doi:10.19920/j.cnki.jmsc.2023.05.001

面向大数据管理决策研究的全景式 PAGE 框架①

陈国青[1], 张 维[2], 任之光[3], 管 悦[4*], 卫 强[1]

(1. 清华大学经济管理学院, 北京 100084; 2. 天津大学管理与经济学部, 天津 300072;
3. 国家自然科学基金委员会管理科学部, 北京 100085; 4. 中国传媒大学经济与管理学院, 北京 100024)

摘要: 大数据和人工智能的飞速发展增加了管理决策领域的复杂性和动态性,也为研究和应用提供了广袤的拓新空间. 全景式 PAGE 框架作为大数据管理决策研究的方向性框架,在大数据问题特征的映射下形成了一个 4×3 要素矩阵,在理论范式(P)、分析技术(A)、资源治理(G)、使能创新(E)的研究探索中发挥着重要作用. 本文结合 NSFC 大数据重大研究计划的开展,进一步阐释全景式 PAGE 框架的内涵和关系逻辑,并面向领域情境讨论重要科学问题、求解路径和新知贡献. 本文旨在从大数据管理决策研究的视角为大数据等颠覆性信息技术应用的相关研究提供前沿方向和创新突破上的启迪;同时为管理学科及其相关学科领域开展交叉属性强、应用情境差异化特点突出的科研攻关提供参考.

关键词: 大数据; 管理决策; 全景式 PAGE 框架; 决策范式; 研究范式

中图分类号: C931 **文献标识码**: A **文章编号**: 1007-9807(2023)05-0004-19

0 引 言

近年来,大数据、人工智能等新兴技术蓬勃发展,并与实体经济和社会生活各个层面深度融合. 电子商务、平台经济等领域不断涌现出新模式与新业态,工业互联网、智能制造等领域全面加速,数字经济成为我国新发展格局的重要形态和新动能. 我国 2022 年数字经济发展取得新的突破,规模达到 50.2 万亿元,占 GDP 比重达到 41.5%[1].

大数据是推动技术发展的重要引擎,是数字经济发展的关键生产要素. 我国近年来针对大数据发展这一议题持续推出系列纲领性文件并做出重要部署. 2015 年,党的十八届五中全会提出实施“国家大数据战略”,国务院印发《促进大数据发展行动纲要》;2017 年,推动互联网、大数据、人工智能和实体经济深度融合在中国共产党十九大报告中被进一步强调;2019 年,第十三届全国人大第二次会议强调深化大数据、人工智能等研发应用,为制造业转型升级赋能,壮大数字经济;2020 年,数据作为与土地、劳动力、资本、技术等传统要素并列的生产要素之一被写入中央政策文件;2021 年,《十四五规划和 2035 年远景目标纲要》提出打造数字经济新优势,要求“充分发挥海量数据和丰富应用场景优势”,“催生新产业新业态新模式”,并将大数据、人工智能等设定为数字经济重点产业. 2022 年,中国共产党二十大报告进一步强调要加快建设网络强国和数字中国.

大数据在给社会经济生活带来深刻变革的同时,也给管理与决策研究带来一系列新的重要科学问题. 许多传统的管理变成或正在变成对于数据的管理,许多传统的决策变成或正在变成基于数据分析的决策[2]. 面向国家重大战略发展和民

① 收稿日期: 2021-12-31; 修订日期: 2022-12-26.
基金项目: 国家自然科学基金资助项目(92246001; 72202220; 91846000).
通讯作者: 管 悦(1993—), 女, 黑龙江双鸭山人, 博士, 讲师. Email:yueguan@cuc.edu.cn

生需求牵引，结合大数据研究和应用的跨学科多领域的交叉融通特点，国家自然科学基金委员会（NSFC）于 2015 年启动了“大数据驱动的管理与决策研究”重大研究计划（简称大数据重大研究计划），旨在围绕大数据驱动的管理与决策所呈现出的新特性，充分发挥管理、信息、数理、医学等多学科合作研究的优势，着重研究大数据驱动的管理与决策理论范式、大数据分析方法与支撑技术、大数据资源治理机制与管理、大数据管理与决策使能及价值创造等重要课题.

“大数据”和“管理决策”概念涉及多学科门类（如哲学、历史、经济、管理、工程、教育、农学、艺术、理学、法学、医学等）、多领域情境（如商务业态、经济金融、医疗健康、公共管理、交通、能源、教育、制造等）、多数据模态（如数值、文本、图像、语音、视频等），使得大数据管理决策相关的基础研究面临着严峻挑战. 以大数据重大研究计划为例[3]，首先，从认识论和方法论的视角，什么样的问题可称之为“大数据问题”？针对大数据问题是否产生了新的研究方法论需求？其次，从决策科学的视角，决策要素及其范式面临着哪些转变？第三，从展开研究的顶层设计视角，如何通过整体布局使得研究活动既体现全局性统一目标和研究方向引领，又具有充分的开放式探索空间？第四，从价值创造的视角，大数据驱动的管理决策研究能够在学术和应用领域取得哪些重要创新？

针对上述挑战，大数据重大研究计划汇集了一大批国内高水平团队进行攻关，并取得了大量重要进展. 概括说来，大数据重大研究计划 1）构建了一个大数据管理决策的研究体系，该研究体系涵盖大数据及其问题的认识论特征、新型“大数据驱动”方法论范式、以及开展大数据管理决策研究的方向性框架（即：全景式 PAGE 框架）；2）形成了一系列在全景式 PAGE 框架下的研究创新；3）设计了一个面向多领域情境的平台集成架构.

大数据的认识论特征可以从三个维度来审视，即数据特征，问题特征以及管理决策特征[2]. 首先，大数据的数据特征体现为体量大、多样性、（价值）密度低、（迭代）速率高等数据属性，分别具有超规模、富媒体、深挖掘、流数据等意味. 其次，一个研究问题可以被称作为“大数据问题”，应该至少具备粒度缩放、跨界关联和全局视图三个特点，粒度缩放是指问题要素的数据化，并能够在不同粒度层级间进行缩放；跨界关联是指问题的要素空间外拓，即在问题要素空间中引入外部视角与传统视角联动，将内部数据与外部数据予以关联；全局视图是指问题定义与求解的全局性，强调对相关情景的整体画像及其动态演化的把控和诠释[2]. 再者，从管理决策的视角来看，大数据管理决策应该反映全景式“关联 + 因果”诉求. 结合管理决策中管理者关心的“What/Why/Will”三个核心问题，大数据可以从业务层面、数据层面和决策层面分别刻画并形成管理决策大数据问题特征框架.

大数据相关研究同时催生了新型研究方法论范式的转变，即从传统的“模型驱动范式”或简单的“数据驱动范式”转变为“数据驱动 + 模型驱动”的新型融合范式，即“大数据驱动研究范式”，更为强调“关联 + 因果”的诉求，并呈现出显著的外部嵌入、技术增强和使能创新特点[2, 3]. 外部嵌入是指通过引入外部视角，将传统模型视角之外的一些新的重要变量引入到模型的变量集合中. 通常新变量的引入将导致构建新型变量关系. 技术增强是指通过机器学习等新型技术方法和工具将引入的新变量（如视频、语音、文本、图片等多模态变量）融入变量集合，并支持新型变量关系的构建. 使能创新是指通过构建新型变量关系，形成大数据驱动的价值创造. 在建模方法论上体现为由新变量引入（$X' = X \cup X_{new}$）和新型变量关系（F'）形成的 $Y' = F'(X')$. 这个过程往往也需要技术增强，如因果推断、函数拟合、可解释性人工智能等方面的技术使用和创新. 这里，新型变量关系的形式可能是原模型的简单变量增拓（如简单线性增拓），或需要探索与原模型不同的变量间映射关系.

将大数据问题特征（粒度缩放、跨界关联、全局视图）映射到四个研究内容方向（即 PAGE：决策范式（Paradigm-P）、分析技术（Analytics-A）、资

源治理(Governance-G)和使能创新(Enabling-E)上,便形成一个面向不同领域情境(如商务、金融、医疗健康、公共管理等)的由 4 × 3 要素矩阵所构成的全景式 PAGE 框架[2]. 具体说来,决策范式(P)研究方向重点关注管理决策范式转变机理与理论;分析技术(A)研究方向重点关注管理决策问题导向的大数据分析方法和支撑技术;资源治理(G)研究方向重点关注大数据资源治理机制设计与协同管理;使能创新(E)研究方向重点关注大数据使能的价值创造与模式创新. 每一个研究方向在粒度缩放、跨界关联、全局视图三个问题特征的映射下,分别体现出不同的研究侧重,如图 1 所示. 全景式 PAGE 框架成为整个大数据重大研究计划的方向性框架和项目指南重要内容[4],在相关关键科学问题凝练和贡献新知方面,发挥了重要指导性作用.

