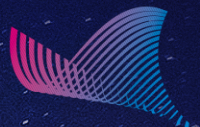




全国科普日



科普中国
CHINA SCIENCE COMMUNICATION

2024年全国科普日活动

提升全民科学素质 协力建设科技强国

9月15日——9月25日

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

情报技术 | INTELLIGENCE TECHNOLOGY

003

基于多源数据的疾病知识图谱构建研究

A Research on the Construction of Disease Knowledge Graph Based on Multi-source Data

孙丝雨 侯跃芳 丁敬达 梅佳月 孙佳

014

基于预训练语言模型的电网检修计划异常检测

A Pre-trained Language Model-based Anomaly Detection Method for Grid Maintenance Plans

吴海洋 吴子辰 吴博科 高华 王莘然 胡伟

025

考虑数据维度与质量维度二重性的数据质量评估研究

Research on Data Quality Evaluation Considering the Duality of Data Dimension and Quality Dimension

刘枏 刘润 李卓城

036

面向民用飞机装配工艺规范数字化建设的标准元数据质量测评研究

Research on the Quality Evaluation of Standard Metadata for the Digital Construction of Civil Aircraft Assembly Process Specifications

苏萍 邹小筑 安鲁陵

情报分析 | INFORMATION ANALYSIS

047

国家学术关注影响因素分析初探

A Preliminary Analysis of Factors Influencing National Academic Attention

刘星涓 伊凡 陆伟

061

国内产业集群研究现状、热点主题及演化路径

Current Status, Hot Topics and Evolution Paths of China's Industrial Cluster Research

魏玲 陈彦均 曾冰 梁立芳

073

基于 DEA 方法的科技企业孵化器绩效评价与对策研究 —— 以安徽省为例

Research on Performance Evaluation and Countermeasures of Technology Business Incubators Based on DEA Method: A Case Study of Anhui Province

杨志发 盛玮芮 张虹霓 刘晶晶

084

山东省科技期刊学术影响力动态研究

Dynamic Study on Academic Influence of Sci-tech Journals in Shandong Province

韦忠明

096

基于 Kano 模型的学术会议视频用户需求研究

Research on User Requirements for Academic Conference Video Based on Kano Model

韩晶

106

省域科技金融效率时空分异及影响因素研究 —— 以山东为例

Research on the Spatiotemporal Differentiation and Influencing Factors of Science and Technology Financial Efficiency in Provincial-level: Evidence from Shandong

姜宁朋 王福康 姜媛 曲妍妍 周立波 张路

114

我国数据要素研究的文献计量与可视化

Research and Analysis of Domestic Data Elements Based on Knowledge Graph

刘桂锋 言怡 刘琮 韩牧哲

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 巍 (澳大利亚)
张智雄	章成志	真 溱	



开放科学
(资源服务)
标识码
(OSID)

基于多源数据的疾病知识图谱构建研究

孙丝雨¹ 侯跃芳² 丁敬达¹ 梅佳月² 孙佳²

1. 上海大学文化遗产与信息管理学院 上海 200444;
2. 中国医科大学健康管理学院 沈阳 110122

摘要: [目的/意义] 基于 PubMed、OMIM 等医学数据库中的多源数据设计疾病知识图谱构建方案, 为疾病的生物学实验研究及诊断治疗提供参考和依据。[方法/过程] 首先利用语义分析工具 SemRep 抽取 SPO 三元组, 通过实体对齐、关系映射等数据处理方法进行知识融合, 然后利用 Neo4j 图数据库实现知识存储及可视化展示, 以多囊卵巢综合征为例进行实证检验和分析, 最终获得 61589 个 SPO 三元组、34697 个实体和 27 种语义关系并归纳总结 7 种语义模式。[局限] 数据处理时, 涉及人工审查, 但由于数据量较大, 审查过程中可能存在些许误差。[结果/结论] 本研究改进现有的知识融合方法, 验证了该疾病知识图谱构建方案的可行性。为后续基于疾病知识图谱进行医学领域知识发现探索奠定基础。

关键词: 疾病知识图谱; SPO 三元组; 知识融合; 语义分析

中图分类号: G302 G35

A Research on the Construction of Disease Knowledge Graph Based on Multi-source Data

SUN Siyu¹ HOU Yuefang² DING Jingda¹ MEI Jiayue² SUN Jia²

1. School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444, China;
2. School of Health Management, China Medical University, Shenyang 110122, China

Abstract: [Objective/Significance] Based on the multi-source data in PubMed, OMIM and other medical databases, the construction scheme of disease knowledge graph is designed to provide reference and basis for biological experimental research, diagnosis and treatment of diseases. [Methods/Processes] Firstly, SPO triples are extracted by SemRep, and knowledge fusion

基金项目 辽宁省教育厅科学研究经费项目(人文社科类基础研究项目)“基于多源数据网络链接预测的知识发现模型构建”(JCRW2020005)。

作者简介 孙丝雨(2000-), 硕士研究生, 主要研究方向为文本挖掘与信息计量; 侯跃芳(1972-), 博士, 教授, 主要研究方向为文本挖掘与知识发现, E-mail: yfhou@cmu.edu.cn; 丁敬达(1978-), 博士, 教授, 主要研究方向为信息计量与科学计量; 梅佳月(1998-), 硕士研究生, 主要研究方向为文本挖掘与知识发现; 孙佳(1998-), 硕士研究生, 主要研究方向为文本挖掘与知识发现。

引用格式 孙丝雨, 侯跃芳, 丁敬达, 等. 基于多源数据的疾病知识图谱构建研究[J]. 情报工程, 2024, 10(4): 3-13.

is carried out by data processing methods such as entity alignment and relationship mapping. Then, knowledge storage and visual display are realized by Neo4j graph database. Taking polycystic ovary syndrome as an example, 61589 SPO triples, 34697 entities and 27 semantic relationships are finally obtained, and 7 semantic patterns are summarized. [Limitations] In the process of data processing, manual examination is involved, but due to the large amount of data, there may be some errors in the examination process. [Results/Conclusions] This study improves the existing knowledge fusion method and verifies the feasibility of the disease knowledge graph construction scheme. It lays a foundation for the follow-up exploration of knowledge discovery in medical field based on disease knowledge graph.

Keywords: Disease Knowledge Graph; Subject-Predication-Object; Knowledge Fusion; Semantic Analysis

引言

随着生物学信息的爆炸式增长,传统的医学知识管理和应用方法已逐渐难以满足当前研究和临床需求。文章创新性地提出一种基于多源医学数据库的疾病知识图谱构建方案,旨在通过先进的语义分析和知识融合技术,为复杂疾病的研究提供一种全新的知识管理和决策支持工具。知识图谱呈现网状的结构,以三元组(头实体,关系,尾实体)的知识构成形式来表达现实世界中的知识或事实,其中图的节点代表实体或者概念,图的边代表实体或概念之间的各种语义关系^[1]。多囊卵巢综合征(Polycystic Ovary Syndrome, PCOS)是一种影响全球许多育龄女性的异质性内分泌疾病^[2],通常与雄激素水平过高、胰岛素抵抗等有关^[3],且增加了引起严重并发症的风险,如心血管疾病等^[4]。

因此,为了促进多囊卵巢综合征医学知识的共享、传播及利用,文章将一体化医学语言系统作为桥梁,实现了PubMed、OMIM等不同医学数据库中数据的标准化,解决了传统方法中数据异构性带来的挑战。采用最新的自然语言处理工具SemRep,通过定制化的语义分析,

精确抽取医学文献中的SPO三元组,为构建知识图谱提供了高质量的原始数据。在知识融合方面,提出一套包含实体对齐、关系映射和属性整合的系统化方法,显著提高了知识图谱中数据的准确性和一致性。利用高性能的Neo4j图数据库,不仅实现了医学知识的高效存储,还通过直观的可视化技术,极大地提升了知识图谱的可访问性和用户的交互体验。

1 相关研究

随着医学信息的日益增加和知识融合技术的不断完善,越来越多的学者利用不同的数据源进行疾病知识图谱的构建研究。蔡妙芝等^[5]利用SemRep在糖尿病专题文献中抽取SPO三元组,利用Neo4j对其进行可视化展示并进行糖尿病知识发现。刘勇等^[6]将电子病历、医学指南等不同数据源中的医学数据进行语义整合,构建了糖尿病医疗知识图谱。Fang等^[7]针对异构数据源之间的知识重复问题,提出了一种基于首尾实体融合模型的实体融合方法。聂莉莉等^[8]利用权威的医学文献和书籍中的医学知识,运用自然语言处理技术,构建以“疾病—症候—特征”三层结构模型为基础的呼吸系统医疗诊

断知识图谱。翟东升等^[9]从 IncoPat 专利数据库和 TCMSP、OMIM 等数据库中获取中医药相关数据,利用深度学习、字符串匹配等方式进行知识抽取、数据规范及实体对齐,构建了基于多源数据的中医药知识图谱。付洋等^[10]以高品质的百科数据及医学文献为基础构建心脏病本体,以自顶向下和自底向上相结合的方法,半自动化地构建了心脏病中文知识图谱。Zhu 等^[11]整合了 34 个不同的生物学数据集中的数据资源,在 Neo4j 中生成了罕见病综合知识图谱,并通过四个案例进行研究分析。

综上所述,近几年医学知识图谱的构建研究取得了一些新的进展。首先,用于构建医学知识图谱的数据来源越来越丰富,包括科研文献、各种医学知识库、电子病历、医疗百科等;其次,医学知识图谱所涵盖的疾病类型越来越多,并逐步从常见疾病发展到罕见疾病。目前,针对多囊卵巢综合征,还鲜见有相关知识图谱的构建研究。此外,虽然知识融合方法和技术得到了快速发展,但目前尚未形成通用性较强的统一融合框架,对多源、异构、大规模医学数据进行知识融合仍存在困难。有鉴于此,

文章基于多源数据设计一套疾病知识图谱构建方案,重点在知识融合方法上进行改进,以一体化医学语言系统(Unified Medical Language System, UMLS)^[12]为标准,实现不同数据源中实体概念对齐,并将不同知识库中的关系类型进行规范化映射。针对各知识库的特点对其审查状态进行分类,再利用该关系属性实现多来源的同一三元组的知识融合,在确保三元组唯一性、消除歧义的同时,提升知识图谱的质量与可信度。为验证该疾病知识图谱构建方案的可行性,文章选取多囊卵巢综合征进行实证检验并深入分析,最后归纳出七种语义模式,以期后续基于疾病知识图谱进行医学领域知识发现、潜在关联预测等方面的研究提供新的方向与思路。

2 疾病知识图谱构建方案

如图 1 所示,文章设计的基于多源数据的疾病知识图谱构建方案主要包括五个部分:知识图谱架构、数据收集与知识抽取、数据处理与清洗、知识融合、知识存储及可视化。

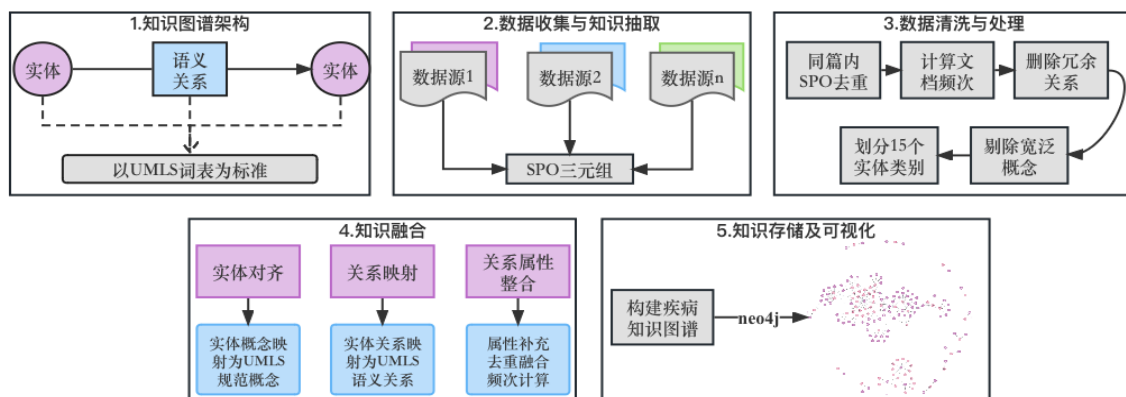


图 1 疾病知识图谱构建方案

SemRep^[13] 由美国国立医学图书馆开发, 是一个从生物医学文本中提取语义关系的 NLP 系统。它是一个基于规则的系统, 以 UMLS 的超级叙词表、语义网络、专家词典和词汇工具为基础, 对生物医学文献进行分词、语义分析、短语映射、语义谓词归一化、语义约束等的处理^[12], 自动提取相关文献中的术语概念和语义关系, 形成主语—谓语—宾语 (Subject-Predication-Object, SPO) 三元组。其中主语和宾语是具有特定语义类型的 UMLS 词表概念, 谓词是 UMLS 语义网络^[14] 扩展版本中的关系类型。目前, SemRep 被广泛应用于生物医学领域的知识图谱构建和知识发现。故文章利用 SemRep 从 PubMed 文献的标题和摘要中抽取 SPO 三元组, 并以 UMLS 词表为标准对知识库中实体和语义

关系进行规范化处理。

2.1 疾病知识图谱架构

知识图谱架构是构建知识图谱的基础, 直接决定了所构建知识图谱的性能与质量。为有效融合医学文献数据和生物知识库数据, 以 UMLS 词表为基础设计知识图谱架构, 具体包括 SPO 三元组的实体、关系命名及其属性定义 (见图 2)。实体指具有可区别性且独立存在的某种事物, 医学领域实体指疾病、药物等; 关系指两个或多个实体间的语义联系, 医学领域实体间的关系包括疾病—疾病、疾病—药物等^[15]。文章定义的实体属性与实体间关系属性如图 2 所示, 其中, 实体类别^[7] 为依据 UMLS 词表提供的语义类型划分的 15 个语义组。

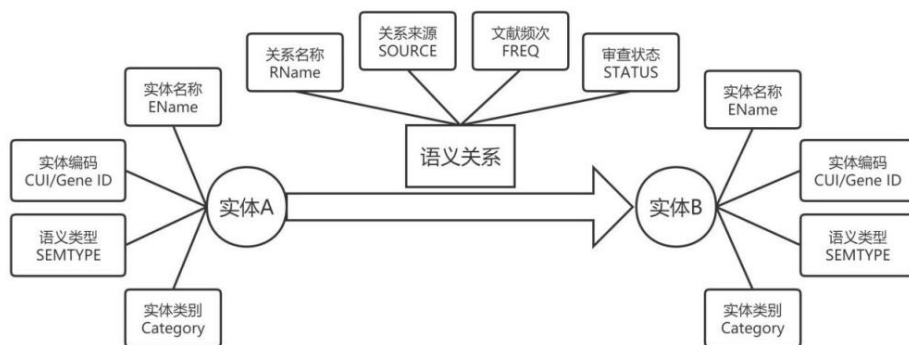


图2 SPO三元组属性定义

2.2 数据收集与知识抽取

相关数据收集是构建知识图谱的首要步骤, 数据质量将直接影响知识图谱的准确性及可用性。知识抽取是指从各种数据源中抽取出实体 (概念)、属性以及实体间的相互关系, 在此基础上形成标准化的知识表达。文章选定 PubMed 文献库和 OMIM、CTD、DisGeNET、

MalaCards 四种知识库作为数据来源。首先, 在 PubMed 文献库中检索多囊卵巢综合征相关文献, 检索式为: "Polycystic Ovary Syndrome" [Mesh], 截止到 2022 年 12 月 31 日, 共得到 17481 篇医学文献。利用 SemRep 对这些文献的标题和摘要进行处理, 得到 SPO 三元组数据表格, 其中包含实体名称、实体 CUI 号 (基

因为 Gene ID)、语义类型、语义关系等。

然后,分别从 OMIM、CTD、DisGeNET、MalaCards 四个知识库中搜集多囊卵巢综合征相关数据,具体过程如下:

(1) 由于 OMIM 数据库中与多囊卵巢综合征相关的数据为文本类型,故对该数据库的数据处理方法与 PubMed 文献库类似,并对处理结果进行人工审核。

(2) 在 CTD 数据库中搜索“Polycystic Ovary Syndrome”,下载化学物质—基因相互作用、疾病相关化学物质、相关基因三部分数据。CTD 对化学物质—基因相互作用提供了三种关系描述,即 increases(增加)、decreases(减少)、affects(影响,未指明方向)。CTD 对化学物质—疾病也提供了三种关系描述,即 marker/mechanism(与疾病相关的化学物质或可能在疾病的病因中起作用的化学物质)、therapeutic(对疾病具有已知或潜在治疗作用的化学物质)和 inferred(推断关系)。推断关系是通过 CTD 审核的化学物质—基因相互作用建立的,如化学物质 A 直接与基因 B 相互作用,基因 B 与疾病 C 相关,则推断化学物质 A 与疾病 C 有关系(通过基因 B)。CTD 对基因—疾病同样提供了以上三种关系。

(3) 在 DisGeNET 数据库中搜索“Polycystic Ovary Syndrome”,下载该数据库中全部基因—疾病关联数据,同时下载该数据库中部分疾病—疾病关联数据,限定依据为经人工审查的数据库推断的疾病间关联。由于 DisGeNET 数据库中的知识关联来源于不同的数据库,因此,文章提供关系来源属性供研究人员进行追根溯源,其中 17 种来源数据库归为四种类型,即人工

审查数据库(CURATED)、动物模型数据库(ANIMAL MODELS)、推断型数据库(INFERRED)和文献数据库(LITERATURE)。

(4) 在 MalaCards 数据库中搜索“Polycystic Ovary Syndrome”,下载该数据库中相关疾病、药物和治疗、基因三部分数据。最后,整理以上各个数据库中的数据,以进行后续数据清洗及知识融合。

2.3 数据处理与清洗

数据处理与清洗是知识图谱构建过程中极为重要的步骤,筛选并保留高质量的 SPO 三元组,才能确保知识图谱的可信度。因此,文章首先对来源于 PubMed 文献库的 SPO 三元组进行处理与清洗,主要包括:同篇 SPO 三元组去重;计算同一 SPO 在不同文献中出现的频次;剔除否定谓词如“NEG_AFFECTS”等;人工剔除无意义的宽泛词,如“Human”等;依据语义类型划分 15 个实体类别。然后,依据各个知识库的数据特点,保留实体名称、文献频次等,补充实体类别、关系名称等,并将冗余数据删除,为后续知识融合阶段打下基础。

2.4 知识融合

知识融合是指通过映射和匹配使不同来源的知识在同一框架规范下进行整合、消歧和加工的过程。知识融合对提升知识图谱的质量、知识复用以及实现异构数据源之间的语义互通都具有重要意义^[16]。因此,文章提出了一套系统化的知识融合方法,包括实体对齐、关系映射和关系属性整合,对不同来源的异构数据进行融合。

（1）实体对齐

为确保同一实体概念标识的唯一性，发现并匹配不同数据源中相同的实体概念^[17]，文章利用 Postman 软件和 UMLS API 将从各个知识库中抽取的实体概念映射到 UMLS 词表中规范化的术语概念，完成实体对齐。但由于 UMLS 词表并不能包含所有的医学实体概念，针对那些未映射成功的实体概念，需要进行人工审查，确定其是否被 UMLS 词表收录。若已收录，则保留该实体概念的规范化名称及 CUI 号，若未收录，则保留该实体概念的原名

称并将其 CUI 号赋予“NULL”。若出现多个映射结果，也需要进行人工审查，以确保映射结果的唯一性。

（2）关系映射

如上文所述，CTD 数据库为实体间关系提供了多种描述，需要进行映射处理，DisGeNET、MalaCards 中的实体间关系也需要定义和补充。文章以 UMLS 语义网络中的语义关系为标准，对知识库中的关系类型进行规范化处理，确保不同来源的 SPO 三元组中的实体间关系表达完整且统一，详情如表 1 所示。

表 1 知识库中的关系映射

数据来源	实体间关系	原关系名称	映射后关系名称
CTD 数据库	化学物质—疾病	marker/mechanism	CAUSES
		therapeutic	TREATS
		inferred	ASSOCIATED_WITH
	基因—疾病	marker/mechanism	CAUSES
		therapeutic	TREATS
		inferred	ASSOCIATED_WITH
	化学物质—基因	increases	STIMULATES
		decreases	INHIBITS
		affects	AFFECTS
DisGeNET	基因—疾病	无	ASSOCIATED_WITH
	疾病—疾病	无	ASSOCIATED_WITH
MalaCards	药物—疾病	无	TREATS
	基因—疾病	无	ASSOCIATED_WITH
	疾病—疾病	无	ASSOCIATED_WITH

（3）关系属性整合

经过实体对齐，已经较好地融合了实体的属性，实现了 SPO 三元组概念表达的一致性。对于关系的属性，由于各个数据库的审查状态不尽相同，文章将审查状态分为两种类型：人工审查（CURATED）、非人工审查（UNCU-

RATED），如表 2 所示。

若出现同一三元组来源于不同数据库的情况（如某一 SPO 三元组既来源于 PubMed 文献库又来源于 CTD 知识库），为保证 SPO 三元组的唯一性，基于审查状态属性将多来源的同一三元组进行去重融合，将关系来源属性和审

表2 审查状态属性定义

数据来源	关系来源情况	审查状态
PubMed	经 SemRep 抽取得到	UNCURATED
CTD	CTD 数据库提供确定的关系（如 therapeutic）	CURATED
	实体间的关系是推断关系	UNCURATED
DisGeNET	来源于经人工审查的数据库或动物模型数据库	CURATED
	来源于推断型数据库或文献数据库	UNCURATED
MalaCards	已批准的药物、具有明确证据的基因	CURATED
	由数据库推断得到的相关疾病和基因、除已批准以外的其余药物	UNCURATED
OMIM	SemRep 抽取后进行人工审核	CURATED

查状态属性进行整合更新（如 SOURCE:CTD, PubMed, STATUS:CURATED），并对文献频次取较高频次，生成最终的 SPO 三元组语义网，避免同一 SPO 三元组因来源不同在知识图谱中多次出现的情况，确保后续展示时的美观度及查询时的高效性。

2.5 知识存储及可视化

选择合理高效的知识存储方式，能够直接提高知识图谱的查询效率。Neo4j 图数据库因其具有高性能、实用性强、轻量级等优点，成为目前最常用的图数据库之一^[18]。基于此，文章利用 Neo4j 图数据库对 SPO 三元组进行知识存储及可视化展示，完成多囊卵巢综合征知识图谱的构建。

2.6 语义模式归纳

SPO 三元组语义模式是用于描述实体之间关系和属性的一种表示方式，文章定义的 SPO 三元组语义模式表示为“语义类型 A—语义关系—语义类型 B”，如 orch-TREATS-dsyn，orch 即 Organic Chemical（有机化学药品），dsyn 即 Disease or Syndrome（疾病或综合征），

该语义模式表示有机化学药品治疗疾病或综合征。参考蔡妙芝等^[5]总结的“诊断治疗”等 5 种语义模式，文章进一步归纳总结 7 种语义模式并进行详解。

3 结果与分析

3.1 PCOS知识图谱

文章从 PubMed 文献库中共获得 29721 个 SPO 三元组，加入四个知识库中收集的数据，并进行知识融合处理后，总计得到 61589 个 SPO 三元组，34697 个实体和 27 种语义关系。在 Neo4j 图数据库中的部分展示结果如图 3 所示。在图 3 中，以“Polycystic Ovary Syndrome”为中心，周围节点是与该疾病有明确语义关系的相关疾病（红色）、基因（绿色）、药物（黄色）等，例如胰岛素抵抗、FTO 基因、维生素 D 等。右侧图示中展示了该节点的四个属性，即 {EName: “Polycystic Ovary Syndrome”, ID: “C0032460”, SEMTYPE: “dsyn”, CATEGORY: “Disorders”}。

图 4 为知识图谱中 SPO 三元组的实体及关系属性示例，节点（实体）集合 E 为 {E1、

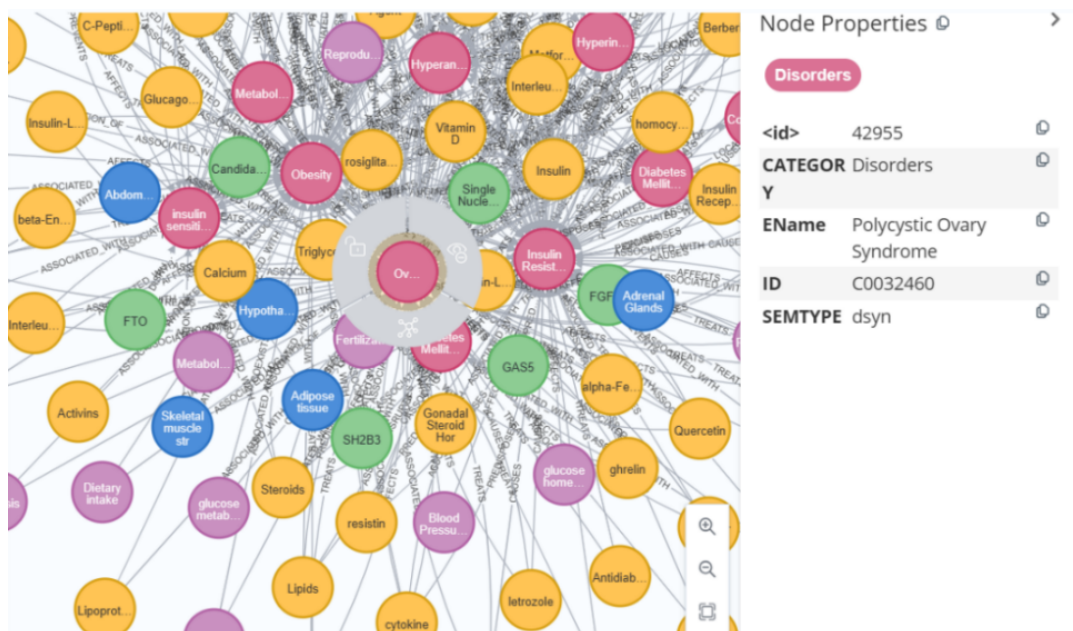


图3 PCOS知识图谱局部展示(实体)

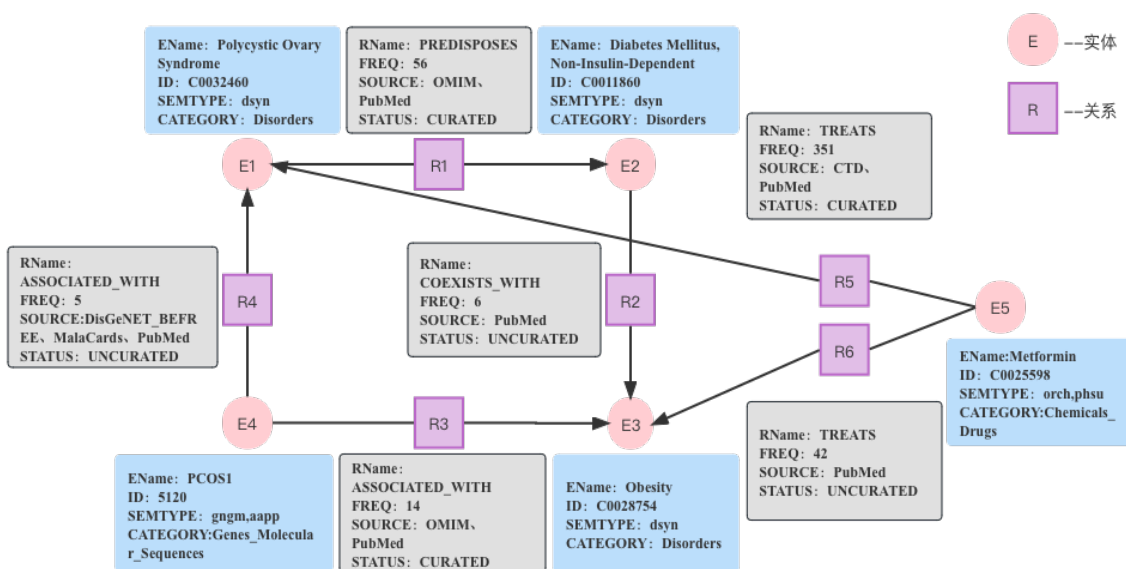


图4 PCOS知识图谱属性图示例

$E2, E3, E4, E5\}$, 边(关系)集合 R 为 $\{R1, R2, R3, R4, R5, R6\}$ 。由图可知,多囊卵巢综合征容易诱发非胰岛素依赖型糖尿病,非胰岛素依赖型糖尿病与肥胖并发,二甲双胍可用于治疗多囊卵巢综合征和肥胖,PCOS1基因与多囊卵巢综合征和肥胖有关。

3.2 语义模式归纳与分析

文章将语义关系总结归纳为相关疾病、诊断治疗、影响因素、药理作用、疾病特征、功能相关、比较关系七种语义模式,具体结果如表3所示。

表 3 PCOS 知识图谱的三元组语义模式

类型	语义关系	语义模式示例	SPO 三元组示例
相关疾病	COEXISTS_WITH	dsyn-COEXISTS_WITH-patf	Spontaneous abortion-COEXISTS_WITH-Insulin Resistance
	COMPLICATES	dsyn-COMPLICATES-dsyn	Hyperinsulinism-COMPLICATES-Disorder of endocrine ovary
	PRECEDES	dsyn-PRECEDES-dsyn	Amenorrhea-PRECEDES-Polycystic Ovary Syndrome
诊断治疗	ADMINISTERED_TO	horm-ADMINISTERED_TO-mamm	estradiol valerate-ADMINISTERED_TO-Rattus norvegicus
	DIAGNOSES	lbpr-DIAGNOSES-dsyn	Follicle stimulating hormone measurement-DIAGNOSES-Polycystic Ovary Syndrome
	PREVENTS	orch-PREVENTS-comd	Metformin-PREVENTS-Oxidative Stress
	TREATS	orch-TREATS-dsyn	Metformin-TREATS-Hirsutism
关联因素	AFFECTS	comd-AFFECTS-dsyn	Oxidative Stress-AFFECTS-Polycystic Ovary Syndrome
	ASSOCIATED_WITH	phsu-ASSOCIATED_WITH-dsyn	Vitamin D-ASSOCIATED_WITH-Obesity
	CAUSES	dsyn-CAUSES-dsyn	Hyperandrogenism-CAUSES-Infertility
	PREDISPOSES	dsyn-PREDISPOSES-mobd	Polycystic Ovary Syndrome -PREDISPOSES-Depressive disorder
	PRODUCES	bpoc-PRODUCES-horm	Ovarian Follicle-PRODUCES-Estrogens
药理作用	AUGMENTS	carb-AUGMENTS-patf	Inositol-AUGMENTS-insulin sensitivity
	DISRUPTS	strd-DISRUPTS-dsyn	Steroids-DISRUPTS-Polycystic Ovary Syndrome
	INHIBITS	horm-INHIBITS-horm	Testosterone-INHIBITS-Progesterone
	INTERACTS_WITH	phsu-INTERACTS_WITH-horm	Aromatase-INTERACTS_WITH-Estradiol
	STIMULATES	horm-STIMULATES-aapp	Gonadorelin-STIMULATES-Luteinizing Hormone
疾病特征	ISA	dsyn-ISA-dsyn	Polycystic Ovary Syndrome-ISA-Endocrine System Diseases
	LOCATION_OF	bpoc-LOCATION_OF-neop	Endometrium-LOCATION_OF-Adenocarcinoma
功能相关	MANIFESTATION_OF	dsyn-MANIFESTATION_OF-dsyn	Acne-MANIFESTATION_OF-Hyperandrogenism
	OCCURS_IN	dsyn-OCCURS_IN-humn	Polycystic Ovary Syndrome-OCCURS_IN-Adolescents, Female
	PROCESS_OF	moft-PROCESS_OF-cell	aromatase activity-PROCESS_OF-granulosa cell
	USES	topp-USES-phsu	Oestrogen therapy-USES-Progestins
比较关系	COMPARED_WITH	phsu-COMPARED_WITH-phsu	Acetylcysteine-COMPARED_WITH-Metformin
	HIGHER_THAN	orch-HIGHER_THAN-phsu	Clomiphene-HIGHER_THAN-Metformin
	LOWER_THAN	phsu-LOWER_THAN-phsu	Metformin-LOWER_THAN-Clomiphene Citrate
	SAME_AS	phsu-SAME_AS-phsu	Letrozole-SAME_AS-Clomiphene Citrate

在“相关疾病”语义模式中, COEXISTS_WITH 表示疾病的并发症; COMPLICATES 表示使其变得更加严重或复杂, 或导致不利影响; PRECEDES 表示在时间上发生得更早。

在“诊断治疗”语义模式中, ADMINISTERED_TO 表示某种药物、治疗或其他医疗干预措施被施用于特定患者或生物体上; DIAGNOSES 表示区分或识别疾病的性质或特征, 即诊断 PCOS 的手段或指标; PREVENTS 表示阻碍或消除病情的药物或手段; TREATS 表示以治愈或控制病情为目的治疗方法。

在“关联因素”语义模式中, AFFECTS 表示产生直接影响; ASSOCIATED_WITH 表示疾病、基因和药物之间的关联关系; CAUSES 表示导致某种状况或结果, 主要揭示了疾病的致病因素; PREDISPOSES 表示对某种疾病、病理或状况存在风险, 主要揭示了疾病的诱发因素; PRODUCES 表示产生或创造, 这包括分泌、生物合成、释放等。

在“药理作用”语义模式中, AUGMENTS 表示扩展或刺激一个过程, STIMULATES 表示增加或促进物质间相互作用, 二者都反映了正向的关系。DISRUPTS 表示改变或影响已经存在的条件、状态或情况, 产生负面影响, INHIBITS 表示减少、限制或阻断物质间相互作用, 二者都反映了负向的关系。INTERACTS_WITH 表示物质间相互作用。

在“疾病特征”语义模式中, ISA 表示实体间的从属关系; LOCATION_OF 表示一个实体的位置、部位或区域或一个进程的部位。

在“功能相关”语义模式中, MANIFESTATION_OF 表示一种现象的一部分可以直接观

察到, 或明显地表现出来, 或为潜在的过程提供证据; OCCURS_IN 表示在一个组或群体中有发病率; PROCESS_OF 表示疾病发生在高等有机体中; USES 表示用于执行某些活动。

在“比较关系”语义模式中, COMPARED_WITH、HIGHER_THAN、LOWER_THAN、SAME_AS 都是比较谓词, 分别表示相较于、高于、低于、相同, 如奥利司他治疗多囊卵巢综合征的疗效与二甲双胍相同 (orlistat-SAME_AS-Metformin)。

4 总结与展望

文章首先设计了疾病知识图谱的构建方案, 以 PubMed 文献库和 OMIM 等四种知识库为数据来源, 以 UMLS 词表为基础定义了实体和关系的命名及其多种属性, 并提出了一套系统化的知识融合处理流程, 以多囊卵巢综合征为例验证了其可行性, 可为其他疾病知识图谱的构建提供一定参考。虽然文章中的知识融合涉及一定程度的人工审核, 但所采用的方法, 如实体对齐、关系映射和关系属性整合, 可以为其他研究者在知识融合自动化方面提供思路和框架。随着自然语言处理 (NLP) 和机器学习技术的发展, 这些人工环节有望被自动化算法所替代, 从而提高效率并减少误差。

通过利用 Neo4j 图数据库, 文章实现了知识图谱的高效存储和直观可视化展示, 这不仅提高了知识图谱的查询效率, 也使得医学专业人员能够更容易地理解和利用知识图谱。在深入分析后, 文章归纳总结了 SPO 三元组的七种语义模式, 这有助于机器更好地理解、处理和

使用知识图谱中的数据,从而实现更高效、智能的语义分析和查询。后续可以尝试开展 PCOS 知识图谱的网络展示及查询,供生物医学相关人员使用,为多囊卵巢综合征的诊断治疗提供相应的参考和支持;同时,基于该知识图谱,尝试进行多囊卵巢综合征知识发现探索,为科研人员提供潜在知识关联与研究思路;该知识图谱提供了规范化的、机器可理解的与多囊卵巢综合征疾病相关的知识,也可尝试将其应用于智慧问答系统的开发,推动智能医疗的发展。

参考文献

- [1] 董文波,孙仕亮,殷敏智.医学知识推理研究现状与发展[J].计算机科学与探索,2022,16(6): 1193-1213.
- [2] ROTHENBERG S S, BEVERLEY R, BARNARD E, et al. Polycystic ovary syndrome in adolescents[J]. Best Practice & Research. Clinical Obstetrics & Gynaecology, 2018, 48: 103-114.
- [3] WITCHEL S F, OBERFIELD S E, PEÑA A S. Polycystic Ovary Syndrome: Pathophysiology, Presentation, and Treatment with Emphasis on Adolescent Girls[J]. Journal of the Endocrine Society, 2019, 3(8): 1545-1573.
- [4] GLUECK C J, GOLDENBERG N. Characteristics of obesity in polycystic ovary syndrome: Etiology, treatment, and genetics[J]. Metabolism: Clinical and Experimental, 2019, 92: 108-120.
- [5] 蔡妙芝,李晓瑛,赵嘉玮,等.基于 SPO 语义三元组的疾病知识发现[J].数据分析与知识发现,2022,6(1): 134-144.
- [6] 刘勇,齐梦霁.基于糖尿病防治的医学知识图谱构建的研究[J].医学信息,2020,33(18): 11-14.
- [7] FANG A, LOU P, HU J, et al. Head and Tail Entity Fusion Model in Medical Knowledge Graph Construction: Case Study for Pituitary Adenoma[J]. JMIR Medical Informatics, 2021, 9(7): e28218.
- [8] 聂莉莉,李传富,许晓倩,等.人工智能在医学诊断知识图谱构建中的应用研究[J].医学信息杂志,2018,39(6): 7-12.
- [9] 翟东升,娄莹,阚慧敏,等.基于多源异构数据的中医药知识图谱构建与应用研究[J].数据分析与知识发现,2023,7(9): 146-158.
- [10] 付洋,刘茂福,乔瑞.心脏病中文知识图谱的构建[J].武汉大学学报(理学版),2020,66(3): 261-267.
- [11] ZHU Q, NGUYEN D T, GRISHAGIN I, et al. An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD)[J]. Journal of Biomedical Semantics, 2020, 11(1): 13.
- [12] 李晓瑛,李军莲,李丹亚.一体化医学语言系统及其在知识发现中的应用研究[J].数字图书馆论坛,2019(9): 24-29.
- [13] RINDFLESCH T C, FISZMAN M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text[J]. Journal of Biomedical Informatics, 2003, 36(6): 462-477.
- [14] KILICOGLU H, ROSEMBLAT G, FISZMAN M, et al. Constructing a semantic predication gold standard from the biomedical literature[J]. BMC bioinformatics, 2011, 12: 486.
- [15] 刘悦悦,李燕.医学知识图谱研究综述[J].软件导刊,2023,22(5): 241-247.
- [16] 范媛媛,李忠民.中文医学知识图谱研究及应用进展[J].计算机科学与探索,2022,16(10): 2219-2233.
- [17] 张晗,安欣宇,刘春鹤.基于多源语义知识图谱的药物知识发现:以药物重定位为实证[J].数据分析与知识发现,2022,6(7): 87-98.
- [18] 孙敏敏,毛雪岷.基于 Neo4j 的肺部疾病知识图谱构建[C]//中国管理现代化研究会,复旦管理学奖励基金会.第十五届(2020)中国管理学年会论文集,2020: 6.



开放科学
(资源服务)
标识码
(OSID)

基于预训练语言模型的电网检修计划异常检测

吴海洋¹ 吴子辰² 吴博科³ 高华² 王莘然² 胡伟⁴

1. 国网江苏省电力有限公司 南京 210000;
2. 国网江苏省电力有限公司信息通信分公司 南京 210000;
3. 国网常州供电公司 常州 213003;
4. 南京大学计算机软件新技术全国重点实验室 南京 210023

摘要: [目的/意义] 电网检修计划异常检测技术可以发现电网检修计划内存在的异常,对于电网通信系统检修活动的正常开展和电力通信系统的稳定运行至关重要。目前,电网检修计划异常检测主要依靠人工判别的方式寻找检修计划的异常,存在准确率低与效率低下的问题。[方法/过程] 提供了一种基于预训练语言模型的电网检修计划异常检测方法,使用电网的检修计划数据对预训练语言模型进行微调,使得预训练语言模型能够获取相关领域的专业知识,同时充分利用预训练语言模型的上下文感知能力和领域通用性,使预训练语言模型能够深入理解检修计划的复杂语境,从而发现检修计划中存在的异常。[结果/结论] 实验证明,相较于传统方法,该异常检测方法能够有效优化电网检修计划异常检测的性能。

关键词: 预训练语言模型; 异常检测; 电网管理; 文本分类

中图分类号: G35; TP391

A Pre-trained Language Model-based Anomaly Detection Method for Grid Maintenance Plans

WU Haiyang¹ WU Zichen² WU Boke³ GAO Hua² WANG Shenran² HU Wei⁴

1. State Grid Jiangsu Electric Power Co., Ltd., Nanjing 210000, China;
2. State Grid Jiangsu Electric Power Co., Ltd. Information & Telecommunication Branch, Nanjing 210000, China;
3. State Grid Changzhou Power Supply Company, Changzhou 213003, China;
4. State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China

Abstract: [Objective/Significance] Anomaly detection methods are designed to detect anomalies in grid maintenance plans, which are crucial for the normal implementation of grid maintenance and the stable operation of the grid system. Existing

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

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

引用格式 吴海洋, 吴子辰, 吴博科, 等. 基于预训练语言模型的电网检修计划异常检测[J]. 情报工程, 2024, 10(4): 14-24.

methods primarily rely on manual discernment, leading to issues of low accuracy and inefficiency. [Methods/Processes] This paper presents an anomaly detection method for grid maintenance plans based on pre-trained language models (PLMs). The method uses grid maintenance plan data to fine-tune the PLMs, so that the PLMs can obtain professional knowledge in related fields, while fully leveraging the contextual awareness and domain adaptability of PLMs. The combination of the two aspects enables the PLMs can deeply understand the complex context of the maintenance plans and detect anomalies in them. [Results/Conclusions] Experimental results demonstrate that the proposed method effectively enhances the performance of anomaly detection in grid maintenance plans.

Keywords: Pre-trained Language Model; Anomaly Detection; Grid Management; Text Classification

引言

在电网通信领域，定期检修电网线路对于保障电网系统的正常运行至关重要。每次线路检修都需要制定详细的检修计划，以确保检修过程能够尽快将电网系统恢复到正常的运行状态。

电网通信系统通常涉及复杂的网络结构和庞大的设备数量，而检修计划通常由不同的部门独立制定并上报进行审批，这使得多个检修计划之间可能存在冲突，即电网检修计划的异常。异常的存在会导致维修效率降低，给电网的可靠性带来威胁。目前，电网检修计划异常检测主要依赖工作人员对申报的检修计划进行审批与排布，以确保检修计划的可行性，优化维修效率。由于检修计划数量大、异常特征不明显，人工检测存在异常识别率低与效率低下的问题。如何高效地识别电网检修计划中的异常，成为当前电网研究中的一个紧迫问题。

本文基于大量的电网检修计划数据，对检修计划中常见的异常进行分析，基于文献[1]并结合对相应检修计划数据的观察，总结出3种检修计划中常见的异常（如表1所示），并对应地提出了3种检修计划异常检测任务。本文将这3种任务建模为自然语言处理任务中的文本分类任务，并设计了一种基于预训练语言模型的电网检修计划异常检测方法，该方法利用预训练语言模型在预训练阶段所获得的通用知识和强大的上下文感知能力，在全面考虑检修计划语义信息的基础上准确识别检修计划的异常。本文对所提出的方法进行了实验验证，并对实验结果进行了详细的分析。实验证明，相较于传统方法，本文提出的电网检修计划异常检测方法在准确性和鲁棒性上都取得了显著的提升，具备较强的泛化能力和实际应用潜力，有望为电网通信领域的检修计划异常检测提供有力支持，为电网智能化发展做出贡献。

表1 常见的检修计划异常

异常类型	检修计划实例	异常说明
检修步骤与检修目标无关	某光缆割接至临时过渡光缆运行：（1）拆除原48#塔；（2）在31#塔处断开原光缆；（3）沿38#杆至31#塔布放临时光缆；（4）将临时光缆与原光缆进行熔接。	检修步骤（1）与检修目标无关
检修步骤顺序错乱	某光缆段改造：（1）开断原光缆；（2）布放新光缆；（3）拆除原光缆；（4）熔接新光缆。	检修步骤（2）和（3）顺序错乱
检修计划间存在冲突	【检修计划1】某光缆中断检修：…（2）从13#电力井至39#电力井断开原光缆；… 【检修计划2】某光缆段迁移至新管道：…（3）从13#电力井至39#电力井拆除原光缆；…	检修与改造无法同时开展，两个检修计划存在冲突

1 相关工作

1.1 文本分类任务

文本分类是自然语言处理（NLP）领域中的一个重要任务，旨在将文本分配到预定义类别或者标签内。传统方法采用朴素贝叶斯^[2]、支持向量机（SVM）^[3]、决策树^[4]等传统机器学习算法进行分类。基于深度学习的神经网络模型包括卷积神经网络（CNN）^[5]、循环神经网络（RNN）^[6]以及更加先进的长短期记忆网络（LSTM）^[7]和门控循环单元（GRU）^[8]等。这些方法难以捕捉全局依赖关系，对上下文理解不足，面对电网检修计划这类长文本时表现不佳。

由于存在异常的检修计划样本量少，电网检修计划异常检测问题是一个非平衡的文本分类任务。为解决非平衡分类问题，传统的机器学习领域通常采用过采样少数类别的样本^[9]、集成多个分类器的预测结果^[10]等技术来提高非平衡分类模型的性能。而在深度神经网络领域，则通过生成对抗网络（GAN）^[11]生成合成的样本、迁移学习^[12]以及优化损失函数^[13]等技术进一步提升模型的分类性能。本文提出的基于预训练语言模型的电网检修计划异常检测方法，将预训练模型蕴含的语义知识迁移到电网领域的检测任务上，有效提高了模型在非平衡数据集上的性能。

1.2 异常检测任务

电网检修领域的冲突检测目前以人工处理为主。有学者^[14-15]提出构建专家系统与知

识库实现检修计划的智能分析，但是需要经验丰富的专家手动完成，构建的知识库规模也较小。随着自然语言处理技术的发展，一些工作依靠领域专家进行冲突规则模板的预设^[16-18]，然后通过判断输入的检修计划文本是否匹配规则进行冲突检测。此类方法人工维护成本高，规则根据需求不断调整。由于深度学习的发展，BERT等预训练语言模型被用于电网通信系统的命名实体识别^[19]，自动化抽取检修计划中的关键信息，降低人工维护的成本，但是将预训练模型进一步应用到难度更大的异常检测任务中，则面临领域内的训练语料不足、模型性能不好等挑战。本文提出的电网检修计划异常检测方法可以高效识别电网检修计划的潜在冲突，推动电网智能化发展。

1.3 预训练语言模型

预训练语言模型基于Transformer^[20]等架构在大规模的无标注语料库上进行自监督学习，使得模型能够学到通用的语言表示，捕获语法、语义和世界知识。代表性的模型包括BERT^[21]和LLaMA^[22]。BERT能够深入理解文本中的上下文信息，面对检修计划这种存在大量专业术语和复杂技术性描述的文本，能够更好地处理这种语境敏感性，提高异常检测的准确性。LLaMA是文本生成的语言模型，可以对输入的问题进行理解和生成回答。

关于预训练语言模型在专业领域中的应用，现有工作涉及电网设备缺陷文本分类^[23]、商品属性抽取^[24]、案件要素识别^[25]等包含大量专有

名词的领域。这些工作表明，通过使用专业领域数据对预训练语言模型进行微调，预训练语言模型能充分理解专业领域的知识，提升模型在专业领域任务上的性能。本文同样借助检修计划语料对 BERT 和 LLaMA 进行微调，使其能根据获取的检修计划知识，结合自身的上下文感知能力和领域通用性，更加深入理解检修计划的复杂语境，从而发现检修计划中存在的异常。

2 任务定义

2.1 检修计划

检修计划的结构与实例如图 1 所示，一个检修计划包含一个检修目标和若干有序的检修步骤。本文将检修计划 P 定义为 $P=(g, s_1, s_2, \dots, s_N)$ ，其中 g 表示检修计划的目标， $s_i(i \in \{1, 2, \dots, N\})$ 表示计划的第 i 个步骤， N 表示该检修计划所包含的检修步骤的数量。检修目标和检修步骤是自然语言构成的文本序列。

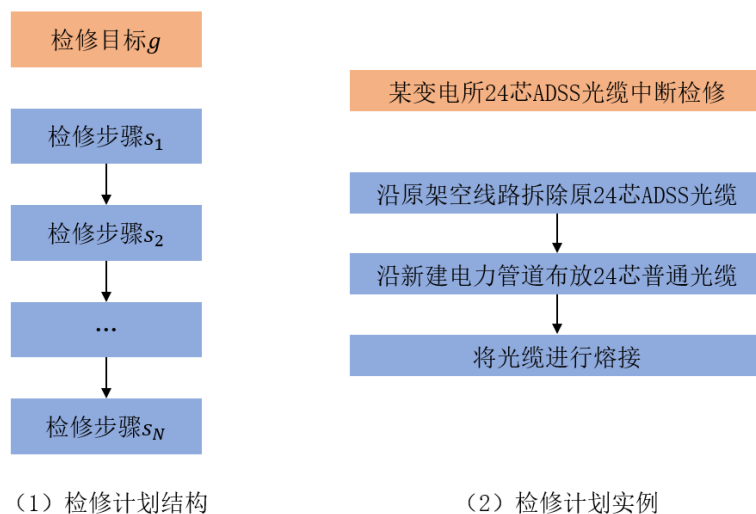


图 1 检修计划结构与实例

2.2 电网检修计划异常

根据研究者提到的步骤推理、顺序推理等任务^[1]，以及结合对检修计划数据的观察，本文总结了 3 种常见的检修计划异常，分别是：

(1) 检修步骤是否与检修目标相关；(2) 检修步骤的顺序是否存在错误；(3) 两个检修计划是否可同时开展。考虑到这 3 种异常相互独立，本文提出分别针对 3 种异常的检测任务，并将每个异常的检测任务建模为自然语言处理任务中的文本分类任务来进行后续的处理。具体如下：

检修步骤相关性检测：第一种常见的检修计划异常是检修步骤中包含与当前检修目标不相关的检修步骤，不相关的检修步骤会对检修计划的开展产生未知的影响。针对这种异常，本文设计了检修步骤相关性检测任务。该任务给定检修目标 g 与检修步骤 s_i ，目标是判断 g 与 s_i 是否属于同一个检修计划。

