中国竞争情报事业30年暨第三十届中国竞争情报年会征文通知

由中国科学技术情报学会竞争情报分会主办的“中国竞争情报年会”是情报和信息领域分享学术研究成果、交流竞争情报实践的盛会，已成为业界品牌，吸引了情报和信息界、咨询界及企业界的专家学者和实践者的积极参与，并引起了社会和媒体的广泛关注。

以 1994 年竞争情报专业委员会（竞争情报分会前身）成立为标志，2024 年是中国竞争情报事业走过 30 年发展历程的一年。回顾这 30 年，正是我国竞争情报业在有关各方的关心指导、推动参与下迅速发展并产生重要影响的 30 年。为纪念竞争情报事业和分会走过的不平凡历程，宣传好行业创造的辉煌成绩，激励广大竞争情报从业者更加热爱竞争情报工作，积极参与国家经济和社会建设，分会兹定于 2024 年 9 月在北京举办“新质新域新篇章—中国竞争情报事业 30 年暨第三十届中国竞争情报年会”，现向会员及同仁广泛征集年会论文，欢迎大家积极投稿。同时希望常务理事、理事、团体会员单位积极组织本单位本部门同事撰写论文。会议期间设论文宣讲并颁发证书，结集发行论文集。本届年会征稿议题包含但不限于以下主题，供投稿作者选题参考：

1. 纪念与回顾文章（使用记叙文、议论文、散文、诗歌、随笔等形式体裁不限

（1）畅谈中国竞争情报 30 年发展与未来展望；

（2）展现竞争情报业在我国经济和社会建设中发挥的重要作用；

（3）讴歌竞争情报从业者新风采；

（4）寄语竞争情报业发展美好明天；

（5）其他议题等。

2. 学术研究论文（格式见下页“征文要求”）

（1）新赛道上竞争情报的新服务；

（2）人工智能在情报工作中的应用；

（3）国家发展战略的竞争情报保障；

（4）新技术对竞争情报技术突破的影响；

（5）国家重大变革时代的竞争情报；

（6）竞争情报的机遇和挑战；

（7）人工智能、大数据、云环境下的情报工作研究；

（8）国家科技创新体系建设与竞争情报；

（9）国家、产业、区域经济竞争情报及其作用探讨；

（10）企业竞争情报实践应用；

（11）新兴产业兴起与竞争情报应对；

（12）兼顾安全与发展的竞争情报研究；

（13）反竞争情报、商业秘密保护与网络信息安全；

（14）国外竞争情报发展研究；

（15）中小企业竞争情报的发展；

（16）竞争情报教育与人才培养；

（17）科技情报机构发展竞争情报的战略思考；

（18）竞争情报理论发展与创新；

（19）竞争情报方法创新与应用；

（20）竞争情报技术创新与实践；

（21）竞争情报工作的保障体系；

（22）竞争情报学科建设；

（23）其他相关议题等。

【温馨提示】1. 请将稿件添加附件发送至分会邮箱：scic-staff@scic.org.cn（主题为“第三十届年会征文”）；2. 投稿截止日期为 2024 年 8 月 15 日；3.2024 年 8 月 16 日 -31 日寄发论文作者录用函；4. 征文要求、论文格式、有关事项及第三十届中国竞争情报年会事宜可登陆分会官网 http://www.scic.org.cn 及公众号：中国竞争情报 及时关注进展情况。

联系人：董老师，010-68962474

中国科学技术情报学会竞争情报分会

二〇二四年四月

目次 Contents

情报分析 | INFRMATION ANALYSIS

003 科学家学术谱系学术影响力扩散因素研究 —— 以情报学家严怡民为例
A Study on the Factors of Diffusion of Academic Influence in Scientists' Academic Genealogy——Take Research Scientist Yan Yimin as An Example
程豪杰

014 科技创新与经济高质量发展耦合协调的时空演变分析及障碍因素诊断——以山东省为例
Spatiotemporal Evolution Analysis and Obstacle Factor Diagnosis of the Coupling and Coordination Between Technological Innovation and High-Quality Economic Development——Take Shandong Province for Example
程铭　袁文琦　姜媛　姜明月

028 创新全过程视角下产学研创新团队潜在合作对象识别方法研究
Researching the Methodology for Identifying Domain Potential Collaborative Entities from the Perspective of Innovation Process
张伟然　王超　许海云　武华维　李欣

049 基于 LDA-ARIMA 的我国智能手机关键技术主题识别与演化分析
Identification and Evolution Analysis of Key Technologies Topics of Smartphones in China Based on LDA-ARIMA
庞庆华　姚玉康　张丽娜

063 基于 ISM-MICMAC 的数据资产价值影响因素研究
Research on Influencing Factors of Data Asset Value Based on ISM-MICMAC
刘栟　周雨　郝雪镜

077 基于文本挖掘的颠覆性技术识别方法综述
A Review of Research on Identification of Disruptive Technologies Based on Text Mining
孙心怡　梁洪振　刘桂锋

088 近 20 年高校图书馆知识服务研究热点与趋势
Research Hotspots and Trends of Knowledge Service in University Libraries in the Past 20 Years
许惠　程川生

文本分析 | TEXT ANALYSIS

104 基于关键特征增强的金融长文本事件分类
Key-Features Enhanced Financial Long Text Event Classification
王洁　李旭晖

114 基于教材文本语料库的自适应主题词表构建 —— 以经济类专业为例
A Textbook Corpus Approach to Constructing a Self-adaptive Subject Word List——Taking the Economics-relevant Majors as an Example
杭建琴　张鸣宇　胡泽文

情报工程
QINGBAO GONGCHENG

2024 年 第 10 卷 第 3 期

编委会 • EDITORIAL BOARD

主任委员：

刘琦岩

副主任委员：

乔晓东　潘云涛　孙建军　陆　伟　姚长青　朱礼军

编委委员（按姓氏汉语拼音为序）：

Ume, Anton（罗马尼亚）	安小米	Coh, Byoung-Youl（韩国）	陈仕吉
陈文广	程永红	迟东训	杜永萍
顾进广	胡春明	Jung, Hanmin（韩国）	黄智生
李广建	刘桂锋	李　贺	Meng, Lin（日本）
刘伟丽	刘玉琴	刘忠宝	卢小宾
罗准辰	孟　遥	Rana, F. Omer（英国）	漆桂林
钱　庆	孙　卫	孙　新	Mandl, Thomas（德国）
汪雪锋	吴晨生	吴广印	吴小兰
夏昊翔	杨　波	杨思洛	杨秀丹
俞立平	翟　云	张德政	张家俊
张　军	张　全	张学福	张　嶷（澳大利亚）
张智雄	章成志	真　溱	

doi:10.3772/j.issn.2095-915x.2024.03.001

科学家学术谱系学术影响力扩散因素研究
——以情报学家严怡民为例

开放科学（资源服务）标识码（OSID）

程豪杰

南昌大学　南昌 330031

摘要：[目的 / 意义] 科学研究的传承演进与科学家学术谱系的繁衍发展密切相连，探索学术谱系影响力扩散因素的重要性已经得到学界的共识。构建科学家学术谱系学术影响力扩散因素模型，有助于进一步丰富科学家学术谱系学术影响力研究，同时为科学家学术谱系的发展提供参考。[方法 / 过程] 以情报学家严怡民学术谱系为例，采用扎根理论研究方法，通过对严怡民学术谱系四代成员的深度访谈，系统性探讨影响学术谱系学术影响力扩散的内外部因素，并分析其中的作用机制。[结果 / 结论] 研究发现影响科学家学术谱系学术影响力扩散的五类因素为科研成果产出、科研成果传播、成果所在领域、学术合作和谱系成长性。

关键词：学术谱系；学术影响力；影响因素；扎根理论

中图分类号：G252.8；G35

A Study on the Factors of Diffusion of Academic Influence in Scientists' Academic Genealogy——Take Research Scientist Yan Yimin as An Example

CHENG Haojie

Nanchang University, Nanchang 330031, China

Abstract: [Purpose/Significance] The inheritance evolution of scientific research is closely connected with the reproduction and development of the academic genealogy of scientists, the importance of exploring the diffusion factors of the influence of academic genealogy has been recognized by the academic community. This study constructs a model of the diffusion factors of academic influence of scientists' academic genealogy, which helps further enrich the study of academic influence of scientists' academic genealogy, and at the same time provides a reference for the development of scientists' academic genealogy. [Methods/

基金项目　南昌大学国家社科基金培育资助项目“基于谱系学进路的人工智能算法生成内容知识产权制度研究”（1015-3101000314）；江西省知识产权保护软科学项目“江西省虚拟现实产业涉外知识产权风险管理研究”（2023-41）。

作者简介　程豪杰（1991-），博士，馆员，主要研究方向为信息分析与应用，E-mail：haojiecheng@ncu.edu.cn。

引用格式　程豪杰 . 科学家学术谱系学术影响力扩散因素研究——以情报学家严怡民为例 [J]. 情报工程 , 2024, 10(3): 3-13.

doi:10.3772/j.issn.2095-915x.2024.03.001

Processes] Taking the academic genealogy of intelligence scientist Yan Yimin as an example, this study adopts the grounded theory research method to systematically explore the internal and external factors affecting the diffusion of academic influence of the academic genealogy and analyze the functioning mechanism of these factors through the in-depth interviews with four generations of members of the academic genealogy of Yan Yimin. [Results/Conclusions] The study found that the five types of factors affecting the diffusion of academic influence of scientists' academic genealogy are scientific research output, dissemination of scientific research results, achievements field, academic cooperation, and growth of the genealogy.
Keywords: Academic Genealogy; Academic Influence; Influencing Factors; Grounded Theory

引言

科学家在科学研究方面作出的贡献，不仅体现在其科研成果所产生的价值，更体现在学术传承者的培养上。科学研究的传承演进与科学家学术谱系的繁衍发展密切相连[1]。不少学者试图通过梳理诺贝尔奖获得者的学术背景，探索科技人才成长的规律。

Chariker 等[2]通过对博士生与其学位论文导师关系的分析，发现诺贝尔奖获得者师生关系的模式是非随机的，诺贝尔奖获得者拥有更多的诺贝尔奖获得者导师和学生。诺贝尔奖得主 Krebs[3] 甚至把他在科学上取得成功的主要因素归于他的导师。科学研究的传承和发展与科学家学术谱系的发展密切相连。近年来，部分学者尝试定量化测量科学家对其学术谱系成员的影响：Rossi 等[4]基于文献计量学的 H 指数，建立了“谱系指数”，以衡量科学家对谱系成员的影响，通过比较菲尔兹奖与沃尔夫数学奖获得者的“谱系指数”，发现后者对谱系成员的影响更大；Lu 等[5]采用一组 H 指数，来衡量一个节点影响，研究证明可以使用系列 H 指数衡量导师对谱系成员的学术影响。杨晓薇等[6]则以马费成学术谱系为例，从 3 个维度构建学术谱系评价指标，以量化分析学术谱系的繁衍与发展。

随着研究的深入，如何计量科学家学术谱系在学界的影响力也成了新的主题：吕瑞花等[7]基于 H 指数建立了个体学术影响力、代际学术影响力、学术谱系学术影响力贡献度等指标，以计量学术谱系学术影响力，并以此比较不同学术谱系的学术影响力；韩天琪等[8]则直接从学术谱系的成果产出情况（谱系论文数量）来反映一个学术谱系的学术影响力。近年来，也有学者关注到科学家学术谱系发展的影响因素，程豪杰等[9]通过扎根理论方法发现科学家学术谱系繁衍的影响因素包括导师因素、学生因素、谱系建设、学术环境和科研机遇等五个方面的因素，常欢[10]在其学位论文中也从学生培养机制、学科、社会环境等外部环境角度论述了学术谱系发展的影响因素。

学术谱系的学术影响力既是学术谱系发展的结果也是学术谱系发展的动力，探索学术谱系影响力扩散因素的重要性已经成为学界的共识。然而前人对学术谱系学术影响力扩散因素的研究较少，也未涉及对学术谱系学术影响力扩散因素的挖掘。本研究以情报学家严怡民学

术谱系为例，采用扎根理论，通过对严怡民学术谱系四代成员的深度访谈，系统性探讨影响学术谱系学术影响力扩散的内外部因素，并分析其中的作用机制。

1 研究设计

1.1 研究方法

扎根理论是在经验资料基础上，运用系统化的程序建立理论的较为科学有效的质性研究方法。其主要应用于管理学领域的研究，尤其适用于尚未成型或者还未达成共识的问题。因此，我们认为采用扎根理论对科学家学术谱系学术影响力扩散的因素进行系统性探索是合适的。本研究借鉴经典扎根理论的三阶段分析方法，首先通过个人深度访谈收集原始数据；然后通过开放式编码、主轴编码及选择性编码，逐步分析数据，直至理论达到饱和；最后从数据中浮现理论，构建科学家学术谱系学术影响力扩散因素模型。该模型主要回答以下两个问题：①科学家学术谱系学术影响力扩散的因素主要有哪些；②科学家学术谱系学术影响力扩散因素之间的关系及作用机制是怎样的。

1.2 访谈提纲设计

本研究的访谈内容涉及学术谱系发表论文的数量、参考文献的选择、合作者的选择、交叉学科研究、学术会议等情况。访谈过程当中，并不完全局限于访谈提纲的顺序和内容，而是根据访谈的实际情况灵活处理。访谈提纲见表 1。

表 1 访谈提纲

问题
1. 您每年发表论文的数量受哪些因素的影响？
2. 您在选择参考文献的时候会倾向于有哪些特点的文献，如发表在较高级别期刊上的文献、师门内部的或者熟人所著的文献等？
3. 您在选择合作者的时候会考虑哪些因素？
4. 有没有跨学科合作的一些项目？能否简要介绍一下？
5. 近五年有没有大型学术会议或者学校的邀请报告，大概有哪些？总数大概是多少，每年平均是多少？
6. 目前影响您学术团队影响力扩散的因素有哪些，这些因素之间有哪些关系？

1.3 样本选择与数据收集

扎根理论研究方法属于目的性非概率抽样。其样本的选择强调代表性与研究深度，选择那些能够为研究问题提供丰富信息的样本。因此，本文在受访对象的选择上，遵循以下原则：①受访对象须是严怡民学术谱系内的活跃成员，对所处的学术谱系的影响力有较深的理解和认识；②尽量保证受访对象分布在学术谱系的不同代际、不同机构，以排除因受访对象分布单一导致的偏差。基于上述原则，通过对工作机构、教育经历、学术后代的数量、学术成就、主持项目、所获荣誉、研究领域等多方面的综合考虑选择受访对象。数据收集与数据分析工作同步进行，并在数据分析中采用不断比较的思路，将数据中的新范畴纳入理论，对理论进行持续修正，直至理论饱和。

原始数据的收集自 2020 年 4 月始至 2020 年 8 月结束。我们依次访谈了严怡民学术谱系内 4 代成员，共有 21 位受访对象接受了访谈邀请，受访对象基本情况统计见表 2。我们对 21 位访谈对象，依据访谈时间顺序依次编码为 A1、A2、A3……A21。每位受访对象的访谈时

doi:10.3772/j.issn.2095-915x.2024.03.001

间为 30~90 分钟。我们首先初步拟好访谈提纲，并在后续的访谈中，根据实际访谈数据的分析结果对访谈提纲及时微调。为确保访谈效果，在正式访谈前，我们将访谈主题以及涉及的相关概念，通过电子邮件、微信等渠道，与受访对象进行事先沟通，以便受访对象深入理解访谈问题，从而有助于访谈的顺利进行。在征求到受访对象同意的情况下进行了录音，并对关键词进行记录。在访谈结束后，我们在 24 小时之内完成访谈记录的整理工作，形成扎根理论分析的原始资料。我们从 21 份访谈记录中随机抽取 16 份访谈记录进行编码分析，剩余 5 份访谈记录留作理论饱和度检验。

表 2　受访对象情况统计

项目	分类	样本数
机构	武汉大学	7
	南京大学	9
	西南大学	1
	中南大学	1
	南京师范大学	1
	湖北大学	1
	兰州大学	1
性别	男	11
	女	10
代际	第一代学生	1
	第二代学生	8
	第三代学生	9
	第四代学生	3
职称	教授	11
	副教授	3
	讲师	2
	其他	5

2　数据分析与模型构建

2.1　开放式编码

开放式编码是指研究者保持开放的心态对收集到的资料进行分析，并采用概念标识与诠释的过程 [11]。基于编码过程严谨性的考虑，我们借鉴了经典扎根理论的编码方法，由 2 名研究者共同编码，并进行不断比较与反复审核。为了保障研究的信效度，在访谈过程中将访谈要点与受访者进行确认，并撰写访谈备忘录。通过备忘录时刻提醒我们对访谈资料中所呈现的概念保持敏感，并注意捕捉新线索。依次通过标签化（b+ 序号）、初步概念化（bb+ 序号）、概念化（B+ 序号）、范畴化（BB+ 序号），对访谈记录逐字逐句分析。由于原始数据较多，本研究只举例说明开放式编码的概念提取过程，如表 3 所示。所有访谈记录共提取了 55 个初始概念，归纳为 22 个范畴（见表 4）。

2.2　主轴编码

主轴编码是在开放式编码的基础上，进一步的编码分析 [11]。主轴编码阶段的内容是通过聚类，整合相似范畴，增加范畴稠密度，发现范畴之间潜在的关系，提炼出主范畴 [12]。根据本研究的特性，通过考察情境关系、类型关系及过程关系，对开放式编码阶段所获得的 21 个范畴进行归纳聚类，最终形成 5 个主范畴（BBB1-BBB5，包括科研成果产出、科研成果传播、成果所在领域、学术合作、谱系成长性），并将主范畴归入谱系科研成果、谱系学术交流和谱系成长性 3 个维度（见表 5）。

2.3　选择性编码

选择性编码是指在主轴编码的基础上，厘清各范畴之间的关系，确定核心范畴，并寻找合适的故事线把核心范畴与其他范畴组织在一

表 3　访谈资料的开放式编码分析举例

受试者 A 的访谈材料	开放式编码			
	标签化……	初步概念化……	概念化……	范畴化 ……
…… 我每年发表论文的数量第一个是受学生的影响。如果我指导的学生比较多（b1），然后他们的产出如果有达到某一种效率（b2）的话，我每年的成果其实大概就会有所保障。因为我们在团队里头经常去分享一些 idea，有的时候是学生在启发我的想法，有的时候老师在启发学生的想法。在这个时期的阶段，学生的想法（b3）是不是丰富或者是有价值就很重要。	b1: 指导学生多 b2: 产出效率 b3: 学生的想法	bb1: 指导学生数量（b1） bb2: 论文产出效率（b2） bb3: 学生的想法（b3）	B1: 学生的数量（bb1） B2: 学生的效率（bb2） B3: 学生的想法（bb3）	BB1: 学生方面的因素（B1、B2、B3）

表 4　开放式编码结果

范畴化		概念化
BB1	学生方面的因素	B1 学生数量、B2 学生的效率、B3 学生的想法、B4 学生的研究能力、B5 学生研究的积极性
BB2	研究项目任务	B6 研究项目任务量、B7 研究项目任务特点、B8 研究项目进展
BB3	科研兴趣	B9 兴趣
BB4	科研的积累和突破	B10 科研积累、B11 科研突破
BB5	科研能力	B12 研究的方向感、B13 研究成果水平
BB6	时间精力	B14 行政事务、B15 用在科研上的时间、B16 教学任务
BB7	研究主题相关性	B17 与研究主题相关的文献
BB8	研究观点的相关性	B18 研究观点支撑
BB9	高质量文献	B19 高水平文献、B20 高级别期刊文献、B21 高被引文献、B22 经典文献、B23 英文文献
BB10	名人的文献	B24 大牛的文献、B25 知名学者的文献
BB11	最新的文献	B26 最近的文献、B27 更新的文献
BB12	自己的文献	B28 自己写的文献、B29 组内所知的文献
BB13	熟人的文献	B30 所知课题组的工作会多留意
BB14	研究方向	B31 研究方向的价值、B32 研究方向的可持续性、B33 研究方向的热度、B34 政策相符度、B35 研究周期、B36 研究方向的渗透性
BB15	契合与互补性	B37 契合性、B38 互补性
BB16	接触交流机会	B39 组内合作、B40 外部合作
BB17	相互的了解和信任	B41 熟悉程度、B42 了解对方
BB18	合作者组合结构	B43 校内合作、B44 校外合作、B45 国际合作、B46 跨学科合作
BB19	团队的开放性	B47 学术交流的态度、B48 学术交流的程度
BB20	团队领导的影响力	B49 导师的名气、B50 导师的学术职务、B51 导师的学术称号
BB21	团队规模	B52 团队在读人数、B53 团队总人数
BB22	团队繁殖速度	B54 团队新增成员数量、B55 团队新增成员比例

doi:10.3772/j.issn.2095-915x.2024.03.001

表 5　主范畴及其与范畴之间的关系

类别	主范畴		范畴		范畴内涵
谱系科研成果	BBB1	科研成果产出	BB1	学生方面的因素	学生的 idea，学生的研究能力，学生研究的积极性（意愿）等影响学生论文发表的因素。
			BB2	研究项目任务	指研究项目任务量和研究项目任务的特点等，科研任务量比较重的时候，学者的科研成果也会相对多一点。研究项目任务的特点，也会影响到学者科研成果的产出，比如研究项目的难度，项目的跨学科程度等。
			BB3	科研兴趣	科学研究的兴趣有的时候受到别人的启发，有的时候受环境的影响，有时候就是自己的无端的感悟。
			BB4	科研的积累和突破	每一个研究领域，或者说在做每一个课题的时候，都需要一定的积累期。随着课题项目的推进，科研工作累积到一定的程度，就会有一种突破，论文的数量和质量都会有大大的提高。
			BB5	科研能力	作为科研者，作为一个课题的主持人，决定一个课题好坏、成败的关键因素在于研究者本人的科研的能力。
			BB6	时间精力	学者投入科研的时间与他的科研产出还是有一个正向的关系。
	BBB2	科研成果传播	BB7	研究主题相关性	在参考文献的选择上，首先要选择与自己研究主题，或者研究问题相关的文献，一定要切题。
			BB8	研究观点的相关性	在文章写作的过程中，尤其是文章讨论部分，如果作者想要表达某一种观点，他往往会根据写作的要求去查找相应的文献。
			BB9	高质量文献	高级别期刊上的文献、经典的文献、高被引文献、英文文献等往往是高质量文献集中的地方。
			BB10	名人的文献	学者都有自己一定的研究领域，而且往往在某一领域研究了很久，很容易知道在某一领域有哪些著名的学者，那么这些著名的作者所著的文献可能会得到关注，从而获得更多的引用。
			BB11	最新的文献	最新的文献是在前人研究基础上的研究，它可能已经包含了旧文献的知识贡献，“所以从省时省力的角度来讲的话，还是会先看先引用那些比较新的文献。”
			BB12	自己的文献	如果学者自己在某一研究方向上持续做下来的话，学者自己的研究也是有延续性的。所以他们也会引用自己或者合作者的一些文献。
			BB13	熟人的文献	学者在情报学的学术范围当中，不难接触到领域内其他学者的研究工作，如果某一学者所熟悉的人，有一项研究是有价值的，而且与自己研究的问题比较相关的话，那么这一项研究可能还是首先获得了解和引用。
	BBB3	成果所在领域	BB14	研究方向	一个正确的研究方向是很重要的。如果选题和政策相符，满足了宏观上的需要，在此基础上做出来的成果就容易接受，容易发表。研究方向还决定研究的周期。
谱系学术交流	BBB4	学术合作	BB15	契合与互补性	契合性，主要指研究方向的契合，合作者双方研究的方向相近。在具体的研究工作中，合作者双方还需要一定的互补性。这样更能实现研究分工明确，进行优势、能力等方面的互补，这也是合作的意义所在。
			BB16	接触交流机会	多数学者与课题组成员之间，或者说与师门内部的合作比较多，是因为组内成员平时接触交流机会比较多，而且有课题在合作，所以这个合作肯定是有的。组外合作可能要找机会。
			BB17	相互的了解和信任	“只有信任他才会跟他合作，只有了解他才知道怎么跟他合作。”科研项目的合作，不是短期的合作关系，不是简单的二者合伙做论文。
			BB18	合作者组合结构	最常见的情况是在申报课题的时候，课题申请人都会考虑研究团队不同成员的构成比例。
			BB19	团队的开放性	团队的开放性与学术谱系的影响力呈正向关系。跟外部的交流合作的机会多不多，形式多不多，团队是不是乐于去接受外部的这种向外部去推广自己的研究成果，同时又乐于接受外部一些研究者的访问交流。
谱系成员	BBB5	谱系成长性	BB20	团队领导的影响力	学术影响力体现在各个方面，如导师的学术品质、导师的科学精神等。
			BB21	团队规模	“团队规模越大，它的影响力就会越大，接触的人越多，它影响力就会越大。”
			BB22	团队繁衍速度	“团队繁殖速度越快，它的学术影响力越大。”

起，以构建理论模型[13]。依据访谈记录的故事线，梳理5个主范畴之间的关系，挖掘出核心范畴，揭示主范畴中的典型关系结构（见表6）。

表6　选择性编码典型关系结构

典型关系结构	关系结构	关系结构的内涵
科研成果产出→学术谱系学术影响力	决定因素	科研成果产出是科学家学术谱系学术影响力扩散的决定因素。
科研成果传播→学术谱系学术影响力	重要因素	科研成果传播是科学家学术谱系学术影响力扩散的重要因素。
成果所在领域→学术谱系学术影响力	重要因素	成果所在领域是科学家学术谱系学术影响力扩散的重要因素。
学术合作→学术谱系学术影响力	重要因素	学术合作是科学家学术谱系学术影响力扩散的重要因素。
谱系成长性→学术谱系学术影响力	重要因素	团队领导的影响力、团队规模、团队繁衍速度等谱系成长性是科学家学术谱系学术影响力扩散的重要因素。

通过上述“故事线”的梳理，本研究确定“科学家学术谱系影响力”这一核心范畴，围绕核心范畴的故事线可概括为：科研成果产出、科研成果传播、成果所在领域、学术合作和谱系成长性5个主范畴对科学家学术谱系的影响力存在显著影响（图1）。其中谱系科研成果方面的因素包括科研成果产出、科研成果传播与成果所在领域三个主范畴。谱系学术交流方面的因素主要指学术合作。谱系成长性则取决于谱系的规模、繁衍速度等。科研成果产出包括研究数量、研究质量、科研的积累和突破、学者科研能力及时间精力，是影响科学家学术谱系影响力的决定性因素。科研成果传播包括研究主题相关性、研究观点的相关性、高质量文献、名人的文献、最新的文献、自己的文献以及熟人的文献，是影响科学家学术谱系学术影响力的重要因素。成果所在领域方面的因素包括研究方向的可持续性与研究方向的热度，是影响科学家学术谱系学术影响力的重要因素。学术合作包括契合与互补性、接触交流机会以及相互的了解和信任，是影响科学家学术谱系学术影响力的重要因素。谱系成长性则指团队领导的影响力、团队规模及团队繁衍速度，也是影响科学家学术谱系学术影响力的重要因素。

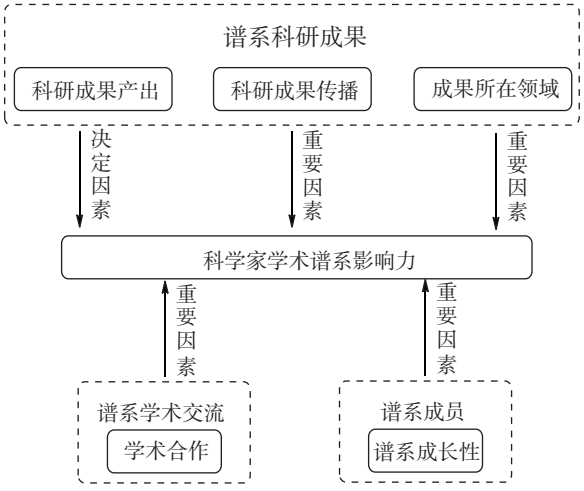

图1　科学家学术谱系影响力扩散因素模型

2.4　理论饱和度检验

利用开放式编码阶段预留的近1/4的访谈记录进行理论饱和度检验，通过开放式编码、主轴编码和选择编码分析，结果并未产生新的重要范畴和关系，未发现新的主范畴关系结构，5个主范畴内部也未发现新的构成因子。由此，上述“科学家学术谱系学术影响力扩散因素”的内在机理在理论上已达到饱和。

doi:10.3772/j.issn.2095-915x.2024.03.001

3 研究结果讨论

3.1 科研成果产出

科研成果产出的数量与质量是科学家学术谱系学术影响力扩散的决定性因素，科研成果产出包括学生方面的因素、研究项目任务、科研兴趣、科研的积累和突破、学者的科研能力以及时间精力 6 个范畴。

科研工作需要积累，也需要分工。随着团队规模的壮大，学术团队之间会经常分享学术想法，有时候是学生在启发导师，有时候是老师在启发学生。情报学本身和传统社会科学的研究并不相同，其学科交叉特征明显，开展相关研究需要数据采集、技术、团队协作等方面的人员分工协作完成，很难依靠团队成员个人完成项目。团队负责人也很少独自完成一篇论文或一个科研项目，如访谈对象 A19 表示“如果团队负责人指导下的学生数量较多、有研究能力、有想法、有热情，那这样的团队的科研成果产出情况必然是乐观的”。

团队科研活动一般具有明确的目标，必须在一定的时间、经费、资源框架内完成相应的科研项目。科研任务重的时候，团队科研成果产量也会相对较多。团队负责人申请到的项目，总是要尽量去完成，在完成项目任务的过程中，就会有相应的科研产出，如访谈对象 A16 表示“根据研究项目的进展，项目完成的中后期成果的数量和质量往往会高一些”。从研究项目任务的特点上来说，研究项目难度、跨学科程度、前期积累等也会影响到团队科研成果的产出。

每一个科研项目，都是一个科研积累的过程。课题开展早期积累阶段，科研成果产出数量可能比较少。随着课题研究的推进，科研积累到一定程度，研究者会逐渐找到课题的突破点，科研论文产出的数量和质量都会有提高。因此，科研成果的产出与科研的积累和突破有相关性。而且这种积累不仅体现在研究者自身科研经验的积累上，还体现在前人研究成果的积累上。

决定一个研究课题好坏与成败的关键因素在于研究者本人的科研能力。这种能力可能体现在很多方面，比如创新能力、独立研究的能力、独立思考的能力等。虽然合作研究很重要，比如在跟合作者沟通的过程中会得到很多启发，但是研究者本人是研究的主体，其研究课题还是需要研究者自己去把握。

此外，研究者的科研兴趣与时间精力无疑也会影响科研成果的产出。科研兴趣的来源可能受别人的启发，可能受环境的影响，也可能是研究者无端的感悟；研究者的工作时间除了用在科学研究之外，还需要用在教学、行政事务等事情上，研究者能否自主把握科研时间，也会影响到科研的进度，进而影响到科研成果的产出。

3.2 科研成果传播

科研成果传播的过程也是科学家学术谱系学术影响力扩散的过程。学术文献是科研成果产出的主要表现形式之一，不同特点的学术文献其学术传播的效果也不尽相同：高质量文献、名人的文献、最新的文献、自己的文献、熟人的文献、相关主题的文献、相近观点的文献更容易获得引用与扩散。

高级别期刊文献、经典文献、高被引文献、英文文献等往往是高质量文献集中的地方。面对同样类型的文献，研究者自然愿意引用质量好的文献。文献质量的高低可以参考文献发表的期刊等级、文献被引频次等，高级别期刊文献、高被引文献质量较高的可能性更大。科学研究需要经典理论、经典文献的支撑，这些是做好学术研究的基础。由于国内的考评机制等因素，高质量的文献往往会先投英文期刊，遭遇拒稿之后再转投中文期刊，因此一般情况下英文文献的质量可能会比中文文献稍好一些。

名人的文献、熟人的文献、自己的文献、最新的文献更容易获得引用与传播。研究者在某一领域研究，很容易知道其所在领域有哪些知名的作者，那么这些知名作者产出的文献可能会更容易得到关注，如访谈对象 A9 表示“研究者也不难接触到领域内其他学者的研究工作并与之产生链接，所熟悉学者的文献更容易获得了解和引用，这种文献获取的方式有时候并不是通过数据库检索，而是通过非正式的交流”。另一方面，研究者的研究往往有一定的延续性，也容易引用自己的文献。此外，最新的文献是在前人研究基础上的研究，它可能已经包含了旧文献的知识贡献，所以从省时省力的角度，最新的文献也会更容易获得引用。因此，如果一个学术谱系内名人辈出、在相关领域不断深耕、积极与内外部学者产生链接并及时高效地产出成果，那么这个学术谱系的学术影响力自然不容小觑。

此外，在文章写作的过程中，尤其是文章的讨论部分，研究者需要一定的观点支撑，往往会根据写作的要求去查阅一些文献。因此经得住检验的结论、观点等也是获得引用，扩大学术影响力的重要因素。

3.3 成果所在领域

成果所在领域对科学家学术谱系学术影响力的影响主要体现在政策相符度、研究周期、研究方向的可持续性、研究方向的热度和研究方向的渗透性等范畴。

正确的研究方向是非常重要的，如果选题和政策相符，满足了宏观上的需要，在此基础上做出来的成果就容易被接受，容易发表；研究方向的可持续性则决定了研究成果的连续性以及可持续性，研究方向随着时代的变化，不断衍生出新的更专深的或者是更广泛的研究成果，那么这个过程会不断地产生影响力；研究方向还决定研究的周期，如访谈对象 A20 表示“周期较长的研究项目在一定时期内研究成果的数量可能会比较少”。

值得指出的是，研究成果渗透到的领域也是其影响力传播到的领域。如果某个研究方向与其他学科有一定的交叉，其实相当于拓展了情报学的研究领域，那么相关研究成果的生命力可能会长久。此外，热门的研究方向，显然会得到更多人的关注，因此，研究方向的热度对学术谱系影响力的扩散也是有帮助的。

3.4 学术合作

科学家学术谱系的学术影响力还体现在谱系成员论文合作者的选择上。学术谱系成员之间的合作有利于提高论文成果的质量，进而提高学术谱系的影响力；学术谱系成员与外部学者的合作，不仅有利于提高论文成果的质量，

doi:10.3772/j.issn.2095-915x.2024.03.001

还直接扩散了学术谱系的影响力。而影响研究者学术合作对象选择的因素主要体现在契合与互补性、接触交流机会、相互的了解和信任、合作者组合结构及团队的开放性等 5 个范畴。

研究者双方前期研究方向的契合度影响双方合作意向，进而影响合作的效率。在具体的研究工作中，为了明确研究分工，实现合作者能力与优势的发挥，合作者双方还需要一定的互补性，这也是合作的意义所在。

接触交流机会影响合作的机会，组内合作或者师门内部的合作比较多，是因为组内成员平时接触交流机会比较多，尤其是导师与学生之间的合作。由于情报学领域的科研项目多半带有交叉或跨学科的性质，需要其他学科的知识贡献，特别是一些重大项目，需要与计算机科学、心理学等学科联合起来共同研究，这些项目的开展加强了跨学科合作，其成果也可能具有更重大的实际价值，这个过程也有助于推动情报学的影响力扩散到其他学科。

只有了解合作者才知道怎样与其合作，只有信任合作者才会与其合作。科研项目的合作，不是短期的合作关系，也不是简单的双方合伙做论文，而是一个长期持续的关系。因此，相互的了解和信任是合作的基础，如访谈对象 A17 表示“更多的合作它实际上是一个长期的持续的一个关系，你知道他可以做什么，能把这个做好，你需要跟他来合作，你如果对别人完全不了解，其实没法合作”。

在申报科研项目时，课题申请人往往会考虑研究团队不同成员的构成比例。越是重大的科研项目，越是需要和来自不同领域、不同机构的专家合作，如访谈对象 A15 表示“校内有哪些老师是相关的，校外有哪些老师是相关的，其实在申请每个科研项目的时候，几乎是一种不成文的一种考虑，你会考虑校内外专家或者是这个研究团队的不同的成员的这种比例，合适的这种构成”。此外，如果团队本身乐于接收外部的信息、积极寻找外部合作机会并积极向外部去推广团队的研究成果，也有利于科学家学术谱系学术影响力的扩散。

3.5 谱系成长性

谱系成长性对科学家学术谱系学术影响力的影响主要体现在团队领导的影响力、团队规模及团队繁衍速度 3 个范畴。

严怡民学术谱系在情报学领域具有很高的学术影响力，从严怡民教授到马费成教授再到大批优秀的学术传承者，不断推动情报学的发展。严怡民教授及其谱系成员对情报学的发展作出了突出贡献，也奠定了严怡民学术谱系在情报学的地位。此外，团队规模与团队繁衍速度对学术谱系学术影响力的扩散具有积极的影响，如访谈对象 A11 表示“学术团队影响力扩散，首先我觉得还是看团队规模，团队规模越大，它的影响力就会越大，接触的人越多，本身团队它的开放性与成长性越大，它影响力就会越大”。团队繁殖速度越快，团队规模越容易形成，它的学术影响力越大。

4 研究结论

本研究选取情报学家严怡民学术谱系为研究对象，运用扎根理论方法，通过深度访谈获取科学家学术谱系学术影响力扩散因素的原始资料，运用经典扎根理论三级编码等步骤，提

炼出科学家学术谱系学术影响力扩散因素模型（如图 1）。研究发现，科研成果产出、科研成果传播、成果所在领域、学术合作、谱系成长性五个方面的因素对科学家学术谱系学术影响力扩散产生重要影响。其中科研成果产出对科学家学术谱系学术影响力扩散的影响主要体现在研究项目任务、科研兴趣、科研的积累和突破、学者的科研能力、时间精力以及学生方面的因素等 6 个范畴；科研成果传播的过程也是科学家学术谱系学术影响力扩散的过程，高质量文献、名人的文献、最新的文献、自己的文献、熟人的文献、相关主题的文献、相近观点的文献更容易获得引用与扩散；成果所在领域对科学家学术谱系学术影响力的影响主要体现在政策相符度、研究周期、研究方向的可持续性、研究方向的热度和研究方向的渗透性等范畴；影响学术合作的因素主要体现在契合与互补性、接触交流机会、相互的了解和信任、合作者组合结构及团队的开放性等 5 个范畴，学术谱系成员之间的合作有利于提高论文成果的质量，进而提高学术谱系的影响力，而学术谱系成员与外部学者的合作，不仅有利于提高论文成果的质量，还直接把学术谱系影响力扩散到外部；谱系成长性对科学家学术谱系学术影响力的影响主要体现在团队领导的影响力、团队规模及团队繁衍速度等 3 个范畴。

参 考 文 献

[1] 盛怡瑾 , 赵勇 . 科学家学术谱系的内涵、构建与测度研究述评 [J]. 图书情报工作 , 2023, 67(14): 109-118.

[2] CHARIKER J H, ZHANG Y, PANI J R, et al. Identification of successful mentoring communities using network-based analysis of mentor–mentee relationships across Nobel laureates[J]. Scientometrics, 2017, 111(3): 1733-1749.

[3] KREBS H A. The Making of a Scientist[J]. Nature, 1967, 215(5109): 1441-1445.

[4] ROSSI L, FREIRE I L, MENA-CHALCO J P. Genealogical index: A metric to analyze advisor-advisee relationships[J]. Journal of Informetrics, 2017, 11(2): 564-582.

[5] LU L, ZHOU T, ZHANG Q, et al. The H-index of a network node and its relation to degree and coreness[J]. Nature Communications, 2016, 7(1): 10168.

[6] 杨晓薇 , 雷鸣 , 高继平 . 杰出科学家学术谱系的指标构建与实证研究——以学者马费成为例 [J]. 图书情报工作 , 2022, 66(23): 115-124.

[7] 吕瑞花 , 常欢 . 基于文献计量的科学家学术谱系学术影响力的研究 [J]. 情报理论与实践 , 2017(1): 76-78.

[8] 韩天琪 , 樊小龙 , 袁江洋 . 唐敖庆谱系与福井谦一谱系比较研究 [J]. 科学与社会 , 2013(1): 110-123.

[9] 程豪杰 , 袁勤俭 . 科学家学术谱系繁衍的影响因素研究——以情报学家严怡民为例 [J]. 现代情报 , 2021, 41(5): 77-86.

[10] 常欢 . 基于文献计量学的科学家学术谱系研究 [D]. 北京 : 北京理工大学 , 2016.

[11] 陈向明 . 扎根理论的思路和方法 [J]. 教育研究与实验 , 1999(4): 58-63.

[12] 陈向明 . 质的研究方法与社会科学研究 [M]. 北京 : 教育科学出版社 , 2000: 24-33.

[13] 朱丽叶·M. 科宾 , 安塞尔姆·L. 施特劳斯 . 质性研究的基础 [M]. 重庆 : 重庆大学出版社 , 2015: 5-10.

doi:10.3772/j.issn.2095-915x.2024.03.002

科技创新与经济高质量发展耦合协调的时空演变分析及障碍因素诊断
——以山东省为例

开放科学（资源服务）标识码（OSID）

程铭[1] 袁文琦[2] 姜媛[1] 姜明月[1]

1. 山东省创新发展研究院 济南 250101；
2. 北京师范大学统计学院 北京 100875

摘要：[目的 / 意义] 科技创新是山东省实现经济高质量发展的核心动力，正确处理好科技创新和经济高质量发展的关系，有利于以高质量创新引领经济高质量发展。[方法 / 过程] 选取 2012—2021 年山东省 16 地市为研究对象，从创新投入、创新产出、创新环境、协调发展、绿色发展、开放发展、共享发展 7 个维度构建科技创新和经济高质量发展评价指标体系，利用熵权法、耦合协调度模型和障碍度模型分析两者的时空耦合关系及障碍因素。[局限] 研究未能从产业等中观角度解释科技创新和经济高质量发展的协同机制。[结果 / 结论] 时序上，山东省科技创新、经济高质量发展水平及其耦合协调度呈逐年上升趋势，从濒临失调到初级协调状态；空间上，两者耦合协调度空间差异较大，济南和青岛为良好和优质协调区，沿海城市处于中级和初级协调阶段，鲁西南城市处于过渡和失调阶段；影响两者耦合的主要障碍因素是创新产出和开放发展，具体指标看，外商投资新设企业数、进口依存度、论文数是两者的关键阻碍因素。根据研究结果，提出针对山东省加强科技创新，提升经济高质量发展水平及两者耦合协调度的具体建议。

关键词：科技创新；经济高质量发展；耦合协调；时空分异；障碍度

中图分类号：F124；F224；G35

Spatiotemporal Evolution Analysis and Obstacle Factor Diagnosis of the Coupling and Coordination Between Technological Innovation and High-Quality Economic Development——Take Shandong Province for Example

CHENG Ming[1] YUAN Wenqi[2] JIANG Yuan[1] JIANG Mingyue[1]

1. Shandong Provincial Institute of Innovation and Development, Jinan 250101, China;
2. School of Statistics, Beijing Normal University, Beijing 100875, China

作者简介 程铭（1994-），硕士，助理研究员，主要研究方向为科技统计和科技管理研究；袁文琦（2000-），硕士研究生，主要研究方向为数据科学与管理；姜媛（1981-），本科，副研究员，主要研究方向为科技政策分析和科技管理研究，E-mail：sdkjtj607@163.com；姜明月（1992-），硕士，助理研究员，主要研究方向为科技产业分析与智库建设。

引用格式 程铭，袁文琦，姜媛，等．科技创新与经济高质量发展耦合协调的时空演变分析及障碍因素诊断——以山东省为例 [J]. 情报工程，2024, 10(3): 14-27.

Abstract: [Purpose/Significance] Technological innovation is the core driving force for achieving high-quality economic development in Shandong Province. Properly handling the relationship between technological innovation and high-quality economic development is conducive to leading high-quality economic development with high-quality innovation. [Methods/Processes] Selecting 16 cities in Shandong Province from 2012 to 2021 as the research objects, this study constructs an evaluation index system for scientific and technological innovation and high-quality economic development from seven dimensions: innovation input, innovation output, innovation environment, coordinated development, green development, open development, and shared development. The entropy weight method, coupled co scheduling model, and obstacle model are used to analyze the spatiotemporal coupling relationship and obstacle factors between the two. [Limitations] The study failed to explain the synergistic mechanism between technological innovation and high-quality economic development from a meso perspective such as industry. [Results/Conclusions] The results indicate that in terms of time series, the level of scientific and technological innovation, high-quality economic development, and their coupling coordination in Shandong Province are showing an increasing trend year by year, from a state of near imbalance to a primary coordinated state; In terms of space, there is a significant difference in the degree of coupling and coordination between the two. Jinan and Qingdao are good and high-quality coordination areas, coastal cities are in the intermediate and primary coordination stages, and cities in southwestern Shandong are in the transition and imbalance stage; The main obstacles to the coupling of the two are innovation output and open development. Specifically, the number of newly established foreign investment enterprises, import dependence, and the number of papers are the key obstacles to the two. Based on the research results, specific suggestions are proposed to strengthen scientific and technological innovation, enhance high-quality economic development, and enhance the coupling and coordination between the two in Shandong Province.
Keywords: Technological Innovation; High Quality Economic Development; Coupling Coordination; Spatial-temporal Differentiation; Obstacle Degree

引言

创新是引领发展的第一动力，科技创新支撑着中国经济的高质量发展。2022 年中央经济工作会议对我国科技创新支撑高质量发展工作提出新要求，将科技政策作为支撑经济高质量发展的“五大政策”之一。立足新发展阶段，要遵循以科技创新为导向的经济发展方针 [1]。反之，增强社会经济实力，有助于为科技创新营造有利条件，塑造创新发展新优势。作为全国创新高地和经济第三大省，山东省经济实力雄厚，聚集了大量优质的科技创新要素。《山东省“十四五”科技创新规划》指出，要全面提升全省科技创新能力，引领支撑经济社会高质量发展。因此，本研究在山东省科技创新水平和经济高质量发展水平评价指标体系基础上，探索科技创新和经济高质量发展的耦合协调关系及障碍因子，对区域实现高质量发展具有重要的战略意义。

1 文献综述

1.1 科技创新及其测度

科技创新是指运用新的知识和技术，开发新产品、提供新服务的过程 [2]。如何建立科技创新评价体系是国内学者的研究热点。一方面考虑不同的评价指标选取。许红丹等 [3] 从创新投入、创新产出和创新效率 3 个方面构建科技创新评价指标体系；蔡文伯等 [4] 基于创新投入、

doi:10.3772/j.issn.2095-915x.2024.03.002

创新产出和技术转让进行综合评价。另一方面考虑不同的评价模型。谷斌等[5]运用熵权法、耦合协调评价模型、探索性空间数据分析与标准差椭圆模型测算我国31个省市科技创新能力发展水平；张凯等[6]利用云模型和证据理论对科技创新能力进行评价。

1.2 经济高质量发展及其测度

自从党的十九大报告提出我国经济迈入高质量增长阶段以后，学者们围绕经济高质量发展在理论和实证层面展开了深入的研究。一方面以经济高质量发展的特征为基础构建评价体系。Mlachila等[7]从经济增长的基本属性和社会成果两个维度构建经济高质量发展模型；袁晓玲等[8]通过投入指标体系和产出指标体系构建省域经济高质量发展评价体系。另一方面以新发展理念为指导构建评价体系。李桂君等[9]立足“五大发展理念”构建城市经济高质量发展指标体系；单勤琴等[10]基于经济、创新、协调、绿色、开放和共享6个指标建立区域经济高质量发展评价指标体系。

1.3 科技创新与经济高质量发展的耦合协调关系

科技创新与经济高质量发展相辅相成，两者之间并非单向线性关系[11]，要深刻把握两者之间的协调关系[12]。一方面考虑两者的耦合协调关系。目前缺乏对山东省16地市的研究，更多以全国作为研究区域。姚振飞等[13]利用耦合协调度模型发现山东省创新能力与经济质量的关系存在同步性和耦合性；祝恩元等[14]得到山东省科技创新与可持续发展处于低度协调耦合阶段，呈现“东高西低，中间高南北低”的空间格局；董弋萱[15]揭示中国城市规模相差越大，两者耦合协调水平差异越大；魏巍等[16]运用耦合协调分析法发现中国创新发展体系与经济高质量发展实现初步协调，分析结果与经济发展水平呈“逆向相关”态势。另一方面研究两者耦合协调的影响因素。已有研究主要集中于两者之间单向影响机制，缺乏对于两者耦合协调影响因素的研究，但障碍度模型可以找出限制两者耦合协调发展的关键因素，厘清两者耦合协调的主要影响因子。刘建华等[17]利用障碍因子诊断模型对黄河流域科技创新与低碳经济的耦合协调及其障碍因子进行了分析；尚英仕等[18]运用障碍度模型厘清东部沿海三大城市群科技创新与绿色发展耦合协调的障碍因素；陈正其[19]分析京津冀地区创新要素配置与经济高质量发展耦合协调的障碍因素。

1.4 研究述评

国内外优秀学者对科技创新与经济高质量发展的研究硕果累累，但仍存在一些问题亟待完善：第一，部分研究设计两者评价指标体系的权重时采用主成分分析或赋值法，存在主观性，有效信息体现不全；第二，已有研究主要阐释了科技创新对经济高质量发展的驱动机制，而忽略了二者的耦合协调关系及各因素作用；第三，较多以全国范围为研究对象，以山东省16地市为研究对象的研究较缺乏。

本文拟做出以下边际贡献：第一，深入理解科技创新及高质量发展内涵，以“创新、协调、绿色、开放、共享”五大新发展理念为基础，结合山东省科技创新和经济发展特色，构建两

者评价指标体系；第二，使用熵权法对两者水平进行科学评价，利用耦合协调度和障碍度模型研究两者耦合协调水平及障碍因子的时空演化格局；第三，在评估省域及地市层面两者发展水平基础上，结合区域发展规划和目标定位，进一步对发展不平衡的关键矛盾进行深入研究，有助于为山东省及全国其他区域科技创新与经济高质量协调发展提供决策启示。

2 研究设计

2.1 研究方法

2.1.1 熵权法

为了实现不同年份的对比，借鉴已有研究中加入时间变量的熵权法，确定各评价指标权重并计算各年份科技创新与经济高质量发展水平[20]。具体计算步骤如下：

（1）为消除各指标量纲和单位的影响，需要对原始数据进行标准化处理，式（1）和式（2）分别为正向和负向指标标准化过程：

$$Y_{ijt}=\frac{x_{ijt}-minx_j}{maxx_j-minx_j}+A \quad (1)$$

$$Y_{ijt}=\frac{maxx_j-x_{ijt}}{maxx_j-minx_j}+A \quad (2)$$

式中，x_{ijt} 为第 t 年 i 地市第 j 个指标值；Y_{ijt} 为 x_{ijt} 标准化后的数据；$maxx_j$、$minx_j$ 分别表示研究时期内所有地市第 j 项指标原始数据最大值和最小值；i=1,2,...,n，j=1,2,...,m，t=1,2,...,r；同时由于对数运算中对底数的要求，需要进行平移处理，为降低平移对数据的影响，A 取值为 10^{-5}。

（2）计算各指标贡献度：

$$p_{ijt}=\frac{Y_{ijt}}{\sum_{t=1}^{r}\sum_{i=1}^{n}Y_{ijt}} \quad (3)$$

（3）计算各指标熵值：

$$e_j=-k\sum_{t=1}^{r}\sum_{i=1}^{n}p_{ijt}\ln p_{ijt} \quad (4)$$

其中，$k=\frac{1}{\ln rn}$。

（4）计算各指标差异系数：

$$g_j=1-e_j \quad (5)$$

（5）计算各指标权重：

$$w_j=\frac{g_j}{\sum_{j=1}^{m}g_j} \quad (6)$$

（6）计算各方案综合得分：

$$s_{ti}=\sum_{j=1}^{m}w_j\times Y_{ijt} \quad (7)$$

2.1.2 耦合协调度模型

本文用耦合协调度模型评价山东省科技创新和经济高质量发展之间的耦合协调程度，具体步骤如下：

（1）耦合度计算。耦合度可以衡量两系统间的相互作用程度[21]。耦合度计算公式如下：

$$C=2\frac{\sqrt{u_1u_2}}{u_1+u_2} \quad (8)$$

其中，C 是耦合度，取值范围为 (0, 1]，值越大说明系统间的耦合度越高，相互影响程度越明显；u_1 为科技创新发展水平；u_2 为经济高质量发展水平。

（2）耦合协调度计算。耦合度仅能表明各系统间相互作用的强弱，但无法区分系统是相互促进还是相互制约，因此，以耦合协调度进一步衡量两系统间的一致性程度[22]。耦合协调度计算公式如下：

$$D=\sqrt{C\left(\alpha u_1+\beta u_2\right)} \quad (9)$$

其中，D 是耦合协调度，取值范围为 (0, 1]，取值越大说明两系统耦合协调发展程度越好；α

doi:10.3772/j.issn.2095-915x.2024.03.002

和 β 分别为科技创新和经济高质量发展系统的权重，本文认为科技创新和经济高质量发展一样重要，因此取 $\alpha=\beta=0.5$。

《山东省区域科技创新能力评价报告》显示山东省16地市科技创新综合水平差距较大，研发经费支出中有5个地市超过100亿元，但有2个地市不到30亿元，每万名就业人员中研发人员数济南、青岛2市超过了100人年，仍有8市尚未达到50人年。近五年，科学技术支出仅占一般公共预算支出的3%左右，地区生产总值年均增长率近4%，高于研发经费支出年均增长率。参考已有研究[22-23]，结合山东省区域科技创新发展不平衡、科技投入与经济发展不匹配等现实问题，提出山东省科技创新与经济高质量发展协调度划分标准，如表1所示。