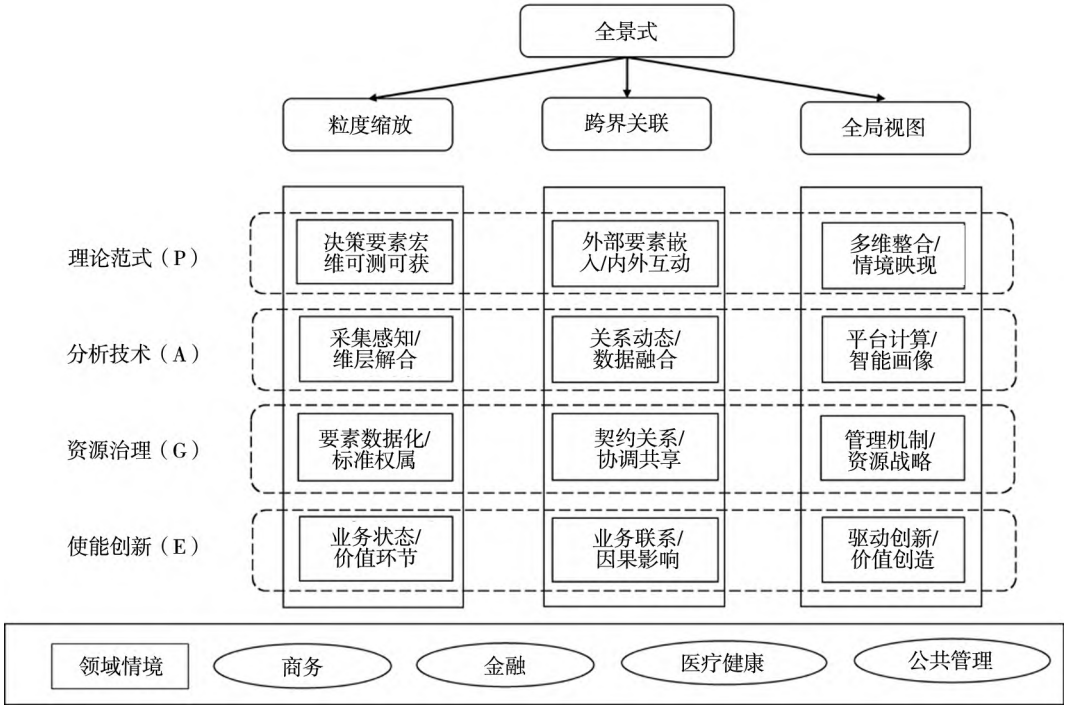


图 1　全景式 PAGE 框架

Fig. 1 Panoramic PAGE framework

本文下面的讨论将进一步阐释和解构全景式 PAGE 框架的内涵及其内在关系,同时围绕 PAGE 方向上的研究,讨论大数据重大研究计划的设计开展、创新思路和重要新知,并概述近年来国内外若干重要探索. 本文的意义主要体现在两个方面:一是从大数据管理决策研究的视角呈现学科领域探索中的重要方向、问题探识和创新突破,为大数据等颠覆性信息技术应用的相关研究提供前沿方向和创新突破上的启迪;二是通过整体研究体系的擘画以体现全局目标、方法论范式和重点议题,为管理学科及其相关学科领域开展交叉属性强、应用情境差异化特点突出的科研攻关提供参考.

1　PAGE 内涵与解构

如前所述,面向不同的领域情境和应用场景,全景式 PAGE 框架涵盖了大数据管理决策研究的四个重要研究方向(即决策范式(P)、技术分析(A)、资源治理(G)、使能创新(E)),并在三个大数据问题特征的映射下形成了一个 4 × 3 要素矩阵(图 1). 这里,将围绕 PAGE 进行内涵描述和结构解析,以获得对于全景式 PAGE 框架的更深入阐释.

首先,大数据基础设施(如通讯网络、物联网、云存储、计算平台等)作为底层架构,是整个

大数据生态的基础层，旨在为 PAGE 研究方向提供数据感测、通讯、获取、存储、计算等基本支撑. 在此基础上，PAGE 研究方向构成了大数据生态的应用层，并相互依存和彼此影响. 沿着 PAGE 间主要关系脉络进行分析，可以刻画出下列影响路径.

第一，决策范式(P)主要面向决策理论建模及其范式转变，涉及决策要素和过程的刻画乃至决策问题的求解策略. 鉴于资源治理(G)、技术分析(A)以及使能创新(E)中存在着各式各样的决策问题，涉及领域情境、决策主体、理念假设、方法流程等方面[4]，故而，决策范式(P)的变化将深刻影响着决策问题建模及其求解. 也就是说，方向 P 对于方向 G、方向 A、方向 E 产生影响.

第二，技术分析(A)主要面向数据和算法的方法创新，是大数据能力构建的核心内容，影响着使能创新(E)的效果和驱动力，即方向 A 对于方向 E 产生影响.

第三，资源治理(G)主要面向大数据能力构建中相关的“数据 + 算法”获取、处理及使用方式，并对相应的环境和生态进行营造、规范和治理，影响着技术分析(A)和使能创新(E)的效果和模式. 换言之，方向 G 对于方向 A、方向 E 产生影响.

第四，使能创新(E)主要面向大数据驱动的经济和社会活动，影响着为个人、组织和社会进行产品/服务创新和商业模式创新、乃至价值创造的效果. 同时，使能创新(E)也以不同形式影响着决策要素和过程，可能导致决策范式(P)的转变. 例如，基于数据采集和分析方法应用的使能创新，驱动决策要素进一步可测可获，使得跨域信息和宽松假设成为可能；基于智能系统/智能机器人应用的使能创新，驱动机器可以自主决策，使得机器充当决策主体成为可能；基于相关机器学习算法应用的使能创新，驱动要素深度交互和环节路径重塑，使得非线性决策流程成为可能. 也就是说，方向 E 对于方向 P 和价值创造产生影响.

最后，构成大数据环境的三个大数据问题特征(粒度缩放，跨界关联和全局视图)体现了不同视角层面上的数据与问题要素的联系：粒度缩放从深度视角出发，在宏微观垂直层面进行问题要素维度内的数据粒度细化和纵向拓展；跨界关联从广度视角出发，在视域水平层面进行问题要素维度间的跨界联结和横向拓展；全局视图从整体视角出发，在垂直和水平层面进行问题要素维度的综合表征与全局呈现.

概括说来，PAGE 的环境要素、基础支撑、内涵逻辑形成了一个相互联系的整体，以服务于大数据驱动的管理创新和价值创造. 图 2 描述了相应的 PAGE 内部关系逻辑.

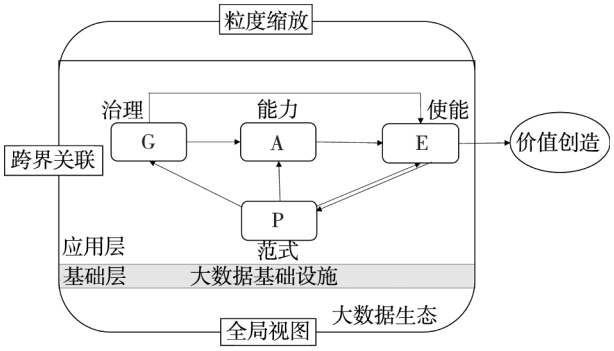


图 2　PAGE 框架内部关系逻辑

Fig. 2 Internal logic of PAGE framework

此外，PAGE 的 4 × 3 矩阵作用于不同的领域情境/应用场景(如：商务、金融、医疗健康、公共管理等)，进一步构成了 PAGE 研究框架的立体图景(图 3). 这里，PAGE 4 × 3 要素矩阵在每个领域情境内形成同构映射，使得 PAGE 要素特征在情境场景中得以体现，即 PAGE 要素的抽象特征在情境场景中实现具象化，进而达到大数据管理决策研究在问题特征、研究方向以及领域情境上的全景式融合呈现.

近年来，大数据重大研究计划沿循全景式 PAGE 框架的研究探索，进一步丰富和深化了 PAGE 的内涵和外延，为促进大数据管理决策领域的发展、进而贡献人类新知以及服务国家战略发挥重要作用. 同时，可以看到国内外许多大数据管理决策方面的研究都呈现出在 PAGE 维度上的方向性. 下面将从 PAGE 的视角出发，以大数据重大研究计划下形成的若干前沿性成果为例讨论和阐释探索新知的重要进展，包括在相关理论方法和场景应用方面的创新贡献. 此外，也概览国内外若干相关的研究工作.

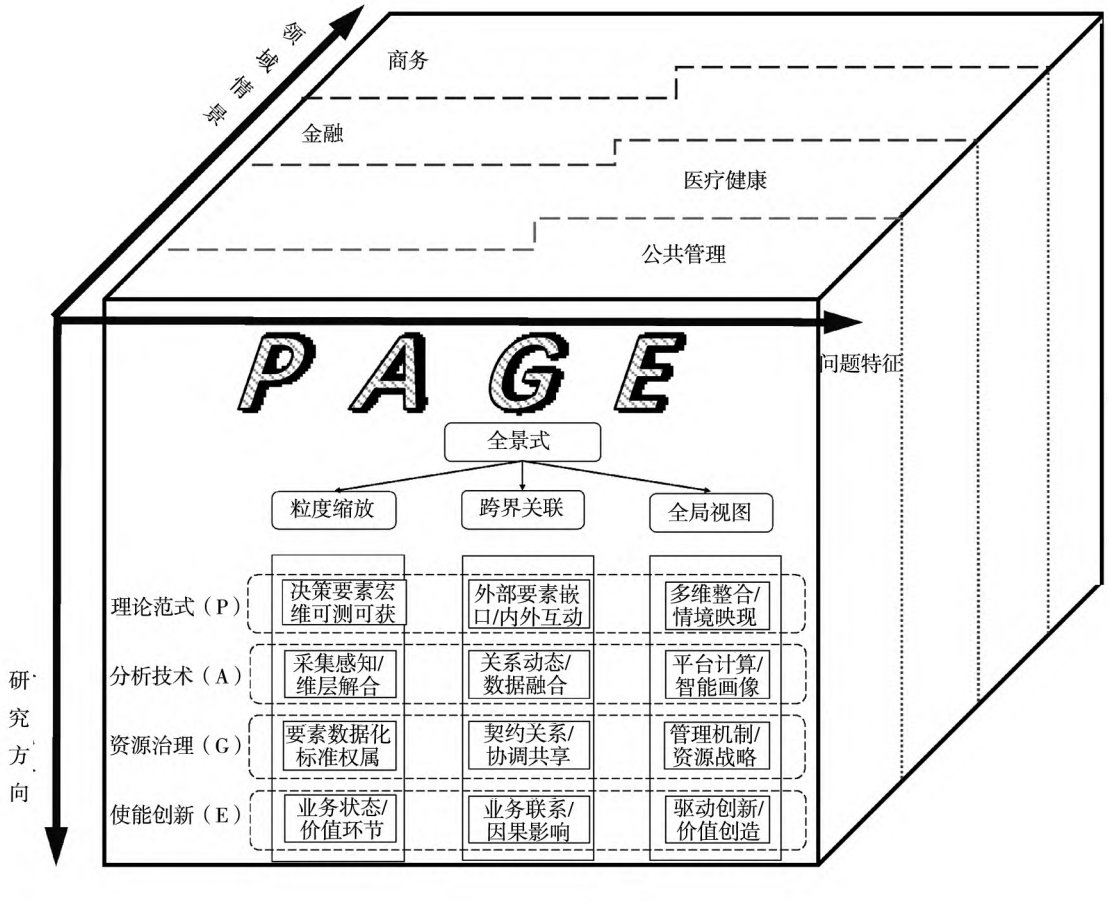


图 3　PAGE 与领域情境融合

Fig. 3 PAGE framework mapping onto specific application domains

2　探索新知

围绕大数据管理决策研究,全景式 PAGE 框架将复杂性高、交叉属性突出的多学科领域的科研探索汇聚到具有统一目标的、体现特定问题特征、探索方向、领域情境的重要视域内,以期既在决策科学层面上具有范式拓新,又在理论、方法、应用层面上具有新知贡献. 一些主要科学发现和重要创新包括:1)在 P 方向上,提出管理决策要素的四大转变(跨域转变、主体转变、假设转变、流程转变)以及新型“大数据决策范式”的跨域型、人机式、宽假设、非线性的特征[3, 5];2)在 A 方向上,构建大数据统计推断与降维、多源异构数据融合与分析、大数据知识图谱等重要理论、方法和关键技术;3)在 G 方向上,设计大数据共享机制与应用模式、大数据标准化和评估体系、大数据隐私分析与影响机理等治理框架与策略;4)在 E 方向上,形成基于大数据的行为洞察、大数据背景下的风险评估与监测、大数据驱动的模式创新等重要价值创造路径与应用示范. 下面通过大数据重大研究计划的若干具体研究实例②,讨论在 PAGE 不同方向上探索新知的相关思路和路径策略.