检修步骤顺序性检测：第二种常见的检修计划异常是检修步骤之间顺序错乱。检修步骤之间通常具有先后顺序，按序执行才可确保检修计划的正常实施。针对这种异常，本文设计

了检修步骤顺序性检测任务。该任务给定检修计划 P 的检修目标 g 以及两个检修步骤 s_i 、 s_j ，目标是判断 s_i 和 s_j 执行的先后顺序。值得注意的是，检修目标 g 对于判断检修步骤之间的先后顺序是必须的，因为不同的检修计划对于同样两个检修步骤的顺序要求可能不同。

检修计划冲突检测：与前两种检修计划异常不同，第三种常见的检修计划异常发生在两个检修计划之间。不同的检修计划可能涉及同一线路同一时间的不同操作，这些操作之间是互斥的。即使两个检修计划都是准确无误的，检修计划也有可能无法正常开展。针对这种异

常，本文设计了检修计划冲突检测任务，该任务给定两个检修计划 P_1 和 P_2 ，目标是判断两个检修计划是否可以同时开展。

3 电网检修计划异常检测方法

本文设计了基于 BERT 的异常检测模型 GAD(Grid Anomaly Detection)-BERT 以及基于 LLaMA 的异常检测模型 GAD-LLaMA，如图 2 所示。其中，GAD-BERT 整体采用编码器与分类器的结构；GAD-LLaMA 对 LLaMA 进行指令微调。本节将分别介绍提出的两个检修计划异常检测模型。



图2 检修计划异常检测模型框架图

3.1 基于BERT的电网检修计划异常检测方法

GAD-BERT 主要包含编码器与分类器两个模块，编码器通过 BERT 编码输入文本并获取向量表示。分类器是一个多层感知机，用于计算异常存在与否的概率。三种任务输入不同，编码器的设计也略有不同。本节首先介绍 GAD-BERT 进行电网异常检测的通用流程，之后分别介绍三种任务不同的编码方式。值得注意的是，三种任务对应的异常检测模型不共享

任何参数，彼此独立。

GAD-BERT 采用 BERT 作为基础的编码器，给定一段输入文本 T ，通过分词得到序列 $x=[[-CLS], t_1, t_2, \dots, t_n, [SEP]]$ ，其中 $t_i(i \in \{1, 2, \dots, n\})$ 表示构成 T 的第 i 个词语， t_i 是预定义在 BERT 词表中的词语， $[CLS]$ 和 $[SEP]$ 是 BERT 预训练时使用的特殊词语，分别用于表示输入序列的开始和结束，这两个特殊词语并不包含在输入文本 T 中。

将序列 x 输入到 BERT 中, 可以获取序列中每个词语对应的向量表示:

$$[e_{[\text{CLS}]}, e_1, e_2, \dots, e_n, e_{[\text{SEP}]}] = \text{BERT}(x) \quad (1)$$

其中, $e_{[\text{CLS}]} \in R^d$ 是特殊词语 [CLS] 对应的向量表示, 其中 d 是 BERT 输出的向量表示的维数。将 $e_{[\text{CLS}]}$ 视为输入文本 T 的向量表示 e_T , 输入到分类器中完成分类:

$$p = \text{Softmax}(W_2 \text{ReLU}(W_1 e_T + b_1)) + b_2 \quad (2)$$

其中, $p \in R^2$ 是输出的概率向量, $p[0]$ 表示输入的检修计划存在异常的概率, $p[1]$ 表示不存在异常的概率, Softmax 函数保证了 $p[0] + p[1] = 1$ 。本文所使用的分类器是由两个全连接层构成的感知机, 其中 $W_1 \in R^{d_1 \times d}$, $W_2 \in R^{2 \times d_1}$, $b_1 \in R^{d_1}$, $b_2 \in R^2$ 是多层感知机的参数, d_1 是一个超参数, 表示感知机中间表示的维数, ReLU 是感知机采用的非线性激活函数。

在获得预测概率后, 通过交叉熵函数计算分类损失 L :

$$L = \text{CE}(p, l) \quad (3)$$

其中, CE 表示交叉熵损失函数, l 是输入文本 T 的标记信息, $l = [1, 0]$ 表示 T 中存在异常, $l = [0, 1]$ 表示 T 中不存在异常。基于分类损失, 通过梯度反向传播算法计算模型参数的梯度, 对模型进行优化。

3.1.1 检修计划相关性检测编码器

检修计划相关性检测任务的输入是检修目标 g 与检修步骤 s_i , 对 g 和 s_i 同时进行分词, 得到序列:

$$x_r = [[\text{CLS}], t_1, t_2, \dots, t_n, [\text{SEP}]], \\ t_{n+1}, t_{n+2}, \dots, t_{n+m}, [\text{SEP}]] \quad (4)$$

其中 x_r 表示经过分词处理后的序列, $t_1, t_2, \dots, t_n$ 和 $t_{n+1}, t_{n+2}, \dots, t_{n+m}$ 分别是检修目标 g 与检

修步骤 s_i 对应的分词结果。

将序列 x_r 输入到 BERT 中, 获取序列中每个元素对应的向量表示:

$$[e_{[\text{CLS}]}, e_1, e_2, \dots, e_n, e_{[\text{SEP}]}, e_{n+1}, e_{n+2}, \dots, e_{n+m}, e_{[\text{SEP}]}] = \text{BERT}(x_r) \quad (5)$$

选择 $e_{[\text{CLS}]} \in R^d$ 作为检修目标 g 与检修步骤 s_i 统一的向量表示, 记作 e_r 。

将 e_r 输入到检修计划相关性检测模型的分类器中, 得到预测的概率 p_r 。 p_r 与标记信息 l_r 通过交叉熵损失函数计算得到检修计划相关性检测的损失 L_r 。

3.1.2 检修计划顺序性检测编码器

检修计划顺序性检测任务的输入是检修目标 g 以及检修步骤 s_i, s_j , 对这三部分文本分别进行分词, 得到对应的序列:

$$x_g = [[\text{CLS}], t_1^g, t_2^g, \dots, t_{n_1}^g, [\text{SEP}]] \quad (6)$$

$$x_{s_i} = [[\text{CLS}], t_1^{s_i}, t_2^{s_i}, \dots, t_{n_2}^{s_i}, [\text{SEP}]] \quad (7)$$

$$x_{s_j} = [[\text{CLS}], t_1^{s_j}, t_2^{s_j}, \dots, t_{n_3}^{s_j}, [\text{SEP}]] \quad (8)$$

其中, x_g, x_{s_i} 和 x_{s_j} 分别表示 g, s_i 和 s_j 分词后的结果, n_1, n_2 和 n_3 分别是 g, s_i 和 s_j 分词后的词语数量。使用 BERT 对 x_g, x_{s_i} 和 x_{s_j} 分别进行编码, 并选择 [CLS] 对应的向量作为检修目标 g 和检修步骤 s_i, s_j 的语义表示, 记作 e_g, e_{s_i} 和 e_{s_j} 。

将 e_g, e_{s_i} 和 e_{s_j} 拼接后作为编码器的输出 $e_s \in R^{3d}$ 。将其输入到分类器中, 得到预测的概率 p_s 。 p_s 与标记信息 l_s 通过交叉熵损失函数计算得到检修计划顺序性检测的损失 L_s 。

3.1.3 检修计划冲突检测编码器

检修计划冲突检测任务的输入是两个检修计划 $P_1 = (g_1, (s_1^1, s_2^1, \dots, s_{N_1}^1))$ 和

$P_2 = (g_2, (s_1^2, s_2^2, \dots, s_{N_2}^2))$ ，其中 N_1 和 N_2 分别是 P_1 和 P_2 所包含的检修步骤的数量。本文首先利用 BERT 对两个检修计划的检修目标和检修步骤分别编码，同样选择 [CLS] 对应的向量作为文本整体的向量表示，记作 $[e_g^1, e_{s_1}^1, e_{s_2}^1, \dots, e_{s_{N_1}}^1]$ 和 $[e_g^2, e_{s_1}^2, e_{s_2}^2, \dots, e_{s_{N_2}}^2]$ 。之后引入 LSTM 网络进一步进行编码：

$$[\hat{e}_g^2, \hat{e}_{s_1}^2, \hat{e}_{s_2}^2, \dots, \hat{e}_{s_{N_2}}^2] = \text{LSTM}([e_g^2, e_{s_1}^2, e_{s_2}^2, \dots, e_{s_{N_2}}^2]) \quad (9)$$

$$[\hat{e}_g^1, \hat{e}_{s_1}^1, \hat{e}_{s_2}^1, \dots, \hat{e}_{s_{N_1}}^1] = \text{LSTM}([e_g^1, e_{s_1}^1, e_{s_2}^1, \dots, e_{s_{N_1}}^1]) \quad (10)$$

利用 LSTM 网络的长距离依赖特性，保证每一部分的语义信息可以传达到其他部分。最后，对 LSTM 网络更新后的表示进行平均池化的操作，得到整个检修计划的向量表示：

$$e_{P_1} = \text{Pooling}([\hat{e}_g^1, \hat{e}_{s_1}^1, \hat{e}_{s_2}^1, \dots, \hat{e}_{s_{N_1}}^1]) \quad (11)$$

$$e_{P_2} = \text{Pooling}([\hat{e}_g^2, \hat{e}_{s_1}^2, \hat{e}_{s_2}^2, \dots, \hat{e}_{s_{N_2}}^2]) \quad (12)$$

其中， $e_{P_1} \in R^d$ 和 $e_{P_2} \in R^d$ 分别是检修计划 P_1 和 P_2 的向量表示，Pooling 表示平均池化操作。

将 e_{P_1} 和 e_{P_2} 拼接后输入到分类器中获取两个检修计划之间是否存在冲突的概率 p_c 。 p_c 与标记信息 l_c 通过交叉熵损失函数计算得到检修计划冲突检测的损失 L_c 。

3.2 基于LLaMA的检修计划异常检测方法

GAD-LLaMA 使用电网检修计划对 LLaMA 进行指令微调，实现对异常的识别。本文针对三种异常检测任务设计不同的提示 (prompt)，并使用 LoRA^[26] 实现参数高效的微调^[27]。表 2 是本文针对三种检修计划异常检测任务设计的提示模板，针对具体的检修计划，只需将其对应的文本内容填充到提示模板中。

表 2 GAD-LLaMA 的提示模板

异常检测任务	输入	提示模板
检修计划相关性检测	检修目标 g 检修步骤 s_i	给定检修目标和检修步骤，判断二者是否相关。[检修目标] g ；[检修步骤] s_i 。[输出]：
检修计划顺序性检测	检修目标 g 检修步骤 s_i 检修计划 s_j	给定如下检修目标和两个检修步骤，判断检修步骤顺序是否异常。[检修目标] g ；[检修步骤] s_i ；[检修计划] s_j 。[输出]：
检修计划冲突检测	检修计划 $P_1 = (g_1, (s_1^1, s_2^1, \dots, s_{N_1}^1))$ ； 检修计划 $P_2 = (g_2, (s_1^2, s_2^2, \dots, s_{N_2}^2))$	给定如下两个检修计划，判断二者是否可以同时开展。[检修计划]： P_1 ；[检修计划]： P_2 。[输出]：

4 实验研究

4.1 数据集与实验设置

4.1.1 数据集构建

电网检修计划异常检测目前没有成熟的数据集，本文收集了大量来源于电网通信领域的检修计划表用于构建具有丰富检修步骤信息的数据集。首先对现有检修计划语料进行人工标

注，识别可能存在的异常。之后对文本进行数据清洗，包括去除敏感信息、处理噪声和规范化文本格式等，确保每个样本的结构尽可能保持一致，提高数据集的质量。

我们按照 8 : 1 : 1 的比例将数据集划分为训练集、验证集和测试集，以便在模型训练和评估中进行有效监控和调整。考虑到现实中异常的情况往往是少数，我们对数据集进行了

非均衡采样，即训练集、验证集和测试集中不包含异常的样本与包含异常的样本比例为 5 : 1。

三个子任务的样本是从同一个数据源（即原始的检修计划表集合）中抽取得到的，由于每个子任务的输入上下文存在差异，例如顺序性检测的多个样本可以从同一个检修计划的若干步骤构造得到，而冲突检测的样本则只能拼接两个完整的检修计划，因此数据集的样本数量是不相同的。三个检修计划异常检测子任务数据集的统计信息如表 3 所示。数据集的规模足以微调 BERT 与 LLaMA 等预训练语言模型，不存在过拟合问题^[28]。

表 3 数据集统计信息

	训练集	验证集	测试集	无异常样本数	异常样本数
相关性检测数据集	37032	4834	4475	38618	7723
顺序性检测数据集	52175	6219	6108	53752	10750
冲突检测数据集	4885	610	610	5087	1018

4.1.2 评价标准

本文使用准确率 P、召回率 R 和 F1 值作为评价指标。评价指标的计算方法如公式（13）所示：

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F1 = \frac{2 \times P \times R}{P + R} \quad (13)$$

其中 TP 代表存在异常的样本被判定为存在异常，FP 表示不存在异常的样本被判定为存在异常，FN 表示存在异常的样本被判定为不存在异常。

4.1.3 实现细节

本文所有实验均在公共服务器上进行，所用显卡为 NVIDIA RTX3090，CPU 为 Intel Xeon Gold 6326，内存 512GB，操作系统为 Ubuntu

18.04.3 LTS。使用的编程语言为 Python3.6，模型主要基于 PyTorch 1.10.2 和 Transformers 4.11.3 实现。预训练语言模型 BERT 选择 BERT-base-chinese，LLaMA 选择 LLaMA-2-7B。对于 GAD-BERT，其中 BERT 的学习率范围是 {1e-5, 3e-5, 5e-5, 1e-4}，LSTM 和分类器的学习率范围是 {1e-4, 3e-4, 5e-4, 1e-3}，分类器中间表示的维度范围是 {32, 64, 128, 256, 512}。对于 GAD-LLaMA，指令微调通过调用 Transformers 的 Trainer 类实现，LoRA 学习率范围是 {1e-4, 2e-4, 3e-4, 4e-4, 5e-4}。

4.2 对比方法

本文选择的对比方法包括 Rule-Based、TextRNN^[29] 和 LSTM。Rule-Based 是异常检测任务的传统方法，利用正则表达式抽取检修步骤的关键词，判断是否命中预设的规则模板来进行检测。然而，Rule-Based 需要领域专家针对数据进行规则的设计，成本比较昂贵，并且难以处理特殊情况。TextRNN 是一种基于 RNN 的模型，能够有效地处理变长序列数据，捕捉文本中的上下文信息，但是容易出现梯度消失或梯度爆炸问题，限制了其对长距离依赖关系的建模能力。LSTM 旨在解决传统 RNN 中的长距离依赖问题，通过引入门控机制更好地捕捉和保持长期记忆，能够有效地处理序列数据中的长距离依赖关系，并缓解梯度消失和梯度爆炸的问题，使得模型在学习长文本时更为有效。然而，TextRNN 和 LSTM 等传统序列模型面对具体的电网领域文本时，仍面临语义理解有限的问题。本文提出的电网检修计划异常检测方法，利用了语言模型在预训练阶段获取的语义

理解能力与常识,可以有效避免由于电网领域训练语料较少导致的模型分类性能较差。

对于 Rule-Based 的方法,我们分别针对 3 项电网检修计划的异常检测任务,总结相应的响应规则,设计正则表达式进行判断。对于 TextRNN 和 LSTM,本文选择 Glove^[30]作为预训练的词向量表示,将异常检测任务对应文本的词向量输入到 TextRNN 或 LSTM 神经网络中

进行分类。

4.3 实验结果与分析

本节主要对比不同方法在电网检修计划异常检测任务上的表现。结果如表 4 所示,本文提出的方法(GAD-BERT 和 GAD-LLaMA)明显优于对比方法(Rule-Based、TextRNN 和 LSTM),其中 GAD-LLaMA 取得了最好的效果。

表 4 检修计划异常检测的实验结果

	相关性预测			顺序性预测			冲突检测		
	准确率	召回率	F1 值	准确率	召回率	F1 值	准确率	召回率	F1 值
Rule-Based	0.782	0.779	0.781	0.747	0.681	0.712	0.655	0.621	0.637
TextRNN	0.878	0.801	0.838	0.841	0.711	0.771	0.773	0.689	0.728
LSTM	0.901	0.826	0.862	0.863	0.764	0.810	0.814	0.711	0.759
GAD-BERT	0.939	0.841	0.887	0.892	0.806	0.847	0.856	0.743	0.795
GAD-LLaMA	0.952	0.856	0.901	0.901	0.811	0.854	0.861	0.748	0.801

通过观察不同方法在三个电网检修计划异常检测子任务上的表现,有以下结论:(1) Rule-Based 准确率最低。考虑到规则是硬性定义的,输入文本与规则略有差异或存在噪音,就会导致样本无法命中规则;(2) BERT 通过双向上下文建模能够更好地理解检修计划中的语义信息,而 TextRNN 和 LSTM 在处理电网检修计划这类长文本时,其对上下文的理解不及 BERT;(3) GAD-BERT 和 GAD-LLaMA 利用预训练语言模型的通用知识,在有限的标注数据上更好地泛化,而 TextRNN 和 LSTM 受限于数据规模和标注质量;(4) 注意力机制使本文提出的方法在处理变长序列方面更为灵活,能够更好地捕捉文本中的关键信息,而 TextRNN 和 LSTM 存在信息损失;(5) GAD-LLaMA 相比 GAD-BERT 有着明显的提升,这说明预训练语言模型的参数量与其在下游任务上的表现呈正相关。

综上所述,在电网检修计划异常检测任务中,基于预训练语言模型的方法更具优越性。

4.4 模型参数实验

本文设计了两个参数实验,分别探究 GAD-BERT 分类器的中间表示的维度对异常检测效果的影响,以及在检修计划冲突检测任务中,GAD-BERT 中使用的 LSTM 网络对冲突检测效果的影响,其结果如图 3 所示。

由图 3(1)可知,分类器中间表示的维度整体与异常检测模型的性能呈正相关。对于相关性检测和冲突检测任务,中间表示的维度选择 256 时表现最好,而顺序性检测在中间表示维度为 128 时表现最好。由图 3(2)可知,添加 LSTM 模块的 GAD-BERT 在三个评价标准上都取得了更好的效果,说明 LSTM 可以将语义信息传播到整个检修计划各个部分中,对于充分

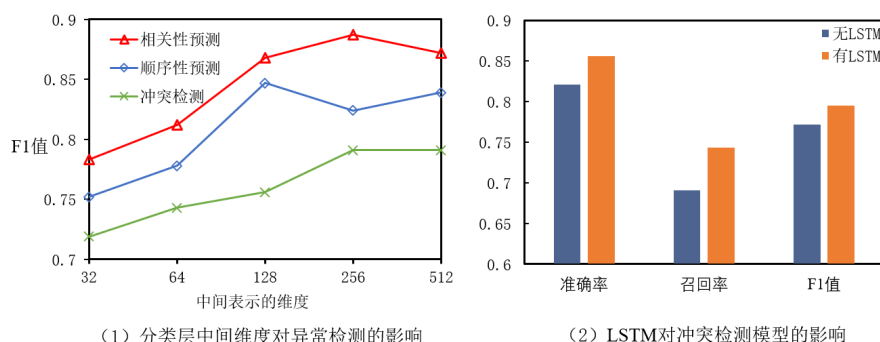


图3 模型参数实验

编码整个检修计划具有重要的作用。

4.5 错误分析

本节对 GAD-LLaMA 分类错误的情况进行分析。针对 3 种电网检修计划的异常，分别挑选一个分类错误的测试样例进行分析。对于相关性预测任务，GAD-LLaMA 错误地将发生在同一条线路（读车 5249 线）两处不同位置（002# 塔和 001# 塔）的检修计划识别为同一个检修计划。对于顺序性检测任务，GAD-LLaMA 错误地认为“新旧 OPGW 光缆熔接”应当在“拆除

临时 ADSS 光缆”之前实施。对于冲突检测任务，GAD-LLaMA 错误地认为发生在同一条线路（晋家 5269 线 / 晋港 5270 线）两处不同位置（139#-140# 档、152#-164# 段）的检修计划不能同时开展。通过分析得知，LLaMA 具备区分不同线路与不同光缆的能力，但是缺乏同一条线路不同路段的检修计划相互独立、不同光缆之间存在检修顺序的常识。本文认为上述分类错误是 LLaMA 缺乏电网检修相关的领域知识造成的，在后续工作中可尝试在电网检修计划语料上进行预训练缓解该问题。

表 5 GAD-LLaMA 的错误样例

检修任务	检修计划实例	GAD-LLaMA 预测	实际结果
检修计划相关性检测	[检修目标] 读车 5249 线木湊变 002# 塔 OPGW 光缆恢复搭接 [检修步骤] 读车 5249 线 001# 塔迁改	检修目标与检修步骤相关	检修目标与检修步骤无关
检修计划顺序性检测	[检修目标] 陆桥加强 500kV 线路工程 500kV 斗常 5267 线升高改造 [检修步骤 1] 新旧 OPGW 光缆熔接 [检修步骤 2] 拆除临时 ADSS 光缆	顺序无异常	顺序异常
检修计划冲突检测	[检修计划 1] 晋家 5269 线 / 晋港 5270 线在 139#-140# 档开展相关线路迁改工作（1）晋家 5269 线在 134# 塔和 144# 塔处断开 OPGW 光缆（2）在 134# 塔和 144# 塔处将原光缆和 ADSS 临时光缆熔接 [检修计划 2] 晋家 5269 线 / 晋港 5270 线 152#-164# 段塔基、线路开展线路迁改工作（1）在晋家 5269 线 152# 塔~164# 塔处架设 2 根 72 芯 OPGW 光缆（2）开断晋家 5269 线 151# 塔、165# 塔临时过渡 ADSS 光缆（3）搭接晋家 5269 线 151# 塔和 165# 塔处 OPGW 光缆（4）搭接 500kV 晋家 5269 线 36 芯 OPGW 光缆	存在冲突	无冲突

5 结论

本文提出了基于预训练语言模型的检修计划冲突检测方法 GAD-BERT 和 GAD-LLaMA，

旨在发现检修计划中存在的异常。实验证明，本文的方法成功地提高了检修计划异常检测的准确性，在电网通信领域的场景下表现出色，并具有较强的泛化能力和实际应用潜力。相较于传统方

法, 基于预训练语言模型的方法显著提高了计划冲突检测的效率和准确性, 为推动电网通信领域的自动化和智能化发展迈出了重要的一步。

参考文献

- [1] ZHANG L, LYU Q, CALLISON-BURCH C. Reasoning about goals, steps, and temporal ordering with WikiHow[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4630-4639.
- [2] RISH I. An empirical study of the naive Bayes classifier[C]//IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, 3(22): 41-46.
- [3] HEARST M A, DUMAIS S T, OSUNA E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.
- [4] SONG Y, LU Y. Decision tree methods: applications for classification and prediction[J]. Shanghai Archives of Psychiatry, 2015, 27(2): 130-135.
- [5] GU J, WANG Z, KUEN J, et al. Recent advances in convolutional neural networks[J]. Pattern recognition, 2018, 77: 354-377.
- [6] MEDSKER L R, JAIN L C. Recurrent neural networks[J]. Design and Applications, 2001, 5(64-67): 2.
- [7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
- [8] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.
- [9] WANG L, HAN M, Li X, et al. Review of classification methods on unbalanced data sets[J]. IEEE Access, 2021, 9: 64606-64628.
- [10] SUN Z, SONG Q, ZHU X, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48(5): 1623-1637.
- [11] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: An overview[J]. IEEE signal processing magazine, 2018, 35(1): 53-65.
- [12] WEISS K, KHOSHGOFTAAR T M, WANG D D. A survey of transfer learning[J]. Journal of Big data, 2016, 3: 1-40.
- [13] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
- [14] 缪相林, 孙超, 李彦, 等. 电网检修计划设计的智能分析与可视化实现[J]. 西安交通大学学报, 2005, 39(6): 582-585.
- [15] 程丽, 缪相林, 张培海, 等. 基于专家系统编排电网设备检修计划的求解策略[J]. 河北工业大学学报, 2006, 35(3): 63-67.
- [16] 丁毓峰, 胡业发, 盛步云, 等. 基于规则和事例混合推理的冲突解决技术研究[J]. 机械科学与技术, 2005, 24(3): 256-260.
- [17] 施荣华, 莫锐, 赵文涛. 一种基于冲突检测的无关联规则集匹配算法[J]. 计算机工程与科学, 2010, 32(10): 1-4.
- [18] 刘洁群. 基于谓词推理的冲突检测算法[J]. 吉林大学学报(理学版), 2016, 54(1): 112-115.
- [19] 孙玉芹, 肖静婷, 王海超. 基于多模型融合的电价运检命名实体识别[J]. 科学技术与工程, 2024, 23(36): 15545-15552.
- [20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017.
- [21] 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.
- [22] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.
- [23] 田园, 原野, 刘海斌, 等. 基于 BERT 预训练语言模型的电网设备缺陷文本分类[J]. 南京理工大学学报, 2020, 44(4): 446-453.
- [24] 张世奇, 马进, 周夏冰, 等. 基于预训练语言模型的商品属性抽取[J]. 中文信息学报, 2022, 36(1): 56-64.
- [25] 刘海顺, 王雷, 孙媛媛, 等. 基于预训练语言模型的案件要素识别方法[J]. 中文信息学报, 2021, 35(11): 91-100.
- [26] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.
- [27] DING N, QIN Y, YANG G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5(3): 220-235.
- [28] DOR L E, HALFON A, GER A, et al. Active learning for BERT: an empirical study[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7949-7962.
- [29] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[J]. arXiv preprint arXiv:1605.05101, 2016.
- [30] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.



开放科学
(资源服务)
标识码
(OSID)

考虑数据维度与质量维度二重性的数据质量评估研究

刘枏¹ 刘润¹ 李卓城²

1. 重庆交通大学经济与管理学院 重庆 400074;
2. 北京邮电大学玛丽女王海南学院 北京 572400

摘要: [目的/意义] 针对数据开发中关联性差、完整性不足所引发的高成本问题, 本文构建了考虑数据维度与质量维度的数据质量评估模型, 旨在评估数据的各维度质量并进行价值分析。[方法/过程] 首先通过数据质量相关的文献分析, 利用修正德尔菲法构建了数据质量维度指标体系; 之后使用层次分析法选取指标, 采用蒙特卡洛采样计算 Shapley 值进而构建了针对数据不同维度的质量评估模型。在房地产数据质量评估案例中, 对其 12 个数据维度运用质量评估模型进行评估, 结果显示其中三个维度质量较差, 建议删除, 同时针对其他维度提出了修改意见。[局限] 未解决数据质量维度指标难以量化问题, 也未对数据进行分类, 仅提出笼统的评估模型。[结果/结论] 将数据维度和数据质量维度进行融合, 并基于数据质量维度指标体系及 Shapley 构建了一个较为全面的数据质量评估模型, 能更为详细地对数据库进行质量评估。

关键词: 质量评估模型; 数据维度; 数据质量维度

中图分类号: G35; F49; TP391.1

Research on Data Quality Evaluation Considering the Duality of Data Dimension and Quality Dimension

LIU Nan¹ LIU Run¹ LI Zhuocheng²

1. School of Economics and Management, Chongqing Jiaotong University, Chongqing 400074, China;
2. Queen Mary Hainan College, Beijing University of Posts and Telecommunications, Beijing 572400, China

Abstract: [Objective/Significance] This article constructs a dimensional evaluation model based on Shapley value to address the high cost issues caused by poor relevance and insufficient integrity in data development. The purpose is to evaluate the quality of

基金项目 教育部基金项目“大数据资产的定价机理、方法和规范研究”(20XJAZH007); 重庆市社会科学规划基金项目“西部陆海新通道跨境数据互联互通机制研究”(2023ZDLH11)。

作者简介 刘枏(1965-), 博士, 教授, 主要研究方向为数据分析、工程管理信息化, E-mail: liunan@cqjtu.edu.cn; 刘润(2001-), 硕士研究生, 主要研究方向为数据分析; 李卓城(2004-), 本科生, 主要研究方向为数据挖掘。

引用格式 刘枏, 刘润, 李卓城. 考虑数据维度与质量维度二重性的数据质量评估研究[J]. 情报工程, 2024, 10(4): 25-35.

each dimension of data and conduct value analysis. [Methods/Processes] The research first analyzes the literature related to data quality, and uses the modified Delphi method to construct a data quality evaluation index system. Then, the analytic hierarchy process is used to select indicators, and the Monte Carlo sampling method is used to calculate the Shapley value and obtain the dimensional evaluation results of data. In the case study of real estate data quality evaluation, the quality evaluation model is applied to evaluate 12 data dimensions. The results show that three dimensions have poor quality and are recommended for deletion. At the same time, suggestions for modification are proposed for other dimensions. [Limitations] This article does not address the difficulty of quantifying data quality dimensions, nor does it classify data, only proposing a general evaluation model. [Results/Conclusions] The research integrates data dimensions and data quality dimensions, and constructs a relatively complete data quality evaluation model based on the data quality evaluation index system and Shapley value. It can conduct more detailed quality evaluation of the database.

Keywords: Quality Evaluation Model; Data Dimension; Data Quality Dimension

引言

随着人工智能、机械学习、深度学习等技术的快速发展,各种数据共享、交换、重复利用活动被广泛应用^[1],数据已成为驱动经济发展的重要资源和新型生产要素^[2];在大数据时代的背景下,充分发掘潜在数据价值,提高数据利用效率的需求愈发迫切。然而,关联数据质量低、数据结构差、完整性不足等数据质量问题导致的开发成本较高影响着数据资源的可持续发展与利用^[3]。

数据价值开发的基础主题之一是数据质量评估^[4],相关研究也得到了国内外相关学者的广泛关注。刘桢等^[5]、张小伟等^[6]从数据质量效用的角度出发,分别基于 Stackelberg 模型及 shapley 值构建了新的数据交易定价模型。Taleb 等^[7-8]创建了一个涵盖质量纲要、质量需求、质量维度及其计分规则的大数据质量框架,并将其用于大数据全生命周期的质量评估工作。Woodall 等^[9]研究了可以动态配置的质量评估模型,并提出了针对不同数据集特征将评价指

标进行移除和关键性排序的评估方法。Martín 等^[10]研究了物联网数据流质量的定义和评估机制以及数据管理解决方案,并将评估模块集成到操作数据富集工具链中,通过智能算法提高数据质量。Buelvasj 等^[11]针对空气质量检测系统提出了一个多维模型,该模型集成了相关质量维度的研究和领域专家的意见,可以生成数据质量综合评分指数。Guerra-garcía 等^[12]提出了在信息系统开发项目中数据质量的标准化方法,并将数据质量需求与软件数据质量需求关联。Mirzaie 等^[13]将数据流质量控制领域的研究进行了总结,将相关的质量研究方法分为质量评估、质量控制方法、质量控制技术、质量应用四个维度,并对相关方法进行了批判性总结。

数据集由众多样本构成,每个样本又包含多个特征,这些特征各自代表着数据的一个维度。上述研究综合考虑了数据质量和数据容量对于数据价值的影响,但缺乏对多维数据下单一维度数据的数据质量对数据库价值影响的思考。在多元化应用背景下,数据各维度具有差

异性价值。考虑维度的数据质量评估工作能够识别各维度在业务中的重要程度，从而明确数据质量关注重点；评估结果也能为数据的存储和管理方式提供针对性的优化建议，进而提升数据利用效率。

1 研究概述

数据库不同维度数据之间的质量问题交互性强，单一维度的质量问题常常在多维度数据的组合中显现，因此必须选择合适的方法对单维度数据质量进行评估。本研究试图通过对过往文献中数据质量维度的概念及计算规则进行总结，构建一个较为齐备的数据质量维度指标

体系。在此基础上，针对不同类型的数据及其特性，使用专家评议法从评估体系选择适宜的评估指标和评分方法，并运用层次分析法对质量维度指标进行权重划分。针对不适用于单一维度的评分方法，通过对数据各种维度之间的所有任意组合进行评分，利用 shapley 值计算出单维数据的评分，并将其作为质量评估的依据（研究框架见图 1）。研究的主要贡献在于提出了一个综合考虑数据维度和质量维度的质量评估体系，并借助 shapley 值的方式对单维数据的质量进行判断，进而帮助研究人员更好地评估数据质量及价值，为数据的有效应用提供支持。

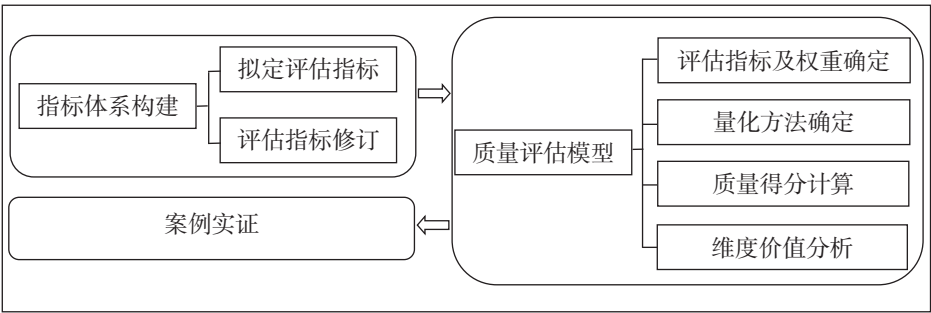


图 1 研究框架图

2 数据质量维度指标体系

数据质量是数据的固有特性，被定义为“在特定条件下满足使用者需求的能力”^[8]。是衡量数据适合使用的指标。而数据质量维度用于衡量、量化和管理数据质量，每个质量维度代表数据质量的一个属性。为了方便管理，我们将质量维度指标分为四类，分别是固有质量、情景质量、表征质量、可达质量。其中，固有质量是指数据本身所固有的、与

使用条件和使用环境无关的质量；情景质量用于衡量数据使用条件和使用环境的匹配程度；表征质量是衡量数据被有效表示的难易程度；可达质量用于衡量获取数据难易程度。本文选取 30 篇有关数据价值影响因素研究的标准与文献，并从中对数据质量维度指标进行提炼。使用修正德尔菲法征询专家意见，提出数据质量维度指标。

拟定数据质量维度指标并对其进行提取分析，构建数据质量维度指标体系（见表 1）。

表 1 数据质量维度指标体系

质量维度	指标说明	指标来源
固 有 质 量	数据容量	数据集中单位数据的数量
	数据完整性	数据集中各单位数据的完整程度
	数据一致性	数据集中各单位数据无矛盾且与其他数据保持一致的程度
	数据准确性	数据集所记录内容的正确性和精确度
	数据规范性	数据集符合数据标准、业务规则或权威参考数据的程度
	数据真实性	数据集反映其蕴含属性和相关实例的真实性和可靠程度
	数据全面性	数据集反映其预期属性和相关实例的完备程度
情 景 质 量	数据相关性	数据集与业务任务的关联匹配程度
	数据及时性	数据集实时更新及传输同步的效率
	数据时效性	数据集与当前状态的关联匹配程度
	重复使用率	数据集在设置、替换、传输等重复使用中保持现有质量的程度
	独特性	数据集在特定背景下的独特价值或独特性质
表 征 质 量	可理解性	使用者读取和解释数据集所蕴含属性的难易程度
	可机读性	数据集被计算机程序自动读取、解析和处理的难易程度
	数据代表性	数据集对其预期属性和相关实例的代表程度
	易操作性	对数据集进行替换、移动、分析等操作并保持其现有质量的难易程度
可 达 质 量	可获取性	数据集查找、检索和获取的难易程度
	非歧视性	针对他人访问、获取、使用 and 分享数据集的潜在约束程度
	数据安全性	保护数据集不被窃取、盗用或破坏的难易程度

其次，使用修正德尔菲法对评估指标进行修正。共发放了《“数据质量维度指标体系”专家咨询问卷》35份，回收有效问卷20份，问卷的回收率为57.1%，问卷的有效率为100%。研究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，分数越高，代表被试者认为该项指标对数据质量影响越大。对回收的问卷进行数据处理和分析，计算每个指标专家意见的平均值、标准差和专家意见的Kendall协调系数。平均值用于衡量指标的取舍，标准差和Kendall协调系数用于衡量专家意见的一致性。平均值低于1.5分的，说明专家意见处于影响程度较小，可以将其从指标体系中剔除。数据处理并分析后发现：“数据代表

性”和“非歧视性”平均分均为1.38902，说明专家认为数据代表性和非歧视性对数据质量影响较小。专家对所有指标评价的标准差均小于1，这意味着专家的意见基本趋于一致和稳定。Kendall协调系数检验呈现出显著性（ $P \approx 0.000 < 0.01$ ），意味着20位专家的评价具有关联性，即说明评价具有一致性。同时Kendall协调系数为0.875，大于0.5，进一步说明评价一致性较强，因此不再进行第二轮的专家调查。

3 质量评估模型

质量评估是根据确定的质量评估对象、质量范围、测量及其实现方法实现质量评测的活

动过程^[30]。基于此,本文的质量评估模型由评估指标选定与权重确定、评分方法确定、质量指标得分计算、维度价值分析四部分组成。

本节所用的符号描述如表2所示:

表2 符号描述

符号	描述
Z	需进行质量评估的数据集
V_n	数据质量指标
D_i	数据集 Z 的第 i 维度所组成的数据集
M_n	可直接计算的 V_n 评分方法
S	评估模型所需的由多个 D_i 所组成的数据集
SV_i^n	由 Shapley 值计算的指标得分
L_n	需借助 Shapley 值计算的评分方法
SC_i^n	需借助 Shapley 值计算的指标得分

3.1 数据预处理

鉴于半结构化数据及非结构化数据质量评估工作所固有的复杂性,在启动评估工作之前,应根据具体的数据集和任务需求对数据进行定制化的处理,以便于为后续的数据评估工作奠定坚实的基础。

数据预处理的两个重点是样本选取及维度提取。由于目前数据库容量大、格式多样,对数据整体进行评估工作可能存在困难,因此可使用随机抽样、等距抽样、分层抽样等抽样方法从数据源中选择少量样本 Z 来表示整个数据,但应保证样本与预评估数据库水平保持一致;针对数据维度特征不明显的的数据,应根据数据应用场景及应用需求通过文本分析、图像分析、音频分析等技术结合数据特征对数据进行维度提取。质量评估模型需要基于维度将样本数据 Z 垂直拆分为多个 D_i ,针对半结构化及非结构化数据,还需综合考虑数据特性和技术实现的

复杂性设计合理的拆分策略与方法。

3.2 数据质量评估指标选定与权重确定

评估模型采用专家评议法确定 n 个数据质量指标 V_n ($1 \leq n \leq n'$)。在评标过程中,评估小组应预先收集确定数据特征、质量需求、技术方案及技术成本等评审内容,并对质量评估指标进行共同的分项分析、比较,然后进行评议,最终确定指标。

层次分析法(AHP)是用于指标权重确定的常用方法。质量评估模型中,在评估小组内部发放调查问卷,通过成员对评估指标两两间重要程度的判断确定评估指标权重,最后求取权重分配的平均值作为最终权重。

质量评估指标是数据质量评价的关键,它反映了数据的多方面属性。在确定评估指标时要将组织战略、数据需求及业务需求和数据质量指标关联起来,明确数据质量与数据需求之间的关系。要根据具体业务需求选择适当的数据质量维度评价指标并进行合适的权重分配。此外,在数据全生命周期中,同一质量维度在不同时间维度上的含义和内容可能有所不同,需要根据评价阶段和使用范畴来确定需要评定的质量评估指标。

3.3 指标评分方法确定

在确定指标评分方法时,数据质量指标量化方法与数据业务需求的关系亟需进一步明确。指标评分方法应具有一定的前瞻性,即结合已经确定或预估的业务发展方向,将未来的数据预期融入测量框架中。对于备选的评分方法,还应考虑成本及可用性,主要包括是否能

够量化、技术上能否实现、初步的资源消耗评估等。

3.4 基于shapley值的质量指标得分计算

3.4.1 shapley值的定义

shapley 值法是由 shapley 提出的为解决多人合作对策及合作受益分配的一种数学方法^[31]。shapley 值的优势是按照成员对最终结果的边际贡献率将收益进行分配,即成员 i 所分得的收益等于其所参加合作所创造的边际利益的平均值。将 shapley 值法应用于质量评估模型之中可描述为:对于需进行质量评估的 i 维数据库 Z 的第 i 维数据单独组成的数据库 D_i ,若 V_n 的评分方法 L_n 不能计算只包含单一维度的 D_i ,则令 S 代表数据库 Z 除 D_i 外其他维度数据集组合的所有可能的任意子集 ($S \subseteq \{D_1, \dots, D_i\} \setminus D_i$),再通过计算所有 S 、 $S \cup \{D_i\}$ 的 V_n 指标得分 $L_n(S)$ 、 $L_n(S \cup \{D_i\})$ 及其两者差值 $L_n(S \cup \{D_i\}) - L_n(S)$,进而得到 D_i 的边际贡献即 D_i 加入时各 S 指标得分的增量,并经过加权平均最终得到 D_i 的 V_n 指标得分 SV_i^n ,如式(1)所示。

$$SV_i^n = \frac{1}{i'} \sum_{s \subseteq \{D_1, \dots, D_i\} \setminus D_i} \frac{L_n(S \cup \{D_i\}) - L_n(S)}{\binom{i'-1}{|s|}} \quad (1)$$

3.4.2 质量指标得分计算

对于需进行质量评估的 i 维数据库 Z 及其第 i 维数据单独组成的数据库 D_i ($1 \leq n \leq n'$),若 V_n 的评分方法可直接对 D_i 进行单独评估,则以 M_n 代指该评分方法。即 M_n 可直接对 D_i 进行评估并得到 D_i 的 V_n 指标得分 SC_i^n 。若 V_n 的评分方法不能计算只包含单一维度的 D_i ,则以 L_n 代指该评分方法。即通过计算 D_i 对所有 S

的边际贡献 ($L_n(S \cup \{D_i\}) - L_n(S)$),再通过加权平均得到 D_i 的 V_n 指标得分 SV_i^n 。

如式(1)所示,若要直接计算所有 D_i 的 shapley 值,则需计算出所有可能的边际贡献,这随着数据的增长呈指数级增长。因此我们采用蒙特卡洛采样^[32]的方法来计算近似的 shapley 值。通过对不同维度的数据集的随机排列进行采样,然后扫描所有排列并计算所有数据集的边缘得分贡献,如图2所示。

算法框架:质量得分计算	
输入:	数据库 Z 、 D_i ($1 \leq i \leq i'$), 评分方法 M_n 、 L_n ($1 \leq n \leq n'$)
输出:	各维度指标得分 SC_i^n 、 SV_i^n
1	if $M_n(D_i)$ is ValueError then
2	Initialize $SV_i^n = 0$, $t = 0$;
3	While convergence criteria not met do
4	Let π^k be a random permutation of $\{1, \dots, i'\}$;
5	$t = t + 1$;
6	for $j = 1$ to i' do
7	$SV_{\pi^{(j)}}^n = L_n(D_{\pi^{(j)}(1)}, \dots, D_{\pi^{(j)}(i')}) - L_n(D_{\pi^{(j)}(1)}, \dots, D_{\pi^{(j)}(i'-1)})$
8	$SV_{\pi^{(j)}}^n = \frac{t-1}{t} SV_{\pi^{(j)}(t-1)}^n + \frac{1}{t} SV_{\pi^{(j)}(t)}^n$
9	end for
10	end while
11	else $SC_i^n = M_n(D_i)$
12	end if
13	return SV_i^n , SC_i^n

图2 算法框架

3.5 维度价值分析

根据质量得分可以衡量各 D_i 的质量水平并支持决策的制定和调整,对于低质量的 D_i ,若该维度存在使用必要性,则对 D_i 进行清洗和修正使其满足使用要求;若 D_i 为非必要维度或数据处理成本较高,则可以考虑舍弃该维度数据。同时,在数据生命周期较长的情况下,则质量得分可以作为依据辅助数据业务人员判断哪些

D_i 及质量维度需要重点关注，以便于更好地维护数据库，满足相关业务需求。

4 案例实证

本文选取二手房数据集验证本文所构建的数据资产维度化评估体系的适用性。案例背景为某房产中介平台拟开展企业数据库构建工作，现已初步收集了部分数据，但发现存在维度过多、部分维度质量较差等问题。为方便后续的数据集收集整理及数据维度管理，需要对已收集的部分数据进行维度化质量评估并给出质量管理意见。

4.1 质量评估需求及数据预处理

拟建的房产交易数据库旨在支持企业发展决策、房产定价、房产信息优化、数据交易等活动。但线上收集的数据在某些维度上可能存在较为严重质量问题。同时，企业也需要通过评估了解不同维度的关键程度以及它们的质量问题，以便于优化成本并更好地完成数据库构建工作。

案例数据为结构化数据，包含户型、建筑面积、朝向、楼层、装修、建筑年代、电梯、产权性质、住宅类别、建筑结构、建筑类别、学校、单价、成交周期、浏览量等十六个维度，无需进行维度提取及标准化处理工作。对于户型、装修档次、朝向等非数值维度，采用赋值法来实现指标的数值化。维度的量化情况见表3。目前已收集 8942 条数据，由于数据为结构化数据且各个体可视为等概率分布，为方便质量评估工作，采用简单随机抽样，随机抽取 300 条数据作为分析样本。

表3 数据维度表

序号	指标	量化方法
1	房屋户型	一室一厅记为 1，两室一厅记为 2，两室两厅记为 3 三室一厅记为 4，三室两厅记为 5，四室两厅记为 6
2	建筑面积	直接使用该套房产的建筑面积（单位：平方米）
3	朝向	房屋朝向为南、西南、东南均记为 1，其他朝向记为 0
4	楼层	低楼层、中楼层、高楼层分别记为 1、2、3
5	装修档次	毛坯、简装、中档、高档、豪华装修分别记为 1、2、3、4、5
6	建筑年代	直接使用该套房产所处楼栋的建成年代（单位：年）
7	电梯	有电梯记为 1，无电梯记为 2
8	产权性质	个人产权记为 1，商品房记为 2
9	住宅类别	普通住宅记为 1，非普通住宅记为 2
10	学校	非学区房记为 1，学区房记为 2
11	单价	房屋售出时每平方米单价（单位：元）
12	总价	房屋售出时交易总价（单位：元）

4.2 评估指标及权重选定

经数据质量评估专家与数据业务相关人员讨论与投票，在充分考虑数据库潜在应用价值的同时，结合房地产行业相关特点，围绕数据处理工作中的实际情况，根据现有评价能力和评价手段，从数据质量维度指标体系中选取并确定数据完整性、数据一致性、数据准确性、数据独特性四个数据质量评价指标。向专家发放《“二手房数据质量评估指标体系”AHP 专家咨询问卷》10 份，以确定一般情况下数据质量评估指标权重。根据专家反馈意见分别将专家对于评估指标权重的分配进行计算并进行一致性检验，最后求取权重分配的平均值作为数据质量评估指标的权重参考。

表 4 维度指标表

类别	质量维度	指标权重
固有质量	数据完整性	0.306860859
固有质量	数据一致性	0.251900705
固有质量	数据准确性	0.300769442
情景质量	数据独特性	0.140468993

其中，数据完整性和数据独特性可通过直接对数据库各维度进行单独评估得到不同维度的指标得分，数据一致性及数据准确性的评分方法只针对数据集而不能只计算单一维度，因此需要通过算法对各维度之间任意组合的所有数据集进行评估，并利用 shapley 值得到各维度的指标得分。各指标度量方法如下。

数据完整性：完整性应衡量各维数据被完整赋值没有空缺的程度。

$N = \text{单维数据被赋值元素个数} / \text{单维数据中元素个数}$ （得分范围：0~1 分）

数据独特性：数据集在特定背景下的独特价值或独特性质。

$M = \text{由评估专家对不同维度进行主观评分}$ （得分范围：0~1 分）

数据一致性：一致性应衡量数据集根据相关规则要求保持一致的程度。

$U = \text{各数据集 Kendall } W \text{ 协调系数得分}$ （得分范围：0~1 分）

数据准确性：准确性应衡量数据集真实、准确客观地反映现实情况的程度。

$A = \text{数据集中异常值个数} / \text{数据集中元素个数}$ （得分范围：0~1 分）

4.3 房地产数据质量评估

4.3.1 完整性

将全部数据导入 spss 软件，通过描述统计得到各维度缺失值及得分情况如下（见表 5）：

表 5 完整性得分表

统 计							
个 案 数	数据维度	户型	建筑面积	朝向	楼层	装修	建筑年代
	有效	8468	8942	8933	8935	6946	6861
	缺失	474	0	9	7	1996	2081
	N	0.947	1	0.99899	0.999	0.777	0.767278
	数据维度	电梯	产权性质	住宅类别	学校	总价	单价
	有效	7083	7041	6087	8942	8942	8942
	缺失	1859	1901	2855	0	0	0
	N	0.7921	0.787407739	0.68072	1	1	1

由统计表可知，户型、建筑面积、朝向、楼层、学校、总价、单价的缺失值较少，拥有较高的完整性。而装修、建筑年代、电梯、产权性质、住宅类别的缺失值较多，其中住宅类别的完整性最差。

4.3.2 数据独特性

经评估人员与数据业务人员讨论，针对房地产数据集各维度独特性难以量化的特点，选取数据特征、数据类型、数据规模作为数据独特性的评估因素，采取专家打分法对各维度指

标得分进行量化，并求取各专家打分情况的平均值得到各维度得分情况如下（见表6）：

表 6 独特性得分表

统 计						
数据维度	户型	建筑面积	朝向	楼层	装修	建筑年代
M	0.745	0.855	0.678	0.642	0.784	0.798
数据维度	电梯	产权性质	住宅类别	学校	总价	单价
M	0.722	0.678	0.633	0.542	0.874	0.912

表 7 一致性得分表

统 计						
数据维度	户型	建筑面积	朝向	楼层	装修	建筑年代
U	0.8208	0.933636364	0.65482	0.727	0.758	0.996909
数据维度	电梯	产权性质	住宅类别	学校	总价	单价
U	0.7205	0.741545455	0.73827	0.762	0.947	0.9645

4.3.4 数据准确性

本案例采用 K 最近邻算法识别数据集中的异常值，该模型是通过搜寻最近的 K 个已知类别样本对未知类别样本进行预判，基于此，我们只需依次计算每个样本点与它最近的 K 个样

4.3.3 数据一致性

本案例中数据一致性较侧重于各维度数据之间的关联性，因此选用 $Kendall W$ 度量数据集一致性。 $Kendall W$ 系数范围为 $[0,1]$ ，当 $W \leq 0.4$ 时，说明数据一致性较差；当 $0.4 < W \leq 0.6$ 时，说明数据一致性一般；当 $0.6 < W \leq 0.8$ 时，说明数据一致性较好；当 $0.8 < W \leq 1$ 时，说明数据一致性为优。选用 matlab 软件计算 $Kendall W$ 系数，将样本数据集以矩阵形式导入 matlab 软件中，并将 $Kendall W$ 系数算法嵌套入质量评估模型中得到各维度得分情况如下（见表7）：