表1 科技创新和经济高质量发展耦合协调度划分标准

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

2.1.3 障碍度模型

为挖掘山东省科技创新与经济高质量发展耦合协调关系的影响因素，引入障碍度模型分析其耦合协调的障碍因素[24-25]。具体公式如下：

$$F_{ij}=\alpha_j w_{ij} \quad (10)$$

$$I_{ij}=(1-Y_{ij}) \quad (11)$$

$$H_{ij}=\frac{F_{ij}I_{ij}}{\sum_{j=1}^{m}F_{ij}I_{ij}}\times 100\% \quad (12)$$

其中，F_{ij} 为第 j 个子系统中第 i 个指标的因子贡献度；α_j 为第 j 个子系统的权重；w_{ij} 为第 j 个子系统中第 i 个指标的权重；I_{ij} 为指标偏离度，表示各指标与系统发展目标的差距；Y_{ij} 为经标准化后第 j 个子系统中第 i 个指标的值；H_{ij} 为第 j 个子系统中第 i 个指标对两者耦合协调的障碍度。

2.2 评价指标体系构建及数据来源

2.2.1 评价指标体系构建

进行科技创新研究的前提是构建科学、可比、系统、全面的科技创新水平评价指标体系。本文借鉴已有研究[13, 26-27]，以创新投入—环境—产出研究视角为基础，结合《中国区域创新能力评价报告》，考虑山东省国有企业体量大、经济贡献率高，并且山东省科技创新受到政府、高校和企业等创新主体共同作用，增加规模以上工业企业R&D经费占比、R&D经费中政府支出和企业支出比值、规模以上工业企业R&D人员占比等指标体现山东省科技创新特色，如表2所示。

经济高质量发展是对经济发展的结构、效率、稳定性和持续性的综合评价，本文从经济高质量发展内涵及“创新、协调、绿色、开放、共享”的新发展理念出发，借鉴已有研究[28-30]，结合山东省基础设施和资源禀赋情况，考虑山东省存在经济大而不强、传统产业占比大、经济发展与环境污染矛盾突出、创新能力与经

济规模不匹配等实际问题，选取着重反映经济稳定性和持续性的指标评价山东省经济高质量发展水平。同时，尽量设计比值、人均等指标，保证区域可比性，如表 3 所示。

表 2　科技创新水平评价指标体系

一级指标	二级指标	三级指标			
		名称	单位	指标属性	权重
科技创新	创新投入	R&D 经费投入强度（S1）	%	正向	7.19%
		科学技术支出占财政支出比重（S2）	%	正向	10.30%
		规模以上工业企业 R&D 经费支出占 R&D 经费总支出比重（S3）	%	正向	5.33%
		每万人 R&D 人员全时当量（S4）	人年 / 万人	正向	7.93%
	创新产出	每万人专利授权量（S5）	件 / 万人	正向	12.66%
		每万人专利受理量（S6）	件 / 万人	正向	9.03%
		每万人拥有科技论文数（S7）	篇 / 万人	正向	19.45%
	创新环境	职业学校师生比（S8）	%	正向	9.50%
		R&D 经费中政府支出与企业支出比值（S9）	%	正向	8.46%
		规模以上工业企业 R&D 人员数占就业人员比重（S10）	%	正向	10.15%

表 3　经济高质量发展水平评价指标体系

一级指标	二级指标	三级指标			
		名称	单位	指标属性	权重
经济高质量发展	协调发展	人均 GDP（E1）	元	正向	4.30%
		第三产业增加值占地区生产总值比重（E2）	%	正向	5.00%
		年末城镇登记失业率（E3）	%	负向	3.50%
		城乡居民人均可支配比值（E4）	%	负向	3.60%
		城乡居民人均消费支出比值（E5）	%	负向	3.00%
	绿色发展	建成区绿化覆盖率（E6）	%	正向	5.50%
		单位 GDP 废水排放量（E7）	吨 / 万元	负向	4.50%
		单位 GDP 二氧化硫排放量（E8）	吨 / 万元	负向	4.90%
		一般工业固体废物综合利用率（E9）	%	正向	4.50%
		环保支出占财政支出比重（E10）	%	正向	6.20%
	开放发展	进口依存度（E11）	%	正向	7.40%
		出口依存度（E12）	%	正向	7.90%
		外商投资新设企业数（E13）	个	正向	8.00%
		外商直接投资额占地区生产总值比重（E14）	%	正向	6.80%
		入境旅游外国人数占入境旅游人数比重（E15）	%	正向	3.50%
	共享发展	城镇职工基本养老保险覆盖率（E16）	%	正向	4.60%
		教育支出占财政支出比重（E17）	%	正向	4.30%
		社会保障与就业支出占财政支出比重（E18）	%	正向	4.50%
		每万人医疗机构床位数（E19）	张 / 万人	正向	4.30%
		移动电话普及率（E20）	%	正向	3.70%

doi:10.3772/j.issn.2095-915x.2024.03.002

2.2.2 数据来源

本文选取 2012—2021 年山东省 16 个城市作为研究对象，研究山东省科技创新和经济高质量发展的耦合协调关系及障碍因子。各城市指标数据来源于 2013—2022 年《山东统计年鉴》《山东科技统计年鉴》。对于个别缺失数据，采用插值法进行处理。

3 实证分析

3.1 科技创新与经济高质量发展的综合评价

3.1.1 全省科技创新与经济高质量发展水平评价

利用 2012—2021 年山东省 16 地市的相关数据，在通过熵权法得出科技创新和经济高质量发展体系各指标权重的基础上，采用综合加权法得到两系统综合指数代表山东省整体水平，如图 1 所示。

从整体上看，山东省科技创新和经济高质量两系统的综合指数在 2012—2021 年基本呈现增长趋势。2012 年两系统水平相差较大，皆处于较低水平，之后，科技创新水平均值从 2012 年的 0.194 升至 2021 年的 0.320，增长了 64.95%，提升幅度大于经济高质量发展水平，表明山东省加快构建科技创新体系，科技创新实力显著提升。经济高质量发展水平均值从 2012 年的 0.251 上升到 0.353，增长了 40.64%，表明山东省经济高质量发展态势良好。

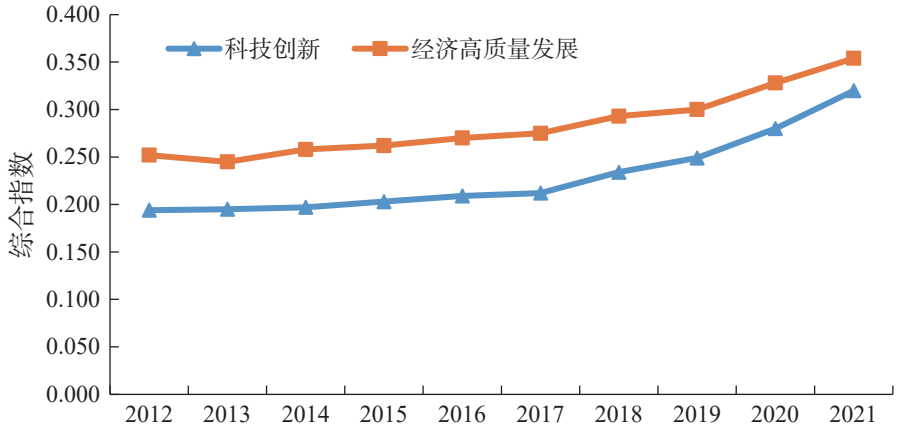

图 1 山东省科技创新与经济高质量发展综合指数变化趋势

3.1.2 区域科技创新发展水平评价

山东省 16 地市科技创新发展水平呈波动上升趋势，地区间差异显著，各城市之间发展不均衡。一方面，青岛、济南和烟台 3 市的科技创新水平较为突出，科技创新水平均值分别为 0.644、0.631 和 0.303。其中，青岛的科技创新实力遥遥领先，由 2012 年的 0.516 升至 2021 年的 0.832，增长了 61.24%，说明青岛作为副省级沿海城市，是海洋科技创新、人才集聚和技术转移转化高地；济南的科技创新水平由 2012 年的 0.653 升至 2021 年的 0.748，说明济南作为省会城市，在研发投入、成果转化和创新环境方面领先其他省内城市；烟台的进步幅度最为明显，科技创新水平由 0.235 升至 0.451，增长近 1 倍，说明烟台全市科技创新综合实力实现新突破。

另一方面，威海、淄博和东营 3 市的科技创新的水平均值较为接近，分别为 0.265、0.250 和 0.240，低于上述城市，但高于省内其他城市，三个城市科技创新水平分别增长了 82.41%、82.76% 和 50.45%，说明 3 市紧紧围绕产业链部署创新链，科技创新水平提升较有成效。而山东省内某些区域存在科技创新“洼地”，如菏泽、枣庄和聊城，这些城市科技创新水平均值低于 0.1。菏泽、枣庄和聊城科技资源欠缺、研发投入不足，应进一步加强与核心城市的联动发展，加快培育高新技术企业，提升关键技术研发能力。

3.1.3 区域经济高质量发展水平评价

山东省 16 地市经济高质量发展水平持续性稳定提升，副省级城市及沿海城市经济高质量发展水平综合实力较强。一方面，青岛、济南、烟台、日照、威海的经济高质量发展水平均值在 0.3 以上，高于省内其他城市，经济发展态势良好。一是在协调发展方面，青岛和济南分别作为胶东和省会经济圈的中心城市，经济体量和产业发展存在相对优势，经济协同发展程度高于其他地区；烟台、日照和威海 3 市加快特色产业优化升级，全面提升城乡统筹水平。二是在绿色发展方面，青岛、济南将“双碳”工作作为推动经济社会发展全面绿色转型的总牵引；烟台、日照和威海 3 市加快构建新型能源体系，节能环保工作成效显著。三是在开放和共享发展方面，青岛、济南、烟台、日照和威海 5 市是黄河流域开放门户，进出口依存度、医疗机构床位数等方面优于其他地区。

另一方面，潍坊、淄博和东营 3 市的经济高质量发展水平略低于上述城市，但也优于省内其他城市，3 个城市的经济高质量发展水平分别增长了 53.72%、32.79 和 43.27%，说明 3 市社会经济基础较为优越，满足了人民日益增长的美好生活需要。而菏泽和德州两市经济高质量发展水平均值不足 0.2，处于失衡状态。其中，菏泽是“鲁西崛起”的核心阵地，产业发展、生态环境等方面与省内其他城市差距大，应抓住黄河流域生态保护和高质量发展这一机遇，重点谋划打造特色产业体系；德州市农业人口众多，第三产业发展落后，应充分利用位于华北、华东经济区连接带位置，推动物流产业发展，释放经济发展活力。

3.2 耦合协调的时空分异特征分析

3.2.1 时间维度：全省耦合协调水平分析

基于上文介绍的耦合协调度模型测算山东省 2012—2021 年科技创新与经济高质量发展的耦合度和耦合协调度。如图 2 所示，2012—2021 年山东省科技创新与经济高质量发展的耦合度较为平稳，趋近于 1，说明两者相互联系紧密；耦合协调度呈持续上升趋势，由 2012 年的 0.494 升至 2021 年的 0.641，从濒临失调到初级协调状态，说明两者相互促进、协同发展，但是科技创新发展进程较慢，两系统整体未达到高水平协同状态。

3.2.2 空间维度：区域耦合协调水平分析

根据上文介绍的耦合协调度模型和划分依据测算山东省 16 地市 2012—2021 年科技创新与经济高质量发展的耦合协调度和所属耦合协调区域，以各地市 2012—2016 年耦合协调度均值表征研究期内各地市耦合协调的平均水平，

doi:10.3772/j.issn.2095-915x.2024.03.002

以 16 地市耦合协调度均值表征该年山东省耦合协调的整体水平 [21]。利用 ArcGIS 10.2 软件绘制主要年份科技创新与经济高质量发展耦合协调类型时空分布图，如图 3 所示。

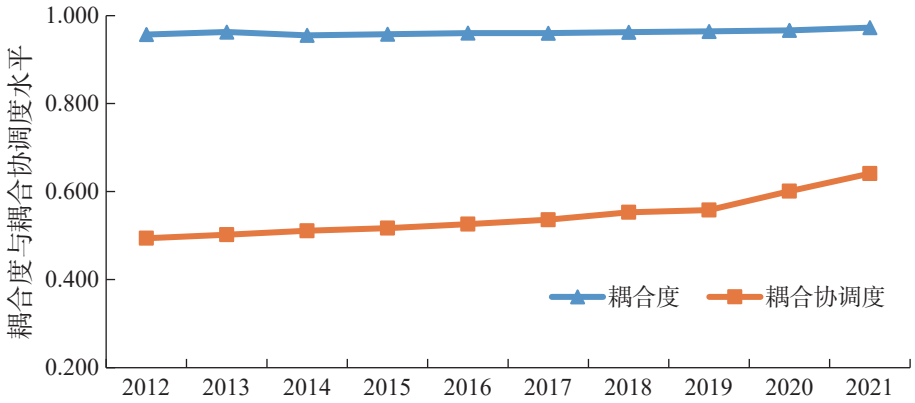

图 2　山东省科技创新与经济高质量发展的耦合协调水平变化趋势

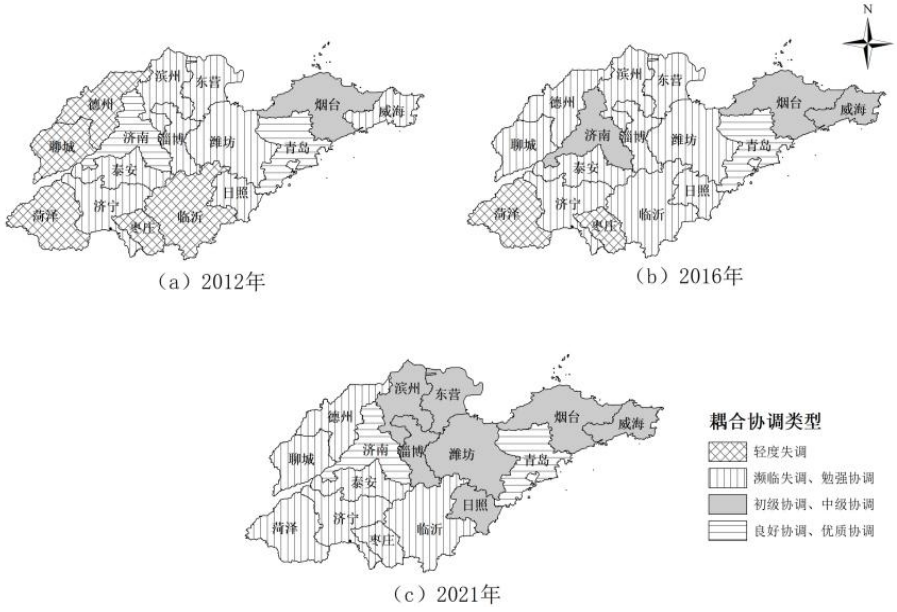

图 3　山东省科技创新与经济高质量发展耦合协调类型的分布格局

（1）优质和良好协调区。2012 年良好协调区城市为青岛和济南，2021 年济南仍旧为良好协调区，青岛演化到优质协调区。该类型城市一直以来都是经济圈核心城市，也是国家中心城市。青岛建设绿色低碳高质量发展先行区，在国家创新型城市排名中位列第十；济南市作为省会城市，聚集着省内优质科技资源，为高质量发展提供了源源不断的动能。该类型城市要发挥辐射带动作用，利用“溢出效应”将优质技术、人才、资源向周边地区扩散，促进区域一体化发展。

（2）中级和初级协调区。2012 年该类型城市只有烟台，2016 年增加到烟台、威海 2 个城市，2021 年达到 7 个城市，占山东省城市数

量的 43.75%。该类型城市多集中在沿海和核心城市周边，空间分布格局由分散到集中连片。烟台、威海、日照等城市均为沿海城市，建设现代海洋体系，科技力量全面起势，赋能经济高质量发展。淄博毗邻济南，受到济南的创新溢出效应影响。该类型城市科技创新与经济高质量发展协同程度稳步上升，但与高水平协调区还有一定差距，要发挥区域特色产业优势，打造区域协同创新共同体。

（3）勉强协调和濒临失调区。该类型城市数量占比较高，2012 年、2016 年、2021 年占比分别为 50%、62.5%、43.75%，且多分布于鲁西南地区。长期位于该区域的德州、聊城、泰安、济宁、临沂等城市实施产学研协同创新，高附加值的新兴产业成为高质量发展的重要支撑。同时，这些城市基础设施较为落后，受到济南虹吸效应的影响，高质量发展相对滞后。该类型城市一方面要通过深化科技体制改革，吸引人才和资本流入；另一方面要充分利用省会城市和沿海城市的辐射带动能力，学习先进经验和技术，推进各类载体平台共建共享。

（4）轻度失调区。2012 年轻度失调区城市有 5 个，2016 年下降为 2 个，到 2021 年该类型城市数量趋于零，表明各地市耦合协调度的空间差异减小，全省科技创新与经济高质量发展逐渐协同。长期位于该区域的菏泽、枣庄 2021 年人均 GDP 分别为全省倒数第一、第四，R&D 经费支出占比为全省倒数后两位，发展质量和效益低下。该类型城市政府应给予政策扶持，大力培育创新主体，集聚创新资源力量，以高水平创新推动高质量发展。

3.3 耦合协调度的主要障碍因素分析

3.3.1 时间维度：全省耦合协调发展障碍因素分析

运用障碍度模型分析影响山东省科技创新与经济高质量发展耦合协调度的障碍因子，如图 4 所示。按照障碍度均值大小排序：创新产出（25.379%）> 开放发展（19.756%）> 创新投入（14.866%）> 创新环境（13.939%）> 绿色发展（9.226%）> 共享发展（8.822%）> 协调发展（8.013%）。

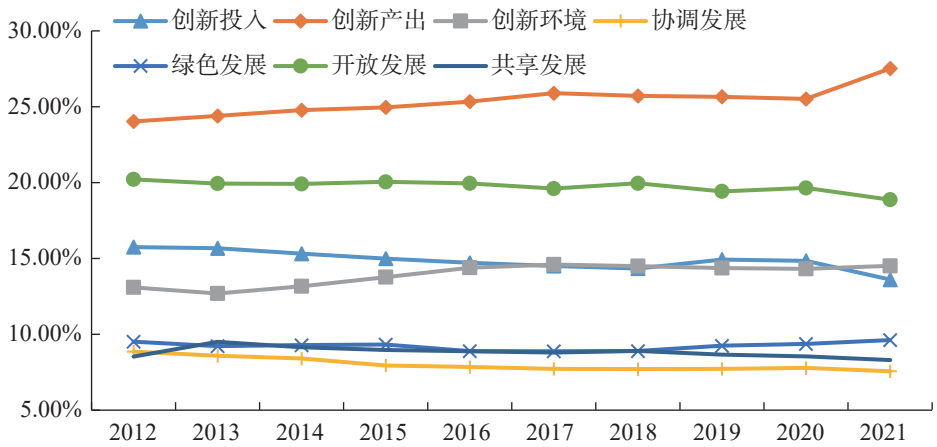

图 4　山东省科技创新与经济高质量发展障碍因素变化趋势

doi:10.3772/j.issn.2095-915x.2024.03.002

在科技创新系统中，创新产出障碍度呈上升趋势，创新投入障碍度呈下降趋势，表明创新投入和产出呈现出不协调的状态，产业与资源要素融合度低，技术创新扩散体系不完善，导致创新成果转化效率低下。在经济高质量发展系统中，开放发展是首要障碍因子，每年障碍度均超过 20%，进出口、外商投资等问题关系着对外贸易，反映了山东省在对外开放上的短板，未来应扩大高质量招商引资，深度融入“一带一路”建设。绿色发展是第二大障碍因素，山东省某些地区产业结构不合理和资源浪费对环境质量影响大，应推动传统产业提质升级，培育绿色低碳产业，有序实施碳达峰碳中和。共享发展和协调发展障碍度最低，且呈下降趋势，但增进民生福祉、促进区域协调发展仍应引起重视。

3.3.2 空间维度：区域耦合协调发展障碍因素分析

根据 2012—2021 年各地市的耦合协调度均值将城市分为五类：良好协调区城市为济南和青岛，初级协调区城市为烟台和威海，勉强协调区城市为淄博、东营、潍坊、日照和滨州，濒临失调区城市为济宁、泰安、临沂、德州和聊城，轻度失调区城市为枣庄和菏泽。通过进一步明确五类城市影响科技创新与经济高质量发展耦合协调度的主要障碍因素，筛选出前 5 位的障碍因子，如表 4 所示。

表 4 区域耦合协调类型主要障碍因素及障碍度

相对发展类型		良好协调区	初级协调区	勉强协调区	濒临失调区	轻度失调区
第 1 位	障碍因子	E11	E13	E13	E13	E13
	障碍度（%）	16.327	19.474	20.333	18.651	18.558
第 2 位	障碍因子	E13	S7	S7	S7	S7
	障碍度（%）	12.920	17.465	16.809	15.312	14.726
第 3 位	障碍因子	S5	E11	E11	E11	E11
	障碍度（%）	8.463	14.548	10.959	12.020	11.796
第 4 位	障碍因子	E12	E14	E12	E12	E12
	障碍度（%）	7.074	5.735	7.001	7.799	7.932
第 5 位	障碍因子	S5	E6	E14	E14	E14
	障碍度（%）	7.041	5.364	6.714	6.790	6.900

这五类城市共有的障碍因子包括外商投资新设企业数、进口依存度、论文数，表明对外贸易和科技成果转化是山东省各地市科技创新与经济高质量耦合协调发展的关键制约因素。良好协调区城市的主要障碍因子包括进出口依存度、外商投资新设企业数、论文数、专利授权量，因此构建良好创新生态、加快构建国际物流大通道是该类城市发展的重要任务。初级协调区城市的主要障碍因子包括外商投资新设企业数、进口依存度、外商直接投资额占比、论文数和建成区绿化覆盖率，因此该类城市要提升贸易便利化水平、优化知识产权服务、改善居住环境。勉强协调区、濒临失调区和轻度失调区城市的主要障碍因子都是外商投资新设企业数、论文数、进出口依存度和外商直接投资额占比。因此应加大对外贸易政策性改革、

促进产学研协同创新。

4 结论及建议

4.1 研究结论

本文以 2012—2021 年山东省 16 地市为研究对象，构建科技创新与经济高质量发展指标体系，利用熵权法、耦合协调度模型和障碍度模型分析两者时空耦合关系及障碍因素。主要结论如下：

（1）山东省科技创新与经济高质量发展水平持续提升，科技创新水平均值由 0.194 升至 0.320，经济高质量发展水平均值由 0.252 升至 0.354，科技创新水平增速高于经济高质量发展水平，两个系统均呈现良好的发展趋势。但山东省内部发展不均衡，济南、青岛和烟台科技创新和经济高质量发展水平较高，菏泽、枣庄和德州在科技创新和经济高质量发展领域实力相对薄弱。

（2）山东省科技创新与经济高质量发展耦合协调度增长稳定，由 0.494 升至 0.641，从濒临失调过渡到初级协调状态，说明两系统已实现初步协调，但距离高水平协调阶段还有差距。在空间演化上，优质和良好协调区城市一直为 2 个，中级和初级协调区城市由 2 个增加到 7 个，勉强和濒临协调区城市基本维持在 7 个，轻度失调区城市由 5 个下降为 2 个，说明各城市耦合协调度呈上升趋势，城市之间差异减小，区域一体化发展成效显著。

（3）山东省科技创新与经济高质量发展水平耦合协调的障碍因素依次为创新产出 > 开放发展 > 创新投入 > 创新环境 > 绿色发展 > 共享发展 > 协调发展，说明山东省在创新成果产出和对外发展上存在短板，后续要提高创新转化效率，促进外贸稳中提质。从具体障碍指标看，外商投资新设企业数、进口依存度、论文数是两者的关键阻碍因素。

4.2 政策建议

（1）优质和良好协调区，要加快科技成果转化，推动区域协调发展。该类城市科技创新、经济高质量发展水平及两者耦合程度均处于全省领先地位，障碍因子中对外发展和创新产出成为制约发展的主要因素。一是要完善科技成果转化服务体系，建设成果转化专业化服务机构；二是要充分发挥投资、外贸对经济发展的关键支撑作用，助力外贸优势企业保订单、稳份额、拓市场；三是要发挥核心城市的辐射带动作用，借助“一群两心三圈”格局，促进区域一体化发展。

（2）中级和初级协调区，要推动特色产业高质量发展，打造绿色发展示范区。该类城市基本为沿海城市，科技创新与经济高质量发展耦合协调度较好，但开放发展、创新产出和绿色发展成为遏制当地发展的瓶颈。一是要不断完善全生命周期招商推进机制，在货物贸易、原产地规则、服务贸易等方面扩大开放；二是要充分发挥海洋经济产业优势，建设高能级重点实验室等创新平台；三是要把绿水青山就是金山银山的理念落实在行动上，大力实施绿色低碳技术国际合作等转型发展举措。

（3）勉强和濒临协调区，要打通内外贸易，构建双循环。该类城市多分布于鲁西南地区，

doi:10.3772/j.issn.2095-915x.2024.03.002

科技创新与经济高质量发展耦合协调程度处于过渡阶段，开放发展和创新产出是两者耦合协调的关键阻碍因素。一是要聚焦制约科技创新的体制机制瓶颈，深入推进科技改革攻坚行动，释放全社会的创新活力，为各类人才施展才华搭建载体平台，提升技术研发和成果转化能力；二是要营造国内一流营商环境，融入长三角、对接粤港澳，突出龙头企业作用，深度融入国内国际双循环。

（4）轻度失调区，要加强区域交流合作，实施差异化发展战略。该类城市多是粗放型经济增长方式的典型代表，科技创新与经济高质量发展呈现失调状态，开放发展和创新产出是两者协调发展的首要障碍因素。一是要继续加大政策倾斜力度，鼓励经济发达城市与该类地区开展技术指导、人才交流等活动，同时深度融入 RCEP、黄河流域生态保护和高质量发展等重大国家战略，深化和拓展国际友好合作关系；二是要因地施策，改造传统老工业基地，推动粗放型经济向集约型转型。

参 考 文 献

[1] 谢泗薪 , 胡伟 . 经济高质量发展与科技创新耦合协调：以京津冀地区为例 [J]. 统计与决策 , 2021, 37(14): 93-96.

[2] 刘亚迪 , 刘晓慧 , 董艳蕊 . 基于集对分析的副省级城市科技创新评价模型及实证 [J]. 科研管理 , 2018, 39(S1): 100-110.

[3] 许红丹 , 杨武 . 中国科技创新驱动经济高质量增长时序性分析——基于 TVP-VAR 模型的实证研究 [J]. 科学学研究 , 2023, 41(11): 2073-2085.

[4] 蔡文伯 , 陈念念 . 长江经济带城市群高等教育、科技创新和经济增长的耦合协调效应 [J]. 现代教育管理 , 2022(11): 33-42.

[5] 谷斌 , 廖丽芳 . 新基建投入与科技创新能力耦合协调发展水平测度及时空演进 [J]. 科技进步与对策 , 2023, 40(11): 60-70.

[6] 张凯 , 张明慧 . 基于云模型和证据理论的科技创新与持续发展能力评价 [J]. 运筹与管理 , 2022, 31(4): 109-115.

[7] MLACHILA M, TAPSOBA R, TAPSOBA S J A. A quality of growth index for developing countries: a proposal[J]. Social Indicators Research, 2017, 134(2): 675-710.

[8] 袁晓玲 , 王军 , 张江洋 . 中国省域经济高质量发展水平评价与比较研究 [J]. 经济与管理研究 , 2022, 43(4): 3-14.

[9] 李桂君 , 田宗博 , 宋砚秋 . 地方政府债务对城市经济高质量发展的影响研究 [J]. 中央财经大学学报 , 2023(1): 3-14.

[10] 单勤琴 , 李中 . 经济高质量发展水平的地区差异及时空收敛性研究 [J]. 经济地理 , 2022, 42(9): 50-58.

[11] MCCLAIN N. Caught inside the black box: Criminalization, opaque technology, and the New York subway MetroCard[J]. The Information Society, 2019, 35(5): 251-271.

[12] 任晓刚 , 刘菲 . 坚持科技创新推动经济高质量发展 [J]. 人民论坛·学术前沿 , 2022(13): 101-104.

[13] 姚振飞 , 张学波 . 山东省地区创新能力与经济质量的耦合协调及影响因素 [J]. 资源开发与市场 , 2021, 37(10): 1192-1199.

[14] 祝恩元 , 李俊莉 , 刘兆德 , 等 . 山东省科技创新与可持续发展耦合度空间差异分析 [J]. 地域研究与开发 , 2018, 37(6): 23-28.

[15] 董弋萱 . 中国城市创新与经济高质量发展耦合协调研究 [J]. 技术经济与管理研究 , 2022(7): 42-47.

[16] 魏巍 , 符洋 , 杨彩凤 . 科技创新与经济高质量发展测度研究——基于耦合协调度模型 [J]. 中国科技论坛 , 2020(10): 76-83.

[17] 刘建华 , 王慧扬 . 黄河流域科技创新与低碳经济耦合协调及障碍因子研究 [J]. 人民黄河 , 2023, 45(1): 6-12.

[18] 尚英仕 , 刘曙光 . 中国东部沿海三大城市群的科技

创新与绿色发展耦合协调关系 [J]. 科技管理研究 , 2021, 41(14): 46-55.

[19] 陈正其 . 京津冀创新要素配置与经济高质量耦合发展研究 [J]. 工业技术经济 , 2022, 41(11): 113-122.

[20] 周筱扬 , 左国存 . 我国中部地区科技创新与经济高质量发展耦合协调度的时空演化 [J]. 科技管理研究 , 2022, 42(22): 77-85.

[21] 魏奇锋 , 徐霞 , 杨彩琳 , 等 . 成渝地区双城经济圈科技创新与经济高质量发展耦合协调度研究 [J]. 科技进步与对策 , 2021, 38(14): 54-61.

[22] 李素峰 , 冯鸿雁 . 科技创新与数字经济高质量发展动态耦合及时空分异——基于中国四大战略区域视角 [J]. 当代经济管理 , 2023, 45(6): 41-50.

[23] 华坚 , 胡金昕 . 中国区域科技创新与经济高质量发展耦合关系评价 [J]. 科技进步与对策 , 2019, 36(8): 19-27.

[24] 刘建华 , 王慧扬 . 黄河流域科技创新与低碳经济耦合协调及障碍因子研究 [J]. 人民黄河 , 2023, 45(1): 6-12.

[25] 张旭 , 袁旭梅 , 魏福丽 . 县域经济高质量发展内部耦合协调水平评价与障碍因子诊断——以国家级创新型县 (市) 为例 [J]. 统计与信息论坛 , 2020, 35(2): 59-67.

[26] 郭爱君 , 杨春林 , 钟方雷 . 我国区域科技创新与生态环境优化耦合协调的时空格局及驱动因素分析 [J]. 科技管理研究 , 2020, 40(24): 91-102.

[27] 贾春光 , 程钧谟 , 谭晓宇 . 山东省区域科技创新能力动态评价及空间差异分析 [J]. 科技管理研究 , 2020, 40(2): 106-114.

[28] 任保平 , 付雅梅 , 杨羽宸 . 黄河流域九省区经济高质量发展的评价及路径选择 [J]. 统计与信息论坛 , 2022, 37(1): 89-99.

[29] 金昌东 , 张宝雷 , 康洁铭 . 经济高质量发展水平及其耦合协调性研究——基于山东省 17 地市面板数据 [J]. 生态经济 , 2021, 37(7): 65-72.

[30] 屈小娥 , 马黄龙 , 王晓芳 . 省域经济高质量发展水平综合评价 [J]. 统计与决策 , 2022, 38(16): 98-103.

doi:10.3772/j.issn.2095-915x.2024.03.003

创新全过程视角下产学研创新团队潜在合作对象识别方法研究

开放科学（资源服务）标识码（OSID）

张伟然[1] 王超[2] 许海云[2] 武华维[3] 李欣[2]

1. 齐鲁工业大学（山东省科学院）科技发展战略研究所 济南 250000；
2. 山东理工大学管理学院 淄博 255000；
3. 西北师范大学档案馆 兰州 730070

摘要：[目的 / 意义] 识别领域不同创新环节中的潜在合作对象，形成贯穿基础研究、应用研究与产业应用的多环节与多主体的产学研创新协同团队，助力科技创新及成果转移转化。[方法 / 过程] 基于创新全过程视角，从产业机构出发，构建产业—技术—科学机构的多层网络分析模型，识别不同创新层中产学研机构的潜在合作对象。以中国干细胞研究领域为对象进行实证研究，验证了方法的可行性与有效性。[结果 / 结论] 创新全过程视角下潜在合作对象识别方法突破了单一视角识别合作对象的局限性，实现了不同创新环节间创新主体潜在合作对象的识别，为创新生态下各主体识别与自身需求更匹配的合作伙伴提供了理论支撑与实践启示。

关键词：合作对象；创新过程；创新生态系统；多层网络；产学研

中图分类号：G353

Researching the Methodology for Identifying Domain Potential Collaborative Entities from the Perspective of Innovation Process

ZHANG Weiran[1] WANG Chao[2] XU Haiyun[2] WU Huawei[3] LI Xin[2]

1. Institute of Science and Technology Development Strategy, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250000, China;
2. School of Management, Shandong University of Technology, Zibo 255000, China;
3. Archives of Northwest Normal University, Lanzhou 730070, China

Abstract: [Purpose/Significance] This study aims to facilitate scientific and technological innovation and the transfer of

基金项目 国家自然科学基金项目“基于弱信号时效网络演化分析的变革性科技创新主题早期识别方法研究”（72274113）；山东省自然科学基金项目“基于弱信号分析的变革性创新主题早期识别方法研究”（ZR202111130115）。

作者简介 张伟然（1998-），硕士研究生，主要研究方向为数据管理与决策分析；王超（1988-），博士，副教授，主要研究方向为科技与产业情报分析，E-mail: kingtaoist@yeah.net；许海云（1982-），博士，教授，主要研究方向为科学计量学；武华维（1985-），博士，副研究馆员，主要研究方向为知识图谱与知识计算；李欣（2000-），硕士研究生，主要研究方向为科学计量学。

引用格式 张伟然 , 王超 , 许海云 , 等 . 创新全过程视角下产学研创新团队潜在合作对象识别方法研究 [J]. 情报工程 , 2024, 10(3): 28-48.

achievements by establishing a collaborative ecosystem that spans basic research, applied research, and industrial applications within the industry-university-research framework. The primary objective is to identify potential partners across various stages of innovation. [Methods/Processes] Employing a holistic innovation process approach, this research constructs a multi-layer network analysis model encompassing industry, technology and scientific institutions. The model facilitates the identification of potential partners within the industry-university-research nexus across diverse innovation tiers. To validate the feasibility and efficacy of the proposed method, an empirical study centered on stem cell research in China is conducted. [Results/Conclusions] The innovative approach of identifying potential cooperative partners across the entire innovation process surpasses the limitations of one-dimensional analyses, enabling the recognition of potential partners amidst different innovation stages. The resulting partnerships, rooted in a deep understanding of collaborative needs, offer both theoretical and practical support for advancing scientific and technological innovation.

Keyword: Cooperation Partners; Innovation Process; Innovation Ecosystems; Multi-layer Network; Industry-Academia-Research

引言

在国家全新发展格局引领下，产业创新发展面临新形势。为全面提升产业技术创新能力，实现中国产业科技自立，科技部提出要聚焦产业创新关键环节，面向产业需求加强产学研用和大中小企业融通创新[1]。习近平总书记强调我国要加强创新链和产业链对接，关键是要确保产业占据创新主体地位，围绕产业链部署创新链，以消除科技创新的“孤岛现象”，围绕创新链布局产业链，为产业提供创新驱动力，实现高水平自立自强[2-3]。产学研协同创新是提升企业自主创新能力、加速科技成果转移转化的重要路径。因此，围绕产业机构创新需求，构建以基础研究为依托，以技术供给为核心，基于基础研究到应用研究再到产业应用的创新全过程视角下潜在合作对象的识别方法，对实现产学研主体深度融合创新、创新链中资源的流通、激发各创新主体活力、加速产业升级转型具有重要现实意义。

当前，关于潜在合作对象识别的相关研究众多，但多数研究局限于单一创新环节内，如基于基础研究范围内以论文为信息载体的潜在合作对象识别研究[4-5]，基于应用研究范围内以专利为信息载体的潜在合作对象识别研究[6-7]，但基于多创新环节的潜在合作对象识别更符合当前国家协同创新战略，更具有现实应用价值。本文在创新全过程视角下，以具有合作创新需求的产业机构为服务对象，围绕产业机构创新需求，综合考虑多种合作影响因素，构建了基于多层网络分析模型展开创新全过程下潜在合作对象识别与选择的方法框架，使创新主体能够高效、准确地定位创新全过程视角下的潜在合作对象，为形成产业创新需求与知识技术供给良好互动的创新团队提供了思路。最后以干细胞领域为例进行实证研究，验证了本文提出方法的可行性与有效性。

1 相关研究

1.1 创新过程理论研究进展

经济全球化和信息技术的发展加速了知识技术的交叉渗透，创新活动日益复杂，创新过

doi:10.3772/j.issn.2095-915x.2024.03.003

程理论不断革新。自 20 世纪以来，学者们对创新过程进行了大量研究，其中三段式创新过程理论和创新生态系统理论最具代表性。

三段式创新过程理论将创新过程划分为三个阶段，如 Timmers[8] 将创新过程分为“基础研究—技术应用研究—开发应用部署”三个阶段，Turkenbur[9] 提出了“研究—示范—扩散”三段创新过程理论。创新生态系统是创新主体与环境相互作用构成的开放性复杂系统，旨在实现创新过程中的知识流动与经济价值转换 [10-11]。当前主要聚焦于企业、产业、区域和国家等维度的创新生态系统研究 [12-13]。其中，企业创新生态系统强调围绕核心企业而相互依存的创新主体所形成的复杂关系 [14-15]。Moore[16] 将企业创新生态系统分为诞生、扩张和领导三个阶段。李军凯等 [17] 认为产业创新生态系统包括理论构想—知识技术孕育孵化—技术产业化—产业规模化等阶段。

综上，无论是三段式创新过程理论，还是创新生态系统理论，均强调创新主体间的合作以及创新过程中知识价值到经济价值的转化。实际上，创新全过程的基本内核是科学知识、技术转化为实际产业应用的过程。

1.2 基于核心机构的潜在合作对象识别

核心机构是指在科研中影响力较大，取得创新成果较多的机构。它们在合作网络中展现出较强的资源配置与信息传递能力，具备高质量的创新资源和最先进的知识技术体系，是创新活动的中坚力量 [18-20]。与其合作建立具有核心竞争力与控制优势的创新团队，能够提高资源搜索与知识转移效率，创造更稳定的创新产出环境，并吸引更多优质合作者 [21]。

社会网络分析法通过可视化图形展现合作关系与合作网络结构，在分析主体合作关系、识别关键合作对象等方面表现出色 [22-23]。王超等 [24] 利用中介中心性指标在合作网络中识别潜在关键核心机构；王谦等 [25] 结合普赖斯定律与社会网络分析法识别儿科人群药物临床试验领域的核心牵头机构；李纲等 [26] 通过凝聚子群分析和滚雪球等方法分析合著网络、共词网络，识别核心团队的研究主题。核心机构在其领域内具有不可替代性 [27]，筛选创新全过程中各创新环节的核心机构强强联合，能形成优势互补、引领协同创新的高水平创新团队。

1.3 基于研究内容相似性的潜在合作对象识别

知识技术的相似性是影响合作对象选择的关键因素 [28-30]。核心机构的研究内容对其领域创新活动具有引领作用 [31]。具有相似研究内容的机构合作能够促进合作网络内知识融合与创新，更易形成高效的合作创新关系 [32]。

基于研究内容的对象识别方法仅考虑研究内容而不涉及社会关系，能够实现不连通网络间潜在合作对象的识别 [33]。其基本思路是：从各主体所发表成果的文本内容出发，运用文本挖掘和自然语言处理等方法计算研究内容相似度，根据相似度识别潜在合作对象。丁敬达等 [4] 融合关键词耦合分析法和链路预测法识别潜在合作关系。Chuan 等 [34] 使用 LDA 模型研究不同个体研究内容相似性，但该模型存在工作量大以及耗时长等问题。陈卫静等 [35] 通过 TF-IDF 关键词加权方式计算作者间研究内容相似

度，根据相似度得分挖掘作者间潜在合作关系。但此方法只考虑词的频率而未考虑语义，结果准确率有待商榷。Word2vec 模型在自然语言处理中表现良好，但忽略了词汇语序对文本内容的影响。为完善这一问题，Le 等 [36] 提出了 Doc2Vec 文档向量模型，文本相似度计算更为精确。

1.4 基于地理距离的潜在合作对象识别

隐性知识具有较强的本地依赖性 [37]，尽管当前通信与交通十分便捷，但由于隐性知识的存在，知识溢出仍具有本地化特征。因此，合作对象的选择有必要考虑地理邻近性。Katz 等 [38] 研究了地理位置等因素对合作的影响，发现主体合作可能性随地理距离的增大而减小。选择邻近机构作为合作对象不仅能降低合作风险和成本，还有利于各方知识技术的转移 [39-40]。贺超城等 [41] 发现尤其是在生物学领域，复杂的实验设备等使主体间的合作表现出地理聚集倾向。地理距离近的机构在相似的制度环境下更易进行密集的沟通，有助于合作效率的提升和合作双方信任的建立 [32,42]。

综上，由于指标融合困难或数据源单一等原因，现有潜在合作对象识别方法多基于单一角度或指标，且局限于单一创新活动范围内。例如仅考虑合作网络关系而忽略主体的研究内容关系与地理距离 [43]、仅考虑研究内容相似性却未考虑合作网络关系及地理距离 [44]、考虑合作网络关系与主题相关性却忽略主体距离 [45]。现实中的合作关系建立受多种因素影响，基于单一指标或方法识别合作对象存在局限性，综合各类方法进行潜在合作对象的识别更具有现实应用价值。

鉴于此，本文尝试在创新全过程视角下结合多个指标识别不同创新环节中机构的潜在合作对象。一方面，立足于创新活动全过程，旨在识别包含科学、技术、产业多元化主体，囊括基础研究、应用研究与产业应用多环节，构建科学、技术、产业上下贯通发展的创新团队；另一方面，综合各类指标与方法，对潜在合作对象层层考察与挖掘，以识别出最合适的潜在合作对象。

2 创新全过程理论模型研究

2.1 创新全过程概述

各领域内完整的创新活动不仅需要发散思维的科学基础研究，也需要解决技术问题的创新应用研究，更离不开将科技成果转化为具有现实应用价值的产品或服务的产业应用研究。总体来看，创新包括从理论到应用、从应用到产业的全过程，其核心在于科学知识、技术的流动，以及其产业价值的实现 [46]。由此可见，科学、技术、产业是实现高质量创新的三个关键要素 [47]。本文将创新全过程定义为涵盖从问题假设到理论知识、从理论知识到技术专利、从技术专利到最终应用于产业的一系列创新目标集合的三段式创新过程。根据各主体的创新目标，将创新全过程划分为基础研究、应用研究与产业应用三个创新环节。根据创新环节中的关键要素，使不同的创新环节对应不同的创新层，不同的创新层中又包含不同类型的创新主体，如表 1 所示。

doi:10.3772/j.issn.2095-915x.2024.03.003

表 1 创新全过程

创新环节	创新目标	创新层	创新主体
基础研究	科学问题—理论知识	科学层（S）	高校、科研院所
应用研究	理论知识—专利技术	技术层（T）	高校、科研院所、企业
产业应用	科技成果—产品或服务	产业层（I）	以企业为主

2.2 创新过程中的知识载体

本文以不同创新层的机构为研究对象，科技论文、专利文献和以产品说明书为主的产品数据分别作为科学层机构、技术层机构、产业层机构的创新内容表现载体，如表 2 所示。

表 2 创新过程中的知识载体

创新环节	基础研究	应用研究	产业应用
创新内容表现载体	科技论文	专利文献	以产品说明书为主的产品数据
数据来源	论文数据库	专利数据库	产品市场数据统计网站

2.3 基于创新全过程的多层网络分析模型

在创新全过程中，不同创新层的主体有不同的创新目标。只有识别贯穿不同创新层的潜在合作对象，形成立足于创新全过程的合作团队，才能实现各创新环节的串联与创新目标的上下衔接。韩菁等 [48] 基于多层网络分析模型，采用链路预测法识别不同领域间的潜在合作对象；刘晓燕等 [49] 通过多层网络探索技术融合机理，为不同类型创新机构合作提供了参考。多层网络模型为本研究基于创新全过程视角下产学研创新机构潜在合作对象的识别提供了思路，有助于展示处于不同创新环节机构间的潜在合作关系识别过程 [50]。不同创新层的主体合作会形成复杂的合作关系，因此，本文借鉴多层网络分析模型，利用多个指标构建涉及产业、技术、科学三个创新层机构的多层合作网络，展现创新全过程视角下潜在合作对象的识别过程。

产业层机构作为创新过程中知识产出的需求方，在科技成果转化中起关键作用。目前中国科技创新成果大多集中于科学层和技术层机构，易导致科技成果与产业需求不匹配以及产业机构创新活力的下降，各主体无法有效对接与合作，科技成果转化困难，阻碍创新活动进程。只有以产业机构的发展趋势和创新需求为导向，围绕产业进行科技创新，才能创造出符合市场需求的创新成果 [51]。提高产业创新绩效需要与应用研究成果形成良好对接，提高技术吸收与转化能力，而应用研究又以基础研究所产生的科学知识积累为支撑来解决现实问题 [52]。因此，本文基于创新全过程视角，综合考虑多种合作影响因素，按照产业—技术—科学（I-T-S）的顺序构建多层合作网络（图 1），展现贯穿产业应用、应用研究、基础研究的创新全过程下的潜在合作对象识别过程。其中，层内连线代表创新层内现有合作关系，层间连线代表不同创新层之间的潜在合作关系，层间箭头代表多层合作网络的构建顺序（I-T-S）。

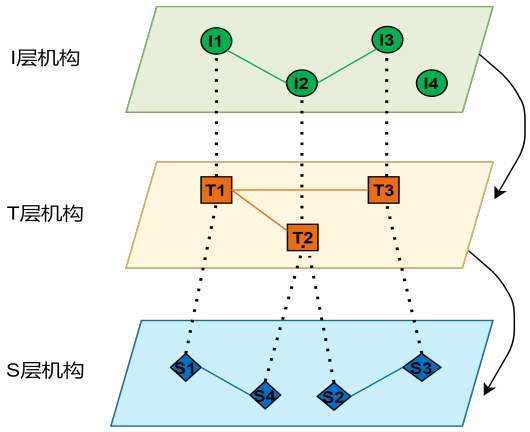

图 1　多层合作网络示意图

3　基于创新全过程的潜在合作对象识别方法

本文基于创新全过程视角构建了从产业机构出发的潜在合作对象识别方法，主要包括以下 5 个环节，识别方法及流程如图 2 所示。

（1）确定候选机构集合。本文将科技论文的发文机构、技术专利的申请机构、产品的生产发起机构分别作为科学层、技术层、产业层的候选机构。

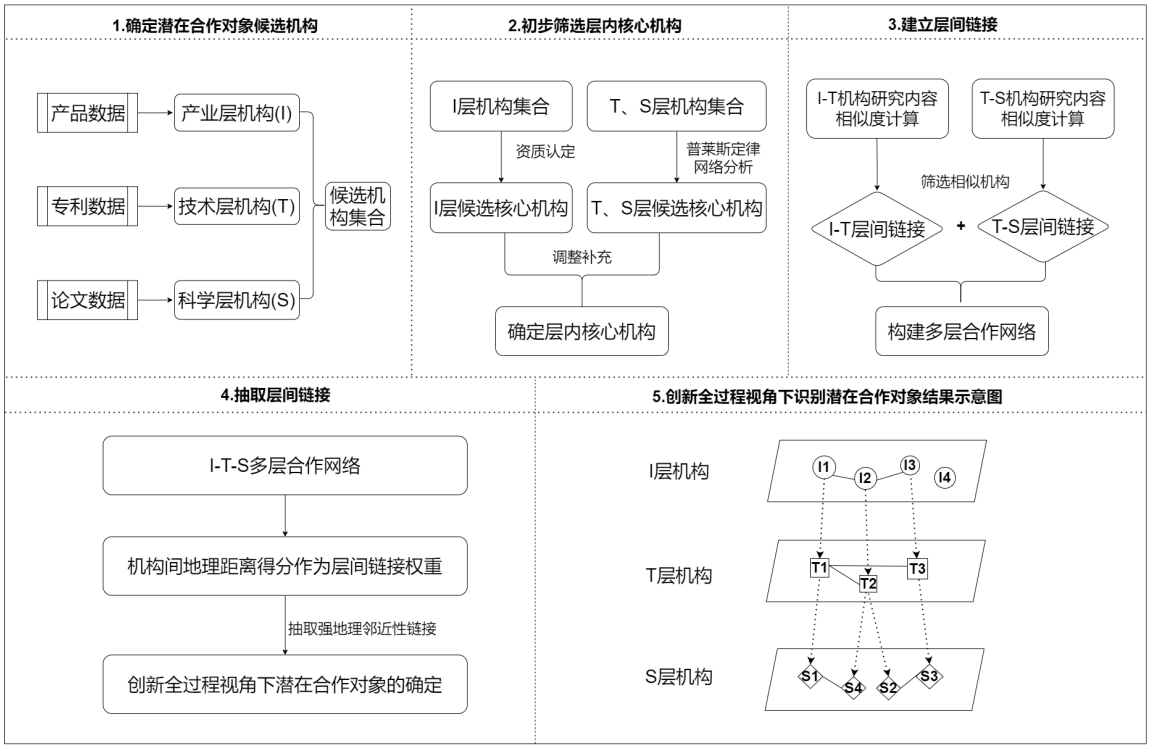

图 2　创新全过程视角下潜在合作对象识别方法及流程

（2）筛选层内机构。根据机构的自身特性、合作关系以及产出能力等指标分别筛选出产业（I）层、技术（T）层、科学（S）层的核心机构，完成层内核心机构的筛选。

（3）建立层间连接。从产业层机构出发，按照产业—技术—科学（I-T-S）的顺序计算不同层核心机构间的研究内容相似度，为不同层中有合作可能性的机构建立层间连接，构建从产业层出发产生连接的三层合作网络。

（4）抽取层间连接。将机构之间的地理距离得分作为层间连接权重，设定权重阈值抽取

doi:10.3772/j.issn.2095-915x.2024.03.003

层间连接，完成创新全过程视角下潜在合作对象的筛选与识别。

（5）选择创新合作对象。识别出潜在合作对象后，进入合作伙伴选择阶段。产业机构根据自身创新战略需求选择最合适的合作对象以形成立足于创新全过程的合作团队。

3.1 核心机构的遴选

核心机构对其领域内其他机构起到引导作用，更有能力与其他创新层中的机构合作以实现高效创新[19,53]。本文筛选各层中具有高产出、高影响力的核心机构作为创新全过程视角下潜在合作对象的初步候选机构，具体步骤见图3。

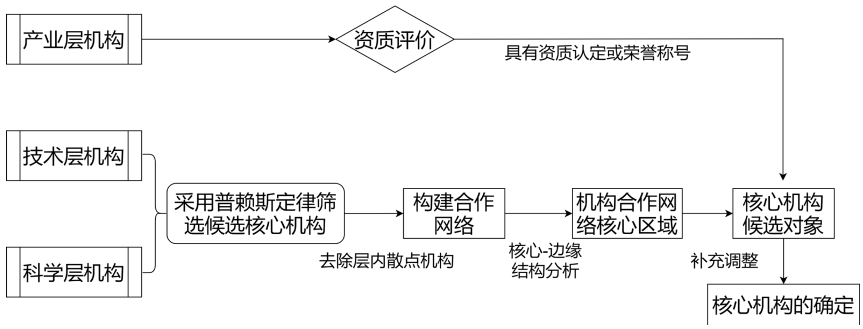

图3 核心机构遴选方法及流程

（1）核心机构的初步筛选

①初步筛选I层核心机构。在I层中，核心机构的筛选主要针对该层内的主要创新主体，即企业机构。对于非企业核心机构，将在后续步骤中进行补充。由于各企业产品的生产在同领域内具有竞争排他性，且目前未有适合的方法来表示企业的合作关系[54]。因此本文暂不对I层内的企业构建合作网络，而是通过判定各企业是否具有相应的资质认定或荣誉称号来筛选产业层内核心机构。原因如下：首先，政府对企业的资质认定或荣誉评定有着严格的审核标准，通过审核的企业在创新表现与经济绩效均优于普通企业[55-56]；其次，国家的扶持政策对企业的研发投入和创新水平均有显著正向激励效应[57]；最后，其所拥有的创新产出条件又为科技成果转化与产业创新提供有力支撑。

综上，获得资质认定或荣誉的企业在各方面均优于普通企业，更适合作为领域内的核心机构。目前，企业资质认定或荣誉称号有高新技术企业、瞪羚企业、雏鹰企业、独角兽企业等。由于中国大多数资质认定以及称号评定针对在大陆注册的企业，因此对于港、澳、台地区的企业，本文通过评估企业官网信息以及咨询专家等方式来判断其是否能够作为核心企业。

②初步筛选T、S层核心机构。随着科研领域创新难度的提升，机构很难独自开展科技创新研究，合作已成为科研创新的主要方式[58-59]。T、S层中机构的创新成果多以合作发明专利、合著论文等形式呈现，机构间合作关系密切，竞争性较弱。因此，本文采用社会网络分析法来筛选T、S层内的核心机构。