2.1　新型决策范式创新

如上所述,传统管理决策范式在大数据情境下受到严重冲击,在决策要素上发生着四个重大转变,进而催生了新型“大数据决策范式”[3, 5].

以大数据重大研究计划的研究进展为例,首

② 注:这些研究实例是迄今为止大数据重大研究计划获得的大量重要成果中的一小部分. 这里讨论这些实例旨在进一步阐释 PAGE 的特定内涵,并以若干具象的形式示意性地呈现对于 PAGE 框架的丰富和贡献,而不是试图在此对于大数据重大研究计划的所有成果进行总结和评价.

先，从决策范式（P）中决策要素的跨域转变视角出发，一个重要课题是探索新型财务报告体系，通过对于数据资产的价值测量和评估，构建“第四张报表”，丰富和扩展现有财务决策范式. 具体说来，传统范式采用的财务三张报表（即资产负债表、现金流量表和利润表）旨在反映企业的运营能力、偿债能力和盈利能力. 然而，在大数据时代，建立在三张报表基础上的企业估值模型在反映诸多企业（如不同类型 IT/数据企业、新型创业企业、新业态企业等）的经营状况和业务变化时存在局限，主要面临测量和评估数据资产方面的严峻挑战. 通过引入“第四张报表”，可以对以品牌、口碑、忠诚度、公允价值/无形资产等形式反映的企业数据资产进行价值测量，进而和传统报表一起为财务决策提供更多可以发掘利用的信息集和新的管理洞见.

第四张报表			
项目	结构化指标	项目	结构化指标
广域经营信息：		**环境保护：**	
电商信息		温室气体排放	
经营用电信息		有毒有害气体排放	
经营用水信息		废水排放	
海关申报信息		危险废弃物	
企业税务信息		绿色低碳产品/服务	
产品与品牌：		环境认证/表彰	
产品相似度		环境处罚	
产品质量		**社会责任：**	
产品口碑		社会投入	
客户忠诚度		社区贡献	
无形价值		慈善公益	
质量认证		员工发展	
创新质量：		性别平等	
研发投入		诚信经营	
专利申请		**治理水平：**	
专利相似度		股东权益	
新产品开发		债权人权益	
科技获奖		董监高履职	
企业风险：		薪酬激励	
资金占用		管理层分析与讨论	
内部控制		中介机构质量	
诉讼处罚		**投资者关系：**	
关联交易		投资者沟通渠道	
政策冲击		投资者调研情况	
供应商与客户：		投资者评价	
客户质量		投资者保护	
客户集中度		**行业与竞争对手：**	
供应商质量		产业政策	
供应商集中度		行业风险	
数字化水平：		竞争对手情况	
数字资产			
数字化能力			
应用系统			
更多要素扩展与数据化			

图 4　“第四张报表”要素框架

Fig. 4 Key elements in the fourth statement

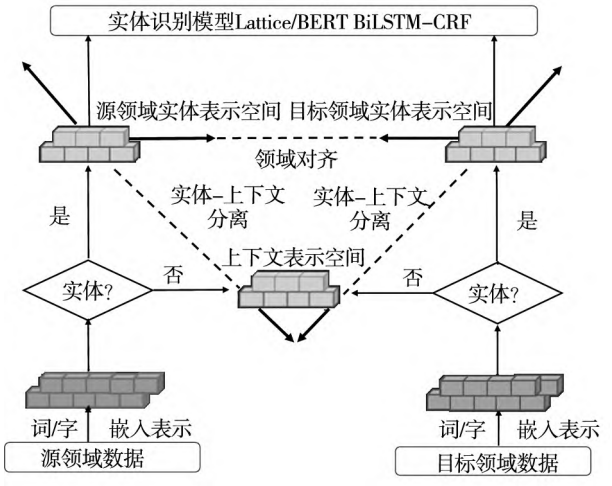
围绕企业价值要素，相关研究提出了一个全局性、系统化的“第四张报表”理论框架，如图 4 所示[6]. “第四张报表”由基础要素和拓展要素构成，前者主要提供以多源异构数据表征的企业日常经营活动相关价值信息，包括广域经营信息、产品与品牌、创新质量、企业风险、供应商与客户、数字化水平等；而后者具有显著的跨域特征，通过突破企业的边界、考虑企业内外部广域主体的关系，提供企业日常经营活动之外的影响企业价值的信息，包括环境保护、社会责任、治理水平、投资者关系、行业与竞争对手等. 例如，在广域经营信息这个维度，大数据中的大量非财务信息，如海关申报等数据，可以更及时反应企业的经营状况，为信息使用者提供关于企业的多维度细粒度画像；在产品与品牌维度，产品口碑和客户忠诚度这些以往不可测不可获的变量，通过对电商、社交网络数据的挖掘分析工作，也可被转化为结构化指标加入报表[6]；在数字化水平维度，企业拥有的数字资产和数字化能力能够帮助企业优化产品设计与生产模式、提升服务质量、促进商业模式创新，从而在数字时代获取竞争优势.

值得一提的是，对于数据资产的引入和测量，在方法论上具有显著“技术增强”特征，这里机器学习和数据分析方法发挥重要作用. 例如，构建“第四张报表”的关键技术包括基于迁移学习的跨领域的实体识别模型（图 5（a））、基于 Bootstrapping 和强化学习的实体关系识别模型（图 5（b））、以及基于图神经网络的篇章级财经领域事件抽取识别模型、基于知识标准化的“第四张报表”指标生成方法等[6].

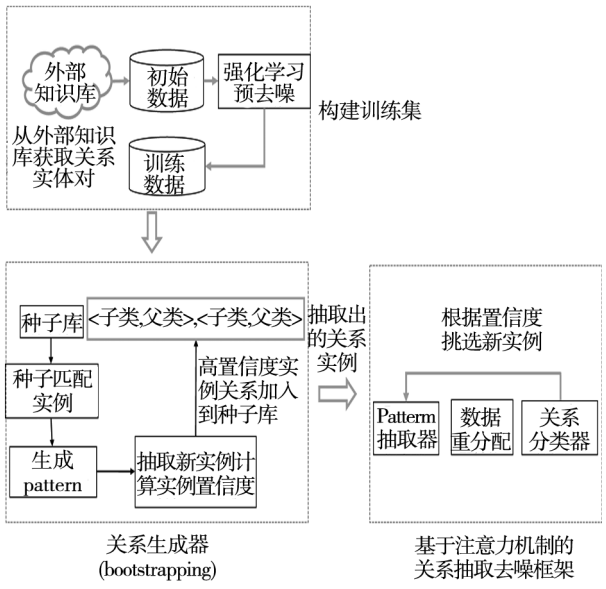
第四张报表的构建是对于传统财务决策范式的重要拓新，与其它三张报表一起形成新型财务决策范式，可更好地描绘全景式的公司经营画像，提供更全面及时的信息，并为会计信息在公司估值、契约设计以及资本监管等方面的应用创新赋能.

再者，从决策范式（P）中决策要素的假设转变视角出发，一个重要课题是探索在大数据情境下，可测性和可获性的提升使得许多传统理论假设被放宽乃至打破，进而对于建模基础、变量关系以及求解策略产生严峻挑战. 例如，对于 O2O（online-to-offline）即时物流决策问题，传统物流调度模型假设在供给端配送员对于同一订单的配送速度相同，在需求端订单需求服从某一先验分布，然

而以上假设并不能完全刻画现实情况. 一方面,配送员的配送速度根据个体能力、经验等有所不同,也受到平台政策、路况和天气状况的影响;另一方面,订单数量和分布随时间的波动性较大,不同区域的订单分布也存在各自的特点,在即时物流配送的场景下,配送调度模型的设计需要对未来的订单量有精准的把握.



(a)



(b)

图 5　构建“第四张报表”关键技术:实体与关系识别

Fig. 5 Key technologies in building the fourth statement: Entity and relationship identification

鉴于此,相关研究通过对于现实应用场景的大数据感测,放宽对于统一配送速度和固定需求分布的假设,结合多领域的细粒度数据,建立个性化配送时间预测模型以及订单需求预测模型[7,8]. 具体说来,对于配送时间个性化预测模型,考虑使用多个维度的数据,包括轨迹相关特征、配送员相关特征、外部环境和区域相关特征. 同时针对数据稀疏情形,进行数据特征聚类. 整体目标为通过选择合适的聚类数量,使得整体预测误差最小

$$\min_L \min \sum_{k\in K} e_k = \min_L \min \sum_{i=1}^{L} \sum_{k\in K_{l,L}} \| g_{l,L}(\Pi_k) - T_k \| \tag{1}$$

其中 K 表示全部配送员的集合, e_k 表示配送员 k 的配送时间预测误差, T_k 代表配送员 k 的实际配送时间,L 代表聚类的类目总数,l 代表聚类后的第 l 个配送员集合, $K_{l,L}$ 为聚类的类目总数为 L 时,第 l 个配送员集合中的全部配送员, $g_{l,L}(\Pi_k)$ 代表将配送员 k 的配送时间特征矩阵 Π_k 输入机器学习模型中得到的配送员 k 对应配送时间的预测值.

对于未来订单数量的预测采用时空相似性度量的方法,融合外部特征和内部特征,从而找到和目标场景之间加权时空距离最小的场景来进行未来订单量的预测. 根据个性化配送时间预测值以及未来订单需求预测值,即时订单调度优化目标是最小化延误成本以及配送员旅行成本

$$\min_{X,Y} \sum \widehat{c_p} (\widetilde{B_i} - l_i^t)^+ + \sum \widehat{c_d} \widetilde{L_k^t} \tag{2}$$

其中 X 表示配送员的路线选择,Y 表示订单分配情况, $\widetilde{B_i}$ 表示订单 i 的配送完成时间, l_i^t 表示订单 i 的最晚送达时间, $\widetilde{L_K^t}$ 表示配送员 k 的总旅行时间, $\widehat{c_p}$ 和 $\widehat{c_d}$ 分别表示延误时间和旅行时间的单位成本.