本的平均距离。再利用计算的距离与阈值进行比较，如果大于阈值，则认为是异常点。

将样本数据集以矩阵形式导入 matlab 软件中，并将 K 最近邻算法嵌套入质量评估模型中得到各维度得分情况如下（见表8）：

表 8 准确性得分表

统 计						
数据维度	户型	建筑面积	朝向	楼层	装修	建筑年代
A	0.8924	0.9765	0.8454	0.7741	0.8324	0.8042
数据维度	电梯	产权性质	住宅类别	学校	总价	单价
A	0.7132	0.5585	0.6547	0.8433	0.9874	0.9872

4.4 评估结果分析

从数据维度层面的得分结果看，总得分在三分以下的维度是住宅类别、产权性质、电梯，分别为 2.706693、2.765453、2.947759；从质量维度层面对住宅类别、产权性质、电梯三个维

度进行分析，它们的数据完整性与数据准确性存在严重的质量问题，建议业务人员在考虑数据获取与处理成本的同时，重点关注这三个维度的数据质量以提高数据库整体质量水平；同时其独特性和一致性评分也均处于较低标准，建议相关业务人员优化这三个维度数据的结构

及处理流程。

从质量维度的得分结果来看,数据库整体的一致性、独特性得分均处于较好水平,说明数据库整体潜在应用价值及与业务需求匹配程度均较好。针对数据完整性而言,除住宅类别、产权性质、电梯外,仅楼层维度评分为 0.7741,其他维度评分均在 0.8 以上,建议业务人员考察相关数据来源及收集流程,重点提高楼层的数据准确性。针对数据完整性而言,除住宅类别、产权性质、电梯外,装修和建筑年代的数据完整性评分也较低,同时装修和建筑年代的一致性、独特性、准确性质量维度评分较高,说明其数据准确性及潜在价值较好,因此在后续的数据库构建及维护工作中应重点关注装修和建筑年代的数据缺失问题。

5 结论

本文通过使用文献调查法和修正德尔菲法建立了兼具科学性、系统性的数据质量维度指标体系;并以此为基础,利用层次分析法、shapley 值法及蒙特卡洛采样构建了一个较为完备的数据质量评估模型;并以某二手房数据集质量评估案例作为实证应用样本,进一步探讨了所构建质量评估模型的有效性和可操作性。本文的创新体现在:将数据维度和数据质量维度进行融合,尤其是从维度层面考虑了数据的质量问题,以此来保证数据的可用性,在一定程度上弥补了当前数据质量研究的不足;针对部分数据质量维度评分方法不适用于单一维度评分的情况,通过对数据各种维度之间的所有任意组合进行评分,再利用蒙特卡洛采样计算 shapley 值得出单维数据的质量评分。

虽然本文为单维数据质量评估提供了新思路,但也还存在一些不足,例如,没有解决上文所提及的部分数据质量维度指标难以量化的问题,同时实证案例中选取的是结构化数据,针对非结构化数据的维度提取及基于维度的数据库拆分等预处理工作未进行详细说明。另外本文没有对数据进行分类,所以提出的是一种笼统的数据质量评估模型,不能为不同类型的数据结构提出定制化的评估方案等,这都是下一步研究需要解决的问题。

参考文献

- [1] JIAN P. A Survey on Data Pricing: From Economics to Data Science[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(10): 4586-4608.
- [2] 中华人民共和国国务院. 中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要 [EB/OL]. (2021-03-13) [2024-01-23]. https://www.gov.cn/xinwen/2021-03/13/content_5592681.htm
- [3] 孙嘉睿, 安小米. 开放政府数据质量评估指标体系研究 [J]. 情报理论与实践, 2023, 46(6): 94-100, 78.
- [4] CABALLERO I, GUALO F, RODRIGUEZ M, et al. BR4DQ: A methodology for grouping business rules for data quality evaluation[J]. Information Systems, 2022, 109: 102058.
- [5] 刘桢, 徐程程, 陈俞宏. 基于效用理论的数据定价方法研究 [J]. 价格理论与实践, 2022(11): 164-167, 211.
- [6] 张小伟, 江东, 袁野, 等. MaSS: 基于单位数据贡献的模型定价框架 [J]. 计算机科学与探索, 2023, 17(9): 2252-2264.
- [7] TALEB I, SERHANI M A, BOUHADDIOUI C, et al. Big data quality framework: a holistic approach to continuous quality management [J]. Journal of Big Data, 2021, 8(1): 41.
- [8] TALEB I, SERHANI M A, DSSOULI R, et al. Big Data Quality: A Survey [C]// proceedings of the IEEE International Congress on Big Data (IEEE BigData) Part of the IEEE World Congress on Services, San

- Francisco, CA, F Jul 02-07, 2018. Ieee: NEW YORK, 2018: 166-173.
- [9] WOODALL P, BOREK A, PARLIKAD A K. Data quality assessment: The Hybrid Approach [J]. Inf Manage, 2013, 50(7): 369-382.
- [10] MARTÍN L, SÁNCHEZ L, LANZA J, et al. Development and evaluation of Artificial Intelligence techniques for IoT data quality assessment and curation [J]. Internet of Things, 2023, 22: 100779.
- [11] BUELVAS J H, MÚNERA D, GAVIRIA N. DQ-MAN: A tool for multi-dimensional data quality analysis in IoT-based air quality monitoring systems [J]. Internet of Things, 2023, 22: 100769.
- [12] GUERRA-GARCÍA C, NIKIFOROVA A, JIMÉNEZ S, et al. ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design [J]. Data & Knowledge Engineering, 2023, 145: 102152.
- [13] MIRZAIE M, BEHKAMAL B, ALLAHBAKHS M, et al. State of the art on quality control for data streams: A systematic literature review [J]. Computer Science Review, 2023, 48:100554.
- [14] VALENCIA-PARRA Á, PARODY L, VARELA-VACA Á J, et al. DMN4DQ: When data quality meets DMN [J]. Decision Support Systems, 2021, 141: 113450.
- [15] 全国信息技术标准化技术委员会. 信息技术数据质量评价指标: GB/T 36344-2018[S]. 北京: 中国标准出版社, 2018: 13.
- [16] VALVERDE C, MAROTTA A, PANACH J I, et al. Towards a model and methodology for evaluating data quality in software engineering experiments [J]. Information and Software Technology, 2022, 151: 107029.
- [17] GONZÁLEZ-VIDAL A, RAMALLO-GONZÁLEZ A P, SKARMETA A F. Intrinsic and extrinsic quality of data for open data repositories [J]. ICT Express, 2022, 8(3): 328-333.
- [18] WOOK M, HASBULLAH N A, ZAINUDIN N M, et al. Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling [J]. Journal of Big Data, 2021, 8(1): 49.
- [19] LIU C, NITSCHKE P, WILLIAMS S P, et al. Data quality and the Internet of Things [J]. Computing, 2019, 102(2): 573-599.
- [20] LIONO J, JAYARAMAN P P, QIN A K, et al. QDaS: Quality driven data summarisation for effective storage management in Internet of Things [J]. Journal of Parallel and Distributed Computing, 2019, 127: 196-208.
- [21] 张夏子钰, 周林兴. 大数据时代档案数据质量: 评估与优化 [J]. 北京档案, 2023(5): 15-18.
- [22] 吴小娥. 物联网数据质量评估应用 [J]. 宝鸡文理学院学报 (自然科学版), 2022, 42(4): 50-54, 62.
- [23] 赵文轩, 李春旺. 关联数据质量评价方法研究述评 [J]. 情报理论与实践, 2016, 39(2): 134-138, 128.
- [24] 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.
- [25] HARYADI A F. Requirements on and Antecedents of Big Data Quality: An Empirical Examination to Improve Big Data Quality in Financial Service Organizations [D]. Delft: Delft University of Technology, 2016
- [26] 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.
- [27] 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.
- [28] WANG R, STRONG D. Beyond Accuracy: What Data Quality Means to Data Consumers [J]. Journal of management information systems, 2016, 12(4): 5-33.
- [29] LUGMAYR A, STOCKLEBEN B, SCHEIB C. A Comprehensive Survey on Big-Data Research and Its Implications—What Is Really ‘New’ in Big Data?—It’s Cognitive Big Data! [C]//Proceedings of the 10th Pacific Australasian Conference on Information Systems Taiwan, 2016: 45.
- [30] 李华锋, 辜汝桐, 胡海青, 等. 基于熵权法的航行通告数据质量评价方法 [J]. 民航学报, 2022, 6(4): 1-5.
- [31] 吴黎军, 项海燕. 基于信息熵的n人合作博弈效益分配模型 [J]. 数学建模及其应用, 2013, 2(Z2): 50-54.
- [32] CASTRO J, GÓMEZ D, TEJADA J. Polynomial calculation of the Shapley value based on sampling[J]. Computers & Operations Research, 2009, 36(5): 1726-1730.



开放科学
(资源服务)
标识码
(OSID)

面向民用飞机装配工艺规范数字化建设的标准元数据质量测评研究

苏萍¹ 邹小筑² 安鲁陵³

1. 南京航空航天大学 人文与社会科学学院 南京 210016;

2. 南京航空航天大学 科技信息研究所 南京 210016;

3. 南京航空航天大学 机电学院 南京 210016

摘要: [目的/意义] 国内民用飞机装配工艺规范数字化建设相对国外缓慢, 实践过程中存在很多阻滞。目前鲜有从数字化需求对标准元数据质量进行测评的研究, 相关研究既能从内容层面对元数据描述提出建设性意见, 提高标准信息资源的数字化质量, 又能推进民用飞机装配工艺数字化进程, 助力智能制造。[方法/过程] 对工艺规范与标准信息资源概念进行对比, 分析民用飞机装配工艺规范数字化及标准元数据进展; 理清民用飞机装配工艺规范数字化需求, 基于调整后的 Bruce Hillmann 测评指标确定了工艺规范数字化需求下的标准元数据质量测评方案; 以国家军用标准全文服务体系为例, 对标最新版国家《标准文献元数据》, 分析元数据对标准内容的揭示程度与数字化需求之间的差距, 提出建议。[局限] 单案例研究, 典型性和全面性仍有待验证。[结果/结论] 分析民用飞机装配工艺规范数字化进展和需求发现: 数字化质量亟待提高, 面向航空工艺规范的元数据标准是学界及业界需要共同解决的问题; 通过测评数据发现: 在描述上, 元数据类型划分针对性不强, 内容描述粗粒度, 具体对象需求数据分散化, 图表数字化程度低。因此建议: 建立行业工艺规范元数据标准, 提升工艺规范数据描述精度, 以对象为中心推进工艺规范数据模块化, 协调标准界、数据库商、业界共同推进工艺规范制定与同步数字化。

关键词: 工艺规范; 标准; 数字化; 元数据质量

中图分类号: G35; V262; G254

Research on the Quality Evaluation of Standard Metadata for the Digital Construction of Civil Aircraft Assembly Process Specifications

SU Ping¹ ZOU Xiaozhu² AN Luling³

1. College of Humanities and Social Sciences, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China;

基金项目 南京航空航天大学智库创新基金(人文社科类)项目“智库在航空制造工艺规范数字化建设中的作用研究”(NK2022012)。

作者简介 苏萍(1997-), 硕士研究生, 主要研究方向为信息组织与利用; 邹小筑(1970-), 双学士, 副研究馆员, 主要研究方向为信息素养、信息组织与利用、信息资源评价与建设, E-mail: zxzlib@nuaa.edu.cn; 安鲁陵(1962-), 博士, 教授, 主要研究方向为数字化制造、工装快速设计、飞机装配技术等。

引用格式 苏萍, 邹小筑, 安鲁陵. 面向民用飞机装配工艺规范数字化建设的标准元数据质量测评研究[J]. 情报工程, 2024, 10(4): 36-46.

2. Institute of Scientific and Technological Information, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China;
3. College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract: [Objective/Significance] The digital construction of domestic civil aircraft assembly process specification is relatively slow, and there are a lot of hindrances in the process of practice. At present, the research of measuring the quality of standard metadata from the digital demand is relatively lacking. Its analysis can not only put forward constructive opinions on the description of metadata from the content level to improve the digitized quality of the standard information resources, but also promote the digitization process of civil aircraft assembly process and assist the intelligent manufacturing. [Methods/Processes] Compare the concepts of process specification and standard information resources, analyze the progress of digitization of civil aircraft assembly process specification and standard metadata; clarify the digitization demand of civil aircraft assembly process specification, on the basis of the adjusted Bruce Hillmann assessment indicators, a programme for assessing the quality of standards metadata based on the digitization needs of process specifications was identified; using the national military standard full-text service system as an assessment case, benchmarking against the latest version of the national “Standard Literature Metadata”, analyzing the gaps between the degree of metadata revelation of standard content and digitization needs, and putting forward suggestions. [Limitations] Single case study, its typicality and comprehensiveness still need to be verified. [Results/Conclusions] Analysis of the progress and demand for digitization of civil aircraft assembly process specifications found that the quality of digitization needs to be improved urgently, and metadata standards for aviation process specifications is a problem that needs to be solved jointly by the academia and the industry; through the measurement data, it is found that in the description, the division of metadata categories is not targeted, the description of the content presents coarse-grained, the data of the specific object demand is diffused and the digitization of charts and diagrams is low, etc. Therefore, it is suggested that the establishment of industry process specification metadata standards, improve the accuracy requirements of process specification data description, promote the modularization of process specification data with the object as the center, and coordinate the industry, standardization community, and database vendors to jointly promote the formulation and synchronization of process specification and synchronous digitization.

Keywords: Process Specification; Standards; Digitization; Metadata Quality

引言

随着数字全球化趋势的深入，标准数字化成为近年来国家标准化发展的重要目标。航空工业作为国防建设的重要组成部分，也是高端制造业中技术最集中的行业，其数字化建设深受关注。2020年，国家标准化委员会在航空航天领域启动了机器可读标准试点工作。航空工艺工作正经历数字化工艺工作模式变革^[1]。民用飞机装配工艺规范作为一类重要的标准，其

数字化质量直接影响产品的质量。

元数据方案是各种文献数字化质量的关键和基础，其作用是对文献外部及内部特征进行描述，从而满足用户对文献的查询和利用需求。作为一类特殊标准文献的工艺规范，民用飞机装配工艺规范的数字化需要围绕企业用户需求对其内容进行深入和准确的描述，才能最大程度发挥其价值。而工艺规范的充分利用，能生成适应性加工工艺，引导制造过程向全局优化的方向发展^[2-4]。

随着数字孪生、虚拟现实、CAD/CAM 技术的推进,国内外就飞机装配工艺数字化展开了一系列工作,涌现了大量有关工艺平台开发和仿真方面的文章。如:工艺开发中的工艺集成计算平台^[5]、工艺设计与管理平台^[6]、工艺工具管理配送平台^[7]、工艺参数控制平台^[8]、飞机加工工艺的仿真探究^[9]等。部分学者就工艺标准数字化进行了探究。如:孙升等^[10]对数字化制造工艺进行分析;国际标准化组织 ISO^[11]成立机器可读标准工作组;ISO 与 IEC 合作绘制了机器可读标准成熟度分类模型^[12];工业领域提出并实践了面向机器可读的工业通用语义知识库^[13];汪烁等^[14]论述了实现机器可读的共性技术。上述研究主要集中在数字化制造工艺内容和机器可读技术应用方面,标准数字化质量测评方面的研究很少,标准内容描述及标准元数据质量测评方面的研究更少;从实践看,航空工业数字化设计制造技术应用较早,但标准建设与工程需求存在较大差距。缺少有效的元数据方案,成为当前影响其数字化质量与进程的主要问题。

故本文从飞机装配工艺数字化需求出发,以国家军用标准文献服务系统为例,对标标准文献元数据国家标准开展测评,力求从元数据层面,为实现民用飞机装配工艺规范内容要素结构化、语义化提供建议和数字化建设思路,增强工艺规范服务飞机装配数字化的能力。

1 工艺规范与标准信息资源的关系

企业数字化建设使标准数字化这一概念在研究中普遍出现。虽部分文献对其概念做了阐

述,但标准与工艺规范概念混用、理解不一现象仍然存在。因此有必要对工艺规范和标准信息资源的关系进行梳理。

《航空工业标准化基础》中指出:工艺规范是规定产品加工环节的制造技术及制造要求的一类标准,是对工艺技术的总结、提炼和提高^[15]。而图书情报视角下的标准信息资源有广狭义之分。狭义上包括标准草案、现行标准、历史标准、作废标准;广义上还包括标准化期刊、标准化专著、技术法规等^[16]。标准信息资源(下文简称标准)有广狭之分。

工艺规范是规定正确的生产方法,使用适宜的设备和工装,使生产过程固定、稳定,生产出符合设计图纸产品的一种工程文件^[15],主要包括工艺参数和工艺条件。工艺规范具备一般规范的有效性、正确性、可操作性和法规性,但对于正确性和可操作性要求更高。从波音工艺规范体系看,工艺规范由内部标准和外部标准组成,内部标准涉及制造工艺(如密封、喷涂、热处理),而外部标准涵盖了波音几乎所有的技术和工艺领域,来自 120 多个组织,包括政府机构、行业协会、专业学会、区域联盟、国际和国家标准化组织、其他公司^[17]。

综上所述,工艺规范是一种特殊的标准,因此工艺规范的数字化研究,可基于标准信息数字化研究理论体系来开展。

2 民用飞机装配工艺规范数字化及标准元数据研究现状

2.1 民用飞机装配工艺规范数字化研究现状

国内,航空工艺规范数字化的研究,可追

溯到工艺信息标准化,部分学者在此基础上展开了数字化实践和技术应用的探索。陈嵩等^[18]以超七飞机研制为例,指出工艺信息标准化是数字化工艺设计技术的基础和保证;许增辉等^[19]分析了航空制造企业飞机精益制造工艺标准化的贯彻力度;孙升等^[10]从设计工艺协同、工艺标准化、工艺数据管理等五大方面分析了数字化制造的工艺工作;航天制造业数字化标准体系框架早在2011年就有研究^[20]。随后,陈润^[17]分析了波音数字化的工作;汪烁等^[14]基于机器可读标准的特点和能力等级模型,分析了6项实现标准机器可识别、可执行、可决策的关键共性技术。

国外的研究分为平台建设、新表达形式和机器可读试点几个方面。波音以PSDD为主体形成了标准数字化平台建设整体解决方案。2019年,SMART标准出台^[21]。立足新的标准表达形式,美国在无人机系统、商业航天工业等领域开展试点工作^[22]。此外,国外在机器可读、语义信息模型分析上取得了阶段性成果。定义了SMART标准的发展阶段及特征,提出灵活迅速的BSI Flex标准形式;分析语义模型对数字化转型的重要性,构建了可在信息物理系统直接使用的标准集成模型^[23-25]。

2.2 标准文献元数据标准及元数据质量评价指标研究现状

2008年标准化研究院起草了国家标准GB/T22373-2008《标准文献元数据》,正式从国家层面对标准元数据的数据格式进行规范。为增强适用性,2021年10月实施GB/T22273-2021《标准文献元数据》,扩充了平台建设标准实

体管理所需元数据,新增发布单位代码、前言、引言、卷期号、出版周期、标引依据、标准历史、参建单位8项公共元数据。航空标准管理方面,1989年航天工业部发布《标准文献著录规则》(QJ 1640-1989),对著录项及标识符、级次、格式、细则、来源等做出说明。1995年《航天文献编目规则标准文献编目规则》(QJ 2777.3-1995)获批,涵盖主题内容与适用范围、引用标准、术语、接收、查重、标引、著录等十部分。上述规则对航空领域标准的管理起到了重要作用,但问题也较明显。第一,大部分著录项仅限于封面、首页和版权页,内容揭示程度有限,描述深度不足。第二,我国标准制定周期平均为30个月^[26],关于周期的规定至今未更新,导致很多内容滞后。第三,主题标引对内容的反映程度受到选词规则的制约。规定中主题词一般为标题中反映标准类型的通用词,正文词汇只能视情况标引;标准中各种材料牌号、工艺加工符号、温湿度等数据均被排除在外。规则参考的词表均出版年代较早,且更新不及时,很难满足当前民用飞机装配工艺规范主题标引的需求。

国内外学者从不同角度对元数据质量评价指标提出了自己的观点。Qin等^[27]指出元数据质量评估是个多维度的问题,需要考虑资源集合、记录和元素三个层次的正确性、完整性、重复性和一致性等指标。William等^[28]、Besiki^[29]、Ochoa等^[30]、黄莺等^[31]学者在此基础上进行丰富,提出了不同的指标体系。众多指标体系中,准确性、一致性和完整性高度重合。Bruce和Hillmann指标体系操作性强、应用可推广,最受推崇。

国内对民用飞机装配工艺规范数字化的研

究相对丰富,包括工艺规范本身及工艺规范数字化的关键技术。国外则更深入,特别是在机器可读标准方面。但目前鲜有从元数据质量角度对工艺规范数字化进行的研究,而元数据方案与质量会直接影响工艺规范数字化进程。故研究结合民用飞机装配工艺数字化需求开展元数据质量测评,以期为民用飞机装配工艺规范数字化提供参考。

3 基于工艺规范数字化需求的标准元数据质量测评方案

Bruce 和 Hillmann 指出:元数据质量的高低取决于元数据满足需求和目标的程度^[32]。即需求是测评的重要参考因素,因此先明确民用飞机装配工艺规范数字化需求。

3.1 民用飞机装配工艺规范数字化需求

数字化环境下工艺规范不只局限于以规范格式编写、出版或网上传递,更需要使内容与业务流程、应用场景融合,让用户无需查阅全文便可完成对所需内容的精准查询与执行。通过笔者对项目合作企业的实地调研和观察发现,企业的主要需求如下:

工艺规范应用者的需求包括:(1)能够以零件为中心,及时、快速、准确地找到所需工艺规范中细粒度的关键信息(如:加工、装配、检验流程中的参数、示图);(2)关键信息在系统中方便调用,便于形成工作成果(如:文档、模型、程序等的一部分);(3)工艺规范升版后能及时推送,以提醒正在使用、引用的人员;(4)同一模块的相关工艺规范信息能以脉络化、条目化的方式全面展示在工作系

统中。例如,对于“高锁螺栓安装”,能够从“人”“机”“料”“法”“环”等角度,一体化输出通用要求、一般要求与特殊要求。

工艺规范管理者的需求包括:以版本为中心,建立工艺规范的信息集成,形成工艺规范重组后的、根据制定的标准范式来表达工艺规范文件,以便对其进行维护、管理和修改。重组后的工艺规范要便于向使用者、设计者展示一本完整的工艺规范。

不难看出,以工艺规范结构化实现知识的快速精确检索是其数字化首要目标和基础。要提高检准率,关键是每个工艺规范的内容描述深度,即标引深度要符合用户需求。而内容描述主要依据元数据方案,分析现有的标准元数据描述与上述需求之间的差距如何。考察现有标准数据库中民用飞机装配标准的元数据描述情况是研究的突破所在。

3.2 测评方案

3.2.1 测评目的

抽取国内典型标准数据库中民用飞机装配相关的标准文献元数据,对标 2021 版国家《标准文献元数据》(下文简称国标),分析元数据描述与要求以及现实需求之间的差距。

3.2.2 测评指标的选取和测评策略

Bruce 和 Hillmann 指标权威性强,包含准确性、一致性、完整性三个高度重合指标,故研究采用该指标。考虑到标准之间存在引用与被引用的关系,而来源和可获得性对既存的数据标准评价性不强,因此对指标进行调整,最终选取完整性、准确性、一致性、有效性、关联性及需求满足度 6 个指标。完整性评估元

数据对目标资源全面、详尽描述的程度；准确性评估元数据是否正确、客观地反映了目标资源的内容；一致性指元数据符合资源库中采用的元数据规范的程度；有效性评估元数据信息是否随变动及时更新和调整；关联性评估与其他相关资源、数据的关系；需求满足度评估元数据基于特定用户群体满足其功能需求的能力。

采用抽样分析法，遵循以下原则：（1）随机性，即数据记录相互独立，任何一条元数据信息被抽中的机率均等；（2）数据具有代表性且覆盖面较广，每个类型的数据都能够被抽到^[33]。分析方法如下：（1）标准记录包含的全部数据项对标国标核心元数据和公共元数据，验证完整性；（2）准确性与一致性都与元数据字段定义和内容描述相关。根据国标数据项的定义与值域要求，重点比对标内容描述项，检测其准确性及一致性；（3）从有效性、关联性挂钩数据项的缺失及描述情况检测其有效性及关联性；（4）将需求中与工艺规范直接相关的内容对标所有记录项，分析元数据对标准内容揭示程度与数字化需求之间的差距。

4 评测研究

4.1 测评数据库选择与抽样规模

《国家军用标准全文服务系统》收录国内航空航天民航等行业标准全文，能极大程度地满足用户对国内军用标准信息的需求，数据具有权威性、典型性和专业性。

以行业分类为标准，选择现行的航空标准，共 11901 条。数据分为航空航天、机械等 18 类，每一类样本采用分层抽样。以 95% 的置信水平

和 5% 的误差幅度为基准，抽取不少于 373 条记录的样本。因数据库元数据方案较为固定，取数据项最全的记录作为对标结果，其余数据作为案例和测评分析的依据。

4.2 测评数据分析

元数据完整性、准确性、一致性、关联性、有效性、满足度对标数据如表 1 所示。

4.2.1 完整性测评数据分析

由表 1 知，记录中核心元数据记录状态、记录识别符、确认日期、补充件 4 项缺失，均为必选项；公共元数据 45 项缺失（受字段限制数据未一一列举），包括记录状态、记录识别符、确认日期、第二标准号和补充件 5 个必选项，2021 版新增 8 项数据均缺失。可见，标准文献数据建设与国家要求存在差距，存在裁剪必选数据项和执行标准滞后等问题。

4.2.2 准确性及一致性测评数据分析

据定义和值域要求知，国标反映内容的项有：（中文、原文、英文）标准名称、（中文、英文）文摘、（中文、原文、英文）主题词、中文自由词、附注、标引依据、（中国、国际、原）分类号、前言和引言 16 项，数据库补充 6 项，相对丰富。但测评发现：除标题、分类号等纲领式内容项，摘要、主题词、前言等揭示内部信息及辅助分析检索的项均缺失。数据库方虽扩充了数据，问题却比较明显，如名称与内容描述不符。以标准“焊接工艺质量控制”为例，其适用范围描述为“本标准规定了航空焊件…要求，本标准适用…控制”，适用范围仅为内容后半段。此外，数据项“全文说明”也名不副实，实为获取全文方式的说明。

表 1 测评数据对标表

比较内容	《标准文献元数据》2021 版 国家标准	国家军用标准全文 服务系统	比较内容	《标准文献元数据》2021 版 国家标准	国家军用标准全文 服务系统
核心元数据	记录状态	×	内容描述有关项（扩充）		文档描述
	记录识别符	×			全文说明
	标准号	标准号			适用范围
	发布日期	发布日期			标准说明
	发布机构	发布单位			关键词
	标准状态	标准状态	状态有关项	记录状态	×
	实施或试行日期	实施日期		标准状态	标准状态
	确认日期	×		被代替标准	被替代标准
	中文标准名称（条件必选）	标准名称		代替标准	替代标准
	原文标准名称（条件必选）	—		修改件	×
	英文标准名称（条件必选）	英文名称		被修改件	×
	被代替标准	被替代标准		补充件	×
	修改件	×		被补充件	×
	补充件	×	关联性有关项	代替标准	替代标准
内容描述有关项	中文标准名称（条件必选）	标准名称		被代替标准	被替代标准
	原文标准名称（条件必选）	—		相关法律	×
	英文标准名称（条件必选）	英文名称		引用文件	引用标准
	中文文摘	×		一致性程度	×
	英文文摘	×		标准历史	×
	中文主题词	×	元数据对标准内容揭示程度		国家军用标准全文 服务系统
	原文主题词	—	需求满足度	有要求、指标等关键信息的数据项	×
	英文主题词	×		同一模块相关信息一体化、 条目化呈现	×
	中文自由词	×		图表内容直接检索	×
	附注	×		关键信息的调用、导出	×
	标引依据	×			
	中国标准分类号（条件必选）	标准分类号			
	国际标准分类号（条件必选）	标准分类号			
	原分类号	—			
	前言	×			
	引言	×			

注：× 代表缺失，—代表空项，带下划线的项为必选数据项。

显然，由于摘要、主题词、前言等涉及标准内容深度以及辅助判定检索的项均为可选项，大部分标准不会标注。数据库扩充行为缺乏有效的规范，元数据的准确性及一致性不够，降低了检索的效率。

4.2.3 有效性及关联性测评数据分析

据定义和值域要求知，国标中状态相关项有：记录状态、标准状态、被代替标准、代替标准、修改件、被修改件、补充件、被补充件 8 项；关联性相关项有代替标准、被代替标准、

相关法律、引用文件、一致性程度、标准历史 6 项。由表 1 可知，有效性相关项有 4 项缺失，3 项必选项；相关法律、一致性程度、标准历史关联性数据项均缺失。像“热喷涂工艺质量控制”“飞机装配工艺”等多数标准，只有标准状态和被代替标准两项状态数据，而记录标准修改、删除、新增的关键项记录状态却存在缺失。

综上知，标准信息的变更缺少记录，并未对入库、修订等进行标识，同一资源存在冗余；状态项与关联项体系庞大，与其他数据、机构组织关联，不同资源融合时数据繁杂，且涉及权益协议问题，成本高，凭借单方力量难以有效解决。

4.2.4 需求满足度测评数据分析

由需求分析知，需求中与工艺规范关键信息直接相关的有：描述要求、指标等，一体化展示同一模块相关内容，图、表内容直接检索，内容增、删、改的及时提醒，标准内容的导出和调用（见表 1），由于内容增、删、改与有效性测评重合，不再重复。

从测评知，记录中元数据项对要求、指标等关键信息缺乏提取，内容描述粗粒度，尚未实现多个标准相关内容的调用和条目化展示，同一对象的数据项独立且分散。此外，标准内图片、表格数字化程度尤其不足，图片及表格中的文字仅能在标题、段落中进行全文检索且不准确。例如，“铰孔”在 HB/Z 223.12-2018 表 1 出现，而在全文范围内检索“铰孔”一词时，并未检索到 HB/Z 223.12。最后，标准关键内容调用、导出有难度，数据库只提供邮箱获取整体原文，并不能实现部分关键信息的调用。

5 民用飞机装配工艺规范数字化建议

5.1 建立行业工艺规范元数据标准

测评结果显示，数据库商进行标准文献数据建设时，存在裁剪关键必选字段和执行国家标准滞后现象。当通用型标准不能精准或完整支持数据描述时，裁剪和扩充一定程度上可以提高应用性，但需要有统一的体系和对照规则以保障数据的准确性和一致性。测评中修订版本与修改件、适用范围和标准说明、有效区域和标准来源，以及高级检索中“标准简介”字段检索结果却在“适用范围”中命中，都表明缺少统一对应规范，将直接影响标准获取与利用效率。同时，国标中摘要、主题词、前言等带有一定内容深度的数据项多为可选项，大部分标准不会标注，导致不查看标准原文便无法了解标准的关键内容；对于工艺规范的应用者，更多需求则是工艺参数、操作要求等深入性的内容，文摘、主题词等一般内容作用十分有限。

由于国家标准元数据是各行业标准通用，对内容信息无法过多规定，而行业内制定专业性标准元数据，可以根据行业特性将元数据深入内容。因此，在民用飞机装配工艺规范数字化过程中，首先应建立行业工艺规范元数据标准。明确字段名称，建立元数据名称对应表，避免因名称不同影响工艺规范应用者检索获取有效信息的效率。其次，对于国标中的内容数据项，如摘要和主题词，以及数据库收录的唯一标识，如标准记录状态和记录识别符，可设为必选项，以提高信息的完整性和标准化程度。此外，要结合民用飞机装配的特性，提升数据

项内容深度。例如,必选项可涵盖材料、工具设备和质量控制等内容,而可选项可包括人员、环境 and 安全防护等一般要求。可增加指标项,形成对工艺参数的有效关联和一体化展示,以便工艺规范应用者不阅读全文即可获取关键信息,能更好地满足其对“人”“机”“料”“法”“环”要求一体化输出的需求。同时,对工艺规范升级、更改等信息的有效标识,可使管理者更好地对工艺规范进行维护、管理和修改,确保数据准确、一致、有效。总之,行业元数据标准修订是基础,只有行业元数据标准修订到位,才能为工艺规范全文数字化、机器可读、检索使用等提供有力支撑。

5.2 提升数据描述精度

由前文分析知,尽管内容相关数据项相对丰富,但在满足工艺规范应用者需求上仍存在较大差距。为充分发挥数据项的作用,有效满足工艺规范应用者及时、快速、准确地找到所需工艺规范中细粒度关键信息的需求,需要进一步提升数据项的描述精度。具体而言,要对“引言、前言”等用于辅助分析的数据项做内容要素限定,确保值域明确恰当,切实发挥辅助作用。统一指标项形式,强化其自明性。实现图表内容的数字读取与转换。对涵盖所有属性的表格进行表格框架归纳提取,然后将表格框架转换为关系数据库的关系表实现表格信息抽取,最后通过SQL语言实现表格元组属性值的查询;图片内容则可以通过标注描述插图的条件或场景来实现,例如,对于插图(NAS1321 无头铆钉的制孔——NACA 方法)便可以标注为“应用条件: NAS1321、NACA 方法、制孔和安装”。

对于技术要求、操作指标则要深入工艺过程,对工艺操作、工序名称、工艺内容做术语提取,明确设备、工装、物料、参数等具体要素。如飞机制造工艺中热处理工艺就含有“真空热处理”工序,有加热炉、加热介质、装夹等工具;加热温度、保温时间、淬火转移时间等工艺参数;淬火油的运动粘度、闪点、水分等内容术语。同时,应针对民用飞机装配和检索语言不断优化更新的需要,结合术语提取内容,建立更加专业的主题词表,满足主题标引的需要。

5.3 以对象为中心推进数据项模块化

模块化主要体现在根据实际使用的需要,将资源对象区分为若干个实体,对资源的描述,即对多个不同实体的组合和描述^[34]。在民用飞机装配和生产加工流程中,数据模块化体现为对同一模块相关内容的一体化输出需求,特别是在工艺规范数据的复用和集成需求方面。通过数据项的模块化,可以形成面向工作流的实体属性集,从更深层次描述实体属性,进而更好地实现内容与业务流程、应用场景的融合。如发布日期、实施或试行日期、确认日期可归为日期元素集;发布机构、批准单位、出版单位等可划为机构单位元素集。日期元素集和机构单位元素集既可独立又能关联用于资源的描述,各个元素集可进一步细分。如机构单位元素集面向不同标准出版单位描述不同,电子资源和实体资源的出版将分为电子出版社和实体出版商。

以对象为中心进行模块化,可以把关联数据项进行封装,满足工艺规范应用者对同一模块的相关工艺规范信息脉络化、条目化展示的需要,也方便管理者对工艺规范信息进行集成。

5.4 协同推进工艺规范制定与同步数字化

航空生产制造活动具有多方性, 工艺规范的应用同样呈现链条化、多方关联状态。民用飞机装配工艺的数字化, 需全面考虑工艺规范制定中的多方关联因素。首先, 工艺规范与其他资源的关联, 例如相关法律。这要求制定工艺规范时, 需充分考虑这些资源, 确保规范的有效性和合规性。其次, 工艺规范内部的关联也需要重视。例如, 历史版本与现版本的一致性程度, 以及新旧规范之间的衔接。解决这些问题有助于保持工艺规范的连贯性和完整性。此外, 企业、数据库商、组织机构间的元数据规范也要协调一致。三方共同努力实现术语、数据元素和数据模型的统一。只有这样, 不同主体间的资源才能直接有效集成, 关键信息在工艺规范应用者的工作系统中才能方便调用, 形成工作成果。同时, 也为管理者重组工艺规范、根据标准范式表达工艺规范文件奠定数据基础。此外, 技术更新对民用飞机装配工艺规范的有效性提出了更高的要求, 必须加强业界、标准界、数据商间的同步数字化。一方面, 标准界与业界要对工艺规范进行复审, 确保其内容能够及时对飞机装配行业的新零件、新设备、新机床等作出相应规定。另一方面, 数据商要充分利用记录状态、记录识别符数据项对工艺规范新增、修订、废止信息的标识作用, 及时推送、提醒使用者, 保障有效性的发挥。即, 各主体通过更好的流程和方法来简化工艺规范的应用, 以更大的灵活性和最低的成本交付产品, 在合作中生产出更符合数字时代运行方式的新的、更好的工艺和服务^[35]。

6 结语

《国家标准化发展纲要》在落实提高标准数字化水平部署中明确: 优化标准制定修订流程、平台和工具, 提高标准质量水平^[36]; 提升标准化技术支撑水平, 推动标准化工作向网络化、数字化、智能化转型^[37]。本研究就上述目标进行探索, 但本文聚焦于需求和现有的元数据方案, 以国家军用标准全文服务系统作为单案例来探讨, 受企业数字化需求实践资料和时间限制, 其普适性和全面性有待通过更多案例进一步验证并丰富研究结论。

参考文献

- [1] 胡雪梅. 基于数字化工艺的标准化要求贯彻[J]. 航天标准化, 2020(4): 19-22.
- [2] TAMURA T, NOGUCHI T, SEKIMOTO Y, et al. Performance and Uniformity of Mass-Produced SIS Mixers for ALMA Band 8 Receiver Cartridges[J]. IEEE Transactions on Applied Superconductivity, 2015, 25(3): 1-5.
- [3] HE Q, XUAN H, LIU L, et al. Perforation of aeroengine fan casing by a single rotating blade[J]. Aerospace Science & Technology, 2013, 25(1): 234-241.
- [4] BILLADE B, NYSTROM O, MELEDIN D, et al. Group of Adv. Receiver Dev. (GARD), Chalmers Univ. of Technology. Gothenburg, Sweden[J]. Terahertz Science & Technology IEEE Transactions on, 2012, 2(2): 208-214.
- [5] 魏鹏啸, 郭钊, 丁正一, 等. 熔模铸造工艺集成计算平台与应用[J/OL]. 特种铸造及有色合金, 1-8[2024-04-18]. <http://kns.cnki.net/kcms/detail/42.1148.TG.20240315.2050.014.html>.
- [6] 邱燕平, 熊光利, 吴兴杰, 等. 航空制造企业工艺设计与管理平台研究与应用[J]. 航空制造技术, 2022, 65(6): 82-89.
- [7] 李志雨, 孙文霞, 黄友欢. 数控机床刀具管理和配送平台的研究与应用[J]. 热力透平, 2024, 53(1): 47-52.
- [8] 王桂铝, 张清, 任淑本, 等. 基于模糊自学习的卷

- 烟品质工艺参数控制平台设计[J]. 机械设计与制造工程, 2024, 53(3): 39-43.
- [9] 孙小琪, 王大宏, 杨超, 等. 航空金属材料热加工仿真研究现状[J]. 金属加工(热加工), 2024(4): 24-29.
- [10] 孙升, 任伟, 林海波, 等. 面向数字化制造的工艺工作分析[J]. 航天工业管理, 2021(6): 44-47.
- [11] ISO, TMBG SAG MRS. Questionnaire on ISO TCs' experience of working with SMART standards[Z]. Geneva: ISO, 2019.
- [12] 马超. 中国标准数字化转型: 认知阐释、现实问题及发展路径[J]. 图书与情报, 2023(4): 50-63.
- [13] 马小雯, 孙红军, 刘彦林, 等. 标准知识数字化表达通用模型与自动抽取技术研究[J]. 标准科学, 2024(1): 83-87.
- [14] 汪烁, 卢铁林, 尚羽佳. 机器可读标准——标准数字化转型的核心[J]. 标准科学, 2021(S1): 6-16.
- [15] 毕国楦. 航空工业标准化基础[M]. 北京: 航空工业出版社, 2002: 4-5.
- [16] 郭德华, 李景. 标准信息资源检索与实用指南[M]. 北京: 科学技术文献出版社, 2015: 5-8.
- [17] 陈润. 波音公司的数字化标准[J]. 航空标准化与质量, 2013(6): 30-33.
- [18] 陈嵩, 陈向东. 加强工艺信息标准化工作 促进数字化工艺设计发展[J]. 航空标准化与质量, 2004(2): 28-29.
- [19] 许增辉, 杜坤鹏, 陈建奎. 面向飞机精益制造的工艺标准化管理[J]. 科技创新导报, 2020, 17(1): 163-164.
- [20] 何薇, 冯叶素, 梁丹. 航天制造企业数字化标准体系的研究与实践[J]. 航天制造技术, 2011(2): 48-53.
- [21] ISO. Standards in a digitalized society-Draft report [EB/OL]. (2018-09-07) [2023-08-15]. <https://lisotc.iso.org/livelink/livelink/open/tc59>.
- [22] 崔静, 杨建军. 标准数字化服务探索[J]. 信息技术与标准化, 2022(10): 6-12.
- [23] 魏来. 标准数字化推动行业数字化转型——标准数字化, 数据标准化[J]. 工程建设标准化, 2022(9): 21-25.
- [24] BSI. New BSI Flex standards enable dynamic consensus[EB/OL]. (2020-09-16) [2024-02-08]. <https://www.bsigroup.com/en-GB/insights-and-media/media-centre/press-releases/2020/september/new-bsi-flex-standards-enable-dynamic-consensus>.
- [25] ZVEI. Details of the asset administration shell part 1t: The exchange of information between partners in the value chain of industrie 4.0(version 2.0)[S]. Berlin: Federal Ministry for Economic Affairs and Energy(BMWi), 2019: 123-128.
- [26] 央视网. 国家标准委: 国家标准制定周期将缩短到 30 个月[EB/OL]. (2019-01-10) [2022-08-03]. <http://news.cct.com/2019/01110/ARTIGRNSUdZCDHHhCeIKOPid190110.shtml>.
- [27] QIN J, ZENG M L. Metadata[M]. New York: Neal-Schuman Publisher, Inc, 2004: 247-249.
- [28] WILLIAM E M, ERIN L STEWART, C R MCCLURE. Assessing Metadata Quality: Findings and Methodological Considerations From an Evaluation of the Us Government Information Locator Service (gils) [C]//Research and Technology Advances in Digital Libraries, 1998: 246-255.
- [29] BESIKI S. A Framework for Information Quality Assessment [EB/OL]. (2012-07-20) [2023-10-09]. <http://stvilia.cci.fsu.edu/wp-content/uploads/2011/03/IQAssessmentFramework.pdf>.
- [30] OCHOA X, DUVAL E. Automatic evaluation of metadata quality in digital repositories[J]. International Journal on Digital Libraries, 2009(10): 67-91.
- [31] 黄莺, 李建阳. 元数据质量评估方法及模型研究[J]. 图书馆学研究, 2013(12): 52-56, 51.
- [32] 曹月珍, 马建玲. 国内外元数据质量控制的研究进展与发展趋势[J]. 图书与情报, 2013(6): 101-104.
- [33] 胡永健, 周寄中. 科技资源信息元数据质量审核方法研究[J]. 管理评论, 2011, 23(1): 41-47.
- [34] 于倩倩, 张建勇, 黄永文. 大数据环境下的文献元数据标准设计特点分析[J]. 图书馆杂志, 2018, 37(11): 35-39, 46.
- [35] EVANS J, MCKEMMISH S, REED B. Making Metadata Matter: Outcomes from the Clever Recordkeeping Metadata Project[J]. Archives & Manuscripts, 2009, 37(1): 30-56.
- [36] 国家标准化管理委员会, 中央网络安全和信息化委员会办公室, 中华人民共和国科学技术部, 等. “十四五”推动高质量发展的国家标准体系建设规划[EB/OL]. (2021-12-30) [2024-02-07]. <https://www.chinadesign.cn/23/202112/2037.html>.
- [37] 工业和信息化部. “2021 十四五”大数据产业发展规划[EB/OL]. (2021-11-30) [2024-02-07]. https://wap.miit.gov.cn/zwgk/zcwj/wjfb/tz/art/2021/art_c4a16fae377f47519036b26b474123cb.html.



开放科学
(资源服务)
标识码
(OSID)

国家学术关注影响因素分析初探

刘星湄 伊凡 陆伟

武汉大学信息管理学院 武汉 430072

摘要: [目的/意义] 将以某个国家为研究对象的科学研究过程称为对该国的学术关注, 并从国家层面开展学术关注度量与影响因素分析; 旨在从宏观上了解国家学术关注行为的影响因素, 并为理解国际关系提供新视角。[方法/过程] 基于 Web of Science 核心合集数据库, 采集 1998—2021 年世界各国有关其他国家研究的论文数据, 开展国家学术关注的度量和描述性统计分析; 并从主客体国家特征和国家联系两个维度, 采用多重线性回归方法, 探索国家学术关注行为的影响因素。[局限] 由于国家学术关注的影响因素非常复杂, 不少因素存在难以获取和难以量化的问题。对于国家学术关注的影响因素的回归分析不够精确, 有待进一步优化。[结果/结论] 从散点图看, 一个国家对于全球的学术关注程度和受到全球的学术关注度均受到该国的地理因素、社会因素、经济因素、科技因素、军事因素的影响。通过回归分析发现, 在其它因素不变的情况下, 军费开支大、国土面积大、人口总数大的国家受到全球的学术关注度更大, 人口年增长率高的国家对于全球的学术关注程度更大; 部分情况下国土面积小、科技论文发文量小的国家对于全球的学术关注程度更高。与中国之间的自然、文化和经济联系均显著影响中国对其他国家的学术关注, 在其他因素不变的情况下, 与中国具有共同官方语言、国土接壤、进出口贸易额越大的国家较全球平均水平受到中国的学术关注更大。以上各种因素的影响随着国家范围和时间而变化。

关键词: 国家学术关注; 影响因素; 多重线性回归; 回归分析

中图分类号: G35

A Preliminary Analysis of Factors Influencing National Academic Attention

LIU Xingmei YI Fan LU Wei

School of Information Management, Wuhan University, Wuhan 430072, China

Abstract: [Objective/ Significance] This paper refers to the scientific research process of a certain country as national academic

作者简介 刘星湄 (1998-), 硕士研究生, 主要研究方向为科学计量、知识管理, E-mail: 2020201040041@whu.edu.cn; 伊凡 (1999-), 博士研究生, 主要研究方向为科学计量与文本挖掘; 陆伟 (1974-), 博士, 教授, 主要研究方向为信息检索与可视化、数据智能与创新评价等。

引用格式 刘星湄, 伊凡, 陆伟. 国家学术关注影响因素分析初探 [J]. 情报工程, 2024, 10(4): 47-60.

attention of that country, conducts academic attention measurement and influencing factor analysis at the national level, and intends to explore the influencing factors of national academic attention behavior from a macro perspective and provide a new perspective for understanding international relations. [Methods/Processes] This paper searches Web of Science core collection database and collects data on global research papers related to other countries from 1998 to 2021. It conducts measurement and descriptive statistical analysis of national academic attention, and explores the influencing factors of national academic attention behavior using multiple linear regression methods from two dimensions: the characteristics of the subject and object countries and the connections between countries. [Limitations] Due to the complex influencing factors of academic attention between countries, many factors are difficult to obtain and quantify, the regression analysis in this article is not precise enough and needs further optimization. [Results/Conclusions] Judging from the scatter plot, both the amount of academic attention a country receives globally and the ratio of academic attention it sends globally are influenced by its geographical, social, economic, technological, and military factors. Through regression analysis, it was found that countries with higher military spending, larger land area, and larger population receive greater global academic attention, while other factors remain unchanged. Countries with higher annual population growth rates have higher ratio of global academic attention. In some cases, countries with smaller land areas and less scientific and technological papers have higher ratio of global academic attention. The natural, cultural and economic connections with China have significantly affected China's academic attention to the country. Countries that share the same official language with China, border with China, and have larger import and export trade volume, are receiving more academic attention from China than global average. The impact of various factors above varies with changes in national scope and years.