首先，采用普赖斯定律对机构进行初步筛选。各层中机构合作网络节点较多且连接复杂，为改善可视化效果，需对机构进行初步筛选。本文根据普赖斯定律[60]，将发文量大于或等于N篇的机构分别作为T、S层内的候选核心机构。

普赖斯定律计算公式如下：

$$N=0.749*\sqrt{n_{max}} \quad (1)$$

其中，n_{max} 为各层内拥有成果最多的机构发表成果的数量，N 表示入选候选核心机构的成果最低数量。

其次，针对 T、S 层内候选核心机构的合作关系分别构建技术层、科学层机构合作网络。未建立合作关系的机构一般合作意向较弱或排他性强，因此去除合作网络中的散点机构。

最后，对机构合作网络采用核心—边缘模型分析。合作网络中的核心节点在合作关系中占据重要地位，核心区可以带动非核心区的发展。通过分析节点联系的紧密程度将机构合作网络划分为核心区域和边缘区域，将核心区域内的机构确定为核心机构。

（2）调整补充并确定核心机构数量

同时涉及多层的机构拥有更丰富的创新资源，与这类机构合作能够加强集成创新优势。例如产业层中隶属于高校或科研院所的机构能实现产品或服务的生产，说明其不仅具有理论与技术创新能力，还具有科技成果转化的实力与集成创新能力。其中，涉及多层的机构主要分为以下几类，如图 4 所示：

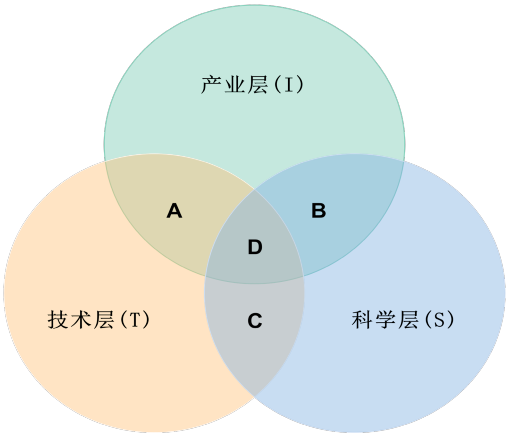

图 4　涉及多创新层的机构

①涉及两个创新层的机构。

A 表示跨产业、技术层，但不涉及科学层的机构；

B 表示跨产业、科学层，但不涉及技术层的机构；

C 表示跨技术、科学层，但不涉及产业层的机构；

②同时涉及三个层的机构（D）：跨产业层、技术层、科学层的机构。

因此，本文筛选各层内涉及两个及以上创新层的机构并将其补充至各层的核心机构中，最终确定各创新层内的核心机构。

3.2　根据机构合作可能性构建层间连接

合作双方具有知识技术相似性是高效开展合作的基础。本文旨在为处于不同创新环节中的核心机构识别跨创新层的潜在合作对象，因此，处于不同创新层内的核心机构应具有相似的研究内容或方向，才能形成创新全过程视角下在某一研究领域囊括三大创新环节、创新目标上下衔接的创新协同组。根据前文理论研究，本文假设内容相似度排名在 TopN 的机构组合有合作可能性，为不同层中有合作可能性的机构建立层间连接。从产业层机构出发，按照产业—技术—科学（I-T-S）的顺序，构建从产业层出发产生连接的三层合作网络。具体步骤如下：

（1）基于 Doc2Vec 算法的文本向量化

本文使用 Doc2Vec 模型对机构的文本内容进行向量化表示。Doc2Vec 算法改进与完善了 Word2Vec 算法，常用于对文档进行分类及向量化应用。提取产业层、技术层、科学层核

doi:10.3772/j.issn.2095-915x.2024.03.003

心机构的文本内容后，对其进行预处理，采用 Doc2Vec 模型对文本信息进行向量化表示，并以此代表各层核心机构的研究内容。

（2）计算机构研究内容相似度

采用向量余弦值计算不同机构研究内容相似度。余弦相似模型通过计算向量间的夹角余弦值来测量向量相似性，性能较好且响应速度较快 [61-62]。采用 Doc2Vec 模型对机构研究内容进行向量化表示后，通过向量余弦值计算不同层间核心机构研究内容的相似度。

（3）构建层间连接

根据需求设定 N 值，选取内容相似度排名在 TopN 的机构组合，为具有合作可能性的不同层核心机构建立层间连接。具体步骤如下：

①筛选与 I 层机构具有合作可能性的 T 层机构。从产业层出发，计算产业层机构与技术层机构的研究内容相似度并排序，筛选与产业层机构内容相似度排名 TopN 的技术层机构。

②筛选与 T 层机构具有合作可能性的 S 层机构。将上一步骤中筛选出的所有技术层机构去重后，计算所筛选的技术层机构与科学层机构的研究内容相似度并排序，筛选与技术层机构内容相似度排名 TopN 的科学层机构。

③为不同层具有合作可能性的机构建立层间连接。将 I 层机构与①中筛选的 T 层机构建立 I-T 连接。对已与 I 层建立连接的 T 层机构，将其与②中筛选的 S 层机构建立 T-S 连接。构建从产业机构出发的 I-T-S 三层合作网络（图 5），识别了层间合作的可能性，完成对创新全过程视角下潜在合作对象的进一步筛选。

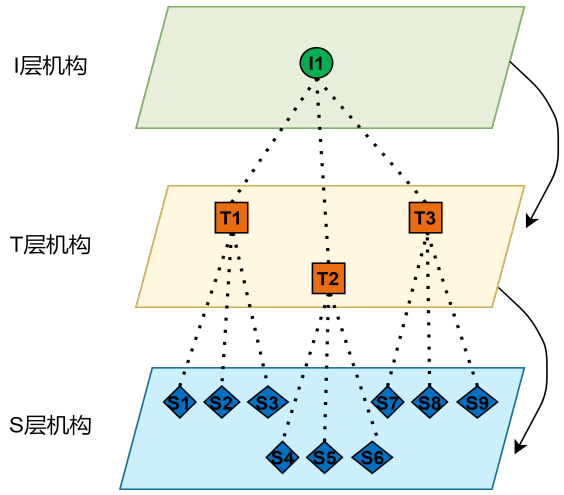

图 5　根据机构相似性构建层间连接示意图（N=3）

3.3　根据地理距离得分抽取层间连接

与具有地理邻近性的对象合作更易实现协同创新，更有利于知识交流与相似性技术的融合 [63-64]。结合现有研究，地理位置越近的机构，成为合作对象的可能性越大。本文根据机构间的地理距离得分为多层合作网络中的层间连接赋权，设定阈值，抽取权重超过阈值的层间连接，完成创新全过程视角下潜在合作对象的最终识别。具体步骤如下：

（1）地理空间距离的计算

不同机构所在城市间的地理距离可以用干线公路距离、航线距离、导航路径距离等方式测度，也可以综合多种方法测度。尽管上述方法的测度结果更接近城市间的真实距离，但存在路径选择困难、数据规模大等问题。本文借鉴嵇正龙等 [65] 关于地理空间距离的表征，采用机构所在城市经纬度计算城市间距离，该方法简化计算更适用于对多个距离进行分析比较。根据机构城市经纬度计算两点近似距离的公式如下：

$$D_{AB}=R*arcos[cos(Y1)*cos(Y2)*cos(X1-X2)+sin(Y1)*sin(Y2)] \quad (2)$$

其中，$X1$、$X2$ 分别为机构 A、B 的经度，$Y1$、$Y2$ 分别为机构 A、B 的纬度，R 为地球的半径，R=6371.0 km。

（2）为层间连接赋权

根据现有研究，假设机构间合作可能性随地理距离增大而减小。本文用 Q_{AB} 表示机构间的地理距离得分，作为多层合作网络层间连接的权重。Q_{AB} 越大，机构间合作的可能性就越大。计算公式如下：

$$Q_{AB}=\frac{1}{D_{AB}} \quad (3)$$

其中，D_{AB} 表示机构 A 与机构 B 所在城市中心点经纬度距离，当机构 A 与机构 B 在同一城市时，Q_{AB}=1。

为多层合作网络的层间连接赋权后，根据需求设定距离得分阈值，抽取层间连接，如图 6 所示，最终完成创新全过程视角下潜在合作对象的识别。

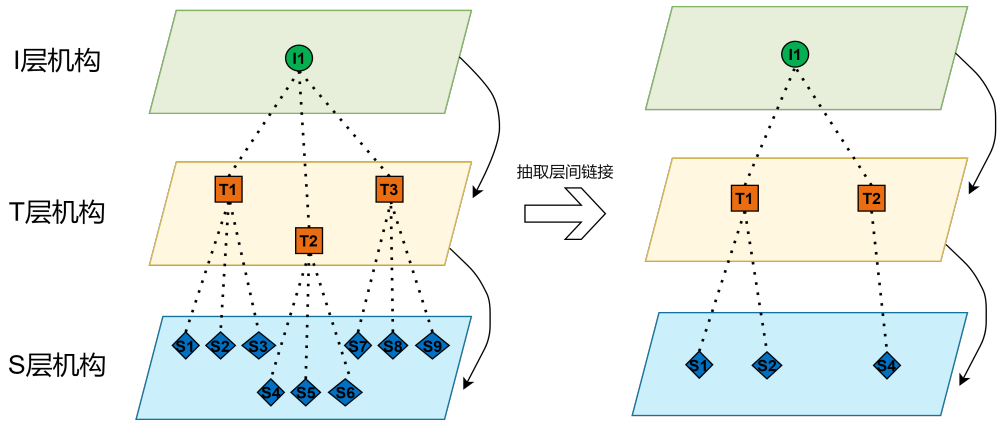

图 6　根据距离得分抽取层间连接

4　实证分析

干细胞研究是当代最前沿的生命科学交叉学科之一，在多种重大疾病治疗方面潜力巨大。世界各国对干细胞研究高度重视，中国也将干细胞研究提升至国家战略高度，加大基础研究和人才储备投入，倡导干细胞领域产业链、创新链、政策链融合发展，打破创新闭环，推动干细胞应用转化，打造科技创新生态。相较于传统学科，干细胞研究领域更能有效验证本文提出的创新全过程视角下潜在合作对象识别方法的可行性与有效性。

4.1　数据处理

4.1.1　数据来源

本文科技论文、专利文献、产品数据分别来源于 Pubmed 论文数据库、Derwent 专利数据库、Cortellis 产品数据库。检索时间为 2016 年 1 月 1 日至 2020 年 12 月 31 日，检索式为 ((ALLD=((“stem cells” OR “stem cell”) NOT (“stem cellulose” or “stem cellular” or “cello” or “cellar” or “cell-phone”)) OR ALLD=((ESC or ASC or iPS or PGC or MSC or CS5C or LSC or TSC or ADSC

doi:10.3772/j.issn.2095-915x.2024.03.003

or HSC) near (cells OR cell))，检索获得 63892 篇干细胞领域相关的科技论文，涉及国内 4033 个科学层机构；获得 7912 篇干细胞领域相关专利文献，涉及国内 2169 个技术层机构；获得 91 个与干细胞领域相关的产品数据，涉及国内 66 个产业层机构。3 名研究人员独立筛选文献及产品数据、提取资料信息，并进行交叉核对，最终获得以上数据。

4.1.2 数据清洗

由于机构数据信息存在标注格式不统一的问题，包括同一机构拥有若干标注形式、曾用名与现用名混用、缩写与全称混用等问题。为保证研究结果准确性，本文结合机构官网信息、“企查查”企业信息查询平台对机构名称进行规范化处理，将机构名称统一为标准形式，规范过程如表 3 所示。

表 3 机构标注格式规范问题（部分）

不规范问题类型	规范前	规范后
同一机构拥有若干标注形式	Renji Hospital, Shanghai Jiao-Tong University School of Medicine	Renji Hospital, Shanghai Jiao Tong University School of Medicine
	The Renji Hospital, Shanghai Jiao Tong University School of Medicine	
	Zhejiang A&F University	Zhejiang A&F University
	Zhejiang Agriculture and Forestry University	
机构与机构合并后名称标注不统一	Qilu University of Technology	Qilu University of Technology (Shandong Academy of Sciences)
	Qilu University of Technology, Shandong Academy of Sciences	
名称标注颠倒	University of South China	University of South China
	South China University	

此外，现实中有些机构在全国不同地市拥有分支机构，如山东大学在济南、青岛、威海等城市均有校区。对此类机构，本文选取机构总部所在城市代表其地理位置。

4.2 核心机构的遴选

（1）核心机构的初步筛选

①初步筛选产业层（I）核心机构。对产业层（I）内的企业机构，结合“企查查”企业信息查询平台以及企业官网，筛选具有资质认定及荣誉称号的企业，统计到符合条件的企业一共有 41 家，如表 4 所示。

②初步筛选技术层（T）核心机构。首先，根据普赖斯定律计算公式，在技术层机构数据集中，发文量最高的为广州赛莱拉干细胞科技股份有限公司，其发文量为 361 篇，N 取值约等于 14。将发表 14 篇以上专利文献的机构选为技术层候选核心机构，共计 85 个。其次，构建技术层（T）候选核心机构合作网络，剔除散点机构后，整体网络共包含 22 个节点，30 条边，见图 7。其中，节点大小表示节点的度，节点越大表示机构的合作数量越多，节点颜色深浅表示机构发文量的多少，连线粗细表示机构间的合作次数。

由图 7 可以看出，技术层（T）候选核心机构合作网络中存在明显的核心区域，该区域中机构合作数量较多且较为密集。接着对合作状况进一步分析，构建机构合作共现矩阵并导入到 Ucinet 软件中，对技术层（T）机构的合作关系进行核心—边缘分析，并计算出节点核心

表 4　I 层核心机构资质认定情况（部分）

机构名称	能力资质 / 荣誉
Tianjin AmCellGene Engineering Co., Ltd.	高新技术企业
Stemirna Therapeutics Co., Ltd.	高新技术企业、独角兽企业
Shenzhen iStem Regenerative Medicine Technology Co., Ltd.	高新技术企业、科技型中小企业
Sclnow (Beijing) Bioengineering Co., Ltd.	高新技术企业、瞪羚企业
Qilu Pharmaceutical Co., Ltd.	高新技术企业、国家级企业技术中心、国家级技术创新示范企业

图 7　T 层候选核心机构合作网络

表 5　T 层机构合作关系核心—边缘分析结果（部分）

区域类型	成员机构
核心区	Chongqing Stem Cell Technology Co., Ltd.; China Stem Cell Group Hainan Boao Affiliate Stem Cell Hospital Co., Ltd.; Shaanxi Stem Cell Technology Co., Ltd.; Suzhou Stem Cell Technology Co., Ltd.; Sanya Stem Cell Technology Co., Ltd.; Shanghai Stem Cell Technology Co., Ltd.; China Stem Cell Group Shanghai Biological Technology Co., Ltd.
边缘区	Guangdong Cardiovascular Disease Research Institute; Guangdong Provincial People’s Hospital (Guangdong Academy of Medical Sciences); Yinfeng Biological Group Co., Ltd.; Henan Yinfeng Biological Engineering Technology Co., Ltd. 等

度，从而对合作网络中每个机构节点的位置有一个量化的认识。T 层机构合作关系核心—边缘分析结果见表 5，核心—边缘密度矩阵见表 6，各机构在合作网络中的核心度见表 7。

表 6　T 层机构合作关系核心—边缘密度矩阵

	核心区域	边缘区域
核心区域	24	0
边缘区域	0	0.39

表 7　T 层合作网络机构核心度（部分）

机构名称	核心度
China Stem Cell Group Hainan Boao Affiliate Stem Cell Hospital Co., Ltd.	0.380
Shaanxi Stem Cell Technology Co., Ltd.	0.379
Suzhou Stem Cell Technology Co., Ltd.	0.378
Shanghai Stem Cell Technology Co., Ltd.	0.378
Sanya Stem Cell Technology Co., Ltd.	0.377
China Stem Cell Group Shanghai Biological Technology Co., Ltd.	0.377

doi:10.3772/j.issn.2095-915x.2024.03.003

核心—边缘分析结果显示，在 T 层的 22 个机构中，处于核心区域的机构共 8 个，这些机构间联系密切并处于合作网络的核心位置。处于边缘区域的机构共 14 个，这些机构之间联系程度较低。从密度矩阵来看，核心区域节点间的互动联系较为密集，密度为 24，说明技术层（T）内关键机构之间交流频繁，具有较强合作创新能力。因此本文选取处于核心区域的 8 个机构作为 T 层的核心机构。

③初步筛选科学层（S）核心机构。首先，根据普赖斯定律计算公式，在科学层机构数据集中，发文量最高的为中国科学院大学，其发文量为 864 篇，N 取值约等于 22.016。将发文量大于 22 的机构选为科学层候选核心机构，共计 494 个。

其次，构建科学层（S）候选核心机构合作网络。去除散点机构后，网络共包含 494 个节点，14247 条边。为更直观展示合作关系，本文将机构合著阈值提高到 10，即仅保留合作次数在 10 次及以上的强合作关系，并选择合适的布局，结果如图 8 所示。

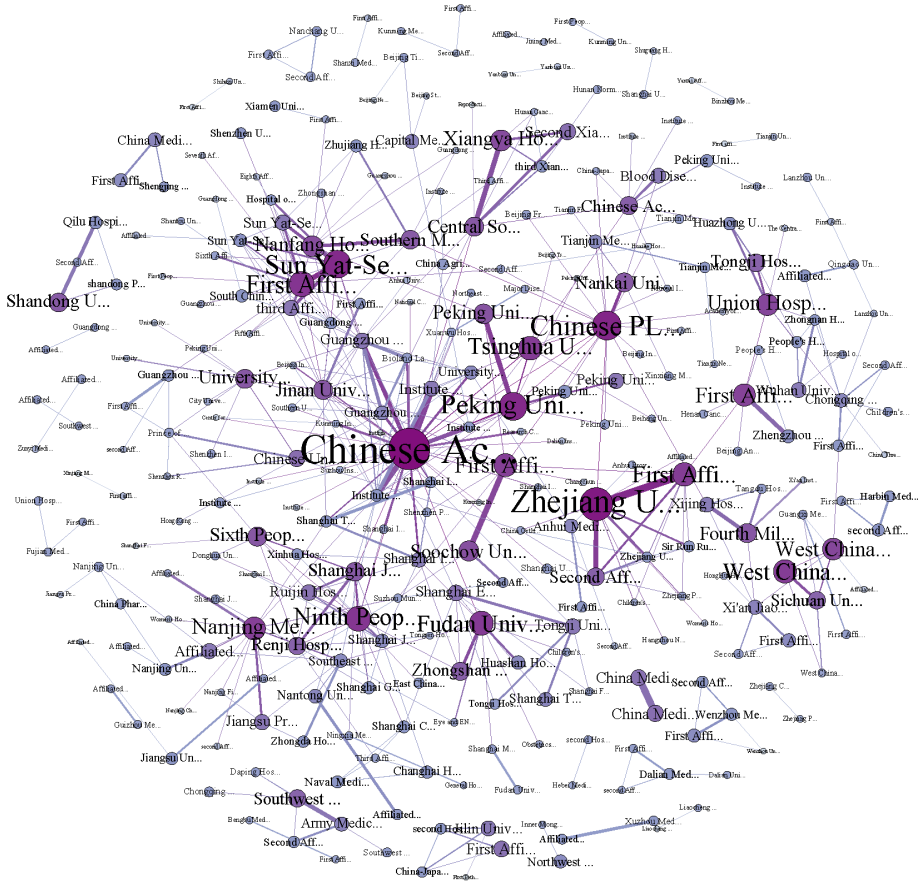

图 8　S 层候选核心机构合作网络

由图 8 可以看出，科学层（S）候选核心机构合作网络中明显存在成果产出较多、合作较为密集的中心合作区域，接着对合作关系进行核心—边缘分析。S 层机构合作关系核心—边缘分析结果如表 8 所示，核心—边缘密度矩阵如表 9 所示，各机构在合作网络中的核心度如表 10 所示。

表 8　S 层机构合作关系核心—边缘分析结果（部分）

区域类型	成员机构
核心区	University of Chinese Academy of Sciences; Sun Yat-Sen University; Zhejiang University; Chinese PLA General Hospital; First Affiliated Hospital, Sun Yat-sen University; Peking University; First Affiliated Hospital, Zhejiang University School of Medicine; Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences; Fudan University; Jinan University; Chinese Academy of Medical Sciences and Peking Union Medical College; Nanjing Medical University; Institute for Stem Cell and Regeneration, Chinese Academy of Sciences; Tsinghua University; First Affiliated Hospital, Soochow University; Guangzhou Medical University; Shanghai Jiao Tong University; Institute of Zoology, Chinese Academy of Sciences; Nanfang Hospital, Southern Medical University; Southern Medical University 等
边缘区	West China Hospital, Sichuan University; Nankai University; Soochow University; Fourth Military Medical University (Airforce Military Medical University) ; Chongqing Medical University; Southeast University; Jiangsu Province Hospital (First Affiliated Hospital of Nanjing Medical University) ; Peking University Third Hospital; Shanghai Tenth People's Hospital, Tongji University 等

表 9　S 层机构合作关系核心—边缘密度矩阵

	核心区域	边缘区域
核心区域	5.792	0.720
边缘区域	0.720	0.169

表 10　S 层合作网络机构核心度（部分）

机构名称	核心度
University of Chinese Academy of Sciences	0.454
Institute of Zoology, Chinese Academy of Sciences	0.261
Institute for Stem Cell and Regeneration, Chinese Academy of Sciences	0.250
Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences	0.232
Guangzhou Medical University	0.177
Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory)	0.162

核心—边缘分析结果显示，在 S 层的 494 个机构中，处于核心区域的机构共 47 个，这些机构之间合作密切并处于合作网络的核心位置。其中，中国科学院大学在网络中核心度最高，发表成果数量最多，在 S 层机构合作网络中合作最为活跃且影响力最大。处于边缘区域的机构 447 个，这些机构成员之间联系程度较低。从密度矩阵来看，核心区域节点之间的联系较为密集，密度为 5.792，说明科学层（S）内关键机构之间合作程度较高，对机构间的合作倾向具有引导作用。因此本文选取处于核心区域的 47 个机构作为 S 层的核心机构。

（2）补充调整各层核心机构数量

筛选各层机构中涉及两个及以上创新层的机构，将其补充到各层核心机构中，调整并最终确定每层核心机构数量。其中，同时涉及三个创新层即跨产业、技术、科学层的核心机构共 7 个，这类机构具备创新的综合实力，选择合作对象时应重点关注。最终确定 I 层核心机构 57 个，T 层核心机构 45 个，S 层核心机构 74 个，如表 11 所示。

4.3　根据机构合作可能性构建层间连接

首先提取干细胞领域内各层核心机构的产品数据、专利数据、科技论文数据。将产品数据中的产品描述、专利的标题与技术信息字段、科技论文的标题与摘要分别作为产业层（I）、技术层（T）、科学层（S）核心机构的研究内容。其次，对研究内容的文本信息预处理，包括分词处理和停用词处理。之后使用 Python 语言中的 Gensim 库进行 Doc2vec 模型的计算，将文本

doi:10.3772/j.issn.2095-915x.2024.03.003

信息向量化，采用向量余弦值计算研究内容相似度。由于本文研究的是机构间的合作关系，因此涉及多创新层的同一机构不计算内容相似度。

从产业层机构出发，计算 I-T 层之间核心机构研究内容的相似度，计算结果如表 12 所示。其中数值越大，表示机构间研究内容相似度越高。

表 11　各层核心机构（部分）

I 层核心机构	T 层核心机构	S 层核心机构
Jiangxi Xlotus Medical Science and Technology Co., Ltd.	Academy of Military Medical Sciences	Chinese Academy of Sciences
Tianjin AmCellGene Engineering Co., Ltd.	China Stem Cell Group Hainan Boao Affiliate Stem Cell Hospital Co., Ltd.	Sun Yat-Sen University
Guangzhou Saliai Stemcell Science and Technology Co., Ltd.	Chongqing Stem Cell Technology Co., Ltd.	Zhejiang University Chinese PLA General Hospital
Nanjing University	Sanya Stem Cell Technology Co., Ltd.	First Affiliated Hospital, Sun Yat-sen University
Chinese PLA General Hospital	Shaanxi Stem Cell Technology Co., Ltd.	Peking University

表 12　I-T 核心机构研究内容相似度计算结果（部分）

I 层 / T 层	Qilu Pharmaceutical Co., Ltd.	TaiGen Biotechnology Co., Ltd.	MabSpace Biosciences Co., Ltd.	Sun Yat-Sen University	…
Jiangsu University	0.709	0.447	0.643	0.797	…
Tongji University	0.674	0.237	0.626	0.722	…
Southeast University	0.662	0.339	0.631	0.743	…
Nantong University	0.654	0.520	0.706	0.718	…
…	…	…	…	…	…

综合考虑相似度及机构数量等因素，设定 N=10，即相似度排名在前 10 的机构组合有合作的可能性。从 I 层机构出发，选取与 I 层核心机构研究内容相似度排名前 10 的 T 层核心机构，为其建立 I-T 层间连接。对已与 I 层产生层间连接的 T 层机构，计算这些机构与科学层核心机构之间即 T-S 层之间的机构研究内容相似度，计算结果如表 13 所示。

表 13　T-S 核心机构研究内容相似度计算结果（部分）

T 层 / S 层	Jiangsu University	Tongji University	Southeast University	Nantong University	…
Southern Medical University	0.554	0.387	0.496	0.489	…
Shandong University	0.470	0.378	0.505	0.423	…
Sichuan University	0.456	0.374	0.401	0.307	…
Zhejiang University	0.449	0.464	0.510	0.456	…
…	…	…	…	…	…

在 T-S 核心机构研究内容相似度计算结果中选取与上述 T 层机构研究内容相似度排名前 10 的 S 层核心机构，为其建立 T-S 层间连接，构建无权 I-T-S 三层合作网络（图 9），初步识别创新全过程视角下不同创新层机构之间的潜在合作关系。

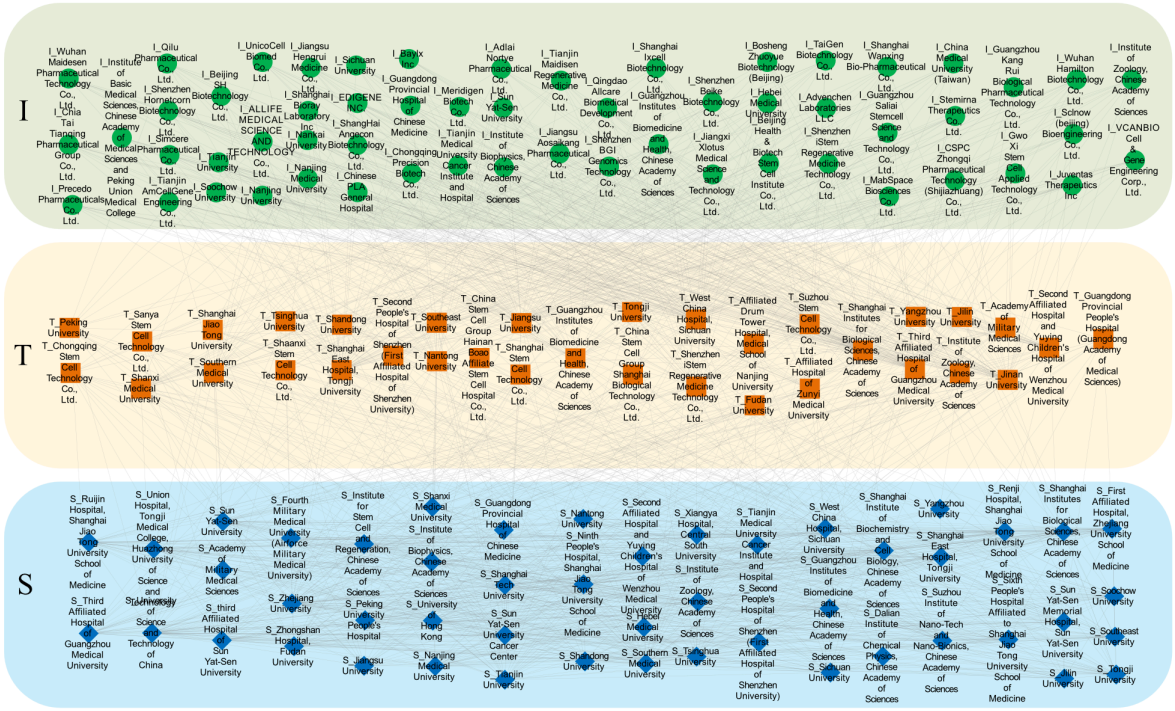

图 9　无权 I-T-S 多层合作网络

4.4　根据地理距离得分抽取层间连接

通过百度地图开放平台获取机构所在城市中心点的 BD09 坐标系经纬度，根据经纬度计算产生连接机构间的地理距离及距离得分。将距离得分作为层间连接权重，构建 I-T-S 有权三层合作网络。将距离得分均值作为阈值，抽取层间连接，筛选产生层间连接的相似机构中的地理邻近机构，完成创新全过程视角下潜在合作对象的最终识别。识别结果如图 10 所示，其中层间连线越粗代表机构间地理距离越近。

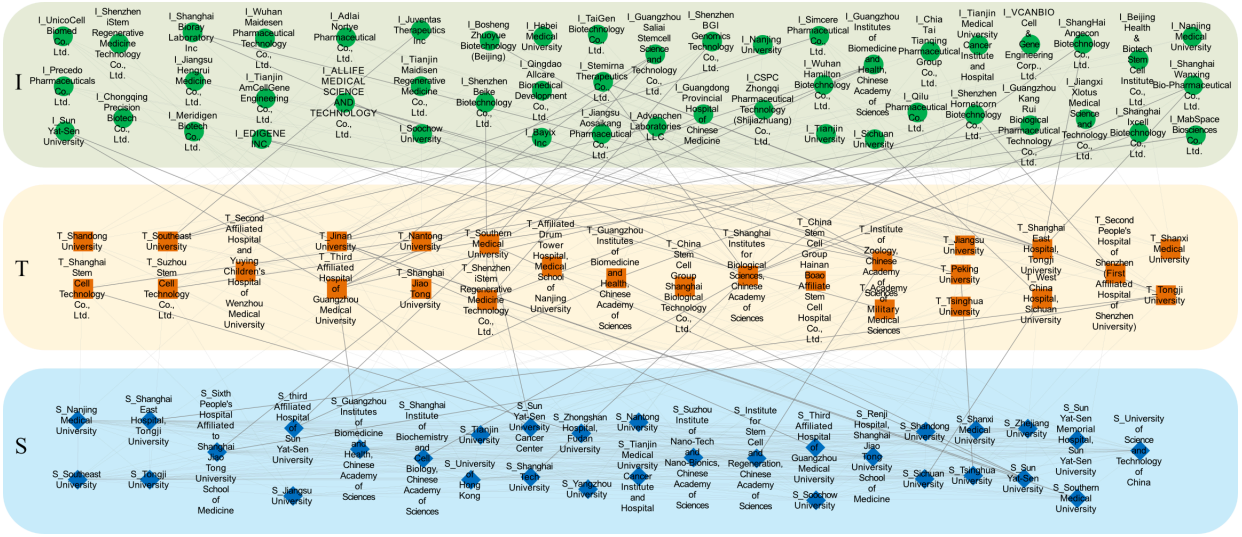

图 10　有权 I-T-S 多层合作网络

doi:10.3772/j.issn.2095-915x.2024.03.003

4.5 结果分析与验证

本文从产业层（I）机构出发，识别跨 I-T-S 多层的潜在合作对象，形成了创新全过程视角下的创新协同组。本文以北京汉氏联合干细胞研究院有限公司（以下简称“汉氏联合”）（Beijing Health & Biotech Stem Cell Institute Co., Ltd.）为例，对识别结果与合作对象选择过程进行解读。选取原因如下：在本文的样本中，①该企业是国家认定的高新技术企业，科技综合实力较强，是中国干细胞产业的先锋企业。为其识别创新全过程视角下的合作对象，围绕其产业链构建创新链，有助于加快干细胞创新产品的研发与生产。②该企业研究生产了关于人胎盘间充质干细胞、脐带来源间充质干细胞的药品，说明其具备较强的产品开发和科技成果转化实力。但结合其现有合作关系来看，该企业的合作对象均为产业层机构，缺乏科学层、技术层的合作机构来构建创新合作团体，因此有必要为其识别创新全过程视角下的潜在合作对象。③本文研究团队与该企业有过合作往来，便于验证本文的研究结果。

在实证分析中，为“汉氏联合”识别出 6 个贯穿产业、技术、科学层的涉及多环节、多主体的创新协同组，分别由 2 个技术层（T）机构、5 个科学层（S）机构组成，使其能够认识到创新全过程下足够多的潜在合作对象，如图 11 所示。

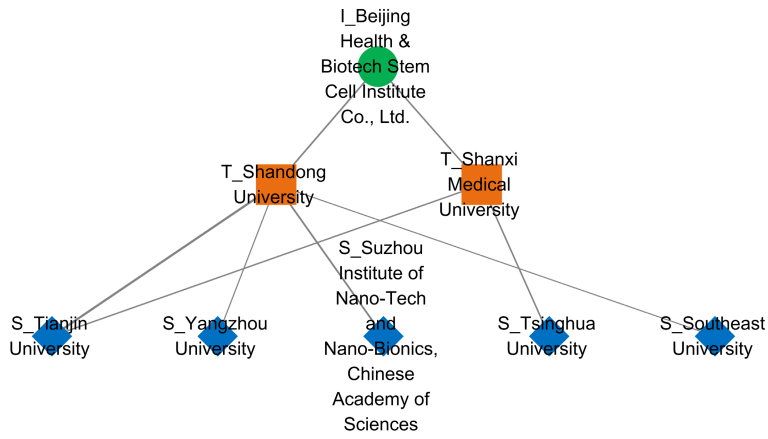

图 11 创新全过程视角下“汉氏联合”的潜在合作对象

以上经过各类指标的层层筛选后识别得到的 6 个创新协同组，是达到可接受预期匹配水平的潜在合作创新团队，所包含的创新活动能够覆盖创新全过程，具备实现基础研究、应用研究、产业应用的能力。

本文提出的潜在合作对象识别方法具有自适应性和可扩展性，产业机构可根据自身创新战略及合作意向在不同创新层中选择具有不同特征的合作对象形成具有不同功能侧重的创新链团体。为支持“汉氏联合”选择最匹配的合作对象，本文对不同协同组中各机构间的产出能力、研究内容相似度、地理距离进行分析，用机构的产出成果在其所在创新层的核心机构产出成果总数中的占比表示机构的产出能力。对“汉氏联合”潜在协同组的 T 层机构能力分析见表 14、S 层机构能力分析如表 15 所示。

表 14 “汉氏联合”创新团体 I-T 机构能力分析

I 层机构	Beijing Health & Biotech Stem Cell Institute Co., Ltd.	
T 层机构	Shandong University	Shanxi Medical University
机构产出成果占比（%）	0.341	0.190
机构间内容相似度	0.666	0.662
距离得分	0.003	0.002

表 15 “汉氏联合”创新团体 T-S 机构能力分析

T 层机构	Shandong University				Shanxi Medical University	
S 层机构	Southeast University	Yangzhou University	Tianjin University	Suzhou Institute of Nano-Tech and Nano-Bionics	Tianjin University	Tsinghua University
机构产出成果占比（%）	0.320	0.191	0.148	0.093	0.148	0.694
机构间内容相似度	0.514	0.512	0.510	0.510	0.494	0.478
距离得分	0.002	0.002	0.004	0.003	0.002	0.003

当“汉氏联合”的创新战略倾向于高产出和研究内容相近时，该企业可优先选择山东大学作为应用研究环节的合作机构，选择东南大学作为基础研究环节的合作机构，形成具有高产出、高竞争力、内容更高关联、目标更加匹配的创新链团体。该合作团队中的创新机构均致力于人类胎盘间充质干细胞应用于慢性伤口修复与炎症的抑制、透明质酸钠用于增强间充质干细胞分泌功能等方面的研究开发，能够实现现有研究领域基础研究、应用研究与产业应用的全面发展，逐步成为引领研究领域发展的中坚力量。

当创新合作需求倾向于地理邻近时，“汉氏联合”可优先选择山东大学作为应用研究环节的合作机构，选择天津大学作为基础研究环节的合作机构，形成具有强地理邻近性、社会文化及经济政策更相似的创新链团体。该团队中各机构间距离较近，便于面对面沟通与交流，提高知识交流和资源互换的效率，为科技成果转移转化提供了便捷渠道，能够更好地提高创新绩效。

为验证本文方法的可行性与识别结果的有效性，本文随机抽取产业层核心机构中产品发明量排名前 20 的机构的专家与机构合作事项负责人，将识别过程与结果（包括潜在合作机构所涉及的创新环节、机构能力资质及特点、研究内容或方向、地理位置等信息）以在线咨询或邮件的方式发送给他们，并请他们回复合作意愿以及其所在机构对于合作伙伴的标准和合作需求，邀请他们探讨本文的实证结果。对回复结果统计分析后，发现大多数识别结果符合机构对其合作伙伴的要求，创新合作能力较强、研究方向与本机构相符、地理距离合适更易于交流是其选择合作对象的重要考虑因素。机构合作项目负责人认为本文为其识别的创新团队具备与其合作创新的综合实力，具有一定合作可行性。少数专家对识别结果存在异议，因为识别的潜在合作对象已有固定且成熟的创新合作团队。上述识别结果是未来合作对象的参考选择，在实际创新活动中，各机构是否真正产

doi:10.3772/j.issn.2095-915x.2024.03.003

生合作，还受诸多现实因素影响。

5 结论与展望

本文基于创新全过程视角，综合考虑合作的潜在规律及影响因素，构建多层合作网络来展现贯穿基础研究、应用研究与产业应用多环节的潜在合作对象识别流程，提出层层递进的识别方法，支撑创新主体高效、准确地定位创新全过程视角下的潜在合作对象。以干细胞领域中的创新机构为研究对象的实证分析验证了方法的可行性与有效性。在识别结果的基础上可根据不同的创新战略及合作意向，选择更符合自身需求的机构形成创新团体，实现干细胞领域更高效的科技创新。本文提出的方法弥补了已有研究中仅基于单一创新环节或单一指标对主体间潜在合作关系识别的不足，为各领域有科技创新需求或科技成果转化困难的机构在开放式创新环境下识别合作伙伴提供了理论支撑与实践启示，且对现有创新团体合作关系的可持续性具有一定预测作用。

本文仍存在一些局限。首先，随着时间推移，机构的研究内容处于动态变化过程中，因此在衡量机构间的研究内容相似度时，有必要考虑机构研究主题随时间的变化；其次，合作对象的遴选过程复杂，影响因素众多，不同创新环节的主体有着不同的创新需求，后续可考虑创新主体的更多属性与合作需求对识别方法进行改进。最后，未来采用可量化的结果验证方法以消除验证方法中的主观性，以期获得更具有科学性与实用性的创新全过程视角下的潜在合作对象识别方法。

参 考 文 献

[1] 光明日报 . 科技部 财政部关于印发《企业技术创新能力提升行动方案（2022—2023 年）》的通知 [EB/OL]. (2022-08-05)[2022-11-08]. http://www.gov.cn/zhengce/zhengceku/2022-08/15/content_5705464.htm.

[2] 本刊综合 . 省科技厅联合 7 部门出台政策推动创新链产业链融合发展 [J]. 今日科技 , 2022(7): 33-34.

[3] 任志宽 . 推动产业链与创新链深度融合 [J]. 支部建设 , 2020(23): 13-14.

[4] 丁敬达 , 郭杰 . 融合内容相似度和路径相似性的潜在作者合作关系挖掘 [J]. 情报理论与实践 , 2021, 44(1): 124-128, 123.

[5] 张金年 , 罗艳 . 基于内容的作者研究相似度与潜在合作网络分析 —— 以图书馆学期刊为例 [J]. 情报科学 , 2021, 39(8): 86-93.

[6] 温芳芳 . 基于专利权人 - 分类号耦合分析的潜在合作关系网络研究 [J]. 情报学报 , 2016, 35(12): 1265-1272.

[7] 陈浩 , 张梦毅 , 程秀峰 . 融合主题模型与决策树的跨地区专利合作关系发现与推荐 —— 以广东省和武汉市高校专利库为例 [J]. 数据分析与知识发现 , 2021, 5(10): 37-50.

[8] TIMMERS P. Building effective public R&D programs[Z]. PICMET Book of Summaries, 1999(1): 591-597.

[9] TURKENBURG W C. The innovation chain: policies to promote energy innovations[J]. Energy for sustainable development, 2002: 137-172.

[10] 焦豪 , 张睿 , 马高雅 . 国外创新生态系统研究评述与展望 [J]. 北京交通大学学报 (社会科学版), 2022, 21(4): 100-112.

[11] ADNER R. Match Your Innovation Strategy to Your Innovation Ecosystem[J]. Harvard Business Review, 2006, 84(4): 98-107.

[12] 王高峰 , 杨浩东 , 汪琛 . 国内外创新生态系统研究演进对比分析: 理论回溯、热点发掘与整合展望 [J]. 科技进步与对策 , 2021, 38(4): 151-160.

[13] KLIMAS P, CZAKON W. Species in the wild: a typology of innovation ecosystems[J]. Review of Managerial Science, Springer, 2022, 16(1): 249-282.

[14] 赵程程 , 秦佳文 . 美国创新生态系统发展特征及启示 [J]. 世界地理研究 , 2017, 26(2): 33-43.

[15] NGONGONI C N, GROBBELAAR S, SCHUTTE C S. Making Sense of the Unknown: Using Change Attractors to Explain Innovation Ecosystem Emergence[J]. Systemic Practice and Action Research, 2022, 35(2): 227-252.

[16] MOORE J F. Predators and Prey: A New Ecology of Competition[J]. Harvard business review, 1999, 71(3): 75-86.

[17] 李军凯 , 高菲 , 龚轶 . 构建面向未来产业的创新生态系统: 结构框架与实现路径 [J]. 中国科学院院刊 , 2023, 38(6): 887-894.

[18] 陈悦 , 陈超美 , 刘则渊 , 等 . CiteSpace 知识图谱的方法论功能 [J]. 科学学研究 , 2015, 33(2): 242-253.

[19] 王其冬 , 李东 , 周艳春 , 等 . 基于项目合作的社会关系网络中核心社团发现 [J]. 计算机应用与软件 , 2020, 37(3): 173-177, 224.

[20] 王衍喜 , 周津慧 , 王永吉 , 等 . 一种基于科技文献的学科团队识别方法研究 [J]. 图书情报工作 , 2011, 55(2): 55-58, 98.

[21] 张克群 , 张文 , 汪程 . 知识组合新颖性、网络特征与核心发明人关系研究 [J]. 情报杂志 , 2022, 41(1): 185-191.

[22] 高楠 , 周庆山 . 基于有向合著网络的核心团队挖掘研究 —— 以图情领域为例 [J]. 图书情报工作 , 2021, 65(20): 81-91.

[23] 黄维 , 陈勇 . 中国教育经济学研究者合作网络的社会网络分析 [J]. 现代大学教育 , 2010(2): 14-19, 111.

[24] 王超 , 武华维 , 董振飞 , 等 . 重大公共危机中科研协同团队识别研究 —— 以诊断试剂为例 [J]. 科技进步与对策 , 2020, 37(9): 21-29.

[25] 王谦 , 丁倩 , 郭春彦 , 等 . 我国儿科人群多中心药物临床试验机构合作关系网络分析 [J]. 中国医院用药评价与分析 , 2024, 24(3): 257-260, 264.

[26] 李纲 , 李春雅 , 李翔 . 基于社会网络分析的科研团队发现研究 [J]. 图书情报工作 , 2014, 58(7): 63-70, 82.

[27] 于永胜 , 董诚 , 韩红旗 , 等 . 基于社会网络分析的科研团队识别方法研究 —— 基于迭代的中间中心度排名方法识别科研团队领导人 [J]. 情报理论与实践 , 2018, 41(7): 105-110.

[28] CAPON N, GLAZER R. Marketing and Technology: A Strategic Coalignment[J]. Journal of Marketing, 1987, 51(3): 1-14.

[29] CRISTIAN G, JORGE H, CHRISTIAN F, et al. Proximity as determinant of business cooperation for technological and non-technological innovations: a study of an agribusiness cluster[J]. Journal of Business & Industrial Marketing, 2017, 32(1): 167-178.

[30] WANG X, LI R, HUANG Y, et al. Identifying R&D partners for dye-sensitized solar cells: a multi-level patent portfolio-based approach[J]. Technology Analysis & Strategic Management, 2019, 31(3): 356-370.

[31] 徐健 , 毛进 , 叶光辉 , 等 . 基于核心作者研究兴趣相似性网络的社群隶属研究 —— 以国内情报学领域为例 [J]. 图书情报工作 , 2018, 62(12): 57-64.

[32] 王超 , 许海云 , 方曙 . 产学研潜在合作对象识别方法研究 [J]. 科学学研究 , 2018, 36(1): 101-113.

[33] 李仲 , 韩红旗 , 吴广印 , 等 . 基于文本稀疏分布式表征的学术合作推荐 [J]. 情报科学 , 2019, 37(6): 113-118.

[34] CHUAN P, SON L, ALI M, et al. Link Prediction in Co-authorship Networks based on Hybrid Content Similarity Metric[J]. Applied Intelligence, 2018, 48.

[35] 陈卫静 , 郑颖 . 基于作者关键词耦合的潜在合作关系挖掘 [J]. 情报杂志 , 2013, 32(5): 127-131.

[36] LE Q, MIKOLOV T. Distributed Representations of Sentences and Documents[J]. Eprint Arxiv, 2014(4): 1188-1196.

[37] GERTLER M S. Tacit knowledge and the economic geography of context, or The undefinable tacitness of being (there)[J]. Journal of Economic Geography, 2003, 3(1): 75-99.

[38] KATZ J S, MARTIN B R. What is research collaboration?[J]. Research Policy, 1997, 26(1): 1-18.

[39] ARINO A, ABRAMOV M, SKOROBOGATYKH I, et al. Partner Selection and Trust Building in West European-Russian Joint Ventures: A Western Perspective[J]. International Studies of Management & Organization, Taylor & Francis, Ltd., 1997, 27(1): 19-37.

[40] 温芳芳 . 基于专利文献计量的我国校企科研合作

doi:10.3772/j.issn.2095-915x.2024.03.003

现状分析 [J]. 情报杂志 , 2014, 33(12): 71-76.
[41] 贺超城 , 吴江 , 魏子瑶 , 等 . 科研合作中机构间科研主导力及邻近性机理 —— 以中国生物医学领域为例 [J]. 情报学报 , 2020, 39(2): 148-157.
[42] AUTANT-BERNARD C, FADAIRO M, MASSARD N. Knowledge diffusion and innovation policies within the European regions: Challenges based on recent empirical evidence[J]. Research Policy, 2013, 42(1): 196-210.
[43] NEWMAN M E J. The structure of scientific collaboration networks[J]. Proceedings of the National Academy of Sciences, Proceedings of the National Academy of Sciences, 2001, 98(2): 404-409.
[44] SCHUEMIE M J, KORS J A. Jane: suggesting journals, finding experts[J]. Bioinformatics, 2008, 24(5): 727-728.
[45] COHEN S, EBEL L. Recommending collaborators using keywords[C]//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro Brazil: ACM, 2013: 959-962.
[46] 吴晓波 , 吴东 . 论创新链的系统演化及其政策含义 [J]. 自然辩证法研究 , 2008, 24(12): 58-62.
[47] 王超 . 基于创新链的区域产业竞争力分析方法研究 [D]. 北京 : 中国科学院大学 , 2018.
[48] 韩菁 , 唐箫 , 余乐安 . 基于多层网络链路预测的潜在合作关系识别研究 [J]. 系统工程理论与实践 , 2021, 41(4): 1049-1060.
[49] 刘晓燕 , 侯文爽 , 单晓红 . 基于多层网络的科技创新政策府际合作机理 [J]. 科研管理 , 2021, 42(3): 97-108.
[50] ZHANG G, TANG C. How could firm's internal R&D collaboration bring more innovation?[J]. Technological Forecasting and Social Change, 2017, 125: 299-308.
[51] XU H, WANG C, DONG K, et al. A Study of Methods to Identify Industry-University-Research Institution Cooperation Partners based on Innovation Chain Theory[J]. Journal of Data and Information Science, 2018, 3(2): 38-61.
[52] 李政 , 王思霓 . 基础研究与应用研究的产业创新效应 [J]. 武汉大学学报 (哲学社会科学版), 2021, 74(5): 91-104.
[53] 王崇锋 , 崔运周 , 王世杰 , 等 . 冠状病毒防疫领域的专利技术合作网络特征研究 [J]. 科学与管理 , 2021, 41(1): 1-9.
[54] 王超 , 张伟然 , 许海云 , 等 . 突发公共卫生危机中科研应急力量的协同整合研究 —— 以诊断试剂应急研发为例 [J]. 情报理论与实践 , 2021, 44(1): 138-146.
[55] 刘家树 , 石洪波 , 齐昕 . 创新链视角下高新技术企业认定效应与机制研究 [J]. 科研管理 , 2022, 43(6): 22-31.
[56] 党兴华 , 郑登攀 . 技术创新网络中核心企业影响力评价因素研究 [J]. 科研管理 , 2007(S1): 19-25, 18.
[57] 雷根强 , 郭玥 . 高新技术企业被认定后企业创新能力提升了吗 ?—— 来自中国上市公司的经验证据 [J]. 财政研究 , 2018(9): 32-47.
[58] 李志宏 , 王娜 , 周广刚 . 国内管理科学领域高校间的学术论文合著网络分析 [J]. 研究与发展管理 , 2012, 24(4): 71-80.
[59] 刘明信 , 李丹丹 , 李荣 , 等 . 专利视阈下产业技术创新合作网络演化分析 —— 以肿瘤疫苗为例 [J]. 图书情报工作 , 2022, 66(9): 117-127.
[60] 丁学东 . 文献计量学基础 [M]. 北京 : 北京大学出版社 , 1992: 220-236.
[61] JEONG Y K, SONG M, DING Y. Content-based author co-citation analysis[J]. Journal of Informetrics, 2014, 8(1): 197-211.
[62] VAN ECK N J, WALTMAN L. Appropriate similarity measures for author co-citation analysis[J]. Journal of the American Society for Information Science and Technology, 2008, 59(10): 1653-1661.
[63] ANSELIN L, VARGA A, ACS Z. Local Geographic Spillovers between University Research and High Technology Innovations[J]. Journal of Urban Economics, 1997, 42(3): 422-448.
[64] 许海云 , 隗玲 , 庞弘燊 , 等 . 产学研潜在合作对象识别方法研究 [J]. 情报学报 , 2016, 35(5): 521-529.
[65] 嵇正龙 , 宋宇 . 融合空间距离测度市场分割指数的方法设计与应用 [J]. 统计与信息论坛 , 2021, 36(7): 10-17.

doi:10.3772/j.issn.2095-915x.2024.03.004

基于 LDA-ARIMA 的我国智能手机关键技术主题识别与演化分析

开放科学（资源服务）标识码（OSID）

庞庆华　姚玉康　张丽娜

河海大学商学院　常州 213022

摘要：［目的 / 意义］专利是技术能力的表现形式，包含关键技术主题信息。以智能手机专利数据为基础，提出关键技术主题的识别和演化分析方法，帮助企业获取行业内的技术信息，调整专利研究的成本与精力投入。［方法 / 过程］首先，选择专利数据库和高级检索内容，下载和导出专利标题和摘要数据，并对数据进行去停用词和 jieba 分词等处理；其次，构建困惑度求解模型，确定最优主题数，再将已经分好词的文本导入 LDA 模型进行主题挖掘，得到每个技术主题下关键词语的分布；再次，将主题热度转化为时间序列，进行平稳性检测和白噪声检验，确定 ARIMA 模型参数后应用预测；最后，依据每个技术主题下提取的特征词确定关键技术主题并进行解读，通过时间序列预测结果对关键技术主题进行演化分析。［结果 / 结论］以智能手机为研究对象，成功识别出屏幕、电池与充电、生物识别系统等 15 个关键技术主题，挖掘出各主题不同的演化发展特征，根据演化趋势分析提出建议，验证了本文主题识别与演化分析方法的可行性与有效性。

关键词：关键技术；主题识别；LDA 模型；主题演化；ARIMA 模型；智能手机

中图分类号：G35; TP391

Identification and Evolution Analysis of Key Technologies Topics of Smartphones in China Based on LDA-ARIMA

PANG Qinghua　YAO Yukang　ZHANG Lina

School of Business, Hohai University, Changzhou 213022, China

Abstract: [Objective/Significance] Patent is a manifestation of technological capability and contains information on key technological topics. Based on smartphone patent data, the identification and evolution analysis method of key technological topics is proposed to help enterprises obtain technological information in the industry and adjust the cost and effort invested in patent research. [Methods/Processes] Firstly, the patent database and advanced search content are selected, and the patent title and abstract data are downloaded and exported. The data is then processed, including deactivating words and jieba-splitting.