这里模型求解存在两个难点:一方面,预测模型的结果存在不确定性,即点估计结果存在偏差,从而影响派单和路径规划决策;另一方面,预测模型和优化模型存在耦合,即个性化配送时间预测模型的输入特征包含跟派单决策有关的变量,如接单数,轨迹距离等,而订单调度优化又需要配送时间的预测结果作为输入. 为了解决宽假设下的新挑战,相关研究[7]提出了两点创新:一是基于模型预测的残差分布,将订单配送时间表征为概率分布形式,并得到优化模型如式(3)

$$\min_{X,Y}\sum\sum_{y} p_{ik}^{t}(y)\,\widehat{c_p}\,(y-l_i^t)^{+}+\sum\widehat{c_d}\,\widetilde{L_k^t} \qquad (3)$$

其中 $p_{ik}^{t}(y)$ 为订单 i 由配送员 k 配送花费时间的概率密度函数；二是针对预测特征和决策方案耦合问题，基于禁忌搜索的思路，提出了一个同步预测和决策的算法进行求解. 概况说来，图 6 给出了一个宽假设下的物流决策建模思路.

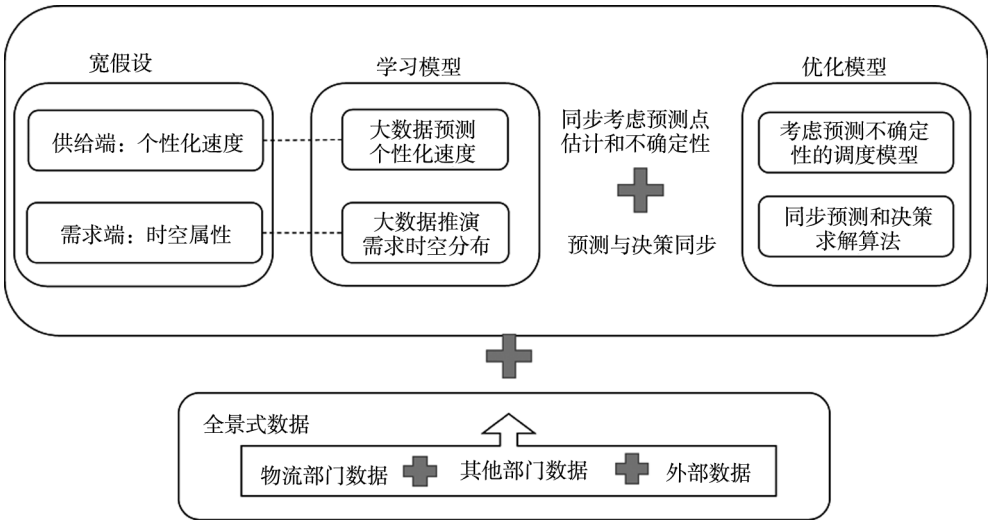


图 6　宽假设下的物流决策建模思路

Fig. 6 Delivery decision modeling under relaxed assumptions

此外，大数据重大研究计划的决策范式（P）研究中的其它重要课题还包括决策主体转变相关研究（例如在人机共生及协同决策情境下的组织行为新理论构建）、流程转变相关研究（例如打破传统线性决策模式的非线性流程建模及其理论扩展）等[9-11].

2.2　大数据方法创新

由于大数据具有体量大、速率高、富媒体、价值密度低等特征，基于大数据的分析技术需要在经典方法上进行改进或重新设计.

以大数据重大研究计划的研究进展为例，从分析技术（A）研究中的统计推断视角出发，一个重要课题是探索缺失数据填补的基础理论问题，以应对大数据环境下管理决策所面临的数据质量和完备性方面的严峻挑战[12]. 具体说来，现实场景中采集到的大数据往往具有较大的缺失比例和稀疏特征，进而可能对于统计推断和机器学习算法的效果产生严重影响，造成管理决策偏差. 一个有效解决思路是对缺失数据进行完备化处理，如根据较少的观测值对原始矩阵进行有效还原. 矩阵完备化问题的形式化定义如下：$\boldsymbol{A}_0=(a_{0,ij})_{n_1\times n_2}$ 表示 n_1 行 n_2 列的原始矩阵，假设矩阵 $\boldsymbol{A}_0$ 的秩是较小整数，$\boldsymbol{Y}=(y_{ij})_{n_1\times n_2}$ 是由 $\boldsymbol{A}_0$ 加上均值为 0 的噪音后形成的可观测矩阵，$y_{ij}=a_{0,ij}+\varepsilon_{ij}$，$\varepsilon_{ij}$ 表示均值为 0 的随机噪音. 同时 $\boldsymbol{Y}$ 只有小部分元素可被观测，例如电影评分场景，每个用户只对少数电影给出了评分. 矩阵完备化问题的目标即通过特定完备化方法来得到真实矩阵 $\boldsymbol{A}_0$ 的估计矩阵 $\widehat{\boldsymbol{A}}$，那么求解 $\widehat{\boldsymbol{A}}$ 可转化为式(4)所示优化问题[12]

$$\widehat{\boldsymbol{A}}=\operatorname{argmin}_{\boldsymbol{A}\in\mathscr{A}}\{L(\boldsymbol{A},\boldsymbol{Y})+R(\boldsymbol{A})\} \qquad (4)$$

其中 $\mathscr{A}$ 表示可能解构成的解空间集合，$L(\boldsymbol{A},\boldsymbol{Y})$ 表示损失函数，衡量 $\boldsymbol{A}$ 与 $\boldsymbol{Y}$ 之间的距离，常见的选择是平方损失函数. $R(\boldsymbol{A})$ 表示正则化项或惩罚项，用于对 $\boldsymbol{A}$ 的结构进行规范，一个自然的想法是使用矩阵的秩（矩阵非零奇异值的个数）作为惩罚项，然而该方法是 NP－难问题，因此可采用矩阵的核范数 $\|\boldsymbol{A}\|_*$（即矩阵奇异值的和）作为矩阵秩的凸近似. 针对大数据的超高维度、多源异质以及时空关联等特点，缺失数据的完备化问题面临着不同的难点与挑战，对损失函数及正则化项也应有不同的具体设定.

例如，在超高维度缺失数据完备化场景中，假设数据遵循随机缺失机制（Missing at Random），即 y_{ij} 被观测到的概率与某些可观测的协变量相关，而与

y_{ij} 取值无关,这里给出一个具体的解决思路. 为了便于分析,引入观测示性矩阵 $\boldsymbol{T}=(t_{ij})$,其中如果 $t_{ij}=1$,则 y_{ij} 可被观测到,否则 y_{ij} 不可被观测到. 与之对应,$\boldsymbol{T}$ 对应的观测概率矩阵可记为 $\boldsymbol{\Theta}=(\theta_{ij})$,其中 θ_{ij} 代表 t_{ij} 取 1 的概率,即 t_{ij} 服从成功概率为 θ_{ij} 的伯努利分布. 那么在随机缺失机制下,可采用 $L(\boldsymbol{A},\boldsymbol{Y})=\frac{1}{n_1 n_2}\times\|\boldsymbol{T}\circ\boldsymbol{\Theta}^{\circ(-\frac{1}{2})}\circ(\boldsymbol{A}-\boldsymbol{Y})\|_F^2$ 作为损失函数,其中$\circ$为矩阵 Hadamard 算子,表示两个矩阵对应元素相乘,$\boldsymbol{\Theta}^{\circ(-\frac{1}{2})}=(\theta_{ij}^{-\frac{1}{2}})$. 然而多数情况下 $\boldsymbol{\Theta}$ 并不可得,因此一个关键环节是需要先构建 $\boldsymbol{\Theta}$ 的估计量 $\widehat{\boldsymbol{\Theta}}$,此时式(4)中的优化问题可表示为式(5)

$$\widehat{\boldsymbol{A}}=\underset{\boldsymbol{A}\in\mathcal{A}}{\operatorname{argmin}}\left\{\frac{1}{n_1 n_2}\|\boldsymbol{T}\circ\widehat{\boldsymbol{\Theta}}^{\circ(-\frac{1}{2})}\circ(\boldsymbol{A}-\boldsymbol{Y})\|_F^2+\lambda\|\boldsymbol{A}\|_*\right\}\tag{5}$$

对于 $\boldsymbol{\Theta}$ 的求解质量直接决定了矩阵 $\widehat{\boldsymbol{A}}$ 的估计准确性. 如果已知影响观测概率 $\boldsymbol{\Theta}$ 的协变量信息 $\boldsymbol{X}$,观测概率矩阵可表示为协变量信息的函数,即 $\theta_{ij}=\Pr(t_{ij}=1\mid\boldsymbol{X})=\frac{1}{1+\mathrm{e}^{-x_i^T\cdot\gamma_j}}$,其中参数 $\boldsymbol{\gamma}_j$ 可用极大似然法进行估计. 进一步,为了降低计算复杂度,可对 $\boldsymbol{A}_0$ 进行列空间分解,将迭代算法改进为奇异值分解算法[13]. 在缺少协变量信息的情况下,可考虑通过低秩缺失机制或者非参数模型对 $\boldsymbol{\Theta}$ 进行估计. 以低秩缺失机制的思路为例,假设 $\boldsymbol{\Theta}$ 由低秩矩阵 $\boldsymbol{M}$ 经过联接函数族映射得到,即 $\boldsymbol{\Theta}=f(\boldsymbol{M})$,对 $\boldsymbol{M}$ 进一步做均值分解,得到 $\boldsymbol{M}=\mu\boldsymbol{Z}+\boldsymbol{J}$,其中 μ 是 $\boldsymbol{M}$ 中所有元素的均值,$\boldsymbol{J}$ 是元素全为 1 的矩阵,$\boldsymbol{Z}$ 是元素和为 0 的矩阵. 那么原优化问题可转化为如式(6)所示的似然函数优化目标

$$F(\mu,\boldsymbol{Z}\mid\lambda)=\sum_{i,j}\left\{t_{ij}\lg(f(\mu+z_{ij}))+(1-t_{ij})\lg(1-f(\mu+z_{ij}))\right\}-\lambda\|\boldsymbol{Z}\|_*\tag{6}$$

得到 μ 和 $\boldsymbol{Z}$ 的估计量后,进一步可求解得到 $\widehat{\boldsymbol{\Theta}}$. 该方法在实际数据场景中相比于采用均匀缺失机制的完备化方法,效果得到显著提升[13]. 概况说来,图 7 给出了上述根据协变量信息解决矩阵完备化问题的一个示意图.