Keywords: National Academic Attention; Influencing Factors; Multiple Linear Regression; Regression Analysis

引言

“关注”是人脑有选择地指向和集中于一定对象或活动的心理状态。“关注”最初是一个心理学概念，被心理学家视为“进入意识的一个重要过程”，也是人类执行有意识行为的关键一步^[1]。诺贝尔奖得主 Simon^[2]指出：由于受到决策者的能力以及信息获取条件等诸多因素制约，决策者往往无法获得决策相关的所有信息，因此对决策有着重大影响的因素便是决策者对事物的关注。陈然^[3]认为社交网络中的关注行为能够反映用户的社会关系及通过这一关系获得的社会资本。学术社交媒体平台中的关

注行为，则体现了用户在学术上关系的建立。严炜炜等^[4]从用户在线学术关系网络的视角出发，分析 ResearchGate 平台上科研用户关注和合作关系的行为模式，发现关注关系网络各用户间联系更紧密。

在学术研究中，研究人员同样无法获取研究相关的所有资源和信息，学术关注对于其研究过程和研究结果同样有着重要影响。从研究者个人层面来说，其学术关注受到研究领域和研究兴趣的影响，如果将研究视角提升到国家层面，会发现研究者群体的学术关注具有一定的共性特点，也会受到国家政策的引导以及社会环境和时代背景的影响。本

文将以某个国家为研究对象的科学研究过程称为对该国家的学术关注，探索性开展国家学术关注度量和影响因素分析。从国家层面开展学术关注研究，围绕研究者所在国家以及研究对象所在国家进行分析，有助于发现隐藏在研究者个性化学术关注背后的特点和规律，从宏观上了解各国之间的学术关注行为及相关机制，也为理解国际关系提供新视角，并丰富区域国别研究的内容。

由于目前鲜见关于国家学术关注的研究，因此本文转向国家间合作等类似科研行为的研究，从中探寻可借鉴的思路和方法。国家学术关注行为和科研合作同属于跨国学术研究行为。一方面，两者有着较为密切的联系，基于对其他国家的学术研究和学者状况的关注和把握，才能成功地实现跨国科研合作。另一方面，在研究内容和国家选择上二者又有所差异；前者侧重于选择具有研究基础和合作可能性的学者和国家展开合作，后者侧重于选择具有研究兴趣和研究意义的国家作为研究对象。

本文分别从主客体国家特征和以我国为代表的国家联系特征两个维度展开，采用回归分析方法探索国家学术关注行为的影响因素，尝试研究以下三个问题：

Q1：具备哪些特征的国家受到其他国家的学术关注度更大，即全球学术被关注度受到客体国家的哪些因素影响？这些影响因素是否随着时间推移发生变化？

Q2：具备哪些特征的国家对于其他国家的学术关注程度更高，即全球学术关注程度受到主体国家的哪些因素影响？这些影响因素是否随着时间推移发生变化？

Q3：具备哪些特征的国家与我国的学术关注关系强度更大，即国家间学术关注关系强度受到哪些因素的影响？这些影响因素是否随着时间推移发生变化？

1 学术关注的概念与度量

1.1 学术关注量与学术关注度

在国家学术关注中，本文将学术关注的发出方称为主体国家，其关注的国家称为客体国家，二者之间产生的关系称为国家学术关注关系，图1为国家学术关注关系的简单示意图。

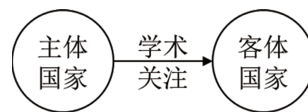


图1 国家学术关注关系

学术关注关系最直接的体现是学术成果。两个国家之间的学术关注关系体现为，在主体国家作者公开发表的学术成果中，将客体国家作为研究对象。关系的强度则体现为学术成果的数量，主体国家以客体国家为研究对象发表的学术成果越多，其关注关系强度越大。值得注意的是，随着国际合作日渐频繁，某个学术成果可能由多个国家合作完成，此时出现多个主体国家关注一个或多个客体国家的情况。本文采用全计数法^[5]来度量国家之间的学术关注，即一篇论文中各主体国家对各客体国家的关注得分均为1。

1.2 研究假设

为研究国家学术关注行为的影响因素，可以借鉴相关的研究成果，比如：有学者认为区

域国别研究的兴衰与国际形势变化密切相关,其发展程度也与主体国综合国力有明显的相关关系。国家综合国力的量化评估广受国内外关注。黄硕风^[6]、胡鞍钢等^[7]学者建立了综合国力评价指标体系,中国社会科学院^[8]和中国科学院相关项目^[9]建立了评价模型,开展了综合国力排序。目前尚未得到国内外公认的综合国力评价体系。国内的量化研究大多基于客观数据进行评价,各类评价指标大都可以分为硬实力和软实力两类,包含政治、经济、科技、文化、教育、国防、外交、资源等构成要素。曹原等^[10]建立了由7个大类指标、50个具体指标组成的综合国力综合评价体系,具体包括国土面积等自然资源指标,劳动力人口数等人力资源指标,GDP等经济发展指标,进出口贸易总额等对外经济指标,军费支出GDP占比、武装部队人员总数等军力水平指标,研发支出GDP占比、科技期刊文章数等科技水平指标等。《美国新闻与世界报道》联合发布的最佳国家排名(Best Countries Ranking)基于问卷调查数据开展,其指标体系包括投资机会、应变能力、文化影响力、企业精神、传统传承、创新能力、商业友好度、国家实力、生活质量和社会责任10个类别,偏重于可能推动贸易、旅行和投资等方面的特征指标^[11]。

综上所述,政治、经济、科技、文化、国防、自然资源、人口资源等都是综合国力的重要构成要素,也是学术关注影响因素中可以纳入考虑的主客体国家特征要素。因此,我们提出如下假设:

H1: 全球学术被关注量与关注行为客体国家经济、军事、科技、自然、社会等因素相关,

即经济繁荣、军事强大、科技发达、人口数量多、国土面积大的国家,在学术上更多地被其他国家关注。

H2: 全球学术关注度与关注行为主体国家经济、军事、科技、自然、社会等因素相关,即经济萧条、军事弱小、科技落后、人口数量少、国土面积小的国家,在学术上更多地关注其他国家。

在国家间的关注行为中,一个国家同时会从自己国家的角度出发,关注那些与本国具有具体关系的国家。国家间存在着地缘、人文、经济、环境、安全等方面超越国界的连接,他们共同影响区域间关系的发展^[12]。顾伟男等^[13]发现地理邻近性、社会邻近性是“一带一路”沿线国家科研合作网络形成的主要机制。桂钦昌等^[14]发现全球科研合作网络受到地理邻近性、认知邻近性、社会邻近性、共同语言、历史联系和人才流动等多重因素的共同作用。张明倩等^[15]发现“一带一路”跨国专利合作网络形成受到地理、宗教、语言、贸易、技术等多种接近性因素影响。基于此,本文做出如下假设:

H3: 主客体国家学术关注关系的强度与经济往来、文化渊源、地缘关系等因素相关,即经济往来密切、文化渊源深厚、地理位置临近的国家之间学术关注关系更紧密。

1.3 变量的操作化

针对上面的三个研究问题,本文分别从主客体国家特征和主客体国家联系两个角度,采用表1和表2所示的变量进行国家学术关注因果关系研究。

表 1 主客体国家特征变量描述表

变量类型		变量名称	变量数据来源	变量说明
自变量	地理因素	国土面积（LandArea）	世界银行世界发展 指标数据库（World Development Indicators）	指一个国家的土地面积
	人口因素	人口总量（PopTotal）		指按年中估算值计算的居民总量
		人口年度增长率（PopGrow）		指按年中估算值计算的居民总量年度增长百分比
		人口净迁移量（NetImm）		移居到国外的移民总数减去每年移居到国内的移民数量
	经济因素	进出口总额（Imp&Exp）		按现价美元计算
		GDP		按现价美元计算
	科技因素	科技论文发文量(STJourArtes)		SCI/SSCI 科技论文发文量
		研发投入占比（R&DRatio）		研发投入 GDP 占比重
	军事因素	军费开支（MilExp）		指军队、国防部委和其他从事防卫项目的政府机构、准军事部队等的经常性支出和资本支出
		武装力量人数（ArmPers）		包括现役军人和可支援或代替正规军队的辅助军队人数
因变量		全球学术关注度（FocRatio _{ig} ）	WOS 核心合集获取论文数据计算得到	i 国作者所发表的关于 i 国之外其他国家的论文计数占 i 国作者发文总量的比例
		全球学术被关注量(FocArtes _{gi})		i 国之外其他国家作者所发表的关于 i 国的论文计数

表 2 主客体国家联系变量描述表

变量类型		变量名	变量说明	变量数据来源
自变量	地缘联系	相邻关系 (Contig)	与中国是否有共同边界	法国国际展望与信息研究中心 CEPII 数 据库
		加权地理距离 (DisTw)	以城市 GDP 作为权重计 算与中国间任意两两城市 之间距离的加权平均值	
	文化联系	语言联系 (LangConn)	与中国是否具有相同的官 方语言	
	经济联系	进出口贸易总额 (Impt&Expt)	与中国间进出口贸易总额	中经网统计数据库
因变量		被关注量国际排名差 (FocedRankDif _i)	WOS 核心合集获取论文 数据计算得到	将特定年份世界各国按被其他国家关 注的论文量降序排列, i 国在其中的排 名作为受全球关注度排名; 将该年份世 界各国按被中国关注的论文量降序排 列, i 国在其中的排名作为受中国关注 量排名; 以上两项的差值为被关注量国 际排名差

1.4 数据获取与处理

本文以 Web of Science 核心合集集中的学术
成果为研究对象, 主要文献类型为论文, 也包

括部分书籍章节、评论等其他学术成果。筛选
1998—2021 年间, 在标题、关键词或摘要中存
在国家名称且存在作者及作者国别信息的数据,

共计 6916395 条；按照全计数方式进行国家关注行为统计。一篇文章只要有一个或多个作者地址字段包含某个国家，则该国作为主体国家的关注量计数算作 1；一篇文章的篇名、关键词、摘要字段中只要提及某一个国家，则该国作为客体国家的关注量计数算作 1。对于别名等各类写法进行名称规范，统一采用官方名称，按照国家统计局标准进行归并，建立各年份主体国家与客体国家间的关注行为矩阵，在此基础上进行国家学术关注度量。

为了进行横向和纵向比较分析，本文选取不同国家范围和不同时段的数据进行回归分析。除全体国家外，本文还希望能聚焦在学术关注方面最活跃的一批国家。因此本文使用国家学术关注数据，以国家为网络节点，学术关注关

系为节点间连线建立了国家学术关注网络。再通过社会网络分析方法中的点度中心度分析，为全球国家在学术关注方面的活跃度进行排名，最终得出最近一年点度中心度最大的 40 个国家，即最活跃的 40 个国家。

最终国家范围上采用了能匹配上数据的全球所有国家（以下简称所有国家）和 2021 年全球关注活跃度前 40 的国家（以下简称 TOP40 国家，详见表 3）两类数据。考虑自变量数据的全面性，选取了数据相对齐全的近 24 年数据（即 1998—2021 年），进行所有国家和 TOP40 国家的对比分析。针对 TOP40 国家则选取了 1998—2005、2006—2013、2014—2021 三个年段数据，进行不同年段的比较，并研究影响因素随时间的变化情况。

表 3 全球关注网络中活跃度前 40 的国家

排名	国家名称	度中心性	排名	国家名称	度中心性
1	美国	1.831	21	尼日利亚	1.345
2	中国	1.672	22	瑞典	1.331
3	印度	1.612	23	比利时	1.327
4	巴西	1.598	24	新西兰	1.327
5	澳大利亚	1.584	25	埃塞俄比亚	1.327
6	英国	1.579	26	巴基斯坦	1.313
7	德国	1.579	27	墨西哥	1.308
8	加拿大	1.556	28	波兰	1.289
9	法国	1.551	29	挪威	1.28
10	南非	1.546	30	马来西亚	1.28
11	日本	1.537	31	沙特阿拉伯	1.28
12	西班牙	1.528	32	埃及	1.28
13	意大利	1.471	33	印度尼西亚	1.28
14	瑞士	1.443	34	孟加拉国	1.266
15	俄罗斯联邦	1.425	35	丹麦	1.261
16	韩国	1.378	36	哥伦比亚	1.233
17	伊朗	1.369	37	智利	1.228
18	葡萄牙	1.364	38	越南	1.219
19	土耳其	1.359	39	捷克	1.214
20	荷兰	1.355	40	奥地利	1.21

2 学术关注的描述性统计分析

本文首先对上述变量进行相关性分析，如图2所示。从各变量与全球学术被关注量和全球学术关注度的影响来看，经济因素（GDP、进出口总额）、科技因素（研发投入占比、科技论文发文量）、军事因素（军费开支、武装力量人数）、人口因素（人口总量、人口净迁移量）、自然因素（国土面积）与全球学术关注度呈负相关，与全球学术被关注量呈正相关。而人口变量中人口年度增长率与全球学术关注

度呈正相关，与全球学术被关注量呈负相关。

对于自变量之间的关系，GDP与进出口总额、军费开支、科技论文发文量等指标具有较强相关性，相关系数均高于0.8。武装力量人数与人口总量相关性也很高，相关系数为0.83。整体来说，经济、科技与军事因素之间都具有较强的相关性。这也非常符合认知，通常经济发达的国家的科技和军事实力也较强。在后续的研究中，会对高度相关的变量进行剔除等处理，尽可能避免多重共线性问题。

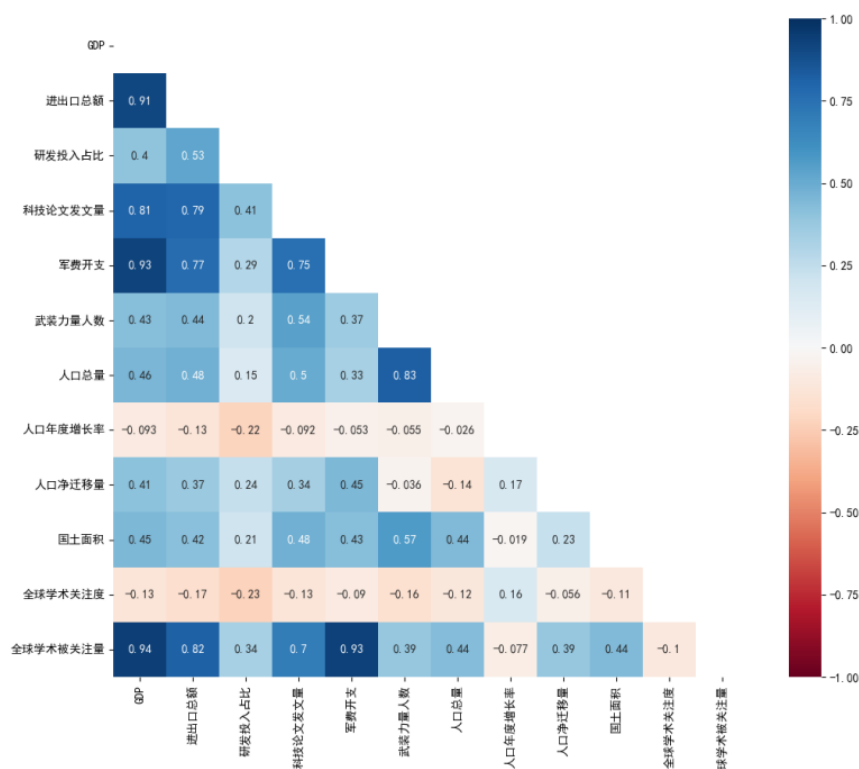


图2 国家特征变量相关性分析

各变量与国家学术被关注量和学术关注度的关系散点图如图3和图4所示。整体来说，图3中各个变量均与全球学术被关注量呈现正相关关系。表明经济、科研、军事、人口规模越大、国土面积越大、人口增长率越高的国家，

受到全球的学术关注量越高。

根据图4来看，各变量与国家学术关注度之间的线性关系并非特别明显。整体来说，全球学术关注度与人口年度增长率呈现正相关，与其他变量均呈现负相关态势。这表明经济、科研、军事、

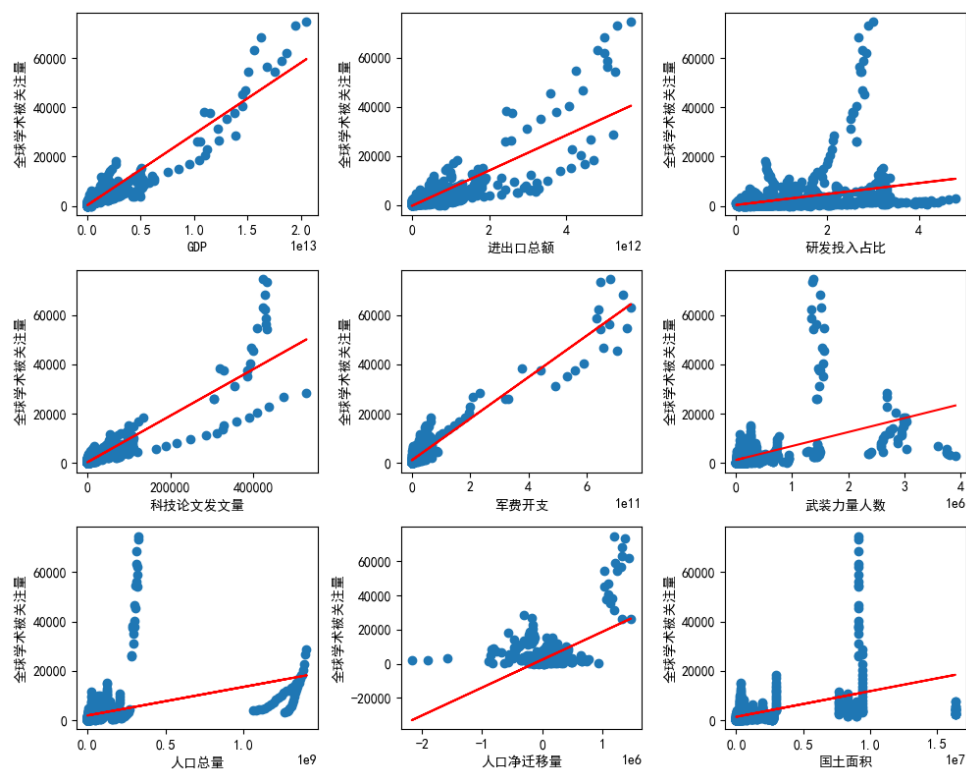


图3 全球学术被关注量散点图

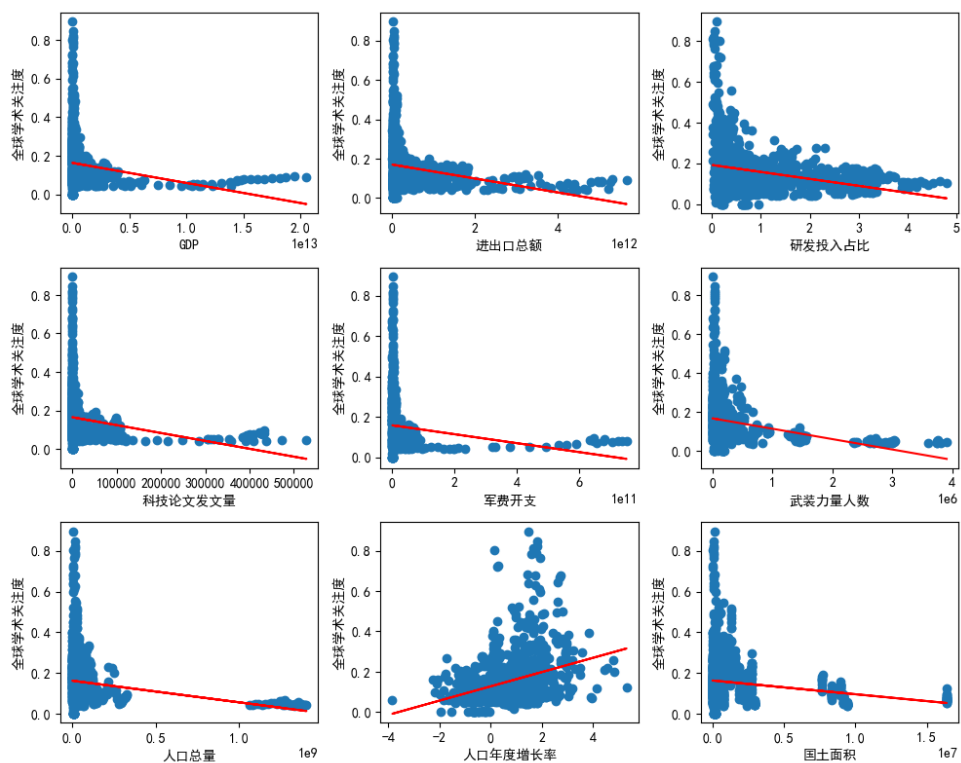


图4 全球学术关注度散点图

人口规模越大,国土面积越大、人口增长率越低的国家,国外相关主题论文比例相对更低;也就是说经济、科研、军事发达,人口多,面积大,人口增长缓慢的国家对于全球的学术关注度相对更低,对本国的学术关注度相对更高。

对于我国为主体国家、其他国家为客体国

家的情况,变量之间的相关性如图 5 所示。我国的关注量与地理相邻关系、语言联系、进出口贸易总额呈现正相关,与加权地理距离呈现负相关,这表示我国更关注地理位置相邻、加权地理距离相近、具有相同官方语言、与我国经济联系更紧密的国家。

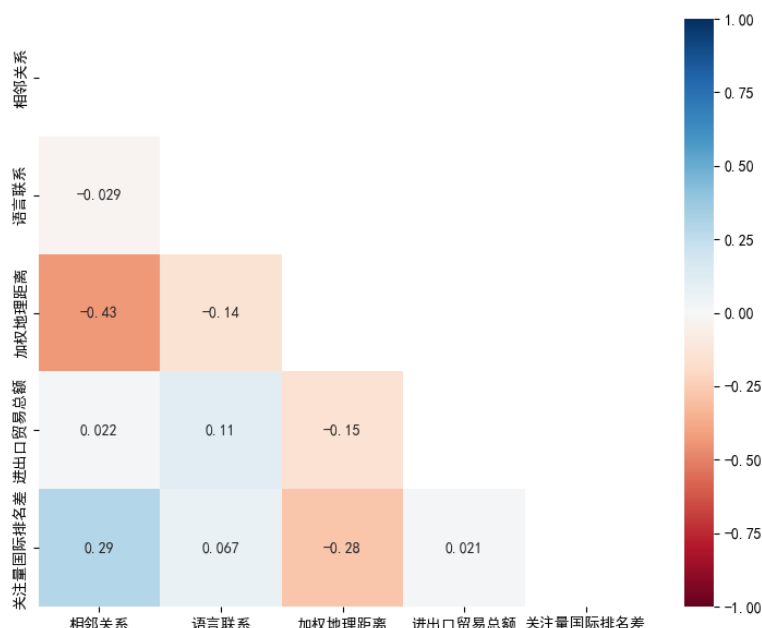


图 5 国家联系变量相关性分析

如图 6 所示,与我国具有相同官方语言、具有相邻关系的国家,我国对其的整体关注量排名差为正,说明我国对这部分国家关注的程度高于这部分国家被世界关注的平均水平,即这部分国家更受我国关注。而随着加权地理距离的增加,我国关注国际排名差下降。在经济方面,进出口贸易总额与我国关注他国的关注量国际排名存在微弱的正相关关系。

3 基于多重线性回归的国家学术关注的因素分析

多重共线性指自变量之间存在高度相关性。

为保障回归模型的科学性和准确性,本文对多重共线性 VIF 值高的变量做剔除处理。保障自变量的 VIF 值小于 5,并在此基础上开展后续的回归分析。

3.1 回归模型

由于文中自变量和因变量都是定距层次的变量,而且是多个自变量和一个因变量的关系,因此本章的统计模型选择了多元线性回归。假设 H1、H2、H3 对应的表达式分别为:

H1:

$$FocedArtcs_{gi} = \alpha_1 + \beta_1 Imp\&Exp_i + \beta_2 GDP_i + \beta_3 Pats_i + \beta_4 R\&DRatio_i + \beta_5 MilExp_i + \delta_i \quad (1)$$

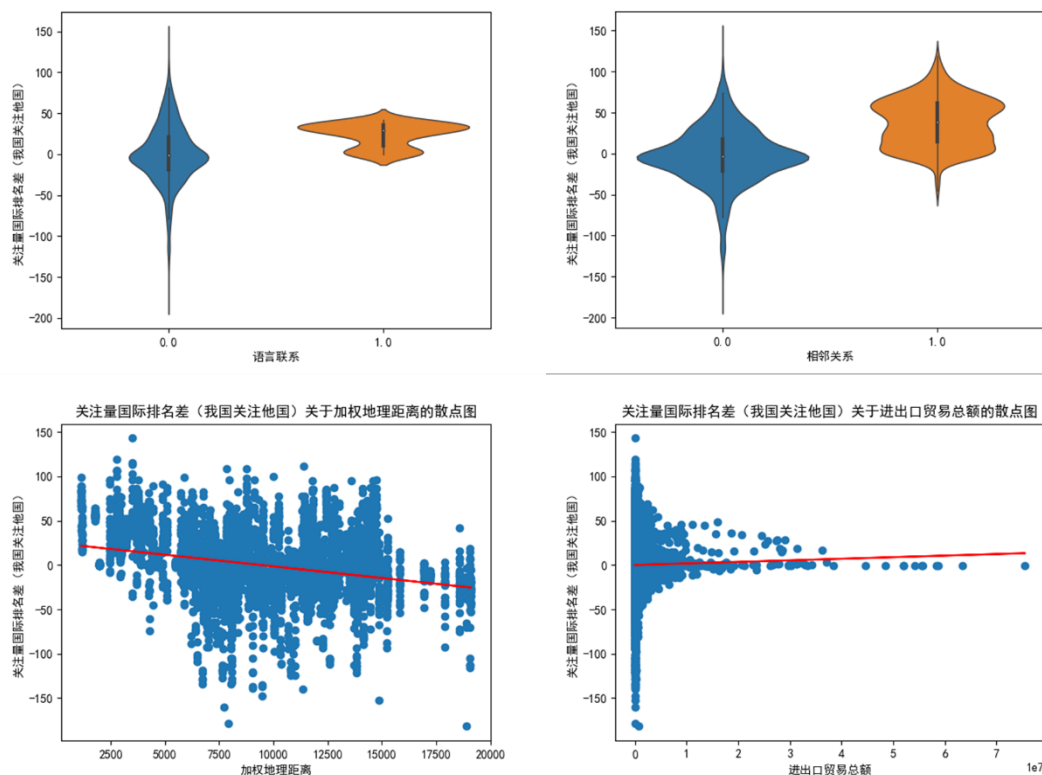


图6 关注量国际排名差与各变量关系图

H2:

$$FocRatio_{ig} = \alpha_2 + \beta_6 Imp \& Exp_i + \beta_7 GDP_i + \beta_8 Pats_i + \beta_9 R \& D Ratio_i + \beta_{10} MilExp_i + \delta_2 \quad (2)$$

H3:

$$FocedRankDiff_i = a_1 + b_1 LangConn_i + b_2 DisTw_i + b_3 Contig_i + b_4 Impt \& Expt_i + c_1 \quad (3)$$

3.2 回归结果

3.2.1 基于主体国家特征的因素分析

表4中调整的 R^2 较低,说明该模型拟合程度不太理想。究其原因可能是因为学术关注度的影响因素更为复杂,存在该模型之外的其它影响因素,比如学者个人的研究兴趣,而这些影响因素不在本文的研究范围内,同时也不好量化。

①地理因素(国土面积)

国土面积在所有国家及TOP40国家部分时

段中通过显著性检验,回归系数在-0.035~-0.142之间;表明其在部分情况下与全球学术关注度低度负相关,即在其他自变量不变的情况下,国土面积越小的国家对于世界其他国家的学术关注度越大。

②社会因素(人口年度增长率)

人口年度增长率在两种国家范围、四个时段中的显著性P值均低于0.01,各类数据中回归系数均在0.136~0.308之间;表明其与因变量呈现较为稳定的低度正相关,即在其他自变量不变的情况下,人口年度增长率越大的国家对于世界其他国家的学术关注度越大。从国家范围上看, TOP40国家的相关性系数相对于所有国家更大。

③经济因素(进出口总额)

在所有国家以及TOP40国家部分时段中进出口总额通过了显著性检验,表现为低度负相

表 4 全球学术关注度多元线性回归结果

	VIF	所有国家	TOP40 国家			
		1998-2021	1998-2021	1998-2005	2006-2013	2014-2021
国土面积	1.315	-0.063*** (0.016)	-0.13*** (0.037)	-0.142** (0.06)	-0.035 (0.072)	-0.085 (0.063)
人口年度 增长率	1.023	0.136*** (0.014)	0.298*** (0.034)	0.272*** (0.06)	0.308*** (0.060)	0.224*** (0.0561)
进出口总额	3.299	-0.19*** (0.026)	-0.024 (0.058)	-0.187 (0.115)	0.049 (0.145)	-0.327*** (0.098)
科技论文 发文量	3.238	-0.008 (0.025)	-0.153*** (0.052)	0.025 (0.114)	-0.446** (0.202)	-0.158** (0.070)
军费开支	2.872	0.096*** (0.024)	0.04 (0.053)	0.032 (0.116)	0.212* (0.108)	0.132 (0.086)
样本数		5088	960	320	320	320
调整的 R ²		0.051	0.151	0.135	0.196	0.247

注: *、** 和 *** 分别表示 0.05<P<0.10, 0.01<P<0.05, P<0.01; 星号前的数据为回归系数, 括号内为相关系数的标准误差。

关。表明进出口总额对因变量有一定的低度负向影响, 但是影响不具有稳定性。

④科技因素 (科技论文发文量)

在所有国家以及 TOP40 国家部分年段中科技论文发文量均通过回归检验, 回归系数在 -0.008~0.446 之间; 表明其与全球学术被关注量低度负相关; 即在其他自变量不变的情况下, 科技论文发文量越大的国家对于世界其他国家的学术关注度越小。

⑤军事因素 (军费开支)

在所有国家以及 TOP40 国家部分年段中, 军费开支的显著性 P 值均低于 0.01, 回归系数较低; 表明其与全球学术关注度低度正相关或者不具相关性。

3.2.2 基于客体国家特征的因素分析

表 5 中多元线性回归模型调整的 R² 超过 0.9, 表明该模型是一个匹配比较理想的模型。

①地理因素 (国土面积)

国土面积在五类数据中显著性 P 值均低于

0.01, 回归系数在 0.019~0.068 之间; 表明其与全球学术被关注量具有稳定的低度正相关关系; 即在其它自变量不变的情况下, 国土面积越大的国家受到世界其他国家的学术关注量越大。从国家范围上看, TOP40 国家的国土面积相关性系数比所有国家略大。

②社会因素 (人口总数)

人口总数在五类数据中显著性 P 值均低于 0.01, 各类数据中回归系数均在 0.05~0.13 之间; 表明其与因变量呈现较为稳定的低度正相关, 即在其它自变量不变的情况下, 人口总数越大的国家受到世界其他国家的学术关注量越大。从国家范围上看, TOP40 国家与所有国家的相关性系数非常接近。

③经济因素 (进出口总额)

在所有国家以及 TOP40 国家部分年段中, 进出口总额的显著性 P 值均低于 0.01, 回归系数在 0.102~0.301 之间。总的来说其对于因变量具有低度正向影响, 1998—2005 年 TOP40 国家

表 5 全球学术被关注量多元线性回归结果

	VIF	所有国家	TOP40 国家			
		1998-2021	1998-2021	1998-2005	2006-2013	2014-2021
国土面积	1.427	0.019*** (0.005)	0.037*** (0.012)	0.068*** (0.014)	0.047*** (0.016)	0.06*** (0.016)
人口总数	1.538	0.129*** (0.005)	0.11*** (0.011)	0.056*** (0.013)	0.094*** (0.016)	0.125*** (0.015)
进出口总额	3.435	0.301*** (0.008)	0.263*** (0.017)	-0.006 (0.023)	0.102*** (0.034)	0.118*** (0.023)
科技论文发文量	3.396	-0.239*** (0.008)	-0.257*** (0.016)	-0.009 (0.025)	-0.0949* (0.054)	-0.177*** (0.018)
军费开支	2.989	0.831*** (0.007)	0.879*** (0.016)	0.947*** (0.025)	0.933*** (0.028)	0.896*** (0.022)
样本数		5088	960	320	320	320
调整的 R ²		0.919	0.922	0.959	0.96	0.952

注：*、** 和 *** 分别表示 0.05<P<0.10, 0.01<P<0.05, P<0.01；星号前的数据为回归系数，括号内为相关系数的标准误差。

的数据未通过检验，可能存在数据异常情况。

④科技因素（科技论文发文量）

除了 1998—2005 年 TOP40 的数据外，其他的四类数据中科技论文发文量均通过回归检验，回归系数在 -0.006~-0.257 之间；表明其与全球学术被关注量低度负相关；即在其他自变量不变的情况下，科技论文发文量越大的国家受到世界其他国家的学术关注量越小。结合图 3 的散点图来看，科技发展水平与全球学术被关注量呈正相关关系，但在表 4 所示的回归模型中回归系数却为负。这可能是因为科技论文发文量与军事、经济发展水平难免存在正相关关系，而相比之下，军事发展水平对因变量的作用更重要，综合作用下使得科技论文发文量对因变量的结果不显著或呈现负向影响。

⑤军事因素（军费开支）

军费开支在全部五类数据中的显著性 P 值均低于 0.01，回归系数在 0.831~0.947 之间；体现出其与全球学术被关注量稳定的高度正相关性；即在其他自变量不变的情况下，军费开支

越大的国家受到世界其他国家的学术关注量越大。从国家范围上看，TOP40 国家的军费开支相关性系数比所有国家略大。

3.3 基于主客体国家联系的因素分析——以中国为例

表 6 中调整的 R² 在 0.75~0.85 之间，拟合程度较为理想。所有自变量的显著性 P 值均小于 0.01，通过了显著性检测，且显著性程度高。

①地缘联系（相邻关系）

相邻关系在两种国家范围、四个时段数据中相关性系数差别不大，均表现为中度正相关；其在 TOP40 国家的三个时间段内相关系数变化不大，具有稳健性。

②地缘联系（加权地理距离）

加权地理距离与因变量受关注度国际排名差中高度正相关，即在其他自变量不变的情况下，加权地理距离越大的国家受到世界关注的论文量排名与受中国关注的论文量排名差别越大。这一分析结果与假设正好相反，有可能是

表 6 学术被关注量国际排名差多元线性回归结果

	VIF	所有国家	TOP40 国家			
		1998-2021	1998-2021	1998-2005	2006-2013	2014-2021
相邻关系	1.245	0.5423*** (0.013)	0.5541*** (0.035)	0.5976*** (0.062)	0.5449*** (0.055)	0.5501*** (0.056)
加权地理距离	1.293	0.9401*** (0.007)	1.1913*** (0.023)	1.1751*** (0.04)	0.9707*** (0.037)	0.7075*** (0.038)
语言联系	1.039	0.4208*** (0.034)	0.5151*** (0.061)	0.5934*** (0.108)	0.4731*** (0.096)	0.3754*** (0.097)
进出口总额	1.034	0.8064*** (0.061)	1.002*** (0.082)	0.983*** (0.125)	0.8683*** (0.094)	0.9201*** (0.099)
样本数		4632	960	320	320	320
调整的 R ²		0.83	0.827	0.815	0.815	0.751

注：*、** 和 *** 分别表示 $0.05 < P < 0.10$, $0.01 < P < 0.05$, $P < 0.01$ ；星号前的数据为回归系数，括号内为相关系数的标准误差。

因为关注行为超越了国家距离的限制。从 TOP40 国家的三个时间段数据来看，加权地理距离对于因变量的影响随着时间变化而减弱。

③文化联系（语言联系）

语言联系与因变量受关注度国际排名差中低度正相关；从 TOP40 国家的三个时段数据来看，其相关性随着时间变化减弱。表明不同的官方语言造成国家学术关注的障碍，但随着时间推移，跨国学术交流不断增强，语言障碍在逐渐减弱。

④经济联系（进出口总额）

进出口总额表现出与因变量中高度正相关。与所有国家数据相比，进出口总额在 TOP40 国家各时段数据中的相关系数更大，表明在关注行为活跃国家中进出口总额的影响更大。

4 结论与讨论

本文基于 Web of Sciece 核心合集数据库，采集 1998—2021 年世界各国关于其他国家的论文数据，采用多重线性回归方法，研究国家学术关注行为的影响因素，主要结论如下：

①关注行为客体特征中，军事因素、社会因素和地理因素对全球学术被关注量具有稳定的影响；其中军费开支表现为高度正相关，国土面积、人口总数为低度正相关，即在其它因素不变的情况下，军费开支大、国土面积大、人口总数大的国家受到全球的学术关注量更大。从散点图来看，经济因素和科技因素与全球学术被关注量同样呈现正相关关系，但在回归模型中其影响不稳定或呈现负向影响，可能是受到其他自变量影响的结果。

②关注行为主体国家特征因素中，从散点图看，地理因素、社会因素、经济因素、科技因素、军事因素均与全球学术关注度呈正相关关系；但在回归模型中，仅有人口年度增长率这一个自变量表现为显著而稳定的正向影响，国土面积、科技论文发文量在部分数据中表现为低度负相关；即在其它因素不变的情况下，人口年度增长率高的国家对于全球的学术关注程度更大，部分情况下国土面积小、科技论文发文量小的国家对于全球的学术关注程度更高。

③从关注行为主客体国家联系因素来看，

与中国之间的地理、文化、经济联系均影响中国的对别国的学术关注。在其它因素不变的情况下,与中国具有共同官方语言、国土接壤、出口贸易额越大、加权地理距离越大的国家相对于全球而言,受到中国的学术关注更大。

④以上各种因素的影响随着时间和国家范围变化有所变化。

本文从国家层面展开学术关注度量与研究,分析了各国之间的学术关注行为的影响因素,尝试揭示国家学术关注行为的机制,为理解国家关系提供新的视角。然而,国家学术关注的影响因素非常复杂,除了本文中所考虑的地理、社会、经济、科技、军事等因素外,还存在政治、历史等影响因素,以及突发事件、研究者个人偏好等其他因素。各种因素对于学术关注会造成错综复杂的影响,而且不少因素存在难以获取和难以量化的问题。受精力和能力所限,本文仅做了探索性研究,存在如下局限。首先,本文所设计的回归模型相对简单,变量选取主要参考前人的研究成果并考虑数据的可获得性。现有变量较为简单,数据源也非常有限。其次,国家学术关注的数据清洗和度量相对简单。目前以标题、关键词或摘要中存在国家名称算作对该国的学术关注,忽略了部分特殊情况,包括:标题、关键词和摘要中出现了国家名称而不以该国为研究对象,英美等国家研究本国的论文而未出现“英国”和“美国”等字眼,针对欧盟等区域组织的研究而未出现特定国家的名称等。今后的研究可以考虑更全面的影响因素和更丰富的数据源,考虑从语义层面实现更加准确的数据清洗和统计,也可以考虑在特定学科内进行回归分析,从而提高回归分析的精准性。

参考文献

- [1] 刘景江,王文星.管理者注意力研究:一个最新综述[J].浙江大学学报(人文社会科学版),2014,44(2):78-87.
- [2] SIMON H A. Designing organizations for an information-rich world [J]. Computers, communications, and the public interest, 1971, 72(2): 221-237.
- [3] 陈然.网络论坛活跃群体社交网络研究——从“关注”行为的视角[J].新闻界,2012(18):51-55,61.
- [4] 严炜炜,温馨.关注与合作:科研用户在线学术关系网络与行为模式分析[J].情报理论与实践,2022,45(4):75-82.
- [5] LINDSEY D. Production and Citation Measures in the Sociology of Science: The Problem of Multiple Authorship[J]. Social Studies of Science, 1980, 10(2): 145-162.
- [6] 黄硕风.综合国力与国情研究[J].中国国情国力,1992(1):13-19.
- [7] 胡鞍钢,郑云峰,高宇宁.对中美综合国力的评估[J].清华大学学报(哲学社会科学版),2015,30(1):26-39,180-181.
- [8] 王涌芬.世界主要国家综合国力比较研究[M].长沙:湖南出版社,1996:65-68.
- [9] 中国科学院可持续发展战略研究组.中国可持续发展战略研究报告[R].北京:科学出版社,2003:20-21.
- [10] 曹原,葛岳静,马腾.2000年以来中国及周边国家综合国力评价与格局变化简[J].经济地理,2018,36(4):10.
- [11] U.S.News Staff. Methodology: How the 2022 Best Countries Were Ranked [EB/OL]. (2022-09-02) [2022-09-27]. <https://www.usnews.com/news/best-countries/articles/methodology>
- [12] 张蕴岭.国际区域学概论[M].济南:山东大学出版社,2022:145-151.
- [13] 顾伟男,刘慧,王亮.“一带一路”沿线国家科研合作网络的多元结构及形成机制[J].地理研究,2020,39(5):1070-1087.
- [14] 桂钦昌,杜德斌,刘承良.全球科研合作网络的动态演化及其驱动机制[J].地理学报,2023,78(2):423-442.
- [15] 张明倩,柯莉.“一带一路”跨国专利合作网络及影响因素研究[J].软科学,2018,32(6):21-25.



开放科学
(资源服务)
标识码
(OSID)

国内产业集群研究现状、热点主题及演化路径

魏玲¹ 陈彦均¹ 曾冰^{1,2} 梁立芳¹

1. 澳门科技大学可持续发展研究所 澳门 999078;

2. 安徽财经大学经济学院 安徽 233030

摘要: [目的/意义] 产业集群是区域经济发展的重要组织形式与载体,揭示国内产业集群的发展脉络与研究现状,对我国现代化产业体系发展具有重要的理论价值与实践意义。[方法/过程] 利用 CiteSpace 软件对中国知网数据库中 2000—2023 年间涉及产业集群研究的 6102 篇文献进行文献计量与网络知识图谱分析。[结果/结论] 通过研究发现,国内产业集群研究在理论体系、研究视角、方法等方面都取得了长足进步,但同时也存在研究合作不足、地域集中、个案研究多但普适性弱等问题。未来,需进一步拓展研究视角,加强合作交流,采用更多元的研究方法,以适应高质量发展的新要求,为我国产业集群转型升级提供有力支撑。

关键词: 产业集群; 文献计量; 研究热点; CiteSpace

中图分类号: F062.9; G353.11

Current Status, Hot Topics and Evolution Paths of China's Industrial Cluster Research

WEI Ling¹ CHEN Yanjun¹ ZENG Bing^{1,2} LIANG Lifang¹

1. Institute for Sustainable Development, Macau University of Science and Technology, Macao 999078, China;

2. School of Economics, Anhui University of Finance & Economics, Anhui 233030, China

Abstract: [Objective/Significance] Industrial clusters are an important organizational form and carrier for regional economic development. Revealing the development context and research status of domestic industrial clusters, which has important theoretical value and practical significance for the development of China's modern industrial system. [Methods/Processes] This article uses CiteSpace software to conduct bibliometric and network knowledge graph analysis on 6102 documents involving industrial cluster research in the China National Knowledge Infrastructure database from 2000 to 2023. [Results/Conclusions] Through research, it has been found that significant progress has been made in the theoretical system, research perspectives,

基金项目 国家自然科学基金地区项目“复杂网络结构下城市群经济合作研究”(72163010)。

作者简介 魏玲(1990-), 博士研究生, 主要研究方向为区域与城市经济; 陈彦均(1992-), 博士研究生, 主要研究方向为产业经济与科技情报; 曾冰(1986-), 博士, 副教授, 主要研究方向为区域与城市经济, E-mail: 120210014@aufe.edu.cn; 梁立芳(1988-), 博士研究生, 主要研究方向为区域与城市经济。

引用格式 魏玲, 陈彦均, 曾冰, 等. 国内产业集群研究现状、热点主题及演化路径[J]. 情报工程, 2024, 10(4): 61-72.

and methods of domestic industrial cluster research. However, there are also problems such as insufficient research cooperation, regional concentration, and a large number of case studies with weak universality. In the future, it is necessary to further expand research perspectives, strengthen cooperation and communication, and adopt more diverse research methods to adapt to the new requirements of high-quality development, providing strong support for the transformation and upgrading of industrial clusters in China.

Keywords: Industrial Clusters; Bibliometric Analysis; Research Hotspots; CiteSpace

引言

产业集群是当今世界经济中颇具特色的经济组织形式，在促进区域经济发展、提升产业竞争力方面发挥着重要作用。自从迈克尔·波特1990年在《国家竞争优势》一书中首次提出“产业集群”（Industrial Cluster）概念以来，产业集群研究在全球范围内引起了广泛关注。目前，学术界对产业集群的研究涵盖了多个视角，如经济地理学派、产业组织理论学派、新制度学派、创新系统学派、社会网络学派及演化经济学派等，为产业集群研究提供了多元化的理论框架。在国际范围内，产业集群研究已经形成了较为完整的研究体系，涵盖了产业集群的竞争优势、影响因素、政府政策、协同创新等方面，代表性理论包括波特的国家竞争优势理论、克鲁格曼的新经济地理学理论、马歇尔的产业区理论和斯托波尔的产业集群空间经济学理论。同时，国外学者在产业集群研究中采用了多种多样的研究方法，常用的有实证研究、案例分析、定性或定量研究、空间分析、社会网络分析、演化博弈论、复杂系统分析、跨学科研究及实验经济学等方法。现有研究成果有利于明确产业集群基本内涵、发展模式、运行机制等问题，初步形成较为完整的研究框架。当前，面对全球化进程加速、新兴技术变革、贸易保

护主义抬头以及地缘政治风险等国际形势的变化，从应对国际形势变化、适应全球化市场需求、重塑传统产业、应对贸易壁垒、实现产业创新发展、推动国内经济转型升级的视角来看，开展产业集群研究对于我国经济高质量发展具有重要的现实意义。面对国外学者在产业集群方面的多样化研究积累，分析并全面掌握我国当前的产业集群研究现状对发现国内学者研究的不足并提出改进措施，为国内产业集群的发展提供更加坚实的理论支撑也至关重要。但是，目前国内关于产业集群的量化研究较为欠缺，缺少对于相关研究成果的系统性梳理。因此，本文利用文献计量分析方法，重点分析和梳理目前国内产业集群研究的总体趋势和研究热点，以加深对国内产业集群研究领域的认识与把握，为全面推进现代化产业体系与经济高质量发展提供参考。

1 研究方法和资料来源

本研究采用文献计量分析法，并利用陈超美教授开发的 CiteSpace 软件对 2000—2023 年间中国知网数据库中篇名为“产业集群”的相关学术论文进行了详尽分析。为确保研究文献的权威性和质量，研究限定在 CSSCI、CSCD 和核心期刊三类文献中进行检索，共收集到 6246 篇文献。

通过对每篇文献标题的仔细阅读和手动筛选，剔除了与研究主题不相关的文献，如人物访谈、领导发言、会议纪要及评述类文献，最终获得有效样本 6102 篇，用于本研究的深入分析。

2 基本面分析

2.1 文献量变化态势

发文量及其趋势总体反映一个领域的受关注程度，并大致反映该领域的研究热度与发展程度。依据涉及“产业集群”研究的重点期刊发文量变化情况（图 1）可以看出，国内对产业集群的研究始于 21 世纪初，过去 24 年间的累计发文量达到 6102 篇，平均每年约有 254 篇相关文献发表。这表明产业集群作为一个研究

领域，在一定程度上保持了持续的学术关注度。特别是在 2000—2007 年间，产业集群研究在起步阶段经历了快速增长，其中 2007 年和 2009 年的发文量达到了峰值。这一时期的高发文量反映了学术界对该领域的极大兴趣，以及产业集群在经济发展中的重要性和影响力的提升。然而，自 2010 年起，发文量开始逐年减少，到 2023 年时已经回落到世纪初的水平。这种下降趋势可能表明，随着研究的深入和领域内知识的积累，产业集群研究的创新点可能减少，或者研究兴趣可能已经转移到其他新兴领域，同时也可能是产业集群发展方面的政策导向或市场动态变化导致的学术关注度下降。这迫切需要通过定量分析明确新的研究增长点，为产业集群研究的转型确定方向。

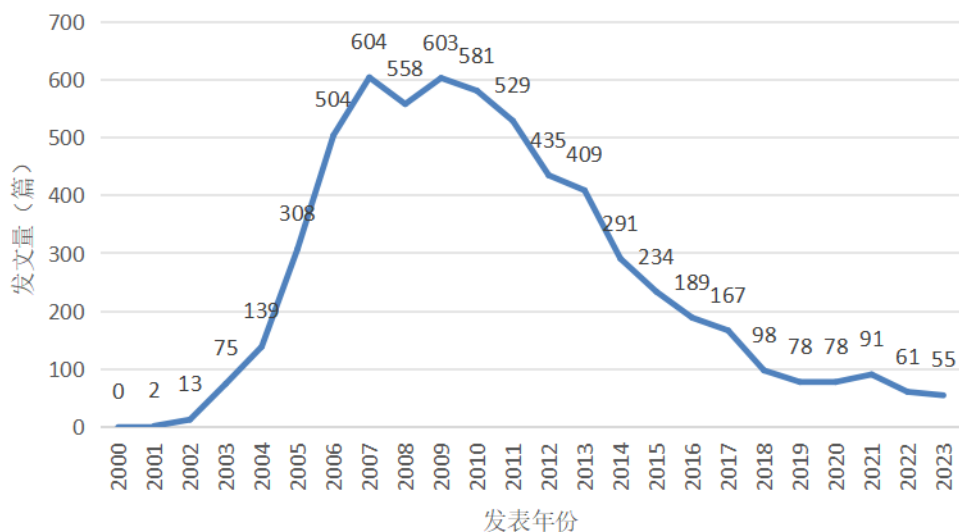


图 1 历年产业集群研究发文量趋势图

2.2 作者特征分析

通过 CiteSpace 中的作者功能对本文文献进行分析，生成了作者共现网络（图 2），整体上呈现出“大分散，小聚合”的特征。共现网络中共有 451 个节点和 124 条连线，整

体网络密度为 0.0012，这说明尽管开展产业集群研究的作者众多，但合作较少，更倾向于独立研究。这种“大分散，小聚合”的现象可能反映了该领域的研究仍处于一个较为分散探索的阶段，或者是因为研究资源和信

息交流的渠道不够畅通,导致研究者之间缺乏有效的合作机制。对于促进学术交流和合

作的现象。未来,通过建立更多的合作平台和交流机制,有望改善这种状况,推动产业集群领域的研究向更深层次发展。

CiteSpace v. 5.2.R2 (64-bit)
Java: 2024-03-28 10:30:39 PM CST
C:\CiteSpace\bin\citespace.jar
Timestamp: 2024-03-28 (Show Logfile)
Network: C:\Users\user\Documents\产业集群研究作者共现图谱
Largest CC: 1.176
Nodes Labeled: 1.0%
Pruning: None

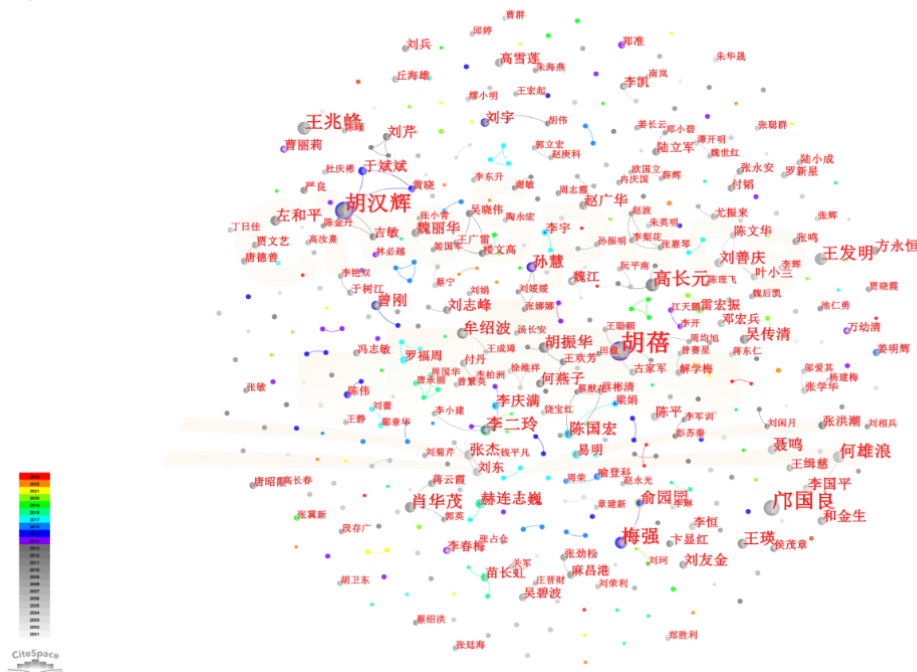


图2 产业集群研究作者共现图谱

另外,根据图2同心圆内外层颜色及表1可见,在排名前十的作者中,发文时间大多始于2006—2010年。胡汉辉、李二玲至今仍在该领域活跃,而胡蓓、梅强、高长元在2015年后停止了发文。邝国良、王兆峰、王发明、何雄浪、肖华茂则更早结束了发文,这表明近年来在该领域内尚未出现新的核心作者。总体而言,在“产业集群”领域持续发文的核心作者较少,核心作者群体增长缓慢。

2.3 机构特征分析

利用 CiteSpace 软件对产业集群研究的文献进行机构共现网络分析,生成的机构共现网络

表1 产业集群研究作者发文量排名

发文量(篇)	首次发文时间(年)	末次发文时间(年)	作者
40	2007	2015	胡蓓
25	2009	2022	胡汉辉
22	2006	2010	邝国良
15	2008	2011	王兆峰
15	2010	2015	高长元
15	2010	2016	梅强
12	2009	2013	王发明
11	2006	2009	何雄浪
11	2009	2012	肖华茂
11	2007	2021	李二玲