基金项目　河海大学 2022 年度“优秀专业硕士学位论文选拔培育”（422003524）。

作者简介　庞庆华（1977-），博士，教授，主要研究方向为竞争情报；姚玉康（1999-），硕士研究生，主要研究方向为数据分析与挖掘；E-mail：yyk13382892903@163.com；张丽娜（1986-），博士，副教授，主要研究方向为资源规划与决策。

引用格式　庞庆华 , 姚玉康 , 张丽娜 . 基于 LDA-ARIMA 的我国智能手机关键技术主题识别与演化分析 [J]. 情报工程 , 2024, 10(3): 49-62.

doi:10.3772/j.issn.2095-915x.2024.03.004

Subsequently, a perplexity solution model is built to determine the optimal number of topics. The preprocessed text is then imported into the LDA model for topic mining, resulting in the distribution of key words under each technical topic. The topic's heat degree is transformed into a time series, followed by smoothness and white noise tests. The parameters of the ARIMA model are determined and applied for prediction. Finally, the feature words extracted under each technical topic are used to identify and interpret the key technical topics. The evolution of these topics is analyzed through the results of the time series prediction. [Results/Conclusions] By focusing on smartphones as the research subject, this study successfully identifies 13 key technological topics such as screen, battery and charging, and biometric system. The different developmental characteristics of each theme are analyzed. Moreover, suggestions are provided based on the analysis of evolutionary trends, thus confirming the feasibility and effectiveness of the theme identification and evolutionary analysis method proposed in this paper.
Keywords: Key Technology; Topic Recognition; LDA Model; Topic Evolution; ARIMA Model; Smart Phone

引言

智能手机作为现代社会不可或缺的一部分，经历了令人瞩目的发展历程。从 2012—2022 年，智能手机行业经历了巨大的变革和创新，全球市场的规模也得到了显著扩大。互联网数据中心（Internet Data Center，简称 IDC）公布 2017 年智能手机出货量达到 15.5 亿台，触及顶峰，而后开始有所下滑。据 IDC 显示，2023 年全球智能手机出货量规模约为 11.7 亿台，同比下降 3.2%。在智能手机市场上，以苹果、三星为代表的美韩企业一直处于领先地位。苹果首次超越三星成为第一位，出货量 2.35 亿台；三星位于第二，出货量 2.27 亿台。值得注意的是，中国的智能手机企业在过去几年中取得了令人瞩目的成绩，在全球市场上崭露头角。2023 年小米、OPPO 分别以 1.46、1.03 亿台的成绩处于第三、第四名，vivo 也在 2023 年第四季度实现同比正增长。根据数据显示，2023 年中国企业的市场份额约占据了全球智能手机市场的三分之一以上，反映了中国企业在技术创新和市场推广方面取得的巨大进步。

关键技术是在产业内起到引导与扩张作用的重要技术，对行业和领域的发展有关键决定作用 [1]。随着智能手机技术的发展迅速，产业变革急速推进，智能手机关键技术的研发和创新是国内各大主流智能手机生产厂商的研究重心。由于芯片、操作系统等受制于国外的技术封锁，我国的智能手机企业存在着短时间难以突破的技术壁垒。在这样的背景下，我国智能手机企业更要加强对关键技术的攻关，合理配置科技资源。

专利是技术能力的表现形式，包含关键技术主题信息。为了更好获取智能手机行业内的重要信息，本文基于专利数据库中的智能手机专利摘要数据，构建了主题识别模型对智能手机关键技术主题进行识别与解读，并将主题热度转化为时间序列进行演化分析。本文旨在帮助我国智能手机企业随着技术市场的转变积极调整技术战略方向，抢夺技术先机，从而在智能手机国际市场的同行业竞争中取得优势。

1 文献综述

1.1 关键技术主题识别相关研究

关键技术主题指对于关键技术的相关特征与内容的总结，关键技术主题识别的过程主要

有两步，分别是：①文本预处理，主要采用 jieba 分词方法 [2] 以及 TF-IDF 算法 [3] 等。其中，传统的分词方法存在识别误差，已有学者对此进行改进，如吕鲲等 [4] 将分词方法相组合，将 jieba 分词（词典分词）和 TF-IDF 算法（统计分词方法）组合运用，提高分词效率，同时也提升了关键词提取的正确率。②文本挖掘，主要使用了 LDA 模型 [5] 以及 SCAN 算法 [6] 等。在文本挖掘方面，伊惠芳等 [7] 改进了增加词汇语境的 Content-LDA 模型，从综述学术论文结束语中挖掘领域技术问题，分析领域主体在生产和创新活动中遇到的各类技术需求，辅助确定关键技术主题。Xu 等 [8] 构建 LDA 主题模型时，在给予响应变量的基础上加入类别标签，提升了中文与外语语料库分类的精细度。王秀红等 [9] 将 BERT 与 LDA 相结合，充分考虑文本的语义信息和上下文信息，明显提高关键技术识别时主题的连贯性及细粒度划分的精准度。

1.2 主题演化分析相关研究

国内外对于技术主题演化分析的方法主要有 ARIMA 模型 [10] 和 SAO 结构法 [11] 等。Bian 等 [12] 组合运用了 ARIMA 模型与 GM 模型、广义神经回归网络模型，研究了趋势预测的组合分析模型。Lv 等 [13] 将 LDA 与 NMF 模型结合，构建双层识别模型，并构建相似性和差异性评价指标，通过关键技术主题演化分析对比主题的发展性和关键性。岳丽欣等 [14] 将 ARIMA 模型的趋势预测功能与人为定性判别方法相结合，利用时间序列模型提高了趋势预测的准确性和科学性，让学科情报分析更加严谨。Hu 等 [15] 从重要性和受欢迎程度两个方面分析了技术话题的演变趋势，并通过构建 ARIMA 模型，预测 2021—2025 年各技术主题在新能源汽车中的流行程度和发展趋势。李勇等 [16] 以生命周期为划分阶段，统计专利摘要词频分析技术热点演变，构建语义网络识别技术关联及技术主题的应用侧重点，更加精准地识别数字文旅关键技术及趋势。

1.3 智能手机专利分析相关研究

智能手机作为互联网时代的产物，在近年来备受关注，对于智能手机技术的研究也有较高的热度。随着专利分析的发展，部分学者尝试将技术识别的方法运用到智能手机技术领域的相关研究中。Sun 等 [17] 提取智能手机技术专利引文数据，根据相关专利参数获取了智能手机核心专利，规划了智能手机相关专利的扩散路径，依据 BCRW 算法实现了专利申请人之中的核心企业的识别。Wang 等 [18] 认为智能手机相关企业需要完善硬件与软件的相关专利布局，同时将智能手机与其他科技融合领域的专利布局进行拓展，将智能手机作为核心，实现了全面的智能科技生态格局。苏敬勤等 [19] 基于德温特数据库中的智能手机专利数据，将专利引用量和被引用量作为指标构建关于时间的变化曲线，分析了颠覆性技术的演化趋势。闯家梁 [20] 基于专利引用量、专利数量等指标构建了智能手机相关技术发展前景的评价体系，分析了我国智能手机专利的技术创新能力，为我国的智能手机相关产业的发展提供了技术指导。彭英等 [21] 制作了专利组合图像来研究智能手机企业的技术竞争力，以专利数量作为指标，分析智

doi:10.3772/j.issn.2095-915x.2024.03.004

能手机企业的技术能力，定义了智能手机专利的相对技术优势进而计算分析了技术强度。

综上所述：一方面，大部分学者将研究重心放在以文本聚类为基础的关键技术主题识别上，关键技术主题识别和演化分析的综合性分析需要进一步深入；另一方面，学者的研究主要集中在研究智能手机的软件层面，而对于硬件技术方面的研究有待加强，另外在智能手机技术领域中关键技术主题识别和主题演化分析方法也需要进一步补充。

针对现有研究所存在的不足，本文将LDA 主题识别模型与 ARIMA 时间序列分析模型相结合，提出一种基于 LDA-ARIMA 的关键技术识别与演化分析方法。本文以专利标题与摘要作为研究数据，构建智能手机专利数据的主题分析模型，挖掘专利数据中的关键技术主题分布情况；将专利研究热度作为时间序列导入趋势预测模型，对智能手机关键技术主题进行演化分析，能够获取智能手机各主题的专利研究热度，帮助企业抓住新的发展机遇。

2 研究设计

2.1 研究思路

本文以智能手机为研究对象，采用 LDA 与 ARIMA 组合方法进行关键技术主题识别与演化分析。本文的研究框架流程分为如下阶段：①数据获取与预处理：选择专利数据库和高级检索内容，下载和导出专利标题和摘要数据，并对数据进行去停用词和 jieba 分词等处理；②基于 LDA 模型的技术主题识别：构建困惑度求解模型，确定最优主题数，再将已经分好词的文本导入 LDA 模型进行主题挖掘，得到每个技术主题下关键词语的分布，并对结果进行聚类以及可视化分析；③基于 ARIMA 模型的技术趋势预测：将主题热度转化为时间序列，进行平稳性检测和白噪声检验，确定模型参数，最后应用预测；④关键技术主题解读与演化分析：依据每个技术主题下提取的特征词确定关键技术主题并进行解读，通过时间序列模型预测结果对关键技术主题进行演化分析，具体研究框架如图 1 所示。

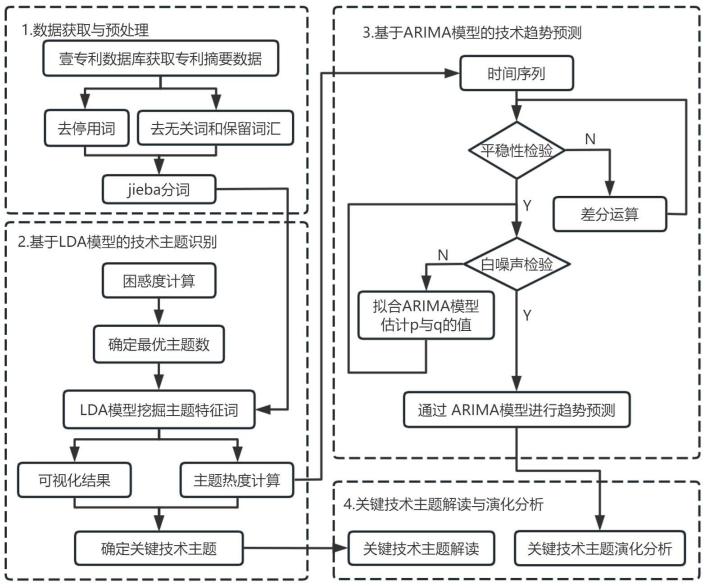

图 1 研究思路

2.2 研究方法

2.2.1 LDA模型

本文利用 LDA 主题模型求得智能手机专利集中每篇专利摘要的主题概率分布，再根据主题分布进行主题聚类和文本分类。其中，一篇专利摘要可以包含多个主题，根据计算产生主题分布概率，可以认为专利摘要中的每个词都由其中的一个主题生成。本文 LDA 模型结构原理如图 2 所示：

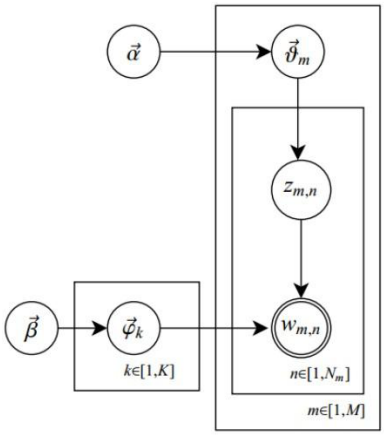

图 2 LDA 模型结构

主题数量通常视不同场景进行调整，本文采用计算困惑度（Perplexity）[22] 的方法来衡量选取智能手机的不同主题数量的优劣：

$$Perplexity(D_{test}) = \exp\{\frac{-\sum_{d=1}^{M}\ln(p(w_d))}{\sum_{d=1}^{M}N_d}\} \quad (1)$$

其中，M 是测试语料的大小（文档的数量），N_d 是第 d 篇文本大小。随着 K（主题个数）值的增大，困惑度逐渐减小。根据手肘法 [23]，困惑度图像存在比较显著的拐点时，选取最佳的 K 值。

本文使用 LDA 模型进行主题识别，利用困惑度确定 LDA 模型的最优主题数，将已经分好词的专利数据导入 LDA 模型，制定好模型参数，选取出现特征词语的频次数，根据困惑度曲线选择主题数量，再确定每个主题提取的特征词数量，将专利标题和摘要以及分词结果存入 EXCEL 结果文件，并输出每条专利数据对应概率最大的主题标号和每条数据对应各个主题的概率，选取各个主题中具有代表性的词作为特征词。

2.2.2 ARIMA模型

ARIMA（自回归积分移动平均）是一种广泛使用的时间序列分析方法。它是一种利用以前的观察结果预测未来趋势的统计模型。ARIMA 模型由三个部分组成：AR（自回归）、I（积分）和 MA（移动平均）。在智能手机专利领域，研究热度趋势具有时间序列特征，ARIMA 模型可以反映智能手机专利研究热度的趋势、季节性特征和随机变化，提供良好的未来趋势预测。

本研究使用的 ARIMA 模型建立步骤如下：首先需要对主题热度时间序列进行平稳性检验，若检验结果不合格，便对时间序列差分运算直至数据达到平稳；其次对时间序列进行白噪声检验；如果主题热度是平稳非白噪声时间序列，就计算 ACF（自相关系数）、PACF（偏自相关系数），确定 ARIMA 模型中参数 p、q 的值，通过拟合精度和参数确定模型的综合最优化配置。

3 研究过程

3.1 数据获取与预处理

3.1.1 数据获取

壹专利数据库覆盖了全球 159 个国家的 1.65 亿多条专利数据，其中囊括专利名称、摘要、

doi:10.3772/j.issn.2095-915x.2024.03.004

专利号、专利价值度、专利状态等权威内容，是国内相对权威的专利数据库，能够成为精准、高效的专利检索和分析工具，因此本文选择壹专利数据库中的智能手机专利信息作为数据来源。

然后构建专利检索式：由于 2011 年为中国智能手机元年，国内智能手机企业如中兴、华为、魅族等开始生产基于 Android 系统的智能手机，于是将检索时间段确定为 2011 年 1 月 1 日—2023 年 7 月 1 日，检索时间为 2023 年 7 月 1 日。由于本文的研究对象主要是国内智能手机企业，故数据范围选择“国内”，申请人类型（中国）选择“企业”且专利状态选择“有效”。本次检索获得 124282 条专利，专利公开的年度趋势如图 3 所示。专利公开量随年份增长急剧增加，其主要原因是智能手机开始普及，国内各大主流智能手机生产厂商加大了专利研发力度。专利数量统计的申请人排名情况如图 4 所示，其中 OPPO、华为、联想、小米、维沃、努比亚是我国知名的通讯设备制造厂商，可见其对专利申请、技术创新的重视程度。

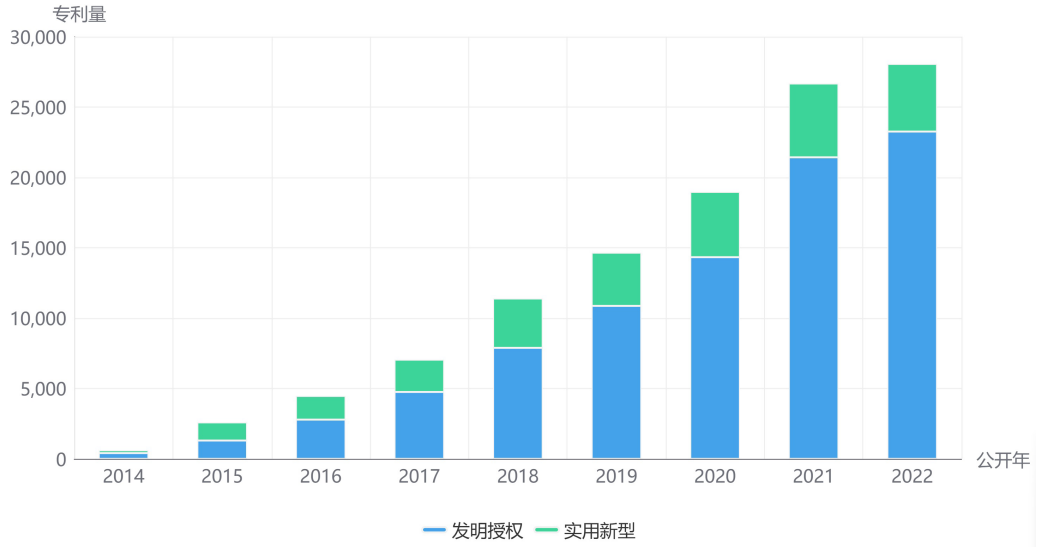

图 3　专利公开的年度趋势

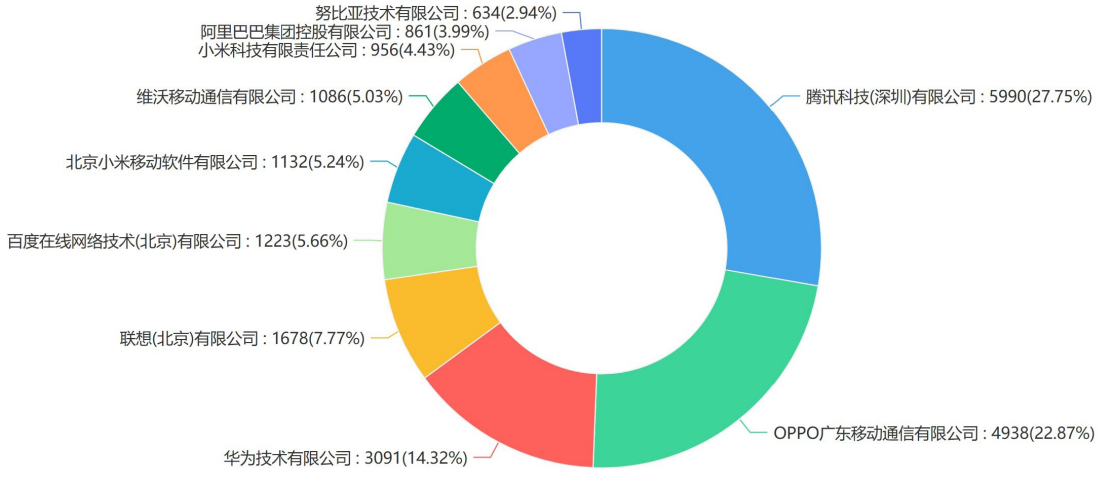

图 4　专利数量统计的申请人排名情况

专利检索数量庞大，其中包含许多干扰性专利，故需要对专利进行筛选。壹专利平台从高专利质量、高技术性、高经济性、发展前景四个维度对专利价值进行评估，专利价值度指标的数值范围为 0~100，数值越高，则专利价值越高。为提高数据的可靠性，模型需采用高质量、高价值的发明专利数据，故本文选取专利价值度大于 50 的专利（共 51057 条）进行导出分析。最终检索式为 TACD=(智能手机) AND 日期 =(2011.01.01:2023.07.01) and selectitem ((申请人类型 =(“企业”) and ipc=(“G” or “H”) and 专利状态 =(“有效”) and 主 ipc=(“G” or “H”) and 专利类型 =(“发明授权” or “实用新型”) and 价值度区间 =(“50~100”)))

3.1.2 数据预处理

本文选择中文分词组件 jieba 对智能手机专利摘要数据进行预处理。先建立停用词列表用于删除停用词，其次创建无关词列表用于去除专利摘要中的无关词，例如“本发明”“方法”等。本文在基本的中文分词基础上做改进：创建自定义保留词典用于保留智能手机新兴短语，例如“安卓”“蓝牙”等。准备工作完成后，对数据进行中文分词处理，分词结果如表 1 所示。

表 1 专利标题与摘要数据 jieba 分词结果

专利编号	分词结果									
1	平行	驱动	线路	平行	线路	传感器	平行	驱动	线路	...
2	恒温器	中央	电子	显示屏	方式	用户	温度	设置	图形	...
3	模组	模塑	感光	组件	方法	模塑	感光	组件	电路板	...
4	光固化	树脂	树脂层	图像	部件	表面	透光性	部件	制作	...
5	信号	方法	电子	设备	保护套	电子	设备 .	保护套	供电	...
6	电极	活性	材料	电极	电池	方法	电极	活性	材料	...
7	液晶显示	装置	显示装置	基板	传感	布线	基板	传感	布线	...
8	基板	聚四氟乙烯	介电	常数	填料	介电	常数	填料	介电	...
9	家用电器	设备	用户	界面	面板	区域	表面	平面	玻璃	...
...					...					

3.2 基于LDA模型的技术主题识别

3.2.1 最优主题数确定

在进行主题识别前，本文通过评估不同主题数模型的困惑度来选择最优的模型主题数。如图 5 困惑度图所示，在主题数增加后，困惑度也随之变化。根据手肘法，当 K 约等于 15 时，困惑度图像存在比较显著的拐点：当 K 小于 15 时，曲线单调递减且斜率较大；当 K 大于 15 时，曲线基本趋于平稳；所以 15 为 K 的最佳值，因此在本文所选定的语料集中，LDA 主题的生成数量选定为 15。

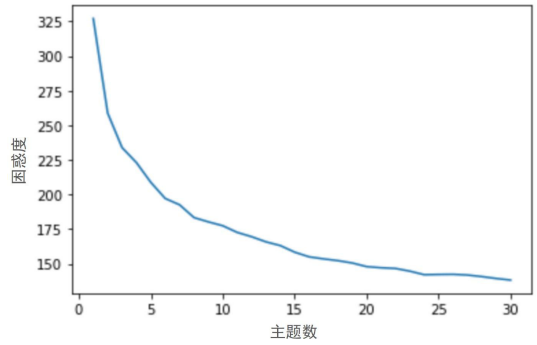

图 5 不同主题数下的困惑度

doi:10.3772/j.issn.2095-915x.2024.03.004

3.2.2 主题识别

本文使用 python 中的 sklearn 工具包进行主题识别工作。将已经分好词的专利数据导入 LDA 模型，提取出现频次排名前 1000 的特征词语，根据困惑度曲线将主题数量定为 15 个，由于数据量较大，可能会出现主题分类效果一般的高频特征词（如“智能”“组件”等），为保证识别效果，每个主题暂提取 25 个特征词，再根据主题特征进行筛选，可视化结果如图 6 所示，其中圆圈大小代表每个专利包含专利特征词数量的多少，圆圈越大则包含特征词越多，反之则越少。专利标题和摘要以及分词结果存入 EXCEL 结果文件，并输出每条专利数据对应概率最大的主题标号和每条数据对应各个主题的概率。各个主题中列取较为典型且具有代表性的特征词如表 2 所示。

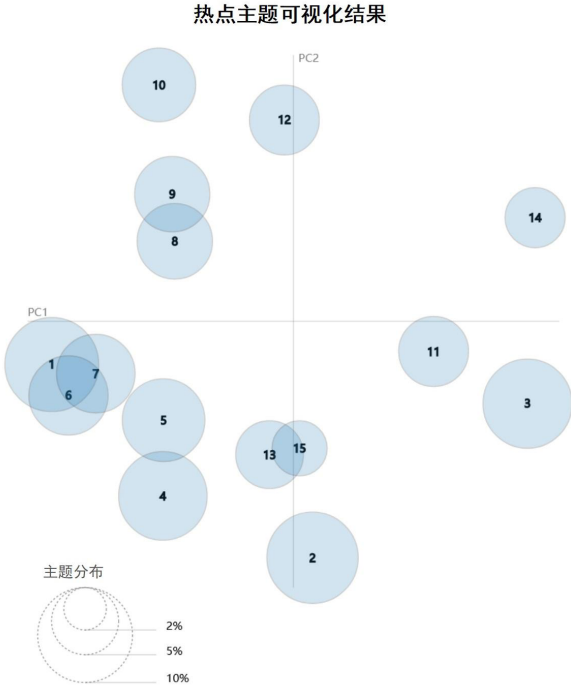

图 6 主题数为 15 时的热点主题聚类可视化结果

表 2 智能手机关键技术主题特征词

主题编号	具有代表性的特征词
主题 1	排列 亮度 分辨率 刷新 柔性 材质 视觉 密度 峰值 功耗 像素点 疏油层 光源 采样 表面
主题 2	协议 电池 电压 装置 状态 温度 电源 调节 散热 电流 供电 处理单元 电力 电量 指令
主题 3	图像 检测 目标 人脸 指纹 摄像头 场景 亮度 画面 短焦 特征 调整 对焦 投影 调节
主题 4	透镜 光学 测试 镜头 表面 折力 光轴 物侧 凸面 材质 传感器 焦度 焦距 镜片 非球面
主题 5	解调 电路 频率 解码 基带 射频 兼容 电波 数字 制式 网络 频段 传输 逻辑 节点
主题 6	频率 处理器 制程 核心 运算 指令 单元 架构 精度 能力 渲染 性能 集成 图形 控制
主题 7	像素 动态 光圈 场景 曝光 环境 画质 防抖 范围 面积 景深 计算 分辨率 插值 画面
主题 8	网络 无线 传输 网页 基站 信道 数据包 系统 链路 对象 地址 接入点 分配 协议 无线网络
主题 9	用户 车辆 屏幕 检测 运动 轨迹 距离 触摸屏 按键 车机 传感器 方向 环境 报警 车载
主题 10	存取 融合 程序 存储器 介质 分区 数据 半导体 系统 电容 组件 读取 稳态 刷新 地址
主题 11	参数 音频 页面 元素 耳机 装置 布局 扬声器 混合 效果 单位 音箱 系数 录音 音频文件
主题 12	终端 用户 指令 账号 密码 通话 体验 联系人 时间 号码 功能 系统 次数 短信 方式
主题 13	互联 管理 家居 服务 语音 用户 控制 指令 监控 感应 平台 智慧 社区 电动 物联网
主题 14	应用 文件 界面 图片 用户 指令 列表 目标 图标 功能 程序 代码 系统 软件 内存
主题 15	材料 体验 质量 玻璃 装置 方向 实体 外观 曲率 属性 按键 材质 盖板 测量 塑料

在 LDA 模型的主题识别结果中，将每篇专利所对应的公开日以时间顺序排列。通过计算各个主题近两年每个月的专利公开数量，可以评估各个主题专利研究的热度。ARIMA 模型结

果的准确性高度依赖于数据的可靠性和差异性，由于专利的申请和授权需要一定的时间，且专利研究的热度在近些年变化幅度明显，故本文选取近两年（2021 年 1 月至检索时间 2023 年 7 月）的专利研究的热度作为原始数据进行模型的拟合和预测，以月作为记录周期，刻画专利研究的热度随时间的变化趋势。

3.3 基于ARIMA模型的技术趋势预测

本文使用 python 中的 statsmodels 包，依次将各个技术主题的研究热度数据导入到 ARIMA（p,d,q）模型。

以主题 1 为例，首先判断数据是否具有平稳性以及是否为白噪声：原数据有单调递增趋势，并不具有平稳性，于是对数据进行差分处理，一阶差分的结果相较于原数据更稳定，使用 python 计算一阶差分的 ADF 检验结果 p 为 0.0225 ＜ 0.05（拒绝原假设），可知一阶差分序列不是非平稳序列，具有平稳性。一阶差分序列的白噪声检验结果 p 值为 0.0007 ＜ 0.05（拒绝原假设），可知一阶差分序列不是纯随机序列，是白噪声。故将模型参数中 d 的值定为 1。

然后确定 ARIMA 模型中参数 p、q 的值，本文利用自相关系数 ACF 和偏自相关系数 PACF 分别确定 p、q 的值，由图 7 可知自相关系数 ACF 在 1 阶后截尾，故将 p 值定为 1。偏自相关系数 PACF 拖尾，故将 q 的值定为 0。

接着检验 ARIMA 模型的残差是否是平均值为 0 且方差为常数的正态分布。利用 QQ 图检验，由图 7 可见结果为线性即为正态分布，故残差符合正态分布检验。使用 Durbin-Watson test（D-W）检验残差的自相关性，DW 值为 1.97，接近于 2 时，则不存在（一阶）自相关性。使用 Ljung-Box test（L-B）检验残差是否为白噪声，LB 值为 0.0039，则残差是白噪声。故残差符合检验，数据适用于 ARIMA 模型。最后对未来 6 个月的专利研究热度进行预测。

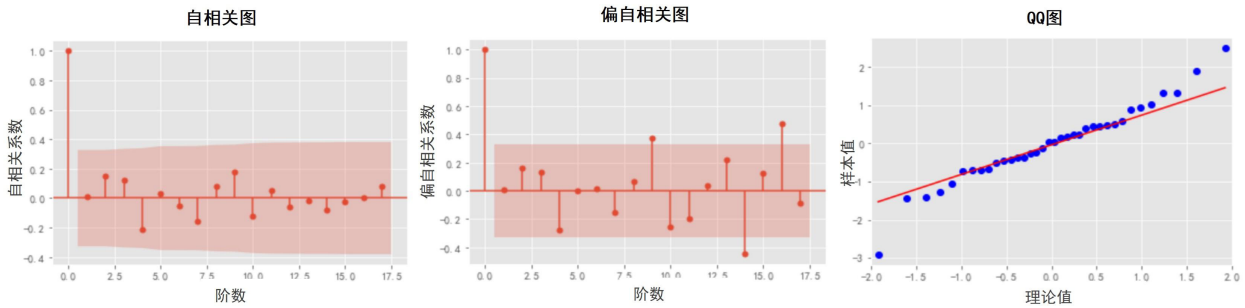

图 7　参数确定与检验结果

3.4 关键技术主题解读

根据表 2 可知主题 1 中的特征词“亮度”“视觉”等关注智能手机的屏幕显示；“分辨率”“像素点”“排列”对应智能手机屏幕相关参数，可以得知主题 1 主要对应的关键技术主题为“屏幕”。

主题 2 中的特征词“电压”“电流”“协议”等对应智能手机充电方面技术特征词；“散热”“功耗”“续航”等关注智能手机电池续航，可以得知主题 2 主要对应的关键技术主题为“电

doi:10.3772/j.issn.2095-915x.2024.03.004

池与充电”。

主题 3 中的特征词“指纹”“短焦”“对焦”则对应智能手机的屏下指纹解锁功能；“人脸”“摄像头”“特征”等与智能手机的前置摄像头相关功能有关，综合分析判断其对应的并非智能手机摄像技术，而是人脸识别解锁技术。可以得知主题 3 主要对应的关键技术主题为“生物识别系统”。

主题 4 中的特征词“透镜”“光学”“镜头”“传感器”“焦距”“镜片”等可以很明显发现该主题与摄影摄像相关，但由于在主题 3 的判断中也曾出现摄影摄像相关技术主题词，而摄影摄像覆盖的技术面较为广泛（包括拍照算法，摄像头传感器硬件等），故将主题 4 暂定为摄影摄像相关技术主题。

主题 5 中的特征词“解调”“基带”“解码”等对应智能手机基带技术特征词；“电波”“网络”“频段”“射频”等对应智能手机射频相关技术。两方面都为智能手机主板中通信功能的关键组成部分，综合考虑可以得知主题 5 主要对应的关键技术主题为“通信”。

主题 6 中的特征词“制程”“核心”“运算”等对应智能手机 CPU；“渲染”“集成”“图形”“性能”等对应智能手机 GPU。可以得知主题 6 主要对应的关键技术主题为“CPU 和 GPU”。

主题 7 中的特征词“光圈”“曝光”“环境”“画质”“防抖”“景深”“计算”等同样与摄影摄像相关。与主题 4 不同的是，主题 7 的特征词更多聚焦于摄影摄像的软件算法；主题 4 的特征词更多关注摄影摄像的硬件参数。故得知主题 7 主要对应的关键技术主题为“摄影摄像算法”；主题 4 主要对应的关键技术主题为“摄影摄像硬件”。

主题 8 中的特征词“传输”“基站”“信道”“链路”“地址”等与主题 5 类似，但频繁出现“网络”“无线”“无线网络”等特征词。可以得知主题 8 主要对应的关键技术主题为“网络”。

主题 9 中的特征词“车辆”“车机”“车载”与汽车关联紧密；同时根据特征词“传感器”“检测”“方向”“环境”等可以得知主题 9 主要对应的关键技术主题为“手机与车机互联”。

主题 10 中的特征词“存取”“融合”“存储器”“电容”“介质”“半导体”等多次提到存储，智能手机中的存储器与计算机类似，分为随机存储器（RAM）和只读存储器（ROM），故主题 10 主要对应的关键技术主题为“RAM 和 ROM”。

主题 11 中的特征词“参数”“扬声器”“耳机”“音箱”“录音”“音频文件”等与智能手机各项音频相关，故主题 11 主要对应的关键技术主题为“音频”。

主题 12 中的特征词“用户”“指令”“体验”等对应智能手机用户操作；“账号”“密码”“系统”等对应智能手机账户功能，可以得知主题 12 主要对应的关键技术主题为“用户与账户”。

主题 13 中的特征词“用户”“控制”“管理”等同样对应智能手机用户操作；而“家居”“服务”“互联”等对应的是家居服务，综合推断可以得知主题 13 主要对应的关键技术主题为“智能家居”。

主题 14 中的特征词“应用程序”“界面”“图片”“图标”等对应智能手机 UI 样式；“程序”“代码”“软件”“系统”等对应智能手机操作系统软件代码，可以得知主题 14 主要对应的关键

技术主题为“操作系统”。

主题 15 中的特征词“材料”“玻璃”“实体”“外观”“曲率”“按键”“材质”“盖板”“塑料”等对应智能手机材质属性以及外观，可以得知主题 15 主要对应的关键技术主题为“外观设计”。

各个主题编号与对应关键技术主题名称如表 3 所示。

3.5 关键技术主题演化分析

由 ARIMA 模型可以得到未来半年内的专利研究热度预测结果如图 8 所示，本文对包含特征词数量前三的主题（“屏幕”“电池与充电”“生物识别系统”）进行演化分析。

表 3 主题编号对应关键技术主题名称

主题编号	关键技术主题名称	主题编号	关键技术主题名称
主题 1	屏幕	主题 9	手机与车机互联
主题 2	电池与充电	主题 10	RAM 和 ROM
主题 3	生物识别系统	主题 11	音频
主题 4	摄影摄像硬件	主题 12	用户与账户
主题 5	通信	主题 13	智能家居
主题 6	CPU 和 GPU	主题 14	操作系统
主题 7	摄影摄像算法	主题 15	外观设计
主题 8	网络		

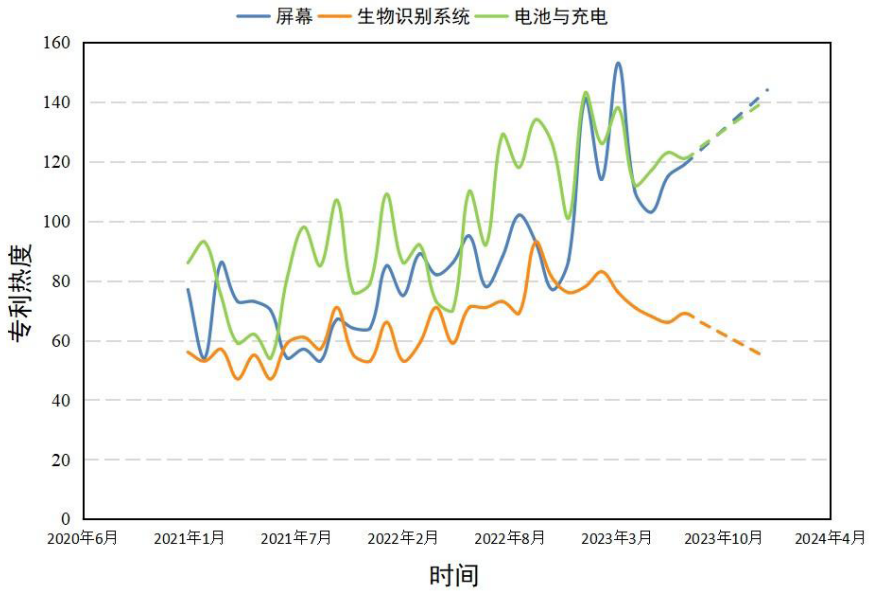

图 8 智能手机专利研究热度预测

3.5.1 智能手机屏幕技术

根据主题“屏幕”专利研究热度进行的预测结果，可以发现智能手机屏幕相关专利在 2021 年至 2022 年上半年间保持平稳，在 2022 年下半年持续增长，从 2023 年开始热度陡然上升，专利数量几乎较 2021 年初翻倍，年底下滑后再次回暖。ARIMA 预测其趋势为逐步上升，虽然上升幅度不如 2023 年大，但是仍保持较高的增长率，分析其原因如下：

智能手机屏幕材质以 OLED（Orangic Light Emitting Diode）[24] 居多，中文名称为有机发光二极管。以往的各家手机厂商的高端旗舰机屏幕供应商基本均为三星，三星屏幕凭借其峰值亮度高、分辨率高、画面清晰、色彩艳丽饱和

doi:10.3772/j.issn.2095-915x.2024.03.004

等优点获得了国内手机厂商的青睐。

国产智能手机屏幕供应商例如京东方、华星光电、天马、维信诺等近年来进步巨大，分别研发了蓝钻、珍珠、风车等屏幕类钻石排列，达到了较高的等效 ppi，在清晰度上与三星 Pentile 排列屏幕接近。在材料方面，我国智能手机屏幕辅助材料国内市场供应占 12%，发光层材料占 5%，关键材料依赖进口。今年来，柯达、UDC 等海外屏幕材料供应商的专利陆续到期，国内对于磷光金属配合物和关键配体等环节正值专利突破的时机。在屏幕制造设备方面，关键设备蒸镀机被日本 Tokki 垄断，高端蒸镀机几乎被三星买断，近年来我国合肥欣奕华首台硅基 OLED 蒸镀机导入客户量生产线，攻克了蒸镀机技术，我国蒸镀工艺蓬勃发展。

综上所述，国产屏幕在材料与设备面临专利发展的机遇，应加大研发攻克难题；在屏幕素质方面接近三星高端屏幕，其他方面例如峰值亮度、功耗等与三星屏幕存在差距，有较大进步空间，另外国内供应商研发了高频 PWM 技术，在护眼程度方面实现了超越，前景光明。

3.5.2 智能手机生物识别系统技术

根据主题“生物识别系统”专利研究热度进行的预测结果，可以发现智能手机生物识别系统相关专利研究热度自 2021 年起保持平稳，并且有缓慢增长的趋势，直至 2022 年底以较快速度提升，从 2023 年开始下滑。ARIMA 预测其趋势为逐步下降，下降速度与 2023 年初保持一致，分析其原因如下：

生物识别技术 [25] 是指通过计算机与光学、声学、生物传感器和生物统计学原理等高科技手段结合，利用人体固有的生理特征（指纹、人脸、虹膜等）和行为特征（笔迹、声音等）进行个人身份鉴定，其中指纹与人脸识别技术常用在智能手机领域。早在 2013 年，苹果公司就在 Iphone5s 上配备了实体指纹识别 TouchID 技术，并且在 2017 全面取消指纹识别，改用更为先进的 FaceID 人脸识别。

近年来部分国内厂商在指纹识别上实现弯道超车：vivo 研发的广域超声波指纹识别技术能实现在湿手或有污渍的情况下，解锁场景十分广泛：利用超声波穿透屏幕面板扫描捕捉手指纹理，由手指不同纹理材质返回超声波的不同时间来计算距离，从而建立对应指纹的 3D 图像来实现指纹识别。在人脸识别方面，华为于 2019 年开始自研 3D 人脸硬件、结构光模组、人脸识别算法，在需要解锁时，红外相机、散斑投射器和红外补光灯捕捉人脸数据，与机主人脸信息对比进行解锁，并且应用于此后发布的旗舰机中。

综上所述，国内手机厂商仅有 vivo 与华为分别在人脸与指纹识别技术上实现了突破，但其他厂商却与其有较大差距。生物识别系统的好坏在消费者的日常感知中十分明显，应该得到各企业的充分重视。ARIMA 模型预测其专利研究热度有降低的趋势，其他企业应该警惕，需要紧跟技术发展趋势，加大研发力度，努力引进 vivo、华为等先行者的相关技术，并进行创新发展。

3.5.3 智能手机电池与充电技术

根据主题“电池与充电”专利研究热度进行的预测结果，可以发现智能手机电池与充电技术相关专利研究热度在 2021 至今起伏较大，但总体趋势是持续上升的，其中 2022 年间增长

率较高，到 2023 年后增长变缓。ARIMA 预测其趋势为逐步上升，仍然维持较高增长率，分析其原因如下：

国内智能手机的充电和电池技术相较于国际大厂（例如三星、苹果）已是遥遥领先。三星在 2016 年发布的智能手机 Samsung Galaxy Note 7 就曾出现过严重的安全隐患，该机型发布一个多月，就已在全球范围内出现了近四十起因电池爆炸而发生的安全事故；苹果手机则因为其低功率充电、电池容量小而被贴上了“充电慢、续航差”的标签。

近年来国内厂商将充电与电池技术作为研发重点，并取得了成效：充电方面，Realme 开发 240W 充电架构，通过定制的三路电荷泵，将高功率充电所带来的高热量进行分散，并引进了 4 ： 2 大功率充电架构，将大电流分解成三路小电流进行充电，降低了充电通路的阻抗，提升了整体的转化效率，同时也减少了热量的堆积；Redmi 全链路自研了 300W 快充架构，定制了 6 ： 2 电荷泵芯片，芯片最高转换效率高达 98%，多颗电荷泵并联后，直接对电池充电。电池方面，Redmi 研发 15C 高倍率电池，使用三明治双电芯堆叠，引入了新型硬碳材料，相比以前的石墨要更加疏松无序，提升充电速度的同时还能兼顾高能量密度；荣耀研发了硅碳负极电池技术，在相同电池体积的情况下提供了更大的电池容量，实现了旗舰手机电池容量的突破。

综上所述，高功率充电技术方面国内的研发已趋于饱和，相关企业可以从温度控制、安全性方面着手突破；在电池方面，锂电池几乎达到化学极限，能量密度提升比较困难，各企业可以朝着新的电池形态（例如固态电池）寻找突破口。未来充电与电池技术将成为国内智能手机的卖点，各厂商应努力维护目前国内充电与电池技术的领先地位。

4 结论

本文通过实证研究确定智能手机关键技术主题分别为屏幕、电池与充电、生物识别系统、摄影摄像硬件、通信、CPU 和 GPU、摄影摄像算法、网络、手机与车机互联、RAM 和 ROM、音频、用户与账户、智能家居、操作系统、外观设计，并根据演化趋势分析提出建议。有助于企业把握技术发展趋势，及时调整战略。

然而，本文仅使用了国内智能手机专利摘要数据进行分析，对于其余重要数据例如论文、舆情数据等并未进行分析；另外对于专利研究热度仅使用各月份的专利公开量作为参考也不够严谨。在未来的研究中将扩大数据源，对国内专利、论文、舆情数据综合分析，并结合国外各项数据分析结果进行对比参考，构建专利研究热度指标进行计算，提高 ARIMA 模型预测的准确性，对智能手机行业起到更好的导向和促进发展的作用。

参 考 文 献

[1] 滕飞，张奇，曲建升，等 . 基于专利竞争力指数和 Doc-LDA 主题模型的关键核心技术识别研究 - 以新能源汽车为例 [J/OL]. 数据分析与知识发 现 ，1-19[2024-04-26]. http://kns.cnki.net/kcms/detail/10.1478.g2.20231118.1750.002.html.

[2] YU C C, CAO L, YIN W B, et al. Automatic word segmentation on Lizu spoken annotation corpus[J]. Application Research of Computers, 2017, 34(5): 1325-1328.

doi:10.3772/j.issn.2095-915x.2024.03.004

[3] 陈稳 , 陈伟 . 科学与技术对比视角下的前沿主题识别与演化分析 [J]. 情报杂志 , 2022, 41(1): 67-73, 163.

[4] 吕鲲 , 陈箫羽 , 靖继鹏 . 基于组合分词方法和 LDA 模型的区块链金融产业关键技术识别研究 [J]. 图书情报工作 , 2022, 66(19): 110-121.

[5] LIU Z Y, WANG Y, WANG W P, et al. Bayesian Parameter Estimation in LDA:[C]//International Conference on Computer Information Systems and Industrial Applications. Bangkok, Thailand, 2015: 837-840.

[6] 郑荣 , 魏明珠 , 高志豪 , 等 . 基于 SCAN-CPM 的产业新兴技术识别与演化路径分析：以新能源汽车产业为例 [J]. 图书情报工作 , 2022, 66(11): 100-109.

[7] 伊惠芳 , 刘细文 , 龙艺璇 . 基于多文本分析的领域关键技术问题挖掘方法研究 [J]. 情报理论与实践 , 2023, 46(1): 187-196.

[8] XU T T, HUANG H J. An Improved Supervised-LDA Text Model and Its Application[J]. Computer Engineering, 2018, 44(1): 69-73, 78.

[9] 王秀红 , 高敏 . 基于 BERT-LDA 的关键技术识别方法及其实证研究——以农业机器人为例 [J]. 图书情报工作 , 2021, 65(22): 114-125.

[10] DE ARAÚJO MORAIS L R, DA SILVA GOMES G S. Forecasting daily Covid-19 cases in the world with a hybrid ARIMA and neural network model[J]. Applied Soft Computing, 2022, 126: 109315.

[11] 马铭 , 王超 , 周勇 , 等 . 基于语义信息的核心技术主题识别与演化趋势分析方法研究 [J]. 情报理论与实践 , 2021, 44(9): 106-113.

[12] BIAN Z L, TANG J Q, NI C H, et al. Analysis on prevalence of pneumoconiosis in Jiangsu Province using ARIMA-GRNN combined model[J]. Chinese Journal of Environmental & Occupational Medicine, 2019, 36(8): 755-760.

[13] LV L C, ZHOU J, WANG X Z, et al. Technology Evolution Analysis Framework Based on Two-Layer Topic Model and Application[J]. Data Analysis and Knowledge Discovery, 2022, 6(2/3): 18-32.

[14] 岳丽欣 , 周晓英 , 陈旖旎 . 基于 ARIMA 模型的信息构建研究主题趋势预测研究 [J]. 图书情报知识 , 2019(5): 54-63, 72.

[15] HU R, MA W, LIN W, et al. Technology Topic Identification and Trend Prediction of New Energy Vehicle Using LDA Modeling[J]. COMPLEXITY, 2022, 2022: 1-20.

[16] 李勇 , 欧志梅 , 黄格 , 等 . 基于专利挖掘的数字文旅关键技术识别和趋势分析 [J]. 图书馆论坛 , 2023, 43(1): 121-130.

[17] SUN B, XU X F, SU X. Identification of the Main Path of Technology Diffusion and Core Enterprises: A Case Study of Patent Citation Network in Mobile Phone Chips[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(2): 201-208.

[18] WANG X W, SONG Y Y. Artificial Intelligence Scientific Management: Artificial Intelligence Patents Layout in the Smartphone Field[J]. Science & Technology for Development, 2018, 14(3): 142-148.

[19] 苏敬勤 , 刘建华 , 王智琦 , 等 . 颠覆性技术的演化轨迹及早期识别——以智能手机等技术为例 [J]. 科研管理 , 2016, 37(3): 13-20.

[20] 闯家梁 . 基于专利分析的中国智能手机企业技术创新能力研究 [J]. 中国集体经济 , 2020(14): 153-154.

[21] 彭英 , 闯家梁 , 杨映雪 . 专利视阈下石墨烯智能手机技术创新研究 [J]. 电子商务 , 2019(10): 42-43, 65.

[22] HASAN M, RAHMAN A, KARIM MD R, et al. Normalized approach to find optimal number of topics in latent dirichlet allocation (lda)[C]//2nd International Conference on Trends in Computational and Cognitive. Savar, Bangladesh, 2021: 341-354.

[23] HUANG L, MA J, CHEN C, et al. Topic Detection from Microblogs Using T-LDA and Perplexity[C]//24th Asia-Pacific Software Engineering Conference. Nanjing, PEOPLES R CHINA. 2017: 71-77.

[24] LIU D X, ZHONG J Y, TANG B, et al. Research progress of flexible and printed OLED[J]. Chinese Journal of Liquid Crystals and Displays, 2021, 36(2): 217-228.

[25] AMIRTHALINGAM G, GOVINDARAJU R. Comparison on Multi-modal Biometric Recognition Method[C]//1st International Conference on Computational Intelligence and Information Technology. Pune, INDIA, 2011: 524-526.

doi:10.3772/j.issn.2095-915x.2024.03.005

基于 ISM-MICMAC 的数据资产价值影响因素研究

开放科学（资源服务）标识码（OSID）

刘枏[1]　周雨[1]　郝雪镜[2]

1. 重庆交通大学经济与管理学院　重庆 400074；
2. 重庆市信息通讯咨询设计有限公司　重庆 400041

摘要：[目的 / 意义] 通过深入剖析数据资产价值的研究现状，指出目前对数据资产价值的定义和理解还比较模糊和笼统。[方法 / 过程] 为了更好地理解和评估数据资产的价值，采用文献调研、专家调查和问卷调查等方法，提取出 57 个影响数据资产价值的因素，并通过因子分析得出 13 个主要因素，通过建立解释结构模型，分析因素之间的层级关系和作用机理。同时，还使用 MICMAC 方法对因素进行驱动力—依赖性分析，实现因素的进一步分层和分类。最后，针对研究得出的主要影响因素，从表层、具体层和关键层三个层面提出了与国家、企业、用户等主体相关的具体对策建议，以促进数据资产价值的实现和提高。[局限] 问卷调查以及专家评估打分存在一定的主观性。同时，数据资产价值影响因素指标也具有一定的可拓展性。[结果 / 结论] 研究结果表明：信息提取要素投入价值因素、数据采集要素投入价值因素以及数据组织管理价值因素为关键层因素，属于独立因素群，具有较高的驱动力和低依赖性，对其他因素发挥支撑和推动作用。

关键词：数据资产价值；影响因素；解释结构模型；MICMAC 分析

中图分类号：G35

Research on Influencing Factors of Data Asset Value Based on ISM-MICMAC

LIU Nan[1]　ZHOU Yu[1]　HAO Xuejing[2]

1. School of Economics and Management, Chongqing Jiaotong University, Chongqing 400074, China;
2. Chongqing Information Communication Consulting and Design Co., LTD., Chongqing 400041, China

Abstract: [Objective/Significance] This paper introduces the research status of data asset value and points out that the definition and understanding of data asset value are still vague and general. [Methods/Processes] In order to better understand and evaluate the value of data assets, 57 factors affecting the value of data assets were extracted by literature research, expert survey and

基金项目　教育部人文社会科学基金项目“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-），博士，教授，主要研究方向为数据资产、数据分析、工程管理信息化，E-mail：liunan@cqjtu.edu.cn；周雨（1999-），硕士研究生，主要研究方向为数据资产、政府数据；郝雪镜（1997-），硕士研究生，主要研究方向为数据资产。

引用格式　刘枏 , 周雨 , 郝雪镜 . 基于 ISM-MICMAC 的数据资产价值影响因素研究 [J]. 情报工程 , 2024, 10(3): 63-76.

doi:10.3772/j.issn.2095-915x.2024.03.005

questionnaire survey, and 13 major factors were obtained through factor analysis. The hierarchical relationship and mechanism of action among factors were analyzed by establishing an explanatory structure model. At the same time, MICMAC method is also used for driving force-dependency analysis of factors to achieve further stratification and classification of factors. Finally, based on the main influencing factors obtained from the research, this paper puts forward specific countermeasures and suggestions related to the country, enterprises, users and other subjects from the surface level, concrete level and key level, so as to promote the realization and improvement of the value of data assets. [Limitations] Questionnaire survey and expert evaluation have some subjectivity. At the same time, the index of influencing factors of data asset value also has certain extensibility. [Results/Conclusions] The results show that the input value factor of information extraction, the input value factor of data acquisition and the value factor of data organization and management are the key factors, belonging to the independent factor group, with high driving force and low dependence, and playing a supporting and promoting role for other factors.