大数据统计推断的另一个重要课题是探索分布式统计推断的基础理论方法,以应对大数据分析技术(A)研究中并行和分布式计算所面临的严峻挑战. 以分位数回归问题为例[15,16],假设 $(X_1,Y_1),\cdots,(X_N,Y_N)$ 代表 N 个独立同分布的随机变量,其中 X_i 是 p 维协变量,Y_1 是数值型变量,分位数回归模型可表示为 $Y=\theta_\tau^T\boldsymbol{X}+\varepsilon_\tau$,其中 τ 代表分位数值,即 $P(\varepsilon_\tau<0\mid\boldsymbol{X})=\tau$,$\theta_\tau$ 是待求解参数,ε_τ 为不可观测随机误差,那么 θ_τ 的估计量 $\widehat{\theta_\tau}$ 可通过式(7)得到

$$M(\theta)=\frac{1}{N}\sum_{i=1}^{N}X_i\psi_\tau(Y_i-\theta^T X_i)=0\tag{7}$$

其中 $\psi_\tau(z)=\tau-I(z<0)$,$I(\cdot)$ 为示性函数. 在大规模/流数据背景下,相关研究提出通过分治策略的思路[14-15],将大数据分为 K 个数据块,每个数据块包含 N_k 个观测值 $(X_{k,1},Y_{k,1}),\cdots,(X_{k,i},Y_{k,i}),\cdots,(X_{k,N_k},Y_{k,N_k})$,并有 $\sum_{k=1}^{K}N_k=N$,对于每个数据块,根据式(7),可得到在每个数据块上的分位数回归估计值 $\widehat{\theta_k}$.

$$M_k(\theta)=\frac{1}{N_k}\sum_{i=1}^{N_k}X_{k,i}\psi_\tau(Y_{k,i}-\theta^T X_{k,i})=0\tag{8}$$

进一步对式(8)运用泰勒展开,可将其转化为可导形式得到各个数据块上的估计值 $\widehat{\theta_{N_k}}$. 结合 $M_\theta=\sum_{k=1}^{K}\frac{N_k}{N}M_k(\theta)$,则可得到整体数据集上对于 θ_τ 估计值的封闭解,如式(9)所示.

$$\widehat{\theta_{NK}}=\left(\sum_{k=1}^{K}\frac{N_k}{N}\boldsymbol{A}_k\right)^{-1}\left(\sum_{k=1}^{K}\frac{N_k}{N}\boldsymbol{A}_k\widehat{\theta_{N_k}}\right)\tag{9}$$

其中 $\boldsymbol{A}_k=E[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}f_{Y\mid X}(\boldsymbol{X}^{\mathrm{T}}\widehat{\theta_{N_k}}\mid\boldsymbol{X})]$. 值得一提的是,在大数据情境下,数据通常是在不同设备上分散存储并以流数据形式动态更新的. 此外,分布式统计理论和计算方法也通常需要对数据进行分块划分、对数据分析结果进行分块集成. 这里需要解决的一个核心难点问题是并行分布式算法如何在子节点(如子机器/单元)和主节点(如主机器/单元)之间进行有效的信息交换和通讯,这在迭代式算法设计中更为重要. 相应的统计推断理论和方法创新体现在构建通讯有效

算法(communication-effective algorithms)使得数据、参数、计算、结果等的迭代和通讯成本显著降低,同时可保持集成结果的良好统计性质. 相关研究进一步围绕分位数回归模型,从分块集成迭代算法(主机和子机多次通讯)和数萃(Meta)算法(主机和子机一次通讯)角度(如图 8 所示),提出了新颖的基于等度连续的通讯有效算法、基于光滑逼近的通讯有效算法、近无限分块的扩展数萃算法等[14,16].

此外,大数据重大研究计划的分析技术(A)研究中的其它重要课题还包括高维数据降维算法[17]、稀疏/微弱信号识别方法[18]、多模态数据融合方法[19]、结合领域情境的知识图谱构建[20]、可解释性人工智能算法等.

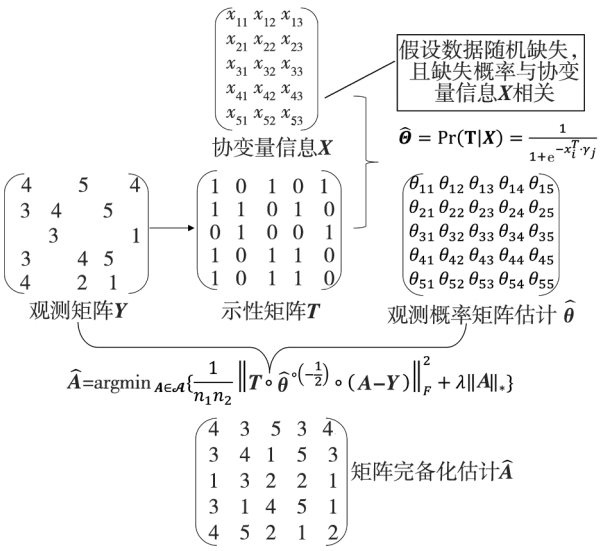
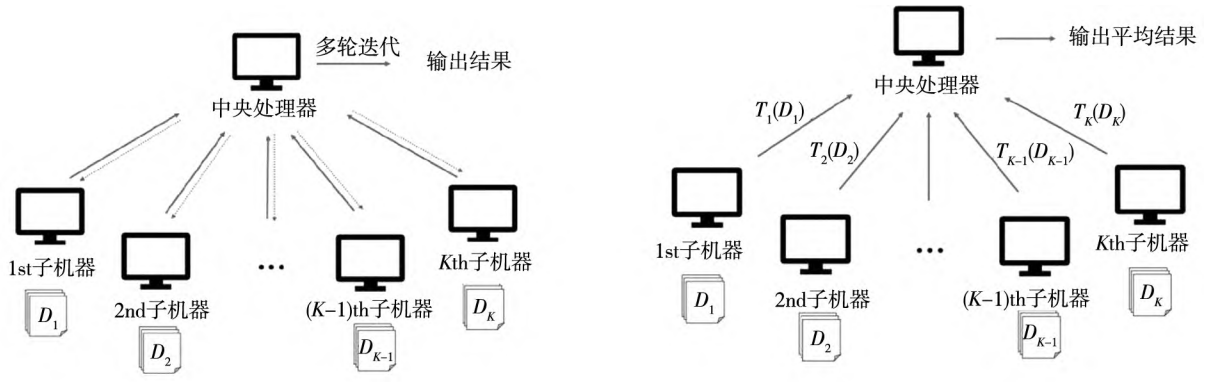


图 7　矩阵完备化思路示意

Fig. 7 A demonstration of matrix completion with Missing at Random data



(a)分块集成迭代算法　　(b) 数萃算法

图 8　分块集成迭代算法与数萃算法示意

Fig. 8 Illustrative divide-and-conquer based algorithm and Meta algorithm

2.3　大数据资源治理创新

在大数据环境下,信息来源多样、任务目标多元、主体关系复杂、决策模型复杂的应用特点,对大数据在共享、开放与监管方面提出了迫切的要求,需要统筹规划共享和治理体系,构建应用模式.

进一步,以大数据重大研究计划的研究进展为例,从资源治理(G)研究中的机制设计的视角出发,一个重要课题是探索构建综合的数据治理、共享和评估体系,并作为示范应用到重要公共服务和政府政策场景中. 具体说来,政府组织内部的跨部门共享是指政府内部共同使用信息资源的一种机制,涉及多元管理主体及多种管理活动. 相关研究构建了权力 – 权益 – 信息的三要素协同管理机制,旨在通过制度建设、标准建构及技术赋权等手段对相关要素进行干预[21]. 这里,行政管理视角下的权力要素协调强调“组织结构调整”在共享中的作用,数据质量视角下的权益要素协调强调“绩效评估”在共享中的作用,信息技术视角下的信息要素协调强调“构建技术平台”在共享中的作用. 进而,凝练出多类大数据综合治理机制,涵盖宏观、中观和微观层次. 在此基础上,相关研究从大数据治理成熟度角度探索政府组织的大数据治理水平,构建了政府大数据治理成熟度评测指标体系[22]. 研究成果应用到作为“全国大数据综合试验区”的某典型省份的大数据治理实践

中,形成了一批重要的大数据治理、共享、规范等方面的标准被政府采纳并产生积极政策影响.

资源治理(G)研究的另一个重要课题是数据标准. 数据标准在数据作为要素的感测、采集、存储、共享、通讯、交换、处理、使用乃至交易、流通、估值等各个环节中发挥着核心作用. 数据标准在很大程度上反映了相关领域的话语体系、知识架构、价值认同和行业规则. 毋庸置疑,大数据相关国际标准的制定意味着大数据治理的国际话语权和影响力. 相关研究在大数据相关国际标准制定方面取得一系列重要进展,主持和参与国际标准化组织(International Organization for Standardization-ISO)、国际电工委员会(International Electrotechnical Commission-IEC)、国际电信联盟(International Telecommunication Union-ITU)、电气与电子工程师协会(Institute of Electrical and Electronics Engineers-IEEE)等国际组织和协会的相关数据标准的制定,涵盖数据概念、术语、本体、测量、审计、迁移、处理、集成、管理、运用等内在及外延属性,涉及大数据相关的物联网、智慧城市、区块链、数据交易、智能系统、管理体系等应用领域[23-25].

从资源治理(G)研究的数据要素市场的角度出发,一个重要课题是探索数据在交易市场中的作用. 在大数据背景下,数据作为一种生产要素,要发挥更为广泛和多元的价值,需要建立有效的数据市场. 虽然大数据价值得到了广泛认可,然而由于数据隐私、数据价值不透明以及定价困难等问题,数据市场的发展处于滞后状态. 相关研究从数据属性的视角出发提出了一个数据交换模型(DADE),如图 9 所示[26]. 首先,从数据生命周期成熟度和数据资产专用性两个维度入手认识数据在交易市场中的作用. 数据生命周期成熟度是指数据能够直接服务于应用需求的程度. 数据从原始数据开始,需要经历清洗、封装、预处理等过程才能够转化为需求方可以使用的数据产品,以上的转化过程主要依赖数据工程来完成. 数据资产专用性是指数据的商业价值取决于特定的应用场景,包括场地专用性、实物资产专用性以及人力资产专用性等,代表了对特定位置、物理配置和人力技能的需求. 从需求出发的数据资产专用性和从供给出发的数据生命周期成熟度相匹配,才能够形成有效的双边市场.

