF* 和 *BERT-BiLSTM-CRF* 等深度学习框架进行训练和学习，并对模型识别性能进行评价，如表 3 所示。

表 3　技术与知识实体识别模型的性能对比评价

模型	实体类型	Precision	Recall	F1
HMM	不区分实体	0.66	0.58	0.62
	知识实体	0.48	0.54	0.51
	技术实体	0.70	0.59	0.64
CRF	不区分实体	0.68	0.85	0.76
	知识实体	0.42	0.80	0.56
	技术实体	0.74	0.86	0.79
BiLSTM-CRF	不区分实体	0.79	0.80	0.79
	知识实体	0.56	0.67	0.61
	技术实体	0.84	0.82	0.83
BERT-Softmax	不区分实体	0.84	0.76	0.80
	知识实体	0.81	0.48	0.60
	技术实体	0.85	0.84	0.84
BERT-CRF	不区分实体	0.84	0.80	0.82
	知识实体	0.78	0.57	0.66
	技术实体	0.85	0.87	0.86
BERT-BiLSTM-CRF	不区分实体	0.81	0.82	0.82
	知识实体	0.77	0.60	0.67
	技术实体	0.82	0.88	0.85

对比各模型的 F1 值，不难发现 *BERT-CRF* 和 *BERT-BiLSTM-CRF* 两者对知识与技术两类实体综合识别性能相当，F1 值均为 0.82。也可看出，数据不均衡对知识实体的识别性能产生一定的影响，为了使得两类实体的识别性能都较为可用，*BERT-BiLSTM-CRF* 模型可作为“技术—知识”关联实体识别的实现路径。

4 “技术—知识”实体关联组合的演化规律分析

专利文本作为科技创新的重要载体，可视为由知识单元实体和技术单元实体按照一定的

doi:10.3772/j.issn.2095-915x.2024.05.007

逻辑和功能组合而成。知识与技术实体在一定程度上反映着“技术—知识”组合的演化规律。本文按照“链接—链条—网络”的分析思路先后对“技术—知识”组合、“技术—知识”演化路径和“技术—知识”关联网络结构演化进行规律揭示。具体而言，运用知识与技术实体识别性能最优的 *BERT-BiLSTM-CRF* 模型从 1985—1998 年和 2010—2022 年跨度内 66665 篇专利文本的第一权利要求、权利要求、独立权利要求和技术功效中识别出知识与技术实体组合共 4769296 对，其各年份关联组合数量分布如图 3 所示。

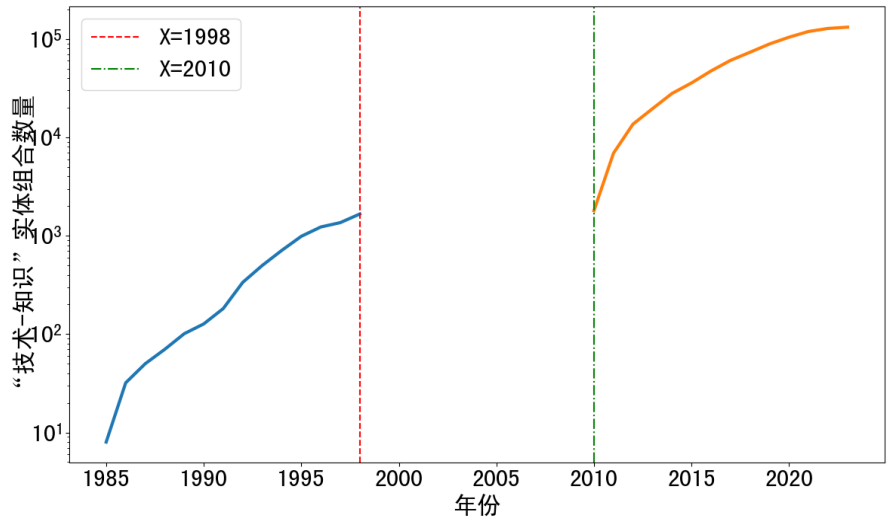


图 3 “技术—知识”实体关联组合数量历时分布情况

观察图 3 中“技术—知识”实体关联组合对数在 1985—1998 年和 2010—2022 年阶段均呈现上升趋势，但 2010—2022 年呈现相对较快的增长率，表明新的“技术—知识”组合创新不断涌现，在出版印刷领域“技术—知识”关联趋势逐渐增强。本文选取 1985—1998 年和 2010—2022 年时间跨度内的“技术—知识”实体关联关系，以时间变量为导引表达不同实体之间的继承关系，进而可视化生成细粒度“技术—知识”实体之间的演化路径，如图 4 所示。

此外，“技术—知识”实体关联对链接形成“技术—知识”关联网络，关联网络的结构变化一定程度也反映了出版印刷领域的技术演化规律。本文为控制网络规模，选取了共现频次大于 2 的知识与技术实体关联对，进而构建技术—知识关联网络，并计算了典型网络结构指标的历时变化情况，如图 5 所示。

观察图 5 可知，1985—1998 年时间跨度内关联网络的平均中心度增长较为缓慢，而 2010—2022 年增长较为迅速。结合图 3 可知，尽管近十多年“技术—知识”关联对不断增加，但“技术”与“知识”节点间的相互关联趋势也在快速上升，表明出版印刷领域技术创新更倾向于“技术”与“知识”相互驱动的模式。中介中心度在初始成长阶段下降较为明显，反弹后仍呈现下降趋势，关联网络中节点平均影响力也在持续下降，表明整个关联网络的连通性变大，也侧面表明了“技术”与“知识”关联趋势增强。

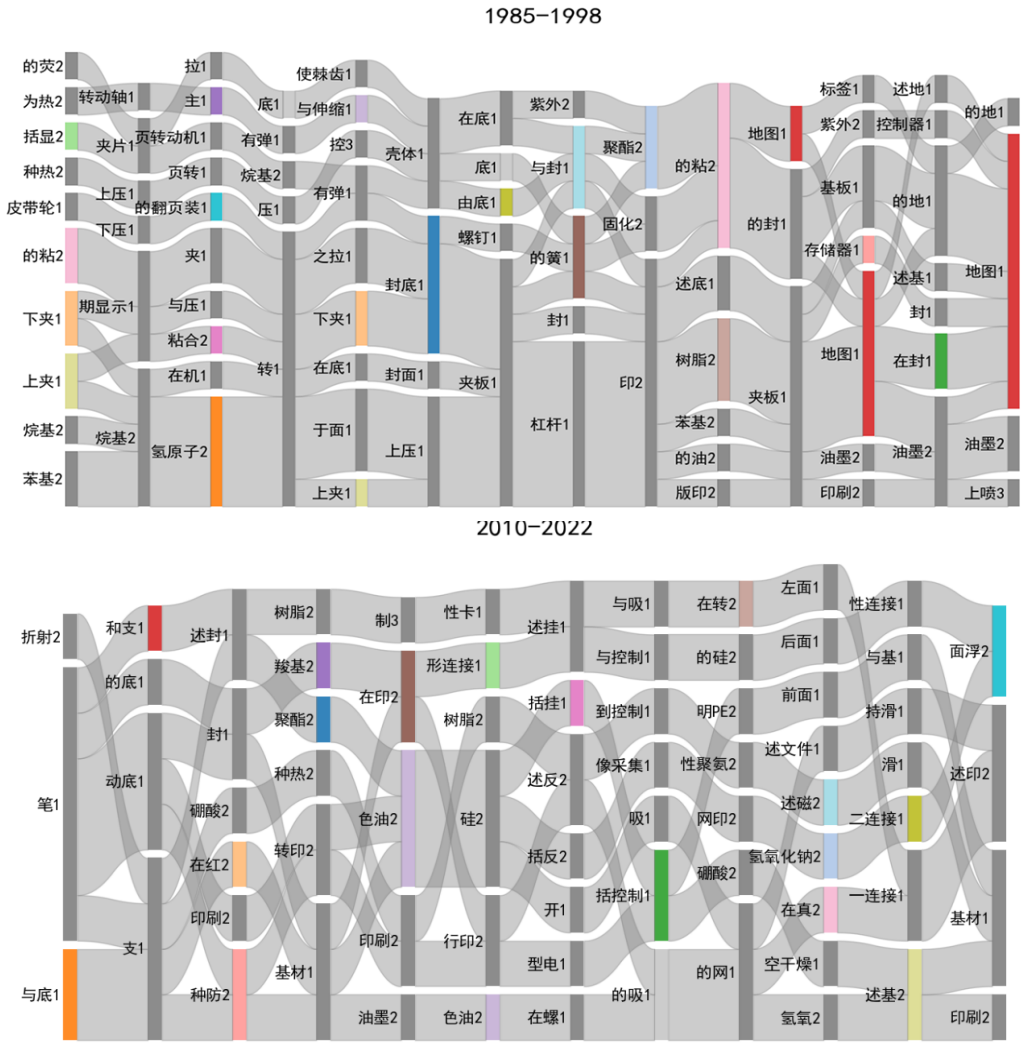


图 4　技术演化路径可视化示例图（1985—1988 年，2010—2022 年）

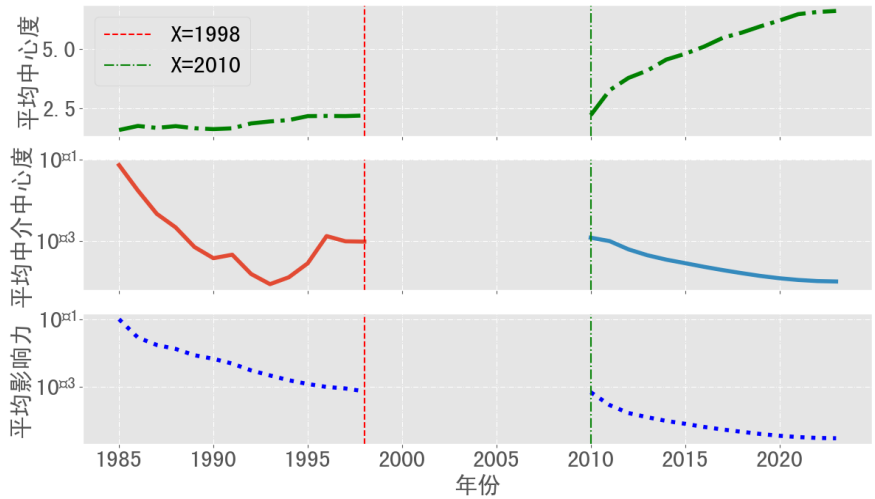


图 5　典型网络结构指标值历时可视化分析

doi:10.3772/j.issn.2095-915x.2024.05.007

5 结语

“技术—知识”关联实体识别对推动科技创新、制定科技政策，以及促进知识的转化具有极其重要的意义。以往科学技术关联研究主要以专利文献引用关系为连接，忽视了文本内容中所蕴含的深层语义信息。针对以上问题，本文从细粒度角度，识别专利文献中的技术与知识要素，通过技术与知识之间原有的内在逻辑关系，深入语义层面挖掘科技之间的关联与互动。为了更为准确识别出技术与知识实体，本文选取了传统机器学习模型 *HMM*、*CRF*，深度学习模型 *BiLSTM-CRF*、*BERT-Softmax*、*BERT-CRF* 和 *BERT-BiLSTM-CRF* 进行任务训练学习，获取性能最优的细粒度技术与知识实体模型框架。实验结果表明以 *BERT* 为底层文本特征编码器的预训练模型识别性能较好，尤其 *BERT-BiLSTM-CRF* 和 *BERT-CRF* 组合框架，对两类实体识别的综合性能 F1 值为 0.82。为进一步探析“科学—技术”关联视角下科学演化规律及模式，本文运用训练出的最优模型 *BERT-BiLSTM-CRF*，从 1985—1998 年和 2010—2022 年两个时间跨度内的专利文本中识别“技术—知识”实体关联组合体，并将各年份关联组合数量分布可视化，分析可知两阶段内实体关联组合体数量呈现上升趋势，2010—2022 年增长率相对较快。在此基础上，本文以时间变量为导引表达不同实体之间的继承关系，构建了细粒度“技术—知识”关联网络结构，计算并可视化显示典型网络结构指标的历时变化情况，结果表明 1985—1998 年期间关联网络的平均中心度增长变化率低于 2010—2022 年期间，中介中心度在初始成长阶段下降较为明显，反弹后仍呈现下降趋势。关联网络中节点平均影响力也在持续下降。实验分析发现尽管近十多年“技术—知识”关联对不断增加，但“技术”与“知识”节点间的相互关联趋势也在快速上升，表明出版印刷领域技术创新更倾向于“技术”与“知识”相互驱动的模式。本文试图从专利本身发掘“技术—知识”实体间的关联，初步分析了“技术—知识”实体关联组合的演化规律，仍存在不足之处，主要包括实验训练数据集存在客观的数据不均衡问题以及技术与知识实体标注过程中需要较强的专业背景知识，未来在继续提升细粒度“技术”与“知识”实体识别准确率的基础上，深入分析技术和知识关联背后的机理。

参 考 文 献

[1] 林苞，雷家骕 . 基于科学的创新与基于技术的创新——兼论科学—技术关系的“部门”模式 [J]. 科学学研究，2014, 32(9): 1289-1296.

[2] 林苞，雷家骕 . 基于科学的创新模式与动态——对青霉素和晶体管案例的重新分析 [J]. 科学学研究，2013, 31(10): 1459-1464.

[3] 刘小玲，谭宗颖，张超星 . 国内外“科学—技术关系”研究方法述评——聚焦文献计量方法 [J]. 图书情报工作，2015, 59(13): 142-148.

[4] 陈其荣 . 自然辩证法导论 [M]. 上海：复旦大学出版社，1995: 188-208.

[5] LI X, WANG T, PANG Y, et al. Review of Research on Named Entity Recognition[C]// SUN X, ZHANG X, XIA Z, et al. Advances in Artificial Intelligence and Security. Cham: Springer International

Publishing. 2022: 256-267.

[6] 马昱堃 . 作为有机体的技术——评《技术的本质：技术是什么，它是如何进化的》[J]. 智能社会研究 , 2023, 2(2): 171-182.

[7] 邬金鸣 , 胡智杰 , 姚茹 , 等 . 面向技术创新过程表征和描述的技术要素关联概念模型 [J]. 图书情报工作 , 2024, 68(10): 1-15.

[8] PAN W, JIAN L, LIU T. Knowledge generation and diffusion in science & technology: an empirical study of SiC-MOSFET based on scientific papers and patents[J]. Technology Analysis & Strategic Management, 2024, 36(7): 1587-1603.

[9] 何玉洁 , 杜方 , 史英杰 , 等 . 基于深度学习的命名实体识别研究综述 [J]. 计算机工程与应用 , 2021, 57(11): 21-36.

[10] LEI J, TANG B, LU X, et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American Medical Informatics Association, 2014, 21(5): 808-814.

[11] ZHAO Y, LI K, WANG T, et al. Joint entity and relation extraction model based on directed-relation GAT oriented to Chinese patent texts[J]. Soft Computing, 2024(28): 7557-7574.

[12] 刘浏 , 王东波 . 命名实体识别研究综述 [J]. 情报学报 , 2018, 37(3): 329-340.

[13] ALI S, MASOOD K, RIAZ A, et al. Named entity recognition using deep learning: A review[C]//2022 International Conference on Business Analytics for Technology and Security (ICBATS). IEEE, 2022: 1-7.

[14] GAO H, WANG T, LUO W, et al. Adversarial Multitask Learning for Technology Entity Recognition[C]//2018 IEEE International Conference on Progress in Informatics and Computing, 2018: 134-139.

[15] MA J X, YUAN H. Bi-LSTM+CRF-based named entity recognition in scientific papers in the field of ecological restoration technology[J]. Proceedings of the Association for Information Science and Technology, 2019, 56(1): 186-195.

[16] 于润羽 , 杜军平 , 薛哲 , 等 . 面向科技学术会议的命名实体识别研究 [J]. 智能系统学报 , 2022, 17(1): 50-58.

[17] CHEN W Q, HAN P, ZHONG Y L, et al. Named Entity Recognition from[C]//Chinese Medical Literature Based on Deep Learning Method2023 China Automation Congress (CAC), 2023: 9332-9337.

[18] 王宇晖 , 杜军平 , 邵蓥侠 . 基于 Transformer 与技术词信息的知识产权实体识别方法 [J]. 智能系统学报 , 2023, 18(1): 186-193.

[19] ZHAO Y, LI K C, WANG T, et al. Joint entity and relation extraction model based on directed-relation GAT oriented to Chinese patent texts[J]. Soft Computing, 2024: 1-18.

[20] 钱时惕 . 什么是科学——科学与人文漫话之一 [J]. 物理通报 , 2009(10): 56-58.

[21] 库恩 . 科学革命的结构 [M]. 北京 : 北京大学出版社 , 2012: 58-72.

[22] BASALLA G. The Evolution of Technology[M]. Cambridge: Cambridge University Press, 1989: 1-20.

[23] 王颖洁 , 张程烨 , 白凤波 , 等 . 中文命名实体识别研究综述 [J]. 计算机科学与探索 , 2023, 17(2): 324-341.

[24] BUNDY A. Australian and New Zealand information literacy framework[EB/OL]. (2004-01-01）[2024-04-23]. http://www. anziil.org.index.htm.

[25] BUSH V. Science: The Endless Frontier[M]. Washington DC: National Science Foundation, 1960: 18-19.

[26] GUAN J, HE Y. Patent-bibliometric analysis on the Chinese science—technology linkages[J]. Scientometrics, 2007, 72(3): 403-425.

[27] VERBEEK A, DEBACKERE K, LUWEL M. Science

doi:10.3772/j.issn.2095-915x.2024.05.007

cited in patents: A geographic “flow” analysis of bibliographic citation patterns in patents[J]. Scientometrics, 2003, 58(2): 241-263.

[28] 韩芳 . 基于专利引文的“科学—技术关系”及技术演化轨迹研究 [D]. 北京 : 北京邮电大学 , 2017.

[29] KE O. An analysis of the evolution of science-technology linkage in biomedicine[J]. Journal of Informetrics, 2020, 14(4): 101074.

[30] 刘雨梦 , 郑稣鹏 . 基础科学、技术科学及工程应用的“链合”演化研究——国家级协同创新中心纵向案例 [J]. 科学学研究 , 2022, 40(10): 1745-1755.