Keywords: Data Asset Value; Identification of Influencing Factors; ISM; MICMAC

引言

随着人工智能、物联网、云计算等技术的发展，全球数据量正以指数型增长。据国际数据公司（IDC）预测，到 2025 年，全球数据量将增至 175ZB。与此同时，数据经济应运而生，大数据正成为数据时代的重要财富[1-2]。2011 年，IDC 和麦肯锡研究院对数据的潜力和关键技术等进行分析，指出数据将成为企业的关键竞争力。为了激活和释放数据价值，美国政府于 2012 年启动“大数据研发计划”[3]。2021 年《“十四五”大数据产业发展规划》提出技术与制度双向引导，加快实现大数据产业发展和数据要素价值释放。2022 年《中共中央国务院关于构建数据基础制度更好发挥数据要素作用的意见》提出搭建数据基础制度体系的“四梁八柱”。目前数据的资产化和数据资产价值受到广泛关注，但对于数据价值的影响因素的研究还不够深入。因此，亟待对数据资产价值的影响因素进行研究，以促进数据资源流通、合理分配以及数据价值释放。

1 文献综述

Gargano 等[4]最早指出可以通过数据挖掘技术来发现数据资产中的隐藏价值，自此，数据具有价值这一提法便逐渐受到国内外研究者的广泛认同。通过文献梳理，发现数据资产价值主要包含了账面价值、市场价值、经济价值以及社会价值四个部分。数据价值在企业中直观的体现是财务报表中日趋扩大的“账面价值”和“市场价值”之间的差距[5]。Heckman 等[6]指出数据可以作为一种资产在市场中进行交易，获取交换价值，体现数据资产的市场价值。目前已有不少研究从数据定价的角度分析了数据的市场价值，包括定价模型和交易策略等。陈筱贞[7]研究了数据交易的市场类型和定价，提出生成级别、信息领域以及应用端用途三类数据价格决定因素。但对于数据的消费者来说，他们更加关心的是通过利用数据改善决策或研发产品等为组织带来的经济效益，即数据资产的经济价值[8]。此外，数据资产的价值很大程度上取决于数据的使用对象和使用方式，一方

面决策者对数据的质量具有主观影响[9]。另一方面，数据价值对应用场景的依赖性很高，同一数据会被应用到不同的场景并产生不同的价值，因此数据的价值便不是由成本所决定的，而是由需求方的使用价值决定的[10]。从数据要素的视角来看，也应该注重数据的有效供给，以用户需求为核心构建数据价值实现体系[11]。数据资产价值的影响因素相关研究从数据质量、数据应用和数据风险三个维度[12]，增加了成本和数量两个维度[13]。李永红等[14]则将运营成本、可比资产市值、未来收益等影响因素考虑进互联网企业数据资产价值评估。同时，数据资产所具有的价值增量特征和大数据特征也被转化为价值影响因素，加入数据资产价值评估中[15]。

IAS（国际会计准则）对于资产的定义是：由于过去事项而由企业控制的、预期会导致未来经济利益流入企业的资源[16]。根据这一准则，可将数据资产定义为企业拥有或者控制的、能够为企业带来未来经济利益的，可读取的网络空间中的数据集资源。目前，有关信息价值的研究较为丰富，专门针对数据资产价值特征的研究较少，而数据是信息的原材料，对信息价值特征的描述一定程度适用于数据价值，同时结合数据与资产两者的价值属性，将数据资产的价值定义为数据资产所蕴含的潜在价值或实际收益。

尽管数据资产的价值已获得广泛认可，对其价值的构成和影响因素也有所研究，但仍存在不足：目前为止，对影响数据资产价值的因素相关研究只涉及某一片面维度，如成本、效用等，缺乏系统全面的数据资产价值评估指标体系；其次，对维度中涉及的各因素，通常只考虑各自独立的影响，缺乏因素间相互作用的分析，体系中各因素较为孤立。因此需要建立一个全面综合的数据资产价值影响因素指标体系，以促进对数据价值的准确量化与评估。

2 主要影响因素识别

论文的研究对象为数据资产价值影响因素，在综合考虑识别方法适用性与独特性的基础上，紧密结合数据资产价值的特性，采用了文献研究、专家咨询、问卷调查以及因子分析 4 种方法对主要影响因素进行识别。主要影响因素的识别过程大致可以分为两步：第一步基于文献研究法和专家调查法识别出与数据资产价值有关的全部影响因素；第二步基于问卷调查法和因子分析法对影响因素进行降维处理，提炼出主要影响因素。

2.1 确定影响因素清单

首先，通过在知网、维普、万方、Google Scholar、Web of science、Science Direct、Springer Link、Engineering Village 等数据库中检索“数据资产”“数据资产价值”“数据资产价值影响因素”“数据价值影响因素”等关键词，搜集符合要求的重要文献，进行分类整理和阅读，并对数据资产价值影响因素进行提取和总结。其次，将因素进行整理交由 20 位专家实施审核判定，对所统计的因素进行取舍和修改，剔除专家累计选择率低于 80% 的因素。最后，根据文献分析和专家评判识别出 57 个大数据资产价值影响因素，具体影响因素指标如表 1 所示。

doi:10.3772/j.issn.2095-915x.2024.03.005

表 1　基于文献检索的数据资产价值影响因素识别

编号	影响因素	来源	编号	影响因素	来源
V1	数据收集成本投入	[17]、[18]	V30	数据服务价值	[25]
V2	数据处理成本投入	[17]、[18]	V31	数据到达率	[28]
V3	数据储存成本投入	[17]、[18]	V32	数据完整性	[18]、[22]
V4	数据运维成本投入	[17]、[18]	V33	数据准确性	[18]、[22]
V5	数据分析成本投入	[19]	V34	数据相关性	[18]、[20]
V6	数据服务成本投入	[6]	V35	数据及时性	[22]、[27]
V7	数据收集技术的先进性	[20]	V36	数据客观性	[20]、[27]
V8	数据处理技术的先进性	[20]	V37	数据安全性	[20]、[27]
V9	数据储存技术的先进性	[20]	V38	数据可验证性	[18]、[20]
V10	数据分析技术的先进性	[20]	V39	数据关联性	[29]
V11	数据可视化技术的先进性	[20]	V40	数据定义一致性	[20]、[21]
V12	数据分析人员专业技能	[20]	V41	数据时间一致性	[28]
V13	跨学科团队的建立与完善	[20]	V42	数据规范性	[18]、[20]
V14	数据供应商的企业信誉	[17]、[21]	V43	数据代表性	[20]
V15	元数据架构	[18]	V44	数据可访问性	[27]
V16	数据质量标准	[18]	V45	数据可读性	[20]、[27]
V17	数据使用规则	[20]	V46	数据易操作性	[20]
V18	数据供应商跨部门协调工作能力	[20]	V47	数据寿命	[17]
V19	数据使用者主观性和判断能力	[22]	V48	数据的信息密度含量	[20]
V20	数据重复使用率	[23]	V49	数据体量	[27]
V21	数据广泛使用程度	[18]	V50	数据维度	[30]
V22	数据使用场景匹配性	[20]、[24]	V51	数据深度	[30]
V23	数据用于战略决策的效果	[26]	V52	数据类型	[18]、[20]
V24	数据用于供应链管理的效果	[25]	V53	数据品种	[17]
V25	数据用于优化销售规划的效果	[25]、[26]	V54	大数据的相关政策及规划	[24]
V26	数据用于企业内部运营管理的效果	[25]、[26]	V55	数据的稀缺程度	[17]
V27	使用数据的负外部性	[22]	V56	数据的权属划分	[31]
V28	使用数据的正外部性	[14]	V57	相似案例历史情况	[20]
V29	数据的品牌效应	[32]			

2.2　提取主要影响因素

2.2.1　问卷数据收集

基于表 1 归纳总结的影响因素清单，制作调查问卷，量表主体部分采用内部一致性较高的五级李克特量表方式进行填答，问卷分为两个部分，第一部分是被调查者的基本信息，主要包括学历、所在单位的性质、从事大数据相关研究的年限以及从事大数据相关研究的类型；第二部分是被调查者对价值因素影响程度测量量表，这部分问题考察调查对象对数据资产价值的影响因素的影响程度评判，共分为 5 个等级，从“1”到“5”五个数字分别表示影响程度由低到高，“1”表示影响程度极小，“2”表示影响程度较小，“3”表示影响程度一般，“4”表示影响程度较大，“5”表示影响程度极大。

为保证问卷选项结构合理且易理解，该问

卷首先实施了小规模的预发放，选取 60 位对数据价值有所研究的老师和同学作为调查对象发放问卷，信度检验结果为 0.810，信度较好，确定以此作正式问卷。

本次问卷调查的对象是大数据交易行业相关的从业人员，大数据软硬件开发从业人员，以及高校科研机构的大数据研究者等。问卷调查针对大数据交易平台、大数据服务供应商、大数据软硬件开发企业、电子商务等数据密集型企业，以及高校等进行。问卷主要以线上形式进行，在专业数据分析平台 Credamo（见数）上进行发放回收，共发问卷 260 份，回收问卷 260 份，回收率 100%，对问卷进行人工筛选，剔除有明显错误的无效问卷 17 份，得到有效问卷 243 份，有效回收率为 93.5%。针对学历、企业类型、研究类型和研究年限对样本进行描述性统计分析，如表 2 所示。

2.2.2 信度、效度检验

在信度方面，论文采用 Cronbach α 系数进行内部一致性检验，经 SPSS 计算得出问卷数据的 Cronbach α 系数为 0.876，这表明所设计的调查问卷具有较高的稳定性和可靠性，如表 3 所示。

效度包括内容效度、标校效度和建构效度。本研究的量表题项来自较为细致的文献调研，并且经过严格的专家审核，认为具有良好的内容效度和标校效度。基于此，以因素分析检验构建效度，通过取样适当性数值（Kaiser-Meyer-Olkin measure of sampling adequacy，KMO）的大小和 Bartlett 球形检验来判别原始因素指标变量是否适合进行因素分析，如表 4 所示。结果可以看出，Bartlett 的值为 8499.141，P 值为 0.000 小于 0.001，表明单位矩阵和相关关系矩阵之间的差异显著，相关矩阵之间有共同因素存在，与此同时 KMO 值为 0.755，大于 0.7，说明本次问卷搜集的数据是比较适合因子分析的。

表 2 问卷统计分析表

特征	分类	选择数（个）	百分比（%）
最高学历	专科及以下	15	6.4
	本科	62	26.4
	硕士	155	66
	博士	3	1.3
企业类型	政府机构	11	4.7
	行业协会	1	0.4
	高校及科研院所	155	65.7
	企业	63	26.7
	其他	6	2.5
研究类型	数据分析	192	28.2
	数据交易	145	21.3
	数据安全	101	14.8
	云存储	30	4.4
	云计算	48	7
	数据硬件	15	2.2
	数据软件	104	15.3
	大数据整体解决方案	46	6.8
研究年限	一年以内	119	50.6
	1-3 年	98	41.7
	4-5 年	15	6.4
	6-10 年	2	0.9
	10 年以上	1	0.4

表 3 信度分析

Cronbach 的 α	项目个数
0.876	57

doi:10.3772/j.issn.2095-915x.2024.03.005

表 4 KMO 和巴特利特检验

Kaiser-Mayer-Olkin 测量取样适当性		0.755
Bartlett 球形度检验	近似卡方	8499.141
	自由度	1596
	显著性	0.000

2.2.3 提取特征向量和特征值

本文采取主成分法提取公因子，通常将相关系数矩阵 R 的特征值大于 1 的因子作为公因子，且方差累计贡献率要求至少达到 50%。进一步通过方差最大化正交旋转法得到旋转后的因子荷载量和因子荷载矩阵，将在两个或更多因子上载荷均大于 0.5 和在任何单一因子上载荷不大于 0.5 的两类题项予以剔除，并再次对问卷数据进行信度、效度检验。

在经过剔除数据寿命（V47）、数据储存技术的先进性（V9）、数据易操作性（V46）和数据达到率（V31）共三次指标调整后，结合相关性系数矩阵得出因子的特征值和方差贡献率，如表 5 所示，可以看出共有 13 个公因子的特征值大于 1，且累计贡献率为 69.832%，满足贡献率的要求，表示 53 个原始变量的 69.832% 的本质信息可以被这 13 个公因子所代表。

表 5 方差解释表

公因子	初始特征值			提取载荷平方和			旋转载荷平方和		
	总计	方差百分比 (%)	累积 (%)	总计	方差百分比 (%)	累积 (%)	总计	方差百分比 (%)	累积 (%)
VF1	8.311	15.682	15.682	8.311	15.682	15.682	3.520	6.641	6.641
VF2	5.599	10.564	26.246	5.599	10.564	26.246	3.062	5.777	12.418
VF3	3.640	6.868	33.115	3.640	6.868	33.115	3.036	5.729	18.147
VF4	3.352	6.325	39.440	3.352	6.325	39.440	2.990	5.642	23.789
VF5	2.768	5.223	44.663	2.768	5.223	44.663	2.938	5.544	29.332
VF6	2.312	4.362	49.025	2.312	4.362	49.025	2.934	5.537	34.869
VF7	2.156	4.068	53.093	2.156	4.068	53.093	2.915	5.500	40.369
VF8	1.888	3.563	56.656	1.888	3.563	56.656	2.866	5.407	45.776
VF9	1.735	3.274	59.930	1.735	3.274	59.930	2.829	5.337	51.114
VF10	1.518	2.864	62.794	1.518	2.864	62.794	2.713	5.119	56.233
VF11	1.332	2.513	65.308	1.332	2.513	65.308	2.567	4.843	61.076
VF12	1.316	2.483	67.791	1.316	2.483	67.791	2.363	4.459	65.535
VF13	1.082	2.042	69.832	1.082	2.042	69.832	2.278	4.298	69.832

2.2.4 提取公因子命名

为了使提取的公因子能够更为准确地反映数据资产价值影响因素，本研究采用最大方差法对原始变量进行正交旋转，得到旋转后的因子荷载矩阵，然后根据各公因子具有高荷载指标对公因子进行命名和解释。具体如表 6 所示。

其中，数据合规性价值因素（VF1）反映的是数据在不同操作系统和不同主体之间进行流动时符合相应规范的程度，以及数据的使用是否遵守一定的规则；数据社会性价值因素（VF2）反映的是数据本身生产和使用之外的外部因素，主要表现在社会价值上；数据可用性价值因素（VF3）反映的是用户获取、运用、追溯数据的便捷化程度；信息提取要素投入价值因素（VF4）反映的是数据处理、分析、可视化阶段的资金、人才和技术投入对数据价值

表 6　对公共因子的命名和方差解释

公因子	显著影响因素变量	公因子命名	方差解释 (%)
VF1	V40、V41、V42、V17、V16、V37、V15	数据合规性价值因素	6.641
VF2	V18、V28、V54、V27	数据社会性价值因素	12.418
VF3	V38、V44、V45	数据可用性价值因素	18.147
VF4	V11、V10、V2、V8、V5	信息提取要素投入价值因素	23.789
VF5	V7、V3、V1、V4	数据采集要素投入价值因素	29.332
VF6	V33、V39、V36、V35、V32	数据质量价值因素	34.869
VF7	V23、V24、V26、V25	数据应用效果价值因素	40.369
VF8	V29、V20、V21、V30	数据应用情感价值因素	45.776
VF9	V12、V18、V13	数据组织管理价值因素	51.114
VF10	V51、V49、V48、V50	数据规模性价值因素	56.233
VF11	V19、V22、V34	数据应用匹配性价值因素	61.076
VF12	V57、V56、V55	数据竞争性价值因素	65.535
VF13	V52、V53、V6	数据类别性价值因素	69.832

的影响；数据采集要素投入价值因素（VF5）反映的是数据收集、储存以及运维阶段的资本和技术投入对数据价值的影响程度；数据质量价值因素（VF6）反映的是数据本身内容相关的质量优劣；数据应用效果价值因素（VF7）反映的是用户应用数据产生的具体效果带来的经济价值；数据应用情感价值因素（VF8）反映的是用户在数据交易和应用中感受到除数据本身之外的附加情绪价值；数据组织管理价值因素（VF9）反映的是数据供应商发挥组织主观能动性对数据价值产生的影响；数据规模性价值因素（VF10）反映的是数据样本的跨度和覆盖面等；数据应用匹配性价值因素（VF11）反映的是数据具体的应用对象和应用场景对数据价值的影响；数据竞争性价值因素（VF12）反映的是数据在市场上获取的难易程度和不可替代性；数据类别性价值因素（VF13）反映的是主要供应商提供不同类的数据所带来的价值。

3　ISM 模型构建

3.1　ISM模型简介

ISM（Interpretatine Structural Modeling Method）即解释结构模型，是一种用于理解复杂问题的多标准决策 (MCDM) 方法，工作原理是将系统进行拆分，拆解成影响系统的单个因素，分析因素之间的相互影响关系，并将这种关系以矩阵表现出来，通过布尔逻辑运算，将各单因素分层，最后将分层结果通过有向拓扑图的方式进行展现。

3.2　建立邻接矩阵与可达矩阵

将确定的 13 个因素交由在数据价值研究领域有充足经验的 20 位行业专家进行判定，若因素 V_i 对 V_j 有影响，则关联矩阵中 a_{ij}=1，反之因素 V_i 对 V_j 没有影响，a_{ij}=0，得到 20 个能体现数据间相互影响关系的关联矩阵。

doi:10.3772/j.issn.2095-915x.2024.03.005

对问卷进行统计，若有 80% 以上的专家在因素的相关关系之间标注了“1”，则认为因素“V_i”对因素“V_j”有影响，则认定 a_{ij}=1，反之认定 a_{ij}=0，由此可建立邻接矩阵 A。

$$A=\begin{bmatrix}1&1&1&0&0&0&0&0&0&0&0&0&0\\0&1&0&0&0&0&0&1&0&0&0&0&0\\0&0&1&0&0&0&1&1&0&0&0&0&0\\1&0&1&1&0&1&0&0&1&1&0&1&0\\1&0&0&0&1&0&0&0&0&1&0&1&1\\0&1&0&0&0&1&1&1&0&0&0&0&0\\0&1&0&0&0&1&1&0&0&0&0&0&0\\0&1&0&0&0&0&0&1&0&0&0&0&0\\1&0&1&0&0&1&0&0&1&0&0&1&0\\0&1&0&0&0&0&1&0&0&1&0&0&0\\0&0&0&0&0&0&1&1&0&0&1&0&0\\0&1&0&0&0&0&1&1&0&0&0&1&0\\1&1&0&0&0&0&1&1&0&0&0&0&1\end{bmatrix}$$

在获得邻接矩阵的基础上，进行矩阵转化计算可达矩阵，矩阵的转换形式是将邻接矩阵与单位矩阵相加后进行布尔代数幂运算，当矩阵的幂运算迭代 K 次之后，矩阵中不再有新的“1”出现，此时出现了一个稳定的矩阵 $(A+I)^K$，对其再进行幂运算得到矩阵 $(A+I)^{K+1}$ 与 $(A+I)^K$ 相同，此时这个稳定的矩阵 $(A+I)^K$，便是我们所求的可达矩阵 M。

$$M=\begin{bmatrix}1&1&1&0&0&0&1&1&0&0&0&0&0\\0&1&0&0&0&0&0&1&0&0&0&0&0\\0&1&1&0&0&0&1&1&0&0&0&0&0\\1&1&1&1&0&1&1&1&1&1&0&1&0\\1&1&1&0&1&0&1&1&0&1&0&1&1\\0&1&0&0&0&1&1&1&0&0&0&0&0\\0&1&0&0&0&0&1&1&0&0&0&0&0\\0&1&0&0&0&0&0&1&0&0&0&0&0\\1&1&1&0&0&1&1&1&1&0&0&1&0\\0&1&0&0&0&0&1&1&0&1&0&0&0\\0&1&0&0&0&0&1&1&0&0&1&0&0\\0&1&0&0&0&0&1&1&0&0&0&1&0\\1&1&1&0&0&0&1&1&0&0&0&0&1\end{bmatrix}$$

3.3 绘制多级递阶结构图

进一步利用 MATLAB 进行编程，对可达矩阵进行结构划分，MATLAB 的层级划分输出结果如图 1 所示。

（1）L1、L2 层：该层的因素是较为表层的因素，对价值的实现起着最直接的决定性作用。包括数据社会性价值因素 (VF2)、数据应用情感价值因素 (VF8) 以及数据应用效果价值因素 (VF7)。这三个因素直观地体现着数据资产的最终价值，他们主要代表的是数据资产在实践应用阶段的效用价值以及客户感知价值等。

（2）L3、L4 层：该层是具体的影响因素，受到深层因素的影响同时影响着表层因素，具有过渡性质，因此也可命名为过渡层。该层的因素数量最多，主要涉及数据的合规性价值因素（VF1）、数据可用性价值因素（VF3）、数据内容质量价值因素（VF6）、数据规模性价值因素（VF10）、数据应用匹配性价值因素（VF11）和数据竞争性价值因素（VF12）。这些因素大多代表着数据本所具有的功能特性，数据资产价值主要围绕这些因素形成，这些因素对数据资产价值的产生具有十分重要的作用。

（3）L5、L6 层：该层因素则为深层次的影响因素，在数据资产价值形成的过程中常常被忽视，但这些因素是最为基础的根源因素，对过渡层和表层因素都有着直接或间接影响，对价值的形成起主导作用。尤其是信息提取要素投入价值因素（VF4）和数据采集要素投入价值因素（VF5），是数据资产价值形成的基础和核心，如果忽视在数据采集和挖掘过程中所需的资本、人才和技术等要素投入，那么数据也只能成为信息社会中散乱着的被“遗忘”的资源，而并不具有价值。

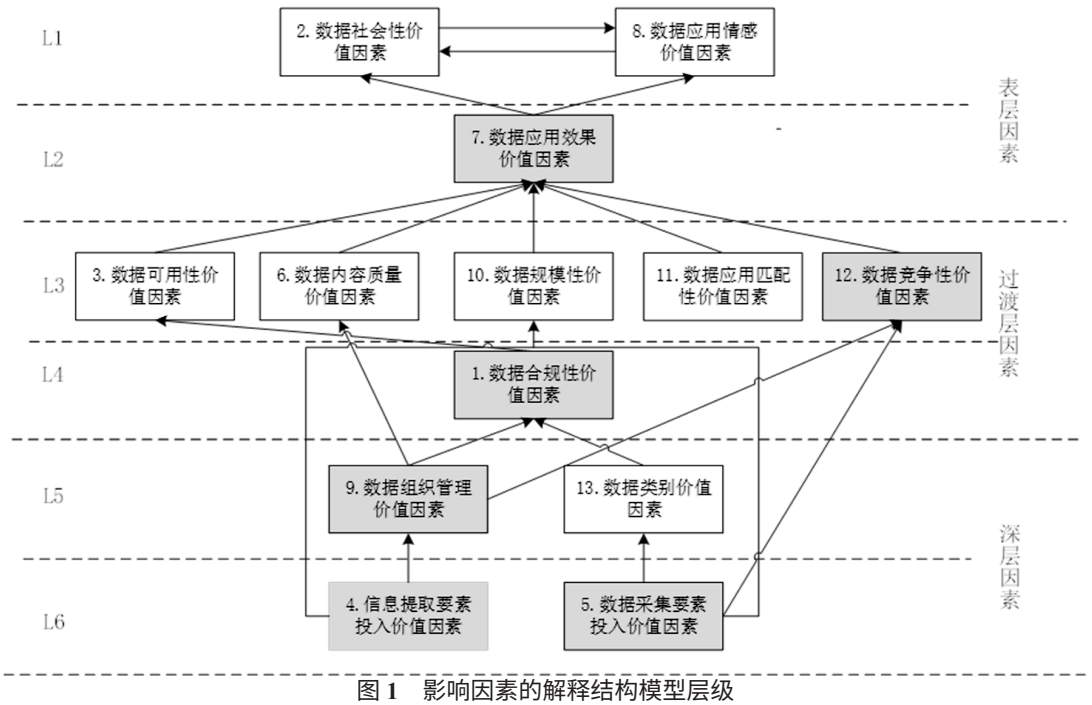

图 1　影响因素的解释结构模型层级

4　主要影响因素的 MICMAC 分析

4.1　MICMAC分析方法

MICMAC 分析是指应用于分类的交叉影响矩阵乘法，以乘法的矩阵特性为基础，允许一种简单的方法来识别和解释复杂的问题。在划分因素层级结构，获得价值影响因素的 ISM 模型之后，为进一步掌握因素对系统的作用途径和效果，将 MICMAC 应用于因素的优先级排序，而这种排序是基于因素的驱动力和依赖性来进行的。在可达矩阵中，因素的驱动力值可以通过行的水平求和的方式来表示，而列的垂直总和可表示依赖性，计算结果如表 7 所示。

以因素的依赖性作为横轴 X 的坐标值，驱动力作为纵轴 Y 的坐标值，在坐标系中标注对应元素所在位置，以此绘制驱动力 - 依赖性分布图，如图 2 所示。根据驱动力和依赖性的大小可将所有影响因素分为四个集群，即链接群（第Ⅰ象限）、独立群（第Ⅱ象限）、自主群（第Ⅲ象限）、依赖群（第Ⅳ象限）。借助 MICMAC 分析构造的矩阵如图 2 所示，用于进一步理解影响因素之间的相互作用。

表 7　影响因素的驱动力 - 依赖性结果

因素	驱动力	依赖性
VF1	5	5
VF2	2	13
VF3	4	5
VF4	10	1
VF5	9	1
VF6	4	3
VF7	3	11
VF8	2	13
VF9	8	2
VF10	4	3
VF11	4	1
VF12	4	4
VF13	5	2

doi:10.3772/j.issn.2095-915x.2024.03.005

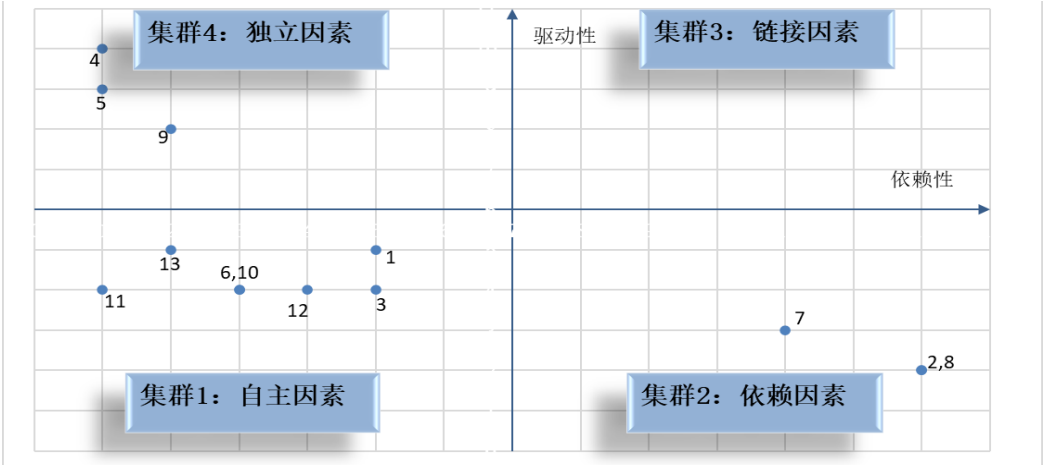

图 2　影响因素的驱动力 - 依赖性分布

4.2　MICMAC结果分析

通过图 2 的影响因素“驱动力 - 依赖性”分布结果可知：

(1) 属于自主群（集群 1）的影响因素有 VF1、VF3、VF6、VF10、VF11、VF12、VF13，这类影响因素通常具有较低的驱动力和较低的依赖性，除数据类别价值因素外，位于解释结构模型的第三、四层，即过渡层。

(2) 属于依赖群（集群 2）的影响因素有 VF2、VF8、VF7，此类因素对其他因素具有较大的依赖性和较小的驱动力，需要下层因素的支撑，位于解释结构模型的第一层级和第二层级，即表层。

(3) 没有属于链接群（集群 3）的影响因素，说明本研究数据资产影响因素独立性强，不存在属于模糊关联概念的链接群因素。

(4) 属于独立群（集群 4）的影响因素有 VF4、VF5 和 VF9，此类因素具有较强的驱动力和较低的依赖性，位于解释结构模型的第五层级和第六层级，即深层因素，也是关键层。

根据 MICMAC 分析结果对因素的层级进行调整，把层次结构划分为三层，即关键层、具体层和表层。如图 3 所示。

第一层级为表面层因素，对应依赖因素群，具有较高依赖性和较低的独立性，因此容易被中层因素和底层因素所触发，容易使得系统处于不安全不稳定状态，因此需要对其进行严格的考察和控制。表面层因素的有效运作往往要以其他因素作为支撑，但所产生的作用也是尤为明显的。

第二层级为具体层因素，对应自主因素群，具有较低的驱动力和依赖性。这类因素同其他因素的关系较弱，一旦形成，便趋于稳定，具有一定的自主性和自发性。这说明在大数据资产价值形成的过程中，直接通过调整其他因素来改善这些因素或通过改进这些因素来改善其他因素都无法取得显著效果。

第三层级为关键层因素，对应独立因素群，具有较高的驱动力和低依赖性。这类因素位于底层，对其他因素的影响较大，但其影响方式通常不够直观，容易受到忽视。而根据帕累托法则，关键的少数会对事物的结果产生绝大部

分的影响，改进这些因素会对其他因素以及整个系统产生较强影响效果；且由于其所受到其他因素的限制较小，所需条件宽松，应对这类因素进行优先考虑，着重改进。

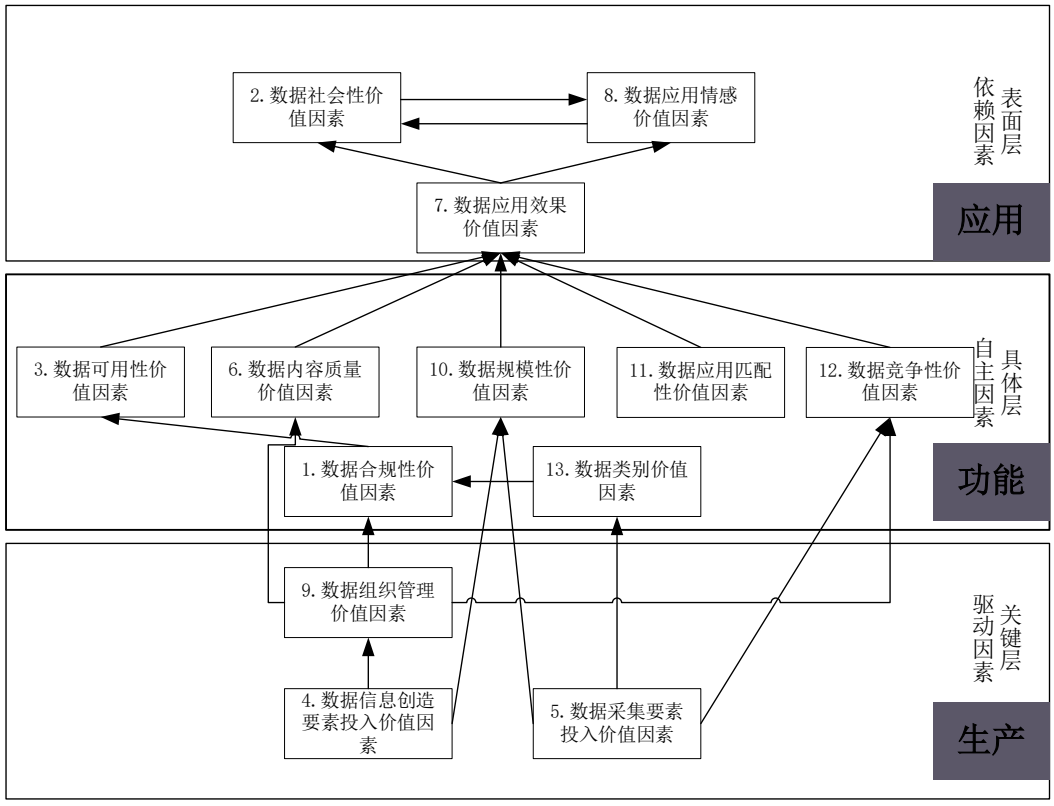

图 3　影响因素解释结构模型的驱动 - 依赖性分层

5　对策建议

5.1　长远布局关键层驱动因素，提供强大动力

IT 基础构架的建立与完善是数据资产实现价值的先决条件。现有的 IT 架构不利于孤立信息的集成，并且难以存储、处理和分析大量的非结构化数据，许多遗留系统都是为批量交付数据而构建的，无法为决策提供实时连续的信息流。其次，由于数据资产价值是一个新的概念，存在人才短缺问题。且大数据涉及跨学科的知识和技术，人力资源培养难度较大，所需时间长。因此，需要对技术、人才和管理等要素进行长远的布局与配置，促进数据资产的价值实现。

人才方面，高校和研究机构是培养大数据资产人才最重要、最直接的场所，应该开设相关课程，提高“跨学科的复合型”人才培养质量，扎实培养包括大数据资产评估师和大数据工程师在内的人才；完善相应的职称评定制度，激励引导数据专业技术人员的职业发展，为他们提供努力的目标与方向；加强政府、高校以及企业之间合作，通过优势互补，促进人才多元化培养，形成产学研结合的育人机制。强调人力资本的共享，建立数据人力资本共享中心，举办人才交流会。

管理方面，企业必须积极调整人才策略和领导力发展方向，加快向数字化转型，确保员工适应并且利用数据相关技术领域的变化。需

doi:10.3772/j.issn.2095-915x.2024.03.005

要团队经理、人力资源、设施管理、IT 和业务部门主管之间进行协调工作，成立跨学科团队；融合数据技术和业务领域的专业知识，明确业务和技术成果责权分配；提供大量以数据技术为中心的职业发展和晋升机会，吸引人才，扩展企业大数据处理能力。

技术方面，数据系统和解决方案（包括开源软件）的供应商和开发商应该持续开发功能更强大的工具，升级 IT 架构和基础设施。开发收集、储存、管理、分析和共享数据的核心技术，促进数据的实时处理；积极探索应用区块链等技术，解决数据交易中的隐私和安全问题。

5.2 着重解决表层依赖因素，提高实施成效

数据资产价值系统的表层因素直接影响着数据的最终价值，因此要充分提高实施成效，激活使用价值。企业应该发展大数据驱动的商业模式，实施大数据战略，促进数据资产在企业运营管理、供应链管理、产品优化以及决策中的应用。

对社会公众，要提高他们的数据素养，包括数据资产相关的知识和开发利用数据的技能，增加公众参与数据资产价值共创的积极性。各地方网信办、大数据管理中心等相关部门等，应举办数据资产相关的知识讲座，普及相关知识，提高公众对数据资产的认知水平；举办数据相关比赛，促进公众对大数据的开发利用；集合筛选优秀应用场景案例，召开交流会，促进数据资产在医疗、工业、交通等领域的融合应用技术的创新突破。

对数据应用企业本身，应该培养数据驱动的文化：首先应加强公司领导者的数据驱动思维模式，公司领袖必须深刻意识到数据作为一种资产，可以利用其来改善决策和提供新产品和服务等，以提高公司效益；其次是建立与数据驱动业务适配的框架体系：用来突破公司现有的业务模型，促进公司对数据的高效利用；重视数据人才的引进，通过“社会招聘 + 企业选拔”的方式，搭建数据资产应用人才队伍。

5.3 合理规划具体层自主因素，稳定支撑保障

尽管可以使用相关工具和技术来评估和提高数据质量，但由于自主因素具有稳定性，因此需要对数据自身质量相关的固有价值因素进行单独考虑，其核心是加强对数据资产的管理，包含以下几个方面：

完善数据资产固有价值的评估体系和数据应用规范，为激活数据资产价值提供良好的政策环境。数据资产固有价值由许多而不是一个属性决定，取决于多个维度及其之间复杂相互作用，并且对于不同的应用场景和使用目的具有不同的属性偏好，需要规范指导如何根据不同需求对这些价值因素的交叉作用以及权重进行平衡和赋值。

增加数据质量管理的能力。数据质量管理是数据资产管理的核心，数据质量管理和数据生产使用的经验是引导公司提高数据能力的无形资产，对数据质量进行测量和评估有助于更好地了解数据资产价值并评估数据质量问题带来的业务风险，降低其不确定性。

促进基于数据资产固有价值因素的市场细分，发挥长尾效应。对于供应商来说，定价是创造价值的一项重要工作。由于自主因素具有的稳定性和过渡性，同时代表了数据的功能性

价值，因此对于指导数据定价意义重大；披露此类因素相关信息可增加用户对价格感知的透明度和公平性，使得数据更具吸引力，提升数据资产价值；由于数据资产具有长尾效应，应不断探索基于数据固有价值的版本化定价策略，来契合数据消费者的个性化需求。这要求能够对用户的需求实现精准识别，对此可结合问卷调查和访谈等方式进行需求调查。

6 结语

本研究对数据资产价值影响因素进行深入挖掘，利用解释结构模型，得出数据资产价值的表层因素、过渡因素和最深层次的因素，结合 MICMAC 分析法，对各个影响因素进行依赖性和驱动力的整合，分别找出了独立群、依赖群和自发群的影响因素，进而提出相应的管理策略。同时本研究也存在一定的局限性：一方面，问卷调查以及专家评估打分存在一定的主观性；另一方面，数据资产价值影响因素指标具有一定的可拓展性。未来的研究中还应考虑对因素个数和维度进行进一步扩充。同时，对于不同类型的数据资产，应分析各个因素之间的具体影响系数，提高研究的深度和精度。

参 考 文 献

[1] GANTZ J, REINSEL D. Extracting Value from Chaos [C] // IDC iview, 2011, 1142(2011): 1-12.

[2] MANYIKA J, INSTITUTE M G, CHUI M, et al. Big Data: The Next Frontier for Innovation, Competition, and Productivity [M]. Washington: McKinsey Global Institute, 2011: 158-164.

[3] Office of Science and Technology Policy Executive, Offrice of the President. Obama Administration Unveils “Big Data” Initiative: Announces $200 Million in New R&D Investment [EB/OL]. (2012-03-29) [2023-02-24]. https://obamawhitehouse.archives.gov/the-press-office/2015/11/19/release-obama-administration-unveils-big-data-initiative-announces-200.

[4] GARGANO M L, RAGGAD B G. Data Mining - a Powerful Information Creating Tool [J]. MCB UP Ltd, 1999, 15(2): 81-90.

[5] MAYER-SCHNGERGER V, CUKIER K. Big Data: A Revolution That Will Transform How We Live, Work and Think [M]. London: John Murray, 2013: 86-98.

[6] HECKMAN J R, BOEHMER E L, PETERS E H, et al. A pricing model for data markets[J]. iConference 2015 Proceedings, 2015(2): 39-49.

[7] 陈筱贞 . 大数据交易定价模式的选择 [J]. 新经济 , 2016(18): 3-4.

[8] NASSER T, TARIQ R. Big Data Challenges [J]. J Comput Eng Inf Technol, 2015, 9307(2): 174-185.

[9] MUGHELE E S, ABAYOMI D. Quantitative and Financial Measurement of the Value of Information [J]. Information systems, 2013, 4(2): 285-296.

[10] 唐要家 , 唐春晖 . 数据要素经济增长倍增机制及治理体系 [J]. 人文杂志 , 2020(11): 83-92.

[11] 夏义堃 . 数据要素市场化配置与深化政府数据治理方式变革 [J]. 图书与情报 , 2020(3): 14-16.

[12] 孙文章 , 杨文涛 . 基于多期超额收益法的互联网金融企业数据资产价值评估研究 [J]. 中国资产评估 , 2023(2): 4-18.

[13] 孙淑萍 . 企业数据资产价值评估影响因素研究 [J]. 商场现代化 , 2022(22): 37-40.

[14] 李永红 , 张淑雯 . 数据资产价值评估模型构建 [J]. 财会月刊 , 2018(9): 30-35.

[15] LI Y, ZHANG S. Research on the valuation method of data assets of Internet enterprises[J]. J. Economic research guide, 2017, 14: 104-107.

[16] DING Y, HOPE O K, JEANJEAN T, et al. Differences between domestic accounting standards and IAS: Measurement, determinants and implications[J]. Journal of accounting and public policy, 2007, 22(1): 1-38.

doi:10.3772/j.issn.2095-915x.2024.03.005

[17] SHEN Y, GUO B, SHEN Y, et al. A pricing model for big personal data[J]. Tsinghua science and technology, 2016, 21(5): 482-490.

[18] IMMONEN A, PÄÄKKÖNEN P, OVASKA E. Evaluating the quality of social media data in big data architecture[J]. Ieee Access, 2015, 3: 2023-2043.

[19] 中华人民共和国国家标准化管理委员会 . 电子商务数据资产评价指标体系 : GB/T 37550-2019[S]. 北京 : 中国标准出版社 2010: 2-7.

[20] VISCONTI R M, LAROCCA A, MARCONI M. Big data-Driven Value Chains and Digital Platforms: from Value Co-Creation to Monetization[J]. SSRN Electronic Journal, 2017(1): 355-372.

[21] LIU J, LI J, LI W, et al. Rethinking Big Data: A Review on the Data Quality and Usage Issues [J]. ISPRS journal of photogrammetry and remote sensing, 2016, 115: 134-142.

[22] 金坚 . 大数据时代信息熵的价值评价研究 [D]. 吉林 : 吉林大学 , 2019.

[23] 张志刚 , 杨栋枢 , 吴红侠 . 数据资产价值评估模型研究与应用 [J]. 现代电子技术 , 2015, 38(20): 44-47, 51.

[24] 李立睿 , 邓仲华 . 面向科研用户小数据的微知识服务研究（上）[J]. 图书与情报 , 2016(3): 49-54.

[25] DAKOVA J, ANTUNES P, CHIU Y T. A Pluralistic Approach to Information Valuation[C]//PACIS. 2018(27): 71-85.

[26] BHADANI A K, JOTHIMANI D. Big data: challenges, opportunities, and realities[J]. Effective big data management and opportunities for implementation, 2016: 1-24.

[27] WANG R, STRONG D. Beyond Accuracy: What Data Quality Means to Data Consumers [J]. Journal of management information systems, 2016, 12(4): 5-33.

[28] 莫祖英 . 大数据质量测度模型构建 [J]. 情报理论与实践 , 2018, 41(3): 11-15.

[29] 韩海庭 , 原琳琳 , 李祥锐 , 等 . 数字经济中的数据资产化问题研究 [J]. 征信 , 2019(4): 72-78.

[30] 赵子瑞 . 浅析国内大数据交易定价 [J]. 信息安全与通信保密 , 2017(5): 61-67.

[31] 王玉林 , 高富平 . 大数据的财产属性研究 [J]. 图书与情报 , 2016(1): 29-35, 43.

[32] 关涛 , 高晶 , 张雪桐 . 顾客感知价值对网络信息产品定价的影响研究 [J]. 财经理论与实践 , 2017, 38(4): 97-102.

doi:10.3772/j.issn.2095-915x.2024.03.006

基于文本挖掘的颠覆性技术识别方法综述

孙心怡　梁洪振　刘桂锋

江苏大学科技信息研究所　镇江 212013

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 对颠覆性技术的基础理论与识别方法进行梳理归纳有助于把握技术发展态势，推动数智时代面向科技创新的智能化情报服务体系构建。[方法 / 过程] 明确颠覆性技术与文本挖掘方法的基础内涵，通过对该研究领域强关联的核心论文进行定性解读，将基于文本挖掘的颠覆性技术识别方法进行分类比较与组合分析。[局限] 当前颠覆性技术识别领域应用的文本挖掘技术尚不成熟，需要进一步改进并与其他方法结合使用来优化识别效果。[结果 / 结论] 基于文本挖掘的颠覆性技术识别方法包括突变词检测、术语词分析、共词分析、文本聚类、主题模型、语义分析。文本挖掘侧重从技术属性角度识别颠覆性技术，深度挖掘科技文本数据隐藏的潜在技术主题。

关键词：颠覆性技术；技术识别；文本挖掘；技术范式创新；突破性技术

中图分类号：G350

A Review of Research on Identification of Disruptive Technologies Based on Text Mining

SUN Xinyi　LIANG Hongzhen　LIU Guifeng

Institute of Science and Technology Information of Jiangsu University, Zhenjiang 212013, China

Abstract: [Objective/Significance] The summarization of basic theories and identification methods of disruptive technologies can help grasp the technology development trend and promote the construction of intelligent intelligence service system towards scientific and technological innovation in the era of digital intelligence. [Methods/Processes] This paper firstly clarifies the connotation of disruptive technology and its related concepts, secondly introduces the concept of text mining method and the difference and connection with bibliometrics, and finally summarizes the identification methods of disruptive technologies based on text mining and carries out comparative analysis. [Limitations] Current text mining techniques applied in the field of disruptive technologies recognition are immature and need to be further improved and used in combination with other methods to optimise recognition. [Results/Conclusions] The identification methods of disruptive technologies based on text mining include mutation word detection, terminology word analysis, co-word analysis, text clustering, topic modelling, and semantic analysis.

基金项目　国家社科基金一般项目“科学数据融合模式设计与体系建构研究”（21BTQ080）；江苏省教育科学规划课题高校重点项目“数智化背景下江苏省高校纪检监察工作运行模式研究”（B/2023/01/194）。

作者简介　孙心怡（1999-），硕士研究生，主要研究方向为文本挖掘与主题识别，Email：1014905154@qq.com；梁洪振（1976-），博士，副研究员，主要研究方向为信息资源管理与高等教育；刘桂锋（1980-），博士，教授，主要研究方向为科学数据管理。

引用格式　孙心怡 , 梁洪振 , 刘桂锋 . 基于文本挖掘的颠覆性技术识别方法综述 [J]. 情报工程 , 2024, 10(3): 77-87.

doi:10.3772/j.issn.2095-915x.2024.03.006

Text mining focuses on identifying disruptive technologies from the perspective of technical attributes, and deeply explores the potential technical themes hidden in scientific and technological text data.

Keywords: Disruptive Technologies; Technology Identification; Text Mining; Technological Paradigm Innovation; Breakthrough Technologies

引言

当前全球科技创新已经进入密集活跃期，以颠覆性技术（Disruptive Technologies）为代表的技术革新为全球政治、社会、经济格局带来深刻影响。对于这些具有颠覆性潜力的技术进行识别具有重要的现实意义，一方面有助于发现创新机会，推动科技进步和经济发展；另一方面及时识别出颠覆性技术可以对潜在风险进行预警从而制定出应对措施，保障社会的稳定和安全；此外，对颠覆性技术识别进行深入研究可以加强学术研究与产业创新的融合，促进资源的优化配置。

目前主流的颠覆性技术的识别方法可以大致概括为以下两种：一是主观识别方法，其主要依据是领域专家经验和学术背景，包括专家意见法、技术路线图法、情景分析法等。这类方法主要通过行业或领域专家的学科知识背景来评估技术属性及潜在价值，从而判断某项技术是否具有颠覆性效应，是使用最早、最广泛成熟的方法。这类方法的操作流程简单、使用成熟，能够综合考虑技术要素和市场要素，分析维度全面，但大多从宏观层面评估技术要素，存在主观性强、过程复杂耗时、实施难度大等问题。二是客观识别方法，主要依赖科技文献信息的相关定量特征，包括文献计量法、技术指标评估法、模型分析法等。这类方法大多将科技文献作为数据来源，挖掘能够表征颠覆性技术的关键信息并进行定量分析，优势在于数据来源科学广泛、数据量大，识别结果更加全面可观，是当前科技情报学界较为推崇的方法。文献计量法是对专利、论文等科技文献中能够反映技术特性信息的计量，如共词分析、词频统计、引文分析等[1]；技术指标评估法主要针对技术特征、市场变化、外部环境等方面将技术量化表示，通过构建综合指标、分配权重从而进行统计识别[2]；模型分析法是以理论准则为基础的技术量化模型，代表性的理论包括 TRIZ 理论、QFD 理论、种间竞争理论、传染病模型等[3-5]。这类方法虽然能够有效避免主观性，但大多是对外部特征的描述，缺少针对多源数据的内容特征分析、关联关系分析、多特征项交叉共现分析等。

近年来兴起并日渐流行的文本挖掘方法，不仅能够有效规避主观识别方法的一系列问题，还能处理非结构化数据，弥补客观识别方法中文本内部特征分析不全面的缺陷。基于文本挖掘的方法可以从科技文本数据中抽取技术术语，通过共词分析、突发词检测、文本聚类、SAO 结构分析、主题模型等方法及相应的软件工具提取科技文本中的关键词、主题词、语义结构等来识别符合某些特征的技术，从而能够更好地挖掘深层次的技术隐性知识。

1 理论与方法基础

1.1 颠覆性技术的内涵

颠覆性技术的概念最早由 Christensen 等 [6-7] 在 1995 年提出，他们认为这类技术“早期以简单、方便、便宜的特征由低端或边缘市场切入，在性能的不断升级中逐渐取代主流技术，从而形成新的市场格局与价值体系”。此后国内外学者也从不同视角提出了各自的见解，然而这一概念在不断地发展演变中尚未统一，本文将其中代表性的观点按不同视角进行归纳，详见表 1。

表 1 不同视角下国内外学者对颠覆性技术的概念界定

视角	侧重点	定义	来源
市场环境	低端市场切入	技术特性及功能与主流市场相似，但初期性能较差，后期性能跃升并取代主流技术。	[8]
	高端市场切入	初期因技术成本过高而导致受众范围较小，后期技术性能显著跃升并取代主流技术。	[9]
技术属性	替代性改进技术	基于现有技术的改进突破，通过不断升级迭代占领主流市场，并形成新的价值体系。	[10]
	原始性创新技术	以全新的概念或原理为基础的原始创新，被接纳后快速传播并给社会带来根本性变化。	[11]
行业角度	企业竞争	一类改变企业竞争基础与能力的绩效指标。	[12]
	产品与服务	现有技术范式、产品与服务功能的变革，并对产业的市场与竞争格局造成破坏性影响。	[13]
其它		一是基于新原理、新发现的原始创新；二是基于现有技术的集成创新与应用；三是科学原理与成熟技术的转移与创新应用。	[14]

从市场环境来看，可根据市场颠覆方式分为高端颠覆性技术和低端颠覆性技术，具体的技术轨迹如图 1 所示。在将市场分为高端、主流、低端三类的基础上，进一步将高端市场和低端市场分别细分为高端分离、高端边缘市场以及低端分离、低端边缘市场，由此可以得到四条技术轨道：高端颠覆性技术从高端分离市场切入，经过高端边缘市场进入并颠覆主流市场；Rafii 等 [8] 认为新市场颠覆性技术轨道从低端分离市场切入，经由低端边缘市场进入并颠覆主流市场；低端颠覆性技术从低端边缘市场切入并随时间推移进入并颠覆主流市场；持续性技术则始终位于主流市场，是对主流技术的持续性创新。

从技术属性来看，一种是替代性改进技术，是对现有技术性能的突破性改进，在经历多次迭代升级后能够提供完全替代甚至超越主流技术性能的产品，最终彻底颠覆占领主流市场。另一种是原始性创新技术，Christensen 等 [11] 将其定义为以全新的技术理念为前提所产生的原始性创新，在技术渗透的过程中逐渐满足主流需求后还能进一步开发新的技术需求与市场需求，最后形成新的价值体系。还有卢光松等 [13] 从行业角度对颠覆性技术概念进行界定，一是从企业竞争角度考虑，关注技术的颠覆性效应对企业竞争能力以及绩效水平的改进；二是基于产品与服务视角，侧重技术创新过程中对产品性能、服务水平、产业变革、社会效益等方面所带来的重大影响和突破。

doi:10.3772/j.issn.2095-915x.2024.03.006

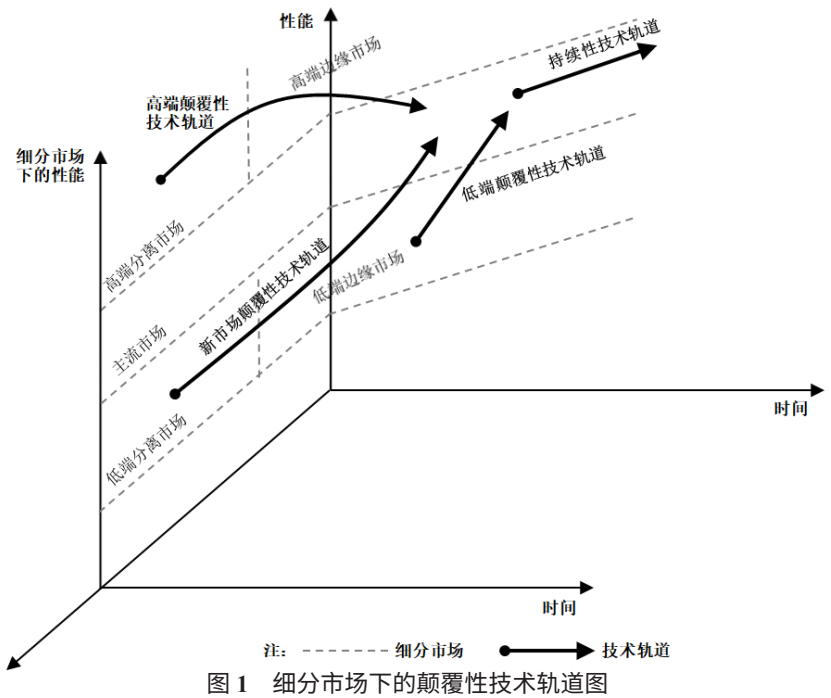

图 1　细分市场下的颠覆性技术轨道图

综合来看，当前研究一致认为颠覆性技术是由现有技术产生替代改进式突破或原始性创新技术动态演变而来的，此类技术初期由边缘市场切入，因技术性能较低而无法满足主流需求而容易被忽视，后期通过性能跃升入侵、威胁、最终彻底破坏主流技术轨道，颠覆主流市场格局，最终塑造全新的价值体系。

通过梳理相关文献可以发现，有些概念与颠覆性技术之间存在密切关联及差异，因而清晰梳理并辨别它们之间的异同有助于为后续相关研究的开展提供理论依据。本文在已有研究基础上对这 5 个相关技术的定义、技术特征、要求进行梳理归纳，如表 2 所示。

表 2　颠覆性技术的关联概念辨析

名称	定义	技术特征及要求	来源
新兴技术	新出现的，具有创造新产业或改造现有产业的潜力的技术。	新颖性、明显发展潜力、持续增长性	[15]
前沿技术	高技术领域中具有前瞻性、先导性和探索性的重大技术。	技术的前瞻性、新兴性、探索性	[16]
突破性技术	通过新知识淘汰现有技术，具有变革性和不连续性的技术。	对已有技术的根本性创新	[17]
破坏性技术	借助低端市场及成本优势对已有技术进行融合创新从而颠覆主流市场。	初始性能较差、结构简单、低成本、高易用性	[18]
“卡脖子”技术	具有高技术价值并同时存在技术垄断的关键核心技术或高技术产品。	垄断性、潜伏性、高技术特征，是国际竞争的威胁性工具，关键核心技术缺乏自主权	[19]

1.2　文本挖掘方法

颠覆性技术具有潜在变革性的影响，对其进行识别与分析对社会、经济、科技等领域的意义重大，近年来文本挖掘技术凭借高效灵活、大数据适应性、信息揭示性、可扩展性等优势被更多地应用到颠覆性技术识别的研究中。文本挖掘指的是从大规模的文本形式的数据集合中提取以往未知的、隐性的、可理解的、实用

的信息的技术，能够处理非结构化数据并常应用于文本表示模型、特征项抽取、模式发现（如关联规则抽取、文本分类、文本聚类）等领域[20-21]。现有的文本挖掘方法主要分为三种：一是基于数理统计的方法，如词频统计、TF-IDF、文本分类等；二是基于人工设定规则的方法，如语法分析、句法分析等；三是基于机器学习的方法，如 LDA、SVM、朴素贝叶斯、深度学习等。文本挖掘方法在技术主题识别的研究中已得到广泛成熟的应用，包括引文聚类、技术术语词频分析、主题模型、异常监测等[22-24]。而数据的增量重组与学科的交叉融合促使文本挖掘方法被更多地应用到计算机科学、图书馆学情报学、教育学、社会学、心理学、技术管理等众多学科领域中，研究对象也逐渐从学术文献扩展到社交网站，如新闻评论、博文帖子、客户评价、语音文本（STT）等，当前的研究趋势聚焦于大数据分析、社交媒体分析、情绪分析、话题建模等方面。随着基础理论与算法的创新优化，文本挖掘方法的应用场景在不断扩张，但在实际应用中也存在诸如文本数据噪声、算法的普适性等问题，因而需要设计更加准确高效的算法加以优化，文本挖掘在不同应用场景下与其他方法的优势组合将成为重要的研究趋势。