图(图3)揭示了产业集群研究领域内的合作模式。网络包含389个节点和88条连线,但整

体网络密度仅为 0.0012，表明国内开展产业集群研究的机构众多但合作稀疏。依据普赖斯定律，发文量超过 6 篇的机构被定义为核心研究机构，共识别出 70 个。据统计，产业集群研究工作主要集中在高等院校，相对聚

集的态势有利于研究的进一步深化。其中，发文量最多的三个核心机构（具体到学院）分别为华中科技大学管理学院、中南大学商学院、武汉大学经济与管理学院，发文量分别为 79、64 和 42 篇。

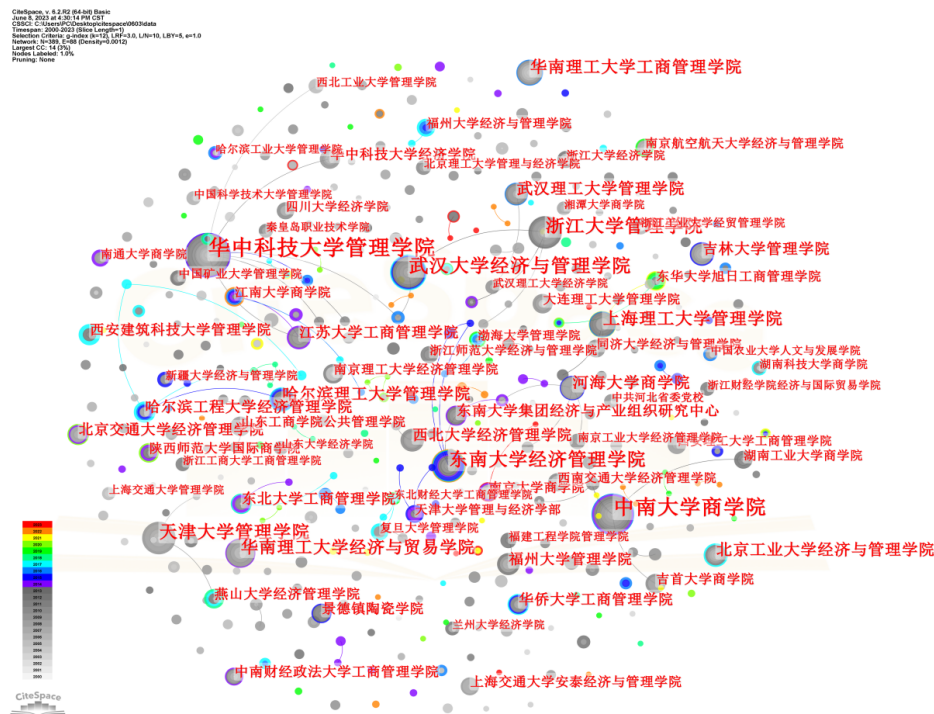


图 3 产业集群研究机构共现图谱

另外，共现网络图（图 3）和机构发文量（表 2）还揭示了机构间合作的地理分布特征，显示了湖北湖南等中部区域的研究活跃度，尤其是以华中科技大学管理学院为代表。同时，江浙地区的研究热度也在上升，近年来东南大学经济管理学院和河海大学商学院的贡献显著。

综合图表信息，可以看出产业集群研究在中国呈现出地域集中的特点，且研究力量分布不均。为推动更广泛的学术合作和交流，促进知识共享与创新，建议加强跨地域合作平台的建立。

表 2 产业集群研究机构发文量排名

发文量 (篇)	首次发文时 间(年)	发文机构
79	2002	华中科技大学管理学院
64	2005	中南大学商学院
42	2007	武汉大学经济与管理学院
41	2003	东南大学经济管理学院
40	2003	浙江大学管理学院
36	2004	天津大学管理学院
29	2006	华南理工大学经济与贸易学院
29	2007	上海理工大学管理学院
26	2005	华南理工大学工商管理学院
23	2010	哈尔滨理工大学管理学院
22	2006	北京工业大学经济与管理学院
22	2002	武汉理工大学管理学院
20	2006	河海大学商学院

3 研究发展路径与演变趋势

3.1 产业集群研究热点识别

关键词是研究者因主旨表达需要而高度凝练,提供给读者获取信息的重要词汇。关键词共现反映了不同主题之间的关系,揭示了信息内容的相关性。而关键词聚类则可以挖掘与主题高度相关的关键词群体,多用于研究热点的识别与分析。本文使用 CiteSpace 的关键词功能生成关键词共现图谱(图4),该图谱中网络节点数量 483 个,连线 1559 条,整体密度为 0.0134,显示产业集群涉及的相关领域广泛,关键词较多,但关联强度不大。关键词节点的大小能反映研究的热点和重点。在产业集群研

究关键词共现图谱中,按照节点从大到小排序,核心的关键词主要包括区域经济、竞争力、技术创新、知识溢出、中小企业、动力机制、创新网络、社会资本、协同创新、产业链、生命周期、影响因素、区域品牌等。其中,“区域经济”“竞争力”“技术创新”“知识溢出”和“中小企业”等关键词节点较大,表明这些是产业集群研究中的核心议题。另外,“动力机制”“创新网络”“产业链”“生命周期”等关键词不仅体现了理论研究的热点,也反映了研究者对产业集群实际运作机制的探索。“演化”“区域品牌”等关键词表明,国内研究者对产业集群的长期发展、演化路径及品牌建设有所关注。

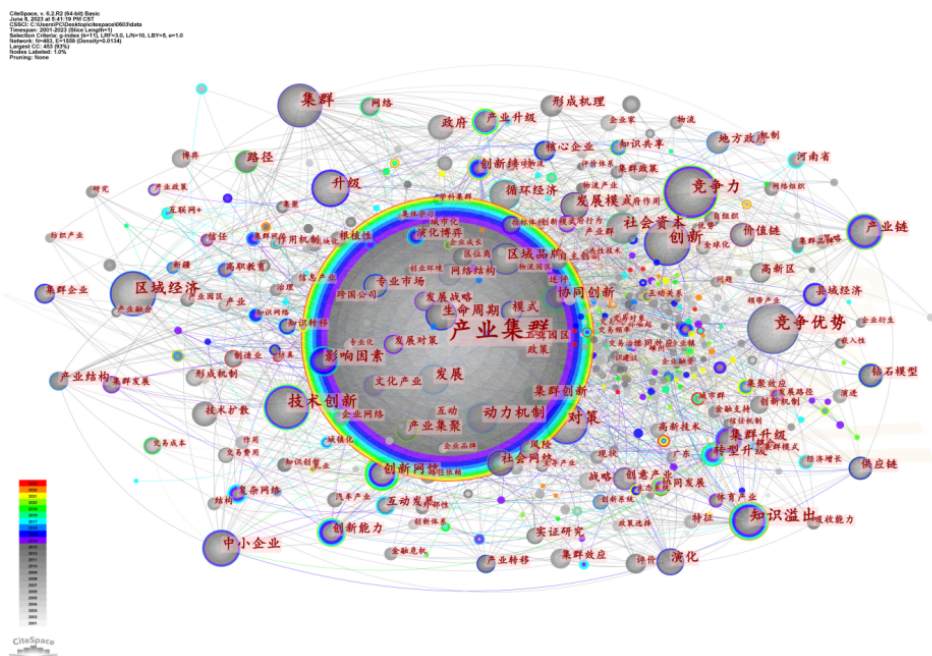


图4 产业集群研究关键词共现图谱

在关键词共现基础上,进一步提取关键词聚类(图5)。关键词聚类是通过中介中心性结合研究文献提炼出研究重点,排名前9的关键词聚类分别是产业集群、技术创新、竞争优势、

集群、竞争力、区域经济、社会资本、影响因素和协同创新。其中,“产业集群”“技术创新”“竞争优势”等聚类显示了研究者对产业集群内部机制、创新过程以及竞争优势来源的深入探讨。

“社会资本”“协同创新”等聚类表明，研究者们开始关注社会网络、合作和协同作用在产业集群发展中的重要性。基于聚类图谱的展示成果，结合人工识读关键文献把握研究热点，目前国内学者对产业集群的研究热点主要集中在竞争力分析、技术创新探讨和社会资本研究等多个维度。另外，产业集群竞争力、技术创新和社会资本是推动我国经济转型升级、促进高质量发展的关键因素。产业集群竞争力的提升可以促进产业升级，带动区域经济增长；技术创新则对产业集群国际竞争力的提升至关重要；而社会资本的积累则有利于降低交易成本，提高运行效率。三者相互促进，共同推动着产业集群的持续发展。通过文献综述，我们发现国内学者在这三方面都取得了丰富的研究成果。

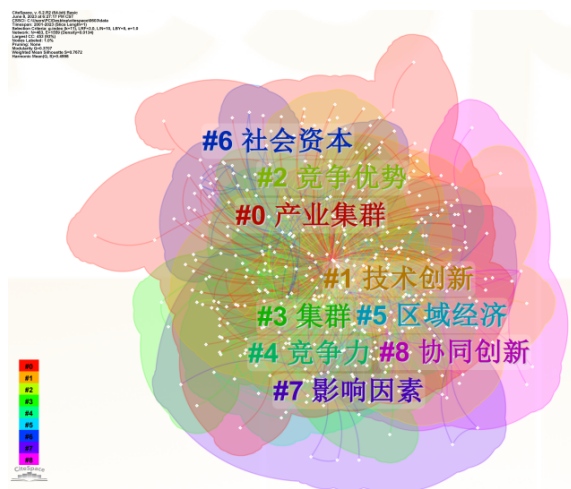


图5 产业集群研究关键词聚类图谱

3.1.1 竞争力分析

产业集群是产业发展的内在规律，也是国家资源配置和提升竞争力的战略选择。学术界对产业集群的竞争力进行了深入研究和探讨，覆盖了国家、区域和城市等多个层面，应用了

钻石模型、GEM模型、熵权法（EWM）、层次因子法（AHP）、结构方程模型（SEM）等多元模型。在研究内容上，关于产业集群竞争力分析的国内研究可以归为以下几类：（1）产业集群竞争力内涵与来源，刘恒江等^[1]总结了产业集群竞争力的内涵主要包含因素、结构和能力三个方面的观点；魏后凯^[2]指出集群的竞争优势来源广泛，如集聚经济效应、灵活专业化、创新型环境、合作竞争机制以及路径依赖等；（2）产业集群竞争力内在机理，张辉^[3]和郑云^[4]从理论上论证了产业集群通过学习效用和对域内资源的自组织能力的正面效应，以及对可能出现“柠檬问题”等经济负外部性和负面效应的规避方面进行了竞争力机制分析；

（3）产业集群竞争力提升路径，魏后凯^[2]认为产业集群可通过提高生产率、刺激企业创新、促进企业进入和成长、形成区域品牌等方式提高竞争力；（4）产业集群竞争力网络视角，吴向鹏^[5]提出以创新网络为核心的经济网络、社会网络相融合的综合网络发展是未来产业集群竞争力网络提升的实现路径；（5）产业集群竞争力案例分析，朱海就等^[6]通过分析浙江和意大利产业集群，总结得出浙江可通过提高企业网络组织化程度，提升产业集群竞争力；赵祥等^[7]通过统计10个产业的数据，得出浙江广东分别在传统制造业、资本与技术密集型制造业具有较高集群竞争力，提出应在产业政策上重点改善集群网络系统的构成和运行；（6）产业集群竞争力评价指标体系，蒋录全等^[8]构建了涵盖12个因素，97个观测指标的五层测评体系；蔡皎洁^[9]从知识的挖掘方面构建评价指标体系。

3.1.2 技术创新探讨

技术创新是科学发展的第一生产力,更是提高产业集群国际价值链地位的关键核心。随着21世纪经济全球化、市场化、信息化的推进,国内学者在产业集群的技术创新方面也进行了广泛探究。从具体研究内容来看,关于产业集群技术创新的研究主要围绕机制、扩散、影响因素和演化路径展开:(1)产业集群技术创新机制,李相银等^[10]认为产业集群从空间集聚、知识交流、文化相似和价值观层面上影响技术创新;黄坡等^[11]认为产业集群提供了技术创新的压力、能力和动力;王琛等^[12]认为产业集群通过沟通、选择、学习和搜寻四个机制协同创新,推动技术创新发展与融合;(2)产业集群技术创新扩散,邝国良等^[13]分析了产业集群与技术扩散的互相促进的自增强关系;邵云飞等^[14]使用数据仿真实验得出空间距离阻碍技术创新的扩散;张秀武等^[15]通过构建重力模型对技术创新扩散的空间进行了研究;(3)产业集群技术创新的影响因素,汤长安^[16]认为在技术创新的早期阶段,政府具有非常重要的协助企业搜寻、选择、传播创新信息等作用;(4)产业集群技术创新演化路径,赵忠华^[17]提出社会联系与信息共享对技术创新多种要素具有正向影响。

3.1.3 社会资本研究

社会资本作为关系资本,与物资资本、人力资本共同构成企业生产的重要资本形式。Bourdieu^[18]定义社会资本为“与群体成员相联系的实际的或潜在的资源总和”。自社会资本概念诞生以来,它受到了国内外经济学、社会学和管理学等多个学科的广泛关注。本文通

过梳理文献发现,国内产业集群社会资本的研究主要集中在其概念、机理、作用机制和评价体系等方面。(1)产业集群社会资本概念和分类。边燕杰等^[19]定义社会资本为“行动主体与社会的联系以及通过这种联系获取稀缺资源的能力”。(2)产业集群社会资本机理。卢林等^[20]认为社会资本通过凝结功能提高交易效率和推动分工演化,提升人与人之间的信任,增强集群的稳定性。周乐欣等^[21]认为产业集群社会资本通过信任机制、流动机制还有创新机制产生作用。金丹^[22]从区域产业集群形成、集群发展和集群的产业链及创新能力上进行社会资本对产业集群的机制分析。(3)产业集群社会资本评价体系。赵国杰等^[23]使用网络层次分析法(ANP)构建了社会资本的评价指标。杨中华等^[24]通过结构、关系和认知三个维度构建进行评价。陈礼林等^[25]通过目标灰色局势评价法进行评价。(4)产业集群社会资本实证研究。朱秀梅等^[26]通过调查问卷实证研究了长春市102家软件企业。颀茂华等^[27]通过168家企业调查问卷数据构建结构方程模型,分析得出社会资本对产业集群竞争力在直接和间接上均发挥了有效作用。还有学者通过回归研究得出知识转移对社会资本和创新绩效具有重要中介效果,提出诸如注重发展集群的资源共享机制、建立深度合作和知识转移的激励机制等对策。

3.2 阶段性前沿研究领域分析

关键词突现反映了短期内出现频次极高的关键词,而关键词突现度可反映某时间段内该研究领域的影响力程度,可用作前沿研究领域

分析。通过文献生成前 25 个关键词突现图谱 (图 6), 从突现强度来看, 协同创新、竞争优势、影响因素和转型升级是排名前四的关键词, 这些关键词代表了产业集群研究的重要方面, 受到了广泛的关注, 是极具影响力的研究前沿热点; 从突现持续时间来看, 发展路径、创新绩效、协同发展和创新网络是排名前四的关键词, 它们在研究领域中影响持续时间最长; 从突现时间序列来看, 2010 年及之前的关键词主要集中在自上而下竞争分析、园区规划、战略分析等前期的规划和资源落地方面, 而 2010 年之后的关键词主要关注影响因素分析、创新升级等发展路径的进一步优化, 这体现了中国产业集群由资源优势向创新优势转变的趋势。最后, 协同发展、创新绩效、发展路径在关键词突现强度上并不太大, 出现的年份较短, 但是它们持续至今, 并且突现性是具有连续性的, 故此认为它们有潜力在未来成为研究的前沿热点, 是值得关注的研究方向。

Top 25 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2001 - 2023
竞争优势	2002	15.84	2002	2007	
工业园区	2003	4.81	2003	2007	
企业家	2004	4.67	2004	2006	
战略	2005	5.65	2005	2008	
创意产业	2007	5.03	2008	2009	
博弈	2005	4.79	2008	2010	
自主创新	2008	4.6	2008	2012	
政策	2005	5.13	2009	2011	
指标体系	2010	5.4	2010	2015	
文化产业	2008	5.74	2011	2014	
高职教育	2011	4.89	2011	2017	
影响因素	2009	11.58	2012	2018	
演化博弈	2012	7.9	2012	2017	
复杂网络	2009	6.43	2012	2017	
集群创新	2009	6.22	2012	2013	
发展路径	2012	5	2012	2023	
协同创新	2008	18.43	2013	2020	
转型升级	2010	10.29	2013	2019	
创新绩效	2008	7.99	2013	2023	
协同发展	2010	5.1	2013	2023	
社会网络	2006	4.97	2013	2014	
创新网络	2004	7.13	2014	2021	
知识网络	2014	6.23	2014	2017	
集群网络	2014	5.89	2014	2016	
创新能力	2006	5.17	2014	2018	

图 6 产业集群研究前 25 关键词突现

3.3 研究主题演变和阶段划分

结合图 6 关键词突现图以及图 7 关键词时区图谱, 可以将中国产业集群研究分为三个阶段。



图 7 产业集群研究关键词时间线

第一阶段为研究摸索期（2000—2006年）：21世纪初期，随着中国市场的进一步开放，资本等要素资源在国际流转速率加快，促进了产业发展与转型，中国产业集群逐渐摆脱“聚而不群”的轻关联阶段，逐步向技术主导而非规模主导发展。此阶段，国内产业集群研究由1999年仇保兴《小企业集群》和2001年王缉慈等人的《创新的空间》拉开帷幕，产业集群研究以定性研究为主，主要探讨产业集群的概念、政策含义、内在机理、动力机制等内涵，以及扩展至区域品牌、生命周期、价值链、供应链、创新体系、产业升级、技术扩散、经济增长、竞争优势、创新网络、战略、对策等外延上。案例区域主要集中于早期产业集群优先发展的浙江省、河南省等省份，以及制造业、产业园区、中小企业等主体对象。

第二阶段为研究成长期（2007—2016年）：随着这一时期传统制造业、高新技术产业、文化创意产业重点集群大量涌现，产业间的前后关联加深，产业链不断优化，区域专业化显现，产业集群逐渐成为区域经济的重要支撑。这一阶段是产业集群研究的成长与扩张阶段，研究主题更为丰富，如创新动力、知识溢出、影响因素、技术创新、转型升级、集群模式、治理与绩效等方面；研究对象也不再限于产业集群本身，而扩展至与其相关的其他对象，如体育、文化、创意、电子信息、物流、金融、汽车、教育、数字出版、农业生态等产业，研究区域扩展至长三角、东北、中部、西部、广东、上海等地区或城市；研究方法也更为多元，如采用复杂网络、演化博弈、仿真、因子分析等。

第三阶段为高质量深化期（2017年至今）：

2017年中国首次提出“高质量发展”，要求推动质量变革、效率变革、动力变革。产业集群增长放缓，更多的是探究技术升级、结构重构的高质量深化发展。在这一阶段，产业集群资源整合能力更强，生产和管理技术更先进，高科技与大数据等前沿技术运用更多。高质量深化期的产业集群研究紧跟产业发展态势，研究方法和专题不断深化，主要使用多种视角对现状加以分析，提出诸如协同、融合发展等解决方案，加大了对创新网络、创新演化、政策效应、交易治理以及价值共享等的分析，同时关注了开发区、一带一路、网络权力、企业价值、产业协同、乡村振兴、中介效应等主题。

总体而言，学者对产业集群的研究随着实践的发展不断推进。随着中国产业高质量发展时代的到来，面对加快建设战略性新兴产业和先进制造业集群的挑战，国家要摆脱以消耗自然资源和利用低成本劳动力为特点的外贸加工导向、依附型集群，转型向创造新知识，生产高附加值、高技术含量的新产品为目标的创新产业集群发展，在复杂多变的国际环境下，产业集群如何让“中国制造”向“中国创造”转型，如何发展具有中国特色的高质量集群，值得进一步研究。

4 结论与启示

本文使用CNKI数据库中2000—2023年6102篇以产业集群为篇名的研究样本，并运用CiteSpace软件进行信息挖掘和整理。通过研究发现，国内产业集群研究在理论体系、研究视角、方法等方面都取得了长足进步，但同时也存在

研究合作不足、地域集中等问题,具体结论如下:

首先,研究热度总体呈增长趋势,2000—2007年间发文量快速增长,之后有回落。这种下降趋势可能表明,随着研究的深入和领域内知识的积累,产业集群研究的创新点可能减少,或者研究兴趣可能已经转移到其他新兴领域,同时也可能是产业集群发展方面的政策导向或市场动态变化导致的学术关注度下降。

其次,产业集群学者合作网络的“大分散,小聚合”现象和研究力量呈现地域集中的特点反映出该领域的研究仍处于一个较为分散探索的阶段,或者是因为研究资源和信息交流的渠道不够畅通,导致研究者之间缺乏有效的合作机制。未来,可通过建立更多的合作平台和交流机制来改善这种状况,推动产业集群领域的研究纵深发展。

最后,从研究热点和演变路径来看,产业集群的研究热点较为分散,包括区域经济、竞争力、技术创新、社会资本等;而协同发展、创新绩效、发展路径逐渐成为未来产业集群学术研究热点;关键词演化具有明显的产业集群演化一致性,中国产业集群的研究由资源优势向创新优势逐渐转变;根据产业集群研究的计量特征,研究大致分为探索期(2000—2006年)、成长期(2007—2016年)和高质量深化期(2017年至今)三个阶段。研究对象较为广泛,个案研究较多,普适性弱,从国内数字化、高端化产业集群发展规划角度出发的理论研究有待加强。

通过系统归纳研究热点与演变趋势,未来应进一步针对现有研究不足拓展研究深度与广度:

第一,未来需进一步拓展研究视角,加强交流,促进相关研究的跨学科合作。采用更多元的研究方法,以适应高质量发展的新要求,完善产业集群理论体系的基本框架,结合现实需求,为我国产业集群转型升级提供有力支撑。

第二,应体现产业集群研究的时代价值,深挖产业集群潜力,结合国家战略,加强对新兴产业、绿色产业等集群的研究,赋能更为丰富的发展场景,为推动经济高质量发展、发展新质生产力贡献力量。

参考文献

- [1] 刘恒江,陈继祥.产业集群竞争力研究述评[J].外国经济与管理,2004(10): 2-9.
- [2] 魏后凯.对产业集群与竞争力关系的考察[J].经济管理,2003(6): 4-11.
- [3] 张辉.产业集群竞争力的内在经济机理[J].中国软科学,2003(1): 70-74.
- [4] 郑云.基于内在经济运行机理的产业集群竞争力分析[J].企业经济,2005(1): 73-74.
- [5] 吴向鹏.产业集群:一个文献综述[J].当代财经,2003(9): 105-108.
- [6] 朱海就,陆立军,袁安府.从企业网络看产业集群竞争力差异的原因——浙江和意大利产业集群的比较[J].软科学,2004(1): 53-56.
- [7] 赵祥,郭惠武.粤苏浙产业集群竞争力的比较研究——基于10大代表性行业的统计分析[J].工业技术经济,2014,33(10): 22-32.
- [8] 蒋录全,吴瑞明,刘恒江,等.产业集群竞争力评价分析及指标体系设计[J].经济地理,2006(1): 37-40.
- [9] 蔡皎洁.基于知识挖掘的产业集群竞争力评价指标体系构建[J].统计与决策,2014(1): 60-63.
- [10] 李相银,胡希.产业集群三个层面的技术创新活动研究[J].甘肃社会科学,2003(6): 69-71.
- [11] 黄坡,陈柳钦.产业集群与企业技术创新[J].新疆社会科学,2006(1): 16-21.

- [12] 王琛,王效俐.产业集群技术创新协同过程及机制研究[J].科学管理研究,2007(5): 22-25.
- [13] 邝国良,方少帆.我国产业集群模式对技术扩散的影响研究[J].工业技术经济,2005(1): 60-62.
- [14] 邵云飞,杜晓明.产业集群内基于时间和距离的技术创新扩散模型研究[J].科技进步与对策,2011,28(20): 67-71.
- [15] 张秀武,林春鸿.产业集群内技术创新扩散的空间展开分析及启示[J].宏观经济研究,2014(11): 114-118.
- [16] 汤长安.产业集群初期技术创新扩散过程的博弈分析[J].科技管理研究,2008(7): 538-540.
- [17] 赵忠华.技术创新扩散为中间变量的创新型产业集群网络关系特征提升创新绩效路径研究[J].科学管理研究,2012,30(6): 37-40.
- [18] BOURDIEU P. The forms of social capital: Handbook of Theory and Research for the Sociology of Education[M]. New York: Greenwood Press, 1986: 21.
- [19] 边燕杰,丘海雄.企业的社会资本及其功效[J].中国社会科学,2000(2): 87-99.
- [20] 卢林,姜滨滨.产业集群与社会资本:一个文献综述[J].商业研究,2009(10): 27-30.
- [21] 周乐欣,李魁.我国产业集群的社会资本理论视角[J].科技创业月刊,2006(9): 23-24.
- [22] 金丹.社会资本对区域产业集群的作用机制分析[J].现代产业经济,2013(Z1): 24-29.
- [23] 赵国杰,关长海,刘翠娥.基于 ANP 的产业集群社会资本评价[J].现代财经(天津财经大学学报),2007(12): 26-29.
- [24] 杨中华,卫武.基于网络层次分析法的产业集群社会资本评价[J].华东经济管理,2009,23(12): 148-151.
- [25] 陈礼林,崔祥民.产业集群社会资本评价体系与方法[J].生产力研究,2012(10): 183-184.
- [26] 朱秀梅,蔡莉.基于高技术产业集群的社会资本对知识溢出影响的实证研究[J].科学学与科学技术管理,2007(4): 117-121.
- [27] 颜茂华,张婧鑫,好日娃,等.社会资本、组织学习能力对产业集群竞争力的影响[J].资源与产业,2019,21(6): 30-38.



开放科学
(资源服务)
标识码
(OSID)

基于 DEA 方法的科技企业孵化器绩效评价与 对策研究 ——以安徽省为例

杨志发¹ 盛玮芮¹ 张虹冕² 刘晶晶¹

1. 安徽省科学技术情报研究所 合肥 230011;
2. 安徽省农村与社会科技发展中心 合肥 230031

摘要: [目的/意义] 对科技企业孵化器开展绩效评价, 不仅有助于提高孵化器自身运营管理水平, 而且对政府部门制定有针对性的政策, 引导孵化器健康发展也具有重要意义。[方法/过程] 采用 DEA 方法将科技企业孵化器投入的人力、物力、财力等资源作为投入变量, 将经济、孵化、社会等效益作为产出变量构建绩效评价指标体系, 选用 2019—2022 年安徽省科技企业孵化器的相关统计数据, 对全省 16 个地市的科技企业孵化器的综合技术效率、纯技术效率与规模效率进行了实证分析。[局限] DEA 方法对投入产出指标的数量有限制, 在后续研究中需要进一步凝练牵动性指标。[结果/结论] 研究发现, 安徽省大部分地市科技企业孵化器处于 DEA 有效状态, 但蚌埠、阜阳、黄山等部分市还未实现最佳投入产出效益。基于研究结论, 结合实地调研, 提出了科技企业孵化器未来发展的对策建议。

关键词: 科技企业孵化器; DEA; 绩效评价; 指标体系

中图分类号: G35; TP391

Research on Performance Evaluation and Countermeasures of Technology Business Incubators Based on DEA Method: A Case Study of Anhui Province

YANG Zhifa¹ SHENG Weirui¹ ZHANG Hongmian² LIU Jingjing¹

1. Anhui Information Institute of Science and Technology Department, Hefei 230011, China;
2. Anhui Rural and Social Science and Technology Development Center, Hefei 230031, China

Abstract: [Objective/Significance] Conducting performance evaluation on technology business incubators not only helps

基金项目 2023 年安徽省创新环境建设专项“新一轮机构改革形势下加强安徽省级科技部门创新资源统筹协调体制机制及对策研究”(202306f01050024)。

作者简介 杨志发(1986-), 硕士, 副研究员, 主要研究方向为科技创新发展战略、规划、政策等; 盛玮芮(1991-), 硕士, 助理研究员, 主要研究方向为科技统计、区域创新评价等, E-mail: 510776095@qq.com; 张虹冕(1986-), 硕士, 副研究员, 主要研究方向为科技创新发展战略、规划、项目管理等; 刘晶晶(1986-), 硕士, 助理研究员, 主要研究方向为科技创新发展战略、产业等。

引用格式 杨志发, 盛玮芮, 张虹冕, 等. 基于 DEA 方法的科技企业孵化器绩效评价与对策研究——以安徽省为例[J]. 情报工程, 2024, 10(4): 73-83.

improve their own operational management level, but also has important significance for government departments to formulate targeted policies to guide the healthy development of incubators. [Methods/Processes] This article uses the DEA method to take the human, material, and financial resources invested in technology business incubators as input variables, and the economic, incubation, and social benefits as output variables to construct a performance evaluation index system. Using relevant statistical data of technology business incubators in Anhui Province from 2019 to 2022, the comprehensive technical efficiency, pure technical efficiency, and scale efficiency of technology business incubators in 16 cities throughout the province were empirically analyzed. [Limitations] The DEA method has limitations on the number of input-output indicators, and further refinement of the impact indicators is needed in subsequent research. [Results/Conclusions] Research has found that technology business incubators in most cities in Anhui Province are in the effective state of DEA, but some cities such as Bengbu, Fuyang and Huangshan have not achieved the best input-output benefits. Based on research conclusions and combined with field research, this article proposes countermeasures and suggestions for the future development of technology business incubators.

Keywords: Technology Business Incubator; DEA; Performance Evaluation; Indicator System

引言

习近平总书记指出,高质量发展是全面建设社会主义现代化国家的首要任务,推进中国式现代化离不开科技、教育、人才的战略支撑,关键在科技现代化。要推动科技现代化,离不开科技服务业的支撑。科技服务业是促进科技与经济融合发展的重要纽带,成为推动以科技创新改造升级传统产业、引领发展新质生产力的关键产业。科技企业孵化器作为科技服务业发展的一支重要新兴力量,为培育科技型企业、促进科技成果转化、推动新兴产业发展等提供了重要的创新创业服务支撑。

加快发展科技企业孵化器,对塑造科技创新创业氛围,推动科技服务业高质量发展,夯实以企业为主体的科技创新基础,以科技创新引领现代化经济体系建设具有重大现实意义。在此背景下,各级政府持续加大对科技企业孵化器的财力资源、基础建设、人员配置的投入力度。随着零基预算制度的全面推广,地方政府部门迫切需要开展财政投入绩效评价,为更好地指导科技企业孵化器健康发展,需要对其

投入产出效率进行科学评价。

1 研究现状

国内外关于科技企业孵化器的研究成果较为丰富,相关学者从多个层面、多种角度对科技企业孵化器进行了不同程度的讨论和分析。Becker 等^[1]针对美国艾奥科技园建立了一个基于职业业绩评估的评价体系,并从其对公司收益的正面影响出发,给出了相关的评估方法和模式。Hackett 等^[2]基于黑箱理论构建了一套关于科技企业孵化器绩效的基模系统,对科技企业孵化器绩效开展了评价研究。Saidi-Mehrabad 等^[3]通过数学模型,采用实证分析的方法对科技企业孵化器的绩效进行了评价研究,并结合案例进行了评价效用的检验分析。国内学者夏星等^[4]首次提出运用三阶段数据包络分析(DEA)模型对科技企业孵化器运行效率进行实证研究,此方法结合了非参数评价方法 DEA 和参数评价方法,克服了外部环境和随机误差的影响。吴文清等^[5]从科技企业孵化器与空间的关联性的角度,研究了国家级科技企业孵化

器的区域差异与集聚度,利用 DEA 方法对我国国家级科技企业孵化器的投入产出效率进行了测算。冯金余^[6]将研究落脚在科技企业孵化器创新驱动效应上,分析了科技企业孵化器促进创新驱动的理论机制,考察了国家级孵化器的创新驱动效应。关成华等^[7]从财政工具的宏观角度探索财政政策对中国科技企业孵化器效率的影响,使用部分共享投入的二阶段网络 DEA 模型,测算中国科技企业孵化器整体效率、创新研发效率和成果转化效率。

关于科技企业孵化器的绩效评价研究,国内已开展相关理论探索。何子佳等^[8]对国内外科技企业孵化器绩效评价相关研究进行了系统梳理,总结了国内外科技企业孵化器的共同点、当前主要流行的分类方式、孵化器绩效的影响因素及绩效评价指标体系的构建。目前,学界对于科技企业孵化器绩效评价体系的研究大多从投入产出角度出发,评价内容主要包括孵化绩效、创新绩效、社会效益等方面。例如,吕明慧^[9]运用 DEA 方法构建评价体系,建立回归模型探讨创新绩效,研究了河南省科技企业孵化器创新综合绩效值及动态演变规律,揭示了河南省科技企业孵化器创新绩效的内在作用机制。郭会斌等^[10]在京津冀协同发展背景下,以人力资源、物力资源、财力资源为投入指标,构建了经济效益、孵化效益和社会效益三维模型,实证检验河北省 2013—2019 年的孵化器绩效。姚梦思^[11]从投入、产出和环境三个维度构建了孵化器创新驱动能力指标体系,对我国孵化器创新驱动能力进行评价,并基于创新地理学理论和空间计量经济学理论,实证分析孵化器创新驱动能力对区域创新绩效的作用机制。

王楠^[12]从投入和产出两个角度,运用 DEA-BCC 模型和 Malmquist 指数方法分别从静态层面和动态层面,对我国科技企业孵化器的绩效进行了全面评估。陈立广^[13]运用三阶段超效率 DEA 模型对 2016—2020 年我国 25 个省份科技企业孵化器的运行效率进行实证分析,运用 Malmquist 指数模型分析了科技企业孵化器运行的动态规律。

综上所述,在现有研究中,运用 DEA 方法进行科技企业孵化器绩效评价的相关研究较为广泛,但关于安徽省科技企业孵化器的绩效评价研究相对较少。因此,本文以安徽省的科技企业孵化器为评价对象,运用 DEA 方法,融合多指标对科技企业孵化器绩效进行全面分析与评价,重点反映安徽省 16 个地市科技企业孵化器相关指标的综合情况,丰富现有研究成果,以为安徽省科技管理部门的监管考核提供可操作的依据,为孵化器运营发展提供可行的对策建议。

2 评价指标体系的建立和评价方法

2.1 评价指标体系的建立

2.1.1 指标选取原则

关于科技企业孵化器绩效评价,国内已形成了一套完整的理论体系,大量学者将评价指标概括为 3 个方面:孵化效率、创新效率和经营效率。本文在《科技部火炬中心关于印发〈科技企业孵化器评价指标体系〉的通知》(国科火字〔2019〕239 号)和《安徽省科技企业孵化器认定、众创空间备案及绩效评价管理办法(试行)》(皖科区〔

2020〕21号)等相关政策文件的基础上,依据以下原则选取指标:

(1) 科学性原则

评价指标体系的建立需要依赖一定的科学理论,这些理论能够从根本上反映评价对象的概念,评价指标需要有一定的衡量标准来对应。同时,科技企业孵化器的绩效评估系统,其层次和属性具有变化性,各项指标的設置必须避免重复,能够全面衡量该系统涉及的各个方面,有机地在各级评价指标和最终的评价目标之间建立联系。评价指标体系要科学、合理、可靠,确保评价结果准确、客观、可信。

(2) 系统性原则

在孵企业作为孵化器的运营主体,将会获取孵化器以及各中介机构经过流动和重组的资源和服务,从而确保创业企业的顺利孵化。同时,科技企业孵化器可以深化机构间的合作,促进科技成果的落地与应用。由此,在构建安徽省科技企业孵化器创新绩效评价体系时,要基于孵化器的功能综合考虑各方面因素的作用,贯彻系统性原则,全面、客观地显示各类投入产出要素在区域内的实际运作效果。

(3) 有效性原则

各个评价指标应在一定统计口径下通过投入产出数据尽可能有效清晰准确地反映安徽省科技企业孵化器的创新水平。因此,在构建安徽省孵化器创新绩效评价指标体系时要充分考虑各指标的有效性。

(4) 可操作性原则

想要得到全面且客观准确的结果,需要纵向和横向地比较不同地区不同时间的具体情况。

因此,要根据可比较以及可操作性选取安徽省孵化器创新绩效指标。指标数据应容易操作,尽量选取具有代表性的指标,减少指标的冗余性,做到既简洁明了又具有可比性。

2.1.2 投入指标

基于上述指标选取原则,本文从人力资源、物力资源、财力资源等三个方面选取投入指标。

(1) 人力资源

人力资源是科研活动重要的投入指标之一,是促进科技创新的中坚力量。科技企业孵化器评价指标中的人力资源一般是指在某一时期内,经营科技企业孵化器的人员为之创造价值所具备的能力。本文选取的人力资源指标是管理机构从业人员数。

(2) 物力资源

科技企业孵化器的物力资源主要包括孵化器自身基础设施建设和提供服务能力两方面的资源投入,孵化器不仅为在孵企业提供基础设施以获取相应的服务收入,还通过开展创新创业教育培训活动等形式对在孵企业进行培训指导,以辅助在孵企业更好成长。本文选取科技企业孵化器场地面积和开展创新创业教育培训活动场次作为物力资源指标。

(3) 财力资源

科技企业孵化器财力资源主要指其拥有的用于正常运营的资本和在资本使用过程中形成的资产,这里主要指资金的投入。本文选取孵化基金总额,其在一定程度上可以代表孵化器对于在孵企业的扶持能力,能够对孵化器创新绩效产生较大影响。

2.1.3 产出指标

基于指标选取原则,本文从经济效益、孵

化效益、社会效益等三个方面选取相应的产出指标。

(1) 经济效益

经济效益是科技企业孵化器的直接回报，体现为孵化器的年度总收入。同时，孵化器培养的企业对促进各地区的社会经济发展也发挥着不容忽视的作用。本文从孵化器总收入和在孵企业总收入两个方面对科技企业孵化器进行绩效分析。

(2) 孵化效益

科技企业孵化器的主要功能就是促进中小企业科技成果转化，助推中小企业茁壮成长，当在孵企业发展到一定规模达到毕业条件后，即可脱离孵化器实现独立发展。因此，孵化器的孵化效益通常以出孵企业或孵化企业毕业率进行评价。本文选取当年毕业企业数，可以直接反映孵化器当年的孵化成效。

(3) 社会效益

科技企业孵化器具有一定的社会效益，社会效益指成功的技术创新活动对社会的积极影响，本文主要考虑孵化器对社会的贡献率。选取孵化器当年知识产权授权数和吸纳应届大学毕业生数，不仅可以反映孵化器的创新水平，还能反映其对当地社会发展的贡献程度。孵化器绩效评价指标体系如表 1 所示。

2.2 评价方法

DEA 是一种运筹学、管理科学与数理经济学交叉研究的方法，用于评价具有可比性的同类型单位的相对有效性，后广泛应用于计算、生产、创新、资源配置等活动的绩效评价中。与传统的效率评价方法相比，DEA 的

表 1 孵化器绩效评价指标体系

类别	一级指标	二级指标
投入指标	人力资源	管理机构从业人员数（人）
	物力资源	孵化器使用面积（平方米） 开展创新创业活动（场次）
	财力资源	孵化基金总额（千元）
产出指标	经济效益	孵化器总收入（千元） 在孵企业总收入（千元）
	孵化效益	当年毕业企业（个）
	社会效益	当年知识产权授权数（件） 吸纳应届大学毕业生数（人）

优势在于它可以处理多输入和多输出的情境，并自动为它们分配最优的权重，避免了主观因素的影响，误差更小、更具客观性，从而进行更全面、更准确的效率评价。

一个决策单元在技术有效的前提下，DEA 是否有效取决于规模效率是否有效。BCC 模型用来评价纯技术效率，增加了假设条件 $\sum \lambda_j$ ，其基本理论为假设决策单元规模报酬是变动的，剔除规模效益因素后进行评估，从而判断投入与产出的比例关系。因此本文选用 DEA-BCC 模型，具体假设如下：

$$\begin{aligned}
 & \min \theta = \min \theta^* \\
 & \text{s.t.} \begin{cases} \sum_{j=1}^n \lambda_j x_{ij} + s^+ \leq \theta x_{jk}, i = 1, 2, 3, \dots, m \\ \sum_{j=1}^n \lambda_j y_{hj} - s^- \geq y_{jk}, h = 1, 2, 3, \dots, s \\ \lambda_j \geq 0 \\ \sum_{j=1}^n \lambda_j = 1 \\ \theta \text{ 无约束} \end{cases} \quad (1)
 \end{aligned}$$

其中， λ_j 表示权向量， x_{ij} 表示第 j 个决策单元第 i 种投入变量， y_{hj} 表示第 j 个决策单元第 h 种产出变量， s^+ 表示投入松弛变量， s^- 表示产出剩余变量， θ 表示整体决策单元对应的效率

值, θ^* 为 θ 的最优解。通过这一模型可以对决策单元是否有效进行判定, 也可以判定决策单元 DEA 是否有效以及 DEA 有效的强弱程度。
 $\theta=1$ 且 s^+ 、 s^- 都为 0 时, 则决策单元强 DEA 有效;
 $\theta=1$ 且 s^+ 或 s^- 不为 0 时, 则决策单元弱 DEA 有效; 若 $\theta < 1$, 则决策单元 DEA 无效。

3 实证分析

本文选取 2019—2022 年安徽省科技企业孵

化器火炬统计年报中全省 16 个地市的投入、产出数据, 以各个地市为基本决策单元, 对其科技企业孵化器绩效进行分析评价研究。

3.1 相关性分析

构建 DEA 评价指标模型要求投入指标和产出指标的数据遵循同向性原则, 本文利用 SPSS 软件对 2022 年相关数据进行 Pearson 相关性检验, 结果如表 2 所示。

表 2 相关性分析

		管理机构从业人员数	开展创新创业活动场次	孵化器使用总面积	孵化器孵化基金总额	孵化器总收入	在孵企业总收入	当年毕业企业数	当年知识产权授权数	吸纳应届大学毕业生数
管理机构从业人员数	皮尔逊相关性	1	0.969**	0.989**	0.923**	0.974**	0.977**	0.977**	0.953**	0.971**
	Sig. (双尾)		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
开展创新创业活动场次	皮尔逊相关性	0.969**	1	0.963**	0.977**	0.989**	0.988**	0.994**	0.987**	0.985**
	Sig. (双尾)	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000
孵化器使用总面积	皮尔逊相关性	0.989**	0.963**	1	0.904**	0.970**	0.967**	0.964**	0.938**	0.970**
	Sig. (双尾)	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000
孵化器孵化基金总额	皮尔逊相关性	0.923**	0.977**	0.904**	1	0.974**	0.971**	0.974**	0.992**	0.945**
	Sig. (双尾)	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000
孵化器总收入	皮尔逊相关性	0.974**	0.989**	0.970**	0.974**	1	0.990**	0.990**	0.988**	0.977**
	Sig. (双尾)	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000
在孵企业总收入	皮尔逊相关性	0.977**	0.988**	0.967**	0.971**	0.990**	1	0.996**	0.990**	0.976**
	Sig. (双尾)	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000
当年毕业企业数	皮尔逊相关性	0.977**	0.994**	0.964**	0.974**	0.990**	0.996**	1	0.991**	0.984**
	Sig. (双尾)	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000
当年知识产权授权数	皮尔逊相关性	0.953**	0.987**	0.938**	0.992**	0.988**	0.990**	0.991**	1	0.969**
	Sig. (双尾)	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000
吸纳应届大学毕业生数	皮尔逊相关性	0.971**	0.985**	0.970**	0.945**	0.977**	0.976**	0.984**	0.969**	1
	Sig. (双尾)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

** . 在 0.01 级别 (双尾), 相关性显著。

分析结果显示, 投入与产出指标之间的相关性均在 1% 的水平上显著, 说明本文选取的投入指标和产出指标之间具备同向性原则, 满足 DEA 分析的前提条件, 可以采用相关指标数据对科技企业孵化器进行绩效评价。

3.2 DEA有效性分析

本文运用 DEAP2.1 软件, 采用投入导向型的 CCR 模型, 计算得到 2019—2022 年安徽省 16 个地市孵化器的综合技术效率、纯技术效率、规模效率和规模收益, 具体如表 3—5 所示。

3.2.1 综合技术效率分析

综合技术效率是指对企业资源配置能力、资源使用效率等各要素能力的全面测量和评估。总体上讲，综合技术效率等于纯技术效率与规模效率的乘积。如果综合技术效率等于 1，即技术效率与规模效率相结合的效率为 1，表明此决策单元的综合技术效率有效。

从表 3 可知，2019—2022 年安徽省科技企业孵化器绩效评价的综合技术效率均值分别为 0.909、0.854、0.891、0.918，4 年总体均值为 0.893，呈小幅波动增长的趋势，但整

体较为稳定。四年来，全省 16 个市中有 12 个市的综合技术效率在 64.9% 以上，说明大部分地市的科技企业孵化器的绩效还是比较理想的，特别是合肥、亳州、淮南、滁州、马鞍山、铜陵、池州等 7 个市连续 4 年综合技术效率为 1，占全省的 43.75%，而且这 7 个市的规模收益也处于不变状态，说明这 7 个市科技企业孵化器绩效全部有效，应当继续保持。黄山、阜阳、蚌埠等 9 个市孵化器的资源配置能力和使用效率需要进一步提升，以实现最佳投入产出效益。

表 3 2019—2022 年安徽省 16 个市科技企业孵化器综合技术效率

地市	2019 年		2020 年		2021 年		2022 年		效率均值
	效率值	规模收益	效率值	规模收益	效率值	规模收益	效率值	规模收益	
合肥	1	-	1	-	1	-	1	-	1
淮北	0.858	drs	0.525	irs	1	-	0.997	irs	0.845
亳州	1	-	1	-	1	-	1	-	1
宿州	1	-	0.676	irs	0.649	drs	1	-	0.831
蚌埠	0.392	irs	0.298	drs	0.458	drs	0.589	irs	0.434
阜阳	0.641	irs	0.748	irs	1	-	0.555	irs	0.736
淮南	1	-	1	-	1	-	1	-	1
滁州	1	-	1	-	1	-	1	-	1
六安	1	-	1	-	0.867	drs	1	-	0.967
马鞍山	1	-	1	-	1	-	1	-	1
芜湖	0.926	drs	0.806	drs	0.844	drs	1	-	0.894
宣城	1	-	1	-	-	无	1	-	≤ 1
铜陵	1	-	1	-	1	-	1	-	1
池州	1	-	1	-	1	-	1	-	1
安庆	0.944	drs	0.846	drs	1	-	1	-	0.948
黄山	0.781	drs	0.767	drs	0.548	irs	0.555	irs	0.663
均值	0.909	无	0.854	无	0.891	无	0.918	无	0.893

从表 3 可以看出，2019—2022 年安徽省各市的科技企业孵化器的综合技术效率均值保持在 0.85~0.92 之间，总体表现良好。从各地市来

看，合肥市等均值最高为 1，蚌埠市均值最低为 0.434，以效率值和排名变化为依据，大致细化成三个梯队：第一梯队包括合肥、亳州、淮

南、滁州、马鞍山、铜陵、池州，效率值为1且规模收益不变；第二梯队包括宣城、淮北、宿州、六安、芜湖、安庆，效率值在0.75~1之间；第三梯队效率值在0.75以下，包括蚌埠、阜阳、黄山。分析发现，安徽省不同地市孵化器的综合技术效率存在显著差异，不同地市间孵化器的互补合作和资源共享仍有待加强。其中，蚌埠市孵化器的规模效率呈小幅波动增长的趋势，4年总体均值为0.853，但其纯技术效率整体偏低，总体均值仅为0.510，说明蚌埠市孵化器的孵化效率和成功率存在较大的提升空间；阜阳市孵化器在2021年综合技术效率为1，但在2022年下降至0.555，观察指标，2022年阜阳市加强了孵化器的资源投入，但其产出不足；黄山市孵化器的综合技术效率在4年间呈下降趋势，2021年以来黄山市孵化器的资源投入不断加大，尤其是孵化基金总额，2022年数量较2021年提升了4倍多，但在孵企业总收入反而下降，黄山市孵化器的专业化服务能力与孵化成功率仍需加快提高。

3.2.2 纯技术效率分析

纯技术效率指企业受到管理、技术水平等因素影响的生产效率，是DMU在规模效率最佳的情况下，投入要素的生产效率。纯技术效率等于1，表明在现有技术水平下，它的资源利用是有效率的。

从表4可以看出，2019—2022年安徽省各市科技企业孵化器纯技术效率平均值分别为0.938、0.888、0.931、0.937，4年总体均值为0.924。四年来，纯技术效率均达到DEA有效的市有9个，分别是合肥、亳州、淮南、滁州、六安、马鞍山、芜湖、铜陵、池州，占总体样

本的56.25%。总的来看，全省科技企业孵化器管理水平较高，但蚌埠、阜阳、黄山等市科技企业孵化器纯技术效率明显偏低，且纯技术效率均小于规模效率，说明这些市综合技术效率很大程度上受纯技术效率的制约，而孵化器服务体系不够完善则是影响其绩效的主要因素。

表4 2019—2022年安徽省16个市科技企业孵化器纯技术效率

地市	2019年	2020年	2021年	2022年	均值
合肥	1	1	1	1	1
淮北	0.921	0.622	1	1	0.886
亳州	1	1	1	1	1
宿州	1	0.771	0.656	1	0.857
蚌埠	0.399	0.419	0.614	0.606	0.510
阜阳	0.806	0.76	1	0.658	0.806
淮南	1	1	1	1	1
滁州	1	1	1	1	1
六安	1	1	1	1	1
马鞍山	1	1	1	1	1
芜湖	1	1	1	1	1
宣城	1	1	-	1	≤1
铜陵	1	1	1	1	1
池州	1	1	1	1	1
安庆	1	0.85	1	1	0.963
黄山	0.878	0.794	0.695	0.723	0.773
均值	0.938	0.888	0.931	0.937	0.924

3.2.3 规模效率分析

规模效率是指受到企业规模因素影响的生
产效率，能真正体现其与最佳生产水平的差异。
规模效率等于1，表明孵化器规模处于规模收
益不变的最佳状态，对于孵化器的投入既不浪
费也不紧缺，投入产出规模恰到好处。

从表5可以看出，2019—2022年安徽省

各市科技企业孵化器的规模效率均值分别为 0.967、0.949、0.949、0.974，4 年总体均值为 0.96。四年来，规模效率均处于 DEA 有效状态的市有 7 个，分别是合肥、亳州、淮南、滁州、马鞍山、铜陵、池州，约占全省的 43.75%，而黄山、蚌埠、阜阳等市科技企业孵化器的规模效率明显低于总体平均水平，说明这些地市的孵化器资源投入存在冗余，缺乏有效的管理机制，无法充分适应市场需求，导致孵化器的规模效率难以提升。