[31] 吴宁 , 杨艳萍 . 基于多层网络的科学 - 技术微观知识流动分析框架研究 [J]. 农业图书情报学报 , 2023, 35(11): 40-52.

doi:10.3772/j.issn.2095-915x.2024.05.008

基于 BERTopic 主题模型融合 RoBERTa 算法的短文本分类方法研究

刘桂锋 陈亦侯 包翔 韩牧哲

江苏大学科技信息研究所 镇江 212013

摘要：[目的 / 意义] 针对短文本分类中的稀疏问题，提出一种基于 BERTopic-RoBERTa-PCA-CatBoost 模型进行主题概率特征扩展的短文本分类方法。[方法 / 过程] 使用 RoBERTa 模型获取短文本的词向量表示，使用 BERTopic 主题模型提取主题概率特征向量，二者融合进行特征扩展，最后通过 CatBoost 算法分类。[局限] 在分类层面，未使用深度学习算法进行验证；在特征融合层面，未来可以考虑其他的特征融合方法。[结果 / 结论] 提出的 BERTopic-RoBERTa-PCA-CatBoost 模型与 LDA-CatBoost 模型相比在准确率上提升 10.90%，精确率上提升 10.91%，召回率上提升 10.68%。基于主题概率特征扩展的短文本分类方法能够克服单一模型的不足，提高短文本分类的效果。

关键词：短文本分类；词向量；BERTopic 模型；RoBERTa 模型

中图分类号：TP39；G35

Research on Short Text Classification Method Based on BERTopic Topic Modeling and RoBERTa Algorithm

LIU Guifeng CHEN Yihou BAO Xiang HAN Muzhe

Institute of Scientific and Technical Information, Jiangsu University, Zhenjiang 212013, China

Abstract: [Purpose/Significance] To address the sparsity issue in short text classification, this paper proposes a short text classification method based on topic probabilistic feature expansion with BERTopic-RoBERTa-PCA-CatBoost model. [Methods/Processes] The RoBERTa model is employed to obtain word vector representations of short texts. Topic probabilistic feature vectors are extracted using BERTopic topic model, which is then fused with word vectors for feature expansion. Finally, the

基金项目 2024 年江苏省研究生科研与实践创新计划项目“基于 BERTopic 主题概率特征扩展的新闻短文本分类方法研究”（2385）；国家社会科学基金一般项目“科学数据融合模式设计与体系建构研究”（21BTQ080）。

作者简介 刘桂锋（1980-），博士，教授，主要研究方向为科研数据管理、大数据分析、专利分析，E-mail：liuguifeng29@163.com；陈亦侯（1998-），硕士研究生，主要研究方向为机器学习、文本挖掘；包翔（1991-），博士研究生，馆员，主要研究方向为机器学习、文本挖掘；韩牧哲（1990-），博士，讲师，主要研究方向为数字人文、知识管理。

引用格式 刘桂锋 , 陈亦侯 , 包翔 , 等 . 基于 BERTopic 主题模型融合 RoBERTa 算法的短文本分类方法研究 [J]. 情报工程 , 2024, 10(5): 85-98.

doi:10.3772/j.issn.2095-915x.2024.05.008

CatBoost algorithm is utilized for classification. [Limitations] In terms of classification, deep learning algorithms have not been utilized for verification. Regarding feature fusion, future work may consider alternative feature fusion methods. [Results/Conclusions] The proposed BERTopic-RoBERTa-PCA-CatBoost model demonstrates improvements of 10.90% in accuracy, 10.91% in precision, and 10.68% in recall compared to LDA-CatBoost model. The short text classification method based on topic probabilistic feature expansion can overcome the limitations of individual models and enhance the effectiveness of short text classification.

Keywords: Short Textbook Classification; Word Vector; BERTopic Model; RoBERTa Model

引言

近年来短文本数据呈现出一种爆发式的增长态势，例如新闻媒体平台每天都会产生海量的数据，导致这些平台对于信息分类的需求也日益提高。传统的人工数据标注流程耗时过长、质量低下，且受到标注人员主观意识的影响，使得自动化短文本分类方法逐渐取代人工方法，但在处理短文本数据时存在准确率不高、适应性不强等问题。一方面，短文本包含的词语较少，导致传统的文本特征提取方法难以有效捕捉到短文本的语义信息。另一方面，短文本可能生成高维稀疏的特征向量，导致“维数灾难”，增加分类模型的计算复杂度，降低分类效果。基于主题词集等特征扩展方式的短文本分类方法由此受到广泛关注。尽管已有不少研究取得了进展，但此类方法容易导致对短文本主题词的提取不直接、不明确，容易引入噪声等问题。

因此，本研究提出一种基于 BERTopic 主题模型融合 CatBoost 算法的短文本分类方法，具体而言，该方法利用 RoBERTa 模型的文本表示能力来捕获短文本的深层语义信息，经主成分分析降维后能够有效处理高维稀疏矩阵，再通过 BERTopic 主题模型对短文本进行主题建模以发掘短文本中蕴含的潜在主题信息，最后将二者拼接得到的特征向量输入到 CatBoost 分类器进行分类学习。本研究的主要创新之处在于提出一种新的特征提取和特征融合方法，旨在将短文本的主题信息和语义信息有效融入分类模型中，构建出效果更好的短文本分类模型。利用该方法对短文本进行自动分类，不仅有助于信息的有效组织和检索，还可以为信息推荐、情感分析、舆情监控等任务提供支持。

1 相关研究

1.1 传统的短文本表示与分类算法

短文本分类是一个按照一定的分类体系或标准，对短文本数据进行自动化分类的过程。这个任务涉及短文本表示和分类算法两大方面[1]。短文本表示通过提取短文本数据的特征信息，生成包含短文本内部信息的表示，随后将这些表示输入到分类器中按照某种规则进行短文本分类[2]。常见的短文本表示方法包括词袋法如 TF-IDF[3]，词嵌入法如 Word2Vec[4]、Glove[5] 和 FastText[6] 等，基于 Transformer[7]

的多层双向编码器 BERT[8] 的方法以及改进的 BERT 方法如 RoBERTa[9]。常见无监督分类算法包括 K 均值聚类、密度聚类、层次聚类等 [10]。常见带监督分类算法包括支持向量机（SVM）、逻辑回归等 [10]。此外，集成分类算法通过将上述多个弱分类器集成来提高预测精度，比如随机森林算法 [10]、LightGBM[11]、XGBoost[12] 和 CatBoost[13]。

1.2 基于外部语料库的短文本特征扩展方法

短文本数据往往长度较少，信息量不多，这导致它们的特征较为稀疏，使得传统的短文本表示方法若直接结合分类算法，效果并不理想 [14]。为了提高短文本分类算法的分类效果，越来越多的学者尝试通过短文本特征扩展方式扩充语义特征空间 [15]。基于维基百科等外部语料库和知识库的特征扩展方法在一定程度上扩充了信息量，解决了短文本内容不足的问题。张巍琦 [16] 提出了一种基于知识图谱的短文本特征拓展方法。刘懿霆 [17] 提出了基于维基百科的文本样本扩展方法。Wensen 等 [18] 提出了基于 Wikipedia 和 Word2Vec 的特征扩展方法。但是通过外部语料和知识库扩展短文本特征耗时较长，且受限于数据库的质量和专业性，所以部分学者尝试利用基于文本内容自身的特征扩展方法来解决上述问题。

1.3 基于文本自身内容的短文本特征扩展方法

基于文本内容自身的特征扩展方法指的是对短文本内容的词频、主题、语义相关性等属性进行分析挖掘的过程，以往的研究较多集中于从高频词集、主题词集和词汇关联词集等内容出发构建内部语料 [19]。比如，Gu 等 [20] 提出了一种基于关键词扩展的短文本分类算法。李心蕾等 [21] 利用 Word2Vec 和 Sent2Vec 算法生成新浪微博的文本向量化表示形式，来提升分类效果并降低计算成本。曾子明等 [22] 利用 LDA 主题模型提取微博文本主题作为谣言识别模型训练的文档 - 主题特征，并结合用户可信度和微博影响力特征，使用随机森林方法进行谣言识别。唐晓波等 [23] 分别使用 TF-IDF 和 LDA 提取关键词，并利用 Word2Vec 对关键词进行特征扩展，应用于医疗问答社区中的健康问题分类。然而，以往基于短文本主题等内容的特征扩展方法更依赖于人工构建的主题词集质量。主题词集通常针对特定任务进行构建，导致模型在泛化能力方面有所欠缺，也就是说当面对新数据或新任务时，需要重新构建主题词集。

针对上述问题，本文主要采用一种基于主题概率特征扩展的短文本分类方法，设计短文本主题信息和语义信息的特征提取方法，以及将主题特征和语义特征进行拼接的特征融合方法，完成对短文本主题和语义层面特征的扩充。

2 基于主题概率特征扩展的短文本分类的研究思路设计

本研究提出的基于主题概率特征扩展的短文本分类的研究框架如图 1 所示，主要工作分为 3 个部分：（1）设计针对短文本语义信息的特征提取方法；（2）设计针对短文本主题信息的特征提取方法；（3）设计融合主题特征和语义特征的特征向量拼接方法。

doi:10.3772/j.issn.2095-915x.2024.05.008

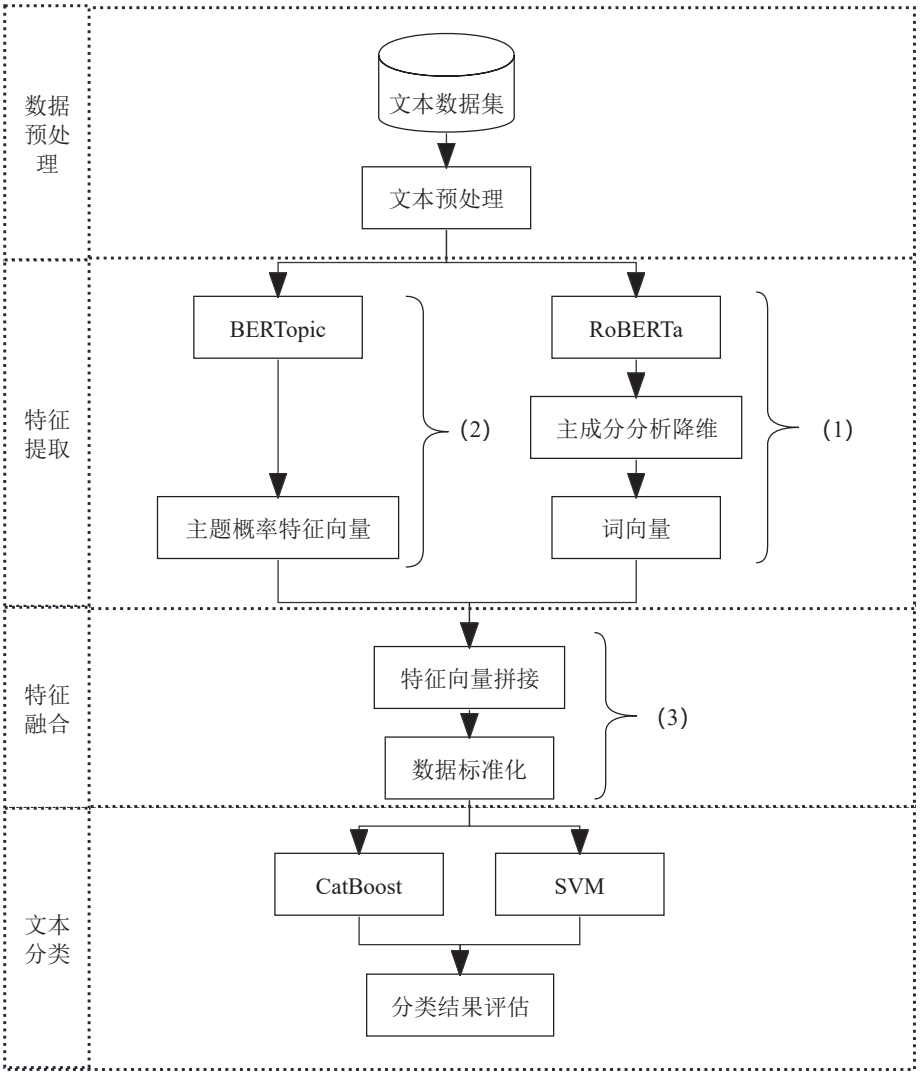


图 1 基于主题概率特征扩展的短文本分类的研究框架

2.1 短文本主题信息和语义信息的特征提取方法设计

本文尝试通过 RoBERTa 预训练模型获取短文本的动态词向量。RoBERTa 预训练模型是一种改进的训练 BERT 模型的方法，在使用 Transformer 进行特征抽取的基础上，结合自注意力机制考虑上下文语义关联，获取动态词向量，从而解决静态词向量无法表示多义词的问题。在此基础上，利用主成分分析将 RoBERTa 模型的输出降维，通过主成分分析法的曲线图确定阈值，确保大部分特征信息得到保留。主成分分析法降维对于下个阶段的特征融合来说确保了主题特征和语义特征的相对平衡，对于分类模型来说则去掉了噪点信息，能够大幅度减少模型的训练时间并且提高模型的效果。针对短文本语义信息的特征提取方法如图 2 所示。

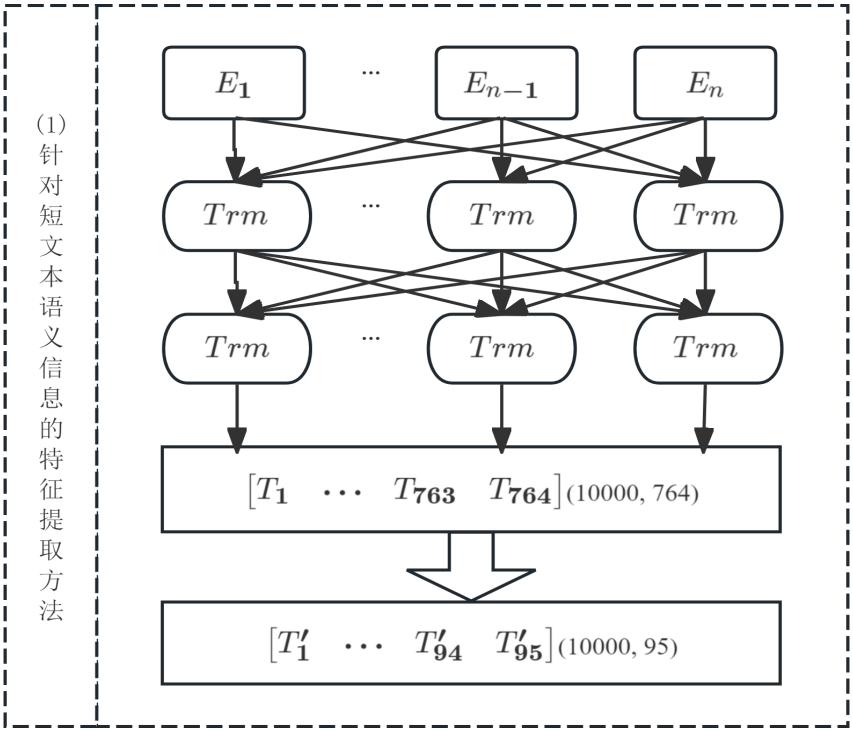


图 2 短文本语义信息的特征提取方法

其中 E_1, …, E_{n-1}, E_n 为词嵌入层的输入，*Trm* 为 Transformer 编码器，[T_1, …, T_{763}, T_{764}] 为 RoBERTa 模型输出的短文本语义特征向量组成的特征矩阵，（10000, 764）是矩阵维度；[T'_1, …, T'_{94}, T'_{95}] 为经过主成分分析降维后的特征矩阵，（10000, 95）是矩阵维度。

本文尝试使用 BERTopic[24] 模型获取短文本的主题向量。首先，通过 BERT 或其他嵌入技术将文档转换为向量表示；再对 BERT 模型处理好的数据使用 UMAP 算法降维后通过 HDBSCAN 算法进行无监督的聚类，以将文档划分为不同的类别；最后结合 C-TF-IDF 方法提取主题词汇，以生成紧密相关的主题群，最后使用模型输出的文档主题概率特征向量来表征短文本。短文本主题信息的特征提取方法如图 3 所示。

其中，BERT 层嵌入文档，UMAP 和 HDBSCAN 层聚类文档，C-TF-IDF 层生成主题表示；[$topic_1$, …, $topic_5$, $topic_6$] 为 BERTopic 模型输出的短文本主题特征向量组成的特征矩阵，（10000, 6）是矩阵维度。

2.2 融合主题特征和语义特征的特征向量拼接方法设计

本文所提出的特征融合方法是使用 BERTopic 主题模型提取出的主题概率特征集合 [$topic_1$, …, $topic_5$, $topic_6$]，结合 RoBERTa 预训练模型提取的经主成分分析降维后的词向量集合 [T'_1, …, T'_{94}, T'_{95}] 进行特征融合，达到一种信息互补的效果。为了得到一个融合多粒度特征的文本表示，通过特征拼接的方式融合 2 种类型的特征，进而得到文本的多粒度特征 F，如式（1）所示。本文选择的特征拼接方法是一种无参数的方法，它不需要预先设定融合权重或进行复

doi:10.3772/j.issn.2095-915x.2024.05.008

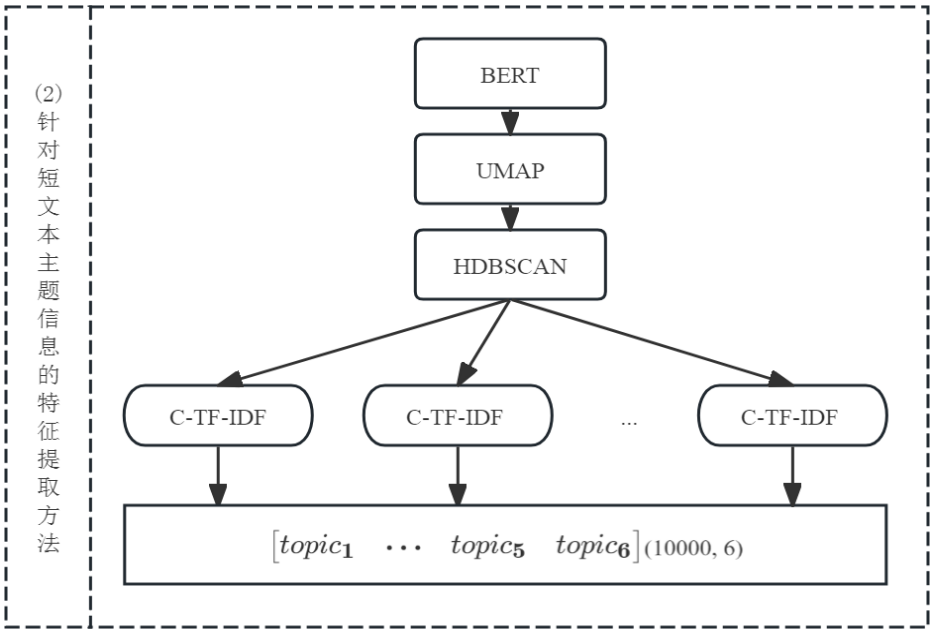


图 3　短文本主题信息的特征提取方法

杂的优化过程，因此更容易实施且能够灵活地适应不同的特征类型和数量。由于这两组特征经过不同的处理，尺度和分布可能会有所不同。为确保模型能够平等地考虑每个特征，对拼接后的特征向量进行归一化处理。

$$F=[topic_1, \ldots, topic_5, topic_6, T'_1, \ldots, T'_{94}, T'_{95}] \quad (1)$$

其中，$[topic_1, \ldots, topic_5, topic_6]$ 为 *BERTopic* 模型输出的短文本主题特征向量组成的主题特征矩阵；$[T'_1, \ldots, T'_{94}, T'_{95}]$ 为经过主成分分析降维后的 *RoBERTa* 模型提取的短文本语义特征向量组成的语义特征矩阵。