2 基于文本挖掘的颠覆性技术识别方法

文本挖掘因其良好的数据挖掘分析能力，被越来越多地运用到技术主题分析中，本文确定检索式：“TI=(颠覆性技术 OR 颠覆性创新 OR 破坏性技术 OR 破坏性创新 OR 突破性技术 OR 突破性创新 OR 不连续技术 OR 技术不连续性) AND TI= 识别”，在 CNKI 中获取到与研究主题强关联的核心期刊论文 40 篇，学位论文 7 篇，并对这 47 篇文献进行解读，将基于文本挖掘的颠覆性技术识别总结为如下六种方法，并对这六种方法进行比较分析。

2.1 方法类型

2.1.1 突变词检测

突变词检测（Burst Items Detection）依据词频增长率的突变情况判断某一领域的发展趋势和研究动向[25-26]，颠覆性技术的发展路径符合突变词的变化规律，颠覆性技术发展之初潜伏于边缘市场，容易被主流市场所忽略，后经过技术性能的更迭升级迅速取代、颠覆主流市场，而突变词则是捕捉一定时期内词频增长率发生异常骤变的词语，并将其定义为突变词。同颠覆性技术一样，突变词的增长率在时段初期不易被关注到，但却能够在某一时间节点发生突变。突变性是颠覆性技术的重要特征，通过观察科技文本数据中发生突变的词语，可以有效发现新兴技术主题和热门话题，结合观察突变词的动态变化情况，有助于挖掘具有颠覆性潜力的词项，为后续的技术识别奠定基础。现有研究中大多基于主题词的突变情况进行异常检测，通过结合不同的机器学习算法来优化识别的精度，如 Miao 等[23] 利用 LDA 主题模型和主题变异模型识别技术演化路径中的关键节点；王秀红等[27] 基于 CBLOF 算法进行关键技术主题的异常监测来表征颠覆性技术的突变属性，从而识别出颠覆性技术。突变词检测常用于新兴前沿技术主题的探测，以此观察颠覆性

doi:10.3772/j.issn.2095-915x.2024.03.006

技术的突变性特征。

2.1.2 术语词分析

文本挖掘的首要任务是命名实体的识别，延伸到颠覆性技术识别的研究中则表现为颠覆性技术术语词的抽取，首先需要借助自然语言处理（NLP）对论文专利等科技文本中的标题、摘要等内容进行分词、去停用词、保留词等操作，进而得到能够体现技术属性或特征的技术术语，并用以反映技术主题，在此基础上进行分析。术语词分析在颠覆性技术的识别研究中表现为对高频词的时序分析，通过术语词词频的年代分布观察技术主题的发展情况，能够发现技术动态演变中的新兴技术术语及其最新动向，观察不同时期的术语词分布有助于发现技术主题的演变规律，从而更好地识别出具有颠覆性潜力的技术，如邢晓昭等[28]根据专利标题和摘要中抽取到的术语关键词构建时间窗下的术语关键词网络并识别出类脑智能领域的颠覆性技术主题。术语词分析主要用于技术实体的抽取，并且更多情况下仅作为一种手段或环节，要想更加准确高效地识别颠覆性技术还需要与其他方法组合搭配进行。

2.1.3 共词分析

共词分析本是文献计量的传统方法，随着方法应用的深度融合逐渐被改进并融入文本挖掘中。共词分析在文献计量学中更多地被用于分析同一篇期刊论文中关键词对的共现情况，从而观察某研究领域的学科结构与主题变化。然而，识别颠覆性技术的数据来源中除了论文，都缺少规范的关键词，文本挖掘能够自动抽取非结构化文本数据中的术语关键词，利用共词分析发现技术主题。目前共词网络分析法使用较成熟，根据技术术语词在同一文本中的共现情况绘制共词网络图，从而更直观地反映出术语词之间的共现关系，揭示技术主题的分布关联情况。如 Dotsika 等[29]使用关键词共现网络揭示颠覆性技术领域的主题内容、结构及时间趋势。此外，共词聚类分析和战略图分析也是共词分析中使用较多的方法，共词聚类常用 K-means 聚类、层次聚类等方法将复杂的术语词汇聚成不同的技术主题，而战略图分析则是在共词聚类的基础上计算主题的中心度和密度来绘制战略图[25]。共词分析法常与文本聚类法相结合，用于技术主题的识别与分析。

2.1.4 文本聚类

文本聚类在颠覆性技术识别的研究中主要表现为技术主题聚类。通过测量已抽取出的术语关键词、突现词等词项各自的语义相似度，将不同文本数据集合中相似度较高的文档聚为一类，并探测相应的技术主题，对科技文本内容进行深度挖掘，从而发现潜在的知识结构，识别出更能够表征颠覆性技术属性的技术主题，进一步优化识别效果。文本聚类一般分为两个阶段：第一个阶段是将术语词根据语义关联或共词关系进行聚类，从而形成多个技术主题团簇；第二阶段是将不同的主题团簇进行命名，生成相应的主题词或构建主题词表。常用的文本聚类方法包括 TF-IDF、K-means、层次聚类、DBSCAN 等，Li 等[30]将文本聚类算法用于定位与技术特征相对应的文本内容来识别颠覆性技术，Liu 等[31]将 K-means 聚类与表示学习相结合从多源文本中识别潜在的颠覆性技术及其特征。综合来看，文本聚类方法可以从宏观上快速直观地把握技术主题的分布，多数情况下

用于技术特征的识别与抽取，但由于目前的文本挖掘技术并不足以深入、精准地捕捉潜在的语义信息，因而聚类效果往往无法切合预期，这就不可避免地需要人为干预辅助分析。要想充分解决这一问题，首要任务在于攻克文本挖掘的技术难关。

2.1.5 主题模型

主题模型是现阶段使用最广泛而普遍的颠覆性技术识别方法，是对大量多源异构的科技文档集合中所识别出的术语词进行统计，从而反映出文本内容中隐藏的潜在技术主题分布、主题结构以及主题演化趋势等。现阶段在识别颠覆性技术的研究中使用最多的主题模型是隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）模型，它的本质是一种概率图模型，根据文档集中的词项分布生成隐藏的技术主题分布从而对文本数据进行降维，发现技术隐性知识。单晓红等[32]利用 LDA 主题模型识别出颠覆性技术，但由于传统的 LDA 模型存在无法把握词项间的语义关系、容易混淆歧义较大的特征词等问题，从而造成识别结果存在误差。因而有学者尝试将 LDA 与其他方法结合使用来优化识别效果。如将 LDA 模型与专利开发路径、K-core 分析相结合，识别与检测光伏行业潜在的颠覆性技术[33]。还有一些常用于识别颠覆性技术的主题模型，如 LDAVec、DTM、STM 等。主题模型法最大的优势在于能够处理大量的文本数据，并对文本内容进行降维处理从而揭示潜在的技术趋势和创新点，但由于在模型的适用性、稳定性、文本数据的质量与处理等方面仍面临一些挑战，因而需要不断进行模型调试与改进来优化识别效果。

2.1.6 语义分析

由于颠覆性技术发展路径存在高度不确定性，仅仅对关键词或主题词进行词频统计无法精准把握颠覆性技术的内在特征，越来越多的学者开始尝试利用语义分析来识别颠覆性技术。语义分析法是一种基于统计的文本数据挖掘技术，常用的方法是 SAO（Subject-Action-Object）结构语义分析，“S”“A”“O”三个字母分别表示的是技术组件的主语、宾语以及表示实体间关联关系的谓语，Li 等[30]使用 TF-IDF 与专利化的 SAO 结构提取技术特征，并结合深度分类模型识别颠覆性技术。除此之外，国外的一些研究机构和政府部门也开始将语义分析用于识别和跟踪颠覆性技术，如美国国防部高级研究计划局（DARPA）通过对科学论文、新闻报道等其他公共数据进行文本主题的语义分析来识别具有颠覆性潜力的技术创新。借助语义分析可以较好地挖掘科技文本中所包含的技术实体及其关联要素，从而更精准地进行技术识别。语义分析主要用于颠覆性技术识别的科技文本分类及技术特征提取等方面。

2.2 方法分析

2.2.1 基于文本挖掘的颠覆性技术识别方法比较

基于文本挖掘的颠覆性技术识别方法类型上以定量为主，主要表现为对文本内容数据的量化统计或结构分析，方法特征上大多从技术特征视角切入，利用文本挖掘技术从科技文档中抽取能够表征颠覆性技术属性的实体，进而识别出颠覆性技术。为了更直观清晰地了解上述六种文本挖掘方法的区别与联系，本文从作用思路、方法内核、优势、局限性、应用范围几方面进行进一步的比较分析，如表 3 所示：

doi:10.3772/j.issn.2095-915x.2024.03.006

表 3　基于文本挖掘的颠覆性技术识别方法比较

方法名称	作用思路	优势	局限性	应用范围
突变词检测	将词频增长率的突变情况作为技术领域研究态势的判断标准	切合技术路径	仅依赖词频增长率判断难以掌握技术本质特征	技术突变性特征计量
术语词分析	借助 NLP 技术从科技文本中抽取技术术语来反映技术主题	抽取技术特征	无法反映术语词或技术主题之间的关系	技术实体抽取
共词分析	对相同文本中术语词的共现情况进行分析	把握技术热点分布及其关联	存在时间滞后性	技术主题分析；技术特征识别
文本聚类	根据计算出的语义相似度将高关联的文本聚为一类探测技术主题与知识结构		难以进行技术主题的横向分析，算法准确度有待提高	
主题模型	从大量科技文档集合中识别出隐藏的潜在技术主题并进行主题分析	处理大量复杂的文本数据	模型准确性和稳定性有待长时间监测	科技文本降维
语义分析	从科技文本中自动抽取语言知识并进行语义计算	更精准理解科技文本含义	需要大量语料库和语言知识的支持，准确性受语言复杂程度影响	技术特征识别；科技文本分类

2.2.2　基于文本挖掘的颠覆性技术识别方法组合

上述六种文本挖掘方法在颠覆性技术识别过程中具体的应用位置如图 2 所示。从整体的识别过程来看，现阶段的学者更倾向于多种方法的组合使用，从而将各个方法的优势发挥到最大。

（1）文本数据来源：识别颠覆性技术的数据大多来源于与技术本质密切相关的科技文本，包括科技论文、专利、科技新闻、政策文本、科技报告等，从数据源上看，基于科技论文或专利的单一数据源进行颠覆性技术识别的研究占比最多，其次是基于论文和专利两种数据源，而随着文本挖掘技术的不断发展，以多种文本数据类型作为数据来源的研究也在不断增多，未来如何基于多源数据来识别颠覆性技术将成为研究热点之一。

（2）数据预处理：在确定数据源之后，接下来则需要对获取到的数据进行预处理，这一步需要借助自然语言处理技术将大量的文本内容进行分词、去停用词、保留词、自定义技术术语词典等操作来完成文本内容的向量化、结构化处理。

（3）技术实体抽取：从处理完的结构化数据中抽取技术实体，这一步可以用到术语词分析的方法，依据词频进行高频术语词分布情况或时间序列上的分析，部分研究会按照具体需要进一步划分时间周期或进行时间切片的操作。

（4）科技文本降维、分类与分析：在识别出技术实体后，则需要对科技文本进行降维，大多数研究会使用主题模型从技术实体中进一步凝练技术主题。在抽取出相应的技术主题后，多数研究会进行技术主题分析，通过使用共词分析和文本聚类观察主题分布情况、演化趋势以及主题间的关联情况等，也有研究利用语义分析进行科技文本的分类后再进行技术主题分析或直接进入到技术特征识别这一步骤。

（5）颠覆性技术特征识别：技术特征识别是整个识别过程的核心，需要将文本内容特性与颠覆性技术特征相匹配，通常将颠覆性技术特征归纳为成长性、融合性、突变性、创新性、突破性等，而突变词检测的方法通常用于识别颠覆性技术的突变性特征，此外，共词分析、文本聚类和语义分析也常被用于颠覆性技术特

征的识别中。

综上，文本挖掘方法或与其他技术识别方法相结合可以提取出符合颠覆性技术属性的技术名称或主题，从而完成某技术领域的颠覆性识别。需要明确的是，现阶段的文本挖掘方法尚不够成熟，且由于文本内容存在语义结构复杂等问题致使语义理解上存在困难，因而无法实现全过程的自动化抽取与分析，在研究过程中往往还是需要依靠技术领域专家进行辅助，从而达到理想的识别效果。

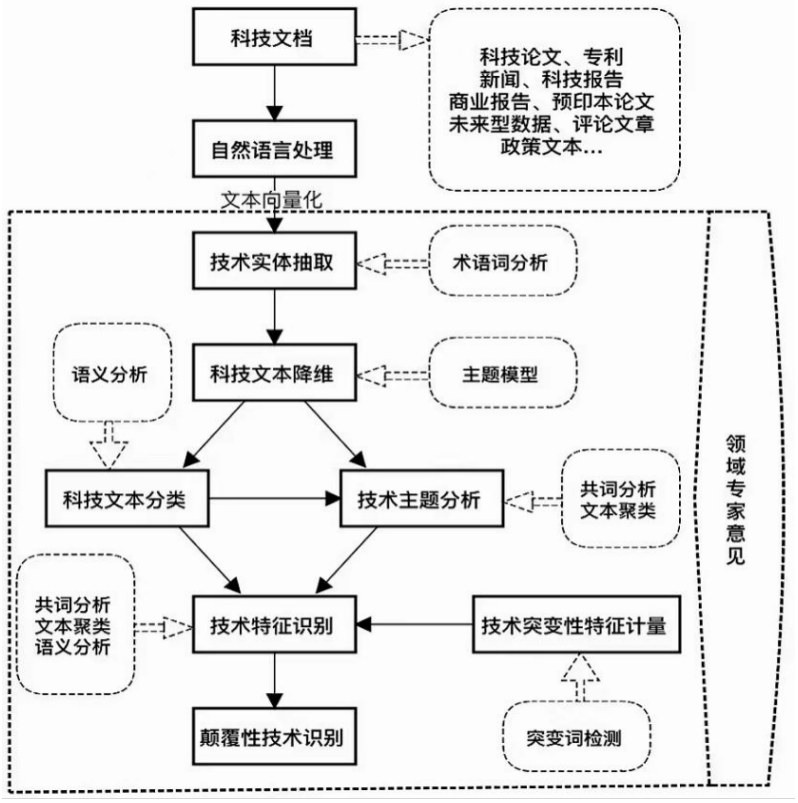

图 2　文本挖掘法在颠覆性技术识别过程中的应用

综合来看，文本挖掘法能够从微观层面探测到颠覆性技术的相关信息，相较于传统识别方法能够更好地发现潜在技术主题，随着机器学习、深度学习等新兴技术的快速发展，多种方法的组合使用成为必然趋势。但由于文本挖掘法在颠覆性技术识别领域的发展时间较短，可直接获取利用的工具有限，自动处理效率不高，需要结合专家意见进行干预。尽管已有研究验证了识别结果的可靠性，但未来如何进行技术优化还有待探索。

3　结语

颠覆性技术对社会和经济发展具有深远的影响，因此其准确识别具有重要意义。识别和研究颠覆性技术，有助于为科研、产业政策等领域提供有力的支持和指导。文本挖掘法能够从大量文本数据中自动提取关键信息，实现对颠覆性技术的准确识别，具有高效、自动化、可扩展性强等优势，可以有效应对大规模数据和多样化的颠覆性技术类型。本文通过文献梳

doi:10.3772/j.issn.2095-915x.2024.03.006

理将基于文本挖掘的颠覆性技术识别方法总结为六种：突变词检测、术语词分析、共词分析、文本聚类、主题模型、语义分析，并对六种方法进行了比较分析。发现文本挖掘法更注重从技术特征角度进行颠覆性技术识别，不少研究也证明了该方法的科学可靠性，然而缺乏市场环境、行业角度等综合视角的分析。其次，虽然文本挖掘方法在颠覆性技术识别中取得了一定的成果，但仍然存在一些挑战。例如，文本数据的质量和多样性，颠覆性技术的时效性和隐蔽性等都会对识别结果产生影响。此外，模型的可解释性、标注数据的获取等问题也需要解决。因而未来研究中需要更加关注改进文本挖掘方法来应对数据特点的变化、结合其他技术手段提高颠覆性技术识别的精度和可靠性、建立更加完善的数据集等方面的问题。此外，跨学科合作、开展实证研究等也是未来发展的重要方向。

参 考 文 献

[1] JIA W F, XIE Y P, ZHAO Y N, et al. Research on Disruptive Technology Recognition of China's Electronic Information and Communication Industry Based on Patent Influence[J]. Journal of Global Information Management, 2021, 29(2): 148-165.

[2] GUO J F, PAN J F, GUO J X, et al. Measurement Framework for Assessing Disruptive Innovations[J]. Technological Forecasting and Social Change, 2019, 139: 250-265.

[3] SUN J, GAO J, YANG B, et al. Achieving disruptive innovation forecasting potential technologies based upon technical system evolution by TRIZ[C]// International conference on management of innovation & technology. Bangkok: IEEE, 2008: 18-22.

[4] NILCHIANI R, EDWARDS C M, GANGULY A. Introducing a tipping point measure in explaining disruptive technology[C]//2019 International Symposium on Systems Engineering. Edinburgh: IEEE, 2019: 1-5.

[5] MULLER J M, KUNDERE R R. Ex-ante prediction of disruptive innovation: the case of battery technologies[J]. Sustainability, 2019(11): 5229.

[6] CHRISTENSEN C M, BOHMER R, KENAGY J. Will disruptive innovations cure health care?[J]. Harvard business review, 2000, 78(5): 102-112.

[7] BOWER J L, CHRISTENSEN C M. Disruptive Technologies: Catching the Wave [J]. Harvard Business Review, 1995, 73(1): 43-53.

[8] RAFII F, KAMPAS P J. How to Identify Your Enemies Before They Destroy You[J]. Harvard Business Review, 2002, 80(11): 115-123, 134.

[9] PARRY M E, KAMAKAMI T. The encroachment speed of potentially disruptive innovations with indirect network externalities: The case of e-readers[J]. Journal of Product Innovation Management, 2017, 34(2): 141-158.

[10] GANGULY A, NILCHIANI R, FARR J V. Defining a set of metrics to evaluate the potential disruptiveness of a technology[J]. Engineering Management Journal, 2010, 22(1): 34-44.

[11] CHRISTENSEN C M, MCDONALD R, ALTMAN E J, et al. Disruptive innovation: An intellectual history and directions for future research[J]. Journal of Management Studies, 2018, 55(7): 1043-1078.

[12] DANNEELS E. Disruptive technology reconsidered: A critique and research agenda[J]. Journal of Product Innovation Management, 2004, 21(4): 246-258.

[13] 卢光松，卢平．企业颠覆性技术路线图制定研究[J]. 科技进步与对策，2011, 28(11): 81-86.

[14] 栾恩杰，孙棕檀，李辉，等．国防颠覆性技术在航天领域的发展应用研究 [J]. 中国工程科学，2017, 19(5): 74-78.

[15] PORTER A L, ROESSNER J D, JIN X Y, et al. Measuring national 'emerging technology' capabilities[J]. Science & Public Policy, 2022, 29(3): 189-200.

[16] 中国政府网．国家中长期科学和技术发展规划纲　要 [EB/OL]. (2006-02-09)[2023-06-08]. http://www.gov.cn/jrzg/2006-02/09/content_183787_5.htm,

2021-07-12.

[17] TUSHMAN M L. Technological discontinuities and organizational environments[J]. Administrative Science Quarterly, 1986, 31(3): 439-465.

[18] 李平 , 臧树伟 . 基于破坏性创新的后发企业竞争优势构建路径分析 [J]. 科学学研究 , 2015, 33(2): 295-303.

[19] 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.

[20] ZHANG Y, CHEN M, LIU L. A Review on Text Mining[C] // IEEE Beijing Section. Proceedings of 2015 IEEE 6th International Conference on Software Engineering and Service Science (ICSESS 2015). Institute of Electrical and Electronics Engineers, 2015: 5.

[21] RONEN FELDMAN, IDO DAGAN. Knowledge Discovery in Textual Databases (KDT) [C]//In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining. Montreal: AAAI Press, 1995: 112-117.

[22] LI X, FAN M, ZHOU Y, et al. Monitoring and forecasting the development trends of nanogenerator technology using citation analysis and text mining[J]. Nano Energy, 2020, 71: 104636.

[23] MIAO Z, DU J, DONG F, et al. Identifying technology evolution pathways using topic variation detection based on patent data: A case study of 3D printing[J]. Futures, 2020, 118(2): 102530.

[24] WANG Y, FENG L, WANG J, et al. Technology Trend Forecasting and Technology Opportunity Discovery Based on Text Mining: The Case of Refrigerated Container Technology[J]. Processes, 2022, 10(3): 551.

[25] KLEINBE RG J. Bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery, 2003, 7(4): 373-397.

[26] CHEN C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377.

[27] 王秀红 , 王欣 , 王少凡 , 等 . 基于 SimCSE-LDA 和异常检测的颠覆性技术识别方法 —— 以农业机器人为例 [J]. 情报理论与实践 , 2023, 46(5): 135-143.

[28] 邢晓昭 , 任亮 , 雷孝平 , 等 . 基于专利主题演化的颠覆性技术识别研究 —— 以类脑智能领域为例 [J]. 情报科学 , 2023, 41(3): 81-88.

[29] DOTSIKA F, WATKINS A. Identifying potentially disruptive trends by means of keyword network analysis[J].Technological Forecasting Social Change, 2017, 119(6): 114-127.

[30] LI R, YU W, HUANG Q, et al. A new identify disruptive technologies algorithm based on technology develop network[J]. Mathematical Problems in Engineering, 2022, 2022(1): 7354535.

[31] LIU X, WANG X, LYU L, et al. Identifying disruptive technologies by integrating multi-source data[J]. Scientometrics, 2022, 127(9): 5325-5351.

[32] 单晓红 , 韩晟熙 , 刘晓燕 . 基于技术主题演化的颠覆性技术识别研究 [J]. 情报理论与实践 , 2023, 46(8): 113-123.

[33] MOMENI A, ROST K. Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling[J]. Technological Forecasting Social Change, 2016, 104: 16-29.

doi:10.3772/j.issn.2095-915x.2024.03.007

近 20 年高校图书馆知识服务研究热点与趋势

许惠　程川生

山东大学图书馆　青岛 266237

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 分析 2001—2022 年高校图书馆知识服务领域的研究现状、研究热点及研究前沿，以期为我国图书馆知识服务研究与实践提供借鉴与参考。[方法 / 过程] 检索 CNKI 平台中 2001—2022 年高校图书馆知识服务领域的相关文献，利用 CiteSpace 软件对其从时间、机构、作者、关键词四个方面进行文献计量，生成可视化知识图谱，呈现出国内高校图书馆知识服务领域研究热点、发展脉络及前沿趋势。[局限]2023 年文献尚未纳入，后续将持续追踪最新研究成果，完善本研究的发展趋势分析；限于文献篇幅，部分研究主题的脉络分析略为粗糙，有待深入细化。[结果 / 结论] 高校图书馆知识服务领域整体发文量以 2019 年为分界点，呈长期上升、近期下降趋势；各作者、机构进行独立研究的现象较为普遍；研究热点为信息素养、学科服务、知识服务、知识产权信息服务、数据管理服务、智库服务、数字学术服务、智慧服务，研究前沿为智慧服务、数字人文、数据素养、数字学术、知识产权、数据管理等方向。

关键词：高校图书馆；知识服务；学科服务；智慧服务；知识图谱

中图分类号：G258.6；G252；G35

Research Hotspots and Trends of Knowledge Service in University Libraries in the Past 20 Years

XU Hui　CHENG Chuansheng

Shandong University Library, Qingdao 266237, China

Abstract: [Objective/Significance] Analyze the current research status, research hotspots, and research frontiers in the field of knowledge service in university libraries from 2001 to 2022 to provide references and guidance for research and practice in the field of library knowledge service in China. [Methods/Processes] Retrieving relevant literature in the knowledge service field of university libraries from 2001 to 2022 on the CNKI platform, using the Citespace software to carry out bibliometrics on it from four aspects of time, institutions, authors and keywords, generating a visualized knowledge map to show the research hotspots, development trends and cutting-edge trends in the knowledge service field of domestic university libraries. [Limitations] The literature in 2023 has not yet been included, and the latest research results will be tracked in the follow-up to improve the analysis

基金项目　山东大学 2023 年度校级教育教学改革研究重点项目“数智时代未来学习生态建设”（2023Z21）；CALIS 2020 年度科研课题“高校用户知识服务新需求特征研究”（2020KT05）。

作者简介　许惠（1979-），硕士，馆员，主要研究方向为图书馆知识服务、情报分析服务、智慧图书馆建设；程川生（1974-），硕士，副研究馆员，主要研究方向为智慧图书馆建设、图书馆管理与服务，E-mail：chengchuansheng@sdu.edu.cn。

引用格式　许惠 , 程川生 . 近 20 年高校图书馆知识服务研究热点与趋势 [J]. 情报工程 , 2024, 10(3): 88-103.

of the development trend of this study. Due to the limitation of the length of the literature, the context analysis of some research topics is slightly rough and needs to be further refined. [Results/Conclusions] The overall volume of articles shows a long-term upward trend, followed by a recent decline. Since 2019, it has been decreasing significantly each year. It is common for authors and institutions to conduct independent research. The research hotspots include information literacy, discipline service, knowledge service, intellectual property information service, data management service, think tank service, digital academic service and smart service. The research frontiers include smart service, digital humanities, data literacy, digital scholarship, intellectual property, and data management.

Keywords: University Libraries; Knowledge Service; Subject Service; Intelligent Service; Mapping Knowledge Domain

引言

张晓林 [1] 于 2000 年提出，知识服务是以信息知识的搜寻、组织、分析、重组的知识和能力为基础，根据用户的问题和环境，融入用户解决问题的过程中，能够有效支持知识应用和知识创新的服务。该观点产生了重大影响，拉开了国内研究图书馆知识服务的序幕 [2]，成为学术界的研究热点并持续至今。依据知识服务的涵义，高校图书馆开展的学科服务、智慧服务、知识产权信息服务、知识管理服务、数字学术服务、科研支持服务等，是知识服务的不同表现形式，均可归入知识服务范畴。广大学者对高校图书馆各类知识服务进行了深入研究，根据内容可将研究成果分为两大类。第一类成果展现了知识服务的概念、内容与实现路径。知识服务、信息服务、学科服务、信息素养研究是较早出现的研究主题，随着信息技术升级迭代及图书馆的不断变革，产生了智库服务、数字学术服务、智慧服务、科研数据服务、数据素养等研究主题。各类知识服务研究主要关注概念界定与服务内容 [3-10]、服务体系 [11-13]、服务模式与服务机制 [14-17]、创新路径 [18-21]、平台建设 [22-25] 等理论研究、实践研究、案例分析与对策研究 [26-29]。第二类成果以宏观视角综述研究内容、分析发展趋势。高晓英 [30]、卢凤玲 [31]、罗星华等 [32] 分别在 2010 年代梳理了学科服务研究进展，初景利等 [33] 深入分析了国内外图书馆嵌入式学科服务相关研究和实践成果；卢章平等 [34] 于 2014 年提出以“资源发现”“知识发现”“用户发现”等服务内容和方式提供学科化服务和知识服务；2019 年詹婧和张仁琼 [35] 探讨了高校图书馆知识服务领域研究的演化特征和发展趋势，指出研究热点主要集中在知识管理、学科服务、学科馆员、数字图书馆、大数据等方面；唐晓波等 [36] 提出人工智能背景下知识服务领域研究热点包括数据安全、用户行为和服务表达等方面；黄蕾 [37]、孟艳丽 [38] 分别于 2015 年、2021 年研究了信息素养教育的研究热点及发展历程；慎金花等 [39] 于 2016 年分析专利信息服务趋势，从馆员队伍、业务规范和服务标准、政策支持和资质认证等方面提出发展建议；朱佳林 [40] 于 2022 年综述了高校图书馆知识产权信息服务研究热点与演化历程；陈晋 [41] 在 2019 年研究了科研数据管理服务的主要研究主题，指出未来研究方向应在科学数据素养教育、数据服务与知识发现、数据管理与知识产权等方面；邓灵斌 [42] 于 2020 年综述了科学数据服务，指出应加强图书馆科学数据服务的法律保障、技术实现和绩效评估等方面的研究。朱珍 [43] 于 2020 年梳理分

doi:10.3772/j.issn.2095-915x.2024.03.007

析我国图书馆智慧服务研究主题，从跨界研究、新技术应用研究、深化智慧服务研究等方面展望未来发展。上述研究产生了众多理论创新，为不同时期高校图书馆知识服务开展提供了强有力的理论依据。

2022 年 11 月，美国人工智能研究机构 OpenAI 发布 AIGC（人工智能内容生成）文本生成产品 ChatGPT，具备强大的对话能力和生成能力 [44]。2023 年，谷歌、百度、科大讯飞等公司相继推出了自己的 AIGC 产品。人工智能应用迎来突破性发展，将对知识生产、科研范式和知识学习产生变革性影响 [45]。此背景下，梳理高校图书馆知识服务研究历程，具有重要意义。调研发现，该领域相关成果多基于新技术、某类平台、特定政策背景研究某种知识服务，从单一类型知识服务视角综述研究进展或探讨一些关键问题，长期以来对高校图书馆知识服务成果开展整体分析不足，也很少通过量化分析全面展示领域综合研究状况。本文以中国知网（CNKI）核心学术期刊数据为研究对象，采用文献计量方法，挖掘高校图书馆所有类型知识服务研究成果数据，生成科学知识图谱，揭示我国高校图书馆知识服务领域近二十年研究脉络、研究热点及研究前沿，展望高校图书馆知识服务未来研究方向，以期对相关研究与实践提供参考。

1 数据与分析方法

1.1 数据来源

为保障文献查全率，本文将高校图书馆开展的所有知识服务类型作为检索词进行数据搜集，以中国知网（CNKI）数据库作为数据源，时间段选取 2001—2022 年，限定 CSSCI 和核心期刊论文进行主题检索。将“高校 + 大学”“图书馆”与“知识服务、学科服务、数字学术服务、智慧服务、学习支持、学术支持、科研支持、知识产权信息服务、出版服务、智库服务、数据服务、科研数据、情报服务、科研服务、信息素养、数据素养”组配成多种检索方案，检索获得文献 4280 条，经筛选去重后得到有效文献 3941 条。检索时间 2023 年 1 月 3 日。

1.2 研究方法

本研究根据文献计量法及科学知识图谱，利用 CiteSpace6.1.R6 对获取的数据集进行共现网络分析，揭示论文数据关键要素之间的关系。科学知识图谱是展示科学知识发展进程与结构关系的一种图像，可以直观显示知识单元或知识群之间网络、结构、互动、交叉、演化或衍生等隐含的复杂关系。CiteSpace 是近年学者在不同学科领域广泛应用的知识图谱绘制工具之一 [46]。

分析过程首先进行数据预处理，利用 CiteSpace 合并同义词及排除意义不明确关键词，比如同义词有“高校学生、大学生与本科生”“国外、美国、英国”“科研数据与研究数据”“慕课与 mooc”等，意义不明确关键词有“发展、特点、挑战、启示、建设”等，之后进行发文作者、发文机构、关键词等的统计分析与网络化分析。

2 文献统计与研究主题分析

2.1 年度发文量分析

根据图 1，整体发文量呈长期上升、

近期下降趋势，2019 年后逐年下降明显。2001—2003 年，曲线呈平缓趋势且每年发文量较少，相关研究此时处于起步阶段。2004—2009 年发文曲线呈逐年稳定上升趋势，说明关注度逐年增加，处于快速发展时期。2010—2017 年波动上升，且上升速度加快，此时研究领域迅速引起广泛关注。2014—2019 年发文量处于最高位，此时段国内高校图书馆知识服务研究处于高产高热时期。2020 年至 2022 年发文量逐年下降加速，可见近年相关领域研究热度有所降低，但发文量仍然可观。

图 1　2001—2022 年历年发文量曲线图

2.2　作者分析

利用 CiteSpace 以每 5 年为一个分区，计算合作关系。分析得到作者共现网络图谱，如“图 2”，图中节点越大，代表发文量越多，节点以年轮形式显示。本次图谱共得到 352 个节点，158 条边，网络密度为 0.0026，孤立节点较多，说明此研究领域作者合作强度较低，密切合作团体较少。根据节点颜色与大小判断，高发文量作者的发文大多集中在 2011—2020 年度。

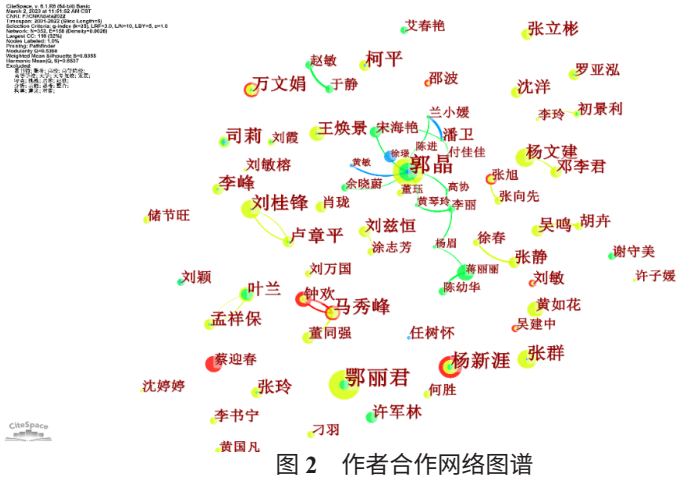

图 2　作者合作网络图谱

分析发现，近 20 年高校图书馆知识服务研究领域有 19 位学者发表文献在 10 篇以上，主要作者的论文数占论文总数的比例较低，发文集中度不高。发文量最多的是上海交通大学图

doi:10.3772/j.issn.2095-915x.2024.03.007

书馆的郭晶，共发表 28 篇论文。其次是燕山大学图书馆的鄂丽君，发表 25 篇论文。通过详细分析各团队的文献内容，归纳发文量 10 篇以上的合作团队或独立作者的研究主题，总结如表 1。

表 1　高产作者或团队及其研究主题（发文 10 篇以上）

研究作者	作者机构	研究主题
郭晶、潘卫、宋海艳	上海交通大学图书馆	学科服务模式、体系、创新及实证研究，学科服务平台构建，学科馆员能力、资质、培训研究，图书馆业务布局重构及理念创新，信息素养教育等。
鄂丽君、马兰	燕山大学图书馆	国外高校图书馆服务实践研究，数字学术服务，科研支持服务，数字学术教育，科研数据管理，开放存取与出版服务等。
刘桂锋、卢章平	江苏大学科技信息研究所	国内外科研数据管理实践与理论研究，数据治理模型研究，研究数据管理服务体系构建等。
柯平、陈媛媛	南开大学商学院	学科服务与学科馆员、图书馆学科发展、科研数据服务、数据馆员职业能力、图书馆服务等。
张群	江南大学图书馆	科研数据服务、科学数据素养研究架构及教育体系构建、数据素养教育模式与路径研究、用户科学数据素养需求分析、学科服务等。
杨新涯	重庆大学图书馆	阅读推广转型研究、信息素养教育模式研究、学科服务创新模式研究、知识产权信息服务设计及教育模式研究、知识管理共享平台研究等。
司莉	武汉大学信息管理学院	科研数据机构库研究、数据素养教育、嵌入式教学服务、科研支持服务、图书馆服务营销等。
万文娟	广西师范大学图书馆	学科服务体系构建及学科馆员、mooc 发展策略分析、智库服务策略、数据素养教育等。
杨文建、邓李君	重庆第二师范学院图书馆	数据素养教育、图书馆社会化服务策略、新环境下图书馆服务转型、数据馆员及智慧馆员培养策略研究等。
李峰	北京大学图书馆	决策支持服务、图书馆服务创新、信息素养教育路径及框架研究、学科服务、国内外数字出版服务调查研究、出版服务模式研究、科研数据管理服务研究等。
马秀峰、钟欢、董同强	曲阜师范大学，天津大学教育学院	智慧化及数据驱动背景下学科服务平台与服务空间构建研究、学科服务模型研究、数据驱动的学习支持服务模型研究、嵌入式创客服务平台设计等。
刘兹恒、苗美娟、涂志芳	北京大学信息管理系	学术图书馆数字出版服务、国内外开放存取出版服务的调查研究、科研数据管理与共享平台研究等。
叶兰	深圳大学图书馆	高校图书馆科研评价服务体系研究、大学图书馆服务变革、科研数据管理研究等。
孟祥保	东南大学图书馆	数据馆员与数据素养教育研究、科研服务与科研数据管理研究等

2.3　发文机构分析

利用 CiteSpace 对发文作者所在机构进行分析，得到机构合作网络图谱与机构发文量统计表，如图 3 和表 2 所示。图谱中，节点越大代表发文量越多，节点之间的连线代表合作关系。

由图 3 和表 2 可知，发文量 30 篇及以上的机构有 14 家，其中大学图书馆占比 57%。武汉大学信息管理学院发文最多，且显著高于排名第二的上海交通大学图书馆，合作机构同样最多，为 23 家，主要合作机构有武汉大学信息资源研究中心、武汉大学图书馆、武汉大学中国科学评价研究中心。其他合作发文较多的机构有，南京大学信息管理学院合作 19 家，中国科学院文献情报中心合作 17 家。南京大学信息管理学院、北京大学信息管理系、中山大学资讯管理学院、江苏大学科技信息研究所与本校图书馆合作发文均在 5 篇以上，合作频次相对较高。观察发现大学图书馆与外单位合作研究成果较少，大多合作仅发生一次。其中，东南大

学图书馆跨校合作成果最多，合作机构有深圳大学图书馆、中山大学图书馆、中山大学资讯管理学院（现更名为中山大学信息管理学院）、江南大学图书馆、东南大学情报科学技术研究所、西安交通大学图书馆。总体上高校图书馆知识服务研究领域高频次合作容易发生在同地域、同专业、同单位系统内部，跨地域、跨专业合作较少。

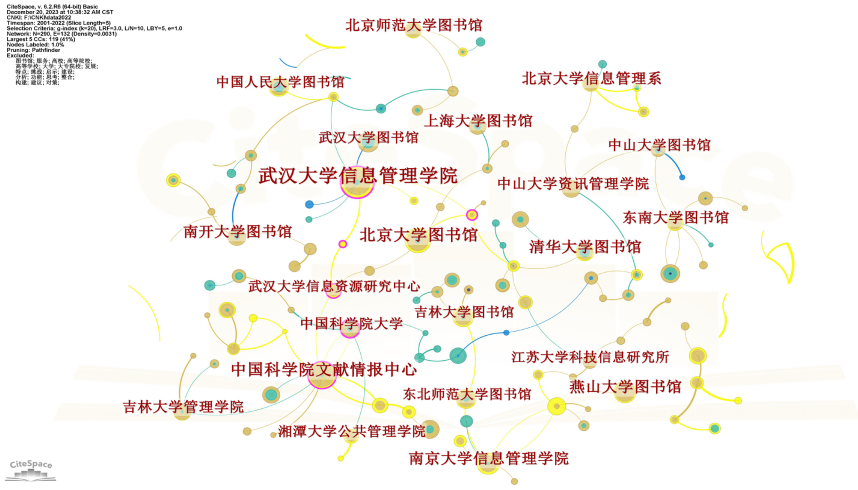

图 3　研究机构合作网络图谱

表 2　高发文研究机构（发文量 30 篇及以上）

发文数量	发文机构	发文数量	发文机构
102	武汉大学信息管理学院	36	上海大学图书馆
68	上海交通大学图书馆	36	北京大学信息管理系
50	北京大学图书馆	36	南开大学图书馆
50	中国科学院文献情报中心	33	北京师范大学图书馆
41	南京大学信息管理学院	33	吉林大学管理学院
41	燕山大学图书馆	30	东南大学图书馆
41	清华大学图书馆	30	中山大学资讯管理学院

3　研究热点及前沿分析

文献的关键词能够揭示和表达文献核心内容，因此，识别关键词共现网络中的核心节点能够确定研究热点。利用 CiteSpace 进行关键词分析，得到关键词共现网络图谱，如图 4 所示，图中节点大小代表关键词出现频次的多少，标签字体大小表明节点在网络中的重要程度。从软件导出高频关键词表，如表 3 所示。

统计分析发现，国内高校图书馆知识服务研究领域关键词非常多，各关键词之间联系紧密，表明研究范围广泛且不同研究方向的相关性较高。关键节点揭示出主要研究内容包括学科服务、信息素养、知识服务、学科馆员、信息服务、科研数据管理、智库、数据素养、大学生用户等，次要研究内容有出版服务、资源管理、知识管理、科研支持、智慧服务、数字人文、数据服务、知识产权、开放获取等。很多研究主题结合网络环境、“双一流”、大数据、人工智能、知识图谱、文献计量、慕课等概念或技术开展实践研究、理论研究或创新研究，在相关主题的制度体系、策略对策、服务模式、实证研究、热点分析、战略规划等方面产出了丰富的研究成果。

doi:10.3772/j.issn.2095-915x.2024.03.007

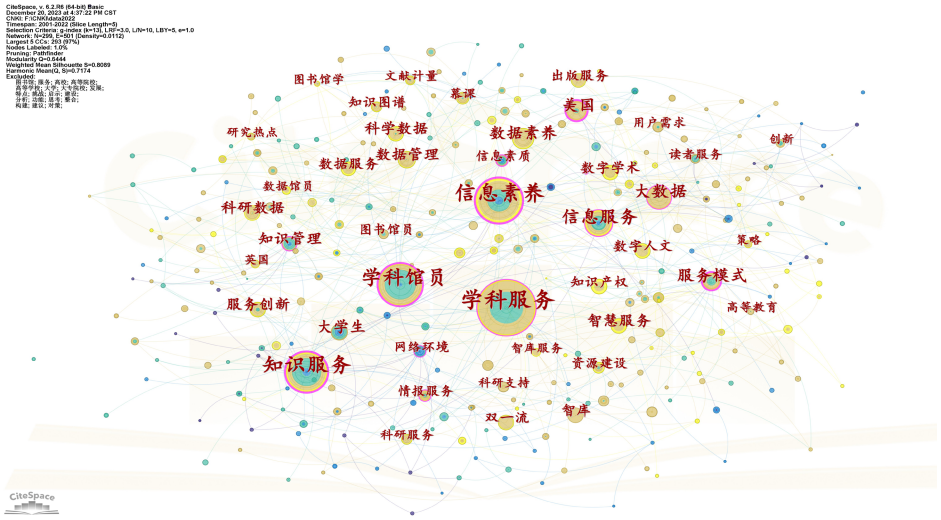

图 4　关键词共现网络图谱

表 3　高频关键词（出现频次 30 及以上）

出现频次	关键词	出现频次	关键词	出现频次	关键词	出现频次	关键词
483	学科服务	87	数据素养	52	科研支持	40	馆员
359	信息素养	83	大学生	45	文献计量	38	数字人文
289	知识服务	77	智库	45	需求分析	37	数字学术
280	学科馆员	70	服务创新	44	实证研究	37	创新创业
158	国外	59	数据管理	43	智慧服务	34	用户服务
133	信息服务	58	资源管理	41	对策	32	用户
106	科研数据	58	出版服务	41	一流高校	32	开放获取
90	服务模式	53	知识管理	40	数据服务	30	知识产权
88	大数据	-	-	-	-	-	-

3.1　以内容为视角的研究热点分析

利用 CiteSpace 软件的 LLR (Log-likelihood rate）算法，在关键词共现网络基础上进行聚类分析，得到聚类图谱，如图 5 所示。聚类产生的网络模块度（modularity）为 0.5313，意味着聚类可以较清楚地界定出高校图书馆知识服务的各个子领域，另一个评价指标平均轮廓值 (Silhouette) 是 0.8409，说明各聚类内部元素紧凑，聚类轮廓清晰度高。因此，本次关键词聚类分析能够反映高校图书馆知识服务领域的研究主题脉络与研究热点。同时，绘制主要研究热点历年发文量曲线对比图，如图 6 所示。

根据主要聚类结果及聚类标签所含关键词，归纳出信息素养、知识服务、学科服务、知识产权信息服务、数据管理服务、智库服务、数字学术服务、智慧服务八个研究热点主题，对本次数据内容进行详细阐述，揭示高校图书馆知识服务研究的发展演变历程。

3.1.1　信息素养研究

分析 #0 聚类，归纳研究热点主题有信息素养、数据素养等。

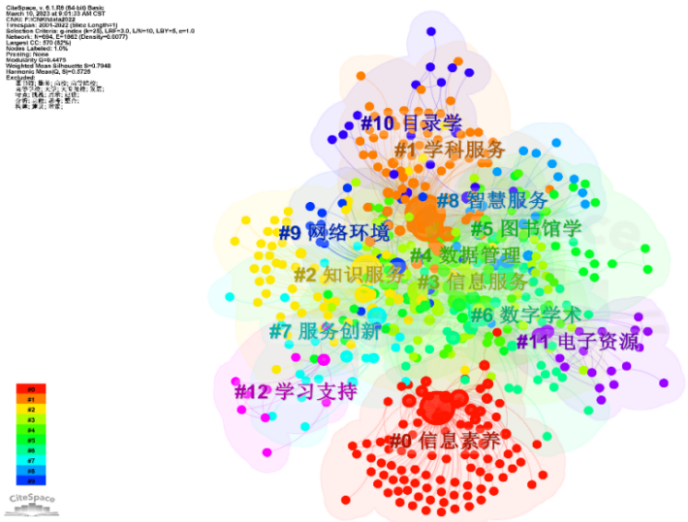

图 5　关键词聚类知识图谱

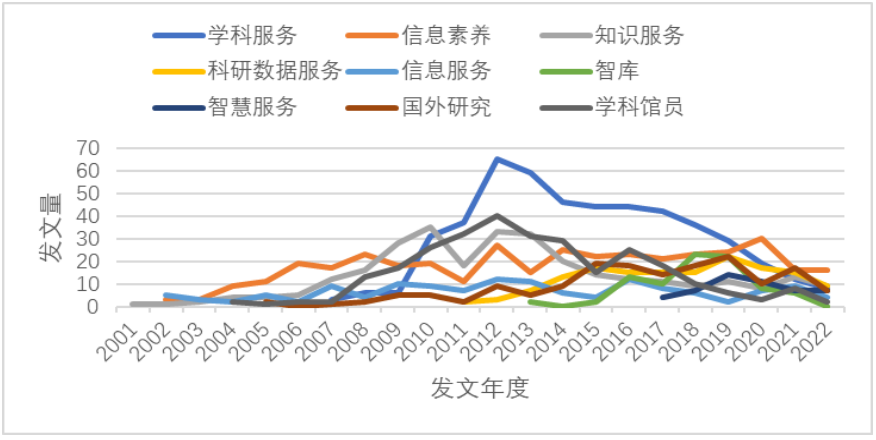

图 6　2001—2022 年主要研究热点历年发文量曲线图

高校图书馆信息素养研究成果于 2001—2006 年逐年快速上涨，2006 年以后发文量曲线缓慢波动持续上升，2020 年达历年最高 30 篇，此后两年有所下降。可见该主题长期引起学者关注，持续保持研究热度。2009 年以前，研究内容聚焦在信息素养概念的界定、信息素养教育内容、大学生的信息素养能力标准、信息素养教学模式探索等方向，2009 年之后，嵌入式教育、课程整合式教育、信息共享空间和学科服务等新概念的兴起，在线教育技术的蓬勃发展，为新阶段的信息素养教育提供了新的研究视角 [37]。此后，与专业结合、嵌入式课程、线上线下结合的培养模式，以及利用新技术的多样化课程模式的实施，有力促进了信息素养教育的发展。高校图书馆数字化进入新阶段，用户需求快速转变，促使信息素养研究进一步向数据素养、多元素养转型，传统信息素养教育开始融入其他素养类型 [38]。2020 年以来，研究主要侧重于支持教学学习的信息素养教育、国外案例调研分析、信息素养教学优化策略、在线信息素养教育研究等方面。其中，数据素养是对信息素养的延伸和扩展，是数据时代高校师生教学科研的必备素养，近几年在研究领域引起较高关注。国内相关研究成果于 2013 年首

doi:10.3772/j.issn.2095-915x.2024.03.007

次出现，2015 年起发文量快速上涨， 2016 年达到峰值 22 篇，之后持续保持较高热度。研究内容主要包括数据素养的涵义与内容、图书馆在数据素养中的角色、数据馆员能力培养、数据素养实践调查与需求分析、数据素养教育开展策略与模式、数据素养能力评价、国外数据素养教育经验研究等。

3.1.2 学科服务研究

#1 号聚类研究热点主题有学科服务、学科馆员、科研支持等。

高校图书馆学科服务研究成果 2006 年首次出现，初期发展缓慢，自 2009 年起研究成果数量增速加快，发文量在 2012 年达到峰值 65 篇，其间处于快速发展阶段，持续保持较高的研究热度。此后 10 年呈缓慢下降趋势，研究热度降低。分析各时期文献内容发现，2011 年以前，研究主题聚焦于高校图书馆学科服务的内容与模式研究、学科馆员制建设与学科服务的关系研究、学科服务平台建设研究、学科服务信息资源建设研究等 [30]。黄筱玲 [47] 指出 2011 年是学科服务重要转型期。近十年学科服务研究主要关注基于新平台新媒体的创新研究、双一流背景下的发展策略及评价研究、嵌入式学科服务研究、智慧图书馆背景下学科服务平台研究、学科化知识服务模式及新型学科馆员能力研究。

学科馆员研究于 2001 年起发文量逐年上升，发展迅速，2012 年发文量达峰值 40 篇，此后研究热度逐年降低，整体上与学科服务的研究成果走势吻合。2013 年以前，研究内容主要有学科馆员制度及管理模式研究、学科馆员服务层次研究、学科馆员角色定位、网络环境中的个性化服务等，此时出现了嵌入式学科馆员、学科馆员素质要求与队伍建设等研究热点 [48]。近十年相关研究内容主要为新技术背景下学科馆员能力及角色定位研究、学科馆员转型研究，比如智慧化学科馆员、数据馆员、数字人文及智库服务馆员研究等。

科研支持服务研究成果出现于 2007 年，至 2013 年发文量保持低位。2014 年起，研究成果显著增多，于 2017 年达峰值 10 篇，随后有所下降。早期研究内容侧重于阐述如何利用信息资源为教学科研服务，做好保障工作，与真正的科研支持服务差别较大。近十年，高校用户科研支持服务需求日益强烈，文献调研、科研数据管理、个性化信息推送、学科或交叉学科分析、知识产权分析、特色资源平台等方面都亟需图书馆提供帮助，因此不少学者对科研支持服务进行了深入探索，研究内容聚焦于国外科研支持服务实践经验与特征分析、国内科研支持实证研究、科研支持服务体系构建及策略研究、数字学术服务中的科研支持实现等 [11,49-50]。

3.1.3 知识服务研究

#2 号聚类研究热点主题以知识服务为主。

高校图书馆知识服务相关研究在 2001—2008 年处于起步阶段，2008—2014 年研究成果丰硕，可视为研究成熟期，2010 年发文量最高，达 35 篇，2015—2022 年间研究成果逐年减少至趋于平稳，近年每年发文 10 篇左右，关注度下降。早期阶段，研究内容侧重于信息服务向知识服务转变、知识服务的概念与特点、知识服务支持技术与运营模式、学科化的知识服务、知识管理等方面 [51]。成熟期主要研究内容聚焦于知识服务创新研究、知识服务与管理平台、知识服务策略与评价、新技术环境下知识服务

模式与体系构建、馆员知识服务能力建设[3]，其他次要研究内容有国外高校知识服务经验调研、用户需求分析、个性化知识服务等。近年，研究内容侧重于知识服务创新与实现路径、适应新形势的知识服务平台、知识服务跨学科发展等，比如微媒体知识服务营销、基于微信公众平台的个性化知识服务、智库知识服务、基于人工智能的知识服务、嵌入科研项目的知识服务、智慧图书馆知识服务、智慧型知识服务平台、用户行为画像分析、知识服务影响因素研究等[21,36,52-53]。

3.1.4 知识产权信息服务研究

分析 #3 号聚类，关键词揭示出主要研究方向为知识产权信息服务。

将知识产权信息服务与专利信息服务合并分析看出，相关研究起步于 2008 年，至 2017 年，每年产出成果较少且增长缓慢，2018 年起显著增多，2020 年急剧上涨至历年最高，为 27 篇，此后发文量持续保持较高水平。近年研究热度迅速升高，与国家知识产权战略发展的政策需要以及相关要求密切相关，比如 2016 年《国务院关于新形势下加快知识产权强国建设的若干意见》、2017 年《高等学校知识产权管理规范》等文件发布，多所“双一流”高校图书馆陆续成立知识产权信息服务中心，强化知识产权信息服务在科研成果产出与转移转化、企业创新等方面的支持作用，引起广大学者广泛关注。知识产权信息服务实践基本以“专利”为重心[54]，研究者主要关注知识产权信息服务内容及范畴、知识产权信息素养教育及培训、知识产权信息服务模式、知识产权信息服务评价及指标、高校图书馆知识产权信息服务实证研究等[55]。