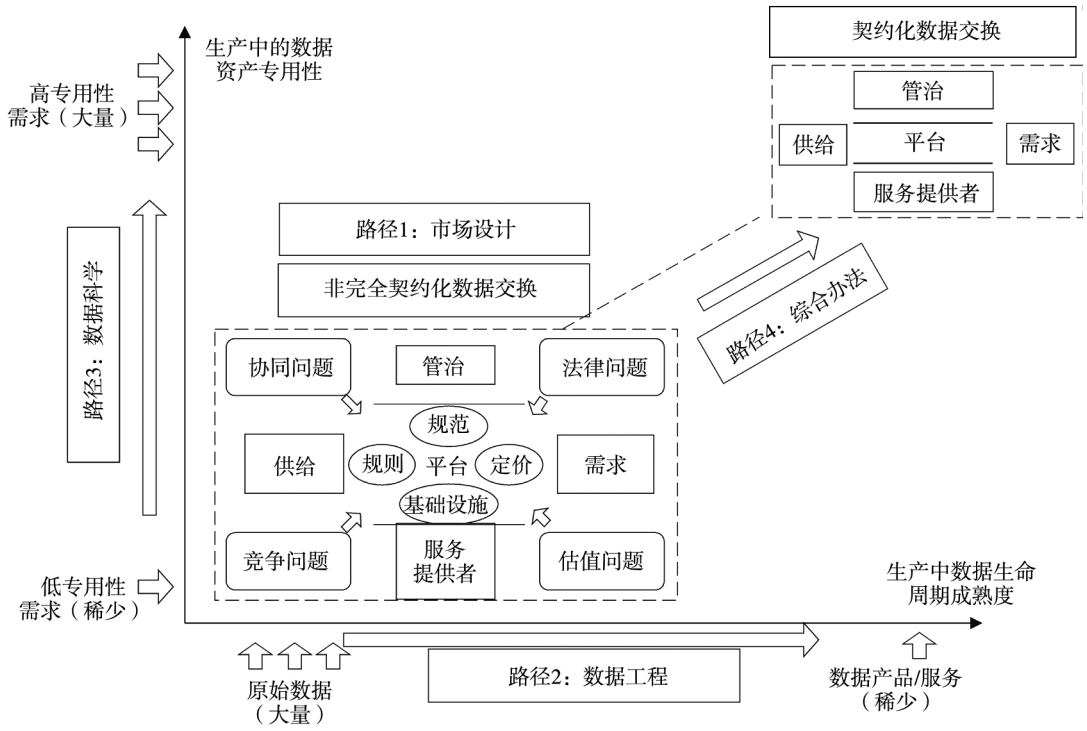


图 9　数据属性影响下的数据交换模型

Fig. 9 The data attributes-affected data exchange model (DADE)

进而，针对供应方提供的原始数据和需求方要求的特定场景化数据产品经常存在不匹配的情况，研究给出了 DADE 模型的四个路径策略[26]：1）市场设计：为了匹配提供原始数据的供应方和具有低专用性需求的数据需求方，需要复杂的机制设计来保证交易的公平合法性，此外，还需要一定的基础设施保证来支持数据处理、整合和评估等活动；2）数据工程：从提升数据生命周期的成熟度角度考虑，通过数据预处理、过滤、再组织等数据工程方法来让数据产品更好的匹配下游分析任务；3）数据科学：从提升数据资产专用性角度来考虑，例如，运用深度学习算法来将数据向特定应用场景转化（如视频和声音中的标签提取等），而且同样的数据集通过数据科学方法可以为不同的数据产品需求来服务；4）综合方法：同时提升数据资产专用性和数据生命周期成熟度.

此外，大数据重大研究计划在资源治理（G）研究中的其他重要课题还包括数据服务体系架构与应用、基于区块链智能合约的数据管理模型与框架、基于贝叶斯网络的数据质量路由机制及评估方法、基于权限的隐私风险量化模型和隐私风险指数评估[27, 28]、数字化转型与政府社会治理模式等.

2.4　大数据使能创新

结合具体环境和场景，构建大数据能力，从多维度动态跟踪行为/事件/行业变化，更系统、精确、及时地进行评估，进而展开相应深度分析，持续驱动面向服务、决策和应用模式的使能创新. 使能创新重点关注如何基于对大数据的分析为不同领域的价值创造赋能.

进一步，以大数据重大研究计划的研究进展为例，在使能创新（E）研究中从不同领域情境（如商务、金融、医疗健康、公共管理等）的视角出发，探索运用大数据理论方法进行赋能/使能③，并实现大数据驱动的价值创造. 首先，如图 2 所示，使能创新（E）受到决策范式（P）和分析技术（A）的影响. 前面讨论的大数据决策范式的跨域型特征（如引入“第四张报表”）和宽假设特征（如打破传统假设建模）分别与金融和商务领域的重要场景（如金融市场、零售平台）相结合，赋能大数据情境下的财务决策和物流决策. 此外，基于数据完备化和分布式统计计算的方法创新，也分别与金融和公共管理领域的重要场景（如金融市场、环境监测）相结合，赋能大数据情境下的风险管理和公共服务决策.

以医疗健康领域为例，流行病传播模式辨识、预测预警和政策管理是医疗防治和公共卫生的重要课题. 大数据重大研究计划相关研究在动力学传播模型的基础上，融合内部数据（如监测数据、队列数据等）以及外部数据（如社交媒体数据、公共管理数据等），形成大数据驱动的全景式信息画像和决策特征集，赋能管理决策/公共干预策略（如图 10 所示）[29].

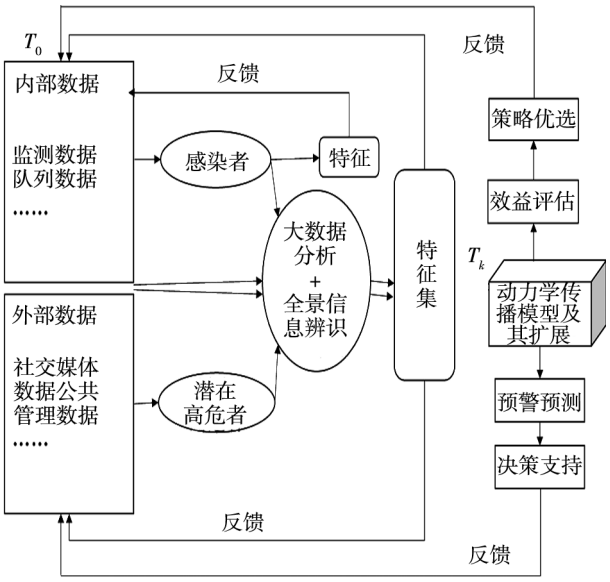


图 10　基于大数据的流行病传播与管理模式示例

Fig. 10 Big data-driven epidemics transmission and management demonstration

具体说来，通过抓取社交网络平台数据，构建了大规模重点观测人群构成的跨领域多源融通的大数据池，然后根据个人图像、行为数据、轨迹位置、社交关系等，结合自然语言处理、计算机视觉模型等大数据分析技术进行用户画像和人群分类，构建传染病高危行为预警模型[30]. 进而设计公共干预策略，形成“发现 – 预警 – 干预”的闭环流程和管理模式，在全国多地得到有效应用. 在此

③　赋能（empowerment）和使能（enabling）二者都有通过能量/能力使得事物/事件延续和发生的含义. 相对而言，使能更强调新能力驱动，使得变化和革新成为可能. 在这个意义上讲，使能是赋能的一种特别情况. 在此小节中，为表述方便，未对二者作特别区分.

基础上,进一步构建了新冠疫情监测策略模型,比较不同场景下流行病的暴发风险和收益,相关监测策略被北京冬奥会组委会采纳. 此外,另有相关研究基于手机定位大数据刻画人群移动行为和迁徙规律,构建了新冠感染人数的指数预测模型[31]

$$y_i = c\prod_{j=1}^{m} e^{\beta_j x_{ji}} e^{\sum_{k=1}^{n}\lambda_k I_{ik}} \tag{10}$$

其中 y_i 代表城市 i 的累积确诊人数, x_{1i} 代表 2020 年 1 月 1 日到 2020 年 1 月 24 日从武汉前往城市 i 的人数, x_{2i} 等代表其他影响因素,如城市 i 的 GDP, λ_k 代表省份 k 的固定效应,如果城市 i 属于省份 k,则 I_{ik} 取值为 1. 通过对上述模型的估计,便可以建立疫情的人群传播模型并计算风险得分,从而可以赋能政府管理决策,提前采取针对性防控策略. 该项研究是较早基于大数据引入外部视角进行新冠流行预测的工作,既在预测性能上具有良好表现,又取得了重要学术影响和社会影响. 再者,相关研究[32]进一步通过对 62 个国家的真实数据分析,显示中国在 2020 年—2021 年处在平衡经济增长和疫情防控的帕累托前沿上,同时揭示了在社区传播风险较高时,提前采取特定防控措施的有效性.

另一个使能创新(E)研究中的例子是知识图谱(Knowledge Graph-KG)赋能商务领域价值创造. 知识图谱描述实体或概念及其之间的关系,提供了一种重要的海量数据组织和领域知识管理的方式,这在大数据时代尤为关键. 知识图谱赋能管理决策源于其在知识表示、融合和推理三大环节上的技术突破. 相关研究基于一系列理论和方法探索,通过与神经网络的融合进行知识学习和推理,构建了综合性商品知识图谱并应用于国内大型知名电子商务平台(如图 11 所示). 基本思路是围绕领域场景以“模型驱动 + 知识增强”为设计原则,通过知识图谱嵌入并融合规则学习的推理方法,在数据 – 知识基础上进行知识图谱模型预训练,从而赋能各主体进行高效实时的商务决策与管理,同时具有决策敏捷化和可解释性的优势[33].

此外,大数据重大研究计划的其它重要课题还覆盖金融领域(如知识图谱和机器学习赋能管理决策、基于重要监测平台和重要金融市场的金融风险管理[34])、环境能源、交通优化[35]、医学诊疗[36]、智能招聘、社交媒体洞察[37]等重要问题情境,并基于重要场景/平台进行应用示范.