3　实验过程

3.1　数据来源以及数据预处理

本文选择清华大学中文文本语料库 THUCNews[25] 数据集作为数据来源，从中抽取财经、家居、教育、科技、社会、时尚、时政、体育、游戏、娱乐共 10 个分类类别，每个类别分别抽取 1000 条样本，共 10000 个样本，按照 8:2 的比例划分训练集和测试集。

新闻短文本数据集的主成分分析 2D 投影图如图 4 所示，图中 x 轴和 y 轴分别代表第一主成分和第二主成分，也就是最大化保留数据特征的两个主成分方向。类别 9（娱乐）在图中呈现出较强的聚集性，对数据点的聚类起到关键作用，而类别 2（教育）等较为分散，对聚类的影响较小。

本实验所用的环境为 Python3.9，数据预处理环节使用 Python 中的 pandas2.2.0 包对数据进行初步的清洗去重后，加载哈尔滨工业大学停用词表、四川大学机器智能实验室停用词库等多个停用词词典，使用 Python 中的 jieba0.42.1 包对摘要文本进行分词，并去除文本中所包含的停用词和其他无意义字符。主题建模环节主

要通过 bertopic0.16.0 和 gensim4.3.2 包来进行模型训练与输出；分类器构建环节选用 scikit-learn1.4.0、catboost2.0 包；结果可视化呈现通过 matplotlib3.7.2 包来实现。

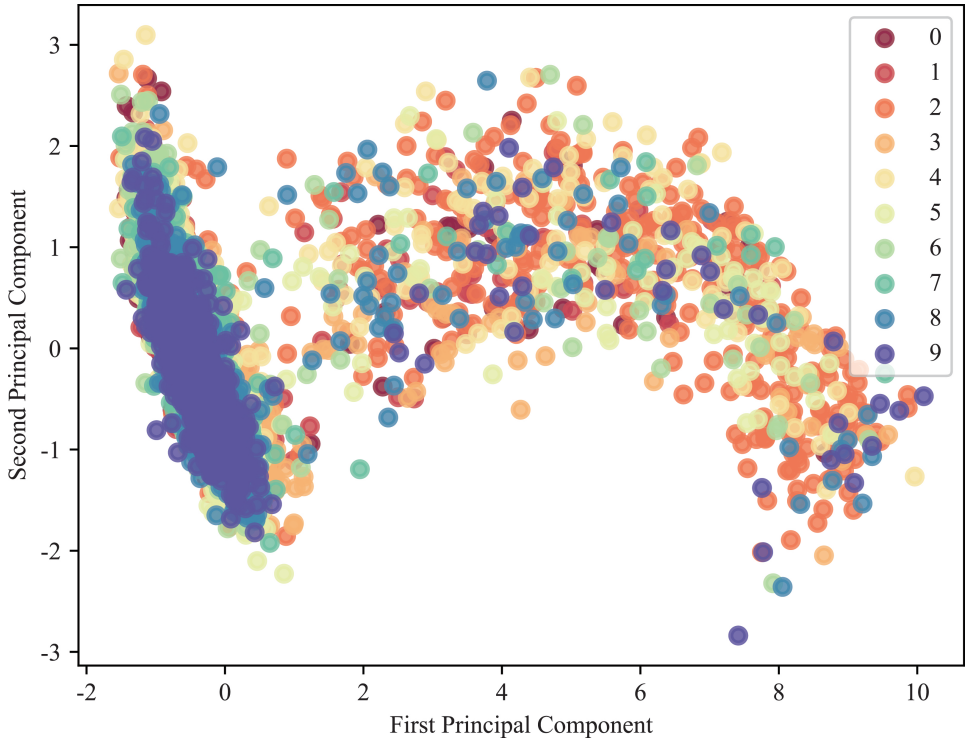


图 4 新闻短文本数据集的主成分分析 2D 投影图

3.2 主题模型参数设置

BERTopic 模型的具体参数如下：嵌入文档，使用中文文本的预训练词向量模型“bert_base_chinese”；初始化 UMAP 和初始化 HDBSCAN 过程使用默认参数；主题数设为 nr_topics =“auto”，在进行主题提取时不限定主题数，这是因为 BERTopic 模型在建模过程中会自动选择最佳参数。

Top2vec 模型的具体参数如下：嵌入文档，使用针对多语言（包括中文）的预训练模型“distiluse-base-multilingual-cased”；设置 min_count=10，表示在构建共现矩阵时，一个词语至少需要出现在 10 个文档中才会被考虑；设置模型训练所需的速度 speed=“learn”，平衡训练速度和训练质量；指定用于并行训练的线程数 workers=8，加快模型训练速度。

LDA 主题模型困惑度的计算通过模型训练时返回的概率分布矩阵得出，LDA 模型在不同主题数下的困惑度指标，如图 5 所示。根据结果选择困惑度最小的主题个数作为模型主题数，LDA 模型最优主题个数 Optimal number of topics = 29；random_state = 42 控制随机数生成的种子，用于确保结果的可重复性。其他参数设置为默认参数。

doi:10.3772/j.issn.2095-915x.2024.05.008

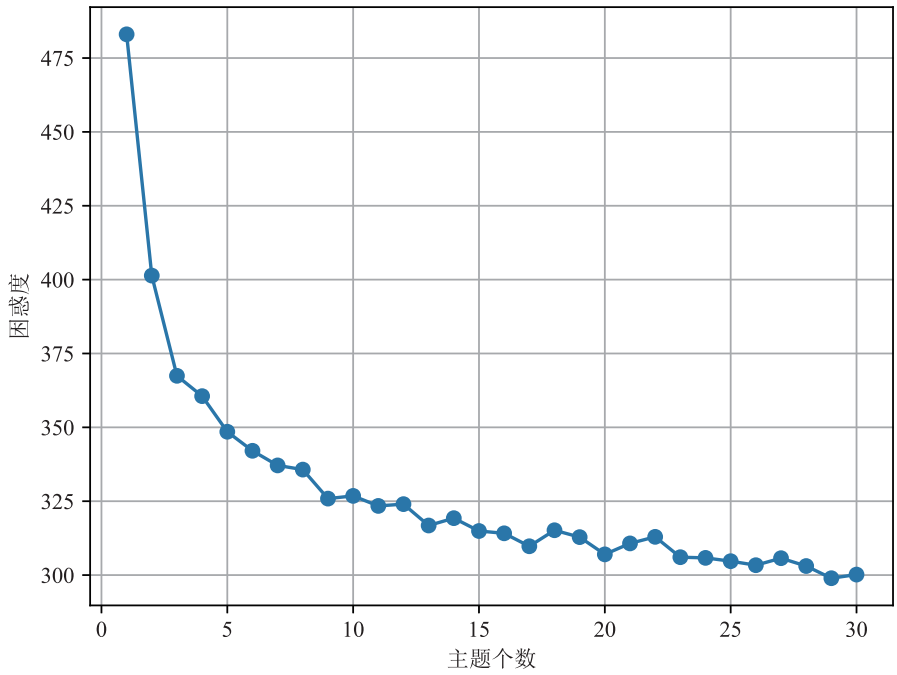


图 5 LDA 模型的困惑度

3.3 其他参数设置

图 6 是主成分分析的方差解释率图。其中每个点表示一个主成分，y 轴值表示该主成分对总方差的解释比例。随着主成分数量的增加，累积方差解释率通常会逐渐增加，但增长速度会逐渐放缓。曲线的拐点通常表示增加更多主成分对总方差的解释增加不再显著。因此，通过选择曲线拐点附近的点，将参数设置为要保留的方差百分比 95% 作为阈值。

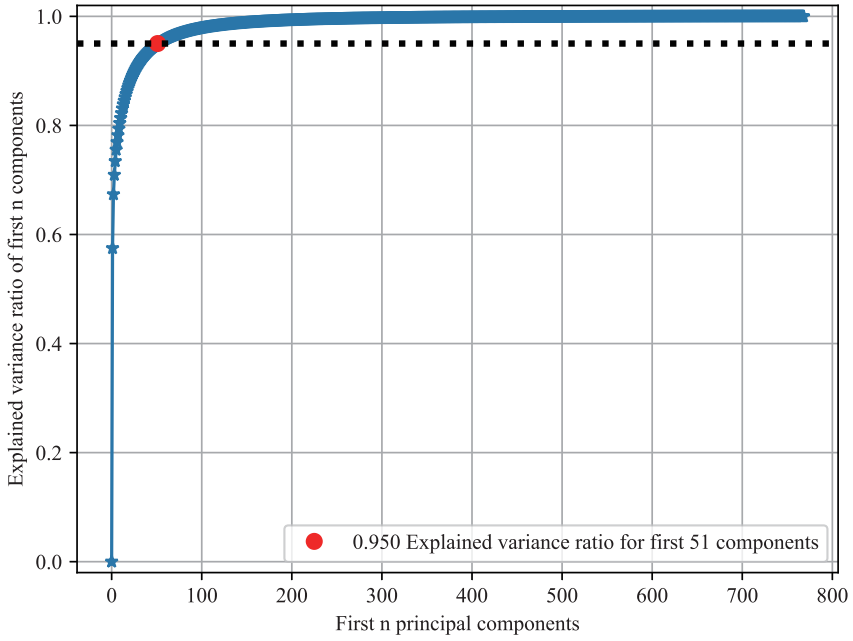


图 6 主成分分析的方差解释率图

CatBoost 分类器参数设置如下：iterations 为迭代次数，默认为 1000；learning_rate 表示模型的学习率，设置 RoBERTa-CatBoost（降维前）为 0.087382，设置其他模型为 0.087979；其余参数均使用默认参数。

4 实验结果分析

4.1 词向量模型与主题模型融合的有效性实验

为验证文本特征融合方法的有效性，将其与前人的分类方法进行对比。由于 LDA[26]、Top2vec[27] 作为经典的主题模型方法被广泛应用，TF-IDF 模型是传统的文本向量表征方式之一，同时支持向量机（SVM）等经典机器学习算法在各种分类任务中展现出优越的性能。因此，分别设置 LDA-CatBoost 和 LDA-SVM 为使用 LDA 模型所表征的文本特征向量，结合 CatBoost 和 SVM 分类器进行分类对比，也就是刘爱琴等 [28] 提出的主题模型融合分类算法的文本自动分类方法。分别设置 Top2vec-CatBoost 和 Top2vec-SVM 为使用 Top2vec 模型所表征的文本特征向量，结合 CatBoost 和 SVM 分类器进行分类对比。分别设置 TF-IDF-CatBoost 和 TF-IDF-SVM 为使用 TF-IDF 模型所表征的文本特征向量结合 CatBoost 和 SVM 分类器进行分类对比。设置 BERTopic-RoBERTa-CatBoost 为文本所提出的方法，即结合经 PCA 降维后的 RoBERTa 模型的词向量表示与 BERTopic 模型的主题向量表示进行特征向量拼接，输入到 CatBoost 分类器进行分类。在 THUCNews 新闻数据集上运用上述方法得到的测试结果的准确率、精确率、召回率、F1 值、AUC 值比较结果如表 1 所示。

表 1 对比实验结果

模型	准确率	精确率	召回率	F1 值	AUC 值
LDA-CatBoost	0.7455	0.7469	0.7480	0.7455	0.9615
LDA-SVM	0.7350	0.7346	0.7375	0.7350	0.9499
Top2vec-CatBoost	0.6885	0.6900	0.6873	0.6866	0.9395
Top2vec-SVM	0.6675	0.6803	0.6660	0.6659	0.9324
TF-IDF-CatBoost	0.6535	0.7384	0.6546	0.6727	0.9325
TF-IDF-SVM	0.6040	0.7685	0.6077	0.6359	0.9406
BERTopic-RoBERTa-PCA-CatBoost	**0.8545**	**0.8560**	**0.8548**	**0.8550**	**0.9772**
BERTopic-RoBERTa-PCA-SVM	0.7840	0.7897	0.7850	0.7865	0.9646

对表 1 中的实验结果进行分析得出如下结论：

本文所提出的 BERTopic-RoBERTa-PCA-CatBoost 模型相对于 LDA-CatBoost 模型在分类的准确率上提升 10.90%，精确率上提升 10.91%，召回率上提升 10.68%，说明本文方法能够解决短文本分类中的稀疏问题，进一步提升分类效果。实验结果表明相较于支持向量机模型，结合集成算法 CatBoost 模型的整体分类效果更高，这说明了

doi:10.3772/j.issn.2095-915x.2024.05.008

使用集成学习算法构建分类器更适用于本文所提出的特征融合场景。根据实验结果，选择效果最优的 BERTopic-RoBERTa-PCA-CatBoost 作为本文的最终分类模型。

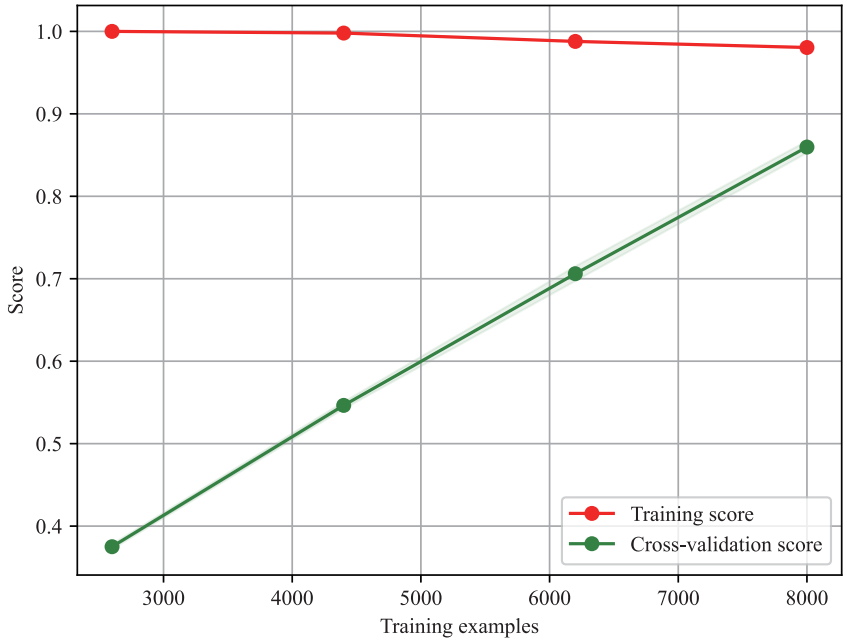


图 7 BERTopic-RoBERTa-PCA-CatBoost 模型的学习曲线图

BERTopic-RoBERTa-PCA-CatBoost 模型的学习曲线图如图 7 所示，随着样本数量的增加，模型在测试集上的交叉验证分数呈明显上升趋势，说明模型的效果在稳步提升，并且在训练完所有样本后达到最高点。另外，随着样本数量的增加，模型在训练集上的分数略下降，说明过拟合的风险降低了。这意味着模型不再过度拟合训练数据中的噪声或细节。从图 7 中可以推断出，若继续扩大样本数量，该模型的效果将进一步提升。

BERTopic-RoBERTa-PCA-CatBoost 模型在短文本分类任务中的优势主要来自三个方面：主题模型 BERTopic 更能有效地捕捉文本中的主题信息；RoBERTa 作为预训练模型，其强大的语义表示能力使得文本的动态上下文语义特征得以充分提取；而 CatBoost 作为分类器，则能基于融合特征进行准确分类。在新闻短文本分类的实际应用中，BERTopic-RoBERTa-PCA-CatBoost 模型提高了分类的准确性。

BERTopic-RoBERTa-CatBoost 模型的 ROC 曲线图以及精确率 — 召回率曲线图分别如图 8、9 所示。精确率 — 召回率曲线展示了不同类别数据在精确率和召回率这两个维度上的综合表现，图 9 中曲线整体下降的趋势明显，因为精准率和召回率是两个相互制约的指标，随着精准率逐渐增大召回率会逐渐地减小，曲线拐点位置是精准率和召回率的平衡位置。从图 8 和图 9 中可以看出，类别 0（财经）、类别 3（科技）和类别 9（娱乐）的曲线较为理想，表明分类器在这些类别上分类效果和性能较好。而类别 8（游戏）的曲线则相对较低，说明分类器在该类别上分类效果和性能均有待提升。

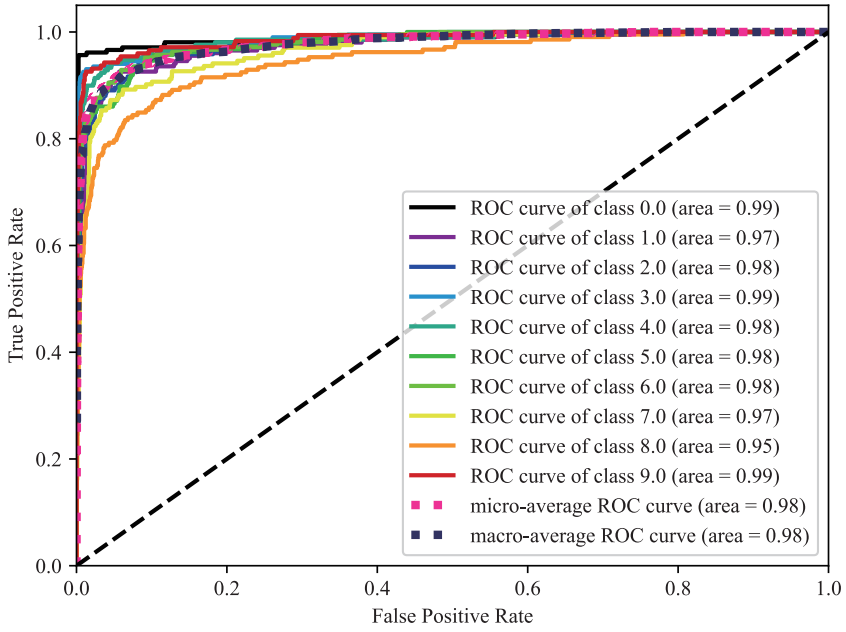


图 8 BERTopic-RoBERTa-CatBoost 模型的 ROC 曲线图

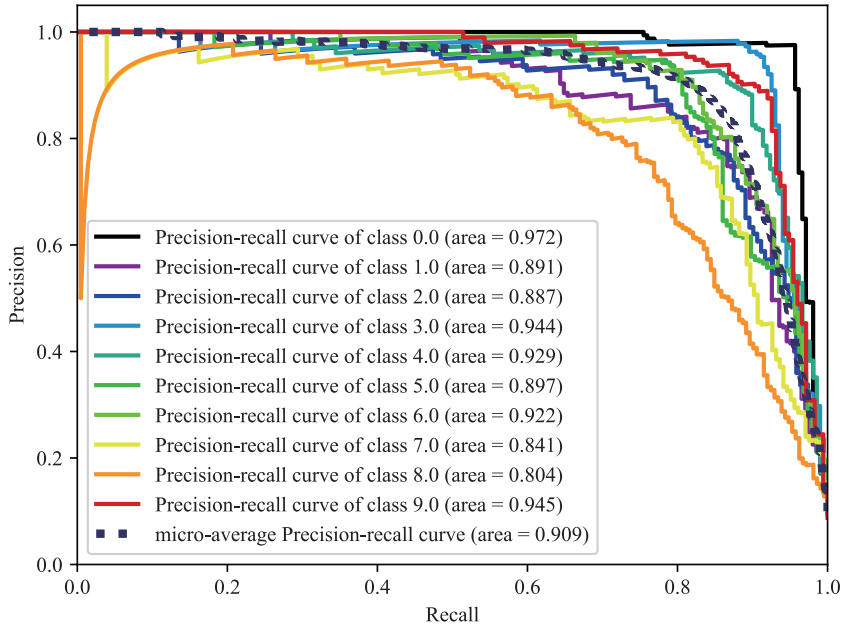


图 9 BERTopic-RoBERTa-CatBoost 模型的精确率 - 召回率曲线图

4.2 基于BERTopic-RoBERTa-PCA-CatBoost模型的消融实验

为了验证 BERTopic-RoBERTa-PCA-CatBoost 模型中各组件的有效性，并评估组件对整体性能的贡献，对模型进行消融实验。设置 RoBERTa 为去除主题模型 BERTopic 和 PCA 模块的模型结构；设置 RoBERTa-PCA-CatBoost 为去除主题模型 BERTopic 的模型结构；设置 BERTopic-CatBoost 为去除预训练模型 RoBERTa 模块的模型结构。基于 BERTopic-RoBERTa-PCA-CatBoost 模型的消融实验比较结果如表 2 所示。

doi:10.3772/j.issn.2095-915x.2024.05.008

表 2　消融实验比较结果

模型	准确率	精确率	召回率	F1 值	AUC 值
RoBERTa-Catboost	0.6340	0.6344	0.6352	0.6328	0.9244
RoBERTa-PCA-CatBoost	0.6235	0.6275	0.6252	0.6247	0.9175
BERTopic-CatBoost	0.8155	0.8398	0.8178	0.8183	0.9687
BERTopic-RoBERTa--PCA-CatBoost	**0.8545**	**0.8560**	**0.8548**	**0.8550**	**0.9772**