表 5 2019—2022 年安徽省 16 个市科技企业孵化器
规模效率

地市	2019 年	2020 年	2021 年	2022 年	均值
合肥	1	1	1	1	1
淮北	0.931	0.844	1	0.997	0.943
亳州	1	1	1	1	1
宿州	1	0.877	0.99	1	0.967
蚌埠	0.982	0.712	0.745	0.971	0.853
阜阳	0.796	0.985	1	0.844	0.906
淮南	1	1	1	1	1
滁州	1	1	1	1	1
六安	1	1	0.867	1	0.967
马鞍山	1	1	1	1	1
芜湖	0.926	0.806	0.844	1	0.894
宣城	1	1	-	1	≤1
铜陵	1	1	1	1	1
池州	1	1	1	1	1
安庆	0.944	0.996	1	1	0.985
黄山	0.889	0.966	0.788	0.768	0.853
均值	0.967	0.949	0.949	0.974	0.960

4 研究结论及对策建议

4.1 研究结论

本文利用 DEA 模型对 2019—2022 年安徽

省 16 个地市的科技企业孵化器的综合技术效率、纯技术效率和规模效率进行测度，比较了不同地市之间的效率差异，并结合实地调研，分析得出安徽省科技企业孵化器在发展过程中存在以下几个问题：

（1）孵化器的资源整合能力还不够强。尽管近年来黄山等市孵化器对孵化基金或孵化器面积等资源上加大投入，但由于缺少与各类专业服务机构、龙头企业、高校院所开展的深度合作，导致孵化器载体平台在技术、市场、人才资源的链接能力偏弱，没有足够的资源向在孵企业输送。如阜阳市孵化器签约中介机构数量较少，说明阜阳市孵化器缺乏与外部资源的有效对接和合作，资源利用效率受到了极大的限制，导致这部分孵化器在创新链、产业链、供应链的资源投入相对匮乏。

（2）孵化器的专业化服务能力有所欠缺。尽管近年来安徽省专业型省级以上孵化器的建设速度不断加快，但就整体结构来看，全省省级以上专业型孵化器占比不足三成，大部分是综合型孵化器，以提供办公场地、咨询服务、融资支持为主要功能。如蚌埠市建有 8 家省级以上孵化器，其中只有 1 家属于专业型孵化器。阜阳市建有 6 家省级以上孵化器，其中只有 1 家属于专业型孵化器。以在综合技术效率中排名第一的合肥市为例，其省级以上孵化器中专业型孵化器占比超过三成。围绕主导产业共性需求和技术难点，适应市场需求培养专业化服务的能力是当下综合型孵化器所欠缺的。

（3）孵化器的管理机制不够完善。一是部分地市孵化器的运营模式较为单一，以政府主导型为主。综合技术效率排名处于第一梯队

的地市,省级以上孵化器中民营企业占比达58.7%,然而第三梯队的地市其省级以上孵化器中民营企业占比仅为37.5%,说明以市场为导向,实现利益共享的孵化体系是孵化器服务功能转变的必然趋势。二是孵化器发展水平区域差异明显,安徽省内不同地市的孵化器间缺少资源共享服务,没有形成组织创新、资源共享的联动发展机制。三是部分地市孵化器孵化创业项目的风险投资制度体系不够健全,如黄山市近2年虽然孵化基金总额大幅提升,然而在孵企业总收入不增反降,说明孵化器投资团队的专业化与体系化建设迫在眉睫。

4.2 对策建议

为提升科技企业孵化器市场化管理运营能力,提高孵化效率,加快推动科技服务业支撑新质生产力发展,针对上文提到的问题,提出以下几点对策建议:

(1) 加强资源整合和共享能力。一方面,孵化器应深入挖掘和整合内外部资源,全面梳理自身的资金、场地、设备、技术、人才等核心资源,深入了解自身已有的资源和能力,明确哪些资源可以直接用于支持在孵企业,哪些资源需要进一步开发或优化,不断增强资源挖掘意识、提升资源整合能力。另一方面,孵化器应积极寻找外部资源,加强与当地政府部门、龙头企业、行业协会、金融机构、高校及科研机构等建立紧密的联系与稳定的合作,可与其他机构合作搭建线上线下资源共享平台,方便在孵企业之间的交流与合作。例如,阜阳市政府管理部门应加快建立安徽科技大市场阜阳分市场,发挥平台优势,集聚优秀科技中介,助

力孵化器引入更多优质资源。

(2) 完善创新孵化服务体系。首先,孵化器应丰富和深化服务内容,尤其是蚌埠市孵化器,除了提供基本的办公场地和咨询服务外,要进一步拓展服务范围,包括创业辅导、资源对接、金融支持等各个方面,以满足在孵企业多元化需求。可以探索将人工智能新技术应用到孵化过程中,通过数据分析优化孵化流程、提供个性化服务,推动创业孵化效能提升。此外,要进一步加强孵化器服务团队专业能力建设,适度增加团队人员数量,引入资深技术经理人团队以及创业导师队伍,定期组织创业指导、行业分析、投融资知识、法律法规等方面的专业培训,提高服务团队的专业素养。同时,建立团队激励机制,鼓励团队成员参与孵化项目的跟投和创业投资活动,形成利益共同体。

(3) 健全孵化器管理机制。孵化器首先要明晰自身战略定位和目标,应根据当地的产业特色、市场需求和资源优势,明确核心业务领域,避免资源分散和战略模糊。例如,黄山市孵化器应紧密结合黄山的特色产业,如旅游、文化、生态农业等,聚焦当地优势资源禀赋与产业基础,并将孵化器的服务内容与市场需求紧密结合,提供有针对性的孵化服务。其次,孵化器可以尝试在运营管理中引入市场化机制,例如,针对淮北市孵化器多年投资收入为0的问题,可以通过探索股权投资、增值服务、品牌输出等多种盈利渠道,提高孵化器自身盈利能力和运营效率,进一步增强孵化器的市场竞争力和可持续发展能力。最后,孵化器的评价体系需适时更新,通过动态的绩效评价管理,在全省

孵化器中树榜样、找差距、补不足，引导全省孵化器形成提质增效的良好局面。

参考文献

- [1] BECKER B, GASSMANN O. Corporate Incubators: Industrial R&D and What Universities can Learn from them[J]. Journal of Technology Transfer, 2006, 31(4): 469-483.
- [2] HACKETT S M, DILTS D M. A Systematic Review of Business Incubation Research[J]. The Journal of Technology Transfer, 2004, 29(1): 55-82.
- [3] SAIDI-MEHRABAD M, SADRAHADI M R, MOHAMMADIAN I. A new method to fuzzy modeling and its application in performance evaluation of tennants in incubators[J]. International Journal of Advanced Manufacturing Technology, 2008, 37(1-2): 191-201.
- [4] 夏星, 张珍. 基于三阶段 DEA 模型的科技企业孵化器运行效率测度 [J]. 统计与决策, 2020, 36(24): 156-160.
- [5] 吴文清, 马赛翔, 刘晓英. 科技企业孵化器集聚及效率与空间关联研究 [J]. 天津大学学报 (社会科学版), 2016, 18(3): 206-210.
- [6] 冯金余. 科技企业孵化器的创新驱动效应研究 [J]. 科研管理, 2017, 38(11): 38-47.
- [7] 关成华, 邱英杰, 袁祥飞. 财政政策工具与中国科技企业孵化器效率 [J]. 财政研究, 2018(12): 48-61.
- [8] 何子佳, 刘琦. 科技企业孵化器绩效评价发展研究 [J]. 财务管理研究, 2022(9): 1-5.
- [9] 吕明慧. 河南省科技企业孵化器创新绩效研究 [D]. 郑州: 郑州航空工业管理学院, 2023.
- [10] 郭会斌, 薛婷婷, 何苗, 等. 基于京津冀协同发展的孵化器绩效实证研究 [J]. 科技进步与对策, 2022, 39(5): 22-29.
- [11] 姚梦思. 孵化器创新驱动能力评价及对区域创新绩效的空间效应分析 [D]. 太原: 中北大学, 2023.
- [12] 王楠. 我国科技企业孵化器绩效评价的实证研究 [D]. 大连: 辽宁师范大学, 2023.
- [13] 陈立广. 我国科技企业孵化器的运行效率评价及预测研究 [D]. 天津: 天津职业技术师范大学, 2023.



开放科学
(资源服务)
标识码
(OSID)

山东省科技期刊学术影响力动态研究

韦忠明

山东省泰山学院图书馆 泰安 271000

摘要: [目的/意义] 学术影响力是期刊的生命, 科技期刊学术影响力评价结果不但影响科研人员的投稿取向, 也会对科技期刊的发展产生重要的导向作用。研究可为科技期刊管理部门提供一定资料, 为提高期刊的学术影响力及制定今后发展策略提供参考, 也可为图书馆馆藏核心期刊的确立奠定基础。[方法/过程] 以中国科学技术信息研究所研究发布的《中国科技期刊引证报告(扩刊版)》(2016—2020年)中13个文献计量指标为基本元数据, 采用文献计量学方法, 综合分析山东省科技期刊的学术影响力动态变化情况。[局限] 缺少对研究问题所在领域相关研究的系统梳理; 对其他省份或者国家的参考价值需要进一步拓展。[结果/结论] 山东省科技期刊的学术影响力不断提高, 期刊的来源指标及被引指标大都呈逐年增长趋势; 山东省各学科科技期刊年均发文量整体呈现减少态势; 天文与地球科学、农林科学和交通运输类期刊的文献老化速度较慢; 环境与安全科学、生物科学和医药卫生类期刊的h指标均值最大, 交通运输类期刊在多项指标上均列末位; 各种期刊在不同指标显示出清晰的学术层级性, 影响力不均衡。

关键词: 山东省; 科技期刊; 学术影响力; 被引指标; 来源指标

中图分类号: G35

Dynamic Study on Academic Influence of Sci-tech Journals in Shandong Province

WEI Zhongming

Library of Shandong Taishan University, Taian 271000, China

Abstract: [Purpose/Significance] Academic impact is vital for the vitality of journals. The evaluation results of academic impact in science and technology journals not only influence the submission preferences of researchers but also play a significant guiding role in the development of these journals. The results of this study can provide certain data for the management departments of science and technology journals, serving as a reference for improving the academic impact of journals and formulating future development strategies, and lay a foundation for the establishment of core journals in library collections. [Methods/Processes]

作者简介 韦忠明(1969-), 学士, 副研究馆员, 主要研究方向为文献计量与期刊评价, E-mail: tsxywzm69@163.com。

引用格式 韦忠明. 山东省科技期刊学术影响力动态研究[J]. 情报工程, 2024, 10(4): 84-95.

Thirteen bibliometric indicators from the “Chinese Science and Technology Journal Citation Reports (Expanded Edition)” (2016-2020) released by the Institute of Scientific and Technical Information of China were used as basic metadata. By employing bibliometric methods, the dynamic changes in the academic impact of science and technology journals in Shandong Province were comprehensively analyzed. [Limitations] There is a lack of systematic review of relevant research in the field of the research problem. The reference value of other provinces or countries needs further exploration. [Results/Conclusions] The academic impact of science and technology journals in Shandong Province continues to improve, with most journals’ source and citation indicators showing an increasing trend year by year. However, the overall annual publication volume of science and technology journals in Shandong Province shows a decreasing trend. Journals in different disciplines exhibit distinct differences in academic impact, such as the slower aging rate of articles in journals related to astronomy and earth sciences, agricultural and forestry sciences, and transportation; the highest mean h-index values in journals related to environmental and safety sciences, biological sciences, and medicine and health; and transportation journals ranking lowest in multiple indicators. Various journals demonstrate clear academic hierarchy in different indicators, indicating uneven impact.

Keywords: Shandong Province; Science and Technology Journal; Academic Influence; Cited Index; Source Indicators

引言

学术影响力是期刊的生命,就科技期刊而言,学术影响力是指某一时期内期刊对其所处领域科研活动的影响深度及广度^[1]。科技期刊学术影响力评价结果不但影响科研人员的投稿取向,也会对科技期刊的发展产生重要的导向作用^[2]。目前关于我国科技期刊学术影响力的研究主要涉及学科维度、单种期刊和区域维度三个方面的研究^[3],如:白雪娜等^[4]的广东省农业类科技期刊学术影响力动态评估,刘青^[5]的提升《河南水利与南水北调》学术影响力的探索,刘艳华等^[6]的江苏省科技期刊学术影响力的影响因素研究分别从这三个维度进行了剖析。而这三个维度中区域维度方面的期刊学术影响力研究最少,关于山东省科技期刊学术影响力的研究更是尚未开展,因此本文拟借助《中国科技期刊引证报告》(扩刊版)所涉及的各类元数据,以山东省科技期刊为研究对象,综

合分析山东省科技期刊的学术影响力状况。本研究缺少对研究问题所在领域相关研究的系统梳理,以及对其他省份或者国家参考价值的分析,但仍可为科技期刊管理部门提供一定资料,为提高期刊的学术影响力及制定今后发展策略提供参考,也可以为图书馆馆藏核心期刊的确立奠定基础。

1 资料来源及方法

为全面、准确、公正、客观地评价和利用期刊,《中国科技期刊引证报告》(扩刊版)在与国际评价体系保持一致的基础上,选取了18项计量指标,这些指标包括期刊被引指标和来源期刊指标,期刊被引指标包括总被引频次和h指标等9项指标,来源期刊指标包括来源文献量和引用半衰期等9项指标^[7]。本研究数据的获取基于2016—2020年版《中国科技期刊引证报告》(扩刊版)中最重要的5项来源指

标和6项被引指标,采用文献计量学方法,按年度对山东省主办的各类科技期刊的元数据进行汇总分析研究。

2 结果与分析

2.1 山东省科技期刊来源指标分析

2.1.1 年度发文量分析

2015—2019年山东省科技期刊各学科年均发文量见表1,发文量按5年发文平均量降序排列(下同)。由表1可知工业技术类期刊的5年年均发文量位列第1,主要是《山东工业技术》《信息技术与信息化》《山东煤炭地质科技》和《山东化工》刊登的科技论文较多。医药卫生类排名第2,《山东医药》和《齐鲁护理杂志》发文量最大,5年平均发文量分别大于1600篇/年和1400篇/年,其余大部分医学期刊平均发文量小于300篇/年。农业科学类期刊发文量比较平均,5年平均发文量大于400篇/年以上的期刊有《山东农业工程学院学报》《山东农业科学》和《山东畜牧兽医》,5年平均发文量小于200篇/年的期刊有6种,其中《花生学报》平均发文量最少,只有51篇/年。天文与地球科学类期刊5年平均发文量高于200篇/年的只有《海洋科学》和《中国海洋大学学报》2种期刊,《海洋气象学报》和《海洋科学进展》平均发文量仅57篇/年和53篇/年,这两种期刊是年均发文量最少的天文与地球科学类期刊。交通运输类期刊发文量普遍较低,超过200篇/年的刊物只有《山东交通科技》,除此之外高于100篇/年的刊物也只有《铁道

车辆》1种,《国外铁道车辆》年均发文55篇,为交通运输类期刊发文最少的刊物。自然科学总论大部分为学报,发文量较为平均,基本在200篇/年以下。山东省的生物科学、数理科学和化学、环境与安全科学以及航空航天类科技期刊较少,只有《中国病原生物学杂志》《中国人口·资源与环境》发文量高于200篇/年,因此它们的总发文量也是最少的。

2015—2019年山东省科技期刊共发文169513篇,其中工业技术类期刊发文量排名第1,发文量高达74650篇,占发文总量的44%。排名第2的是医药卫生类期刊,共发文54403篇,占比32.1%。农林科学类期刊发文17892篇排名第3,占比10.6%。这三种期刊合计146945篇,占山东省科技期刊5年发文总量的86.7%,其余7个学科合计发文22568篇,占发文总量的13.3%。从山东省科技期刊各学科年均发文量5年变化分析,年均发文量由216.3篇/刊降至210篇/刊,整体呈现减少态势。

2.1.2 篇均引文量分析

2015—2019年山东省科技期刊各学科篇均引文量见表2。由表2可知,环境与安全科学、天文与地球科学、生物科学、数理科学和化学等基础性学科的篇均引文普遍高于工业技术、交通运输等工程类学科,该现象说明不同学科之间的引文状况差异较大,可能是由于工程类学科参考来源于试验和实践较多。从篇均引文量5年变化分析,山东省科技期刊在2015—2019年整体呈现上升趋势,篇均引文量从2015年的13.1次/篇增至2019年的19.4次/篇,呈现良好的发展态势。

表 1 2015—2019 年山东省科技期刊发文统计

序号	学科名称	2015 年 / (篇 / 刊)	2016 年 / (篇 / 刊)	2017 年 / (篇 / 刊)	2018 年 / (篇 / 刊)	2019 年 / (篇 / 刊)	5 年平均 / (篇 / 刊)	总发文 量 / 篇
1	工业技术	448.3	455.3	446.5	427.5	491.5	447.0	74650
2	医药卫生	374.8	370.0	343.6	323.6	287.4	340.0	54403
3	生物科学	290.0	283.0	296.0	329.0	328.0	305.2	1526
4	农林科学	266.6	287.8	263.0	230.2	290.2	267.0	17892
5	环境与安全科学	158.0	257.0	152.5	126.5	128.5	154.2	1388
6	天文与地球科学	141.1	129.1	129.4	135.3	119.7	131.1	7736
7	交通运输	128.3	114.5	112.7	123.8	166.7	125.1	3252
8	数理科学和化学	119.5	115.0	115.0	110.5	109.0	113.8	1138
9	自然科学总论	116.6	110.1	109.6	96.0	110.3	108.9	7076
10	航空航天	120.0	100.0	87.0	76.0	69.0	90.4	452
	总计平均值	216.3	222.2	205.5	197.8	210.0	208.3	1695.3

表 2 2015—2019 年山东省科技期刊篇均引文量统计(单位: 次 / 篇)

序号	学科名称	2015 年	2016 年	2017 年	2018 年	2019 年	5 年平均
1	环境与安全科学	20.4	24.6	26.2	33.7	35.6	28.5
2	天文与地球科学	20.7	25.5	28.1	29.7	31.5	27.0
3	生物科学	19.4	20.7	21.7	22.2	21.1	21.0
4	数理科学和化学	13.6	21.3	19.5	21.0	24.0	20.0
5	航空航天	9.0	18.6	18.5	19.0	19.2	16.9
6	自然科学总论	12.8	15.9	15.1	17.3	18.5	16.2
7	医药卫生	11.9	13.5	14.5	15.9	17.1	14.5
8	农林科学	9.9	11.6	10.9	13.9	12.1	11.6
9	工业技术	8.3	9.3	9.5	10.5	10.8	9.7
10	交通运输	4.9	5.9	6.3	7.6	3.9	5.9
	总计平均值	13.1	16.7	17.0	19.1	19.4	17.1

具体到各学科期刊数据, 环境与安全科学和天文与地球科学篇均引文量最高, 大部分在 20~30 次 / 篇, 其中英文期刊 *Chinese Journal of Population and Environment* 篇均引文最高, 为 31.2 次 / 篇。数理科学和化学和生物科学类期刊篇均引文大部分在 15~20 次 / 篇, 其中《复杂系统与复杂性科学》和《中国病原生物学杂志》篇均引文排在前 2 位, 分别为 23.2 次 / 篇和 21 次 / 篇。航空航天和自然科学总论篇均引文大部分不到 20 次 / 篇, 篇均引文超过 20 次 /

篇的只有《鲁东大学学报》《中国石油大学学报》和《中国海洋大学学报》, 分别为 20.1 次 / 篇、24.3 次 / 篇和 29.3 次 / 篇。医药卫生、农林科学和工业技术类期刊篇均引文量差异较大, 篇均引文量排名前 3 的是《渔业科学进展》《中国海洋药物》和《中国烟草科学》, 篇均引文量分别为 27.8 次 / 篇、24.6 次 / 篇和 24 次 / 篇, 篇均引文量最少的期刊是《烟台果树》《氮肥技术》和《山东水利》, 篇均引文分别只有 1.4 次 / 篇、1.3 次 / 篇和 0.8 次 / 篇。交通运输类期

刊只有《山东交通学院学报》，篇均引文达到18.7次/篇，其余大部分期刊篇均引文量都较低，还有较大的提升空间。

2.1.3 篇均作者数量分析

2015—2019年山东省科技期刊各学科发表论文篇均作者统计见表3。由表3可知，2015—2019年山东省科技期刊各学科论文篇均作者数由3.49人/篇增至3.80人/篇，说明山东省科研工作者越来越注重科研协作。虽然各学科篇均作

者数大体呈增长态势，但在各学科间篇均作者数却存在明显差异。5年平均篇均作者数生物科学类期刊为5.46人/篇，排名第1。天文与地球科学和农林科学类期刊相差不大，篇均作者数量为4.5人/篇左右。医药卫生类排名第4，篇均作者数为3.82人/篇，排在其后的航空航天、自然科学总论、工业技术和环境与安全科学类期刊篇均作者数比较平均，大约为3人/篇，交通运输类期刊篇均作者数较低，为2.27人/篇左右。

表3 2015—2019年山东省科技期刊篇均作者统计(单位：人/篇)

序号	学科名称	2015年	2016年	2017年	2018年	2019年	5年平均
1	生物科学	5.30	5.20	5.50	5.90	5.40	5.46
2	天文与地球科学	4.43	4.59	4.73	4.79	4.83	4.67
3	农林科学	3.94	4.17	4.31	4.82	4.65	4.36
4	医药卫生	3.74	3.72	3.81	3.86	3.97	3.82
5	数理科学和化学	3.55	3.55	3.70	3.60	3.75	3.63
6	航空航天	3.20	3.10	3.40	3.30	3.40	3.28
7	自然科学总论	2.78	3.08	3.11	3.19	3.46	3.16
8	工业技术	2.99	3.09	3.03	3.15	3.00	3.05
9	环境与安全科学	2.75	2.70	2.90	3.15	3.20	2.94
10	交通运输	2.18	2.30	2.32	2.20	2.37	2.27
	总计平均值	3.49	3.55	3.68	3.80	3.80	3.66

进一步分析期刊数据发现，篇均作者大于7人/篇的期刊仅1种，即《中国烟草科学》为7.16人/篇，篇均作者少于7人/篇且超过5人/篇的期刊有15种，分别为《花生学报》《渔业科学进展》《山东农业科学》《中国病原生物学杂志》《中国海洋药物》《中国动物检疫》《中华行为医学与脑科学》《海洋地质与第四纪地质》《中华肿瘤防治杂志》《海洋与湖沼》《中国石油大学学报》《中国矫形外科杂志》《山东大学学报医学版》和*Journal of Ocen University of China*以及*Chinese Journal of Occanology and Limnology*。篇均作者少于5人/篇且多于3人/

篇的期刊共66种，其中篇均作者较多的期刊有《海洋科学》《海洋科学进展》《中国蔬菜》《海洋地质前沿》《油气地质与采收率》《生物医学工程研究》《山东科学》《中外葡萄与葡萄酒》《海洋湖沼通报》等，这些期刊篇均作者多于4.5人/篇。剩余的24种期刊篇均作者均少于3人/篇。

2.1.4 篇均基金论文比分析

2015—2019年，山东省科技期刊各学科篇均基金论文比见表4。由表4可知，山东省科技期刊的学科篇均基金论文比差异较大。自然科学总论、环境与安全科学、生物科学、数理

科学和化学、天文与地球科学类期刊的篇均基金论文比5年平均

篇均基金论文比均小于0.49，交通运输类期刊篇均基金论文比最少，仅为0.1427。

表4 2015—2019年山东省科技期刊篇均基金论文比统计(单位：个/篇)

序号	学科名称	2015年	2016年	2017年	2018年	2019年	5年平均
1	自然科学总论	0.6836	0.7622	0.7548	0.8042	0.7747	0.7571
2	环境与安全科学	0.7580	0.9100	0.7740	0.7000	0.6900	0.7504
3	生物科学	0.6100	0.6000	0.7200	0.6800	0.6300	0.6480
4	数理科学和化学	0.6100	0.5450	0.6400	0.6000	0.6300	0.6050
5	天文与地球科学	0.7166	0.6952	0.4725	0.5835	0.4925	0.5878
6	农林科学	0.4745	0.4341	0.4725	0.5835	0.4924	0.4897
7	航空航天	0.4400	0.4300	0.3100	0.4100	0.4500	0.4080
8	工业技术	0.3600	0.3742	0.3709	0.3894	0.4033	0.3800
9	医药卫生	0.2829	0.3115	0.3609	0.3847	0.3974	0.3472
10	交通运输	0.1323	0.1528	0.1410	0.1834	0.0787	0.1427
	总计平均值	0.5068	0.5215	0.5017	0.5319	0.5039	0.5141

进一步研究发现，5年篇均基金论文比值大于等于0.9的期刊有6种，以高校学报居多，分别是《聊城大学学报》《济南大学学报》《中国石油大学学报》《中国海洋药物》《中国人口·海洋与环境》《山东大学学报理学版》，6种期刊的5年篇均基金论文比分别为1、0.982、0.944、0.918、0.906、0.9，高校学报5年篇均基金论文比大于0.5的有25种，占全部学报总数的64.1%。5年篇均基金论文比为0.5~0.9（不包含）的期刊有43种，《鲁东大学学报》《中国海洋大学学报》《复杂系统与复杂性科学》《青岛农业大学学报》《海洋与湖沼》《齐鲁工业大学学报》这些期刊的5年篇均基金论文比均大于0.89，排名比较靠前。5年篇均基金论文比小于0.5的期刊有66种，排名比较靠后的期刊有《煤矿现代化》《居业》《山东冶金》《山东煤炭科技》和《氮肥技术》，这些期刊的5年篇均基金论文比均小于0.035。从整体来看，2015—2019年山东省科技期刊各学科篇均基金论文比变

化不大，呈平稳发展态势。

2.1.5 论文选出比分析

论文选出比是剔除非论文文献后选出的文献数与期刊发表文献数之比，主要体现期刊的规范程度。由表5可见，除数理科学和化学外其余所有学科期刊论文选出比均大于0.91，说明山东省科技期刊总体学术规范程度较高。环境与安全科学类期刊论文选出比排名第1，紧随其后的是生物科学、自然科学总论和航空航天类期刊，这些学科的期刊论文选出比都大于0.95，刊载的论文总体质量较高。医药卫生、天文与地球科学、工业技术和农林科学类期刊论文选出比在0.93~0.95之间，彼此相差不大，交通运输和数理科学和化学类期刊论文选出比排名最后。

进一步研究发现，论文选出比为1的有2种期刊，分别是《中国人口·资源与环境》和《山东师范大学学报》，论文选出比大于等于0.98的期刊有29种，《医学检验与临床》《山东水利》《海洋科学》《山东医学高等专科学校学

报》《医学影像学杂志》《山东农业大学学报》
《山东科技大学学报》《济宁医学院学报》《济
南大学学报》和《泰山医学院学报》等期刊的

论文选出比较大，排名比较靠前，这些期刊以
高校学报居多，证明高校学报刊载的论文总体
质量较高。

表 5 2015—2019 年山东省科技期刊论文选出比统计(单位：%)

序号	学科名称	2015 年	2016 年	2017 年	2018 年	2019 年	5 年平均
1	环境与安全科学	99.00	100.00	96.50	98.50	98.50	98.33
2	生物科学	97.00	98.00	97.00	99.00	98.00	97.80
3	自然科学总论	97.67	96.82	97.08	94.27	96.58	96.52
4	航空航天	97.00	97.00	94.00	97.00	92.00	95.40
5	医药卫生	94.09	93.81	97.76	96.74	93.72	94.67
6	天文与地球科学	93.42	94.83	93.75	94.58	96.45	94.07
7	工业技术	91.56	92.72	92.76	93.17	94.65	93.83
8	农林科学	94.14	90.00	94.67	93.46	94.09	93.25
9	交通运输	91.67	88.00	91.00	94.80	94.67	91.62
10	数理科学和化学	76.50	79.50	81.50	98.00	95.00	86.10
	总计平均值	93.21	93.07	93.60	95.95	95.37	94.16

2.2 山东省科技期刊被引指标分析

2.2.1 总被引频次分析

2015—2019 年，山东省科技期刊各学科总被引频次见表 6。由表 6 可知，各学科期刊的总被引频次差异较大，环境与安全科学、医药卫生、生物科学、农林科学、天文与地球科学和工业技术科学类期刊总被引频次排名靠前。从篇均总被引频次 5 年变化分析，山东省科技期刊在 2015—2019 年总被引频次由 1446.2 次 / 篇增至 1778.3 次 / 篇，呈逐年增长态势。可见山东省科技期刊的影响力不断扩大，其中环境与安全科学和工业技术类期刊被引频次增长较快，5 年分别增长了 1704 次 / 篇和 493.3 次 / 篇。

进一步分析发现，5 年总被引频次均值大于 2000 次 / 篇的期刊共 25 种，其中《山东医药》《齐鲁护理杂志》和《山东工业技术》5 年平均总被引频次高达 18609 次 / 篇、18510.8 次 /

篇、10653.8 次 / 篇，位列前 3 名。《中国人口·资源与环境》《中国矫形外科杂志》和《医学影像学杂志》5 年总被引频次均值大于 5000 次 / 篇，位列 4~6 位。《中华肿瘤防治杂志》《中华行为医学与脑科学》《中国蔬菜》《腹腔镜外科杂志》《现代妇科进展》《居业》《山东农业科学》《油气地质与采收率》《山东中医杂志》和《海洋与湖沼》5 年总被引频次均值大于 2500 次 / 篇，位列 7~16 位。《中国现代普通外科进展》《海洋科学》《中国烟草科学》《泰山医学院学报》《中国海洋大学学报》《中国病原生物学杂志》《家禽科学》《山东化工》和《工程塑料应用》5 年总被引频次均值大于 2000 次 / 篇，分列 17~25 位。

需要说明的是，虽然总被引频次可以显示该期刊被使用和受重视的程度以及在科学交流中的地位和作用，但它与办刊时间、载文规模以及学科领域有较大关系。

表 6 2015—2019 年山东省科技期刊总被引频次统计(单位：次 / 篇)

序号	学科名称	2015 年	2016 年	2017 年	2018 年	2019 年	5 年平均
1	环境与安全科学	4073.0	4853	5181.0	5042.0	5777.0	4985.3
2	医药卫生	2775.6	3000.2	2906.6	3025.8	3077.4	2955.2
3	生物科学	2246.0	2191.0	2217.0	2404.0	2277.0	2267.0
4	农林科学	1431.1	1296.1	1291.8	1545.4	1775.8	1450.5
5	天文与地球科学	1262.8	1280.0	1288.8	1245.6	1196.3	1243.5
6	工业技术	876.0	1061.7	1190.1	1355.0	1369.3	1176.3
7	数理科学和化学	679.0	760.5	768.0	800.5	741.0	749.8
8	自然科学总论	409.8	390.7	382.3	294.7	664.2	456.4
9	航空航天	426.0	442.0	404.0	390.0	459.0	424.2
10	交通运输	282.5	278.7	285.0	335.2	446.3	311.2
	总计平均值	1446.2	155.4	1591.5	1643.8	1778.3	1656.2

2.2.2 他引次数分析

他引次数是指引用被评价期刊的期刊数，它能反映被评价期刊被使用的范围^[7]，可以减少期刊虚假自引带来的统计偏差。2015—2019 年,山东省科技期刊各学科刊均他引次数见表 7。由表 7 可知，山东省科技期刊他引次数呈明显

的学科差异性，各学科刊均他引次数与表 6 的总被引频次排序基本一致。从篇均他引次数 5 年变化分析，山东省科技期刊在 2015—2019 年他引次数由 347.9 次 / 篇增至 383.6 次 / 篇，总体呈现增长态势，可见山东省科技期刊的学术影响力在不断扩大。

表 7 2015—2019 年山东省科技期刊刊均他引次数统计(单位：次 / 篇)

序号	学科名称	2015 年	2016 年	2017 年	2018 年	2019 年	5 年平均
1	环境与安全科学	749.5	857.5	802.5	806.0	873.0	817.7
2	生物科学	483.0	542.0	504.0	523.0	568.0	524.0
3	医药卫生	427.5	451.8	436.8	438.0	453.1	441.3
4	数理科学和化学	322.5	331.0	313.0	317.0	318.5	320.4
5	农林科学	307.5	311.6	291.1	330.5	347.0	315.6
6	天文与地球科学	305.3	314.2	308.0	311.3	286.8	305.4
7	工业技术	261.1	290.5	280.9	283.0	299.2	280.1
8	自然科学总论	279.8	270.4	255.1	234.5	310.4	274.9
9	航空航天	219.0	246.0	197.0	199.0	225.0	217.2
10	交通运输	123.7	136.3	130.3	133.0	155.3	133.6
	总计平均值	347.9	375.1	315.9	357.5	383.6	363.0

进一步研究发现，5 年刊均他引次数前 3 的期刊分别是《中国人口·资源与环境》（1592 次 / 篇），《山东医药》（1091.2 次 / 篇），《山东工业技术》（805.2 次 / 篇）。除这 3 种期刊外大于 400 次 / 篇的期刊共有 30 种，其中医药卫生类期刊占比最多，共 14 种，分别是《中华

行为与脑科学》《中华肿瘤防治杂志》《齐鲁护理杂志》《山东大学学报医学版》《中国矫形外科杂志》《医学影像学杂志》《现代妇科进展》《泰山医学院学报》《药学研究》《青岛大学医学院学报》《中国现代普外科进展》《山东中医杂志》《齐鲁医学杂志》《精神医

学杂志》等；工业技术类有 6 种期刊，分别是《山东化工》《山东大学学报工学版》《食品与药品》《信息技术与信息化》《山东大学学报理学版》《工程塑料应用》等；自然科学总论有 3 种期刊，分别是《中国海洋大学学报》《山东农业大学学报》和《中国石油大学学报》；农业科学类期刊有 3 种，分别是《山东农业科学》《山东农业工程学报》和《中国蔬菜》；天文与地球科学有 2 种期刊，分别为《海洋科学》和《海洋与湖沼》；生物科学类、数理科学和化学类期刊只有《中国病原生物学杂志》和《化学分析计量》各 1 种期刊。

2.2.3 影响因子分析

2015—2019 年山东省科技期刊各学科影响因子见表 8。由表 8 可知，山东省科技期刊的影响因子呈现明显的学科差异性，影响因子 5 年均值大于 1.0 的期刊共有 3 类，环境与安全科学类影响因子最高，为 2.4087，生物科学类 1.3482 排名第 2，医药卫生类 1.0230 排名第 3。航空类和交通运输类期刊影响因子相对较低，为 0.3156 和 0.2132。从篇均影响因子 5 年变化分析，篇均影响因子从 2015 年的 0.6392 上升至 2019 年的 0.9413，呈现良好的增长态势。

表 8 2015—2019 年山东省科技期刊影响因子统计

序号	学科名称	2015 年	2016 年	2017 年	2018 年	2019 年	5 年平均
1	环境与安全科学	1.6215	2.5968	2.6510	2.6120	2.5620	2.4087
2	生物科学	1.2190	1.3350	1.3010	1.5770	1.3290	1.3482
3	医药卫生	0.7409	0.8663	0.9850	1.2087	1.3216	1.0230
4	数理科学和化学	0.6190	0.7530	0.6790	0.7245	0.7150	0.6981
5	天文与地球科学	0.5459	0.5286	0.5994	0.6874	0.8660	0.6417
6	农林科学	0.4757	0.5251	0.5038	0.6858	0.8210	0.5898
7	工业技术	0.3958	0.5463	0.5116	0.5415	0.5574	0.5117
8	自然科学总论	0.2912	0.3215	0.3307	0.3568	0.5836	0.4002
9	航空航天	0.2790	0.2730	0.3640	0.2890	0.3730	0.3156
10	交通运输	0.2035	0.1768	0.2027	0.2384	0.2847	0.2132
	总计平均值	0.6392	0.7922	0.8128	0.8921	0.9413	0.8150

5 年影响因子均值大于 1.0 的期刊有 21 种，其中医药卫生类期刊有 14 种，分别是《齐鲁护理杂志》《中华肿瘤防治杂志》《中国矫形外科杂志》《腹腔镜外科杂志》《中华行为医学与脑科学》《现代妇科进展》《医学影像学杂志》《精神医学杂志》《中国现代普通外科进展》《山东医药》《山东大学耳鼻喉眼学报》《山东大学学报医学版》《青岛医药卫生》《中国中西医结合影像学杂志》；工业技术类期刊有 2 种，分别是《油气地质与采收率》和《居业》；

环境与安全科学、农业科学、生物科学、自然科学总论和天文与地球科学类各有 1 种期刊，分别是《中国人口·资源与环境》《中国烟草科学》《中国病原生物学杂志》《中国石油大学学报》和《海岸工程》。5 年影响因子均值为 0.5~1.0 的期刊有 41 种，0.2~0.5 的期刊有 48 种，剩余的 8 种期刊 5 年影响因子均值小于 0.2。

2.2.4 即年指标分析

即年指标主要描述期刊当年发表的论文在

当年被引用的情况，它能反映期刊对当前热点问题的反应速率^[7]。2015—2019 年，山东省科技期刊各学科即年指标见表 9。由表 9 可知，山东省科技期刊的即年指标学科差异较大，5 年即年指标均值大于 0.1 的学科为环境与安全

科学、生物科学、医药卫生、农林科学、天文与地球科学和工业技术等 6 个学科，其余 4 个学科均小于 0.1。从篇均即年指标 5 年变化分析，山东省科技期刊在 2015—2019 年篇均即年指标由 0.1163 增至 0.1806，整体上也呈现上升趋势。

表 9 2015—2019 年山东省科技期刊即年指标统计

序号	学科名称	2015 年	2016 年	2017 年	2018 年	2019 年	5 年平均
1	环境与安全科学	0.3335	0.4470	0.4225	0.2865	0.5085	0.3943
2	生物科学	0.2790	0.1370	0.1640	0.1550	0.1330	0.1736
3	医药卫生	0.1347	0.1652	0.1848	0.1569	0.1897	0.1654
4	农林科学	0.0904	0.1103	0.1233	0.1660	0.2397	0.1411
5	天文与地球科学	0.0590	0.0959	0.0942	0.1757	0.1647	0.1171
6	工业技术	0.0794	0.0990	0.1069	0.1232	0.1613	0.1146
7	自然科学总论	0.0611	0.0773	0.0939	0.0658	0.1599	0.0996
8	数理科学和化学	0.0555	0.0665	0.1195	0.0655	0.1100	0.0834
9	交通运输	0.0448	0.0195	0.2027	0.0404	0.0520	0.0754
10	航空航天	0.0250	0.0500	0.0460	0.1010	0.0870	0.0618
	总计平均值	0.1163	0.1268	0.1558	0.1336	0.1806	0.1426

进一步研究发现，5 年即年指标均值大于 0.6 的期刊只有《中国人口·资源与环境》1 种期刊，5 年即年指标均值为 0.4~0.6 的有 2 种期刊，分别是《油气地质与采收率》和《齐鲁护理杂志》，5 年即年指标均值为 0.3~0.4 的有 3 种期刊，分别是《山东工业技术》《居业》和《中华行为医学与脑科学》，5 年即年指标均值为 0.2~0.3 的有 15 种期刊，分别是《海洋气象学报》《中国动物检疫》《现代妇科进展》《中国烟草科学》《腹腔镜外科杂志》《泰山医学院学报》《中国蔬菜》《中国矫形外科杂志》《青岛医药卫生》《山东大学学报医学版》《山东国土资源》《海岸工程》《山东科技大学学报》《山东医药》《煤矿现代化》，5 年即年指标均值为 0.1~0.2 的共有 38 种期刊，其余期刊的即年指标均值小于 0.1。

2.2.5 被引半衰期分析

2015—2019 年，山东省科技期刊各学科被

引半衰期见表 10。由表 10 可知，天文与地球科学、农林科学和交通运输类期刊，被引半衰期 5 年均值大于 7.0 年，说明这 3 个学科期刊的文献老化速度较慢，产生的影响较为持久，自然科学总论、航天航空和数理科学和化学类期刊，5 年被引半衰期均值为 6.0~6.6 年，工业技术、环境与安全科学、医药卫生和生物科学类期刊 5 年被引半衰期均值小于 5.8 年。

进一步研究发现，被引半衰期 5 年均值大于 10 年的期刊有 2 种，分别为《国外铁道车辆》和《海洋科学进展》，被引半衰期均值为 8~9 年的期刊共 19 种，分别为《海洋地质与第四纪地质》《海洋与湖沼》《青岛农业大学学报》《海岸工程》《山东师范大学学报》《海洋科学》《海洋湖沼通报》《山东农业大学学报》《中国海洋药物》《家禽科学》《山东陶瓷》《中国石油大学学报》《中国海洋大学学报》《铁道车辆》《中外葡萄与葡

萄酒》《中国铸造装备与技术》《山东冶金》《落叶果树》和《山东林业科技》，被引半衰期均值为6~8年的期刊有43种，4~6年的期刊有42种，剩余15种期刊被引半衰期均值小于4年。

表 10 2015—2019 年山东省科技期刊被引半衰期统计(单位：年)

序号	学科名称	2015 年	2016 年	2017 年	2018 年	2019 年	5 年平均
1	天文与地球科学	7.8667	7.9167	7.7250	7.8250	8.0091	7.8661
2	农林科学	7.4429	7.3429	7.3467	7.1769	6.5818	7.2075
3	交通运输	6.9500	7.0000	7.1000	7.5600	7.7000	7.2000
4	自然科学总论	6.8167	6.6364	6.1667	5.9909	6.9579	6.5677
5	航空航天	5.7000	6.5000	6.0000	6.7000	7.5000	6.4800
6	数理科学和化学	5.8500	6.0500	5.9000	6.0500	6.1500	6.0000
7	工业技术	5.9000	5.7125	5.8500	5.6029	5.7029	5.7515
8	环境与安全科学	5.5000	4.7000	3.7000	3.8500	5.3000	4.6000
9	医药卫生	5.0121	4.7484	4.5061	4.1613	4.1344	4.5163
10	生物科学	3.9000	3.9000	4.0000	3.7000	4.1000	3.9200
	总计平均值	6.0938	6.0507	5.8295	5.8617	6.2136	6.0109

2.2.6 h指标分析

2015—2019 年山东省科技期刊各学科 h 指标见表 11。由表 11 可知，环境与安全科学、生物科学、医药卫生类期刊的 h 指标 5 年均值大于 7.7，说明这 3 个学科的期刊具有较强的影响力。农林科学、数理科学和化学、天文与地球科学和工业技术类期刊也表现不俗，h 指标均大于 5.0，自然科学总论、航空航天和交通运输类期刊的 h 指标均值都小于 4.3。

表 11 2015—2019 年山东省科技期刊 h 指标统计

序号	学科名称	2015 年	2016 年	2017 年	2018 年	2019 年	5 年平均
1	环境与安全科学	10.05	12.00	11.5	13.00	12.00	11.80
2	生物科学	8.00	8.00	8.00	8.00	8.00	8.00
3	医药卫生	6.97	7.58	7.27	8.19	8.59	7.71
4	农林科学	6.36	6.36	5.93	6.31	6.09	6.21
5	数理科学和化学	5.50	6.50	6.50	7.00	5.00	6.10
6	天文与地球科学	5.92	6.08	5.67	5.67	5.64	5.80
7	工业技术	5.13	5.50	5.59	5.66	4.68	5.31
8	自然科学总论	4.42	4.27	4.08	4.09	4.47	4.29
9	航空航天	4.00	4.00	3.00	3.00	5.00	3.80
10	交通运输	3.00	3.17	3.33	3.60	3.33	3.27
	总计平均值	5.94	6.35	6.09	6.45	6.28	6.23

仔细分析数据发现，《中国人口·资源与环境》的 h 指标 5 年均值高达 20.8，位列第 1，h 指标 5 年均值为 8~17 的共 20 种期刊，分别为《山东医药》《齐鲁护理杂志》《中国矫形外科杂志》《山东工业技术》《中华行为医学与脑科学》《医学影像学杂志》《现代妇科进展》《中华肿瘤防治杂志》《居业》《腹腔镜外科杂志》《中国烟草科学》《中国现代普通外科进展》《油气地质与采收率》《中国蔬菜》《海洋地质与第四纪地质》《海洋与湖沼》《中国石油大学

学报》《泰山医学院学报》《精神医学杂志》和《中国病原生物学杂志》，h 指标 5 年均值为 4~7 的有 67 种期刊，其余 29 种期刊的 h 指标 5 年均值小于 4.0。有很多学者认为 h 指标有其不足之处，它会随着载文量的增加而提高^[8]，因此，这 29 种期刊的 h 指标较小可能与载文量低有关。

2.3 结语

综上所述，2015—2019 年山东省科技期刊在学术影响方面呈现如下趋势：

第一，5 年平均作者数、篇均基金论文数、他引次数和即年指标方面低于江苏省科技期刊外^[3]，其余 8 项指标都高于江苏省科技期刊，这证明山东省科技期刊的学术影响力较强，与科技期刊强省相比也毫不逊色。

第二，各学科期刊年均发文量整体呈现减少态势，表明一些期刊为了增加学术影响力而降低了发文量。期刊影响力是一个比较复杂的命题，受多重因素的影响，在现行期刊评价体制下，适度降低发文量确实能够起到作用，但这需要与其他策略综合运用，单纯降低发文量不一定能够发挥很好的作用，合理设置载文量，采取多种措施提升论文质量，才是提高期刊影响力的重要举措^[9]。

第三，就篇均引文量、篇均作者、篇均基金论文比、论文选出比而言，篇均引文量整体呈现上升趋势；科研工作者越来越注重科研协作，合作度在不断加强；期刊的稿源分布地覆盖全国省市自治区总数的一半以上，稿源总体分布情况较好；篇均基金论文比和年均作者机构数量变化并不大，呈平稳发展态势；山东省科技期刊总体学术规范程度较高。

第四，就总被引频次、他引次数、影响因子、即年指标、被引半衰期和 h 指标而言，山东省科技期刊无论是篇均总被引频次和他引次数还是影响因子和即年指标都呈逐年增长态势，可见山东省科技期刊的学术影响力在不断扩大；天文与地球科学、农林科学和交通运输 3 个学科期刊的文献老化速度较慢，产生的影响较为持久；环境与安全科学、生物科学、医药卫生类期刊的 h 指标 5 年均值最高，说明这 3 个学科的期刊具有较强的学术影响力。需要指出的是，交通运输类期刊在 7 项指标上均列末位，应该引起重视。

参考文献

- [1] 何学锋, 彭超群. 科技期刊学术影响力的动态评估模型 [J]. 编辑学报, 2002, 14(4): 238-240.
- [2] 邹海彬, 杜宁, 张京娜, 等. 科技期刊学术影响力保障机制探讨 [J]. 编辑学报, 2018, 30(1): 1-3.
- [3] 钱爱兵. 江苏省科技期刊学术影响分析 [J]. 中国科技期刊研究, 2017, 28(10): 887-899.
- [4] 白雪娜, 陈夏, 张辉玲. 广东省农业类科技期刊学术影响力动态评估 [J]. 科技管理研究, 2017, 37(7): 248-254.
- [5] 刘青. 提升《河南水利与南水北调》学术影响力的探索 [J]. 中国传媒科技, 2021(9): 129-131, 94.
- [6] 刘艳华, 钱爱兵, 华薇娜. 江苏省科技期刊学术影响力的影响因素研究 [J]. 图书情报研究, 2019(3): 10-15.
- [7] 中国科学技术信息研究所. 2020 年版中国科技期刊引证报告 (扩刊版) [M]. 北京: 科学技术文献出版社, 2020: 8.
- [8] 俞立平, 韩亚茹, 张再杰. 考虑载文量影响的 h 指数优化研究—hq2 指数 [J]. 现代情报, 2020(2): 114-121.
- [9] 许力琴, 顾黎, 周英智, 等. 中文科技期刊载文量“瘦身”现象的调查与建议—基于 71 种高校自然科学学报的统计分析 [J]. 中国科技期刊研究, 2019(5): 443-450.



开放科学
(资源服务)
标识码
(OSID)

基于 Kano 模型的学术会议视频用户需求研究

韩晶

中国科学技术信息研究所 北京 100038

摘要: [目的/意义] 在学术会议现场使用摄像设备录制专家的演讲过程,并将录制的视频进行后期剪辑和处理,这种方式可以将专家的发言与演示内容完整地记录下来,并通过剪辑和处理提取出重要观点和亮点,制作成精炼的学术讲解视频,这是目前常见的学术会议视频形式。在时效性、质量、增值服务三个维度上,用户均有多项功能需求。采用 kano 模型分析方法对这些需求进行分类和优先排序,以期为学术会议视频的升级迭代提供决策依据。[方法/过程] 利用 Kano 模型,首先收集用户对学术会议视频的功能需求并编制问卷,然后将汇总的数值填入 Kano 评价表,确定各项功能的 Kano 属性,最后通过计算得到每项功能的 Better-worse 系数,梳理出各项功能需求实施的优先级。[结果/结论] 研究结果表明:九项用户功能需求中,必备型需求三项,期望型需求两项,魅力型需求三项,无差异型需求一项,据此提出了学术会议视频功能升级迭代对策。

关键词: Kano 模型; 学术会议; 视频; 用户需求

中图分类号: G35 G255.51

Research on User Requirements for Academic Conference Video Based on Kano Model

HAN Jing

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

Abstract: [Purpose/Significance] Using camera equipment to record the expert's speech process at academic conferences, and post editing and processing the recorded video, this method can fully record the speeches and presentations of experts, extract important viewpoints and highlights through editing and processing, and produce a refined academic presentation video. This is currently a common form of academic conference video. Users have multiple functional requirement in terms of timeliness, quality and value-added services. This study uses the Kano model analysis method to classify and prioritize these requirements, in order to provide decision-making basis for the upgrading and iteration of academic conference videos. [Methods/Processes] In this paper, Kano model is used to first collect users' functional requirements for academic conference videos and prepare a

作者简介 韩晶 (1975-), 硕士, 馆员, 主要研究方向为情报学、视频数据库、文献检索等, E-mail: 1509213084@qq.com。

引用格式 韩晶. 基于 Kano 模型的学术会议视频用户需求研究 [J]. 情报工程, 2024, 10(4): 96-105.

questionnaire, then fill the summarized values into the Kano evaluation table to determine the Kano attributes of each function. Finally, the Better-Worse coefficient of each function is calculated to sort out the implementation priority of each function. [Results/Conclusions] The research results indicate that among the nine user functional requirements, there are three must-be qualities, two one-dimensional qualities, three attractive qualities, and one indifferent quality. Based on this, this article proposes iterative strategies for upgrading the functionality of academic conference videos.