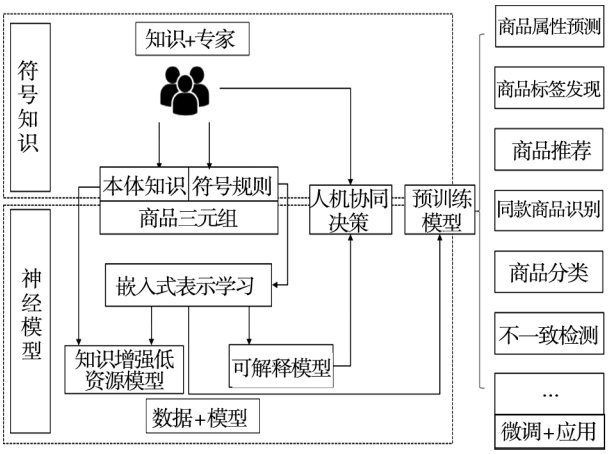


图 11　知识图谱与神经网络融合框架示例

Fig. 11 The integration of knowledge graphs and neural networks

2.5　相关领域动态与未来研究展望

在大数据重大研究计划开展的同时,可以看到国内外各界也不同程度地在 PAGE 方向上取得研究和应用成果.

1)在决策范式(P)方向上,第一,在跨域转变方面,大数据环境下信息收集及处理技术不断提升,决策要素的测量变得更为完善,管理决策所面临的信息环境从领域内部扩展至领域外部,融合内外部信息的决策模型提升了决策的准确性. 例如,基于文本大数据构建的上市公司年报语调和社交媒体情绪指标能够有效预测上市公司的业绩表现,并发挥风险预警功能[38]. 企业还可利用社交媒体上的用户生成内容来分析品牌的市场定位以及识别市场竞争结构[39]. 第二,在主体转变方面,人工智能技术的发展让智能系统深入参与决策过程,决策主体从单一的组织个人转变为由智能系统所主导,或者是人机混合决策的形式. 研究者基于动物行为学的理论框架,从功能、机制、发展和演化等维度出发,提出了机器行为、人机交互行为、人机协同行为等研究方向[40, 41]. 第三,在假设转变方面,传统管理决策需要依赖经典假设构造模型,而在大数据环境下数据驱动增强的新模型代替经典假设,使得模型更加接近现实情境. 例如,在线零售场景下,消费者潜在的购买后悔和迟疑后悔等行为因素的考虑会影响零售商的最优库存和定价决策[42]. 此外,在线上场景,消费者的购

买行为不仅取决于当下时点和单一产品类别，同时受到历史购买以及其他产品类别（如个性化促销）的影响，相关研究借助于条件受限玻尔兹曼机模型，可以对跨时间和跨产品类别关系建立可有效估计的模型结构，增进对消费者决策过程的理解[43]. 第四，在流程转变方面，传统决策通常遵循线性、分阶段的过程模式，而大数据环境下对于全局场景的刻画以及各要素间的动态交互催生了非线性决策流程. 例如，用户在电商平台购物并非遵循营销漏斗的理论，而是在不同阶段状态之间动态转化，基于用户的行为数据可对其所处的购买转化阶段进行建模求解[44]. 在医疗健康管理中，患者就诊 – 疾病诊断 – 疾病治疗的线性流程也经历着重大转变，基于大数据所构建的的医疗疾病路径转化知识图谱，可为用户提供潜在的疾病风险预警[45]，以实现用户全周期健康管理.

2）在分析技术（A）方向上，一些经典的统计推断和降维方法不再适用于体量大、具备相当比例噪音和缺失值的大数据环境，相关研究讨论了高维数据下的统计性质、降维方法、异常模式识别、高效采样方法等[46, 47]此外，大数据来源多样、类型复杂，大数据分析方法需要捕捉数据关系及其动态变化，并能够进行多源异构的内外数据融合，相关研究包括多模态信息表征、跨模态数据转换、多模态信息融合方法、知识迁移以及异构信息网络的构建、分类与推理方法等[48,49]. 再者，大数据价值密度低，需要从海量数据中提取出有用的知识，构建知识图谱，便于高效的知识查询. 以知识图谱构建为例，连接节点较少的稀疏实体的表示学习是一个技术难题，相关研究通过外部知识融合的方式提供了解决思路[50]. 随着模型日趋复杂化，更高效的管理决策不仅需要分析能力的提升，还需要对于模型机理的深入理解，即算法可解释性研究方向. 大数据的数据体量和异质性优势催生了因果推断领域方法的推陈出新，例如，基于随机森林模型构建的因果树可得到异质处理效应的估计值，同时具有良好的统计性质[51].

3）在资源治理（A）方向上，与数据相关的法律法规以及行业标准相继出台，数据治理也日益引发各界的关切. 例如，为了实现数据在多主体间有效、可靠的流动与应用，各国政府层面针对数据监管与个人隐私保护出台了系列法律法规. 2018 年欧盟出台 GDPR（《一般数据保护法》），用于保护欧盟公民个人隐私和数据，同年美国加州通过《2018 加州消费者隐私法案》（CCPA），法案规定消费者有权要求企业删除收集的任何个人相关信息，即消费者拥有“被遗忘权”. 2019 年 5 月，旧金山通过法令，禁止政府机构购买和使用人脸识别技术，成为全球首个禁用人脸识别技术的城市. 2021 年 6 月《中华人民共和国数据安全法》（以下简称《数据安全法》）正式颁布，明确数据处理活动应该依法合规，符合社会公德和伦理，并加强风险监测. 2021 年 8 月，《个人信息保护法》获得通过，该法律是数据安全法则在个人信息保护领域的延伸，细化完善了个人信息保护应遵循的原则和个人信息处理规则，明确了个人信息处理活动中的权利义务边界.

行业层面同样需要统一的标准约束，相关研究工作讨论了数据治理是可信人工智能的基础，提出了数据治理的框架和体系[52]，并就公共部门如何提供更加负责、高效和公平的基于大数据和人工智能算法的服务给出了参考建议[53]. 学术研究方面，大数据资源治理和用户隐私安全也是重要的关注议题. 相关研究从技术角度讨论了本地差分隐私作为一种分布式隐私模型，能够在分析处理数据时提供较强隐私保证，并系统总结了不同的差分隐私模型、算法框架以及针对不同数据统计和分析任务的机制设计[54]. 另有研究表明统计机构在发布统计数据时面临数据发布精准度和用户隐私保护之间的权衡取舍[55]，提升用户隐私保护程度意味着降低数据发布的精准度. 企业在实际应用大数据资源时，需要对大数据相关模型深入理解，确保模型结果在主观和客观层面都做到公正无偏. 例如，相关研究显示性别中立的算法在实际应用于科学、技术、工程和数学专业相关招聘广告投放时，算法的优化结果会自然导致女性看到该类广告的次数更少[56].

4）在使能创新（E）方向上，相关研究讨论了大数据对于不同领域的价值创造. 例如，在金融财务领域，宏微观混合大数据的引入能够对股市低风险定价异象这一传统理论无法解释的问题作出解释[57]. 基于复杂股权网络的金融大数据能够帮助决策者分析关键股权风险结构，实现有效的系统监管[58]. 在营销领域，公司传统上通过访谈来

了解消费者需求,而用户生成内容提供了新渠道来更好地设计营销策略和开展产品研发. 但是海量用户生成内容包含大量无价值信息或者重复信息,相关研究使用卷积神经网络模型过滤掉无价值信息[59],并对句子嵌入表示进行聚类以减少重复内容采样,实现了用户需求的高效识别. 在卫生健康领域,基于线上平台消费者产生的评论文本进行分析,可以识别餐厅卫生检查所存在的健康隐患[60]. 在基础科学研究中,根据氨基酸序列来预测蛋白质的三维结构一直是基础生命科学的研究难点,研究者发现了基于深度神经网络预测蛋白质结构的方法,即使在不知道相似结构的情况下,也能够在原子层面上精确预测蛋白质结构[61]. 此外,相关研究讨论了人工智能技术赋能全球健康问题,尤其是对于资源匮乏的地区和场景[62]. 在公共管理领域,聊天机器人的引入对民众与政府之间的沟通形态转型起到赋能作用[63],基于城市交通出行大数据和气象数据的建模分析[64],为空气污染和交通拥堵等问题的治理提供了有益启示.

展望未来,面向数智化新跃迁[65]的大数据管理决策研究将沿循 PAGE 方向不断深化,并将产生一系列新的重要研究问题.

围绕决策范式(P),数智化新跃迁进一步加速了管理决策向以数据与智能为中心的跨域型、人机式、宽假设、非线性的范式转变,并形成若干重要前沿课题,如:范式转变要素、关系和路径的测度与甄别;融合微观和宏观各层次行为和目标的全景式建模方法;内外部要素嵌入和整合优化的管理决策;决策主体智能化的流程转变与再造等.

围绕分析技术(A),数智化新跃迁使得数据“像素”扩张提升到新高度,而智能“成像”将成为关注焦点,体现为以算法为代表的分析技术的高阶智能化,并形成若干重要前沿课题,如:面向多模态信息表达和处理的智能算法与知识图谱;面向人机融合行为与决策情境的智能算法;面向因果推断和管理可解释性的智能算法;生成式预训练与大模型相关技术及其管理应用等.

围绕资源治理(G),数智化新跃迁强化了数据的生产要素属性,同时数据安全、隐私、权属、共享、要素市场等问题也得到日益关切,数据扩张与数据资源治理的平衡发展将成为一个主基调,并形成若干重要前沿课题,如:大数据标准化与质量;数据安全、隐私、共享及智能化解决方案;组织层面数据治理体系;数字化转型与政府社会治理等.

围绕使能创新(E),数智化新跃迁对业务赋能及价值创造过程形成了更大的驱动力,深刻影响着新价值创造理论的构建和商业模式拓新,并形成若干重要前沿课题,如:机器行为学与智能技术增强的业务赋能;企业数智化能力培养及战略性应用;面向数据市场/平台的商业模式建模与价值评估;数字化转型新情境下的价值创造等.

3 结束语

大数据、人工智能等技术的进步和广泛应用对个人、组织、社会与政府的管理决策带来颠覆性的冲击和挑战. 针对大数据驱动的管理决策这一研究主题,本文系统阐释了全景式 PAGE 框架的组成内涵、解构了其关系逻辑并勾勒了面向多领域情境的整体式图景. 进而,围绕全景式 PAGE 框架讨论了在理论和应用层面的重要研究进展和新知贡献.