从表 2 中可以看出，RoBERTa-PCA-CatBoost 模型比 RoBERTa-Catboost 模型在准确率上降低 1.05%，在 F1 得分方面降低 0.81%，在 AUC 值上降低 0.69%。这说明主成分分析成功保留了大部分的特征，对分类模型效果的影响很小。利用主成分分析法降维的原因在于 RoBERTa 模型输出的特征数量远超主题模型，如果直接进行特征拼接的话，BERTopic 模型析出的主题信息可能被忽略，起不到特征融合的意义。因此，加入主成分分析这一特征提取方法是进行有效特征融合的前提，使分类模型更为均衡地考虑到各个层面的特征，从而提升分类效果。

实验结果表明本文提出的 BERTopic-RoBERTa-PCA-CatBoost 模型相较于 BERTopic-CatBoost 模型在准确率、召回率、F1 得分上分别提升了 3.90%、3.67% 和 0.85%；相较于 RoBERTa-PCA-CatBoost 模型的方法在准确率、召回率、F1 得分上分别提升了 23.10%、23.03% 和 5.97%，验证了本文所提出的词向量模型与主题模型融合方法的有效性。由于新闻短文本篇幅有限，文本中词汇的共现模式较为显著，分类时往往需要对文本进行深层次的语义理解。主题模型如 BERTopic 能够有效地捕捉这些共现模式，所以分类效果相对较优；而依赖文本表示的分类方法在处理具有复杂语义的新闻短文本时，可能对语义理解得不够充分，因此分类效果相对较差。本文使用主题模型 BERTopic 提取无监督聚类结果，结合词向量模型 RoBERTa 经主成分分析降维后做了特征融合实现的监督分类算法，二者融合达到了信息互补的效果，因此效果最优。

5　结论

本文主要提出一种基于主题概率特征扩展的短文本分类方法，通过设计短文本主题信息和语义信息的特征提取方法，以及融合主题特征和语义特征的特征向量拼接方法，完成对短文本主题和语义层面的特征扩展，省去了人工构建主题词集的人力成本。这为情报学提供了一种新的文本分类视角和理论支持。同一文本数据往往蕴含着多重维度的特征分布，包括语义特征、结构特征、情感特征等，特征的多样性和丰富性使得特征融合技术尤为重要。通过本文所提出的特征提取和特征融合方式能够更全面、更准确、更有效地整合多种类型的特征，在情报学信息推荐、信息检索、情感分析、舆情监控等领域具有广泛的应用前景。

实验结果表明，融合主题特征向量和词向

量能够对短文本分类效果产生正向影响，这在一定程度上克服了短文本内容的稀疏问题，提高分类的准确率。此外，引入主成分分析降维不仅有助于特征融合，还能够大幅缩短分类器的迭代时长，提升分类效率。未来可以通过扩大数据集、探索更多的特征提取方法和分类算法，尝试将本文提出的短文本自动分类技术应用于更多的情报分析场景，包括：（1）在分类层面，未来的研究中还可以尝试使用深度学习模型如卷积神经网络 CNN、循环神经网络 RNN、长短期记忆网络 LSTM 等分类器进行验证，比较和评估该方法结合不同的分类算法的分类准确率和性能；（2）在特征融合层面，未来可以考虑其他的特征融合方法，如 TextCNN 的融合门机制等方式。

参 考 文 献

[1] GE J W, LIN S C, FAN Y Q. A Text classification algorithm based on topic model and convolutional neural network[J]. Journal of Physics: Conference Series, 2021, 1748(3): 32-36.

[2] 许和旭 , 王兰成 . 基于语料库文本自动分类算法及应用比较研究 [J]. 图书情报导刊 , 2021, 6(6): 45-53.

[3] JONES K S. Index term weighting[J]. Information storage and retrieval, 1973, 9(11): 619-633.

[4] MIKOLOV T. Efficient estimation of word representations in vector space[J]. arxiv preprint arxiv:1301.3781, 2013.

[5] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[6] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Fasttext. zip: Compressing text classification models[J]. arXiv preprint arXiv:1612.03651, 2016.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, 12: 6000-6010.

[8] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[9] ADOMA A F, HENRY N M, CHEN W. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition[C]//2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE, 2020: 117-121.

[10] 周志华 . 机器学习 [M]. 北京 : 清华大学出版社 , 2006: 73-224.

[11] KE G, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 3149-3157.

[12] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[13] HANCOCK J T, KHOSHGOFTAAR T M. CatBoost for big data: an interdisciplinary review[J]. Journal of big data, 2020, 7(1): 94.

[14] GE J W, WANG H X, FANG Y Q. Short Text Classification Method Combining Word Vector and WTTM[C]// IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference. Chongqing: IEEE, 2020: 1994-1997.

[15] 王连喜 . 微博短文本预处理及学习研究综述 [J]. 图书情报工作 , 2013, 57(11): 125-131.

[16] 张巍琦 . 基于知识图谱的短文本分类研究 [D]. 成都 : 电子科技大学 , 2020.

[17] 刘懿霆 . 基于维基百科的文本样本扩展方法及其应用研究 [D]. 上海 : 上海大学 , 2019.

[18] WENSEN L, ZEWEN C, JUN W, et al. Short text classification based on Wikipedia and Word2vec[C]//2016 2nd IEEE International

doi:10.3772/j.issn.2095-915x.2024.05.008

Conference on Computer and Communications (ICCC). IEEE, 2016: 1195-1200.

[19] 邵云飞 , 刘东苏 . 基于类别特征扩展的短文本分类方法研究 [J]. 数据分析与知识发现 , 2019, 3(9): 60-67.

[20] GU Y, SHEN J. Short text classification based on keywords extension[C]//2019 Chinese Automation Congress (CAC). IEEE, 2019: 2616-2621.

[21] 李心蕾 , 王昊 , 刘小敏 , 等 . 面向微博短文本分类的文本向量化方法比较研究 [J]. 数据分析与知识发现 , 2018, 2(8): 41-50.

[22] 曾子明 , 王婧 . 基于 LDA 和随机森林的微博谣言识别研究——以 2016 年雾霾谣言为例 [J]. 情报学报 , 2019, 38(1): 89-96.

[23] 唐晓波 , 高和璇 . 基于关键词词向量特征扩展的健康问句分类研究 [J]. 数据分析与知识发现 , 2020, 4(7): 66-75.

[24] EGGER R, YU J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts[J]. Frontiers in sociology, 2022, 7: 886498.

[25] 孙茂松 , 李景阳 , 郭志芃 , 等 . THUCTC：一个高效的中文文本分类工具包 [EB/OL]. (2016-01-01) [2023-12-31]. http://thuctc.thunlp.org/.

[26] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[27] ANGELOV D. Top2vec: Distributed representations of topics[J]. arXiv preprint arXiv:2008.09470.

[28] 刘爱琴 , 郭少鹏 , 张卓星 . 基于 LDA 模型融合 Catboost 算法的文本自动分类系统设计与实现 [J]. 国家图书馆学刊 , 2023, 32(5): 84-92.

doi:10.3772/j.issn.2095-915x.2024.05.009

开放科学（资源服务）标识码（OSID）

基于大模型的科研设备成本评估框架

程齐凯[1,2] 雷道宇[1,2] 石湘[1,2] 刘寅鹏[1,2]

1. 武汉大学信息管理学院 武汉 430072；
2. 武汉大学信息检索与知识挖掘研究所 武汉 430072

摘要：[目的 / 意义] 提出一种创新的基于大模型的科研设备成本评估框架，旨在解决传统成本评估方法中的局限性，如成本评估的不精确性和效率低下问题。自动化地从科研论文中抽取实验材料与设备信息，并设计科研设备成本估算模型，从而精准和高效地评估科学研究成本，为实验成本的精确评估和科研资源的有效利用提供了新的工具和方法。[方法 / 过程] 以物理和计算机领域为例，利用 arXiv 数据库与 Paper With Code 网站提供的论文数据构建了一个训练数据集，并采用 LoRA 微调技术在基准模型 LLaMA2-13b 上进行微调，使其能够精确抽取目标领域论文中关于实验设备与材料的详细信息。通过 Wikipedia 进行实体链接消歧，并综合考虑材料设备的价格波动，设计了一种平均情况分析的成本估算公式，以计算机视觉领域为例对科研设备成本评估框架的有效性进行验证。[局限] 只在计算机领域和物理领域进行了实验，同时数据集的构建主要依赖于公开可获取的论文数据，这可能限制了成本评估框架的泛化能力和准确性。[结果 / 结论] 通过对计算机科学与物理学领域的科研论文进行实证分析，展示了基于大模型的科研设备成本评估框架的有效性。通过 LoRA 技术微调的 LLaMA2 模型在信息抽取任务上显示出较高的准确率和召回率，证明了本框架在精准抽取实验材料与设备信息方面的能力。同时，在计算机视觉领域开展了成本估算分析，揭示了计算资源已经成为制约计算机视觉领域科研产出的关键因素之一和特定的算法模型结构或研究范式存在性能上限等结论。这些发现与实际科研活动相吻合，证明了本文提出的成本评估框架能够准确反映科研实践的现实情况，为科研项目的资源优化提供了重要参考。

关键词：信息抽取；大模型；高效微调；成本估算框架

中图分类号：G35；TP391

Scientific Research Equipment Cost Evaluation Framework Based On Large Language Model

CHENG Qikai[1,2] LEI Daoyu[1,2] SHI Xiang[1,2] LIU Yinpeng[1,2]

基金项目 国家自然科学基金面上项目“基于机器阅读理解的科学命题文本论证逻辑识别”（72174157）；国家自然科学基金重点项目“数智赋能的科技信息资源与知识管理理论变革”（72234005）。

作者简介 程齐凯（1989-），博士，副教授，主要研究方向为文本挖掘、信息检索；雷道宇（1998-），硕士研究生，主要研究方向为自然语言处理，E-mail: leidaoyu@whu.edu.cn；石湘（1998-），博士研究生，主要研究方向为信息检索、知识管理；刘寅鹏（1998-），博士研究生，主要研究方向为文本挖掘、自然语言处理。

引用格式 程齐凯 , 雷道宇 , 石湘 , 等 . 基于大模型的科研设备成本评估框架 [J]. 情报工程 , 2024, 10(5): 99-114.

doi:10.3772/j.issn.2095-915x.2024.05.009

1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China

Abstract: [Objective/Significance] This study proposes an innovative framework for assessing the cost of scientific research equipment based on large language models, aiming to address the limitations of traditional cost assessment methods, such as the inaccuracy and inefficiency of cost estimation. By automatically extracting experimental material and equipment information from scientific research papers and designing a cost estimation model for research equipment, this framework provides a new tool and method for accurately and efficiently evaluating the cost of scientific research, enabling precise cost assessment and effective utilization of research resources. [Methods/Processes] Using physics and computer science as examples, this study constructs a training dataset based on the paper data provided by the arXiv database and the Paper with Code website. It employs the LoRA fine-tuning technique on the benchmark model LLaMA2-13b, enabling it to accurately extract detailed information about experimental equipment and materials from papers in the target domains. Entity linking disambiguation is performed using Wikipedia, and a cost estimation formula based on average-case analysis is designed, considering the price fluctuations of materials and equipment. The effectiveness of the research equipment cost assessment framework is validated using the field of computer vision as an example. [Limitations] Experiments were conducted only in the computer science and physics domains, and the construction of the dataset primarily relies on publicly available paper data, which may limit the generalizability and accuracy of the cost assessment framework. [Results/Conclusions] Through empirical analysis of scientific research papers in the fields of computer science and physics, this study demonstrates the effectiveness of the research equipment cost assessment framework based on large language models. The LLaMA2 model fine-tuned using LoRA technology exhibits high accuracy and recall in the information extraction task, proving the framework's ability to accurately extract experimental material and equipment information. Additionally, the study conducts cost estimation analysis in the field of computer vision, revealing that computational resources have become one of the key factors constraining research output in computer vision, and that specific algorithmic model structures or research paradigms have performance limits. These findings align with real-world scientific research activities, demonstrating that the proposed cost assessment framework can accurately reflect the realities of scientific practice and provide important references for optimizing resources in research projects.

Keywords: Information Extraction; Large Language Model; Efficient Fine-tuning; Cost Evaluation Framework

引言

科学实验是科研工作者探索未知、验证假设、积累知识的核心手段，也是推动科学理论发展和技术创新的直接动力。科技文献承载了研究者的思想和成果[1]，实验则是这些成果诞生和检验的实践场域。每项科学实验都伴随着一系列成本的评估，包括实验材料、设备和时间等资源的投入，这些成本不仅会影响研究的可行性和实验的规模，也是科研项目规划和决策过程中不可或缺的考量因素。尤其在一些高度专业化和技术密集的领域[2]，如物理和计算机领域，实验设备的成本和配置问题显得更为复杂，在进行研究之前必须采用更加精细化和动态化的成本评估方法，以确保科研资源的有效利用和项目的顺利进行。

近年来，信息技术的发展促进了成本管理工具的创新，一些研究通过开发软件和算法来帮助科研人员估算项目成本，提高决策效率，例如 Uddin 等[3]的一项研究展示了使用各种机

器学习算法（如 SVM、随机森林等）来探索项目成本超支的原因和频率的方法，并提出了数据驱动的成本超支情况预测系统。虽然当前研究为科研成本管理提供了重要洞见，但多数研究仅关注成本管理的宏观策略或针对特定领域的实施方式，缺少对科研过程中各个具体环节成本的详尽分析。特别是在成本密集的实验阶段，缺乏一种系统的评估与分析方法。为此，本研究提出了一种创新的科研设备成本评估框架，旨在自动化地从科研论文中抽取与实验成本相关的关键信息，根据设计的成本评估公式精准和高效地评估科研设备投入成本，这不仅填补了现有研究的空白，也为实验成本的精确评估和科研资源的有效利用提供了新的工具和方法。

本研究聚焦于计算机科学与物理学两个领域内的科研文献，利用 arXiv 数据库与 Paper With Code 网站所提供的论文数据构建了一个训练数据集，采用 LoRa 微调技术在基准模型 LLaMA2 上进行了微调，使其能够精确抽取计算机科学与物理学领域论文中关于实验设备与材料的详细信息，这一步骤为后续的成本分析奠定了坚实的基础。本研究根据平均情况分析的思想进一步设计了成本估算公式，对科研项目的设备成本进行了系统估算，通过在计算机视觉领域的实证分析，本研究发现论文平均研究成本呈指数级增长的趋势，同时不同模型架构和研究范式的性能提升存在客观上限。这些发现为科研项目的资源配置和预算管理提供了新的思路和依据，本研究的成果有望为科研管理者和决策者提供有价值的参考，推动科研资源的合理配置和科研事业的可持续发展。

1 相关研究

科研活动的成本评估是科研管理中的重要环节，需要综合考虑设备、材料、时间等多种要素，传统的成本评估方法难以有效应对科研活动的不确定性和创新性。近年来，人工智能技术的发展为科研成本评估提供了新的思路，通过自然语言处理等技术，可以从科技论文等非结构化文本中自动抽取与科研成本密切相关的信息，如材料、设备、实验参数等，并将其转化为结构化的知识表示，用于支撑科研项目的成本分析和管理。为此本节将从科研活动的成本评估方法和科技论文的信息抽取方法两个大方向对相关研究进行梳理。

1.1 科研活动的成本评估方法

科研活动的成本评估是科研管理中的重要环节，传统的科研活动成本评估方法主要包括专家评估法、类比估算法和参数估算法等 [4]。专家评估法依赖于专家的经验和判断，容易受到主观因素的影响；类比估算法通过与类似项目的对比来估算成本，但难以应对科研活动的创新性和不确定性；参数估算法建立了成本参数与影响因素之间的数学模型，但对参数的选择和量化存在挑战。这些方法往往依赖于手工收集和分析数据，难以适应科研活动日益增长的复杂性和数据量。

近年来，人工智能技术的发展为科研成本评估提供了新的思路。一些研究者尝试利用机器学习算法建立成本预测模型，通过对历史数据的训练来预测新项目的成本。Sajadfar 等 [5] 提出了一种结合特征工程和数据挖掘算法的成

doi:10.3772/j.issn.2095-915x.2024.05.009

本估算方法，利用线性回归和数据挖掘技术挖掘与企业资源计划（ERP）系统相关的制造过程数据，建立成本估算函数，该方法采用渐进式实施策略，提高成本估算的准确性和实用性；Liu 等 [6] 则基于支持向量回归（SVR）机器和粒子群优化（PSO）算法，建立了材料成本预测模型。该模型首先对实际数据进行预处理，然后利用 PSO 算法优化 SVR 的参数，进行数据挖掘和成本预测。研究表明，该模型能够很好地拟合实际材料成本数据，预测效果令人满意。然而，这些方法主要关注总体成本的预测，缺乏对科研活动中关键成本驱动因素的分析，如实验材料、设备等。此外，这些方法通常需要大量的历史项目数据作为训练样本，在一些新兴和交叉学科领域可能难以满足。

综上所述，现有的科研成本评估方法在应对日益复杂的科研活动时存在诸多局限，需要一种能够充分利用科技文献大数据、自动化获取关键成本信息、适应不同学科特点的新方法。近年来，随着自然语言处理技术的发展，特别是大型语言模型的出现，为材料设备信息的精准抽取提供了新的可能。相关技术如指令工程和高效微调使得大型语言模型能够在特定领域进行信息抽取，为材料设备成本的评估奠定坚实的数据基础。通过对科技文献的深入挖掘和分析，大型语言模型能够快速准确地识别和提取与研究成本密切相关的实验材料、设备等关键信息，并结合领域知识库实现成本的估算，从而为科研项目的成本评估和管理提供更加智能、高效的决策支持。本研究正是基于这一认识，提出了一种创新的基于大型语言模型的科研设备抽取和成本评估框架。