3.1.5 数据管理服务研究

分析 #4 号聚类，归纳研究热点主题为科研数据管理服务、开放获取与出版服务等，且各研究主题均与高频关键词“国外”密切相关。

国内高校图书馆科研数据管理研究成果自 2011 年出现，2013 年起发文量加速上涨，2019 年达历年最高，为 16 篇。研究主题聚焦于科研数据管理描述与构建框架的基础研究、国外经验借鉴研究、政策与理论模型研究、科研数据服务平台与知识库建设研究、面向学科的科研数据引证分析与发表服务、数据馆员及用户教育培训研究、科研数据与知识产权等[41]。

开放获取与出版服务研究密切关联。高校图书馆开放获取相关研究成果最早出现于 2005 年，早期研究领域主要关注开放获取的定义与类型、开放获取对图书馆服务体系或资源建设的影响等内容，近年该领域研究聚焦于国外开放获取政策、数字出版、出版服务、机构知识库、科研数据、学术期刊等主题。数字出版与开放获取运动促使大学图书馆开始探索出版服务[56]，其作为术语第一次出现于 2012 年，十年来研究成果多为国外高校图书馆实践经验研究，调研学术图书馆出版服务的组织形式与合作模式、基于某一类出版服务的分析研究、有代表性的学术图书馆服务内容及开展情况、国外图书馆出版联盟等。国内高校图书馆出版服务的研究内容主要为图书馆参与出版的动因与角色定位、出版服务模式研究、面临的问题与对策等方向[57]。

3.1.6 智库服务研究

分析 #5 号聚类，关键词揭示出主要研究方向为智库服务。

doi:10.3772/j.issn.2095-915x.2024.03.007

高校图书馆智库服务研究成果 2013 年出现，2016 年起发文量加速上涨，2018 年达最高 23 篇，2020—2022 年发文量骤减，可知 2016-2019 年间其关注度较高，但近两年研究热度迅速下降。智库服务的三个主要过程，即开展信息资源的分析整合、进行智库产品的加工生产、实现智库服务的传播利用 [58]。相关研究成果主要聚焦于智库服务的内涵及可行性研究、智库服务模式、智库发展策略、智库建设实践等 [59]，重点关注图书馆如何参与高校智库建设、图书馆与智库智能的关联问题等 [60]。

3.1.7 数字学术服务研究

分析 #6 号聚类，归纳研究热点主题为数字学术服务、数字人文研究。

高校图书馆数字学术服务研究成果自 2017 年出现，2018 年发文量便达历年最高 12 篇，迅速吸引学者关注。研究方向聚焦于数字学术服务具体内容、数字学术馆员能力、数字学术服务模式与实施路径等，研究内容多为国外高校图书馆数字学术服务实践经验分析，侧重于调研数字学术空间、数字人文服务、科学数据服务、学术交流与数字出版、数字学术研讨与培训、数字科研工具与技术支持、相关部门组织结构设置等。有学者指出，数字学术基于数字人文发展起来，是数字人文发展的新阶段 [61]。本文分析过程虽然未将数字人文作为检索关键词，但是数字人文出现频次较高，反映出高校图书馆知识服务研究与数字人文的紧密关系。本次数据集中，高校图书馆领域数字人文相关研究成果自 2016 年出现，2018 年发文量最高，达 14 篇，同样发展迅速，研究内容主要为国外数字人文项目、数字人文服务经验借鉴研究、数字人文教育研究、馆员能力及用户需求分析、数字学术与学科服务主题中的数字人文服务等。

3.1.8 智慧服务研究

分析 #8 号聚类，归纳研究热点主题为智慧服务。

高校图书馆智慧服务研究成果自 2014 年出现，初期发展缓慢，2017 年起发文量加速上涨，2019 年达历年最高 14 篇，短时间内迅速引起关注。本次数据源中与智慧服务相关的研究成果，较多以智慧图书馆建设、智慧空间建设、智慧学科服务平台、智慧背景下的图书馆服务为研究重点，智慧服务作为研究成果之一表述。总体上，研究聚焦于智慧服务概念与内涵、智慧服务技术及服务模式、智慧服务馆员及用户、智慧服务质量评价、国外图书馆经验借鉴等 [62]，从技术应用、服务机制、组织模式、硬件基础等方面提出高校图书馆智慧服务发展路径或实施策略，以实现提升智慧服务质量，为读者提供便捷智慧服务的目的 [63-64]。

3.2 以时间为视角的研究前沿分析

在 CiteSpace 中将关键词聚类图谱转换得到关键词共现网络的时间线图谱，此图将聚类编号作为 Y 轴，文献发表年份作为 X 轴，清晰展现各个聚类发展演变的时间跨度和研究进程，如图 7 所示。此外，关键词突现图谱也与时间特征相关，它观测关键词在某一时段内出现频次的变化率即突现率，突现强度高的关键词叫作突现词。本次突现图谱设置显示 25 个突现词，如图 8 所示。strength 表示突现强度，begin 与 end 表示起始结束年度，年度线上的红色表示时间跨度。突现起始年度较近，结束年份在 2022

年的关键词，可视为当前高校图书馆知识服务领域的研究前沿主题。

分析图 7 和图 8 发现，2001—2022 年间，我国高校图书馆知识服务领域研究历程大致分为三个阶段：

第一阶段，2001—2010 年，在图书馆全面进入数字化、网络化时代的背景下，知识服务研究热点与重点主要是网络环境下的信息服务、大学生与信息素养教育、学科服务内容与模式、学科馆员制度建设、知识服务内容与模式、知识管理等主题。

第二阶段，2011—2015 年，移动互联网从出现到高速发展，大数据、智能化应用在各领域逐渐深入，此时高校图书馆知识服务研究重点与热点主要是新技术环境下的学科服务创新与学科馆员能力研究、慕课与信息素养教育、知识服务与管理平台、知识服务策略与评价、新技术环境下知识服务模式与体系构建、基于用户需求的个性化服务等内容。

第三阶段，2015—2022 年，大数据、人工智能技术的应用促进了智慧图书馆的快速发展，其间教育部发布了建设“双一流”高校、提升高校专利质量与促进转化运用等一系列政策，极大地推动了图书馆知识服务领域的现代化进程。观察图 8 可知，智慧服务、智库服务、科研数据管理服务、数据素养、数字学术服务、数字人文、数据馆员、出版服务、知识产权信息服务、人工智能、应急管理等研究方向近年受到极大关注，是增长势头强劲的研究主题。其中，数据素养、智慧服务、数字学术、知识产权、数字人文、出版服务、科研数据突现时间持续至 2022 年，且突现强度值明显高于其他主题，可视为高校图书馆知识服务领域的研究前沿，代表了国内的研究和发展趋势。

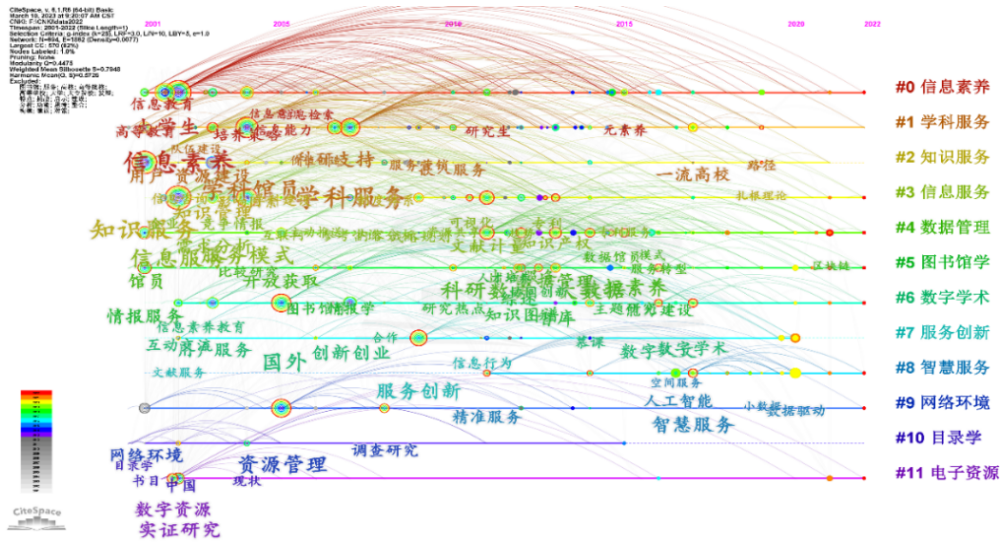

图 7　关键词时间线图谱

4　研究结论与趋势分析

本文利用 CNKI 核心学术期刊数据，从发文量、作者、机构、关键词等多维度量化分析 2001—2022 年国内高校图书馆各类型知识服务研究内容，探索高校图书馆知识服务发展现状

doi:10.3772/j.issn.2095-915x.2024.03.007

与演进趋势。从内容视角揭示出知识服务、学科服务、知识产权信息服务、数据管理服务、智库服务、数字学术服务、智慧服务、信息素养等主要研究热点，并分析展示研究热点的发展轨迹及主题演变历程。从时间视角整体梳理知识服务发展脉络，发现数据素养、智慧服务、数字学术、知识产权、数字人文、出版服务、科研数据等研究前沿。高校图书馆知识服务研究领域紧跟信息时代、数据时代、计算时代的发展步伐，与信息技术不断融合，从简单保守到丰富开放，在知识服务、学科服务、信息素养等研究方向上形成了相对完整的学科体系。研究前沿梯队的知识服务类型，均与数字化、智能化技术应用紧密相关，其研究领域的理论体系、实现路径与发展策略等不成熟不深入，尚处于初步探索阶段。

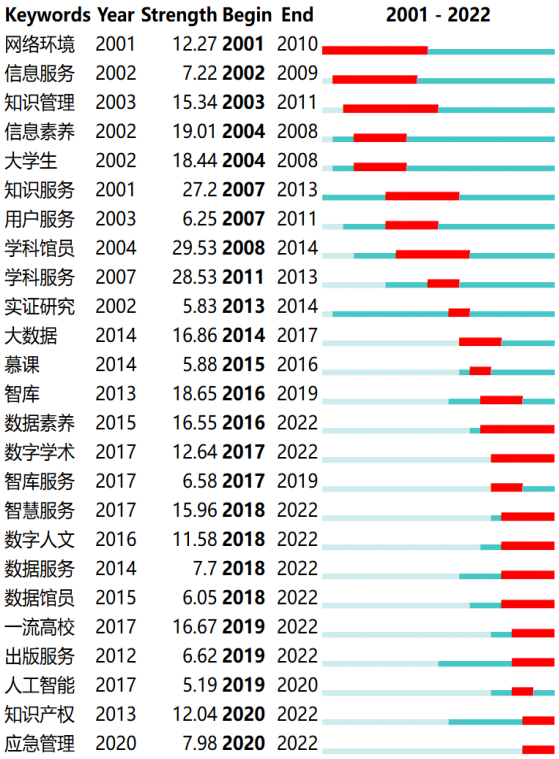

图 8　关键词突现图谱

从上述研究前沿主题可以推测，高校图书馆知识服务研究未来发展方向必将与图书馆正在经历的数字化、智慧化转型发展历程相契合。柯平 [65] 预测，到 2025 年，图书馆将实现基础的智能化，而到 2035 年，将迈入真正的智慧化时代。吴建中 [66] 指出图书馆正在从信息处理与服务向学术交流与研究的方向拓展，向知识基础设施的方向演变。教育部高等教育司 2023 年工作要点中明确提出探索推进未来学习中心试点，未来学习中心是智慧化学习空间，图书馆现有的资源、空间和服务都需要重塑服务场景 [67]。面对 AIGC 应用在各行业爆发的形势，张晓林等 [68] 指出知识服务的作用环境和生存机制都在发生巨变，未来知识服务领域可能的探索方向是如何提高知识服务的感知能力、认知决策能力、支持 AI for Science 的能力及改造发挥传统的文献资源能力 [69]。因此，为了应对正在或即将面临的诸多挑战，促进高校图书馆服务的转型升级，高校图书馆知识服务研究领域将加快探索融合新技术的知识服务，关注新路径、新用户、新需求，注重人、技术、智慧空间三方面互相融合，关注图书馆员的专业素质与能力要求，关注个人信息素养的与时俱进，积极拓展理论研究的广度和深度，开展广泛的实践研究。

5　研究建议

（1）深化 AI 驱动的智慧知识服务研究。新形势下，研究者应该针对学科服务、数字学术服务、数据管理服务、知识产权信息服务、智库服务等不同知识服务实施场景，在科技文献深度挖掘分析、智能化知识服务平台建设、

新型知识服务模式与体系构建、政策与理论模型标准化、数据安全与知识版权问题、服务质量评估与评价、用户需求识别与行为分析、馆员队伍体系与馆员能力研究等细分方向，开展深入探索，形成符合我国国情的理论体系及运行模式。

（2）开展面向知识服务的智慧空间构建研究。人工智能、VR、AR 等技术在图书馆行业的普遍应用，使得图书馆智慧空间建设成为可能 [70]。学者们应广泛开展面向智慧化知识服务的空间服务研究，探索线上虚拟空间、虚拟与物理融合的虚实结合空间等的建设路径，多模态资源关联融合，多元化服务模式等内容 [71]，助力高校图书馆融合空间、资源、平台、服务为一体，充分发掘海量资源，提供精准智慧服务。

（3）加强适应智能知识获取的数据素养研究。高校用户学习过程、科学研究、业务管理等各类环节繁多，数据素养应用场景复杂多样，新技术环境对个人的数据素养提出更高要求。广大学者应重视人工智能视域下的数据素养能力及数据素养教育研究，比如数据素养的概念内涵与特征、数据素养能力构成、数据素养评价体系构建、数据使用行为与需求、数据素养教育模式及体系设计、数据素养教育跨学科多元协作研究等方向 [72]，促进该领域研究的深化与完善。

参 考 文 献

[1] 张晓林 . 走向知识服务 : 寻找新世纪图书情报工作的生长点 [J]. 中国图书馆学报 , 2000(5): 30-35.
[2] 任萍萍 . 国内图书馆知识服务研究综述 (1999-2011)[J]. 图书情报工作 , 2012, 56(7): 5-10.
[3] 胡舒莉 . 高校图书馆知识服务环境及模式探究 [J]. 情报杂志 , 2008(3): 155-157.
[4] 徐恺英 , 刘佳 , 班孝林 . 高校图书馆学科化知识服务模式研究 [J], 2007(3): 53-55, 116.
[5] 张翔 . 基于 SERVICE 的嵌入式学科服务营销——武汉大学图书馆学科服务探索 [J]. 大学图书馆学报 , 2011, 29(5): 73-76.
[6] 张晓娟 , 张寒露 , 范玉珊 , 等 . 高校信息素养教育的基本模式及国内外实践研究 [J]. 大学图书馆学报 , 2012, 30(2): 95-101.
[7] 缪军 . 美国高校图书馆的典型数字学术服务与启示 [J]. 图书与情报 , 2017(4): 34-37.
[8] 宋维维 , 夏绍模 , 李赞 . 基于 SPARK 大数据处理平台的图书馆智慧服务探索与实践 [J]. 情报科学 , 2018, 36(6): 45-49, 56.
[9] 介凤 , 任树怀 . 科研数据联盟及其对我国科研数据管理的启示 [J]. 图书馆杂志 , 2014, 33(10): 49-53.
[10] 吴新年 . 学术图书馆的科研数据管理服务 [J]. 情报资料工作 , 2014(5): 74-78.
[11] 肖珑 , 张春红 . 高校图书馆研究支持服务体系 : 理论与构建——兼述北京大学图书馆的相关实践 [J]. 大学图书馆学报 , 2016, 34(6): 35-42.
[12] 万文娟 . 高校图书馆学科服务体系构建策略研究 [J]. 情报理论与实践 , 2012, 35(9): 50-53, 65.
[13] 吕源 , 陈慧琪 , 刘敏榕 . 高校知识产权信息服务内容体系构建研究 [J]. 图书情报工作 , 2022, 66(12): 59-67.
[14] 凌霄娥 . 数字人文下的图书馆开放数据服务机制分析 [J]. 图书馆 , 2021(5): 68-73.
[15] 徐璟 , 郭晶 . 高校图书馆学科化知识服务模式探究 [J]. 图书情报工作 , 2010, 54(17): 14-18.
[16] 张海涛 , 宋拓 , 刘健 . 高校图书馆一站式知识服务模式研究 [J]. 情报科学 , 2014, 32(6): 104-108, 113.
[17] 闫闵 , 过仕明 . 高校数字图书馆知识服务模式创新研究——基于读者行为偏好 [J]. 情报科学 , 2018, 36(5): 35-40.
[18] 徐璟 . Living library: 高校图书馆学科化知识服务的创新点——以上海交通大学图书馆为例 [J]. 图书情报工作 , 2010, 54(1): 58-61.
[19] 周剑 , 王艳 , XIE IRIS. 世代特征 , 信息环境变迁与大学生信息素养教育创新 [J]. 中国图书馆学报 , 2015, 41(4): 25-39.
[20] 柳益君 , 李仁璞 , 罗烨 , 等 . 人工智能 + 图书馆知

doi:10.3772/j.issn.2095-915x.2024.03.007

识服务的实现路径和创新模式 [J]. 图书馆学研究 , 2018(10): 61-65, 42.
[21] 吴爱芝 . 后疫情时代高校图书馆知识服务模式创新探索——以北京大学图书馆为例 [J]. 现代情报 , 2022, 42(5): 132-140.
[22] 郑峰 . 高校图书馆学科服务平台建设研究 [J]. 图书馆学研究 , 2012(17): 41-44.
[23] 钟欢 , 马秀峰 . 数据驱动下的学科精准服务平台构建研究 [J]. 图书馆学研究 , 2020(8): 50-55.
[24] 熊太纯 , 王勋荣 . 面向图书馆知识服务的智能互动平台研究 [J]. 图书馆学研究 , 2018(14): 38-42.
[25] 董同强 , 马秀峰 . 融入“双一流”建设的高校图书馆智慧型学科服务平台构建 [J]. 现代情报 , 2019, 39(5): 97-103.
[26] 袁红军 , 袁一帆 . “双一流”高校图书馆信息素养教育现状调查与分析 [J]. 图书馆学研究 , 2020(19): 20-28, 35.
[27] 徐春 , 张静 , 卞祖薇 . “双一流”建设背景下高校图书馆科研支持服务现状及发展对策研究 [J]. 图书馆学研究 , 2019(19): 71-78.
[28] 陈红艳 , 章望英 , 孙晶 . 我国“985”高校图书馆学科服务现状调查与分析 [J]. 高校图书馆工作 , 2012, 32(3): 85-89.
[29] 胥文彬 . 国外高校图书馆数字学术服务调查分析 [J]. 情报杂志 , 2021, 40(6): 187-192, 186.
[30] 高晓英 . 我国高校图书馆学科服务研究综述 [J]. 图书馆学研究 , 2012(22): 12-15.
[31] 卢凤玲 . 学科服务研究的现状、热点及趋势 [J]. 图书馆建设 , 2012(7): 55-57, 61.
[32] 罗星华 , 田苍林 . 近十年我国图书馆学科服务研究趋势与热点分析——基于内容分析法和文献计量学定律 [J]. 新世纪图书馆 , 2017(11): 59-64, 69.
[33] 初景利 , 孔青青 , 栾冠楠 . 嵌入式学科服务研究进展 [J]. 图书情报工作 , 2013, 57(22): 11-17.
[34] 卢章平 , 袁润 , 王正兴 . 发现服务 : 大学与研究机构图书馆的趋势 [J]. 中国图书馆学报 , 2014, 40(3): 20-26.
[35] 詹婧 , 张仁琼 . 高校图书馆知识服务研究现状及主题演化分析 [J]. 图书馆工作与研究 , 2019(5): 54-60.
[36] 唐晓波 , 李新星 . 基于人工智能的知识服务研究 [J]. 图书馆学研究 , 2017(13): 26-31.
[37] 黄蕾 . 20 年来国内信息素养教育研究与实践综述 [J]. 图书馆杂志 , 2015, 34(3): 16-22.
[38] 孟艳丽 . 数智环境下高校图书馆信息素养教育研究——基于可视化文献计量统计分析 [J]. 图书馆工作与研究 , 2021(9): 44-52.
[39] 慎金花 , 张更平 . 高校图书馆专利信息服务的趋势与思考 [J]. 大学图书馆学报 , 2016, 34(6): 51-55.
[40] 朱佳林 . 我国高校图书馆知识产权服务研究热点及前沿可视化分析 [J]. 图书馆工作与研究 , 2022(2): 77-84, 95.
[41] 陈晋 . 2008-2018 年我国科研数据管理服务研究述评 [J]. 图书馆工作与研究 , 2019(11): 78-84.
[42] 邓灵斌 . 我国图书情报界科学数据服务研究综合评述与展望 [J]. 新世纪图书馆 , 2020(6): 91-96.
[43] 朱珍 . 国内图书馆智慧服务研究综述 [J]. 图书馆工作与研究 , 2020(6): 62-68.
[44] 陈孝良 . ChatGPT 的现状、影响、启示和建议：大模型和多技能 [EB/OL]. (2023-02-14) [2023-08-20]. https://wap.sciencenet.cn/home.php?mod=space&do=blog&id=1376176.
[45] 张智雄 , 曾建勋 , 夏翠娟 , 等 . 回应 AIGC 的信息资源管理学人思考 [J]. 农业图书情报学报 , 2023(1): 4-28.
[46] 陈悦 , 陈超美 , 刘则渊 , 等 . CiteSpace 知识图谱的方法论功能 [J]. 科学学研究 , 2015, 33(2): 242-253.
[47] 黄筱玲 . 我国图书馆学科服务变革历程及特点 [J]. 高校图书馆工作 , 2019, 39(5): 24-27.
[48] 刘子辉 , 杨应全 , 吴晓英 . 近 30 年我国学科馆员的发展脉络及趋势研究 [J]. 图书馆理论与实践 , 2021(4): 31-36.
[49] 司莉 , 曾粤亮 . 世界一流高校图书馆科研支持服务调查与分析 [J]. 图书情报工作 , 2018, 62(8): 30-41.
[50] 陆美 , 顾烊 , 石明兰 . 高校图书馆科研支撑服务实践探索——以东南大学图书馆“量身定制的学者投稿指南”为例 [J]. 高校图书馆工作 , 2014, 34(5): 3-6.
[51] 靳红 , 程宏 . 图书馆知识服务研究综述 [J]. 情报杂志 , 2004(8): 8-10.
[52] 宋雪雁 , 张祥青 . 基于微信公众号的大学图书馆知识服务质量评价研究 [J]. 现代情报 , 2020, 40(2): 103-113, 152.
[53] 庞莉 . 智库型图书馆知识服务的需求分析及优化

策略 [J]. 图书与情报 , 2018(4): 105-110.
[54] 刘敏 , 喻萍萍 . 我国高校图书馆知识产权信息服务现状及优化策略 [J]. 图书馆学研究 , 2021(12): 51-59.
[55] 刘敏 . 我国高校图书馆知识产权信息服务研究综述 [J]. 图书馆学研究 , 2021(16): 2-8.
[56] 魏蕊 , 初景利 . 学术图书开放获取与美国大学图书馆出版服务 [J]. 大学图书馆学报 , 2014, 32(3): 17-22.
[57] 刘雅琼 . 近十年学术图书馆出版服务研究述评 [J]. 大学图书馆学报 , 2017, 35(2): 62-68.
[58] 陈静 , 陈茫 . 新型智库情境下高校智库服务的过程及机制研究 [J]. 图书馆 , 2020(10): 53-60.
[59] 梁新华 , 田硕硕 . 近十年高校图书馆智库服务研究综述 [J]. 图书馆 , 2020(10): 61-68.
[60] 龚雪竹 . 全球高校智库发展对图书馆服务的启示 [J]. 现代情报 , 2018, 38(10): 146-151.
[61] 王英 , 叶兰 , 赵雅 , 等 . 高校图书馆数字学术服务能力成熟度模型构建研究 [J]. 大学图书馆学报 , 2023, 41(5): 49-61, 68.
[62] 张旋 . 我国高校图书馆智慧服务研究现状分析 [J]. 图书馆工作与研究 , 2020(9): 109-116.
[63] 马静 . 云计算环境下高校图书馆智慧服务转型路径研究——以西安外国语大学图书馆为例 [J]. 图书馆工作与研究 , 2021(9): 69-75.
[64] 左平熙 . 大数据时代高校图书馆智慧服务的逻辑与路径 [J]. 图书馆工作与研究 , 2021(5): 48-54.
[65] 柯平 . 图书馆未来 2035 与“十四五”规划编制 [J]. 图书馆杂志 , 2020, 39(10): 13-17.
[66] 吴建中 . 未来二十年，图书馆是什么模样？ [J]. 图书馆杂志 , 2023, 42(11): 1-9.
[67] 杨峰 , 张雪蕾 , 李娟 . 未来学习中心对智慧图书馆建设的启迪与思考 [J]. 图书馆杂志 , 2023, 42(9): 35-43.
[68] 张晓林 , 梁娜 . 知识的智慧化、智慧的场景化、智能的泛在化——探索智慧知识服务的逻辑框架 [J]. 中国图书馆学报 , 2023, 49(3): 4-18.
[69] 张晓林 . 从猿到人：探索知识服务的凤凰涅槃之路 [J]. 数据分析与知识发现 , 2023, 7(3): 1-4.
[70] 周纲 , 陈晨 , 蔡文杰 , 等 . 虚实结合虚实相生的图书馆智慧空间服务 [J]. 图书馆建设 , 2023(4): 46-58.
[71] 李娇 , 孙坦 , 鲜国建 , 等 . 智能时代图书馆空间功能及服务创新的研究与思考 [J]. 图书情报工作 , 2022, 66(1): 138-145.
[72] 王静茹 , 郑燕林 , 马芸 . 国内数据素养的研究脉络与展望 [J]. 情报科学 , 2022, 40(5): 187-193.

doi:10.3772/j.issn.2095-915x.2024.03.008

基于关键特征增强的金融长文本事件分类

开放科学（资源服务）标识码（OSID）

王洁[1] 李旭晖[1,2]

1. 武汉大学信息管理学院 武汉 430072；
2. 武汉大学大数据研究院 武汉 430072

摘要：[目的 / 意义] 为了解决长文本模型输入长度限制问题，通过抽取事件关键句和事件关键词，对长文本进行关键特征增强，以提高模型的特征表示能力。[方法 / 过程] 基于关键特征增强的模型，在原文的基础上利用 TextRank 算法抽取事件关键句，并利用 TF-IDF 算法抽取事件关键词，将二者作为关键特征对长文本进行特征增强，再利用 BERT 和 Self-Attention 模型进行特征的进一步提取，最后进行事件分类。[局限] 模型仅在金融领域事件分类上进行实验，可以考虑在其他领域内也进行实验并进一步验证模型效果。[结果 / 结论] 在金融长新闻事件分类数据集上，提出的模型准确率达到 88.40%，比基准模型提升了 2 个以上的百分点，表明了模型的有效性。

关键词：事件分类；长文本分类；关键特征；特征增强；自注意力机制

中图分类号：G35；TP391

Key-Features Enhanced Financial Long Text Event Classification

WANG Jie[1] LI Xuhui[1,2]

1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Big Data Institute, Wuhan University, Wuhan 430072, China

Abstract: [Objective/Significance] In order to address the issue of length limitations in long-text models, this study enhances the feature representation capability of the model by extracting event-related key sentences and keywords from long text. [Methods/Processes] The key-features enhanced model utilizes the TextRank algorithm to extract key event sentences and the TF-IDF algorithm to extract event keywords from the original text. These key features are used to enhance the long text, and further feature extraction is performed using BERT and Self-Attention models, followed by event classification. [Limitations] The model in this study was only tested on event classification in the financial domain. It is recommended to conduct further experiments and verify the effectiveness of the model in other domains as well. [Results/Conclusions] On the financial long news event

基金项目 国家自然科学基金重点项目“基于知识关联的金融大数据价值分析、发现及协同创造机制”（91646206）；国家社科基金重大项目“文化遗产智慧数据资源建设与服务研究”（21&ZD334）。

作者简介 王洁（1998-），硕士研究生，主要研究方向为文本挖掘与金融文本分析；李旭晖（1975-），博士，副教授，主要研究方向为知识表示与自然语言处理，E-mail：lixuhui@whu.edu.cn。

引用格式 王洁 , 李旭晖 . 基于关键特征增强的金融长文本事件分类 [J]. 情报工程 , 2024, 10(3): 104-113.

classification dataset, the proposed model achieved an accuracy rate of 88.40%, outperforming other benchmark models by more than 2 percent, which demonstrates the superiority of the model.

Keywords: Event Classification; Long Text Classification; Key Features; Feature Enhancement; Self-Attention Mechanism

引言

金融新闻事件实时影响着市场参与者的情绪和决策，进而影响股票市场走势以及其他金融活动。对金融新闻事件进行分类有助于市场投资者快速掌握影响金融市场的重要事件，帮助其更好地进行分析和决策。在本研究中，一篇金融新闻即是对一个事件的报道，对金融事件进行分类即是根据新闻中报道的事件类别对文本进行分类。一方面，与传统的领域文本分类有所不同，事件分类主要关注的是领域内部的事件类别，根据文本中描述的具体事件将其分为某一事件类别，其分类更依赖事件要素相关词的区分。另一方面， 金融新闻文本中有许多篇幅较长的文章，要想对这类长文本进行分类，传统方法一般采用 TF-IDF 算法或 word-2vec 模型对文本进行向量化表示后再输入模型中训练，但其分类效果存在一定的局限性。以 BERT[1] 为代表的预训练语言模型在多项 NLP 任务上均超过了传统方法，但考虑到模型计算效率问题，BERT 模型的输入长度限制在 512 个 token[2]，这使得长文本无法完全输入 BERT 中进行表示。

基于以上问题和分析，针对长篇幅文本无法完全输入 BERT 模型的问题，本文提出从长文本中抽取出事件关键句的方法对原事件文本进行特征降维，达到将长文本中的关键信息抽取出来的目的。同时考虑到事件文本中的动词和名词往往更有可能携带关键的事件信息和事件要素，比如事件触发词和事件主客体等，本文采用 TF-IDF 算法从原长文本中抽取出关键词，再从中筛选出动词和名词，作为事件关键词。利用事件关键句和事件关键词中携带的关键信息作为原文的特征增强，有效地解决了长文本向量化表示的问题，并使模型学习到了更多关键信息，提升了金融新闻长文本事件分类的效果。本文的主要贡献如下：（1）针对文本过长无法完全输入模型的问题，采用关键句抽取的方法，将长文本进行特征降维，抽取出长文本中重要性更高的句子，作为原文的补充特征；（2）根据事件文本的特性，即动词和名词携带了更多事件信息，利用 TF-IDF 算法抽取出来的关键动词和名词作为事件关键词，对事件特征进行增强；（3）构建了金融领域内长文本事件分类的模型框架，将关键句特征和关键词特征作为原文的增强特征，提高了模型的特征表达能力。

1 相关工作

目前文本分类的常用方法主要可以分为基于传统机器学习的方法和基于深度学习的方法。早期文本分类任务使用的传统机器学习方法包括 NB[3]、SVM[4]、kNN[5] 等，而深度学习方法以 CNN 系列模型 [6] 和 RNN 及其变体 [7] 为主。自 Transformer 提出以来，预训练模型在 NLP 领域逐渐兴起。预训练模型分为以 BERT 为代表的自编码模型和以 GPT[8] 为代表的自回归模

doi:10.3772/j.issn.2095-915x.2024.03.008

型，GPT 主要用于文本生成任务，而 BERT 是一种基于双向 Transformer 构建的预训练模型，能够学习到文本的上下文特征表示。自提出以来，BERT 模型被广泛地应用到各 NLP 任务中，比如谌志群等 [9]、李颖 [10]、孙红等 [11] 均使用 BERT 模型进行文本向量化表示，再结合 CNN、LSTM、GRU 等模型进行下游任务，效果较传统方法也有显著提升。注意力机制因其能提升关键词的权重，也常被应用到文本分类相关研究中。Yang 等 [12] 在词和句级别分别应用注意力机制，使其能够在构建文档表示时关注更重要的内容。Shen 等 [13] 提出了一种用于语言理解的定向自注意力网络，使用注意力机制来学习句子嵌入。Basiri 等 [14] 提出了一种基于注意力机制的双向 CNN-RNN 深度模型用于情感分析。

针对长文本篇幅过长，无法完全输入 BERT 模型进行训练的问题，有相关研究通过句向量压缩或者分段等方式对长文本进行处理。叶瀚等 [15] 提出了一种句向量平均池化法以及注意力机制加权法对分类特征向量进行压缩编码进行长文本分类。卢玲等 [16] 将文本中的句子表示为段落向量，构建段落向量与文本类别的注意力模型计算句子的注意力，将句子注意力的均方差作为其对类别的贡献度，然后输入 CNN 中实现分类。鲍闯等 [17] 按照文本结构划分长文本，融合卷积最大池化特征向量和 BERT 句向量生成最终句向量，最后利用 Bi-LSTM 和注意力机制进行文本分类。长文本分类的相关研究较多采用对长文本切片进行文本表示，再对句向量进行平均等方式进行特征压缩，这种方式虽然能够利用到全局的文本信息，但是无法突出全文的关键信息，且对长文本的计算复杂度较高，对计算资源的要求也比较高。

事件分类和事件要素的抽取是构建事件知识图谱的基础，各大领域都有构建领域事件知识图谱的应用需求，比如自然灾难、体育赛事、历史事件等，金融事件也是其中一个重要的研究领域。Jacobs 等 [18] 构造了一个金融领域的英文新闻事件分类数据集，将金融事件划分为十大类别，包括买入评级、债务、股息、并购、盈利、季报、销售量、股份回购、目标价位、营业额，并在该数据集上分别采用线性 SVM 和 RNN-LSTM 进行实验。Jacobs 等 [19] 利用拥有 18 个类别和 64 个子类别的事件标注系统进行迭代标注，构建了一个英文金融事件分类数据集 SENTiVENT，并在 BERT 和 RoBERT 上进行了实验。Bhokare 等 [20] 对现有研究进行总结后将金融事件分为 11 类，并使用机器学习模型和 BERT 系列模型进行实验。已有研究大多基于英文新闻文本进行金融事件分类，而中文金融事件分类数据集几乎没有；同时，现有的金融事件分类研究重心在于金融事件类别的系统划分上，而没有对金融事件文本的特性进行分析，没有考虑如何提高金融事件分类的准确性。

2 关键特征增强的长文本事件分类

本文提出的基于关键特征增强的长文本事件分类模型整体结构如图 1 所示。整个模型由事件关键特征提取、事件关键特征增强、自注意力机制层、全连接层和 Softmax 分类层四个部分组成。整体流程为：（1）利用 TextRank 算法从原文中抽取事件关键句，再利用 TF-IDF

算法从原文中抽取出关键动词和名词作为事件关键词；（2）将原文、事件关键句、事件关键词分别输入 BERT 模型获得文本表示，再进行特征融合；（3）将增强后的特征输入自注意力层进行重要特征的进一步提取；（4）输入全连接层和 Softmax 激活函数进行事件分类。

2.1 事件关键特征提取

事件关键特征提取主要分为两部分：事件关键句提取和事件关键词提取。本文采用 TextRank 算法进行文本关键句抽取，得到的事件关键句能够表征长文本全局的主要信息，对文本长度过长无法输入模型的新闻文本进行特征补充。针对金融新闻事件分类任务，本文采用 TF-IDF 算法对全文词语计算重要性权重并排名，同时考虑到能够表征领域事件的词汇往往是领域名词和动词，本文仅筛选 TF-IDF 权重较高的动词和名词作为事件关键词。

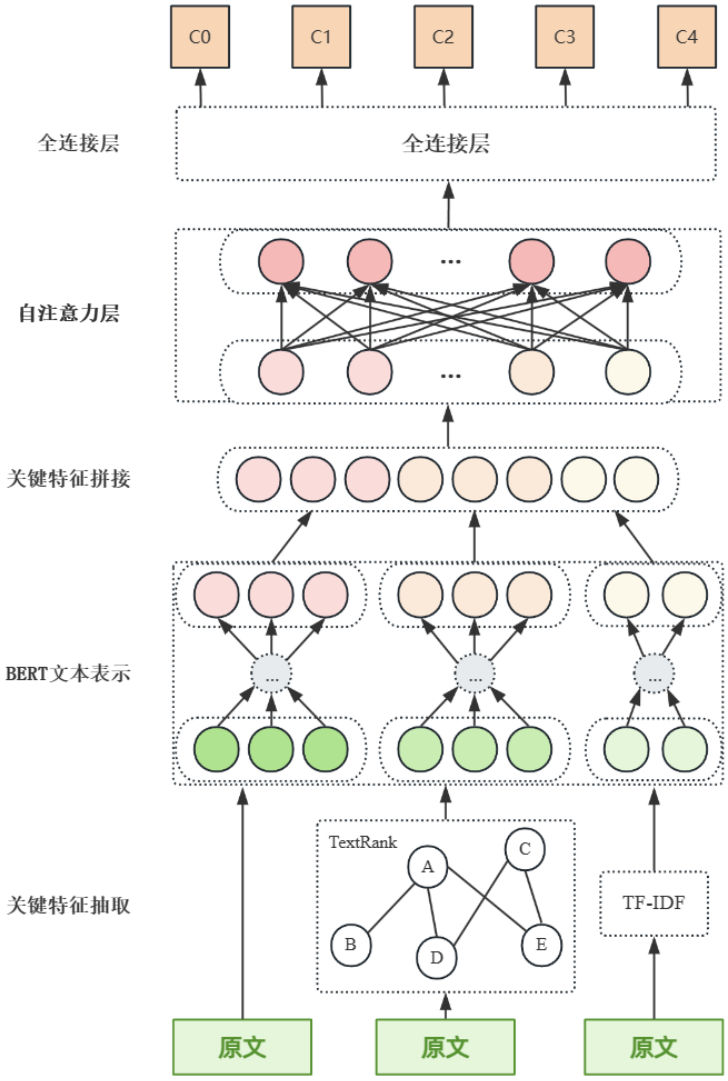

图 1　关键特征融合的金融长文本事件分类模型

doi:10.3772/j.issn.2095-915x.2024.03.008

2.1.1 事件关键句提取

TextRank[21] 算法源自 Google 提出的 PageRank 算法，本文使用的 TextRank 算法将长文本中的句子类比为 PageRank 算法中的网页，通过构建句子间的图结构关系并进行迭代计算，可以得到文本中比较重要的若干个句子作为新闻长文本的事件关键句。TextRank 的计算公式如下：

$$WS(v_i)=(1-d)+d\times\sum_{v_j\in In(v_i)}\frac{w_{ij}}{\sum_{v_k\in Out(v_j)}w_{jk}}WS(v_j) \tag{1}$$

其中，$WS(v_i)$ 表示节点 v_i 的权重，d 为平滑因子，表示在图结构中从一个节点转移到另一个节点的概率。$In(v_i)$ 表示节点 v_i 所有前驱节点的集合，$Out(v_j)$ 表示节点 v_j 所有后继节点的集合。w_{ij} 表示节点 v_i 和节点 v_j 之间的权重。

同时，本文采用以下公式来计算句子之间相似性作为节点之间边的权重：

$$Softmax(z_i)=\frac{\exp(z_i)}{\sum_j\exp(z_j)} \tag{2}$$

其中，$Sim(s_i, s_j)$ 表示句子 s_i 和句子 s_j 的相似度。w_k 表示两个句子的词集合中的第 k 个词，$|s_i|$ 和 $|s_j|$ 分别表示句子 s_i 和句子 s_j 的词的数量。句子 s_i 和 s_j 中相同词的个数越多，两个句子的相似度就越高；分母是对长度较长的句子进行一定程度的遏制，防止其因为长度优势而使句子间的相似度偏高。

2.1.2 事件关键词提取

事件关键词相对关键句而言更为短小精悍，核心语义信息更为集中，对事件文本进行关键词抽取能够从更细的粒度上对原文进行特征提取，提取出来的事件关键词能够进一步地对原文特征和事件关键句特征进行补充。同时，考虑到事件文本中的动词和名词往往携带了更多事件相关的信息，本文仅筛选 TF-IDF 算法排序靠前的动词和名词作为事件关键词。首先将原文经过分词、词性标注、去停用词等预处理步骤，然后采用 TF-IDF 算法对原文进行关键词重要性排序，并从中筛选出排名前 50 的动词和名词，得到事件关键词。

2.2 事件关键特征增强

自然语言文本需要转换成词向量表示，才能输入模型进行训练，本文采用 “bert-base-chinese” 模型进行文本的向量化表示。通过 TextRank 算法进行事件关键句抽取，得到长文本下的全局关键信息；同时，根据 TF-IDF 算法得到能够表征文档事件核心语义信息的事件关键词。至此，我们将原文 T、事件关键句 S 和事件关键词 W 分别输入 BERT 模型，得到词向量嵌入表示 F_t、F_s、F_w，然后将 F_s 和 F_w 拼接到原文的词向量 F_t 后面，得到进行特征增强后的文本表示 F_e。计算公式如下：

$$F_e=F_t\oplus F_s\oplus F_w \tag{3}$$

$\oplus$ 表示向量之间的拼接操作，上述公式将 F_s 和 F_w 拼接到 F_t 后面得到 F_e，得到经过事件关键特征增强之后的向量表示。

2.3 自注意力层特征提取

本步骤使用自注意力机制模型对得到关键特征增强之后的文本表示进行进一步的特征提取。具体而言，对于序列中的每个元素，自注意力机制计算该元素与其他元素之间的相似度，并将这些相似度归一化为注意力权重。然后，通过将每个元素与对应的注意力权重进行加权

求和，可以得到自注意力机制的输出。注意力机制的核心公式为：

$$Attention(Q,K,V)=Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

其中，Q 表示 query 矩阵，K 表示 key 矩阵，V 表示 value 矩阵，d_k 表示 k 矩阵的维度，除以 $\sqrt{d_k}$ 是为了使训练过程中的梯度值保持稳定。

2.4 全连接层和Softmax分类层

在经过上一个步骤的特征提取后，将得到的特征向量输入全连接层，再通过 Softmax 激活函数进行文本分类。全连接层在模型中的作用是可以有效地降低文本特征信息的损失。激活函数采用 Softmax 函数，将全连接层的特征向量输出作为 Softmax 函数的输入，然后进行金融新闻长文本的事件分类。Softmax 函数的具体计算公式如下：

$$Sim(s_i,s_j)=\frac{\left|\left\{w_k \mid w_k \in s_i \ \&\ w_k \in s_j\right\}\right|}{\log(|s_i|)+\log(|s_j|)} \tag{5}$$

3 实验与分析

3.1 实验数据准备

（1）实验环境

本文实验环境配置参数情况如表 1 所示。

表 1 实验环境配置参数

环境	配置参数
处理器	AMD Ryzen Threadripper 3970X 32-Core Processor 3.70 GHz
显卡	NVIDIA GeForce RTX 4090
内存	192G
编译器	PyCharm，Python 3.10

（2）数据准备

本文选取 5 类金融事件（人事变动、经营风险、证券市场风险、股权增减持、盈利能力），利用 python 网络爬虫方法，根据这 5 类事件标签从金融新闻资讯网站中获取实验数据，经过初步整理后得到每类 2000 条，共计 10000 条数据作为本次实验的数据集，并按照训练集：验证集：测试集为 8 ： 1 ： 1 的比例进行划分。同时，本文统计的文本数据长度分布情况如表 2 所示：

表 2 数据集信息

统计项	统计值
平均长度（字符）	1093
中位数（字符）	572
最短长度（字符）	68
最长长度（字符）	28276

由表 2 可见，金融新闻文本长度较长，最长达 28276 个字，平均长度为 1093 个字，中位数为 572 个字，这表明有一半以上的文本其长度超过了 BERT 模型所能处理的上限，超过的部分只能截断。因此，本文有必要对长新闻文本进行关键特征的抽取，以提高模型的事件分类效果。

3.2 实验设置与评价指标

（1）实验设置

实验参数设置如表 3 所示。

表 3 实验参数设置

参数	设定值
max_length	512
事件关键句数量	10
事件关键词数量	50
batch_size	32
epoch	15
optimizer	Adam

（2）评价指标

实验采用文本分类模型较为常用的查准率

doi:10.3772/j.issn.2095-915x.2024.03.008

（Precision）、查全率（Recall）、调和平均值（F1 值）和准确率（Accuracy）作为本文模型的评价指标。其计算公式如下：

$$P=\frac{TP}{TP+FP} \tag{6}$$

$$R=\frac{TP}{TP+FN} \tag{7}$$

$$F1=\frac{2PR}{P+R} \tag{8}$$

$$A=\frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

3.3 实验结果与分析

3.3.1 文本输入长度对分类模型效果的影响

为了验证输入文本长度对分类模型效果的影响，本文以 BERT+Self-Attention 作为基础分类模型，选择文本输入长度为 128、256、512 分别进行对比实验，实验结果见图 2：

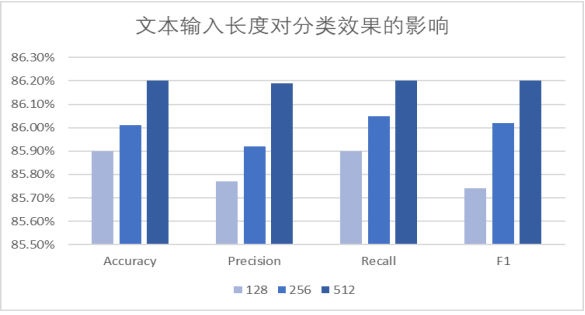

图 2 文本输入长度对比实验结果

由表中不同输入长度得到不同的模型效果，不难看出输入文本长度越长，模型能够从原文中获取到的信息越多，模型的分类效果也就越好。因此，针对 BERT 模型输入文本长度不超过 512 的限制，本文有必要通过关键特征抽取的方式，将全文的主要关键信息尽可能地输入模型中进行训练。

3.3.2 关键特征增强对分类模型效果的影响

在本部分对比实验中，为了探究关键句特征和关键词特征对分类模型的影响，本文以 BERT+Self-Attention 分类模型为基准，设置了 3 组对照实验，分别为：原文、原文＋关键句、原文＋关键词，实验组设置为：原文＋关键特征。同时为了验证关键句和关键词作为特征补充的思路的正确性，本文还设置了对比实验：关键句、关键词。实验结果对比如表 4 和图 3 所示：

表 4 关键特征增强的对比实验结果

模型	Accuracy	Precision	Recall	F1
原文	86.20%	86.19%	86.20%	86.08%
关键句	85.10%	85.17%	85.10%	85.02%
关键词	83.60%	83.52%	83.60%	83.39%
原文+关键句	87.20%	87.12%	87.20%	87.14%
原文+关键词	87.40%	87.46%	87.40%	87.30%
原文+关键特征	88.40%	88.33%	88.40%	88.28%

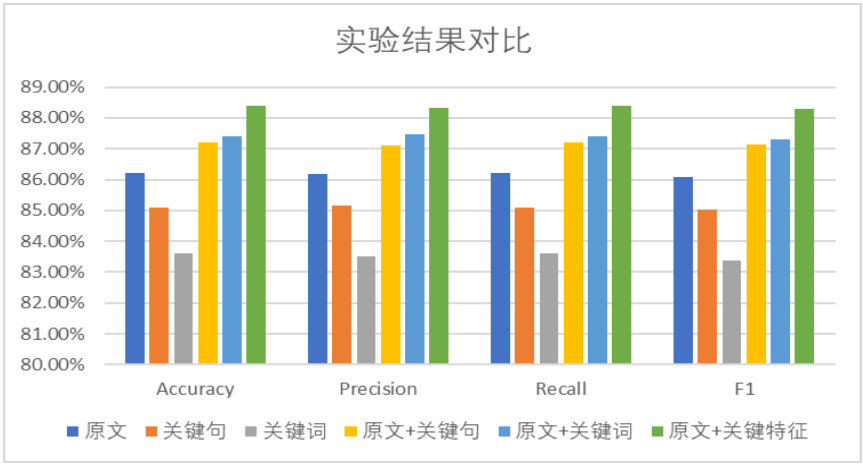

图 3 关键特征增强的对比实验结果

从表中的实验结果对比中可以看到，总体而言，本文提出的基于关键特征增强的长文本分类模型（原文＋关键特征）比基准模型（原文）的准确率提升了 2.2 个百分点，F1 值提升了 2.2 个百分点。从单个关键特征的角度来看，事件关键句特征增强方法（原文＋关键句）比（原文）的准确率和 F1 值均提升了 1 个百分点，表明本文提出的从长文本中抽取关键句进行特征增强的思路是可行且有效的，事件关键句中携带的文本信息能够对原文起到特征补充的作用；事件关键词特征增强方法（原文＋关键词）比（原文）的准确率和 F1 值均提升了 1.2 个百分点及以上，这表明本文抽取出来的事件关键词能够代表事件关键特征，能够携带事件相关信息。在与基准模型的对比之下，本文提出的将两个关键特征作为原文补充特征的模型效果最好，表明了该模型在领域内长文本事件分类任务上的可行性和有效性。

同时，观察到仅用关键句或关键词的分类效果并不如使用截断后的原文，推测可能的原因是事件关键句是从原文中按重要性排序抽取出来的句子，原本的语句顺序被打乱导致句子可能不具有语义连贯性；而仅用关键词输入模型的效果相对关键句的效果更差一些，原因是输入的文本为不连贯的词语，由词语本身携带的语义信息无法代表完整的文档信息，仅能突出部分关键特征。但将关键句和关键词作为原文的特征补充输入模型的效果均比原文有提升，表明以关键特征作为原文补充特征的思路是有效的，同时也表明 BERT 语言模型对文档全局语义信息的处理和利用是较为充分的，相比语义信息较为零散的单个句子和词语，其更擅长从全文中获取完整而连贯的语义信息。

表 5 展示了五个事件类别对应文档中提取出来的事件关键词，原文和事件关键句的详细信息由于篇幅过长不便展示。从表 5 中可以看到，人事变动事件类别文本中抽取出来的关键词包含“副行长”“行长”“监事长”等与企业人事职位变动相关的名词，也包括“担任”“辞职”“辞去”等职位变动相关的动词；经营风险类别中，包括“减值”“爆雷”“风险”等与公司经营风险相关的事件关键词；证券市场风险类别中包含“股价”“下跌”“收盘价”等与证券市场相关的专用词汇；股权增减持类别中包含了“减持”“股本”“持有”“股东”等与公司股东增减持事件相关的词；盈利能力类别中包括了“增速”“净利润”“扣非”“归母”“业绩”等描述企业业绩与盈利能力的关键词。这些事件关键词能够表达文本的核心语义信息，仅仅从几个简单的事件关键词便能勾勒出事件的大致轮廓，能够直观地反映金融事件类型。

3.3.3 与基准模型的对比

本文也与几个基准模型进行了对比。在传统机器学习模型中选择效果最好的线性 SVC 模型进行对比；在深度学习模型中则在 CNN 和 RNN 系列模型中选择文本分类最常用的 TextCNN 模型和 BiLSTM 模型进行对比；预训练语言模型中选择 BERT 分类模型进行对比。实验结果见表 6。

由表 6 可以看出，本文模型的事件分类效果最好。其中，线性 SVC 模型的事件分类准确率在 80.19% 左右，而深度学习分类模型的效果则明显高于传统的机器学习方法。TextCNN 模型和 Bi-LSTM 模型的准确率均在 86% 左右，

doi:10.3772/j.issn.2095-915x.2024.03.008

表 5　事件类别和事件关键词

事件类别	事件关键词
人事变动	银行 副行长 担任 董事会 行长 信息 辞职 充足率 截至 显示 首席 加入 高管 监事长 衷心感谢 股东 被执行人 辞呈 履历 职务 下降 原因 资本 记者 辞去 个人 强于 有限公司 资产 列入 发布公告 名单 董事 增长 做出 贡献 辞任 营收 人游 率为 拨备 换血 表示 资料 质量 负责 转任
经营风险	减值 净利润 爆雷 披露 公司 预告 业绩 投资者 基数 工业 集装箱 界面 增长 研究部 拐点 出现 景气 年报 风险 数据 历史 达到 年度 钢铁 应当 A股 等来 高箱 箱价 转跌 产组 新闻 满足 货箱 收益 商誉 结束 注意 来源 资产 下降 年度报告 股份 会计年度 要求 豁免 计提 超过 隐形 国际
证券市场风险	股价 下跌 产品 网络 企业级 网安 产业 领导者 企业 硬件 榜首 评为 新低 主营业务 触及 位居 收盘价 竞争力 客户 收盘 累计 简称 科技 截至 有限公司 发布 指数 提供 行业 历史 政府 继续 公司
股权增减持	减持 公司 普通股 股本 股份 显示 持有 股东 比例 内容 业务 交易 业务 范围 创意 视频 套现 均价 发布公告 产业链 互联网 儿童 定价 数字 净利润 资料 应用 从事 开发 公告 集中 报告 文化 方式 技术 产品 增长
盈利能力	增速 行业 装饰 净利润 创兴 扣非 归母 均值 排名 业绩 装修 东易 预减 名雕 预告 数据 精装 建筑 中值 资源 证券 下限 平均水平 上限 代码 时空 名称 经常性 损益 预增 低于 类型 公告 减少 来源 股份 年度 披露 公司 科技 名列 缩减 主营业务 指标 预喜 增亏 截至 收益 甲方 免责

表 6　基准模型对比实验结果

	accuracy	precision	recall	F1
线性SVC	80.19%	80.15%	79.58%	79.86%
TextCNN	86.05%	85.90%	86.05%	85.97%
BiLSTM	85.95%	85.93%	85.82%	85.90%
BERT	86.42%	86.38%	86.12%	86.32%
本文模型	88.40%	88.33%	88.40%	88.28%

比线性 SVC 模型高出 5 个百分点以上。BERT 模型准确率在 86.42% 左右，而本文基于关键特征增强的模型效果在 88.40% 左右，比 TextCNN、Bi-LSTM 模型和 BERT 模型均高出 2 个百分点左右，表明本文提出的模型在长文本事件分类任务上有更好的效果。

4　总结展望

现有文本分类研究关注领域之间的文本分类较多，而关注领域内部事件分类的相对较少，本文以金融领域为例，研究该领域下的长新闻事件分类问题。为了解决长文本无法完全输入模型的问题，并根据事件的特性捕获更多事件相关特征，本文提出基于关键特征增强的长文本事件分类模型，通过 TextRank 算法抽取长文本的事件关键句，再用 TF-IDF 算法抽取出表征事件信息的动词和名词作为事件关键词，经过 BERT 文本表示并对原文进行特征补充后，输入 SelfAttention 模型实现金融新闻长文本的事件分类任务。

虽然本文的关键特征增强模型有一定的创新性，也在金融领域长文本事件分类任务上比基准模型表现更好，但是本研究依然存在一定的不足和改进空间。本文的关键句抽取方法采用 TextRank 算法，其中句子节点之间的转移概率为句子间的相似度，而句子间的相似度计算方法可以进行改进以抽取出更准确的事件关键句；事件关键词抽取算法采用传统的 TF-IDF 算法，在抽取算法上也存在较大的改进空间。未来的工作将考虑在这两个方面进行更深入的

研究。

参 考 文 献

[1] 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.