Keywords: Kano Model; Academic Conferences; Video; User Requirement

引言

近年来,随着电子信息技术的不断发展,各种在线会议平台如雨后春笋般出现,为线上学术会议的召开提供了强有力的技术保障^[1]。在新冠疫情之后,学术会议新形式探索的建议越来越多,组建学术视频资源库,形成共享机制^[2]成为共识。在学术会议现场使用摄像设备录制专家的演讲过程,并将录制的视频进行后期剪辑和处理,这种方式可以将专家的发言与演示内容完整地记录下来,并通过剪辑和处理提取出重要观点和亮点,制作成精炼的学术讲解视频,这是目前常见的学术会议视频形式。

现场录制学术会议视频配合后期剪辑的优势包括:(1)现场录制确保了对专家演讲和演示内容的全面记录,捕捉到演讲的每个细节和要点,这使得会议进程可以被准确、全面地记录下来;(2)录制演讲视频能够视觉展示专家的演示内容,包括幻灯片、图表和其他可视化辅助工具,这能够促进观众对演讲内容的理解,提高参与度;(3)通过后期剪辑和处理,可以分析录制的视频并提取重要观点和亮点,这有助于提炼演讲的精华,制作出侧重于关键要点和重要观点的精炼视频,为观众提供有价值的内容;(4)通过录制会议场景,视频可以被更

广泛的观众所接触,未能参加会议的人仍可以从这些视频中受益。此外,观众可以灵活选取时间观看视频,实现自主学习和参考。

总的来说,现场录制学术会议视频为知识传播、研究成果分享和学术讨论提供了宝贵的资源,有利于学者之间的学术交流。学术会议视频平台吸引哪些用户,其观看的频率取决于所提供的“学术视频”的类型、质量等因素^[3]。在视频创作过程中注重对专业知识和精彩片段的剪辑,将用户关切的内容给予重点呈现,强化视频的层次感、节奏感和重点^[4],即视频制作应以用户需求为导向,面向学术会议视频的用户需求研究的必要性体现在:(1)通过深入了解用户的需求,会议视频的制作团队可以针对性地改善视频内容和交流方式,从而提升用户的观看体验,例如根据用户喜好调整视频长度、改进讲者的表达方式以及优化视听效果,都可以增强用户的满意度;(2)了解用户需求有助于设计更具吸引力和互动性的视频内容和形式,从而提高用户的参与度和互动性,如设置合适的问答环节、提供实时在线交流平台以及利用社交媒体进行互动等;(3)通过了解用户的需求和关注点,会议视频可以更好地促进学术交流和合作,例如通过提供在线交流平台和社交媒体互动,促进与其他观众和讲者之间

的交流,有助于激发新的研究思路和合作机会,推动学术领域的发展;(4)满足用户需求并提供高质量的会议视频内容可以增强学术会议的品牌价值和声誉。用户对内容的满意度和认可度将直接影响他们对会议的评价和推荐,进而影响到会议的声誉和吸引力,对于吸引更多的观众和学者参与具有重要意义。

通过多年持续性跟踪和深入研究,我们发现用户会关注视频中呈现的学术内容是否具有深度、广度和权威性,他们希望从视频中获取新知识、见解和研究成果,内容的原创性和创新性是用户所关注的重点之一。同时,用户对讲者的专业知识和表达能力有着很高的期待,他们希望听到有经验、权威的专家分享自己的见解,并且能够清晰、生动地表达出来。用户还希望在视频交流中参与讨论和互动,这可能包括提问环节、在线讨论区或者社交媒体平台上的互动。视频的画质、音质、时长等功能也会直接影响用户的观看体验,用户希望能够清晰地看到演讲者和演示内容,并且能够听清楚讲者的发言。如何将这些用户需求权重进行有

效排序,从而有针对性地进行优先改进,这将直接影响用户对视频的满意度和价值认可。

本文旨在探索一种实用的“用户需求分析方法”,通过深入了解用户对产品或服务的需求,帮助企业更好地理解用户的需求,从而针对性地进行改进和优化,以便设计出符合用户期望的产品或服务,实现提高用户满意度和产品竞争力。Kano 模型是一种经典的用户需求分类和优先排序方法。该模型由日本学者狩野纪昭提出,通过调查用户对产品和服务各项功能的满意度和期望程度,将功能分类到不同的需求类型中,并据此进行优先排序和产品设计。除了 Kano 模型,还有其他一些常用的用户需求分析方法,如 QFD 模型(质量功能展开)、CSI 模型(客户满意度指数)、VOC 模型(客户之声),表 1 为四款模型的优劣势对比。选择 Kano 模型来进行用户需求分类和优先排序,具有一定优势。一方面,Kano 模型能够将用户需求划分为不同类型,并针对不同类型的需求给予不同的重视程度,能够更准确地指导产品设计和改进。另一方面,Kano 模型在实际应用中被广泛验证

表 1 四款模型工具的优缺点对比

名称 内容	Kano 模型	QFD 模型 Quality Function Deployment	CSI 模型 Customer Satisfaction Index	VOC 模型 Voice of the Customer
定义	Kano 模型可以将用户需求分为五种类型:必备型需求;期望型需求;魅力型需求;无差异型需求;反向型需求。有助于制定针对性的产品或服务策略。	质量功能展开是一种将用户需求转化为产品设计特性的方法。它通过建立需求与产品特性之间的关联矩阵,帮助制定产品设计和开发方案。	客户满意度指数是一种定量评估客户满意度的方法,可以通过调查问卷等方式收集用户反馈,然后计算满意度指数。	客户之声是一种系统地收集、分析和应用用户反馈的方法,可以帮助了解用户需求、期望和偏好,从而指导产品或服务的改进和创新。
优点	可对产品或服务的功能性需求进行分类和优先排序,识别用户对不同需求的满意度和期望程度。	能够将用户需求转化为具体的产品设计特性。	可以提供定量的客户满意度指数。	能够系统性地收集和分
缺点	对于某些复杂的需求情况,可能需要额外的细化和深入调研。	在应用过程中可能需要大量的数据和专业	可能受到调查问卷设计和样本选择等因素的影响。	可能需要较长的时间和资源成本,以及对用户反馈的有效处理和应用。

和采用,其在包装设计^[5]、交互设计^[6]、工业设计^[7]、服务设计^[8]、家居设计^[9]等多个领域中展现出广泛的应用价值,具有较强的可操作性和实用性,如在数字产品交互设计上,它揭示关键触点,助力设计师提升界面响应速度、易用性等用户体验要素^[10],已经成为业界的主流方法之一。因此,围绕“通过了解用户对不同需求类型的反应,优先满足和引入对用户最有价值的需求,提高产品或服务的竞争力”的研究目的,本研究最终采用 Kano 模型分析方法对这些需求进行分类和优先排序,以期为学术会议视频的升级迭代提供决策依据。

1 相关研究现状综述

关于利用 Kano 模型分析视频功能的研究论文并不多见。相关的研究包括黄冰毅^[11]通过网络调研总结归纳国内主要短视频平台所提供的拓展信息服务功能,据此设计调查问卷,并依次进行 Kano 模型数据分析、混合类分析、满意度指数分析,对短视频平台拓展信息服务功能进行需求类型及功能优先度界定;张佳燕^[12]通过发放 Kano 调查问卷,洞悉了产品体验点的属性分类和重要度排序,并转化为产品设计要素,进而导出了短视频老龄用户的积极体验模型和短视频平台适老化设计策略;程正^[13]利用 Kano 模型,对 20 个黏性付费的样本进行需求满意度的分析,将各项服务进行分类,最终起到优化各个网络视频平台的会员体验的作用;伍谦^[14]通过问卷调查的方式,面向在川教育部直属高校图书馆的学生用户,综合分析学生对微视频服务的认知,了解他们对所在馆已

有微视频服务的重要度、现状评价、质量需求等认知态度;孙敏夫^[15]梳理并分析了网络直播服务质量和图书馆服务质量评价的相关研究,阐述了公共图书馆网络直播服务概况,总结了其存在的问题,建立了公共图书馆网络直播服务质量优化决策规则;姜冬等^[16]通过 Kano 模型方法进行调查统计,对数字音乐信息获取途径的品质要求进行分类,并采用逻辑分析方法对各品质要素进行解析;于鹏飞^[17]利用 Kano 模型分析了当时很火热的腾讯、搜狐、爱奇艺等视频网站,得出了用户模糊需求的模型,然后再结合用户接受度、反向用户率等一系列指标给出建议。综上,将 Kano 模型应用于学术会议视频功能分析的研究尚少。

2 基于 Kano 模型的学术会议视频用户需求分析方法

Kano (卡诺)模型由日本的狩野纪昭(Noriaki Kano)在 1984 年提出,其设计灵感来自美国心理学家赫兹伯格的双因素理论,Kano 模型常应用在用户需求分类和计算优先级方面,它清晰揭示了质量特性的满足程度投射到用户满意度上的指标含义,其框架见图 1。

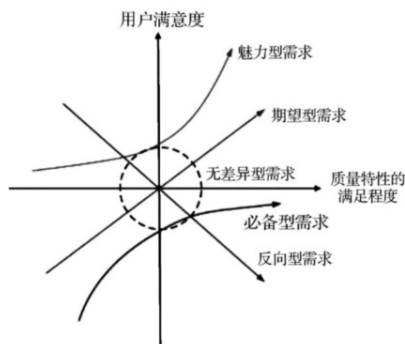


图1 Kano 模型框架

狩野纪昭将影响用户满意度的因素划分为五个类型^[18],包括:用户认为产品“必须具备”的功能需求即为必备型需求(must-be quality, M);用户的满意状况与需求的满足程度成线性正比例关系的需求即为期望型需求(one-dimensional quality, O);不会被用户过分期望的需求即为魅力型需求(attractive quality, A);用户不关心提供与否的需求即为无差异型需求(indifferent quality, I);用户完全没有此需求,即为反向型需求(reverse quality, R)。根据以上定义,将用户对产品每一项功能的正向及反向态度结果按“A”(魅力型需求),“O”(期望型需求),“M”(必备型需求),“I”(无差异型需求),“R”(反向型需求)进行分类,填入Kano评价表,见表2。其中“Q”为可疑性结果,即在用户面对正向问题和反向问题时,同时选择了很喜欢或者很不喜欢,这属于自相矛盾的结果,真实性存疑,故分类为Q。Kano评价表是Kano问卷结果评价的依据。如当正向问题的回答是“很喜欢”,对反向问题的回答是“很不喜欢”,那么在Kano评价表中,这项需求的类型就分类为O,即期望型需求;又如反向型需求,和期望型需求恰好相反,不提供用户就满意,提供了用户就反感,那么这些项的类型就分类为R。

表2 Kano 评价表

正向问题: 提供的态度	反向问题: 不提供的态度				
	很喜欢	理应如此	无所谓	勉强接受	很不喜欢
很喜欢	Q	A	A	A	O
理应如此	R	I	I	I	M
无所谓	R	I	I	I	M
勉强接受	R	I	I	I	M
很不喜欢	R	R	R	R	Q

2.1 样本量计算

依据何蔚珊^[19]对Kano问卷样本量确定的原则,样本量的计算公式为: $N=Z^2 \times \sigma^2/d^2$,其中Z为置信区间、N为样本容量、d为抽样误差范围、 σ 为标准差,一般取0.5。常用的置信水平(90%; 95%; 99%)对应的Z值(1.64; 1.96; 2.58)。本研究设置置信区间为99%,允许误差范围为10%,此时最小样本量为166。

2.2 样本对象选取

利用问卷星,对不同学科背景的科研人员进行问卷调查,发放问卷288份,清洗不合格问卷,最终回收了来自全国23个省、市、自治区的有效问卷231份,覆盖了理、工、农、医、人文、社会及交叉学科等各类学科。

2.3 问卷信度和效度分析

通过SPSS软件进行Cronbach信度分析,其中信度系数值Cronbach α 为0.603,通常大于0.6,说明样本回答真实可靠,信度质量可以接受。在效度方面,使用KMO和Bartlett球形度检验进行效度验证,KMO值为0.609,介于0.6~0.7之间;p值为0.000,小于0.05,说明研究数据通过Bartlett检验,具有效度。

2.4 研究过程和步骤

2.4.1 收集功能需求,编制Kano调查问卷

该部分是对“学术会议视频”产品各项功能的调查,从时效要求、质量要求和增值要求三个维度共整合出9对条目,每对条目按表3所示的形式进行设计。

表 3 调查问卷示意图

如果提供此功能，您会	如果不提供此功能，您会
A. 很喜欢	A. 很喜欢
B. 理应如此	B. 理应如此
C. 无所谓	C. 无所谓
D. 勉强接受	D. 勉强接受
E. 很不喜欢	E. 很不喜欢

具体功能需求分类表见表 4，据此编制 Kano 问卷。

2.4.2 完成Kano评价表

依次把样本数据正反面答题组合人数，填入到 Kano 评价表中，以 A1 功能“第一时间获得学术会议视频以便真实还原会议现场”举例，见表 5。分为正反两方面的问题编制 Kano 问卷，每一个人在面对正向问题时有 5 种选择，同理，面对反向问题也是 5 种选择。所有受访者选择完毕后，就可以基于 Kano 模型对“学术会议视频”该项功能的重要程度进行分析。

利用表 5 将相同类型的数据累加，就可以得到各类需求所占比例的总和，其中占比最大

的就是此项功能的归属类。从表 5 可以看出，“第一时间获得学术会议视频以便真实还原会议现场”功能属于魅力型需求。

表 4 功能需求分类表

维度	功能编号	功能描述
时效要求	A1	第一时间获得学术会议视频以便真实还原会议现场
	B1	采用多机位展示演讲者的形式，通过视觉效果让学术传播更生动
质量要求	B2	使用超清晰的视频技术，呈现更为清晰的图像和声音质量
	B3	对于超过 20 分钟的学术会议视频，提供分集浏览功能
	B4	没有繁琐的广告或者无意义的信息介绍
增值要求	C1	观看或浏览学术会议视频时，系统根据用户兴趣点进行智能推荐，使用户快速得到自己关注领域的其他相关视频
	C2	观看或浏览学术会议视频时，同时提供该论文参考文献的全文链接服务
	C3	观看或浏览学术会议视频时，无缝链接各类沟通软件 / 社交网络平台，增加学术交流的互动性
	C4	观看或浏览学术会议视频时，系统提供实时翻译和字幕，方便跨语种用户观看和理解

表 5 Kano 评价表 -A1 功能

A1		反向问题：不能第一时间获得学术会议视频					合计
		很喜欢	理应如此	无所谓	勉强接受	很不喜欢	
正向问题：第一时间获得学术会议视频以便真实还原会议现场	很喜欢	Q(21)	A(12)	A(27)	A(46)	O(67)	173
	理应如此	R(2)	I(13)	I(7)	I(3)	M(13)	38
	无所谓	R(0)	I(0)	I(14)	I(0)	M(0)	14
	勉强接受	R(0)	I(0)	I(2)	I(3)	M(1)	6
	很不喜欢	R(0)	R(0)	R(0)	R(0)	Q(0)	0
合计		23	25	50	52	81	231

2.4.3 确定各项功能的Kano属性

同理，将所有 9 对类目的数据结果导入

Kano 评价表进行统计分析、汇总，确定各项功能的 Kano 属性，见表 6。

表 6 功能属性归类表

维度	功能编号	功能描述	A	O	M	I	R	Q	合计	归属类
时效要求	A1	第一时间获得学术会议视频以便真实还原会议现场	85	67	14	42	2	21	231	A
	B1	采用多机位展示演讲者的形式，通过视觉效果让学术传播更生动	43	30	109	38	1	10	231	M
质量要求	B2	使用超清晰的视频技术，呈现更为清晰的图像和声音质量	20	146	51	10	4	0	231	O
	B3	对于超过 20 分钟的学术会议视频，提供分集浏览功能	55	56	113	4	2	1	231	M
	B4	没有繁琐的广告或者无意义的信息介绍	12	23	189	2	3	2	231	M
增值要求	C1	观看或浏览学术会议视频时，系统根据用户兴趣点进行智能推荐，使用户快速得到自己关注领域的其他相关视频	102	67	56	5	1	0	231	A
	C2	观看或浏览学术会议视频时，同时提供该论文参考文献的全文链接服务	170	32	28	1	0	0	231	A
	C3	观看或浏览学术会议视频时，无缝链接各类沟通软件 / 社交网络平台，增加学术交流的互动性	75	84	66	4	1	1	231	O
	C4	观看或浏览学术会议视频时，系统提供实时翻译和字幕，方便跨语种用户观看和理解	43	45	120	12	6	5	231	M

2.4.4 功能实施的质量因素判定方法

数值计算公式，计算出每个功能的 Better-Worse 系数，用以确定功能实施的优先级，见表 7。

对功能属性进行归类后，带入 Better-Worse

表 7 Better-Worse 系数

维度	功能编号	功能描述	A	O	M	I	R	Q	合计	归属类	Better	Worse
时效要求	A1	第一时间获得学术会议视频以便真实还原会议现场	85	67	14	42	2	21	231	A	0.73	-0.39
	B1	采用多机位展示演讲者的形式，通过视觉效果让学术传播更生动	43	30	109	38	1	10	231	M	0.33	-0.63
质量要求	B2	使用超清晰的视频技术，呈现更为清晰的图像和声音质量	20	146	51	10	4	0	231	O	0.73	-0.87
	B3	对于超过 20 分钟的学术会议视频，提供分集浏览功能	55	56	113	4	2	1	231	M	0.49	-0.74
	B4	没有繁琐的广告或者无意义的信息介绍	12	23	189	2	3	2	231	M	0.15	-0.94
增值要求	C1	观看或浏览学术会议视频时，系统根据用户兴趣点进行智能推荐，使用户快速得到自己关注领域的其他相关视频	102	67	56	5	1	0	231	A	0.73	-0.53
	C2	观看或浏览学术会议视频时，同时提供该论文参考文献的全文链接服务	170	32	28	1	0	0	231	A	0.87	-0.26
	C3	观看或浏览学术会议视频时，无缝链接各类沟通软件 / 社交网络平台，增加学术交流的互动性	75	84	66	4	1	1	231	O	0.69	-0.66
	C4	观看或浏览学术会议视频时，系统提供实时翻译和字幕，方便跨语种用户观看和理解	43	45	120	12	6	5	231	M	0.4	-0.75

传统 Kano 模型中质量因素判定的方法往往只能选取频率出现次数最多的类别作为最终归属类别, 往往忽视了在其他类别上的数据, 丢失了大量的数据, 从而导致研究结果过于笼统^[20]。为了克服这种弊端, 达到精细化管理的要求, 对于 Kano 模型中质量因素判定, 可以通过计算质量因素归类属性的百分比, 得出 Better-Worse 系数^[21]。

Berger 等人提出的 Better-Worse 系数计算公式为:

$$\text{Better} = (O+A) / (M+O+A+I)$$

$$\text{Worse} = (O+M) / (M+O+A+I) \times (-1)$$

Better 系数的分子为期望型需求和魅力型需求, 分母则为期望型需求、魅力型需求、必备型需求与无差异型需求的总和, 该值一般为

正, 说明提供了某项功能后, 用户也就越满意, 该值越大, 说明满意度提升得越快; Worse 系数的分子为期望型需求和必备型需求, 分母在 better 的基础上乘了 -1, 所以该值一般为负, 表示如果不提供某功能, 用户的满意度就会降低, 其绝对值越大, 代表用户满意度降低的效果会越强, 满意度下降得越快。因此, 根据 Better-Worse 系数, 对系数绝对值较大的功能应当优先实施。

2.4.5 绘制四象限散点图

在 excel 中制表, 取 worse 系数的绝对值为 x 轴, better 系数为 y 轴, 以 worse 系数绝对值的平均值、better 系数的平均值分别作为中心点, 绘制四象限散点图, 直观呈现各功能的属性, 见图 2。

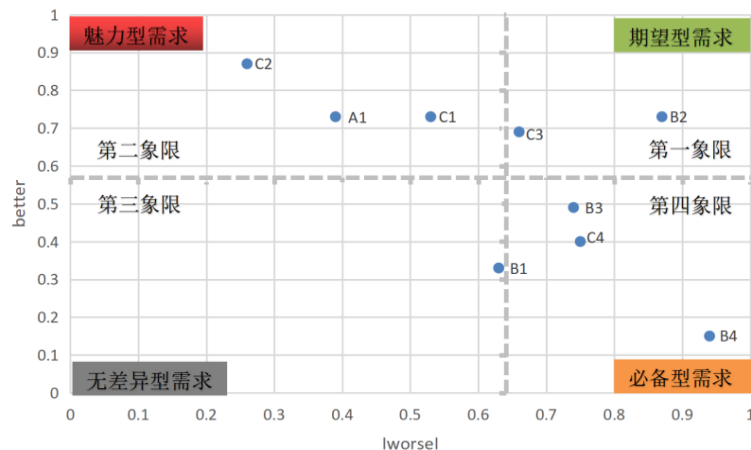


图 2 Better-Worse 系数分析图

2.4.6 研究结果分析

第一象限到第四象限按逆时针顺序分别称为期望型需求、魅力型需求、无差异型需求和必备型需求。观察图 2 发现, 第一象限 better 系数相对较大, worse 系数绝对值也大, 功能 B2、C3 包含在此象限内, 即“高清拍摄”和“增加互动”两项功能是用户的期望型需求; 第二象限

better 系数值相对较大, 但 worse 系数绝对值相对较小, 功能 A1、C1、C2 包含在此象限, 即“第一时间获取视频”“系统智能推荐相关视频”和“提供参考文献全文链接”三项功能是用户的魅力型需求; 第三象限 better 系数值相对较小, worse 系数绝对值也小, 功能 B1 包含在此象限, 即用户对“多机位拍摄”功能并不关心; 第四象

限 better 系数值相对较小,但 worse 系数绝对值相对较大,功能 B3、B4、C4 包含在此象限,即“分集浏览”“去掉广告”和“系统提供实时翻译和字幕”功能是用户的必备型需求。

3 学术会议视频功能升级迭代对策

充分利用 Better-Worse 系数分析,当各项功能服务指标满足与不满足时,可以更清晰地了解服务对象的满意程度^[22]。由此可见,实施这些功能的优先级顺序为:必备型需求>期望型需求>魅力型需求>无差异型需求。结合本文研究内容,分为四个阶段有针对性地提供学术会议视频升级迭代对策。

3.1 必备型需求迭代策略

对于用户关注的“对于超过 20 分钟的学术会议视频,提供分集浏览功能”“没有繁琐的广告或者无意义的信息介绍”和“观看或浏览学术会议视频时,系统提供实时翻译和字幕,方便跨语种用户观看和理解”这三个必备型需求,一定优先确保得到满足,因为如果没有这些功能,用户会感到不满意。所以在去掉广告的同时可以进一步深入调研,如对于超过 20 分钟的学术会议视频提供分集浏览功能,不同学科背景的学者对每一集的时长控制在多久比较合适;对于实时翻译和字幕功能,是提供中文还是中英文对照,抑或中文繁体对照,这些都是功能设计时需要优先考虑的问题。

3.2 期望型需求迭代策略

两个期望型需求,即“使用超清晰的视频技术,呈现更为清晰的图像和声音质量”和“观看

或浏览学术会议视频时,无缝链接各类沟通软件/社交网络平台,增加学术交流的互动性”,这是产品的制胜法宝,与用户的满意度呈正相关,在满足了必备型需求后,可以考虑逐步满足期望型需求,以提高用户的满意度和忠诚度。如 4K 的观看体验与 480P 标清完全不同,再如用户是否可以与发言嘉宾就某一学术问题进行互动和交流,这可以作为产品的控标点在市场上大放异彩。

3.3 魅力型需求迭代策略

对于用户认为的三个增值服务,即“第一时间获得学术会议视频以便真实还原会议现场”“观看或浏览学术会议视频时,系统根据用户兴趣点进行智能推荐,使用户快速得到自己关注领域的其他相关视频”和“观看或浏览学术会议视频时,同时提供该论文参考文献的全文链接服务”,均属于魅力型需求,这是指用户可能没有预期到的,但一旦实现了会极大地提升用户体验。所以在还有余力的时候,也要努力做到精准匹配,如“第一时间”如何界定,是会议召开后的一小时内、当天还是一周内,甚至一月内,把握好用户的兴奋点,就能提升用户对产品的美誉度。

3.4 无差异型需求迭代策略

还应注意无差异型需求,即“采用多机位展示演讲者的形式,通过视觉效果让学术传播更生动”,目前用户对这些特性或功能并不关心,所以并没有必要花大力气去实现,不需要额外投入资源,因为它们对用户满意度没有实质性影响。当然这只是现阶段反映出来的结果,用户的需求也会实时变化,多机位拍摄的效果肯

定优于单机位,如用户希望学术会议视频除了可以展示发言嘉宾的 PPT 内容以外,人文社科的用户会更聚焦于其声音内容、面部表情和肢体语言,理工科用户的兴趣点在于是否可以在视频中涵盖发言中涉及的实验步骤、实验结果、支撑数据和补充材料。

4 结语

Kano 模型是一种质量管理工具,它的优点在于用户导向,将用户的需求和满意度作为核心,有助于企业更好地理解用户对产品或服务的期望。通过将用户需求分为不同类型,可以帮助企业理清优先级,集中精力在对用户最为重要的功能的开发和迭代上,在制定优先策略时,需要根据不同类型需求的特点和对用户满意度的影响程度进行综合考虑,并结合产品或服务的实际情况和资源分配来确定优先级。由于 Kano 模型依赖于用户反馈,缺乏实际使用数据,这些调查结果可能受到样本不足或无法准确获取的情况影响,从而可能导致评估结果的偏差,且用户需求和市场环境都是变化的,Kano 模型需要不断更新和调整以保持有效性,这对企业来说可能是一项挑战。总体而言,Kano 模型是一种有用的工具,可以帮助企业更好地了解 and 满足用户需求,但也需注意其局限性和适应性问题。

参考文献

- [1] VIGLIONE G. How scientific conferences will survive the coronavirus shock[J]. Nature, 2020, 582(7811): 166-167.
- [2] 宁燕,马诗诗,黄鹃,等.新型冠状病毒肺炎疫情常态化防控形势下学术会议的新形式[J]. 上海医学, 2021, 44(4): 262-265.
- [3] 钟颖. 中国诞生“学术视频”网站的可能性[J]. 青年记者, 2007(14): 149-150.
- [4] 王春迎,王曼玉,占莉娟. 媒体融合背景下我国科技期刊 B 站传播现状与提升策略研究[J]. 中国科技期刊研究, 2022, 33(1): 67-75.
- [5] 马雪寒,刘文良. 基于 KJ-KANO 模型的“一站式”水果 APP 设计[J/OL]. 包装工程:1-17[2024-04-28]. <http://kns.cnki.net/kcms/detail/50.1094.TB.20240422.1205.002.html>.
- [6] 王锰,徐磊,孙红蕾. 公共文化云平台游戏化设计研究——以国家公共文化云为例[J]. 图书情报工作, 2023, 67(21): 73-88.
- [7] 张凡煦. 被动式下肢外骨骼运动辅具设计研究[D]. 无锡: 江南大学, 2023.
- [8] 高凡,喻兴佳,陈亚芳. 常态化疫情防控阶段高校学生对图书馆线上服务的需求研究[J]. 图书情报工作, 2023, 67(14): 21-30.
- [9] 刘文良,马雪寒. 基于 KJ-KANO 模型的亲子家居文创产品设计研究[J]. 家具与室内装饰, 2024, 31(3): 72-77.
- [10] 江玉洁,吕杰锋,孙莹. KANO 模型在产品领域中的应用现状研究综述[J]. 设计, 2024, 37(5): 73-75.
- [11] 黄冰毅. 基于 Kano 模型的短视频平台拓展信息服务功能研究[J]. 梧州学院学报, 2023, 33(2): 17-28.
- [12] 张佳燕. 基于积极体验的短视频平台适老化设计研究[D]. 无锡: 江南大学, 2023.
- [13] 程正. 网络视频平台的定制化会员服务设计[J]. 科技传播, 2021, 13(11): 146-150.
- [14] 伍谦. 基于学生用户认知的高校图书馆微视频服务研究——以在川教育部直属高校图书馆为例[D]. 成都: 四川大学, 2021.
- [15] 孙敏夫. 基于 Kano 模型与 IPA 分析的公共图书馆网络直播服务质量优化研究[D]. 武汉: 华中师范大学, 2020.
- [16] 娄冬,周密,王雪莹. 基于 KANO 模型的用户数字音乐信息获取途径品质要求研究[J]. 情报科学, 2020, 38(1): 116-124.
- [17] 于鹏飞. 互联网在线视频的用户需求分析[D]. 上海: 华东理工大学, 2016.
- [18] 魏丽坤. Kano 模型和服务质量差距模型比较研究[J]. 标准科学, 2006(9): 10-13.
- [19] 何蔚珊. 基于 Kano 模型的在线学习 App 交互界面适老化设计研究[D]. 广州: 华南理工大学, 2020.
- [20] 齐向华. 基于 Kano 模型的图书馆电子服务质量要素分类研究[J]. 情报理论与实践, 2015(4): 80-85.
- [21] 邓磊. 基于 KANO 模型的档案信息化服务质量研究[J]. 兰台世界, 2017(8): 17-19.
- [22] 施衍如. 基于 Kano 模型的少儿图书馆服务质量影响因素分类研究[J]. 图书馆, 2019(10): 106-111.



开放科学
(资源服务)
标识码
(OSID)

省域科技金融效率时空分异及影响因素研究 ——以山东为例

姜宁朋¹ 王福康² 姜媛³ 曲妍妍⁴ 周立波¹ 张路²

1. 山东省科技服务发展推进中心 济南 250101;
2. 齐鲁工业大学(山东省科学院)经管学部 济南 250353;
3. 山东省创新发展研究院 济南 250101;
4. 山东产权交易中心有限公司 济南 250101

摘要: [目的/意义] 探讨科技金融效率的评价机制。[方法/过程] 通过构建超效率 SBM 模型和全局 Malmquist 指数,以山东为例测度了山东省各地市科技金融效率及其年际变化和地域差异,并进一步剖析了影响科技金融效率的相关因素。[局限] 未细分到具体的科技金融工具和手段,下一步将继续深入研究。[结果/结论] 山东省科技金融效率整体上升,但地域差异显著,而技术进步有助于缩小地区差异。时间维度上,2014—2017 年间,多数地区效率稳步上升,而 2018—2022 年间,三大经济圈一体化联动使得效率均呈现波动上升趋势。金融市场发展和数字化发展有助于山东省科技金融效率的提高,政府和企业科研投入与科技金融效率呈负相关,科技金融资源的多元化、科学化配置是各级政府和市场需重点考虑的问题。

关键词: 科技金融效率; 影响因素分析; 超效率 SBM 模型; Tobit 模型

中图分类号: G35; F832.48

Research on the Spatiotemporal Differentiation and Influencing Factors of Science and Technology Financial Efficiency in Provincial-level: Evidence from Shandong

JIANG Ningpeng¹ WANG Fukang² JIANG Yuan³ QU Yanyan⁴ ZHOU Libo¹ ZHANG Lu²

1. Shandong Science and Technology Service Development Promotion Center, Jinan 250101, China;
2. Department of Economics and Management, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China;
3. Shandong Innovation and Development Research Institute, Jinan 250101, China;
4. Shandong property rights Trading Center Co., LTD, Jinan 250101, China

基金项目 山东省重点研发计划(软科学)项目“财政科技股权投资对科技企业创新能力影响研究”(2024RZB0305); 山东省重点研发计划(软科学)项目“山东省科技金融体系构建及创新方法研究”(2022RZA03016)。

作者简介 姜宁朋(1980-), 硕士, 副研究员, 主要研究方向为科技金融与信息服务; 王福康(1999-), 硕士研究生, 主要研究方向为科技金融、区块链金融; 姜媛(1981-), 学士, 副研究员, 主要研究方向为科技创新与科技金融, E-mail: 40026646@qq.com; 曲妍妍(1982-), 硕士, 中级经济师, 主要研究方向为产权交易、科技金融; 周立波(1990-), 硕士, 助理研究员, 主要研究方向为科技金融与信息服务; 张路(1982-), 博士, 副教授, 主要研究方向为科技金融、区块链金融。

引用格式 姜宁朋, 王福康, 姜媛, 等. 省域科技金融效率时空分异及影响因素研究——以山东为例[J]. 情报工程, 2024, 10(4): 106-113.

Abstract: [Objective/ Significance] This paper aims to explore the evaluation mechanism of STF efficiency. [Methods/Processes] By constructing a super-efficiency SBM model and a global Malmquist model, the study measured the efficiency of STF and its inter-annual changes and regional differences in various cities in Shandong Province. At the same time, the Tobit model was used to further analyze the factors that influence the efficiency of STF. [Limitations] This study has not been subdivided into specific technology financial instruments and means, and further in-depth research will be conducted in the next step. [Results/Conclusions] The results indicate that the efficiency of STF in Shandong Province has increased overall, but there are significant regional differences. Technological progress helps to narrow regional disparities. In terms of time dimension, from 2014 to 2017, the efficiency of most regions steadily increased, while from 2018 to 2022, the integration and linkage of the three major economic circles led to a fluctuating upward trend in efficiency. The degree of financial market development and digitization contributes to the improvement of STF efficiency in Shandong Province, while government and enterprise research investment is negatively correlated with STF efficiency. The diversified and scientific allocation of STF resources is a key issue that governments at all levels and the market must take into serious consideration.

Keywords: Science and Technology Finance; Analysis of Influencing Factors; Super SBM Model; Tobit Model

引言

2023 年中央金融工作会议指出，要“做好科技金融、绿色金融、普惠金融、养老金融、数字金融五篇大文章”“着力打造现代金融结构和市场体系，疏通资金进入实体经济的渠道”。近期，国家出台了一系列金融支持创新政策，对科技金融结合工作进行了部署。各省应对省内各市科技金融工作实施绩效进行全面评估，进一步落实好中央科技金融工作安排。

山东省作为北方先进省份，在推动金融服务科技创新方面，初步形成以科技信贷、风险补偿、贷款贴息、股权投资、上市培育为主的科技金融服务体系，在缓解企业融资难题、促进科技成果转化方面取得良好成效。然而，山东省各市科技金融工作开展不均衡，与高质量发展和科技创新要求存在差距，如科技金融生态体系尚不健全，政策联动不足，社会资本参与度低，初创科技企业融资难题仍未得到有效

解决。

有学者对山东省以及国内其他省市的科技金融效率差异和动态演进进行了研究^[1-2]，并且从科技信贷、科技保险、科技资本市场等角度分析了制约科技金融发展的因素^[3-4]。但针对山东省各地市不同年份及地市间科技金融效率的对比研究还存在不足。因此，本文以山东省 16 地市为例，深入探讨山东省的科技金融效率，并根据地域特性提出精准的优化资源配置对策，为国家和其他类似省份的科技金融政策布局提供有价值的参考。本文主要研究问题包括：（1）山东省科技金融效率损失情况分析；（2）各市科技金融效率的年际变化和区域差异研究；（3）影响山东省科技金融效率的主要因素研究。

1 研究设计与模型构建

本文通过构建科技金融投入与产出指标，利用超效率 SBM 模型测算山东省内 16 个地市

(除莱芜市)^①的科技金融效率,并运用全局 Malmquist 指数分析各地市科技金融效率的时空变化。数据来自历年出版的《山东统计年鉴》《山东科技年鉴》和《国民经济和社会发展统计公报》。

1.1 指标选取与数据说明

本文参考胡欢欢等^[1]、刘濛等^[5]和张远为^[6]等学者的研究,从科技金融投入和科技创新产出两方面构建了山东省各地市科技金融效率评价指标体系,该体系全面反映了山东省各地市的实际情况,确保了数据的连续性和可得性,具体指标如表 1 所示。

表 1 山东省各地市科技金融效率评价指标体系

指标类型	指标名称	衡量指标	单位
科技金融投入	科研资金投入	地方科学技术支出 (X_1)	亿元
	金融资本投入	金融机构科技贷款额 (X_2)	亿元
	人力资本投入	R&D 人员全时当量 (X_3)	人/年
科技创新产出	知识产出	发明专利授权量 (Y_1)	件
	技术产出	技术市场成交合同额 (Y_2)	亿元

科技金融投入中,地方科学技术支出 (X_1) 体现了政府对科研的资金支持;金融机构科技贷款额 (X_2) 反映了金融机构的资本投入;R&D 人员全时当量 (X_3) 反映了人力资本投入。科技创新产出方面,发明专利授权量 (Y_1) 代表知识产出;技术市场成交合同额 (Y_2) 代表科技活动市场化产出结果。

1.2 山东省科技金融效率测算分析

本文运用超效率 SBM 模型测算了 2014—2022 年山东省内 16 个地市的科技金融效率,并根据相关文件将山东省划分为省会、胶东和鲁南三大经济圈,省会经济圈以济南为中心,涵盖七个地市;胶东经济圈以青岛为中心,包括五个地市;鲁南经济圈包含四个地市。

如图 1 所示,2014—2022 年山东省及三大经济圈的科技金融效率整体上升,其中 2014—2016 年效率较低且波动。2016 年底,山东省出台政策优化金融资源配置,为科技型企业提供多元化融资支持,导致次年科技金融投入增加而产出未即时兑现,因此整体效率有所下降。自 2018 年起,效率增速显著提升。胶东经济圈

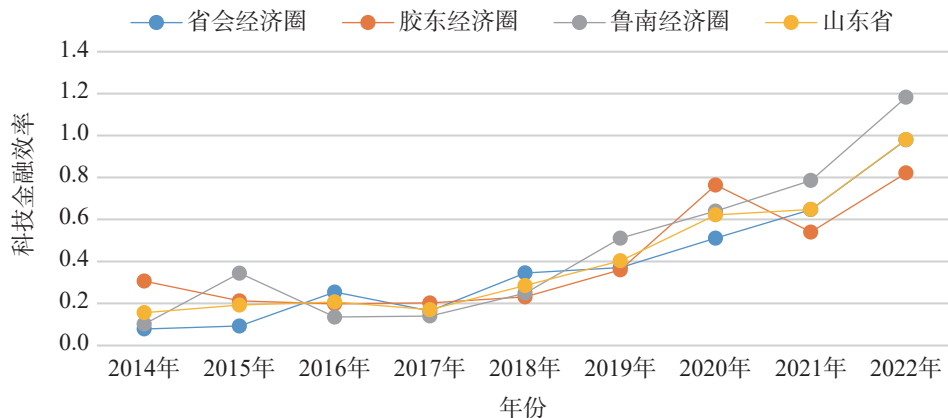


图 1 山东省及省内三大经济圈科技金融效率

①为保证数据的连续性和可比性,本文未将 2019 年 1 月并入济南市的莱芜市纳入研究样本。

科技金融效率在 2019 年和 2020 年翻倍增长，之后虽略有波动但仍保持上升趋势。省会经济圈与鲁南经济圈在效率上交替领先，鲁南经济圈部分年份实现了对省会经济圈的超越。

2 山东省科技金融效率时空差异

为进一步探究山东省科技金融效率的时空异化及其动因，本文采用全局 Malmquist 指数（即 TFP）来分析山东省各地市效率的年际和地域差异，同时将其分解为 EC 和 TC。

2.1 山东省科技金融效率变动的年际变化

如图 2，2016—2022 年山东省与省会经

济圈和鲁南经济圈效率年际变化相似，尽管 2020—2021 年略有下降但效率仍在提升。省会经济圈在 2015—2016 年效率最高，鲁南经济圈在此期间保持稳定。胶东经济圈自 2017 年起效率提升速度加快，尽管 2020—2021 年有所放缓，但仍在提升。2021—2022 年全省及三大经济圈的效率均显著提升。

图 3 显示，山东省及三大经济圈在某些年份的科技金融技术效率变化小于 1，成为阻碍科技金融效率提高的因素。但在其他年份，尽管存在波动，技术效率总体上推动了科技金融效率的提升。

如图 4，从技术进步的角度看，2014—

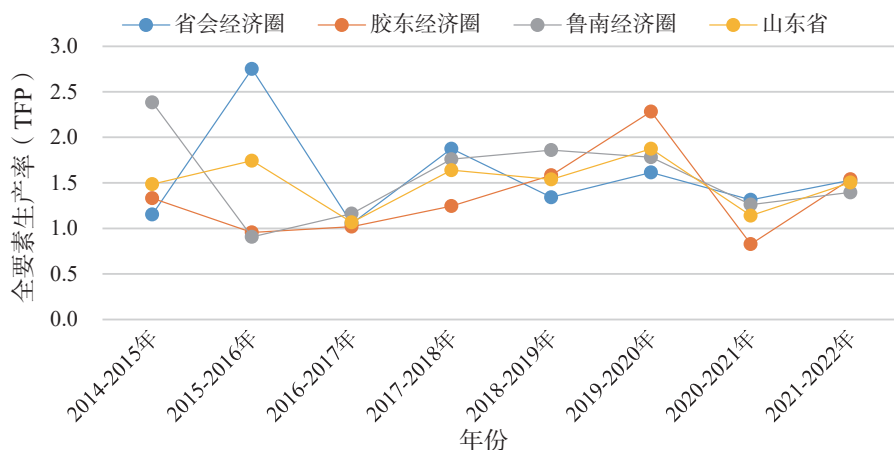


图2 山东省及三大经济圈全要素生产率(TFP)

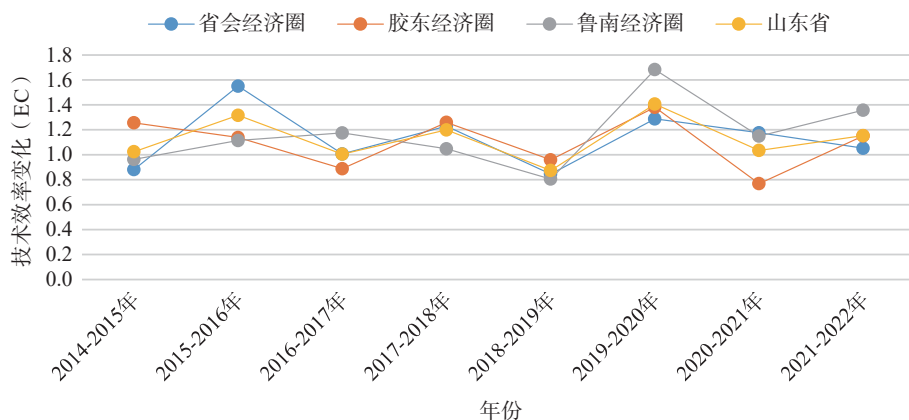


图3 山东省及三大经济圈技术效率变化(EC)

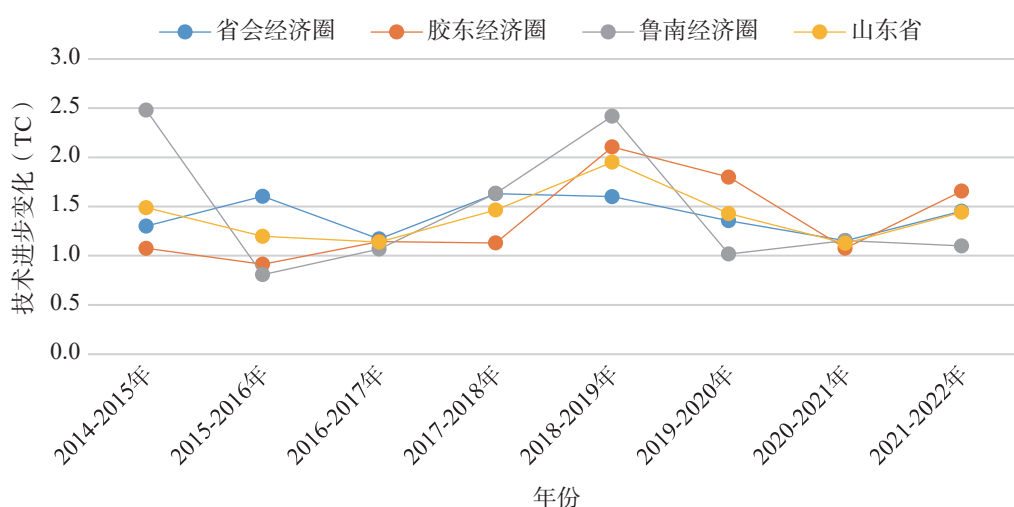


图4 山东省及三大经济圈技术进步变化(TC)

2022年期间,除个别年份外,山东省及三大经济圈的技术进步变化均大于1,对科技金融效率有积极贡献。特别在2018—2019年技术进步对科技金融效率提升起到了决定性作用。

2.2 山东省科技金融效率变动的地域差异

为进一步探究科技金融效率的地域差异,本文计算了2014—2022年山东省各地市的全局Malmquist指数及其分解。表2显示,各地市TFP几何均值均大于1,表明山东省科技金融效率整体上升。其中,聊城市效率增长最显著,年均提升164.6%,主要得益于77.2%的技术进步年均增长率。在全省范围内,16地市的技术进步变化均大于1,而技术效率变化中青岛市、威海市和德州市小于1,因此技术进步是效率提升的主要驱动力。

表2 2014—2022年山东省各地市全局Malmquist指数几何均值及其分解

地市	全要素生产率 TFP	技术效率变化 EC	技术进步变化 TC
济南市	1.377	1.007	1.384
青岛市	1.358	0.986	1.366
淄博市	1.475	1.122	1.345
枣庄市	1.828	1.092	1.876
东营市	1.445	1.064	1.369
烟台市	1.295	1.073	1.290
潍坊市	1.555	1.170	1.441
济宁市	1.342	1.117	1.302
泰安市	1.318	1.242	1.330
威海市	1.246	0.959	1.349
日照市	1.288	1.312	1.366
临沂市	1.332	1.004	1.339
德州市	1.258	0.988	1.314
聊城市	2.646	1.269	1.772
滨州市	1.534	1.208	1.343
菏泽市	1.752	1.434	1.322

3.1 变量选取与数据说明

参考先前学者的研究,将从政府科研资金支持(GA)、企业科研资金支持(RG)、经济发展水平(PG)、金融市场发展程度(DG)、地区技术发展水平(PT)以及数字化程度(DI)

3 山东省科技金融影响因素分析

实证结果显示,山东省科技金融效率整体上升,但各经济圈及地市间存在差异和波动。下面进一步分析影响山东省科技金融效率的因素。

等多个维度，深入探究影响山东省科技金融效率的具体因素。影响因素及量化指标详见表3。

表3 山东省科技金融效率影响因素及量化指标

影响因素	量化指标	单位
政府对科研的资金支持力度（GA）	地方科学技术支出 / 地方财政支出	%
企业对科研的资金支持力度（RG）	R&D 投入 / 地区 GDP	%
经济发展水平（PG）	人均 GDP	万元
金融市场发展程度（DG）	地区各项贷款余额 / 地区 GDP	%
地区技术发展水平（PT）	地区专利授权数 / 全国专利授权数	%
数字化程度（DI）	北京大学数字普惠金融指数中数字化程度指数	/

注：数据来自《山东统计年鉴》《山东科技统计年鉴》及《北京大学数字普惠金融指数》。

3.2 模型构建

本文将科技金融效率作为被解释变量，左侧截断点取0，将上述六个影响因素作为解释变量，构建 Tobit 模型如式（1）所示。

$$TE_{it} = \alpha_i + \beta_1 GA_{it} + \beta_2 RG_{it} + \beta_3 PG_{it} + \beta_4 DG_{it} + \beta_5 PT_{it} + \beta_6 DI_{it} + \varepsilon_{it} \quad (1)$$

其中，i 表示地市，t 表示年份， TE_{it} 为科技金融效率值， α_i 为个体效应， ε_{it} 为随机误差项，其余变量名称含义如表3所示。

3.3 影响因素结果分析

基于 Tobit 模型，对公式（1）进行回归，结果如表4所示。

表4回归结果显示，政府和企业对科研的资金支持力度（GA 和 RG）与科技金融效率负相关，分别在1%和10%水平上显著。金融市场发展程度（DG）和数字化程度（DI）与其正相关，

分别在10%和1%水平上显著。经济发展水平（PG）和地区技术发展水平（PT）回归结果不显著，但对科技金融效率的回归系数为正。

表4 山东省科技金融效率影响因素回归结果

被解释变量	解释变量	系数	标准差	P 值
科技金融效率（TE）	GA	-0.102***	0.032	0.002
	RG	-0.106*	0.055	0.056
	PG	0.012	0.010	0.217
	DG	0.002*	0.001	0.064
	PT	0.047	0.109	0.666
	DI	0.004***	0.001	0.000
	常数项	-0.458***	0.167	0.007

注：***、**、* 分别表示为在1%、5%、10%的水平上显著。

回归结果显示：（1）政府和企业对科研的资金支持均对科技金融效率产生负效应，这表明，单纯增加科研投入并不一定能提升山东省的科技金融效率，可能的原因是投入冗余、资金配置不均及资金使用管理的不完善；（2）金融市场发展程度和数字化程度对科技金融的效率有正面影响，因此完善科技金融市场体系以及利用数字化基础设施对提升山东省科技金融效率至关重要。

3.4 区域异质性分析

为深入探究各要素对科技金融效率影响的区域差异，将山东省分为省会、胶东和鲁南三大经济圈，并分别对这些区域进行科技金融效率与影响因素的回归分析，如表5所示。

表5回归结果显示：（1）在省会经济圈中，政府对科研的资金支持力度（GA）对科技金融效率有负效应，数字化转型（DI）对其有正效应，

但均在胶东和鲁南经济圈中不显著；（2）在鲁南经济圈，企业对科研的资金支持力度（RG）和地区技术发展水平（PT）对科技金融效率负向显著，可能与资源投入冗余或错配有关；金融市场发展程度（DG）对科技金融效率正向显著；（3）金融市场发展程度（DG）在胶东和鲁南经济圈与科技金融效率正相关，分别在 1% 和 5% 水平上显著，但在省会经济圈不显著。

表 5 山东省科技金融效率影响因素的区域异质性			
解释变量	省会经济圈	胶东经济圈	鲁南经济圈
GA	-0.173*** (-4.34)	-0.013 (-0.21)	-0.110 (-0.44)
RG	0.060 (0.63)	-0.096 (-1.23)	-0.673*** (-3.67)
PG	0.005 (0.55)	0.020 (1.20)	0.204** (2.32)
DG	0.000 (-0.00)	0.005*** (3.32)	0.009** (2.30)
PT	0.350 (1.42)	-0.113 (-1.11)	-1.685* (-1.82)
DI	0.004*** (4.48)	0.001 (1.22)	0.001 (0.47)
常数项	-0.808*** (-3.51)	-0.347 (-1.40)	0.047 (0.11)

注：其中，***、**、* 分别表示为在 1%、5%、10% 的水平上显著；括号内为 t 值。

4 研究结论与对策建议

4.1 研究结论

本文探讨科技金融效率的评价机制，通过构建超效率 SBM 模型和全局 Malmquist 指数，测度了山东省各地市科技金融效率及其年际变化和地域差异，为国家各省市科技金融政策布局提供了可借鉴的研究方法。主要研究结论如下。

（1）从年际变化来看，2014—2022 年山东省科技金融效率整体上升，特别是出台相关政策一段时间后，即 2018 年开始科技金融效率增速明显提高，且技术进步对科技金融效率提升起到关键作用。

（2）从地域差异角度，鲁南经济圈效率最高，省会经济圈最低，地区间效率差距显著，经济欠发达的鲁南地区科技金融投入较少但效率却很明显，经济发达的省会和胶东经济圈投入了大量的科技金融资金却未发挥最大的撬动作用。

（3）在影响科技金融效率的因素中，金融市场发展程度和数字化程度对科技金融的效率有正面影响，而投入冗余、资金配置不均及资金使用管理的不完善等原因会造成政府和企业对科研的资金投入没有有效提升科技金融效率。

4.2 对策建议

根据上文研究结论，科技金融效率的提高离不开政策的支持、区域的禀赋优势以及技术的赋能。为了促进国家各省市有效布局科技金融政策，推动金融资源提升地区科技发展水平，本文从以下角度提出对策建议。

4.2.1 统筹科技金融工作，发挥财政资金引导撬动作用

统筹科技金融安排，实现省级统筹、市县个性化联动配合，以降低协作成本并确保政策落地。持续完善科技资金配置机制，转变传统科技计划无偿资助的模式，建立科技金融政策体系，探索建立省级科技金融专项资金，强化市级资金配套保障，通过多样化的手段，如风险补偿、贷款贴息、保费补贴、投资奖补等，

形成省市政策合力,逐步转变政府资金支持方式,以科技金融资金引导撬动社会资本以多元化方式支持企业创新^[7-8]。

4.2.2 加强区域协作,发挥地区优势

强化区域协作、错位发展,避免单打独斗,可提升科技金融效率^[9-11]。以山东为例,发挥济南科创金融改革试验区带动作用,利用省会经济圈依托省会的资源虹吸优势,集聚资源,加强数字化设施建设,搭建综合性科技金融服务平台。发挥胶东经济圈的海洋资源和财富管理中心优势,强化对重点、优势产业的股权、债权等投资支持。鲁南经济圈虽经济总量偏小,产业结构初级,但科技金融潜力大,可加大对该区域的普惠性科技金融支持力度,提升经济和金融市场发展水平可显著提高科技金融效率^[12]。

4.2.3 利用大数据分析,精准匹配企业需求

强化数字化技术手段在科技金融市场中的运用,打造新型科技金融服务平台,打通各类政务信息数据,整合银行、基金、风投、融担等多方资源,通过金融科技挖掘科技型企业的无形资产价值,实现量化信用、多级分险,降低投融资风险,推动科技型企业的“本领”成本钱、“知产”变资产,激发金融资本科技创新的支持。同时,利用数字化技术优化资金配置,避免金融错配^[13-14]。

参考文献

- [1] 胡欢欢,刘传明.中国科技金融效率的区域差异及动态演进[J].统计与决策,2022,38(24):117-122.
- [2] 沈丽,范文晓.我国科技金融效率的空间差异及分布动态演进[J].管理评论,2021,33(1):44-53,67.
- [3] 高扬,王桂葵.山东省科技金融效率影响因素及区域差异研究[J].华东经济管理,2023,37(7):92-99.
- [4] 陈敏,南爱华,刘国华.山东科技金融发展的制约因素及对策研究[J].山东社会科学,2013,220(12):157-161.
- [5] 刘濛,李嫣资.京津冀城市圈科技金融效率测度与优化路径研究——基于模糊集定性比较分析[J].会计之友,2023,698(2):93-100.
- [6] 张远为.湖北省科技金融资源配置效率研究——基于BCC模型和Malmquist指数法[J].湖北社会科学,2021,417(9):58-66.
- [7] NAFF C. Corporate Finance, Innovation, and Strategic Competition[M]. Berlin: Springer Berlin Heidelberg, 2003: 522.
- [8] PREDRAG B. The Financing of Research and Development[J]. Oxford Review of Economic Policy, 2002, 18(1): 35-51.
- [9] 张明喜,魏世杰,朱欣乐.科技金融:从概念到理论体系构建[J].中国软科学,2018(4):31-42.
- [10] 张芷若,谷国锋.科技金融与科技创新耦合协调度的空间格局分析[J].经济地理,2019,39(4):50-58.
- [11] 杜江,张伟科,范锦玲,等.科技金融对科技创新影响的空间效应分析[J].软科学,2017,31(4):19-22,36.
- [12] 温小霓,张哲.基于系统动力学的科技金融支持科技发展研究[J].科学管理研究,2017,35(5):95-99.
- [13] 欧阳卿.科技创新企业投融资联动:国内探索与发展建议[J].科技管理研究,2018,38(3):179-183.
- [14] 周新,马丁.我国科技金融发展效率的影响因素——基于模糊集的定性比较分析[J].科技管理研究,2021,41(6):49-54.