1.2 指令工程下的信息抽取技术

信息抽取任务（Information Extraction）是自然语言处理领域的重要任务之一，旨在从非结构化数据中抽取出结构化信息。传统的基于监督学习的信息抽取技术（如利用 CRF[7]、BERT[8] 等深度神经网络进行实体识别任务）大多遵循预训练 + 下游任务微调的范式，然而，这种传统方法的一个重要限制是对大量高质量标注数据的依赖，这不仅增加了成本，也限制了模型在面对复杂化数据时的适应性和扩展性。在论文材料设备信息抽取的实际任务场景中，往往缺乏大量高质量的训练数据，为此最新的信息抽取研究进一步探索利用指令工程（Prompt Engineering）、少样本学习（Few-shot Learning）等技术，在小样本、低资源场景下实现信息抽取，降低人工标注的成本。

随着 GPT-3.5[9] 模型的推出，开启了自然语言处理大模型元年，一年以来涌现出以 GPT-4[10]、Claude[11]、Gemini[12]、Qwen[13] 为代表的众多大模型，通过多任务训练和统一编码 [14] 的方式让模型展现出强大的语义理解能力和任务泛化能力，打破了自然语言任务之间的壁垒，将多种自然语言任务统一成序列到序列的生成任务，从而摆脱了对标注数据的依赖。指令工程指的是为大型语言模型设计和优化输入指令（或“提示”）的过程，旨在更有效率地借助模型的预训练知识库，完成特定的信息抽取任务。通过采用诸如思维链（Chain-of-Thought，CoT）[15]、语境学习（In-Context Learning）[16] 以及专家提示（Expert Prompting）[17] 等策略，研究人员能够引导模型进行深层次的语义分析，从而使之

精确理解特定信息抽取任务的需求，通过规范模型输出格式，使模型能高效地从非结构化文本中提取所需的结构化信息，而无需额外的标注数据集。指令工程不仅能将大型语言模型的泛用能力迁移到特定任务上，还可以提升其在特定信息抽取领域的性能和准确度。

这种对指令工程的深入研究和应用，已经开始在实际的信息抽取任务中展现其成效。Wei 等 [18] 将复杂的信息抽取任务转化为一个分两阶段进行的多轮问答问题，利用 ChatGPT 构建了一个名为 ChatIE 的零样本信息抽取多轮问答框架。首先，通过一次问答识别文本中的实体、关系或事件类型，然后利用链式抽取模板在多轮问答中进一步抽取与这些类型相关的信息，该方法在实体—关系三元组抽取、命名实体识别和事件抽取三个任务上的效果超过了一些全量训练的模型。Wang 等 [19] 提出了 InstructUIE 框架，利用自然语言指令引导大型语言模型完成信息提取任务，InstructUIE 在监督设置下与 BERT 的效果相当，在零样本设置下显著超过了 GPT-3.5 模型。Xiao 等 [20] 提出了一个名为 YAYI-UIE 的聊天增强型指令调整框架，结合对话数据和信息抽取数据进行训练，利用端到端的方法自动调整指令，使其适应不同的信息提取任务，该框架在中文数据集上达到了 SOTA 水平。

1.3 大模型高效微调技术

尽管提示工程通过设计精细框架和输入指令来引导大型语言模型执行特定任务，展现了大语言模型在零样本或少样本设置下处理信息抽取任务的能力。然而在本研究面临的实验材料与设备信息抽取任务中，目标信息通常包含大量专业术语、数值单位以及复杂的语义结构，对模型的领域知识理解和语义抽取能力提出了更高的要求，仅依赖提示工程很难使通用大模型在这一高度专业化的任务上达到理想的性能表现。如果采用特定领域数据对大模型直接进行微调，虽然可以让大模型充分学习到特定领域知识，提高大模型在领域任务上的能力，但是由于需要调整大量的参数，直接微调往往需要花费高昂的计算资源和可能持续数月的训练时间。为此我们需要采用一些高效微调的技术，如 Adapter Tuning[21] 或 LoRA（Low-Rank Adaptation）[22] 技术。Adapter Tuning 在模型的各层之间插入小型的可训练模块（称为 Adapter），只调整这些模块的参数而不改变原始预训练模型的权重，从而能够在保持预训练知识的同时，对特定任务进行有效的微调；LoRA 的核心原理基于对模型权重的低秩适配，这种方法能够在保持大部分预训练知识不变的同时，通过调整一小部分参数来实现对新任务的快速适应。

在信息抽取任务中，Jiao 等 [23] 结合 LoRA 技术提出了一个新颖的面向需求的信息提取框架 ODIE（On-Demand Information Extractor），这一框架允许模型根据用户指令生成结构化的信息表格，显著提升了在自动化评估和人类评估中提取表头和内容的准确性；Dagdelen 等 [24] 提出了一种从科学文本中提取结构化信息的创新方法，这项研究聚焦于材料科学领域，主要解决固态杂质掺杂、金属有机框架（MOFs）以及通用材料信息提取三项任务。通过使用少量标记数据对大语言模型进行高效微调，并以预定义的结构化模式（如 JSON 文档）输出信息，

doi:10.3772/j.issn.2095-915x.2024.05.009

展现了大模型通过微调展现出的将非结构化科学文本转换为结构化数据的潜力。

2 基于实验材料与设备信息的科研设备成本评估框架

在现代科研活动中，面对资源有限的现实，科研团队需要精确评估项目成本，确保资金的有效利用。传统的成本评估方法依赖于人工收集和分析数据，这不仅耗时耗力，而且容易受到主观判断的影响，难以应对科研活动中的复杂性和动态变化。尤其是在需要大量实验材料与设备的研究领域，如计算机科学和物理学，传统方法的局限性更为显著。这些领域的研究往往涉及高端设备和专业材料，其成本估算不仅需要考虑实验不同流程需要的不同材料设备，还需评估材料设备市场价格、设备运行时间等多个维度，这对成本评估的准确性和效率提出了更高要求。

针对当前科研设备成本评估方法研究的不足，本文从科研过程中重要的实验环节切入，设计了一个基于实验材料与设备信息的科研设备成本评估框架，如图 1 所示。该框架由两大核心模块构成：实验材料与设备信息抽取和

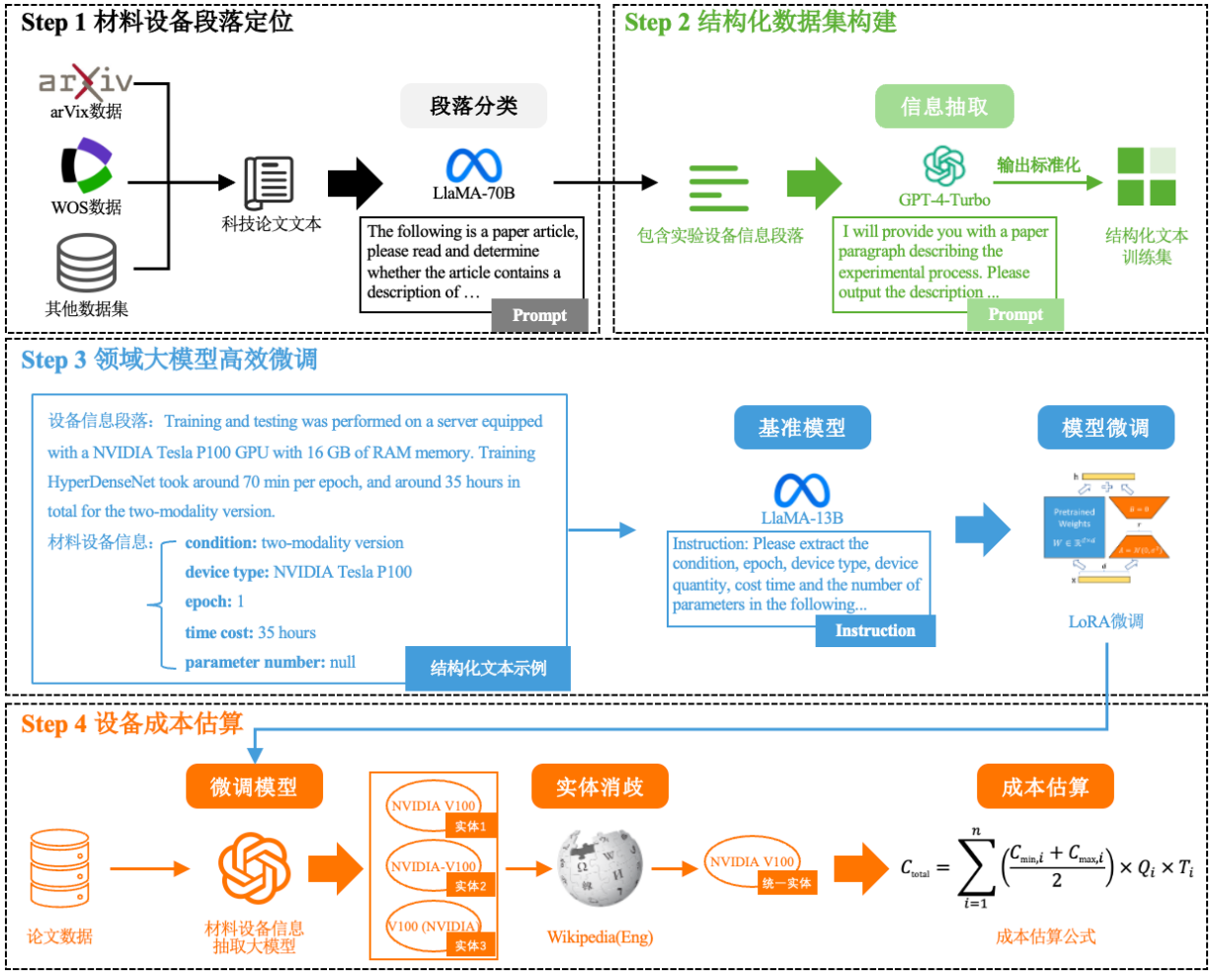


图 1 科研设备成本评估框架

成本评估。首先，实验材料与设备信息抽取模块以大型语言模型为基础，通过流水线策略的设计，实现从科学文献中自动定位并提取关键设备信息的功能。然后，成本评估模块通过对设备功能性与经济属性的分析，实现对多种类型材料与设备的统一价值评估，为科研项目成本提供更全面和精确的评估结果，帮助科研人员和项目管理者更好地规划和控制项目成本。

2.1 实验材料与设备信息抽取模块

实验材料与设备信息抽取模块采用“粗筛选＋精解析”的设计策略，先从海量的文本中定位包含材料与设备信息的潜在段落，然后再从筛选出的文本中进行关键信息的提取，以保证整个信息抽取模块的准确性和效率，为后续的成本评估提供精准的基础信息。

2.1.1 实验材料与设备信息定位

为了尽可能召回包含实验材料与设备信息的文本段落，本研究设计了两种互补的信息定位策略来引导大型语言模型完成该目标。第一种策略采用“粗召回—分类”的两阶段方法实现对成本段落的快速定位。首先，使用正则表达式编写关键词模板，通过匹配段落中所有与材料、设备和成本相关的关键词，包括计算设备如“GPU”“NVIDIA”等，时间成本如“hours”“days”等，从而尽可能多地召回目标句；随后，设计用于文本分类任务的指令，进一步检验通过模板匹配采集到的句子是否满足需求。该指令如图 2 所示，由任务指令（红色）、任务描述（蓝色）以及输出指令（绿色）三个部分组成，保证模型在理解任务需求的同时能够以固定的格式输出结果。

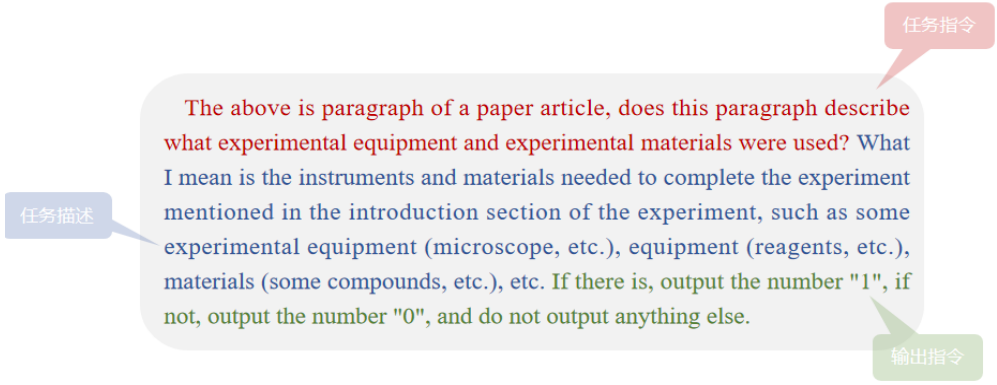


图 2 实验材料与设备信息定位指令模版

第二种策略则面向难以通过既定模板匹配出的句子，如一些具有特殊用途的设备（X-ray）和材料等，在这些未匹配出的文本上进一步进行分类操作。在该策略中，为了尽可能降低误检率，在模型输出结果上作出了更加严格的约束，假设模型在分类句子时生成正例结果“1”的概率为 α，当 α 大于 0.8 时才将检测出的句子视为可信。

2.1.2 实验材料与设备信息识别

在完成目标句的筛选后，本研究采样了少量的句子用于训练数据的构建，完成材料与设备信息抽取模型的设计，实现批量快速的信息抽取，服务于研究最终的成本评价目标。详细的训练数据获取过程以及模型构建过程如下。

doi:10.3772/j.issn.2095-915x.2024.05.009

（1）材料与设备信息识别数据获取

不同学科领域在实验过程中使用的材料和设备会有较大的差异。以本研究关注的计算机领域与物理学领域为例，计算机领域的实验成本主要来源于计算资源的数量、性能及其在训练特定参数规模模型时的轮次和耗时，而对于物理学领域，不同材料都有其固定的成本开销，同时在实验时间上也有一定的耗费，但这一信息较少在文中提及，因此在本研究中没有考虑。不同类型材料和设备需要考虑的成本属性如表 1 所示。

表 1　实验材料和设备的成本属性

	物理材料与设备	计算机设备
成本属性	名字（Name）	设备类型（Device type）
	描述（Description）	设备数量（Device quantity）
	数量（Quantity）	轮次（Epoch）
	—	时间开销（Time cost）

基于对材料和设备重要成本属性的分析，本研究仍采用指令工程技术来引导性能更佳的大型语言模型来对采样的目标句进行结构化解析，如表 2 所示。

表 2　材料与设备信息识别数据获取示例

步骤	角色	文本
指令	SYSTEM	Now you need to help me complete the information extraction task.
	USER	I will provide you with a paragraph describing the training cost of model in some condition. Please out put the condition (fully described),epoch ,device type ,device quantity and cost time of each condition in the list of JSON documents like [{}, {}], If a key does not exist, output value as None. You should pay attention to the model I gave you, don’t output the cost of other model
输入	USER	We use 8 NVIDIA A100 GPUs for model training, and each training is finished within 3 days for large models. During training, we apply SpecAug using 2 frequency-masks with parameter F of 27, and 10 time-masks with parameters p of 0.05.
输出	ASSISTANT	[{ “condition”：“default”， “epoch”:None, “device_type”：“Nvidia A100”， “device_quantity”: 8, “time_cost”：“3 days”， “parameter_number”:None }]

由于大型语言模型性能的局限，解析出的结果会出现一定错误，为了保证训练数据的质量，本研究在模型识别的结果上开展了人工验证。

（2）模型构建

在构建好的训练数据的基础上，本研究采用指令微调的方式来引导通用的大型语言模型在材料和设备信息抽取任务上取得更好的应用效果。具体地，本研究采用了 LORA 微调策略，即在原有的模型上添加小规模的参数来学习任务信息。采用该方法主要原因在于计算资源以及训练数据数量的局限性，以及该方法在类似任务上取得的成功应用。

假设通过大型语言模型的总参数量为 W_{base}，LORA 微调策略的目标是在此基础上添加少量参数 ΔW，而这一新添加的参数由多个升维矩阵 A 和降维矩阵 B 组成，由此：

$$W_{base}+\Delta W= W_{base} + BA \quad (1)$$

若 $W_{base} \in R^{d\times k}$，则 $B \in R^{d\times r}, A \in R^{r\times k}$，其中 r 远小于 d 和 k，因此 ΔW 要远小于 W_{base}，因此使用少量计算资源来微调大规模参数模型成为可能。本研究在 W_{base} 的每一个键—值线性层的旁路都添加了上述可学习的参数，可学习参数的比例占模型全参数比例的 0.0503%。最后，冻结 W_{base}，将抽取实验材料与设备成本属性的指令作为输入，Json 输出结果作为输出对 ΔW 进行微调。

2.2 成本评估模块

利用微调后的大模型对科学论文进行材料设备信息的结构化抽取后，本研究通过链接消歧—价格估算的方式对实验成本进行计算。对于抽取出来的实体先通过 Wikipedia（English）进行查询，如果不同实体指向了相同的 Wikipedia 信息页面，本研究将其视为是同一实体的不同展现形式，并对它们进行合并以消除歧义，然后通过式（2）对合并后的材料设备信息实体进行成本估算，得到最后结果。

2.2.1 实体链接

在科学文献中，对同一材料或设备的描述可能采用多种不同的表达方式。例如，“综合物性测量系统”既可以被完整地表述为“Physical Property Measurement System”，也可以简略地称作“PPMS”，此外还存在其他形式的实体歧义，如大小写差异、是否包含注释等，如表 3 所示。

表 3 设备名称歧义示例

情况	示例
设备名称存在大小写区分	{entity: OSCAR, mention: Oscar}
设备名称简写	{entity: Physical Property Measurement System, mention: PPMS} {entity: Chemical Vapor Deposition System, mention: Chemical Vapor Deposition (CVD) system }
设备名称包含符号	{entity: x-ray diffractometer, mention: x ray diffractometer }, {entity: light scattering measurement equipment, mention: measurement equipment (light scattering)}
设备名称包含注释	{entity: Microprobe, mention: Microprobe (2015)}

为了精确和高效地估算成本，本研究采纳了基于 Wikipedia 的实体链接方法来解决实体的歧义问题。作为一个庞大的在线知识库，Wikipedia 涵盖了从科技术语到实验材料等广泛的主题，几乎所有的材料和设备实体都能在其数据库中找到相应的条目。此外，Wikipedia 的数据高度结构化，包括信息框（infoboxes）、分类（categories）、重定向页面（redirect pages）以及消歧义页面（disambiguation pages）等元素，这些结构化的信息极大地提升了实体识别和消歧的精准度。通过利用 Wikipedia-based Entity Linking[25] 方法，本研究能够有效地识别和统一这些多样化的表述，从而在成本估算中实现更高的准确性和效率。