[2] DING M, ZHOU C, YANG H, et al. Cogltx: Applying bert to long texts[J]. Advances in Neural Information Processing Systems, 2020, 33: 12792-12804.

[3] 丁月 , 汪学明 . 基于改进特征加权的朴素贝叶斯分类算法 [J]. 计算机应用研究 , 2019, 36(12): 3597-3600, 3627.

[4] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[5] 丁正生 , 马春洁 . 改进词向量和 kNN 的中文文本分类算法 [J]. 现代电子技术 , 2022, 45(1): 100-103.

[6] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A Convolutional Neural Network for Modelling Sentences[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 655-665.

[7] 贺波 , 马静 , 李驰 . 基于融合特征的商品文本分类方法研究 [J]. 情报理论与实践 , 2020, 43(11): 162-168.

[8] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. 2018.

[9] 谌志群 , 鞠婷 . 基于 BERT 和双向 LSTM 的微博评论倾向性分析研究 [J]. 情报理论与实践 , 2020, 43(8): 173.

[10] 李颖 . 基于 BERT-DPCNN 的垃圾弹幕识别改进及应用 [D]. 上海 : 上海师范大学 , 2020.

[11] 孙红 , 陈强越 . 融合 BERT 词嵌入和注意力机制的中文文本分类 [J]. 小型微型计算机系统 , 2022, 43(1): 22-26.

[12] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1480-1489.

[13] SHEN T, ZHOU T, LONG G, et al. Disan: Directional self-attention network for rnn/cnn-free language understanding[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1): 5446-5455.

[14] BASIRI M E, NEMATI S, ABDAR M, et al. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis[J]. Future Generation Computer Systems, 2021, 115: 279-294.

[15] 叶瀚 , 孙海春 , 李欣 , 等 . 融合注意力机制与句向量压缩的长文本分类模型 [J]. 数据分析与知识发现 , 2022, 6(6): 84-94.

[16] 卢玲 , 杨武 , 王远伦 , 等 . 结合注意力机制的长文本分类方法 [J]. 计算机应用 , 2018, 38(5): 1272-1277.

[17] 鲍闯 , 乔杰 , 李海斌 . 基于融合特征的长文本分类方法 [J]. 重庆理工大学学报 (自然科学), 2022, 36(9): 128-136.

[18] JACOBS G, LEFEVER E, HOSTE V. Economic event detection in company-specific news text[C]//1st Workshop on Economics and Natural Language Processing (ECONLP) at Meeting of the Association-for-Computational-Linguistics (ACL). Association for Computational Linguistics (ACL), 2018: 1-10.

[19] JACOBS G, HOSTE V. SENTiVENT: enabling supervised information extraction of company-specific events in economic and financial news[J]. Language Resources and Evaluation, 2022, 56(1): 225-257.

[20] BHOKARE P, SONAWANE A, SONAWANE S, et al. Detection and Classification of Financial Events from News Articles[J]. Iconic Research and Engineering Journals, 2023, 6(8): 171-180.

[21] MIHALCEA R, TARAU P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.

doi:10.3772/j.issn.2095-915x.2024.03.009

基于教材文本语料库的自适应主题词表构建
——以经济类专业为例

开放科学（资源服务）标识码（OSID）

杭建琴[1] 张鸣宇[1,2] 胡泽文[3]

1. 华中师范大学语言与语言教育研究中心 武汉 430079；
2. 武汉大学国际教育学院 武汉 430079；
3. 南京信息工程大学管理工程学院 南京 210044

摘要：[目的 / 意义] 构建一套面向汉语非母语学习者的专业词表对专业学习和国际中文教育学科建设及发展具有重要意义。[方法 / 过程] 针对当前外向型专业词表较少及构建方法单一问题，本文首先从网站爬取小说、新闻和论坛留言构建参照语料库，根据教育部专业课程设置目录，选取专业教材构建专业教材语料库，运用 *TF-IDF-TF* 算法遴选专业主题词并构建词共现矩阵，利用凝聚聚类法实现专业主题词聚类。在此基础上，计算词簇内主题词的语义相关性，选取语义共现度最大的词作为词簇中心词，并根据语义相关性编排词表。最后，以经济学类专业为例构建面向留学生的专业主题词表。[结果 / 结论] 结果表明，本文构建的经济类专业主题词表能够较好地提取专业词汇且有效地将语义关联度紧密的专业主题词聚类在同一词簇内，学习者能够快速有效获取相关词簇进行专业自适应学习，并为其他专业主题词表的构建提供了依据。

关键词：主题词表；凝聚聚类算法；语义共现度；词簇中心词

中图分类号：G35

A Textbook Corpus Approach to Constructing a Self-adaptive Subject Word List——Taking the Economics-relevant Majors as an Example

HANG Jianqin[1] ZHANG Mingyu[1,2] HU Zewen[3]

1. Research Centre for Language and Language Education, Central China Normal University, Wuhan 430079, China;
2. School of International Education, Wuhan University, Wuhan 430079, China;
3. School of Management Science and Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China

基金项目 国家社会科学基金项目“面向海量科技文献的潜在‘精品’识别方法与应用研究”（20CTQ031）；国家社会科学基金一般项目“鄂西北四省市过渡地带方言语法调查与比较研究”（20BYY039）；江苏省高校哲学社会科学一般项目“‘一带一路’国家来华留学生学习焦虑情绪对汉语学习的影响及对策研究”（2023SJYB0753）。

作者简介 杭建琴（1988-），博士研究生，主要研究方向为专门用途汉语教学与语料库语言学；张鸣宇（1980-），博士研究生，讲师，主要研究方向为对外汉语教学；胡泽文（1985-），博士，副教授，主要研究方向为科学计量与科技评价、数据科学与知识工程，E-mail：huzewen915@163.com。

引用格式 杭建琴 , 张鸣宇 , 胡泽文 . 基于教材文本语料库的自适应主题词表构建——以经济类专业为例 [J]. 情报工程 , 2024, 10(3): 114-127.

Abstract: [Purpose/Significance] Building a specialized word list for non-native Chinese learners is of great significance for specialized learning and the construction and development of International Chinese Language Education discipline. [Methods/Processes] In response to the current shortage of Chinese specialized word list for foreign learners and the single construction method, this paper first crawls novels, news, and forum comments from websites to construct a reference corpus. Based on the specialized curriculum directory of the Ministry of Education, textbooks are selected to construct a corpus of specialized textbooks. Algorithms are used to select specialized subject words and construct a word co-occurrence matrix. Cohesive clustering is used to achieve subject words clustering. On this basis, calculate the semantic correlation of the subject words within the word cluster, select the word with the highest semantic co-occurrence as the central word of the word cluster, and arrange the word list based on the semantic correlation. Finally, taking economics major as an example, a specialized subject word list for foreign students is constructed. [Results/Conclusions] The results showed that the economic subject word list constructed in this paper can greatly extract the specialized vocabulary, and effectively cluster closely related specialized subject words within the same word cluster. Learners can quickly and effectively obtain relevant word clusters for adaptive learning. What's more, this method also provides a basis for the construction of other subject word list as well.

Keywords: Subject Word List; Cohesion Clustering Algorithm; Semantic Co-occurrence; Central Word of the Word Cluster

引言

当前，中国已成为亚洲最大的留学目的国，学历生人数占比已远超非学历生，进入专业学习的人数不断增加。伴随汉语国际化的传播发展，专业领域汉语人才的需求不断扩大，专业汉语正成为国际中文教育发展的新方向[1]。根据学者前期调研发现，国际中文教育，学习者通过预科考试进入专业学习后，专业词汇是其遇到的首要问题[2]。因此，专业词表的研制对后续专业学习具有重要意义。目前，国内已研制了一系列外向型通用词表，形式多样，种类繁多，但专业词表种类十分有限并存在词汇数量大、词汇编排不合理、使用不经济等问题，面向汉语非母语学习者的专业词表则更少。

词汇是构成篇章的基本单位，在留学生专业学习中占有举足轻重的地位，只有掌握专业词汇才有助于其理解专业内容，并为后续深入学习和科学研究做好充足的词汇储备。但针对留学生遇到的词汇问题，目前还没有一套系统的专业词表帮助克服这一困难。因此，建立一个面向汉语非母语学习者的外向型专业词表，能够有效帮助学习者消除词汇障碍，提升学习者的专业术语理解能力和学术表达的规范性。在词汇习得领域，反映学生学习认知规律的心理词典（mental lexicon）已经受到学者的普遍关注，认为学习者所习得的词汇知识有组织地存储在大脑中[3]。而且，词与词之间的意义是以网状形式相互联系，最终形成一个语义网络层级体系，而现有的词典及词表都是从便于检索的角度出发编排，而忽视学习者的词汇认知和习得规律。因而，将学习者的心理词典理论应用到词表构建层面，应以某一核心词为中心，找到与其在意义上相联系并处在同一个连续统中的专业词汇，共同构成一个专业主题词词汇语义网，从核心意到边缘意，按照认知规律去编排专业主题词表，实现将专业中能够表达某一主题中心思想或核心内容的主题词汇聚集在一起。留学生可以根据自身的学习需求选取某一类主题词进行针对性学习，从而实现留学生

doi:10.3772/j.issn.2095-915x.2024.03.009

学习记忆词汇的最佳效果。鉴于此，本文首先从网站爬取小说、新闻和论坛留言构建参照语料库，根据教育部专业课程设置目录，选取专业教材构建教材语料库，运用 *TF-IDF-TF* 算法和凝聚聚类方法实现专业主题词聚类，在此基础上，根据语义相关性确定词簇中心词，最后，以经济学类专业为例构建面向留学生的自适应专业主题词表。

1 相关研究

本文旨在研究专业主题词表的专业词汇识别、提取、聚类及主题词的确定。因此，与本文相关的研究包括外向型词表的设计、主题词表的构建方法及主题词表的应用。本节分别对这三个方面的研究进行总结。

1.1 外向型词表设计相关研究

外向型词表是留学生进行专业学习和科学研究的必备工具书。目前，国内学者面向留学生的外向型词表研制展开了一系列的研究。姜德梧 [4]、李红印 [5]、赵金铭等 [6] 就《汉语水平词汇与汉字等级大纲》通用词表从候选词的抽取、遴选、收录等角度展开了深入的研究，为通用型词表的优化和后续开发提供了参考。专业词表研制上，刘华等 [1] 提出了词语聚类算法，将具有文化内涵的中医词语进行合理分类，构建了一个按照词语的语义关联度进行聚类的中医汉语主题词表。这里基于算法的方法是指利用多种聚类方法，最后产生不同的聚类结果，或者通过不同的参数设置，从而使一种方法产生不同的聚类效果 [7]。该方法在一定程度上改进了外向型专业词表的设计，但是，现有方法并未考虑到文本之间的相似度、主题词簇中心词的界定，进而学习者无法从宏观层面了解学科框架和知识体系脉络间的联系，也无法快速获取相关主题词簇展开针对性学习。词汇是科学知识的载体 [8]，而核心专业词是核心内容的浓缩与提炼，能直接反映专业领域知识点的分布与发展脉络 [9-10]。此外，词表的体系结构设计包括自上而下和自下而上两种方法，前者根据词表的编制目的，先建立顶层宏观架构再分解出细分类别，而后者是对领域概念和术语的归类，梳理底层细分类别，逐层聚类获得更高级更抽象的类目 [11]。

1.2 主题词表构建方法研究

主题词表的编制提升了相关主题信息检索的效率和准确性，有利于科研工作人员和学习者及时调整相关策略，从而做出科学的决策。国内学者在主题词表研制领域开展了丰富的研究工作。例如，邵卫等 [12] 使用无监督方法构建了一种基于依存句法分析的科技政策文本关键词抽取算法，并在此基础上提取了科技政策领域文本的主题词，该方法可以帮助解决领域未登录词切分、主题词之间关系缺乏等问题。杜慧平等 [13] 设计出一套专题主题词表的自动构建方案，并结合民国文献语料构建了民国时期抗战史主题词表。基于该方法设计的主题词表构建方法降低了编表时间和成本，有效推动了主题词表的推广和应用。类似研究还包括仲云云等 [14]、曾文等 [15]，这些研究为面向留学生的外向型专业主题词表的构建提供了借鉴和参考。此外，Kondo 等 [16]、Gupta 等 [17]、Tsai 等 [18]、

罗卓然等[19]、王亚楠等[20]就学术词汇的抽取与识别、归纳与聚类的算法提出了新颖的解决方案，这些研究为专业主题词表模型的构建提供了新的思路。

1.3 主题词表的应用研究

主题词表在图书情报领域的应用及重要性已得到广泛认识。孙立媛等[21]从高校应急决策角度系统论述了主题词表应用于高校应对突发事件及制定科学决策层面起到的积极作用。另有学者从南海文献资料的开发及生物、医学、农业术语系统发展需要层面深度剖析了主题词表在实现文献资料自动标引、文献资料深度挖掘及科研实践研究应用上的重要作用[11,22]。在国际中文教育领域，苏新春[23]认为，“词表研制工作是当代词汇研究的一项重要任务，它在语言学习、语言测试、词典编纂、中文信息处理、语言翻译、心理认知、语言推广等众多方面都有重要作用”。苏新春[24]运用词汇计量研究手段探讨不同目的之下进行的词汇计量和词表制作对教材编写的影响，强调了词表在教材编订、大纲设计中的应用价值，并建议应根据教学和学习需求编制通用型词表与专业型词表，但文中并未对相关词表的构建展开论述。因此，主题词表对留学生的专业学习及科学研究同样具有重要的实践价值，但有关其应用设计还未得到充分挖掘。

综上所述，虽然国内外学者关于主题词表的研制问题已做了深入细致的研究，但有关外向型主题词表的研究主要集中于通用型词表，鲜有学者就面向留学生的外向型专业词表研制展开探讨；此外，从词表设计方法来看，现有的主题词表构建方法的改进主要是面向图书情报领域，提升了决策的效率和准确率，但忽视了通用词汇和专业词汇的平衡、取舍问题，而且，在国际中文教育领域，有关主题词表应用价值的认识还需进一步加强。因此，本文将结合前人研究成果，从学习者的专业学习需求出发，自建语料库，计算语料库文本之间的语义相似度，界定主题词的分类，改进 *TF-IDF* 算法，遴选专业主题候选词，构建词共现矩阵，采用凝聚聚类法实现主题词聚类，并在此基础上进一步计算出主题词簇中心词，构建一个自适应的专业主题词表，并进行量化评估，证实其科学性、有效性，提升国际中文教育领域有关专业主题词表作用的关注度。最后，本文以经济类专业为例，选取 8 本专业通用教材自建语料库，设计一个面向经济类专业留学生的主题词表。

2 专业主题词表构建思路与方法

2.1 专业主题词表构建思路

本文设计的专业主题词表是为初入专业学习的汉语非母语学习者而设计的外向型词表，根据所选取教材按不同主题进行分类、标注和挖掘，提供元语义的识记与理解，以期帮助学习者加强对专业术语的理解，提升后续的专业学习效果。目前，自动获取专业词语的方法主要有两类：基于规则和基于统计的，基于规则的方法利用“人工建构好的领域词语特征词库、经常与领域词语共现的指示词库和指示领域关系的关系词语库，在大规模语料中利用模版匹配的方式获取领域词语”[25]。本研究主要采用基于统计的模型计算法，专业主题词表设计思路如图 1 所示。（1）按照课程设置要求，选择

doi:10.3772/j.issn.2095-915x.2024.03.009

专业通用教材并从网络爬取包含小说、新闻、论坛评论等通用语料之后，进行预处理，自建一个专业教材语料库和参照语料库；（2）利用 *TF-IDF-TF* 算法遴选出专业主题词表候选词；（3）构建专业主题词的词共现矩阵并在此基础上采用凝聚聚类分析法实现主题词聚类；（4）计算每个词簇中词语的语义共现度，选择共现度最大的词作为词簇中心词；（5）按照词簇间中心词的语义关联度编排专业主题词表；（6）对主题词表进行量化评估。

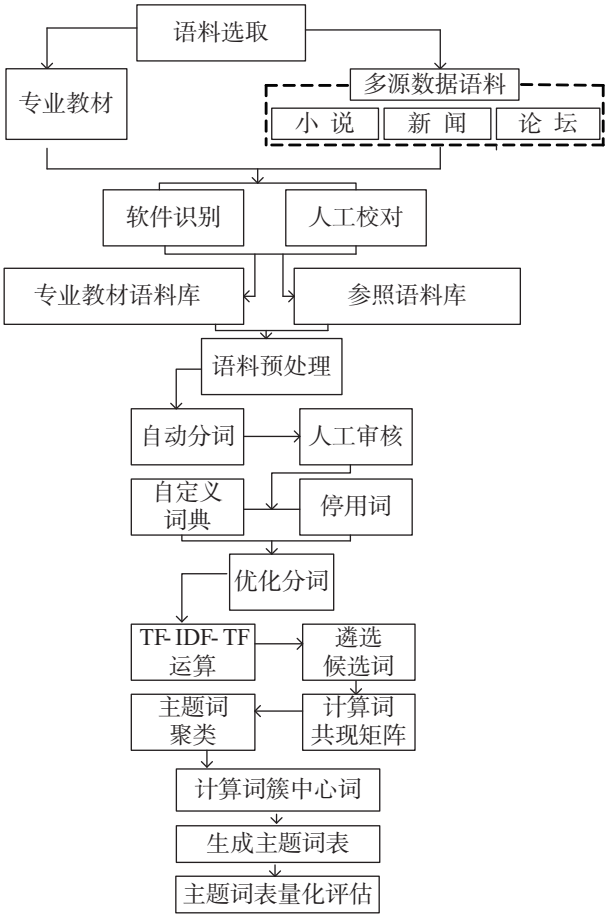

图 1　专业主题词表构建流程图

2.2　专业主题词表构建方法

（1）测算专业教材的语义相似度

为进一步优化主题词表候选词的语义相似度，了解专业学科特点，首先利用文本语义相似性测度方法对教材进行分析。理论上，两个向量的夹角越小，说明这两个向量之间的关联性越大，所以，向量的余弦值可以反映向量的相似性。因此，本文使用余弦相似度方法计算专业类教材的语义相似性。

假定 A 和 B 是两个 n 维向量，则 A 和 B 的相似性 $S(A,B)$ 可以用其向量夹角 θ 余弦来表示，其计算公式表示为：

$$S(A,B)=\cos\theta=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}=\frac{A\cdot B}{|A|\times|B|} \quad (1)$$

计算专业教材的相似度就是将文本的相似度转化为代表两个文本的向量夹角的余弦值。文本相似性的计算过程包括：分词、词频统计、构造词频向量、计算余弦相似度。两个文本之间的余弦相似度值越接近 1，两个文本间的相似度越高，而越接近于 0，表示两个文本的相似性越低。

（2）基于 *TF-IDF-TF* 算法遴选专业主题词表候选词

TF-IDF 算法主要思想是：假设某个词汇在某一类型文本中出现的频率较高，而在其他类文本中出现的频率较低，则可以认为该词具有较好的文本区分度 [26]。因此，在专业主题词表的构建中可以采用这一方法。建立一个专业教材语料库与参照语料库，运用 *TF-IDF* 的运算公式提取专业主题候选词。根据自建的专业教材语料库，确定文本领域集 $D=\{d_j|j=1,2,3,\cdots,n\}$，教材文本语料分词 $(w_1, w_2, w_3, \cdots, w_i)$。由于本文构建的教材语料包含两个层次，即单个教材层次和专业总体教材层次，仅仅使用 *TF-IDF* 方法仅能获取单个教材的主题词，为此，本文考虑

使用 *TF-IDF-TF* 方法遴选经济类专业主题词，包含两个步骤：第一，运用 *TF-IDF*[27] 方法分别获取每一本教材的主题候选词；第二，合成 *n* 本教材，运用 *TF* 方法对第一步获取的候选词进一步遴选，确定最终的专业主题词。

TF（Term Frequency）是指某词汇在文本内的频率，即某一经济类词汇在某本教材中出现的频率，计算公式为：

$$TF_{ij}=\frac{f_{ij}}{\sum_{i=1}^{T}f_{ij}} \quad (2)$$

其中，f_{ij} 表示词汇 w_i 在文本 d_j 中出现的频率，$\sum_{i=1}^{T}f_{ij}$ 指在 d_j 文本中所有词汇的统计频数。*DF*（Document Frequency）用来计算词汇在参照语料库中的重要性。其计算公式为：

$$DF_i=\frac{N}{n_{ij}} \quad (3)$$

其中，n_{ij} 表示在参照语料库中包含文本 d_j 中的词汇 w_i 的段落数，N 表示参照语料库的数量，此处用参照语料库的段落总数表示。*IDF*（Inverse Document Frequency）指词汇的逆文本频率，其计算公式为：

$$IDF_{ij}=-log\frac{N}{1+n_{ij}} \quad (4)$$

计算文本 d_j 中主题词 w_i 的 *TF-IDF* 值，即专业语料库中词汇区分专业语料库文本和参照语料库中文本的能力，计算公式为：

$$TF-IDF_{ij}=TF_{ij}\times IDF_{ij}=\frac{f_{ij}}{\sum_{i=1}^{T}f_{ij}}\times\left(-log\frac{N}{1+n_{ij}}\right) \quad (5)$$

分别计算出专业教材词汇的 *TF-IDF* 数值之后，需要进一步确定专业主题候选词。鉴于本文构建词表的双重性，①本文考虑整合 *n* 本专业教材，使用 *TF* 方法对确定的候选词进行筛选，最终遴选出专业主题词。

（3）构建专业主题词共现矩阵

通过 *TF-IDF-TF* 的迭代计算，最终确定了专业主题词表候选词。为了能够将专业主题词中语义相关或语义度联系比较紧密的主题词聚类在一起，首先要构建一个经济类专业主题词词共现矩阵。所谓词共现（co-occurrence）是指两个词项在一定的文本窗口范围内共同出现[28]。如果某一个词 *x* 与另一个词 *y* 共现的频率明显高于 *x* 与其他词汇的共现频率，则可以认为 *x* 与 *y* 之间存在某种语义相关性。计算词共现的计算公式为：

$$co_occurence\left(w_i,w_j\right)=\frac{1}{2}\left(\frac{f\left(w_i,w_j\right)}{f\left(w_j\right)}+\frac{f\left(w_j,w_i\right)}{f\left(w_i\right)}\right) \quad (6)$$

其中，$f(w_j)$ 表示专业主题候选词 w_j 在语料库文本中出现的频次，$f(w_i)$ 表示专业主题候选词 w_i 在专业教材语料库文本中出现的频次，$f(w_i, w_j)$ 为专业主题候选词 w_i 和 w_j 在专业教材语料库自然语言段落中共现的频次，即同时包含 w_i 和 w_j 的段落数。$co_occurence\left(w_i,w_j\right)$ 作为词汇 w_i 和词汇 w_j 共现度值。

（4）专业教材主题词聚类及词簇中心词选取

为了生成专业主题词表，需要对测度的候选词共现矩阵进行进一步分析。本文使用凝聚聚类分析方法进行词簇聚类分析。凝聚聚类是一种无监督学习算法，是一种基于相同原则构建的聚类算法。本文在候选词共现矩阵的基础上，通过测算不相似系数矩阵，通过如下公式测度：$dissimilarity_{ij}=1-\overline{co_occurrence}_{ij}$。其中，$dissimilarity_{ij}$ 表示不相似系数矩阵，

①本文构建的主题词表不仅含有单本教材的课程特征，而且考虑了经济类教材的专业特征，因此具有双重性特征。

doi:10.3772/j.issn.2095-915x.2024.03.009

$\overline{co_occurrence}_{ij}$表示主对角元素为 1 的词共现矩阵。令 I、J 为两个组成新类的两个簇，K 是另一个类，定义 n_I、n_J 和 n_K 为类 I、J、K 大小。本文使用“average”间距离进行聚类，其聚类更新方程为：$\frac{n_I d(I,K)+n_J d(J,K)}{n_I+n_J}$。类 A 和 B 的不相似聚类的计算公式表述为：$d(A,B)=\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B}d[a,b]$。

凝聚聚类算法的具体步骤包括：（1）计算 n 个样本之间的距离作距离矩阵 D；（2）根据不同的扰动样本构建一个间距序列，并对序列进行排序；（3）对排序后的序列绘制一个样本图，而最接近阈值的样本配对组将合并为一个新的类别，如果所有样本都是该图的元素，则停止，否则，重复此步骤；（4）输出算法是一个嵌套的层次结构图，其以欧式距离相似度的阈值作为分类边界，并产生一个分类子集。

在此基础上进行专业主题词簇中心词选取，计算公式步骤如下 [29]：

首先计算某一个专业主题词 w_i 与该词簇 W_i 的语义关联度 $C(w_i)$：

$$C(w_i)=\sum_{w_j\in W_i} co_occurence(w_i,w_j) \quad (7)$$

将上述公式进行归一化处理，得到 w_i 对专业主题词簇 W_i 语义共现度值：

$$C(w_i\,|\,W_i)=\frac{C(w_i)}{\sum_{w_i\in W_i}C(w_i)} \quad (8)$$

选取词簇 W_i 中语义共现度值最大的 w_i^* 作为词簇中心词：

$$w^*_i=\max_{w_i\in W_i}C(w_i\,|\,W_i) \quad (9)$$

（5）主题词表构建效果评估

为了量化评估本文所构建的主题词表效果，本文进一步采用主题文本分析中常用的主题语义连贯性和离散度方法进行检验。为比较本文构建的模型在主题词表构建中是否具有优势，本文进一步选取其他与本文方法较为接近的主题词表构建方法进行词表构建，并做进一步的对比分析。本文参考居亚亚等 [30] 的研究，选取三类主题语义连贯性量化指标，即 C_{UMass}、C_{NPMI} 和 C_{UCI}。三个检验指标的统计量分别表示为：

$$C_{UMass}=\frac{2}{N\cdot(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\log\frac{p(w_i,w_j)+\varepsilon}{p(w_j)} \quad (10)$$

$$C_{NPMI}=\frac{2}{N\cdot(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{PMI(w_i,w_j)}{-\log p(w_i,w_j)+\varepsilon} \quad (11)$$

$$C_{UCI}=\frac{2}{N\cdot(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}PMI(w_i,w_j) \quad (12)$$

其中，$PMI(w_i,w_j)=\log\frac{p(w_i,w_j)+\varepsilon}{p(w_i)\cdot p(w_j)}$。$PMI$ 为点互信息指数（Pointwise Mutual Information）。通过计算不同主题词簇间的互信息 PMI，再根据主题词簇的数量求平均值，得到模型的量化评价指标。三个指标的值越大，则表示该主题的语义连贯性越强。

此外，本文还选取主题离散度测度指标反映主题词之间的离散程度。由于标准方差过于敏感，本文选择常用的离散指标测度方法，即变异系数。变异系数通过平均值除以偏差得到，能够消除量纲影响。变异系数 CV 越大，意味着主题词之间的离散程度越高。在通过聚类得到主题词簇的基础上，分别基于词共现矩阵计算其对应的变异系数，最后取平均值得到模型的变异系数，以反映主题词表整体的离散程度。

3 词表构建的语料来源与语料预处理

3.1 语料来源

为了剔除非专业的通用主题词，本文首先

建立一个参照语料库。参照语料库的内容主要有从网络爬取的《人民日报》新闻，包括独家特稿、热点新闻以及涵盖港澳台、国际等区域和教育、经济科技、军事、社会法制、文旅体育等领域的主题要闻；小说有曹禺的《雷雨》、老舍的《骆驼祥子》和《老张的哲学》、刘慈欣的《三体》等通用文本以及论坛留言。

此外，为构建专业主题词表，还需要构建专业教材语料库。结合留学生人数的专业分布情况，本文考虑以经济类专业为例，构建经济类专业主题词表。经济类专业主要涵盖国际经济与贸易、管理学、金融学以及会计等专业。按照教育部的课程设置规定，确定语料的选择范围，上述经济类专业的基础课程包括“微观经济学”“宏观经济学”“政治经济学”“计量经济学”“统计学”“会计学”“财政学”“金融学”，共 8 门课程。这些课程的教材品类较多，我们对开设相关课程的部分财经类院校和“双一流”高校进行了调研，从教材的权威性、认可度及广泛性三个方面进行全面综合考察，最终选定如下教材:《马克思主义政治经济学概论》编写组编写的《马克思主义政治经济学概论》（第二版）（以下简称《政治经济学》）、陈共编写的《财政学》（第十版）、李子奈、潘文卿编著的《计量经济学》（第五版）、贾俊平、何晓群、金勇进编著的《统计学》（第七版）、《西方经济学》编写组编写的《西方经济学》（上）（即《微观经济学》）和《西方经济学》（下）（即《宏观经济学》）、周正云编著的《基础会计》（第三版）及黄达、张杰编著的《金融学》（第五版），共计 8 本教材构成经济类专业教材语料库的语料。

3.2 语料预处理

两个语料库内容确定之后将对其进行预处理。删除图片、数字表格、脚注、参考文献、前言后记，去除语料中除中文字符以外的所有标点符号、图片、公式、数字等信息。另外删除教材转化过程中存在的一系列广告等无关字段，并进行人工校对，纠正转码中出现的文字错误和遗漏信息。最终，得到研究所需要的经济类专业教材语料库和参照语料库。因汉字与英文等西文字符不同，是连续书写的字符串，词与词之间没有分隔，因此我们在语料预处理完后需要进行分词处理，分词方法是利用 Python3.7 的编程程序，在 PyCharm 集成开发环境下调入 Jieba（结巴）分词包进行自动分词。为了能够使分词效果达到满意的结果，自动分词之后进行人工校对，发现两类问题，一是部分经济类专业词汇切分颗粒度过小，如“纳什均衡”被切分成两个词“纳什”和“均衡”。此外，部分分词存在歧义，如“大部分业务”切分成了“大部”和“分业务”。针对上述问题，加载一个自定义词典，将切分中存在歧义和切分颗粒度较小的词放在自定义词典中进行二次分词从而解决上述问题。另外，有些通用词汇的文本区分度较小，如虚词和代词，对主题词表构建研究意义较小，我们通过停用词词表过滤相关词汇。预处理后的语料库基本信息如表 1 所示。

表 1 经济类专业教材语料库与参照语料库基本信息表

语料库	行 数	类 符	形 符
经济类专业教材语料库	25 712	58 215	720 320
参照语料库	39 249	87 208	937 000
合 计	64 961	145 423	1 657 320

doi:10.3772/j.issn.2095-915x.2024.03.009

4 经济类专业主题词表构建结果

4.1 经济类专业教材的语义相似度结果

本文构建的经济类专业主题词表涵盖 8 门基础课程。为此，我们对所选教材的相关性进行统计分析，以便在后期的学习中学习者可以根据自身实际学习需求快速定位相关主题词簇，锁定核心要义，辅助专业学习。文章通过自然语言处理结果计算教材之间的余弦相似性，初步了解所选教材的文本特征，语义相似度的测算结果如表 2 所示。从表中可看出，《宏观经济学》和《金融学》两本教材余弦相似性平均值最高，因为这两本教材中有一些章节的内容是重叠的，说明这两门课程教材在经济类专业基础课程教材中占据主导地位。《计量经济学》与《基础会计》两本教材余弦相似性平均值较低，分别为 0.378、0.321，说明这两本教材的专业性较强，语义相似性较小，与其他课程教材的关联度也较低。根据课程教材之间的语义相似度和学习者专业学习需求，结合专家意见，进行内容整合，将其分为 5 个主题词目录，分别为：经济学、金融学、统计学、会计学、计量学。

表 2 教材的语义相似性分析结果

教材名称	《政治经济学》	《微观经济学》	《宏观经济学》	《财政学》	《金融学》	《基础会计》	《统计学》	《计量经济学》
《政治经济学》	1							
《微观经济学》	0.512	1						
《宏观经济学》	0.571	0.670	1					
《财政学》	0.635	0.520	0.628	1				
《金融学》	0.638	0.582	0.783	0.657	1			
《基础会计》	0.296	0.337	0.341	0.351	0.383	1		
《统计学》	0.421	0.512	0.525	0.457	0.538	0.333	1	
《计量经济学》	0.292	0.371	0.443	0.324	0.382	0.208	0.630	1
平均值	0.481	0.501	0.566	0.510	0.566	0.321	0.488	0.378

4.2 确定经济类专业主题词表候选词

根据上述的 *TF-IDF* 计算公式步骤，将其编入 Python 程序，分别导入 8 本经济类专业教材和参照语料库，计算每本教材分词的 *TF-IDF* 值，如下表 3 是《政治经济学》的部分 *TF-IDF* 值。通过观察下列数值发现，值越小，该词的文本区分度越弱，对经济类专业主题词表构建的意义也越小。在此基础上，结合专家审核意见，选取候选词共计 1876 个，《政治经济学》142 个，《微观经济学》151 个，《宏观经济学》200 个，《基础会计》230 个，《统计学》161 个，《计量经济学》141 个，《财政学》300 个，《金融学》551 个，作为构建主题词表的候选词。

在此基础上，为进一步优化经济类专业词汇候选词，将 8 本教材所遴选的候选词合并至一个文档中，再次计算候选词的 *TF* 值，剔除出现频次较低的候选词，最终选取 1260 个主题词构建词表，部分结果如表 4 所示。

表 3 《政治经济学》*TF-IDF* 数值表（部分）

序号	词汇（w_i）	$TF\text{-}IDF_{ij}$ 值	序号	词汇（w_i）	$TF\text{-}IDF_{ij}$ 值
1	资本家	0.0476272096143862	4153	维护者	0.0000770666822239258
2	剩余价值	0.0467024094276991	4154	万邦	0.0000770666822239258
3	利润率	0.0122536024736042	4155	和睦	0.0000770666822239258
4	生产资料	0.0110590688991334	4156	建立	0.0000768514121618478
5	经济危机	0.00959480193687877	4157	统筹	0.000076797218300066
6	所有者	0.00901680182019932	4158	影响	0.0000765596645777158
7	可变资本	0.00867000175019166	4159	健康	0.0000762103857547711
8	商品经济	0.00855440172685577	4160	其次	0.000075862515314177
9	非公有制	0.00855440172685577	4161	特征	0.000075391319566884
10	平均利润	0.00809200163351221	4162	大大	0.000075391319566884
……	……	……	4163	多元化	0.0000751400151683277

表 4 *TF-IDF-TF* 数值表（部分）

序号	词汇（w_i）	TF_{ij}	序号	词汇（w_i）	TF_{ij}
1	货币	0.00776738116392714	8	边际	0.00200605286539316
2	资本	0.00409262549977788	9	支出	0.00195885162150155
3	投资	0.00336933585073301	10	资产	0.00179642381163927
4	利率	0.00321801421590404	11	财政	0.00153265215459796
5	金融	0.00295007774322523	12	税收	0.00111339404709018
6	成本	0.00279320302087961	13	国债	0.00106896934695691
7	变量	0.00275988449577965	……	……	……

4.3 经济类专业主题词共现矩阵网络图

根据词共现矩阵的计算方法，对所有教材按照段落进行主题词的共现矩阵测度。8 本教材的段落合计 25712 行。根据公式 $co_occurence(w_i,w_j)=\frac{1}{2}\left(\frac{f(w_i,w_j)}{f(w_j)}+\frac{f(w_j,w_i)}{f(w_i)}\right)$，计算得到共现矩阵，利用 Gephi9.2 软件，通过两两候选词之间的关系刻画拓扑网络图，结果如图 2 所示[②]。可以看出，筛选后的关键词具有较高的组间区分度，例如，图中左下角为计量经济学专业名词，包括“均值”“分布”“参数估计”“显著性”等候选词，其与其他教材的区分度较大，说明本文测度的词共现矩阵较为合理。

4.4 经济类专业主题词聚类和词簇中心词选取

根据本文测度词共现矩阵，使用凝聚聚类算法对选取的 1260 个经济类专业主题词进行聚类分析，共计得到 98 类词簇。在此基础上，根据公式 $C(w_i \mid W_i)=\frac{C(w_i)}{\sum_{w_i \in W_i} C(w_i)}$ 计算出词簇中每一

②由于本文的主题词有 1260 个，因此，无法清晰地反映在一张图表上，本文选取其中的 200 个节点进行绘制，在绘图时根据边的大小进行筛选，去除了部分边权重较小的部分，以使图表展示得更加清晰。

doi:10.3772/j.issn.2095-915x.2024.03.009

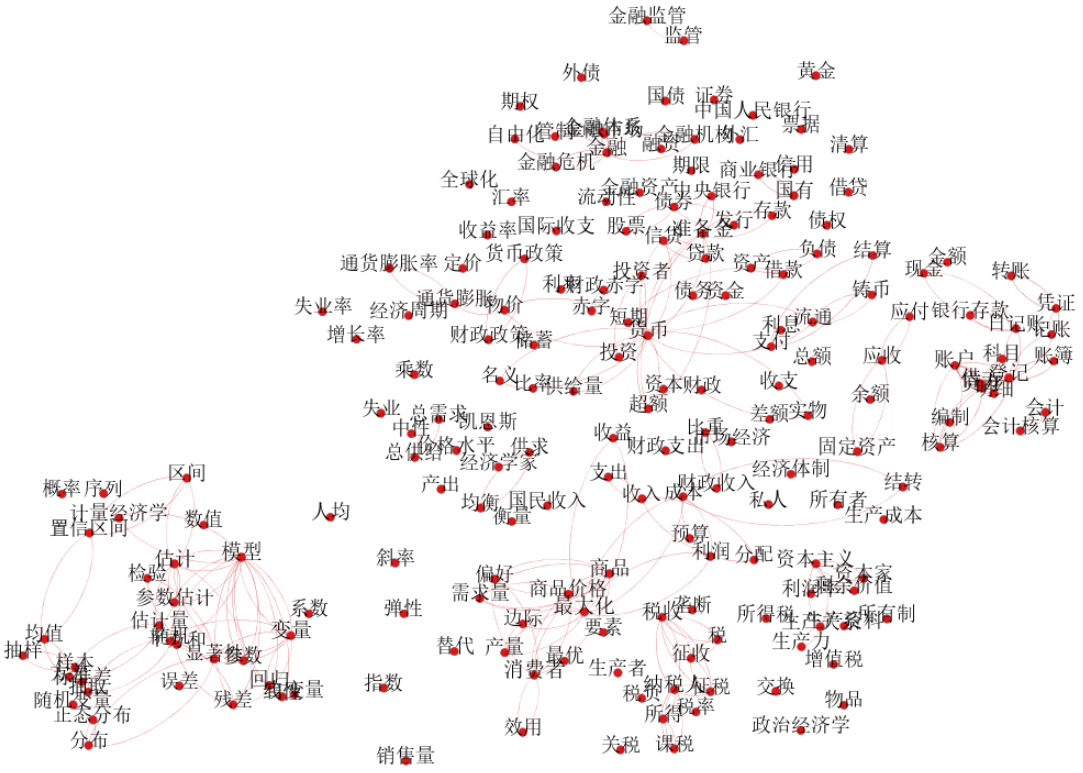

图 2　主题词共现图（部分）

个词的语义共现度，从中选取语义共现度最大的词作为主题词簇中心词，词簇联接词及词簇中心词如表 5 所示。

随机抽取词簇中心词“生产力”和“中央银行”进行可视化展示，结果如图 3 所示。词簇内词与词之间通过线性方式连接，线条粗细反映了词语之间的关联性大小，关联性越大，线条越粗，反之亦然；线条的疏密反映了词语之间的语义关联性的强弱，线条越密集，词簇内部词语间关联性越强，反之亦然。

表 5　经济类专业主题词簇表（部分）

序号	词簇中心词	词簇联接词
1	账款	借款，应付，应收，货款，款项，折旧，预付，薪酬，税金……
2	盘盈	清查，盘点，账面，损溢，盘亏，账实，账面，核对……
3	生产力	私有制，公有制，基本矛盾，商品经济，自然经济……
4	中央银行	再贴现率，抵押，公开市场，担保，银行法，抵押品……
5	增值税	进行税额，销项税额，运输费，收讫，抵免，价款，抵扣……

4.5　经济类专业主题词表可视化结果

主题词完成聚类之后按照词簇中心词之间的语义关联度构建可视化图谱，结果如图 4 所示。其中，词簇中心词网络节点大小与词簇中专业主题词数量成正比，节点越大，词汇数量越多。连接线的粗细反映了词簇中心词间语义度的相关性，连线越粗，语义联系越紧密。胡文飞 [31] 认为词库各网络节点连接强度的差异性

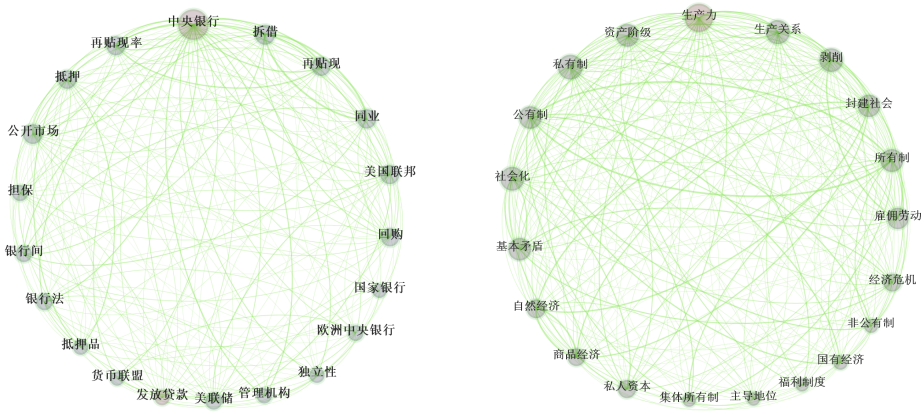

图 3　经济类专业主题词可视化结果（部分）

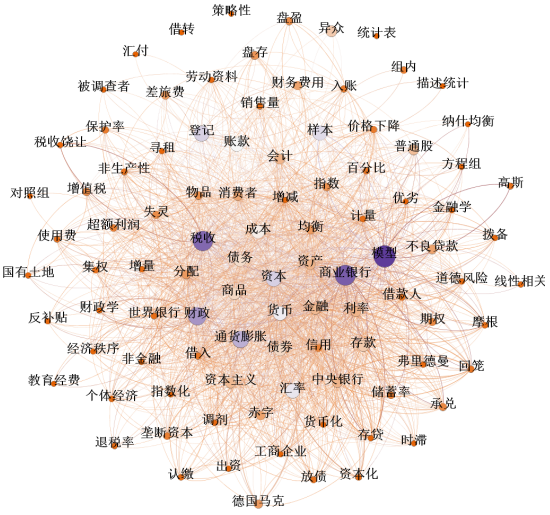

图 4　主题词簇间的网络关系

会影响信息提取、加工的速度和效率，这客观上要求编者在词典表征过程中不断增强表征组构成分的关联性。据此，本文结合主题词目录，根据词簇中心词之间的语义关联度编排经济类专业主题词表。学习者可按照自身的学习需求，结合教材之间的相似度特征，围绕主题词目录项下的中心词快速获取与本中心词语义相关的一系列经济类专业词汇，加大学习者心理词库的网络节点连接强度，有助于其不断深化专业的学习和理解。因此，通过上述方法生成的主题词表能够将语义关联度较大的专业主题词聚类在一起，并且词簇之间有明显的语义区分度。

4.6　主题词表构建效果评估结果

在得到主题词表构建结果的基础上，本文进一步运用主题语义连贯性和离散度两类方法进行效果检验。表 5 描述了主题词表构建模型的评价结果。从主体连贯性来看，使用 *TF-IDF* 和词共现矩阵和聚类分析的方法得到主题词表具有较高的统计指标，说明模型具有较好的主题连贯性，而基于 *TF-IDF-TF* 的模型具有显著的主题离散优势。通过对比发现，不考虑词共现矩阵，仅仅使用 *TF-IDF* 进行主题词表构建，其主题的连贯性效果显著降低。总体来看，本文构建的模型（“*TF-IDF-TF*+共现矩阵+聚类”）

doi:10.3772/j.issn.2095-915x.2024.03.009

和模型（“*TF-IDF*+ 共现矩阵 + 聚类”）均具有较好的词表效果。

表 5　模型评估结果

模型	主题连贯性			主题离散度
	C_{UMass}	C_{NPMI}	C_{UCI}	*CV*
“*TF-IDF-TF*+ 共现矩阵 + 聚类”模型	-2.5770	0.1709	1.4095	0.0081
“*TF-IDF*+ 共现矩阵 + 聚类”模型	-1.8891	0.2046	1.6676	0.0174
“*TF-IDF*+ 聚类”模型	-3.3884	0.0201	0.2050	0.0071

为了进一步分析本文构建词表的优势和构建方法的稳健性，本文进一步选取其他类别的文本数据进行分析。具体来说，本文选取新闻类数据，使用“新闻联播”的文本数据构建新闻类主题词表。此外，本文还从构建的语料库中选取小说文本数据，进一步构建小说类主题词表。本文构建的新闻类主题词表和小说类主题词表的量化评估结果如表 6 和表 7 所示。从表中可以看出，本文构建的“*TF-IDF-TF*+ 共现矩阵 + 聚类”模型和“*TF-IDF*+ 共现矩阵 + 聚类”模型均具有较好的词表生成效果。新闻类主题词表的主题语义连贯性较低，且其主题离散度较高，而小说类主题词表的语义连贯性较高，主题离散度的波动较大，特别是“*TF-IDF*+ 聚类”模型的主题离散度显著增高。综合来看，本文构建的主题词表方法具有较好的适应性，可以广泛应用于其他特定领域。

表 6　新闻类主题词模型评估结果

模型	主题连贯性			主题离散度
	C_{UMass}	C_{NPMI}	C_{UCI}	*CV*
“*TF-IDF-TF*+ 共现矩阵 + 聚类”模型	-0.7463	-0.1036	-0.4922	0.0185
“*TF-IDF*+ 共现矩阵 + 聚类”模型	-0.7078	-0.0823	-0.3194	0.0208
“*TF-IDF*+ 聚类”模型	-0.9223	-0.1570	-0.6286	0.0005

表 7　小说类主题词模型评估结果

模型	主题连贯性			主题离散度
	C_{UMass}	C_{NPMI}	C_{UCI}	*CV*
“*TF-IDF-TF*+ 共现矩阵 + 聚类”模型	-1.8055	0.4875	5.0627	0.0037
“*TF-IDF*+ 共现矩阵 + 聚类”模型	-2.0930	0.2273	2.3626	0.0017
“*TF-IDF*+ 聚类”模型	-0.8223	0.3912	3.9930	0.0514

5　结论与启示

本次构建的经济类专业主题词表，从学习者的专业需求出发，按照教材的语义相似度特征，结合主题词目录，依循主题词的分类，以词簇中心词为基础，掌握语义相关联的一系列专业词汇，这不仅符合学生的认知习得规律，也能够最大化地降低机械记忆，提升意义学习的效果。此外，本文构建的经济类专业主题词表中的词汇是一种原型度较高的语块，学习者掌握相关词汇之后，在今后的学习中易于提取和运用，有利于提升专业术语表达的准确性和专业论文撰写的学术规范性。但是，本次构建的词表所选教材是经济类课程的专业公共课，如果学习者后续进入到专业学习，还需要对专业课程的教材进行调查分析，构建一个面向某一具体专业的主题词表，在后续的研究中将不断对词表进行完善和更新。

参 考 文 献

[1] 刘华，李晓源 . 基于语料库的中医汉语主题词表构建 [J]. 华文教学与研究，2022(2): 77-85.
[2] 高彦德，李国强，郭旭 . 外国人学习与实用汉语情

况调查研究报告 [R]. 北京 : 北京语言学院出版社 , 1993.

[3] 邢红兵 . 词语搭配知识与二语词汇习得研究 [J]. 语言文字应用 , 2013(4): 117-126.

[4] 姜德梧 . 关于《汉字水平词汇与汉字等级大纲》的思考 [J]. 世界汉语教学 , 2004(1): 81-89.

[5] 李红印 .《汉语水平词汇与汉字等级大纲》收“语”分析 [J]. 语言文字应用 , 2005(4): 73-79.

[6] 赵金铭 , 张博 , 程娟 . 关于修订《（汉语水平）词汇等级大纲》的若干意见 [J]. 世界汉语教学 , 2003(3): 73-83.

[7] 张颖怡 , 章成志 , 陈果 . 基于关键词的学术文本聚类集成研究 [J]. 情报学报 , 2019, 38(8): 860-871.

[8] COURTIAL J P. Comments on Leydesdorff's Article[J]. Journal of the American Society for Information Science, 1998, 49(1): 98.

[9] SU H N, LEE P C. Mapping Knowledge Structure by Keyword Co-occurrence: A First Look at Journal Papers in Technology Foresight[J]. Scientometrics, 2010, 85(1): 65-79.

[10] HU J M, ZHANG Y. Research Patterns and Trends of Recommendation System in China Using Co-word Analysis[J]. Information Processing and Management, 2015, 51(4): 329-339.

[11] 齐小英 , 丁子仪 , 杨海平 . 多源数据融合的南海领域主题词表 [J]. 国家图书馆学刊 , 2022, 31(5): 82-94.

[12] 邵卫 , 化柏林 . 基于依存句法分析的科技政策领域主题词表无监督构建 [J]. 情报工程 , 2020, 6(6): 33-44.

[13] 杜慧平 , 薛春香 . 民国抗战史主题词表自动构建研究 [J]. 图书馆杂志 , 2020, 39(8): 75-81.

[14] 仲云云 , 侯汉清 , 杜慧平 . 电子政务主题词表自动构建研究 [J]. 中国图书馆学报 , 2008, 34(3): 97-102.

[15] 曾文 , 王惠临 . 跨语言主题词表自动构建技术研究 [J]. 图书情报工作 , 2011, 55(4): 106-109.

[16] KONDO T, NANBA H, TAKEZAWA T, et al. Technical Trend Analysis by Analyzing Research Papers' Titles[C]// Proceedings of the 4th Conference on Human Language Technology: Challenges for Computer Science and Linguistics. Heidelberg: Springer, 2011: 512-521.

[17] GUPTA S, MANNING C D. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers[C]// Proceedings of the 5th International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing, 2011: 1-9.

[18] TSAI C T, KUNDU G, ROTH D. Concept-based Analysis of Scientific Literature[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2013: 1733-1738.

[19] 罗卓然 , 陆伟 , 蔡乐 , 等 . 学术文本词汇功能识别——在论文新颖性度量上的应用 [J]. 情报学报 , 2022, 41(7): 720-732.

[20] 王亚楠 , 马春鹏 , 曹海龙 , 等 . 面向科技语料的短语结构句法分析器 [J]. 情报工程 , 2017, 3(3): 10-20.

[21] 孙立媛 , 苏新宁 . 面向高校应急决策的领域主题词表构建研究 [J]. 情报科学 , 2019, 37(4): 137-143.

[22] 吴霞 , 曾建勋 , 吴雯娜 . 汉语主题词表生物、医学、农业领域顶层本体语义类型框架研究 [J]. 情报科学 , 2022, 40(1): 94-101.

[23] 苏新春 . 词典与词汇的计量研究 [M]. 上海 : 上海辞书出版社 , 2010: 253.

[24] 苏新春 . 对外汉语教学词汇大纲与两种教材词汇状况的对比研究 [J]. 语言文字应用 , 2006(2): 103-111.

[25] 刘华 . 词语计算与应用 [M]. 广州 : 暨南大学出版社 , 2010: 22.

[26] SALTON G, BUCKLEY C. Improving Retrieval Performance by Relevance Feedback[J]. Journal of the American Society for Information Science, 1990, 41(4): 288-297.

[27] SALTON G, YU C T. On the Construction of Effective Vocabularies for Information Retrieval [J]. Acm Sigplan Notices, 1975, 10(1): 48-60.

[28] 丁国栋 , 白硕 , 王斌 . 一种基于局部共现的查询扩展方法 [J]. 中文信息学报 , 2006(3): 84-91.

[29] 安亚巍 , 操晓春 , 罗顺 . 面向语料的领域主题词表构建算法 [J]. 计算机科学 , 2018, 45(S1): 396-397, 410.

[30] 居亚亚 , 杨璐 , 严建峰 . 基于动态权重的 LDA 算法 [J]. 计算机科学 , 2019, 46(8): 260-265.

[31] 胡文飞 . 二语习得的涌现特征与新型汉英学习词典表征模式的构建 [J]. 外国语文 , 2015, 31(2): 70-76.