开放科学
(资源服务)
标识码
(OSID)

我国数据要素研究的文献计量与可视化

刘桂锋^{1,2} 言怡¹ 刘琼² 韩牧哲¹

1. 江苏大学科技信息研究所 镇江 212013;

2. 江苏大学图书馆 镇江 212013

摘要: [目的/意义] 数据要素作为关键生产要素,对于推动我国经济社会发展具有重要意义。通过对数据要素研究的可视化分析,旨在为政策制定者、企业管理者和学者提供有关数据要素研究的全面认识,以为数据要素的发展和应用提供智力支持。[方法/过程] 选择了中国知网(CNKI)数据库作为主要的文献检索平台,以“数据要素”“数据要素市场”“数据生产要素”“数据交易”为关键词进行检索。在获取相关文献后,对相关文献的年度发文量趋势、作者和机构的发文分布情况等进行了分析。此外,通过 VOSviewer 软件绘制了图谱,以直观地展现研究领域的研究主题和演变趋势。[结果/结论] 研究发现,数据要素研究在我国已经取得了丰硕的成果,发文量逐年上升,研究热度不断攀升。数据要素研究涉及多个学科,如经济学、法学和管理学等,研究方法日趋多样化。根据关键词共现分析和关键词时区图分析,数据要素和数字经济是研究的核心概念,研究不仅丰富了理论框架,还对实践应用产生了影响,为政策制定和行业实践提供了重要的指导和支持。

关键词: 数据要素;数据科学;数据管理;大数据;信息可视化

中图分类号: G35

Research and Analysis of Domestic Data Elements Based on Knowledge Graph

LIU Guifeng^{1,2} YAN Yi¹ LIU Qiong² HAN Muzhe¹

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

2. Jiangsu University Library, Jiangsu University, Zhenjiang 212013, China

Abstract: [Purpose/Significance] As a key production factor, data factor plays an important role in promoting China's economic and social development. This paper aims to provide a comprehensive understanding of data factor research for policy makers, enterprise managers and scholars through visual analysis of data factor research, in order to provide intellectual

基金项目 国家社会科学基金一般项目“科学数据融合模式设计与体系建构研究”(21BTQ080)。

作者简介 刘桂锋(1980-),博士,教授,主要研究方向为数据管理与数据科学;言怡(1995-),硕士研究生,主要研究方向为数据管理与文本挖掘;刘琼(1986-),硕士,馆员,主要研究方向为阅读推广与学科服务;韩牧哲(1990-),博士,讲师,主要研究方向为数字人文与信息组织。

引用格式 刘桂锋,言怡,刘琼,等.我国数据要素研究的文献计量与可视化[J].情报工程,2024,10(4):114-127.

support for the development and application of data factor research. [Methods/Processes] In this paper, CNKI database is selected as the main literature retrieval platform, and the keywords “data elements” “data elements market” “data production factors” and “data transaction” are used for searching. After obtaining the relevant literature, the annual publication trend of the relevant literature and the distribution of the authors and institutions were analyzed. In addition, the VOSviewer software is used to draw maps to visually show the research themes and evolving trends in the research field. [Results/Conclusions] It is found that the research on data elements has achieved fruitful results in China, the number of papers is increasing year by year, and the research heat is rising. The research of data elements involves many disciplines, such as economics, law and management, and the research methods are increasingly diversified. According to keyword co-occurrence analysis and keyword time zone graph analysis, data elements and digital economy are the core concepts of the research, which not only enriches the theoretical framework, but also has an impact on practical application, providing important guidance and support for policy making and industry practice.

Keywords: Data Elements; Data Science; Data Management; Big Data; Information Visualization

引言

大数据时代催生着数字经济的蓬勃发展，数据已成为世界各国高度重视和争相获取的重要战略性资产^[1]，数据要素作为推动经济社会发展的关键要素，其重要性日益凸显。中国信通院的《数据要素白皮书》指出，数据要素一词面向数字经济，指根据特定生产需求汇聚、整理、加工而成的计算机数据及其衍生形态^[2]。在全球范围内，各国都在积极探索如何充分利用数据要素，促进经济增长、提升产业竞争力，以及改善社会治理能力。而我国作为全球最大的互联网市场和数据生产国，对数据要素研究的重视程度更是日益增强。

在此背景下，文献计量与可视化成为了解数据要素研究进展的重要工具。文献计量方法能够对大量文献进行系统性分析，揭示研究的热点和趋势，为学者们提供研究方向和思路。与此同时，可视化技术则能够直观地展现文献之间的关联和演化过程，为研究者们提供更深入的认识和理解。当前我国在数据要素研究领

域的文献计量与可视化方面还存在一定的不足和挑战，尽管近年来该领域的研究逐渐增多，但相关文献的数量和质量仍有待提升。同时，对于文献计量与可视化方法的运用也尚未充分发挥其潜力，缺乏系统性和深度的研究。因此，本文旨在通过对我国数据要素研究文献的计量分析与可视化呈现，探讨当前研究的现状与趋势，挖掘研究的热点和难点，为学者们提供更深入的研究视角和方法论支持。同时，也旨在为政策制定者和企业管理者提供科学依据和决策参考，推动我国数据要素研究的深入发展，助力数字化转型和经济社会的可持续发展。

1 数据来源与研究方法

1.1 数据来源

本文综合借鉴了现有文献的研究成果，并选择了中国知网（CNKI）数据库作为主要的文献检索平台，旨在全面了解我国“数据要素”领域的研究进展。检索式为“数据要素” or “数

据要素市场” or “数据生产要素” or “数据交易”，自 1985—2024 年共检索出 7272 篇相关文献（见表 1）。为确保文献研究的严谨性和准确性，本文进一步限定了文献来源与时间范围，

选取了收录在 CSSCI 期刊的文献，剔除非完整年份 2024 年的期刊论文，排除了会议记录和科技成果等非期刊文章后，获得了适合进行分析的 2301 篇文献。

表 1 文献检索策略

数据库	检索表达式	时间	期刊来源	文献类型	结果（篇）
CNKI	主题 = “数据要素” or “数据要素市场” or “数据生产要素” or “数据交易”	1998—2023	CSSCI	期刊论文	2301

此外，本文利用 Excel 软件对这些文献进行了文本统计分析，分析内容包括年度发文量趋势、作者和机构的发文分布情况等。为了更直观地展现研究领域的知识结构和演变过程，本文使用了 VOSviewer 软件来进行可视化分析，以展示出关键词的共现聚类结果，揭示研究热点和主题演化的路径。通过上述方法的综合运用，本文旨在从量化和可视化的角度，对“数据要素”领域的学术发展脉络进行深入分析，以期为后续研究提供参考和启示。

1.2 研究方法

VOSviewer 是一种用于可视化和分析科学文献、专利和其他类型的文本数据的软件工具，它可以通过计算文献之间的共现关系或者其他相关性指标，将文献转化为网络或地图，并提供多种可视化和分析功能，帮助用户理解和发现文献之间的关联性、研究热点和趋势，通常用于科学研究领域，特别是文献计量学和科学可视化研究中。在 VOSviewer 中，用户可以根据自己的需求，生成不同类型的网络图或地图，以探索文献之间的关系、发现相关主题和关键词、识别领域内的重要作者或机构等信息。这有助于用户更好地理解研究领域的关联性、研

究热点和趋势。

本文对选中的 2301 条基础数据，通过中国知网“文献导出”选项中的 Refworks 格式，将其导入至 Endnote 软件并创建数据库。由于 VOSviewer 软件无法读取中国知网导出的文件格式，因此使用 Endnote 软件将数据另存为可读取的 RIS 文件格式，并导入 VOSviewer 软件，选中所有数据，进行年度发文量及高频关键词统计。此外，通过软件自带的主题词剔除与合并功能，对意思相近的主题词进行合并处理，并生成发文作者、共词分析、关键词时间图、关键词聚类四类可视化图谱。

2 数据要素研究的可视化分析

2.1 发文时序分析

本文对数据要素研究的发文数量进行统计，旨在把握我国数据要素研究领域的热度和趋势，具体研究方法是利用 Excel 进行文本统计，分析该领域在国内的研究发展状况。通过对数据要素发文时序分析，可以发现我国的研究热度逐渐攀升，随着大数据时代的到来，数据要素的重要性日益凸显，吸引了越来越多的研究者加入该领域。

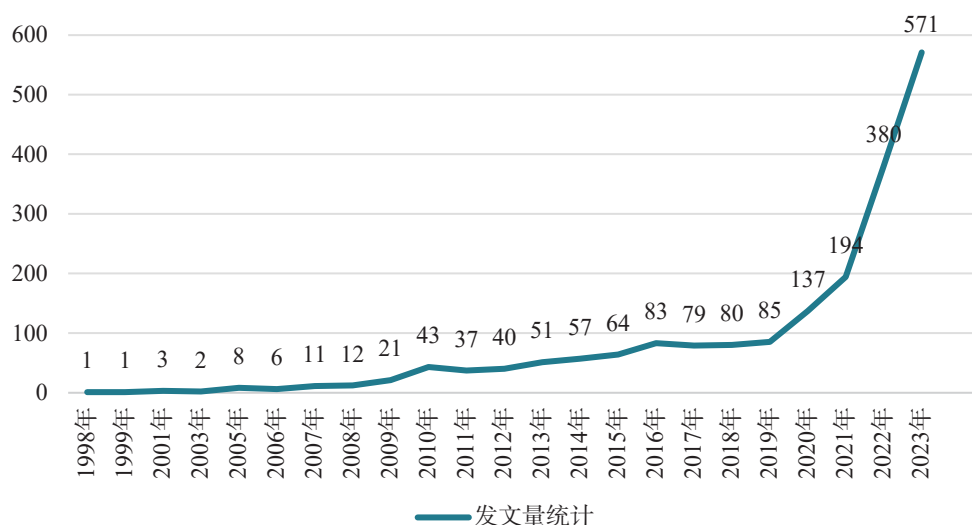


图1 数据要素研究发文量统计图

由图1所示,自2010年以来,我国数据要素研究的文献发表数量呈现稳步上升趋势,尤其在2022年和2023年期间,发文量的激增标志着该领域吸引了更多学者的兴趣,研究热情达到了前所未有的高度,反映出数据要素作为研究主题的普及和深入。数据要素研究横跨多个学科,如计算机科学、经济学、法学与管理学等,这种跨学科的特性丰富了研究的视角和内容,为学者们提供了各自专业领域内的独特研究点。在多学科交叉的推动下,研究方法也日趋多样化,包含理论探讨、实证分析和案例研究等,这些方法相互补充,共同推动了该领域知识体系的深度发展和方法论的创新。但是与数据要素相关的计量与可视化研究文献相对较少,特别是以数据要素为研究对象的文献计量与可视化方面的文献,目前有相关研究的学者包括王妍晴^[3]、周明生等^[4]。

随着大数据产业的蓬勃发展,数据要素的商业价值和社会作用日益凸显,越来越多的企业和机构开始将数据要素作为核心资源进行投资和管理。在实际操作中,它们不断尝试和完

善数据的收集、处理、分析、交易等环节,为数据要素的理论研究提供了实践案例和经验教训。综合来看,我国数据要素研究领域正处于一个快速发展阶段,其研究热度和发展趋势均显示出明显的上升势头。鉴于当前研究的广泛性和深入性,未来在该领域的理论创新和实践应用有望继续增长,这将为我国的社会经济发展提供更加坚实的数据支撑和智力支持。

2.2 发文作者及机构分析

2.2.1 发文作者

本文结合 Excel 及 VOSviewer 软件对文献作者发文数量进行统计(表2),任保平^[5-7]以15篇的发文量位列第一,是研究领域中的主要贡献者之一,其研究成果涵盖了从数据要素市场化配置的法治推进、数字经济与实体经济的深度融合,到新型工业化与经济高质量发展的新动能培育等多个方面。其他主要作者包括陈兵、欧阳日辉、杨东、王建东等,他们的发文量也相对较高,展现了他们在该领域的重要地位和影响力。例如,欧阳日辉^[8]深入分析了数

据要素在数字经济中的理论逻辑和市场配置问题，提出了数据估值定价的新方法和评估指标；陈兵^[9]则集中研究了数据要素权益的分类和竞争法治在数据市场中的推进，相关研究工作不仅丰富了学术理论，也为政策制定提供了实践指导。这些发文作者的贡献不仅体现在发文量上，更重要的是他们通过研究作为数据要素研究领域的发展和进步做出了积极的贡献，他们的研究成果为该领域的学术交流和进步提供了重要支持。

表 2 作者发文数量前 10 排名

排序	作者	发文量（篇）
1	任保平	15
2	陈兵	14
3	杨东	11
4	王建东	9
5	欧阳日辉	9
6	谢康	8
7	夏杰长	8
8	夏义玺	7
9	林春	7
10	张会平	6

Total Link Strength（总的共现次数）是 VOSviewer 软件用来衡量作者之间合作关系强度的指标，一个高的 Total Link Strength 表明该作者与其他研究者有较多的合作，且在学术网络中占有重要位置。通过对文献作者 Total Link Strength（表 3）进行统计，尹西明、张会平、林镇阳、廖祖君等作者在数据要素市场化配置、数字经济、数据基础设施建设等领域内具有显著的学术影响力，他们的研究工作在推动理论研究和实践应用方面表现出色。

表 3 VOSviewer 的作者 Total Link Strength 前 10 排名

排序	作者	Total Link Strength
1	尹西明	27
2	陈劲	26
3	张会平	25
4	谢康	23
5	林镇阳	23
6	窦悦	23
7	杨晓光	22
8	黄倩倩	21
9	廖祖君	20
10	王建东	19

尹西明作为该列表中 Total Link Strength 最高的作者，与多篇文献的其他作者有合作。陈劲也是一个活跃的作者，他的研究集中在科技创新、新质生产力发展、数据基础设施等领域，他与尹西明在多篇文献中共同出现，表明他们在研究上有紧密合作^[10-11]。张会平^[12-14]的研究更侧重于公共数据的治理、运营和市场化流通，以及国家治理现代化的角度。谢康^[15-17]的研究涉及企业数字化转型、数据化互动创新、数据驱动的农业数字化转型等领域，在企业层面的数据应用和创新研究领域有着丰富的贡献。

深入分析这些作者的文献，可以发现几个关键主题和趋势：第一，数据要素市场化配置涉及数据要素如何在市场中流通、定价和交易，以及政策、法规和技术框架如何支持这一过程；第二，数字基础设施建设研究如何构建支撑数字经济的基础设施，例如数据中心、网络等，并探索这些基础设施如何促进社会经济发展；第三，企业数字化转型探讨企业如何通过利用数据和数字技术转变其业务模式、生产流程和管理方式，以提高效率和创造新的价值；第四，

公共数据治理研究政府和公共机构如何管理和运营公共数据资源，以及如何在保护隐私和安全的同时实现数据的开放和共享；第五，数字经济与国家治理分析信息技术如何推进国家治理现代化，以及数字经济在全球价值链中的趋势和角色。

综合上述信息，可以看出这些作者在推动数据要素理论与实践方面发挥了关键作用，特别是在数据要素的市场化配置和数字化转型管理方面的研究贡献显著。通过阅读这些作者的论文，可以进一步了解每个研究主题的深度和广度，以及他们如何通过学术合作推动领域进步。

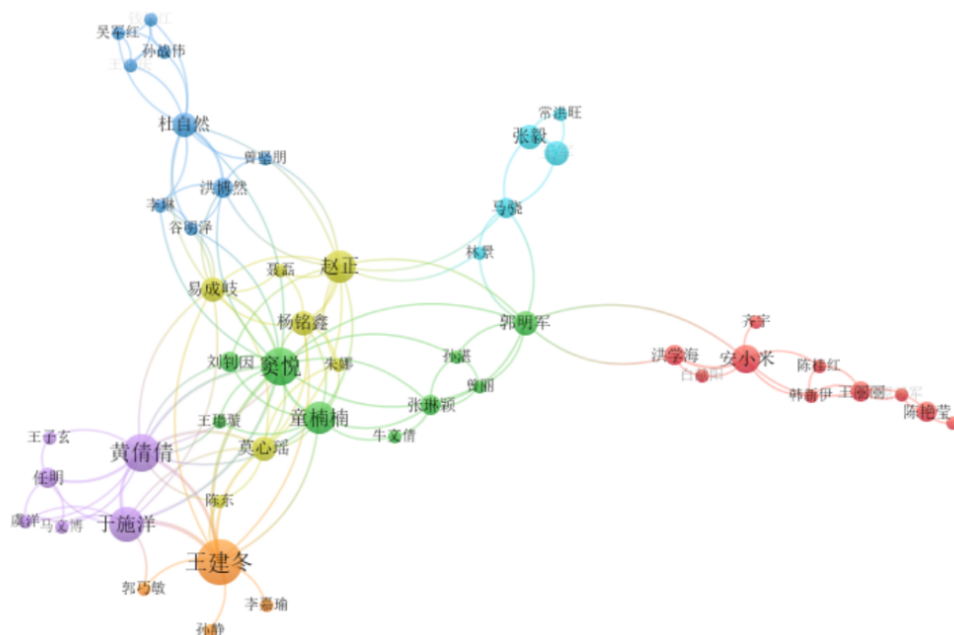


图2 数据要素研究作者的合作网络

需要注意的是，本文在之前的分析中提到了尹西明、陈劲、张会平、谢康等作者，他们在数据要素市场化配置、数字经济、数据基础设施建设等领域的贡献显著，这些作者在VOSviewer前期分析中显示出较高的Total Link Strength，说明他们与其他研究者有较多的合作，并在学术网络中占有重要位置。而在图2中，出现了王建冬、黄倩倩、于施洋等作者作为数据要素研究领域的代表人物，这些作者名字的字号较大意味着他们在该领域发表的文献数量较多，而作者之间的粗连线则表明他们之间的合作关系较为紧密。

总体而言，通过VOSviewer软件的作者共现分析，可以识别出这些领域内的关键研究者，以及他们之间的合作关系，揭示了研究网络内的主要人物和关键联结，有助于理解学术社区的结构和研究趋势。此外，新出现的研究者如王建冬、黄倩倩、于施洋等人也在他们的研究领域内贡献了大量的文献，并与其他研究者建立了紧密的合作关系，表明数据要素研究领域是活跃且不断发展的，新的研究者和新的研究成果不断涌现。通过对这些作者的文献进行深入分析，可以了解到数据要素市场化配置、数字基础设施建设、企业数字化转型、公共数据

治理以及数字经济与国家治理等关键主题的研究动态和进展,这些研究不仅丰富了理论框架,也对实践应用产生了影响,为政策制定和行业实践提供了重要的指导和支持。

2.2.2 发文机构

从表4可见,高校和研究机构的论文发表数量反映了他们在数据要素研究领域的学术活跃度和影响力。特别是,西安交通大学经济与金融学院以27篇的发文数量位居榜首,显示了该院在数据要素研究领域的领先地位。这些研究机构分布在我国不同地区,包括西北地区的西北大学经济管理学院、华北地区的北京大学法学院、东部地区的南开大学经济学院以及南部地区的暨南大学经济学院等,表明数据要素研究在中国各地均得到了重视,并且形成了较为全面的研究网络。

表4 发文机构前15排名

排序	发文机构	发文数量(篇)
1	西安交通大学经济与金融学院	27
2	西北大学经济管理学院	16
3	北京大学法学院	13
4	南开大学经济学院	13
5	南开大学法学院	12
6	中国政法大学民商经济法学院	12
7	湖南大学经济与贸易学院	11
8	清华大学法学院	11
9	中国人民大学法学院	11
10	中国社会科学院法学研究所	11
11	西南政法大学民商法学院	10
12	中国社会科学院财经战略研究院	10
13	武汉大学法学院	9
14	暨南大学经济学院	8
15	北京师范大学经济与工商管理学院	7

西安交通大学经济与金融学院不仅发文数量多,而且涉及的主题也广泛,从数实融合、数据产权、数字化转型等方面均进行了深入研究^[18-20]。

在论文主题方面,可以看出数据要素的流动性、数据财产权、数字经济与实体经济的融合是当前研究的热点。例如,高郅梅^[21]、师博等^[22]探讨了数据交易和数字经济发展的法律和经济问题。

此外,北京大学法学院和中国政法大学民商经济法学院的研究聚焦在数据要素的法律规制上,例如张鼎^[23]研究了数字要素流动的责任规则,冯晓青^[24]研究了数据产权法律构造,表明在数字时代背景下,法律领域对于数据流动、数据财产权以及个人信息保护等问题的深入思考。

总的来说,这些机构通过发表论文在学术界和实践领域推动了对数据要素研究的认识,为我国乃至全球数字经济的发展提供了理论支持和政策建议,他们的研究不仅是学术讨论的产物,更对产业实践和政策制定产生深远影响,这些机构之间的合作网络促进了知识的积累和创新,共同推动了数据要素研究领域的发展。

3 研究热点与趋势分析

3.1 共词分析

共词分析是一种文献计量学方法,用于探究科研文献中关键词之间的关系,尤其是它们在文献中共同出现的次数和模式,有助于识别科研领域的研究热点和趋势,以及不同研究主题之间的联系。

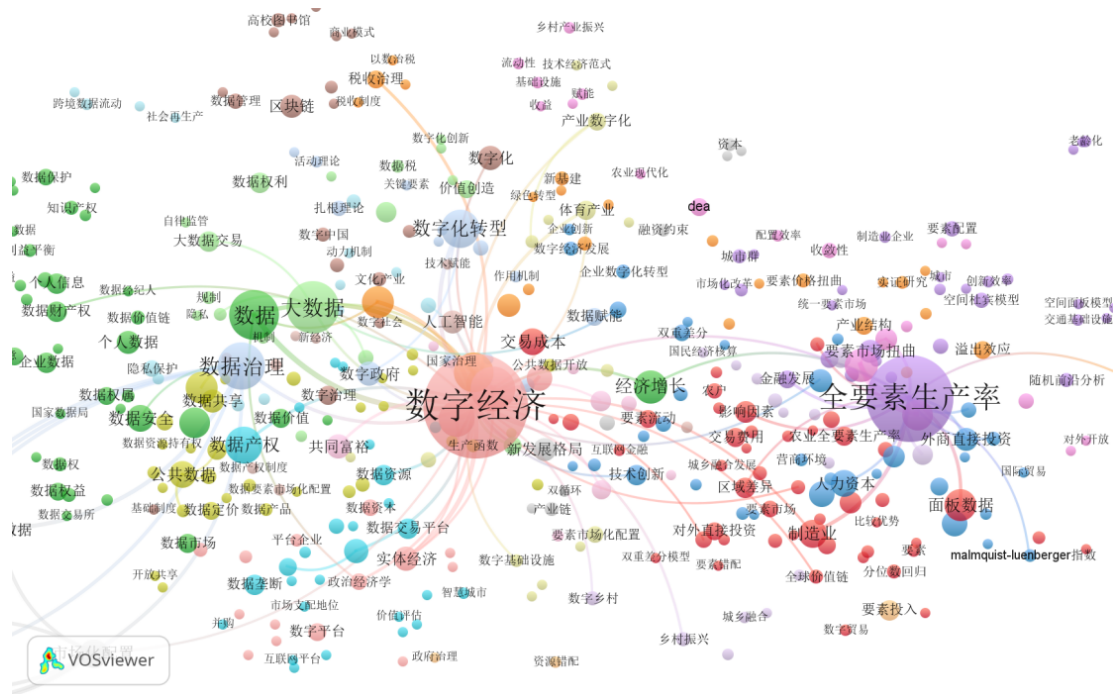


图 3 数据要素研究的关键词共现分析

通过 VOSviewer 软件剔除主题词数据要素并生成共现图谱（图 3），在该视图中，圆点面积越大、越居中，代表该关键词的热度与重要程度越高。通过共现图谱可以直观地看出在国内数据要素研究领域中，出现了“数字经济”“全生产要素”“数据治理”等主要关键词。为更为清晰地把握核心热点，本文对数据要素研究的前 5 位高频关键词进行了顺序排列，如表 5 所示。

表 5 关键词类团

序号	类团	Total Link Strength
1	数字经济	658
2	全要素生产率	296
3	数据治理	192
4	数据交易	167
5	大数据	148

（1）数字经济类团

数字经济类团主要研究数字经济的发展趋

势、影响因素以及数字化转型对经济的影响，涉及经济学、信息技术和数据科学等学科。在经济学领域，涉及宏观经济政策、产业结构调整等方面的研究；在信息技术领域，涉及大数据处理技术、数据分析方法等方面的研究；在数据科学领域，涉及数据挖掘、机器学习等方面的研究。例如，司聪等^[5]研究了数字经济培育中国装备制造业高质量发展新动能的路径。谢康等^[15]研究了大数据驱动的农业数字化转型与创新。

（2）全要素生产率类团

全要素生产率类团主要研究生产要素的组合和配置对经济增长和生产效率的影响，主要与经济学和管理学相关。在经济学领域，涉及生产函数模型、全要素生产率计量方法等方面的研究；在管理学领域，涉及企业管理、生产效率提升策略等方面的研究。例如，孙建军等^[25]

研究了信息资源管理学科视角下的数据要素。尹西明等^[26]研究了平台企业加速数据要素向新质生产力转化的逻辑与进路。

（3）数据治理类团

数据治理类团主要研究数据治理的规范和机制，探讨数据市场建设和数据产权保护的问题，涉及信息法律、政治经济学等学科。在信息法律领域，涉及数据隐私法规、数据保护法律等方面的研究；在政治经济学领域，涉及数据市场竞争政策、数据产权保护政策等方面的研究。例如，范佳佳^[27]探讨了公共数据在数据市场中的应用模式，提出了相关的治理机制。赵蕊菡等^[28]研究了公共数据治理的政策框架，对数据治理的规范机制提供了分析。

（4）数据交易类团

数据交易类团主要研究数据流通和数据安全对数据交易过程的影响，探讨数据交易的模式和规范，主要涉及信息技术和网络安全等学科。在信息技术领域，涉及数据传输技术、数据交易平台开发等方面的研究；在网络安全领域，涉及数据隐私保护、数据加密技术等方面的研究。例如，段巍等^[29]探讨了数据交易中的安全问题，提出了激励机制和治理模式。张勇等^[30]研究了数据交易中的安全和合规问题，提出了软法治理的模式。

（5）大数据类团

大数据类团主要研究大数据的应用场景和数据确权的问题，探讨大数据交易的模式和技术，主要涉及信息技术和数据科学等学科。在信息技术领域，涉及大数据存储技术、数据处理技术等方面的研究；在数据科学领域，涉及数据挖掘算法、大数据分析方法等方面的研究。

例如程琳等^[31]研究了基于区块链的大数据定价机制及交易模式，胡燕玲^[32]研究了大数据交易现状与定价问题。

3.2 关键词时间图

通过分析我国数据要素研究的关键词时间图，可以进一步理解该领域的历史发展和当前的研究态势。结合图1的年度发文量趋势图和图4的关键词时间图，可以观察到以下几个研究阶段：

（1）初始探索期（1998—2009年）

数据要素作为一个完整词汇第一次出现是在《制订美国国家标准前期工作中数字制图数据要素的研究》一文中，但是并非广泛意义上的数据要素。此后陆续几年的研究中，出现了数据或要素等词。在1998年至2009年间，数据要素研究处于较为初期的探索阶段，发文量极少，反映出当时该领域的研究基础较弱，社会和经济对数据价值认识有限。关键词集中在“信息化”“知识管理”等较为广泛的概念上，缺少对“数据要素”本身深入的讨论和研究。

（2）逐步发展期（2010—2019年）

随着互联网和大数据技术的快速发展，数据的重要性开始被广泛认识，数据要素研究逐渐受到关注。2015年我国发布《促进大数据发展行动纲要》，标志着数据成为国家基础性战略资源。这一时期，关键词的出现开始多样化，包括“大数据”“云计算”“物联网”，以及与行业应用相关的词汇，如“智能制造”“电子商务”。

（3）高速发展期（2020—2023年）

从2020年开始，数据要素研究进入高速发

展期，发文量显著增长，这与全球经济数字化转型的大趋势密切相关。关键词围绕“人工智能”“机器学习”“区块链”等新兴技术领域展开，同时“数据治理”“数据安全”“隐私保护”等问题也受到重视。与数字经济相关的关键词，如“数字化转型”“数字货币”“数据交易市场”等开始出现，反映了对经济领域数据要素角色的深入探讨。

在图 4 中，可以观察到不同时间段内的热

点关键词的变化。通过图谱可以看到早期的研究更加集中在经济学的传统领域，例如要素生产，涉及生产要素如劳动、资本、土地和企业管理等，反映了对经济增长、劳动生产率和资源分配等基础议题的关注。随着技术的发展和社会的变革，新的研究领域随之出现并成为焦点，例如数字经济、数据产权等。这些领域的出现和扩张代表了经济研究的新方向和新挑战。

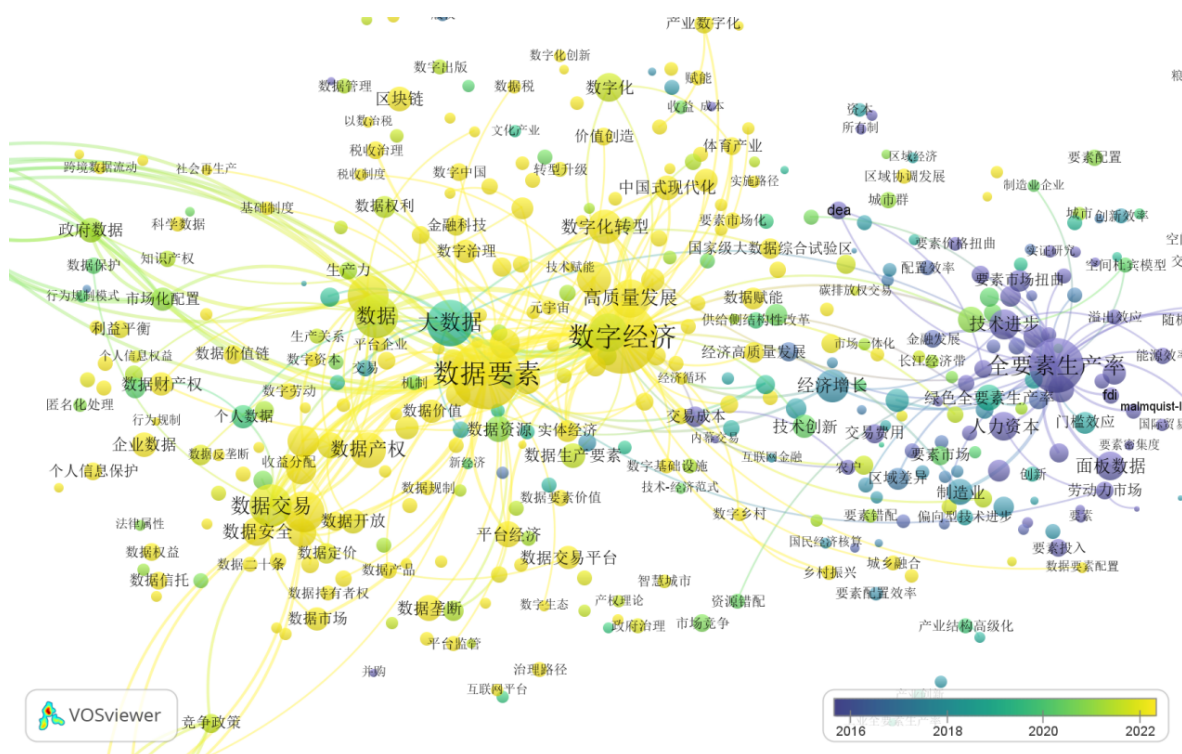


图 4 数据要素研究的关键词时间图

3.3 关键词聚类与研究热点分析

关键词聚类与研究热点分析对于把握数据要素领域的研究动态具有重要意义，通过运用 VOSviewer 对关键词的共现关系进行可视化分析，研究人员能够掌握不同研究主题之间的关联和演进趋势。基于此进行聚类图谱的绘制，得到图 5。在关键词共现聚类图中，同种颜色的关键

词标签即代表一个聚类，也是研究的一个主题。

将国内数据要素相似主题合并后，得到五大类研究主题，分别是：数据要素类、数字经济类、全要素生产率类、数据交易类、数据治理类（表 6）。需要注意的是，由于导出后的图为静态图片，受平面空间限制，部分关键词在图中未能显示。

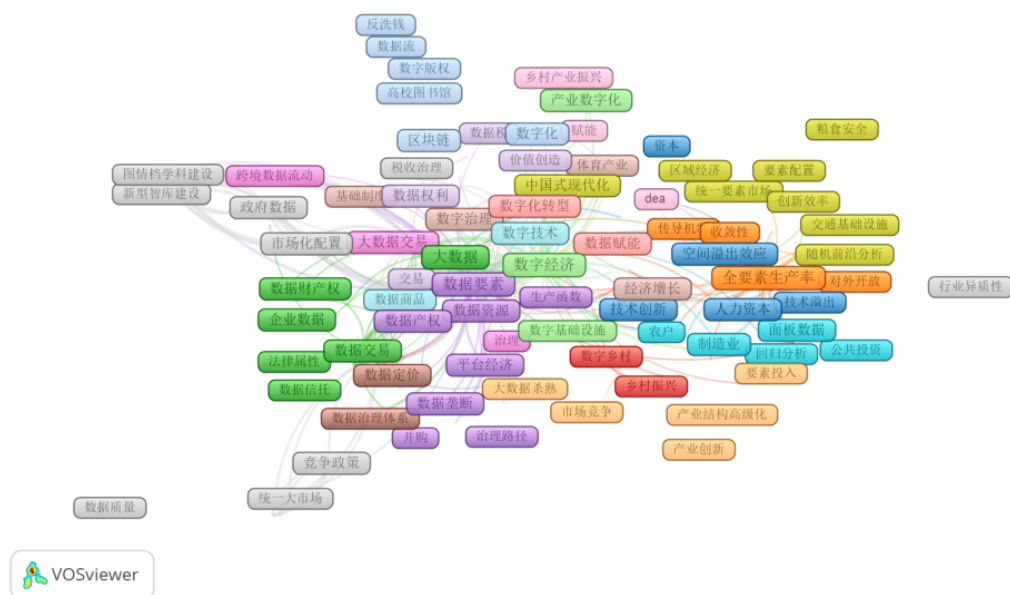


图5 数据要素研究的关键词共现聚类

表 6 数据要素研究的关键词聚类分析

聚类编号	聚类名	主要关键词
1	数据要素	数据产权、数据资源、数据垄断、数据治理等
2	数字经济	数据价值、数字基础设施、高质量发展、经济循环等
3	全要素生产率	生产要素配置、配置效率、技术效率、要素禀赋结构
4	数据交易	数据财产权、数据保护、数据权属、数据市场等
5	数据治理	政府数据、数据资产、数据供给、数据质量等

(1) 数据要素类

数字经济的发展使数据成为关键生产要素，对经济增长和社会发展至关重要。数据要素作为可独立使用的最小数据单位，其管理、交易机制和流通是当前研究的重点。数据要素被视为新的资产类别，其价值认知和计量标准尚在探索中。新结构经济学和信息资源管理学科为数据要素评估和管理提供了理论基础。数据要素的市场化配置是提高生产效率的关键，政策制定者和企业需了解其交易机制和流通过程，

以优化数据资源利用。数据要素通过数据驱动的创新能显著促进实体经济发展，尤其在医疗健康领域能改变治疗和管理方式，提升服务质量并降低成本。然而，数据要素的合法和伦理使用面临制度和法律挑战，需要政策和法律的支持和指导。

（2）数字经济类

数字经济是利用数字计算技术进行生产、销售和交易的经济活动，已成为现代经济的核心。它强调知识和信息作为重要的生产要素，并推动创新和效率提升。数据要素在此框架下被视为关键生产力，促进经济数字化和智能化发展。有效的数据治理策略确保数据安全、隐私保护和合理利用，进而促进数字经济的可持续发展。大数据分析能力对市场趋势、消费者行为和供应链优化至关重要。市场化配置指数数据通过市场机制有效收集、交易、处理和应用，与新的生产力密切相关。研究涵盖了服务业发展、经济效率提升以及数据要素市场化配置等

多个方面。这些研究突出了数据要素对经济增长和社会福祉的重要性,同时也提出了数字经济发展中的挑战,如数据安全、隐私保护和数字鸿沟的问题。

(3) 全要素生产率类

全要素生产率衡量经济体在技术和生产效率方面的进步,数据要素成为其关键组成部分。研究表明,通过数据分析和应用,数据要素能够提升全要素生产率,促进经济增长。相关研究包括数据要素促进产业结构优化升级、农业领域的绿色全要素生产率提升、合作社对家庭农场的影响、数字普惠金融对农业全要素生产率的正面影响、大数据发展对企业全要素生产率的影响、制造业服务化的影响、人工智能和技术复杂度的作用等,数据要素在农业、制造业、服务业等各领域均显著提升全要素生产率。

(4) 数据交易类

数据交易市场的发展引起了广泛关注,涉及数据的买卖、共享、隐私保护、安全性、交易规范和价值评估等多个方面。相关研究包括数据交易市场的基本架构、数据交易平台现状及其不足、数据交易政策的发展趋势和议题、公共数据进入市场的有效模式和挑战。数据交易中的违约问题、刑事合规管理、数据安全激励和平台治理等方面。总体而言,数据交易市场的发展涉及多学科领域,需要综合考虑经济学、法学、信息科学等各方面因素,特别是在确权、定价、安全和合规性方面面临着重大挑战和讨论。

(5) 数据治理类

我国数据治理研究取得了显著进展。主要研究内容包括公共数据治理政策的协同机制、

公共数据授权运营的立法路径、大数据时代的数据垄断问题、县域公共数据开发利用的治理困境、数字乡村建设中的数据治理、美国企业数据共享实践的借鉴、欧盟数据中介制度的比较研究、高校数据治理的关键要素和作用机制,以及产业数据安全治理。研究涵盖了从宏观政策到微观实践、国际比较、区域应用及具体行业问题。尽管取得了许多成果,但仍面临跨部门、跨层级协同机制和标准化与国际接轨等挑战。

4 研究结论与展望

本文运用VOSviewer和Excel等相关计量、统计工具对国内数据要素研究的文献进行深入分析,得出以下结论:

第一,数据要素研究在我国已经取得了丰硕的成果,发文量逐年上升,显示出该领域的活跃性和重要性。同时,研究数据要素的学者和机构遍布全国,形成了全面的研究网络,这种跨学科的研究特点为数据要素领域提供了丰富的研究视角和方法论创新。第二,数据要素研究涉及多个学科,如计算机科学、经济学、法学和管理学等,这些学科的交叉融合不仅丰富了数据要素研究的理论框架,还推动了研究方法的创新。从理论探讨、实证分析到案例研究,各种方法相互补充,共同推动了该领域知识体系的深度发展和方法论的创新,但是与数据要素相关的计量与可视化研究文献相对较少。第三,数据要素的商业价值和社会作用日益凸显,越来越多的企业和机构开始将数据要素作为核心资源进行投资和管理,这些实践为数据要素的理论研究提供了实践案例和经验教训,同时也为政策制定和企业实践提供了重要的参

考。第四,在数字化转型的背景下,研究人员对于如何有效开发和利用数据要素的浓厚兴趣,关键词“数据要素”和“数字经济”作为研究的核心概念,显示了这两个概念在当下研究中的普遍性和流行程度,进一步证实了数据要素和数字经济在研究领域的重要地位,也预示着这两个方向将继续成为未来研究的热点。第五,关键词共现聚类结果表明,目前我国数据要素研究分为五大类研究主题,分别是数据要素类、数字经济类、全要素生产率类、数据交易类、数据治理类,这些研究主题相互联系,相互促进,共同推动数据要素的蓬勃发展。

数据要素作为数字经济时代的重要组成部分,其研究仍处于不断发展之中。首先,数据要素研究涉及多个学科领域,如计算机科学、经济学、法学和管理学等。未来,应该进一步加强跨学科合作,促进不同学科领域之间的交流与合作,这样的合作可以丰富数据要素研究的理论框架,推动研究方法的创新,形成更加完善的数据要素研究体系。其次,随着数字化转型的加速推进,研究人员对于数据要素的商业价值和社会作用的关注不断增加。未来的研究应该更加注重理论与实践的结合,充分挖掘数据要素在实际应用中的潜力和效果。同时,高校应该加强与企业和政府的合作,深入探讨数据要素的商业模式和政策支持,促进数据要素的有效利用和管理。最后,数据要素研究的最终目的是促进数字经济的发展和社会进步。因此,未来的研究应该更加注重研究成果的社会影响力和实际应用效果。研究者们应该积极参与社会实践,为政策制定者和企业提供理论支持和实践指导,推动数据要素研究成果的转

化和应用,为数字经济的发展和社会治理提供有力支持。

参考文献

- [1] 王雪,夏义堃,裴雷.国内外数据要素市场研究进展:系统性文献综述[J].图书情报知识,2023,40(6):117-128.
- [2] 中国信息通信研究院.数据要素白皮书(2022)[R/OL].[2023-03-20].www.caict.ac.cn/kxyj/qwfb/bps/202301/t20230107_413788.htm.
- [3] 王妍晴,罗有敢,胡胜男.我国数据治理研究领域的知识图谱可视化分析[J].国家图书馆学刊,2023,32(5):101-113.
- [4] 周明生,周珺.我国数据要素研究的热点与演进趋势[J].河南大学学报(社会科学版),2023,63(4):13-19,152.
- [5] 司聪,任保平.数字经济培育中国装备制造业高质量发展新动能的路径探析[J].贵州社会科学,2024(1):131-138.
- [6] 任保平,王思琛.新发展格局下我国数据要素市场治理的理论逻辑和实践路径[J].天津社会科学,2023(3):81-90.
- [7] 洪银兴,任保平.数字经济与实体经济深度融合的内涵和途径[J].中国工业经济,2023(2):5-16.
- [8] 欧阳日辉.数据要素促进数字经济和实体经济深度融合的理论逻辑与分析框架[J].经济纵横,2024(2):67-78.
- [9] 陈兵,赵秉元.数据要素市场高质量发展的竞争法治推进[J].上海财经大学学报,2021,23(2):3-16,33.
- [10] 尹西明,陈劲,林镇阳,等.数字基础设施赋能区域创新发展的过程机制研究——基于城市数据湖的案例研究[J].科学学与科学技术管理,2022,43(9):108-124.
- [11] 尹西明,陈劲,王华峰,等.强化科技创新引领加快发展新质生产力[J/OL].科学学与科学技术管理,1-10[2024-05-08].<http://kns.cnki.net/kcms/detail/12.1117.G3.20240221.1012.002.html>.
- [12] 聂耀昱,尹西明,林镇阳,等.数据基础设施赋能碳达峰碳中和的动态过程机制[J].科技管理研究,2022,42(18):182-189.

- [13] 张会平, 马太平. 城市全面数字化转型中数据要素跨界流动: 四种模式、推进逻辑与创新路径 [J]. 电子政务, 2022(5): 56-68.
- [14] 杨庐峰, 张会平. 数字经济与实体经济深度融合发展的着力向度与治理创新——以贵州省的融合发展实践为例 [J]. 理论与改革, 2021(6): 140-150.
- [15] 谢康, 肖静华, 王刊良, 等. 企业高质量数字化转型管理: 理论前沿 [J]. 管理学报, 2024, 21(1): 1-9.
- [16] 谢康, 易法敏, 古飞婷. 大数据驱动的农业数字化转型与创新 [J]. 农业经济问题, 2022(5): 37-48.
- [17] 谢康, 吴瑶, 肖静华. 基于大数据合作资产的适应性创新——数字经济的创新逻辑 (二) [J]. 北京交通大学学报 (社会科学版), 2020, 19(2): 26-38.
- [18] 王建冬, 童楠楠. 数字经济背景下数据与其他生产要素的协同联动机制研究 [J]. 电子政务, 2020(3): 22-31.
- [19] 韩奇, 杨秀云. 公共数据开放能否促进企业数字化转型? [J]. 现代经济探讨, 2024(4): 44-59.
- [20] 马黄龙, 屈小娥. 数字经济赋能全球价值链韧性: 理论机制和实践路径 [J]. 现代经济探讨, 2023(11): 65-71.
- [21] 高邴梅. 数据交易违约可得利益的类型化研究 [J]. 财经法学, 2024(2): 37-51.
- [22] 师博, 魏倩倩. 数字经济与实体经济融合的制度安排 [J]. 经济与管理评论, 2024, 40(2): 17-28.
- [23] 张鼎. 数字要素流动的责任规则: 实践与原理 [J]. 甘肃政法大学学报, 2024(2): 76-93.
- [24] 冯晓青. 数据产权法律构造论 [J]. 政法论丛, 2024(1): 120-136.
- [25] 孙建军, 李阳. 信息资源管理学科视角下的数据要素研究 [J/OL]. 图书情报知识, 1-7 [2024-05-08]. <http://kns.cnki.net/kcms/detail/42.1085.G2.20240407.1731.004.html>.
- [26] 尹西明, 钱雅婷, 武沛琦, 等. 平台企业加速数据要素向新质生产力转化的逻辑与进路 [J]. 技术经济, 2024, 43(3): 14-22.
- [27] 范佳佳. 公共数据进入数据要素市场模式研究 [J]. 信息资源管理学报, 2024, 14(2): 68-81.
- [28] 赵蕊菡, 陈俊蕾, 张潇月. “周期 - 工具 - 主体”协同视角下我国公共数据治理政策解析 [J/OL]. 情报资料工作, 1-19 [2024-05-08]. <http://kns.cnki.net/kcms/detail/11.1448.G3.20240408.0930.003.html>.
- [29] 段巍, 袁健. 数据安全视角下数据交易的激励机制——兼论数据交易平台治理模式 [J]. 上海经济研究, 2024(2): 54-69.
- [30] 张勇, 王杰. 数据交易安全刑事合规的“软法”治理及清单模式 [J]. 河南社会科学, 2024, 32(3): 10-19.
- [31] 程琳, 黄广玉, 王政, 等. 基于区块链的大数据定价机制及交易模式探究 [J]. 全国流通经济, 2019(22): 132-135.
- [32] 胡燕玲. 大数据交易现状与定价问题研究 [J]. 价格月刊, 2017(12): 16-19.