2.2.2 成本估算

精确计算科学实验中材料与设备的成本是一个高度复杂的任务。即使是同一名称的设备或材料，其价格也可能因品牌、型号等多种因素而产生显著波动，这种价格波动给预算制定和成本控制带来了不确定性。为了应对这一挑战，本研究借鉴了计算机科学领域中平均情况分析（Average-Case Analysis）的概念。平均情

doi:10.3772/j.issn.2095-915x.2024.05.009

况分析是一种算法性能评估方法，不同于最坏情况分析（Worst-Case Analysis）和最佳情况分析（Best-Case Analysis），它关注的是算法在所有可能的输入上的平均表现。这种分析方法提供了一种更加贴近现实的性能预测，因为该方法考虑了算法在正常各种输入情况下的性能，而不仅仅是在极端输入情况下的表现。

将平均情况分析的核心理念应用到科学实验材料与设备的成本估算中，意味着在计算成本时将考虑到价格波动的整体分布，而不仅仅是最高价或最低价，从而提供一个更加准确的成本估算。为此，本研究提出了如下成本估算公式。

$$C_{\text{total}} = \sum_{i=1}^{n} \left(\frac{C_{\min,i} + C_{\max,i}}{2} \right) \times Q_i \times T_i \quad （2）$$

假设一个实验需要 n 项材料或者设备完成，其中 Q_i 为在实验中第 i 项材料或设备的数量；T_i 为使用第 i 项材料或设备的实验时长（如果使用），如果某项材料或者设备的使用成本不涉及时间长度，则 T_i 可从公式中省略；$C_{\min,i}$ 和 $C_{\max,i}$ 分别代表第 i 项材料或设备的最小和最大成本。

此公式旨在综合考虑同一材料或设备在不同情况下的价格波动，通过计算其平均成本来提供一个既实用又具有代表性的成本估算。这种方法不仅允许本研究在面对价格信息不完全或变化大的情况下进行有效的成本规划，而且还提高了成本估算的准确度和可靠性。

3 实验结果和分析

3.1 实验设置

（1）实验数据

本研究所依托的原始数据集经由 Paper With Code 和 arXiv 数据库获得，覆盖计算机和物理学两大领域，涉及自 2015—2023 年发表的学术论文总计 11319 篇，其中计算机领域论文 3192 篇，物理学领域 8127 篇。基于此原始数据集，本研究进一步采用基于大型语言模型的预处理技术，筛选出含有实验材料与设备信息的段落共 5228 条。此外，通过精心设计的指令模板，本研究构建了包含 5228 条数据的数据集，同时该数据集进一步细分为训练集 4928 条、验证集 100 条以及测试集 200 条。

（2）模型选择

在构建科研设备成本评估框架的过程中，针对实验材料与设备信息的准确定位、识别及训练数据的收集，本研究采纳了三种差异化的计算模型以实施相应任务。首先，为了筛选含有材料及设备信息的描述性文段，本研究使用了具有良好语言理解能力的 70B 参数 LLaMA2 模型。其次，通过信息抽取提示模板的设计，利用性能更卓越的 GPT-4-turbo 模型对文段进行精细化解析，旨在构建一个高质量的训练数据集。最后，本研究采用了一个规模较小（13B 参数）的 LLaMA2 模型进行模型微调优化，实现将大规模参数模型的信息抽取能力迁移到较小规模模型上，从而提高实验材料与设备信息的识别效率。

（3）参数设置

训练模型的参数设置详见表 4，训练批次大小为 8，训练轮次为 10，学习率为 1e-4。

（4）实验设备

为完成实验材料与设备信息抽取以及成本评估等关键研究内容，本研究采用了 2 块 A100 GPU 来执行大模型的提示工程与指令微调工作，具体的实验设备与环境如表 5 所示。

表 4　大型语言模型微调参数设置

参数名称	参数含义	参数设置
batch_size_training	训练批次大小	8
num_epochs	训练轮次	10
lr	学习率	1e-4
gamma	伽马值	0.85
val_batch_size	验证批次大小	1
peft_method	调优框架	lora

表 5　实验设备与环境

实验环境	环境配置
操作系统	Ubuntu 22.04 LTS
GPU	2 × NVIDIA A100 SXM4 80GB
内存	1.5 T
编译器	Python 3.10.11
深度学习框架	Pytorch 2.0.0

3.2　评价指标

本研究为了验证基于实验材料与设备信息的科研设备成本评估框架的可靠性，将首先着重对实验材料与设备信息抽取的性能进行评价，使用信息抽取研究通用的准确率（precision）、召回率（recall）和 F1 值进行评测。

3.3　实验结果与分析

3.3.1　材料设备信息抽取性能评估

本研究选择了 LLaMA2-13b-base、GPT-3.5-turbo-0125 和 LLaMA2-13b-lora（本实验调优后的模型）三个大语言模型在测试集上进行实验，性能评估覆盖了计算机科学和物理学两个领域，重点关注从各自领域的科技论文中抽取材料和设备细节的能力，实验结果如表 6 所示。

实验结果表明，在进行计算机科学和物理学领域内的材料设备信息抽取任务时，LLaMA2-13b 的基础模型表现较为一般，这在某种程度上反映了该基础模型在理解这两个领域知识的深度与广度方面存在的局限。然而，通过使用 LoRA 技术进行高效调优，LLaMA2-13b-lora 模型在这两个领域内展现出显著的性能提升，部分性能指标甚至超越了 OpenAI 推出的 GPT-3.5-turbo-0125 模型。这一成果突显了 LoRA 技术在特定应用领域调优大型语言模型的潜力，尤其是在精确抽取材料设备方面的应用，并为后续的成本评估工作奠定了坚实的基础。

表 6　实验结果

研究领域	计算机				物理			
评价指标	P	R	Macro F1	Micro F1	P	R	Macro F1	Micro F1
LLaMA2-13b-base	0.3871	0.4953	0.4038	0.4345	0.5335	0.5929	0.5437	0.5725
GPT-3.5-turbo-0125	0.5042	0.6453	0.5017	0.5661	0.6568	0.7035	0.6487	0.6793
LLaMA2-13b-lora	0.6221	0.5966	0.4939	0.6091	0.7284	0.6679	0.6890	0.6969

3.3.2　科研设备平均成本变化分析

本研究选择了 Paper With Code 网站计算机视觉领域下的高 star（收藏数大于 10000）科技论文 292 篇作为典型样例进行小样本成本估算。这一策略确保了样本论文在学术与工业界的广泛认可度和应用价值，反映出其影响力和代表性。

如图 3 所示，本研究按照时间顺序进行了

doi:10.3772/j.issn.2095-915x.2024.05.009

论文数量和平均论文成本的初步统计。数据表明，无论是典型论文数量还是平均论文成本，都随着年份的增加而呈现上升趋势，特别是在2020年，可以明显看到这一趋势的加速。本研究注意到，这与CV界的Transformer架构ViT[26]的提出时间相吻合。自ViT的提出以来，计算机视觉领域的研究开始更多地采用Transformer架构，而这一架构模型训练所需的数据量也通常比之前的模型多，这就导致科技论文实验需要更高性能的计算资源，成本开销大幅度提升，这与本研究的成本估算的实验结果也保持一致。

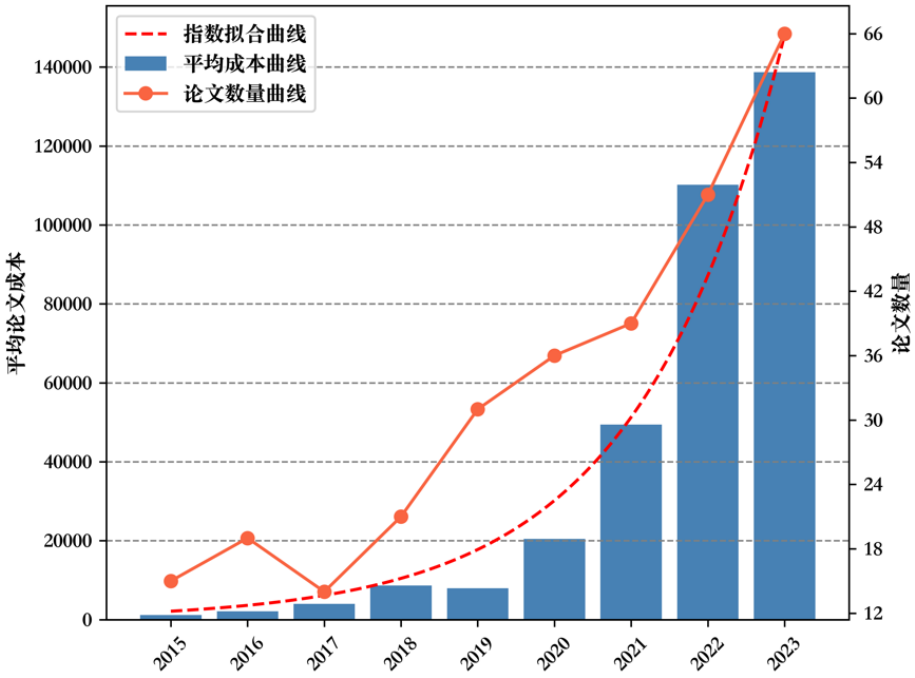


图3 样例成本估算结果

OpenAI于2022年末发布的ChatGPT语言模型，更是引发了学界对大模型研究的高度关注，这进一步推动了2023年计算机视觉领域平均论文成本的增长。综合以上分析可以看出，在当前大模型主导的研究范式下，论文的平均研究成本正呈现出指数级的增长态势，这表明计算资源已经成为制约计算机视觉领域科研产出的关键因素之一。

3.3.3 科研设备投入与模型性能关联分析

为更细致地刻画科研设备对关键任务性能的影响，本小节聚焦目标检测这一计算机视觉的核心任务，对历年SOTA模型的成本效益进行了实证分析。本小节选取了Paper With Code网站中目标检测任务在COCO test-dev数据集上的SOTA模型对应的论文作为研究对象，对其进行了成本估算分析。图4展示了不同时期的SOTA模型在COCO数据集上的性能表现（BOX mAP分数）与其对应论文的研究成本之间的关系。散点图中每个散点代表一个SOTA模型，其中散点的颜色代表模型所采用的基本架构。

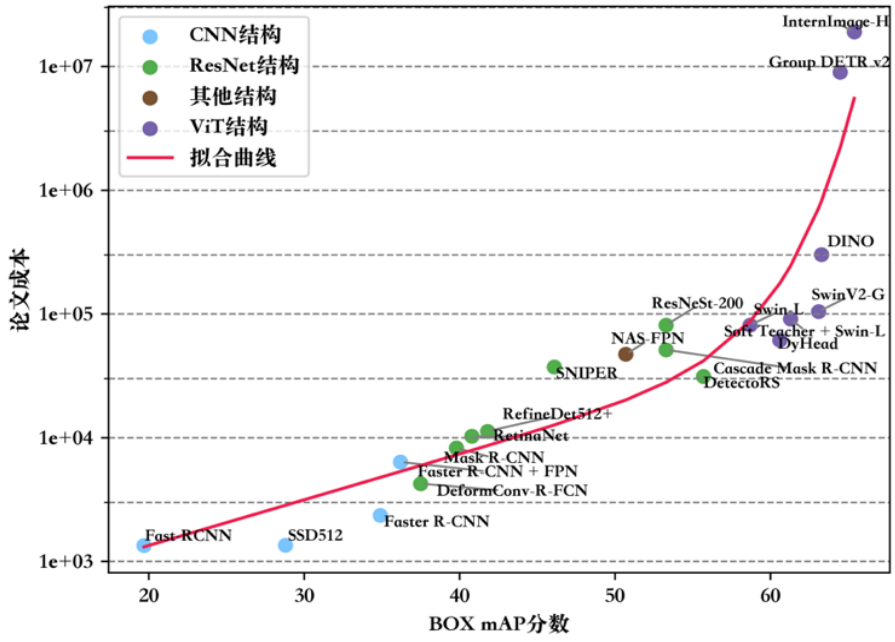


图 4　模型效果与设备关系散点图

如图 4 所示，随着科研设备成本的增加，目标检测模型的性能整体上呈现出上升的趋势。这表明在目标检测任务上，投入更多的计算资源和设备成本，通常能够带来性能的提升，这一点与直觉相符，也与前文分析论文设备投入整体成本变化趋势的结论一致；同时，本研究也注意到不同模型架构的性能表现和成本效益存在明显差异。以传统的 CNN 结构的模型为例，其性能随着成本的增加而提升，但当达到一定程度后，继续增加设备成本带来的性能提升变得非常有限；相比之下，以 ResNet 为代表的更先进的模型架构，在相同设备成本下能够取得更优的性能表现，这得益于残差网络更强大的特征提取和表示能力，使其能够更好地捕捉和建模图像中的关键信息。ResNet 结构的出现，在一定程度上突破了传统 CNN 模型的性能瓶颈，通过构建更深、更复杂的网络，可以进一步推动目标检测任务的性能提升。然而，ResNet 结构的性能提升也并非没有上限，当模型复杂度和计算成本达到一定程度后，其性能增益同样会趋于饱和。

近些年来，ViT 架构的模型为目标检测任务的发展注入了新的活力，其强大的特征学习能力使其在目标检测任务上取得了突破性的进展。从图中可以看出，ViT 模型的出现进一步推高了目标检测任务的性能上限，但与此同时，其对计算资源的需求也大幅增加。为了充分发挥 ViT 模型的性能优势，需要在海量数据上进行大规模的模型训练，这无疑会带来更高的设备投入成本。同时本研究注意到，早期的深度学习模型能力较为单一，往往只作用于单一数

doi:10.3772/j.issn.2095-915x.2024.05.009

据集的单一任务上；而现在的大模型一方面具备处理在多个数据集上进行多种任务的能力，另一方面大模型的评测往往需要大量的对比消融实验，这也进一步加剧了科研设备成本呈指数型增长的现象。

综上所述，对于特定的模型架构或研究范式，客观上存在一个性能上限，当设备投入成本和模型复杂度超过某个临界点后，继续增加设备投入所带来的性能提升将变得非常有限，此时设备投入的边际效益将显著降低。这对于科研人员和项目管理者优化资源配置、控制成本风险具有重要的指导意义，通过及时发现和预警低效益或高风险的研究方向，可以更加科学地规划和调整项目预算，提高科研投入的效益。

4 讨论与结论

本研究旨在实现对物理和计算机领域学术论文中的材料设备信息进行低成本、高效率及高质量抽取，为此，本研究针对不同的学科领域设计了特定的指令集，通过这些指令集引导大型语言模型深入挖掘不同领域学术论文对于材料设备描述的特征，同时为不同领域论文中材料设备信息的抽取设计了专门的抽取范式。本研究构建了材料设备信息抽取任务的训练数据集，利用低秩微调技术对大模型进行精细调优，这种方法显著提高了大模型在特定领域内对材料和设备信息进行结构化抽取的能力，为后续的材料设备成本计算工作提供了坚实的基础，进而为科研人员和项目管理者在资源规划和成本估算方面提供有力的支持。

本研究通过对计算机视觉领域的实证分析，验证了基于大型语言模型的成本评估方法的有效性。研究结果表明，本文提出的框架能够准确估算不同时期论文的平均研究成本，揭示出计算资源对科研产出的重要影响。同时，通过对目标检测任务的成本效益分析，本研究还发现，不同模型架构的性能提升和设备投入之间存在一个客观的临界点，超过这一临界点后，继续增加投入的边际效益将显著降低。这些发现与实际科研活动的规律相吻合，证明了本文提出的成本评估方法能够为科研项目的资源配置和风险管理提供有价值的决策参考。尽管当前的分析还主要集中在计算机视觉领域，但本研究提出的方法论具有一定的普适性，未来可以进一步拓展到更多学科和任务场景中，为科研管理提供更加全面和精准的成本评估工具，全面提升科研项目管理的质量和效率。

参 考 文 献

[1] 罗准辰，赵赫，叶宇铭，等．基于文献链接信息分析的科技资源风险评估 [J]. 中文信息学报，2020, 34(5): 64-73.

[2] KAIL E, KACSUK P, KOZLOVSZKY M. New aspect of investigating fault sensitivity of scientific workflows[C]//2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES). Bratislava, Slovakia: IEEE, 2015: 185-188.

[3] UDDIN S, ONG S, LU H. Machine learning in project analytics: a data-driven framework and case

study[J]. Scientific Reports, 2022, 12(1): 15252.

[4] LOCK D. Cost estimating[C]//In The Essentials of Project Management. Routledge, London, 2001: 31.

[5] SAJADFAR N, MA Y. A hybrid cost estimation framework based on feature-oriented data mining approach[J]. Advanced Engineering Informatics, 2015, 29(3): 633-647.

[6] LIU S, QI G, ZHEN L, et al. Research on the Prediction Model of Material Cost Based on Data Mining[J]. The Open Mechanical Engineering Journal, 2015, 9(1): 1062-1066.

[7] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2019.

[9] BROWN T, MANN B, RYDER N, et al. Language Models are Few-Shot Learners[C]//Advances in Neural Information Processing Systems: Vol. 33. Curran Associates, Inc., 2020: 1877-1901.

[10] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

[11] WU S, KOO M, BLUM L, et al. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology[J]. arXiv preprint arXiv:2308.04709, 2023.

[12] ANIL R, BORGEAUD S, WU Y, et al. Gemini: A family of highly capable multimodal models[J]. arXiv preprint arXiv:2312.11805, 2023.

[13] BAI J, BAI S, CHU Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.

[14] MISHRA S, KHASHABI D, BARAL C, et al. Cross-Task Generalization via Natural Language Crowdsourcing Instructions[C]//Muresan S, Nakov P, Villavicencio A. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 3470-3487.

[15] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[16] DONG Q, LI L, DAI D, et al. A survey on in-context learning[J]. arXiv preprint arXiv:2301.00234, 2022.

[17] XU B, YANG A, LIN J, et al. Expertprompting: Instructing large language models to be distinguished experts[J]. arXiv preprint arXiv:2305.14688, 2023.

[18] WEI X, CUI X, CHENG N, et al. Zero-shot information extraction via chatting with chatgpt[J]. arXiv preprint arXiv:2302.10205, 2023.

[19] WANG X, ZHOU W, ZU C, et al. Instructuie: Multi-task instruction tuning for unified information extraction[J]. arXiv preprint arXiv:2304.08085, 2023.

[20] XIAO X, WANG Y, XU N, et al. YAYI-UIE: A Chat-Enhanced Instruction Tuning Framework for Universal Information Extraction[J]. arXiv preprint arXiv:2312.15548, 2023.

[21] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]// International conference on machine learning. PMLR, 2019: 2790-2799.

[22] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-

doi:10.3772/j.issn.2095-915x.2024.05.009

rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

[23] JIAO Y, ZHONG M, LI S, et al. Instruct and Extract: Instruction Tuning for On-Demand Information Extraction[C]//2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023. Association for Computational Linguistics (ACL), 2023: 10030-10051.

[24] DAGDELEN J, DUNN A, LEE S, et al. Structured information extraction from scientific text with large language models[J]. Nature Communications, 2024, 15(1): 1418.

[25] HACHEY B, RADFORD W, NOTHMAN J, et al. Evaluating Entity Linking with Wikipedia[J]. Artificial Intelligence, 2013, 194: 130-150.

[26] DOSOVITSKIY A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

doi:10.3772/j.issn.2095-915x.2024.05.010

基于机器学习的科技人才综合学术水平评价指标与模型研究

开放科学（资源服务）标识码（OSID）

李勃慧　王运红　杨代庆　郑楚华　陈国娇

中国科学技术信息研究所 北京 100038

摘要：[目的 / 意义] 从学术诚信、科研产出、科研活动和学术声誉 4 个维度，打破单一指标局限性，为人才评价中“破四唯”“立新标”提供新的思路和参考。[方法 / 过程] 针对科技人才的特征，构建了 98 项定量指标与定性特征指标。选取研究样本，进行数据集构建和数据预处理，形成最终识别预测模型的输入数据；通过机器学习获得样本特征指标与评价结果之间的隐含关系，比较 9 种模型算法下的人才学术能力表现，综合得出最优模型进行后续训练和调优，通过对比模型输出与实际结果进行模型评价。[局限] 模型仅基于实验样本数据量和特征，推广应用还需要更多的样本数据进行训练调优。[结果 / 结论] 实证研究表明，模型对高水平科技人才评价具备较好的适用性，为科技人才评价提供新的视角和方法。

关键词：科技人才；人才评价；人才识别；评价模型

中图分类号：G350

Research on Comprehensive Academic Level Evaluation Indicators and Models for Scientific and Technological Talents Based on Machine Learning

LI Bohui　WANG Yunhong　YANG Daiqing　ZHENG Chuhua　CHEN Guojiao

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

Abstract: [Objective/Significance] This study breaks the limitations of a single indicator from four dimensions: academic integrity, research output, research activities, and academic reputation, providing new ideas and references for breaking the “4 Single dimension” and “setting new standards” in talent evaluation. [Methods/Processes] For the characteristics of scientific and technological talents, 98 quantitative indicators and qualitative characteristic indicators were constructed. This study selected research samples for dataset construction and data preprocessing to form the input data for the final recognition and prediction

基金项目　国家社会科学基金项目“大数据环境下同行评议方法模式研究”（19BTQ082）。

作者简介　李勃慧（1994-），女，硕士，助理研究员，研究方向为科技评价、数据挖掘，E-mail：libh@istic.ac.cn；王运红（1971-），女，硕士，研究员，研究方向为科学计量、科技评价；杨代庆（1975-），男，硕士，研究员，研究方向为科学计量、科技评价。郑楚华（1991-），女，博士，助理研究员，研究方向为科技人才评价；陈国娇（1992-），女，博士，助理研究员，研究方向为科技人才评价。

引用格式　李勃慧，王运红，杨代庆，等 . 基于机器学习的科技人才综合学术水平评价指标与模型研究 [J]. 情报工程，2024, 10(5): 115-127.

doi:10.3772/j.issn.2095-915x.2024.05.010

model; Obtain the implicit relationship between sample feature indicators and evaluation results through machine learning, compare the academic performance of talents under 9 model algorithms, comprehensively determine the optimal model for subsequent training and optimization, and evaluate the model by comparing the model output with actual results. [Limitations] This research model is only based on experimental sample data size and features, and further training and optimization of sample data are needed for its widespread application. [Results/Conclusions] Empirical research has shown that the model has good applicability for the evaluation of scientific and technological talents, providing new perspectives and methods for the evaluation of scientific and technological talents.

Keywords: Scientific and Technological Talents; Talent Evaluation; Talent Identification; Evaluation Models

引言

科学、合理的科技人才评价体系是人才工作的“指挥棒”，不仅影响着科技人才的质量和水平，同时影响科技人才引进、管理、培养和使用等。近年来，中共中央办公厅、国务院办公厅、科技部、教育部等部门针对科技人才评价连续发文；2022 年，科技部会同教育部等 7 部门联合召开科技人才评价改革试点启动会，部署推进《关于开展科技人才评价改革试点的工作方案》，推动人才评价改革落地见效，着力克服“唯论文、唯职称、唯学历、唯奖项”倾向。由此可见，对科技人才的评价问题不仅仅是学术界研究和探讨的热点，更是影响国家人才培养、使用的重要命题。

科技人才评价改革提出“破四唯”，但是以“立新标”为突破口的研究还在探索中。在“破四唯”的要求下，科技人才能力的评价必须打破原来的标准，对科技人才的科研综合能力能够进行全方位、客观、合理和可操作的评价。当前科技人才评价工作还存在一些问题，如缺乏科学合理、各有侧重的人才评价标准，评价标准动态更新调整机制不够及时有效；在采用成果影响力评价科研能力时，将 SCI 和国内核心期刊论文发表数量、引用榜单和影响因子排名等作为评价参考的重点，忽视了标志性成果的质量、项目能力、学术活跃度、学术声誉贡献等综合指标；评价指标的主观性较强，不利于操作性和公平性。

本研究针对科技人才的个体特征和学术特性，构建了 4 个维度，包含 98 项定量指标与定性指标。同时，选取了 531 名工程院院士候选人作为研究对象，进行数据集构建和数据预处理，形成最终识别预测模型的输入数据；通过机器学习获得样本特征指标与评价结果之间的隐含关系，比较 9 种模型算法下的人才学术能力表现，综合得出最优模型进行后续训练和调优。实证结果表明，该方法可应用于科技人才的评价和识别工作实践中，为多维度评价指标下，应用机器学习模型进行科技人才的评价识别提供参考。

1 国内外研究现状

1.1 科技人才评价方法

学术界呼吁当前亟须创新人才评价机制，建立健全以创新能力、质量、贡献为导向的科

技人才评价识别体系，形成并实施有利于科技人才潜心研究和创新的评价制度。科技人才评价则是对个体能力、表现和潜力进行系统性分析和评估，在科技人才评价方法研究上，国内学者开展的主要研究如下。

一是同行评议、专家访谈、调查问卷等定性研究方法。这些定性方法在揭示科技人才评价的隐性因素和复杂关系等方面具有明显优势，但存在过度依赖主观分析和判断影响、获取和处理专家意见带来的人力成本、不同环境或条件下的适用性和可迁移性差等问题。

二是基于科学计量学的定量研究方法。通常采用“构建评价指标体系—设置各项参评指标权重或赋分规则—计算得分”路径，研究重点和创新点多集中在新型评价指标的选取和指标权重确定方法两方面。在评价指标选取方面，刘璇等[1]通过合著网络的中心度指标发现能够带领整个领域发展方向的学术带头人，宋培彦等[2]提出颠覆性指数，基于文献引用的聚集效应评估了文章的创新程度，常缪缪等[3]使用调和 H 指数综合合著者数量和合著者顺序影响，实现了合著贡献程度的量化。在指标权重设置方面，常见的权重确定方法包括利用数字的相对大小信息进行权重计算，通常与专家打分或问卷调研等方法结合使用，常见方法包括层次分析法和优序图法等；根据各项指标值的变异程度进行权重计算，常见方法包括熵权法[4]、CRITIC 等；基于变量之间的相关性进行权重计算，常见方法包括主成分分析法、因子分析法[5]等。此类科学计量学方法通过构建指标体系和计算得分，提供了一种相对客观和标准化的评价过程，减少了主观判断的干扰；强调评价指标体系的系统性构建，能够从多个维度全面反映被评价对象的综合素质和能力。然而，依赖线性模型的方法忽视了非线性的复杂关系，丢失信息的丰富性，可能导致评价结果的简化和失真。

三是基于数据挖掘及机器学习模型构建的研究方法。该方法能够自动从大规模数据集中识别复杂非线性关系，降低了人工判断的主观性以及传统计量分析的约束性。例如基于神经网络[6]的动态社交分析高潜力人才早期识别，基于支持向量机[7]的高、低层次人才二分类研究，基于决策树 C4.5 分类器[8]的人才绩效模式预测，基于动态神经网络[9]的人才流动和工作绩效的建模，基于递归神经网络[10]的人才招聘场景下求职者经验与匹配度评价，基于决策树[11]的人才类型与人才结构划分等。尽管上述基于机器学习的人才评价方法在多个数据集和应用场景中较为有效，但此类方法在指标维度设计和数据集的构建方面具有高度依赖性，在科技人才评价场景下的适应性及效能尚未得到验证。

1.2 科技人才综合学术水平评价维度

单一的评价维度已不能反映人才的综合学术能力和水平，不再是学者研究的重点。随着人才评价体系研究的不断完善和深化，对科技人才综合学术水平的评价呈现出以下三方面的特征。

一是评价维度更加全面多元，突破了传统维度中对于评价学术历程和经验的偏向性，考虑创新能力、团队领导力、产学研结合能力等。例如，田军等[12]针对陕西省科技人才评价的实

doi:10.3772/j.issn.2095-915x.2024.05.010

际需要，提出创新知识、道德素质、创新动机、影响力、创新能力、产出绩效 6 维度评价模型；丁宁等 [13] 基于医学人才的学科特点，提出临床、科研、教学相结合的指标体系；张熠等 [14] 从创新驱动对人才的需求出发，建立了以基本素质为基础、以创新能力为核心、以创新成果为导向的评价体系。余波等 [15] 在传统高校人才评价指标基础上提出了社会影响力和可持续发展、国际化和国际交流等指标。王运红等 [16] 在对科技人才科研综合能力分析的基础上，构建科技人才科研综合能力评价模型，设计基于科技信息大数据的评价指标体系，包括基本素养、科研产出影响力、科研管理能力、学术潜力和学术地位 5 个维度，并研究其在人才引进工作中的应用，其评价指标维度的设计重点关注科研成果的数量，如论文和专利数量；关注成果的质量，如高影响期刊发表、高被引论文；科研成果的实际应用和转化，如技术许可和商业化成功案例。

二是评价维度更注重标准化、规范化与连续性。科睿唯安引文桂冠奖 [17] 通过统计过去 30 年里 4 个学科领域内各位高被引作者的总引用量、单篇引用量、高被引文章总数，以及领域内论文的篇均被引次数，或每位作者的篇均被引次数等多个指标，以衡量这些作者是否是该领域的开创者、是否曾经因该项工作多次获奖等，最终对引文桂冠奖获得者做出判定。

三是新兴评价维度基于国家需求和政策导向，更加重视创新能力、贡献、科学家精神等层面。例如，井润田 [18] 采用多案例研究方法归纳总结出科技人才能力特征的三个较为特殊的维度，包括学术热情和学科敏感性、战略规划及其实现路径、团队氛围构建。冯粲等 [19] 基于社会背景、教育背景、工作经历、科技贡献、科学精神等五个维度归纳科技人才的规律性特征，其中科技贡献引入了对科研影响力的更广泛评估，如社会影响、政策制定、科学传播等，科学精神则集中体现科学的理想、信念、态度、思维和视野。

国内外学者构建了全面多元、标准化规范化、与国家需求和政策导向紧密结合的科技人才综合学术水平评价维度，为人才识别和评价研究提供了更为全面和深入的认知。本研究基于前述工作，补充战略性科技人才特征指标，分析不同学科特性，评估指标的表征性和可获取性，结合定量与定性指标，提出科技人才综合学术水平评价维度。

综上所述，同行评议方法具有依赖主观判断的问题，基于科学计量学的人才评价方法过于依赖科学成果产出维度的指标。同时，在科研数据呈爆炸式增长的情况下，简单的赋权重的定量评价方法在辨识复杂模式、泛化能力以及处理高维数据方面具有很大局限性。因此，科技人才的遴选评价体系需要融合更多的数据挖掘和机器学习方法，以提升评价的准确性和适应性。

2 综合学术水平评价指标与模型构建

2.1 评价指标与模型构建流程

本研究运用机器学习方法进行科技人才评价的模型构建，具体流程如图 1 所示。

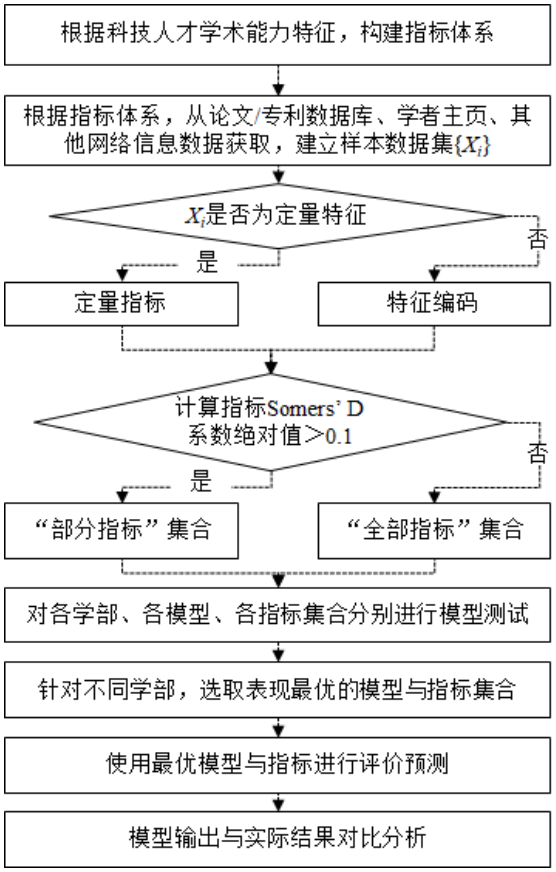


图 1　科技人才评价模型构建流程

第一步，评价指标体系构建。科技人才评价与识别指标体系设计的原则一方面要着重考虑指标的有效性，另一方面也要兼顾指标的可获得性，确保指标体系的建立简单易行。在指标的有效性方面，根据科技人才学术能力的特征研究构建指标体系；在指标的可获得性方面，考虑从论文 / 专利数据库、学者主页、其他公开网络渠道获取数据。

第二步，数据集构建和数据预处理。首先根据指标体系从多来源公开渠道收集并清洗、整合数据。在数据预处理阶段，根据数据类型区别，针对分类变量，采用独热编码方式，将每个分类变量转换为独立二进制特征；针对数值变量，使用 Somers'D 系数衡量各项自变量 X_i 与因变量 Y 之间的关联程度，筛选出系数值介于 -1 和 1 之间的变量以构建“部分指标”集合，并与“完全指标”集合共同构成形成最终科技人才识别预测模型的输入数据。

第三步，构建机器学习模型并进行模型评价。考虑不同类型的人才指标集的特征差异，对各学部、各机器学习模型、各指标集合分别进行模型测试，选取表现最优的模型与指标集合进行后续训练和调优。预测完成后，通过对比模型输出与实际结果进行模型评价。

2.2　综合学术水平评价维度

推动人才评价改革落地的重要目标就是着力克服“唯论文、唯职称、唯学历、唯奖项”倾向，本研究基于科技人才具有较高学术能力的特征，对其综合能力设计多维度的多元评价指标体系，破除“四唯”带来的片面性。

除科研人员具备的综合学术能力外，科技人才还要具备科技创新能力强、科研任务组织领导能力强等特征，其职责和目标与普通科研人员相比有更高层面的要求，需要根据科技人才的特征来设计其评价维度。本研究选择学术诚信、科研产出、科研活动、学术声誉 4 个维度来衡量，如表 1 所示。

表 1　科技人才综合学术水平评价维度

评价维度	说明
学术诚信	道德诚信，含成果真实性和履历真实性
科研产出	包括学者各类科研产出数量、质量、影响力等
科研活动	包括学者参与各类科研项目，参加学术会议情况
学术声誉	学术界对于学者的认可，包含科研奖励、学术荣誉等

doi:10.3772/j.issn.2095-915x.2024.05.010

学术诚信主要从人才的履历真实性和产出成果真实性两个方面考察。履历真实性考察教育背景和工作经历，从学历、工作任职等个人基本信息判断是否有造假和瞒报；产出成果真实性从学术道德上进行判断，如论文存在抄袭、造假而被撤稿，专利有侵权而被追诉，专著或者标准等任何一项产出成果存在的抄袭、造假等问题。

科研产出直观反映人才的研究成果影响力，体现了人才在特定时间内的学术贡献。该维度可以具体量化人才的研究能力和创新程度，同时反映了其研究的深度和广度。常见指标包括在高质量期刊上发表的论文数量与质量、独立研究成果、学术专著的出版、专利的申请和授权情况等。

科研活动展示了人才参与学术交流和合作研究的情况，反映了其在学术界的活跃度和影响力，它不仅体现了人才的学术实践能力，还显示了其领导与合作的能力。常见指标包括参与国家级或国际合作的科研项目、与其他领域专家和知名学者的学术合作，以及组织学术会议和进行学术演讲等。

学术声誉是学术界和社会上对人才学术成就及其贡献高度认可的体现，体现了学术界对个体学者整体贡献的认可和尊重。科技人才获得的科研奖励、学术荣誉等反映了学者在学术界的声望。

2.3 综合学术水平评价指标体系

基于学术诚信、科研产出、科研活动、学术声誉 4 个维度来衡量的科技人才评价指标体系，将定性维度细化为可测度的指标，具体如表2所示。

表 2 综合能力定量指标列表

维度	指标名称	说明
学术诚信	成果真实性、履历真实性	学术诚信（2 项），若任何一项发现有问题，可“一票否决”
科研产出	论文总数、核心期刊论文数、第一作者论文数、总被引频次、篇均被引频次、单篇最高被引频次、被引论文量、被引率、h 指数	中文论文（9 项）
	论文总数、通讯作者论文数、第一作者论文数、高被引论文数、总被引频次、篇均被引频次、单篇最高被引频次、被引论文量、被引率、h 指数	英文论文（10 项）
	论文总数、第一作者论文数、通讯作者论文数	SCI 论文（15 项），分别统计 Q1, Q2, Q3, Q4 分区情况及 SCI 总体情况
	论文总数、第一作者论文数、总被引频次、单篇最高被引频次、被引论文量、h 指数	CPCI 论文（6 项）
	论文总数、第一作者论文数、通讯作者论文数	EI 论文（3 项）
	专利总量、第一发明人专利数、发明专利数量、实用新型专利数量、外观设计专利数量、专利有效量、质押数、许可数、转让数、实质审查数、公开数、授权数、权利终止数、撤回数、驳回数、放弃数、专利合享价值度、被引证次数、h 指数	专利（19 项）
科研活动	863 计划、973 计划、自科基金、国家重点研发计划、国际合作项目、科技成果转化计划、科技基础性专项工作、科研院所研究开发专项、星火计划、火炬计划、支撑计划、重大仪器专项、重大专项、富民强县、公益性行业科研专项计划、创新方法专项、项目总计	项目（17 项）仅统计项目负责人
	国家自然科学奖一等奖、国家自然科学奖二等奖、国家技术发明奖一等奖、国家技术发明奖二等奖、国家科学技术进步奖创新团队、国家科学技术进步奖特等奖、国家科学技术进步奖一等奖、国家科学技术进步奖二等奖、奖项总计	奖励（9 项）统计全部获奖人
学术声誉	杰出青年、长江学者	荣誉称号（2 项）

本研究还补充了定性指标 8 项，补充定性指标主要从人才个人特征和背景信息，包括性别、年龄、所在机构等作为评价的背景变量指标，定性指标经特征编码后可以形成虚拟变量。因本研究对象为工程院院士候选人，提名渠道也反映肯定其科研能力和贡献度的重要指标，如表 3 所示。

表 3　补充定性指标

维度	指标名称	说明
基本信息	性别	男性 =1，女性 =0
	年龄	
	机构类型是否为军工单位	是 =1，否 =0
	机构类型是否为高校	是 =1，否 =0
	机构类型是否为科研院所	是 =1，否 =0
	机构类型是否为企业	是 =1，否 =0
	提名渠道是否为院士提名	是 =1，否 =0
	提名渠道是否为科协提名	是 =1，否 =0

以上两个表中，反映学术诚信、科研产出、科研活动和学术声誉的可量化评价指标共计 92 项，学术诚信的 2 个指标中，若任何一项发现有造假或者瞒报等问题，可“一票否决”，不参与学术综合水平的定量计算，参与模型计算的定量指标与描述性指标共计 98 项。

3　指标分析与模型结果验证

3.1　研究样本与数据来源

本研究以中国工程院 9 个学部，2019 年 531 名院士候选人为研究样本。对 531 名院士候选人的综合学术水平进行评价，以是否入选作为命中率判定模型效果。

中文论文数据来源于北京万方数据股份有限公司的万方数据；英文论文数据来源于科睿唯安 Web of Science 核心合集（含 SCI 及 CPCI）、爱思唯尔 EI 数据库；专利数据来源于 IncoPat 数据库；基金项目、科技奖励、荣誉称号等信息均来源于互联网公开信息；基本信息包括性别、年龄、所在机构、提名渠道等，均来源于互联网公开信息。

3.2　特征指标分析

定量指标与描述性指标等各项特征指标共计 98 项，记为自变量 X_i；因变量 Y 为候选人实际通过第二轮评选的情况。为避免过拟合，需去除冗余特征，对 98 项指标（X_i）进行降维。由于工程院的 9 学部的学科特点不同，各项学术成果产出或学术活动对于能否当选院士的影响程度也不同，故对于每个学部分别进行特征选择。

使用 Somers’ D 系数计算各项特征 X_i 对因变量 Y 的关联程度，Somers’ D 的绝对值越大，表明该 X_i 与 Y 之间的序数关联越强，意味着该特征对因变量 Y 具有较大的影响。针对 9 个学部，分别计算并筛选 Somers’ D 绝对值大于 0.1 的 X_i，得到指标数量在 24 项至 67 项之间，形成“部分指标”集合。每个学部选择指标具体情况如表 4 所示。

由表 4 可以看出，不同学部中，各项特征 X_i 与因变量 Y 的关联程度差异较大。除了工程管理学部和信息与电子工程学部外，提名方式对于能否当选院士均存在较大关联；专利类指标在化工、冶金与材料工程学部，机械与运载工程学部，信息与电子工程学部关联性较为明显；在医药卫生学部，中文论文类指标存在显著的负面影响。

doi:10.3772/j.issn.2095-915x.2024.05.010

表 4　不同学部指标选择情况

学部	筛选指标数量	筛选指标平均 Somers' D 绝对值	指标关联度 TOP5
工程管理学部	39	0.17	Q1 论文总数（0.3644***） Q1 通讯论文数（0.3401***） 英文通讯论文数（0.3178***） SCI 论文总数（0.3158***） Q4 论文总数（0.3097***）
化工、冶金与材料工程学部	46	0.19	院士提名（0.4390***） 科协提名（-0.3659***） 专利权利终止数（0.3608***） 实用新型专利数（0.3100***） 第一发明人专利数（0.2998***）
环境与轻纺工程学部	49	0.20	奖项总计（0.4023***） 院士提名（0.3530***） 科协提名（-0.3530***） Q3 一作论文数（0.3465***） Q3 论文总数（0.3350***）
机械与运载工程学部	67	0.25	院士提名（0.4366***） 科协提名（-0.4366***） 第一发明人专利数（0.3918***） SCI 一作论文数（0.3860***） 专利有效量（0.3782***）
能源与矿业工程学部	51	0.24	院士提名（0.5122***） 科协提名（-0.5122***） SCI 通讯论文数（0.4129***） SCI 论文总数（0.4077***） EI 论文总数（0.3782***）
农业学部	57	0.26	院士提名（0.5153***） 科协提名（-0.4565***） Q2 论文总数（0.4437***） Q2 通讯论文数（0.4309***） 奖项总计（0.4207***）
土木、水利与建筑工程学部	60	0.23	项目总计（0.4335***） CPCI 论文总计（0.3780***） 奖项总计（0.3687***） 自科基金项目数（0.3651***） 院士提名（0.3370***）
信息与电子工程学部	24	0.16	实用新型专利数（-0.2831***） 国家科学技术进步奖二等奖（0.2308***） SCI 一作论文数（0.2099***） CPCI 论文总计（0.1995***） 项目总计（0.1864***）
医药卫生学部	26	0.22	中文论文篇均被引频次（-0.3937***） 中文论文 h 指数（-3805***） 院士提名（0.3714***） 中文论文总被引频次（-0.3714***） 中文论文单篇最高被引频次（-0.3490***）

3.3 模型选择与训练

（1）模型选择

本研究运用机器学习方法进行模型构建，机器学习算法能够有效处理具有大量特征的高维数据，擅长揭示非线性关系，在多学科整合应用中具有较强灵活性和泛化能力。

变量 Y 代表候选人实际是否通过第二轮评选，取值只有“1”（通过）或“0”（未通过），故本模型处理的是一项二分类问题，根据数据分布情况选取合适的机器学习模型进行训练。鉴于不同机器学习模型对数据集的适应性各有差异，本研究中的样本数据集的特征概括如下：①数据集规模较小（样本数量 531 项），无需考虑计算资源因素；②数据集维度较高（全部指标共计 98 个），需重点考虑模型复杂度和过拟合问题；③数据缺失值较少，各项指标数据均通过公开渠道收集获得，无需考虑机器学习模型对缺失值的敏感度；④指标特征既包括分类变量，又包括数值变量，需要模型支持混合型的数据输入。

为了找到最佳拟合本研究数据集的模型，选择多种类型的机器学习方法进行测试，包括：传统统计模型，如逻辑回归（LR）；基于边界的模型，如支持向量机（SVM）；基于树的方法，如随机森林（Random Forest）、梯度提升树（GBDT）、极端梯度提升（XGBoost）和轻型梯度提升（lightGBM）；神经网络方法，如多层感知机（MLP）；基于实例的方法，如 K 近邻（KNN）；贝叶斯方法，如朴素贝叶斯（Naive Bayes）。

在深入分析数据分布特性的基础上，遴选与数据内在结构相匹配的模型，以确保学习过程的最优性能，具体模型选择说明如表 5 所示。

表 5　模型选择说明

序号	模型名称	选择原因
1	逻辑回归（LR）	最常见的二分类模型；计算效率高；易于解释和理解
2	支持向量机（SVM）	对于数据集的特征维度和样本量都有较好的适应性；在复杂非线性问题上表现良好
3	多层感知机（MLP）	常见深度学习模型；具有很强的表达能力和学习能力，适用于各种类型的数据集
4	朴素贝叶斯（Naive Bayes）	计算速度快；对小规模数据集表现良好
5	K 近邻算法（KNN）	适用于本项目小样本量数据集
6	随机森林（Random Forest）	能够处理高维数据和大规模数据集；对于特征的重要性排序具有较好的解释能力
7	梯度提升树（GBDT）	具有较强的拟合能力和预测精度
8	极端梯度提升（XGBoost）	在零膨胀数据集中表现出色
9	轻型梯度提升（lightGBM）	适用于高维特征数据集

（2）模型训练

将每个学部的部分指标和全部指标作为特征参数输入模型，通过模型训练找到 $Y=f(X_i)$ 的隐含关系，从而根据 X_i 值计算出候选人当选院士的概率，可视为院士候选人评价的综合得分。对 9 个学部分别构建模型，每个学部均采

doi:10.3772/j.issn.2095-915x.2024.05.010

用 9 种机器学习模型评价预测，每个模型分别使用全部指标和部分指标作为数据输入，即共构建 9×9×2=162 个机器学习模型，预测准确率如表 6 所示。

表 6　模型预测准确率（单位：%）

各模型预测准确率		LR	SVM	MLP	Naive Bayes	KNN	Random Forest	GBDT	XGBoost	Light GBM	综合预测准确率
工程管理学部	部分指标	60.6	59.2	57.6	61.8	55.7	58.2	54.4	57.7	57.8	58.1
	全部指标	55.8	53.5	52.4	44.3	50.6	54.9	52.5	49.5	57.8	52.4
化工、冶金与材料工程学部	部分指标	66.8	59.4	60.2	66.8	64.0	61.0	56.0	58.1	60.8	61.5
	全部指标	61.3	60.7	58.2	64.8	63.5	60.3	57.9	58.4	59.9	60.6
环境与轻纺工程学部	部分指标	66.6	64.7	58.2	57.2	60.5	57.7	53.2	54.9	54.8	58.6
	全部指标	59.3	59.9	52.1	40.4	49.4	58.2	50.1	49.8	53.3	52.5
机械与运载工程学部	部分指标	66.8	56.5	58.4	65.2	56.3	61.1	62.3	59.8	60.7	60.8
	全部指标	66.0	56.0	57.9	55.8	58.3	60.7	60.5	58.7	58.7	59.2
能源与矿业工程学部	部分指标	67.9	73.0	65.7	69.6	70.1	70.9	68.6	68.8	65.1	68.9
	全部指标	62.9	66.0	63.4	63.9	63.9	66.4	66.6	68.3	62.3	64.9
农业学部	部分指标	71.2	66.0	65.2	65.4	62.9	61.0	63.3	63.1	58.8	64.1
	全部指标	68.2	62.7	62.9	57.6	62.3	61.1	65.6	63.9	58.4	62.5
土木、水利与建筑工程学部	部分指标	57.1	62.5	58.8	61.4	56.0	61.1	63.3	60.8	62.3	60.4
	全部指标	57.7	63.2	59.5	60.5	60.8	62.1	63.6	62.2	61.9	61.3
信息与电子工程学部	部分指标	56.8	59.8	53.5	57.0	58.4	56.7	56.1	54.7	58.8	56.9
	全部指标	52.8	55.7	54.7	60.0	52.6	52.0	49.2	47.3	58.4	53.6
医药卫生学部	部分指标	70.1	66.9	65.5	68.9	64.4	61.2	63.1	63.1	64.9	65.3
	全部指标	63.5	61.1	57.5	61.1	52.9	58.3	56.9	56.2	61.8	58.8

针对每个学部，根据测试结果选取表现最好的模型及指标集合用于后续模型预测，具体选择模型以及指标情况如表 7 所示。

表 7　各学部模型及参数选择

学部名称	选择预测模型	选择指标数量	模型准确率
工程管理学部	Naive Bayes	39	61.80%
化工、冶金与材料工程学部	LR	46	66.80%
环境与轻纺工程学部	LR	49	66.60%
机械与运载工程学部	LR	67	66.80%
能源与矿业工程学部	SVM	51	73.00%
农业学部	LR	57	71.20%
土木、水利与建筑工程学部	GBDT	97	63.60%
信息与电子工程学部	Naive Bayes	97	60.00%
医药卫生学部	LR	26	70.10%

3.4 模型评价效果分析

本研究借鉴推荐系统思想，使用 hits 指标评价模型效果，hits@n 即模型评价综合分值的前 n 个候选人在不同年份增选时累计命中的数量。针对样本群体，使用表 7 中的模型及参数，分别对不同学部候选人通过评审的概率进行计算，去除实际未进入第二轮评选的人员后，输出每个学部入选概率最高的人员名单。将模型输出与 2019、2021、2023 年工程院院士增选实际结果对比，如表 8 所示。

表 8　模型评价结果在不同年份的累计命中情况

hits@n	年份	工程管理学部	化工、冶金与材料工程学部	环境与轻纺工程学部	机械与运载工程学部	能源与矿业工程学部	农业学部	土木、水利与建筑工程学部	信息与电子工程学部	医药卫生学部
hits@3	2019	2	1	0	1	0	1	2	1	0
	2021	2	2	1	2	1	1	3	1	2
	2023	2	3	1	3	2	1	3	2	2
hits@5	2019	2	2	0	1	2	2	4	1	0
	2021	2	4	1	3	3	2	5	1	4
	2023	2	5	1	4	4	2	5	2	4
hits@10	2019	3	4	1	4	3	4	5	4	3
	2021	4	6	3	8	5	5	7	4	7
	2023	4	7	4	9	6	6	9	5	8
hits@15	2019	4	5	4	5	5	5	7	6	6
	2021	7	8	6	11	7	6	10	9	11
	2023	7	11	8	12	8	8	13	10	12

通过对比模型输出与 2019、2021、2023 年院士增选的实际情况，结果如下：①模型在所有学部中评价结果较为准确。在全部学部中 hits@3 指标均大于 1，表示模型在各学部推荐的前 3 名候选人，截至 2023 年均有至少 1 人增选成为工程院院士。②不同学部的模型命中率存在差异，整体命中率为 46.67%~86.67%。其中，表现最好学部为土木、水利与建筑工程学部，模型评价并推荐的 15 位学者，有 13 人在未来 3 届内成功当选为工程院院士。③三年新增选结果对比发现，模型的评价结果与实际增选结果的一致性逐年提升，说明模型在院士增选评价预测场景下具有一定稳定性和持续性。

4 研究结论与展望

本研究基于科技人才的定量和定性指标数据，以中国工程院院士候选人作为研究对象，采用机器学习方法，训练并构建形成适用于不同学科领域的人才评价模型，进行实证研究。

研究结果表明，模型在中国工程院院士评价预测场景中具有较好适用性，在土木、水利与建筑工程领域，模型输出的 15 人名单中，高达 86.67% 的学者在未来 3 届内成功当选为工程院院士。机器学习模型输出的评价结果基于客观数据，不涉及专家打分或指标赋权，具有客观性；模型能够适应不同学科领域的特点，快

doi:10.3772/j.issn.2095-915x.2024.05.010

速地对大量数据进行分析和处理，为科技人才评价工作提供了客观、准确、高效的方法。总体而言，本模型在识别出具有院士科研能力和潜质的高层次科技人才上具有很高的适用性，能够在大范围内快速协助识别潜在人才，或为同行评议提供参考和支持，在人才规划、人才储备、人才引进等工作中具有现实意义。此外，本研究所采用的机器学习方法具备较强的泛化能力，通过挖掘样本特征与评价结果之间的隐含关系，能够灵活应用于不同领域和层次的人才评价场景，尽管目前的研究样本主要集中在工程院院士及候选人，但通过相同的方法和流程，可以将这些模型和指标推广应用于更广泛科技人才评价中。

本研究尚存在一定的局限性和不足之处，一方面，限于时间和数据采集成本等因素，模型的训练数据时间跨度较小，若能补充更长时间范围的数据进行模型训练，则能够识别到更多历史趋势和变化特征；另一方面，模型选择的研究对象为院士候选人，输出结果仅在高层次人才层面展现出较好的适用性，为确保其在更广泛的人才群体中保持高效性和准确性，对于某些特定领域、不同类型的人才，需要补充采集相应数据进行训练和调优。

为提升当前模型在不同学科领域中的科技人才评价准确率，未来计划在以下两方面进一步对模型进行优化和扩展。一是增加更多反映科技人才科研产出和学术活动的指标，使评价的维度更加全面、多元，提升模型成功拟合实际影响因素的概率；二是扩充样本数据量，增强模型的训练基础，从而提高其评价结果的精度和稳定性。此外，本研究所采纳的指标体系与模型方法具备良好的扩展性，能够适用于不同的数据集进行分析，未来可以扩大至其他科技奖励或荣誉的潜在人才识别应用。

参 考 文 献

[1] 刘璇 , 段宇锋 , 朱庆华 . 基于合著网络的学术人才评价方法研究 [J]. 情报杂志 , 2014, 33(12): 77-82.

[2] 宋培彦 , 冯超慧 , 龙晨翔 , 等 . 基于颠覆性指数优化的细分领域优秀科技人才发现研究 [J]. 情报杂志 , 2022, 41(5): 61-65.

[3] 常鏐鏐 , 高继平 , 师丽娟 . 调和 H 指数视角下的技术创新人才识别方法与实证研究 [J]. 中国科技论坛 , 2021(11): 104-112.

[4] 孙峰 . “双一流”建设背景下我国高校教学科研人才多维度评价体系构建探析 : 以华南理工大学为例 [J]. 科技管理研究 , 2022, 42(14): 79-84.

[5] 王喆 , 杨国栋 . 高校博士后科技人才培养绩效评价指标体系构建研究——基于平衡计分卡方法 [J]. 科技管理研究 , 2021, 41(15): 127-133.

[6] YE Y, ZHU H, XU T, et al. Identifying high potential talent: A neural network based dynamic social profiling approach[C]//2019 IEEE International Conference on Data Mining. Beijing: IEEE, 2019: 718-727.

[7] YE J, HU H, CHAI C. A talent classification method based on SVM[C]//2009 International Symposium on Intelligent Ubiquitous Computing and Education. Chengdu: IEEE, 2009: 160-163.

[8] JANTAN H, HAMDAN A R, OTHMAN Z A. Human talent prediction in HRM using C4.5 classification algorithm[J]. International Journal on Computer Science and Engineering, 2010, 2(8): 2526-2534.

[9] SUN Y, ZHUANG F Z, ZHU H S, et al. The impact of person-organization fit on talent management: a structure-aware convolutional neural network approach[C]// In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: Association for

Computing Machinery, 1625-1633.

[10] QIN C, ZHU H, XU T, et al. An enhanced neural network approach to person-job fit in talent recruitment[J]. ACM Transactions on Information Systems, 2020, 38(2): 1-33.

[11] NIJS S, DRIES N, VAN V, et al. Reframing talent identification as a status-organising process: Examining talent hierarchies through data mining[J]. Human Resource Management Journal, 2021, 32(1): 169-193.

[12] 田军 , 刘阳 , 周琨 , 等 . 陕西省科技人才评价指标体系与评价方法构建 [J]. 科技管理研究 , 2022, 42(4): 89-96.

[13] 丁宁 , 苏颖 , 胡豫 , 等 . “ 破四唯 ” 背景下我国高校附属公立医院人才评价体系分析 [J]. 中华医院管理杂志 , 2021, 37(12): 953-957.

[14] 张熠 , 倪集慧 . 科技创新人才评价指标体系构建 [J]. 统计与决策 , 2022(16): 172-175.

[15] 余波 , 陈仕吉 , 赵嘉懿 . 中国高校科技人才评价的影响因素及指标体系构建研究 [J]. 农业图书情报学报 , 2023, 35(7): 63-74.

[16] 王运红 , 潘云涛 , 赵筱媛 . 基于科技信息大数据的科技人才科研综合能力评价及应用研究 [J]. 医学信息学杂志 , 2017(38): 7-14.

[17] DAVID PENDLEBURY. Researchers of Nobel class: Citation Laureates 2023 [EB/OL]. (2023-09-19) [2024-01-28]. https://clarivate.com/blog/researchers-of-nobel-class-citation-laureates-2023/.

[18] 井润田 . 高校科研团队管理与战略科学家能力建设 [J]. 上海交通大学学报（哲学社会科学版）, 2022, 30(4): 43-56.

[19] 冯粲 , 童杨 , 闫金定 . 关于培养使用科技人才的思考——基于中外 100 位科技人才的履历分析 [J]. 科技导报 , 2022, 40(16): 38-45.