近年来,大数据相关技术仍在快速迭代和不断拓展,相应地,在方法论范式更新、新颖特征感测、问题情境适配、跨学科交叉等方面不断涌现新挑战、新问题和新应用. 融合领域情境的全景式 PAGE 框架凝练了网格化要素和创新点聚焦的方向性体系,为大数据重大研究计划乃至整个大数据管理决策研究领域的创新篇章贡献了重要一页. 在今后的大数据管理决策的研究探索中,将和管理科学以及相关学科领域(如信息、数理、医学等)的广大科技工作者一起继续积极进取,为贡献人类新知和服务国家战略而不断努力④.

④ 感谢国家自然科学基金“大数据驱动的管理与决策研究”重大研究计划的指导专家组、管理组、顾问组和研究团队对于项目布局、实施以及本文相关内容所完成的出色工作、提供的大量帮助和贡献的真知灼见.

参 考 文 献：

[1]中国信息通信研究院.《中国数字经济发展研究报告(2023 年)》. http://www.caict.ac.cn/kxyj/qwfb/bps/202304/t20230427_419051.htm.
China Academy of Information and Communications Technology. China Digital Economy Development Annual Report (2023). http://www.caict.ac.cn/kxyj/qwfb/bps/202304/t20230427_419051.htm. (in Chinese)

[2]陈国青, 吴　刚, 顾远东, 等. 管理决策情境下大数据驱动的研究和应用挑战——范式转变与研究方向[J]. 管理科学学报, 2018, 21(7): 1-10.
Chen Guoqing, Wu Gang, Gu Yuandong, et al. The challenges for big data driven research and applications in the context of managerial decision-making: Paradigm shift and research directions[J]. Journal of Management Sciences in China, 2018, 21(7): 1-10. (in Chinese)

[3]Chen G, Li Y, Wei Q. Big data driven management and decision sciences: A NSFC grand research plan fundamental research[J]. Fundamental Research, 2021, 1(5): 504-507.

[4]国家自然科学基金委员会. 大数据驱动的管理与决策研究重大研究计划年度项目指南(国科金发计〔2018〕71 号; 国科金发计〔2020〕71 号; 国科金发计〔2021〕48 号), 2018-2021.
The National Natural Science Foundation of China. The guide on big data driven research and applications in the context of managerial decision-making (Ref.: [2018]#71; [2020]#71; [2021]#48), 2018-2021. (in Chinese)

[5]陈国青, 曾大军, 卫　强, 等. 大数据环境下的决策范式转变与使能创新[J]. 管理世界, 2020, 36(2): 95-105.
Chen Guoqing, Zeng Dajun, Wei Qiang, et al. Transitions of decision-making paradigms and enabled innovations in the context of big data[J]. Management World, 2020, 36(2): 95-105. (in Chinese)

[6]陈信元, 何贤杰, 邹汝康, 等. 基于大数据的企业"第四张报表": 理论分析、数据实现与研究机会[J]. 管理科学学报, 2023, 26(5): 23-52.
Chen Xinyuan, He Xianjie, Zou Rukang, et al. The oretical analysis, data realization, and research opportunities[J]. Journal of Management Sciences in China, 2023, 26(5): 23-52. (in Chinese)

[7]代宏砚, 陶家威, 姜　海. 大数据驱动的决策范式转变——以个性化 O2O 即时物流调度为例[J]. 管理科学学报, 2023, 26(5): 53-69.
Dai Hongyan, Tao Jiawei, Jiang Hai. Paradigm shift for big data-driven decision making-New paradigm for O2O on-demand logistics[J]. Journal of Management Sciences in China, 2023, 26(5): 53-69. (in Chinese)

[8]Tao J, Dai H, Chen W, et al. The value of personalized dispatch in O2O on-demand delivery services[J]. European Journal of Operational Research, 2023, 304(3): 1022-1035.

[9]He X, Kothari S P, Xiao T, et al. Long-term impact of economic conditions on auditors' judgment[J]. The Accounting Review, 2018, 93(6): 203-229.

[10]Li Y, Xie Y, Zheng Z. Modeling multi-channel advertising attribution across competitors[J]. MIS Quarterly, 2019, 43(1): 263-286.

[11]Shen D, Zhu H, Zhu C, et al. A joint learning approach to intelligent job interview assessment[C]. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2018: 3542-3548.

[12]陈松蹊, 毛晓军, 王　聪. 大数据情境下的数据完备化: 挑战与对策[J]. 管理世界, 2022, 38(1): 17-27.
Chen Songxi, Mao Xiaojun, Wang Cong. Missing data completion in the big data era: Challenges and solutions[J]. Management World, 2022, 38(1): 17-27. (in Chinese)

[13]Mao X, Chen S X, Wong R K W. Matrix completion with covariate information[J]. Journal of the American Statistical Association, 2019, 114(525): 198-210.

[14]周　勇, 张澍一, 李子洋. 大数据下分位数回归模型的数据融合与通讯有效算法及其应用[J]. 管理科学学报, 2023, 26(5): 70-102.
Zhou Yong, Zhang Shuyi, Li Ziyang. Data fusion and communication-efficient algorithm with applications in the framework of big data analysis[J]. Journal of Management Sciences in China, 2023, 26(5): 70-102. (in Chinese)

[15]Chen L, Zhou Y. Quantile regression in big data: A divide and conquer based strategy[J]. Computational Statistics & Data Analysis, 2020, 144: 106892.

[16]Song X, Kim D, Yuan H, et al. Volatility analysis with realized GARCH-Itô models[J]. Journal of Econometrics, 2021, 222(1): 393 – 410.

[17]Chen S X, Li J, Zhong P S. Two-sample and ANOVA tests for high dimensional means[J]. The Annals of Statistics, 2019, 47(3): 1443 – 1474.

[18]Qiu Y, Chen S X, Nettleton D. Detecting rare and faint signals via thresholding maximum likelihood estimators[J]. The Annals of Statistics, 2018, 46(2): 895 – 923.

[19]Zhu Y, Zhao H, Zhang W, et al. Knowledge perceived multi-modal pretraining in e-commerce[C]. Proceedings of the 29th ACM International Conference on Multimedia, 2021: 2744 – 2752.

[20]Zhang N, Jia Q, Deng S, et al. AliCG: Fine-grained and evolvable conceptual graph construction for semantic search at Alibaba[C]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2021: 3895 – 3905.

[21]安小米, 白献阳, 洪学海. 政府大数据治理体系构成要素研究——基于贵州省的案例分析[J]. 电子政务, 2019, (2): 2 – 16.
An Xiaomi, Bai Xianyang, Hong Xuehai. Research on constituent elements of government big data framework: A case study based on Guizhou province[J]. E-Government, 2019, (2): 2 – 16. (in Chinese)

[22]洪学海, 王志强, 杨青海. 面向共享的政府大数据质量标准化问题研究[J]. 大数据, 2017, 3(3): 44 – 52.
Hong Xuehai, Wang Zhiqiang, Yang Qinghai. Research on the quality control of sharing big data for government[J]. Big Data Research, 2017, 3(3): 44 – 52. (in Chinese)

[23]IEC SRD 63235: 2021, Systems Reference Deliverable (SRD)-Smart City System-Methodology for concepts building[S]. IEC SyC Smart Cities, 2021.

[24]TU-T FG-DPM Technical Report D0.2, Data Processing and Management for IoT and Smart Cities and Communities: Methodology for data processing and management concept systems[S]. ITU-T FG DPM, 2019.

[25]ISO 30300: 2020, Information and Documentation: Records Management: Core Concepts and Vocabulary[S]. ISO, 2020.

[26]Huang L, Dou Y, Liu Y, et al. Toward a research framework to conceptualize data as a factor of production: The data marketplace perspective[J]. Fundamental Research, 2021, 1(5): 586 – 594.

[27]孟小峰, 朱敏杰, 刘俊旭. 大规模用户隐私风险量化研究[J]. 信息安全研究, 2019, 5(9): 778 – 788.
Meng Xiaofeng, Zhu Minjie, Liu Junxu. Quantitative research on privacy risk of large-scale mobile users[J]. Journal of Information Security Research, 2019, 5(9): 778 – 788. (in Chinese)

[28]朱敏杰, 叶青青, 孟小峰, 等. 基于权限的移动应用程序隐私风险量化[J]. 中国科学: 信息科学, 2021, 51(7): 1100 – 1115.
Zhu Minjie, Ye Qingqing, Meng Xiaofeng, et al. Privacy risk quantification of mobile application based on requested permissions[J]. Scientia Sinica Informationis, 2021, 51(7): 1100 – 1115. (in Chinese)

[29]Zhang B, Lu Z, Wang L, et al. Tracking HIV infection and networks of drugs users in China: A national series, cross-sectional study[J]. The Lancet, 2018, 392: S48.

[30]Zhang B, Yan X, Li Y, et al. Association between drug co-use networks and HIV infection: A latent profile analysis in Chinese mainland[J]. Fundamental Research, 2021, 1(5): 552 – 558.

[31]Jia J S, Lu X, Yuan Y, et al. Population flow drives spatio-temporal distribution of COVID-19 in China[J]. Nature, 2020, 582(7812): 389 – 394.

[32]Zhu Z, Chen B, Chen H, et al. Strategy evaluation and optimization with an artificial society towards a Pareto optimum [J]. The Innovation, 2022: 100274.

[33]余 艳, 张 文, 熊飞宇, 等. 融合知识图谱与神经网络赋能数智化管理决策[J]. 管理科学学报, 2023, 26(5): 231 – 247.
Yu Yan, Zhang Wen, Xiong Feiyu, et al. Integrating knowledge graph and neural network to empower data-intelligence management decisions[J]. Journal of Management Sciences in China, 2023, 26(5): 231 – 247. (in Chinese)

[34]洪 亮, 马费成. 面向大数据管理决策的知识关联分析与知识大图构建[J]. 管理世界, 2022, 38(1): 207 – 219.
Hong Liang, Ma Feicheng. Knowledge association analysis and big knowledge graph construction for big data management and decision-making[J]. Management World, 2022, 38(1): 207 – 219.

[35]Zeng G, Li D, Guo S, et al. Switch between critical percolation modes in city traffic dynamics[J]. Proceedings of the Na-

tional Academy of Sciences, 2019, 116(1): 23 – 28.

[36]Long E, Lin H, Liu Z, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts[J]. Nature Biomedical Engineering, 2017, 1(2): 1 – 8.

[37]Abbas A, Zhou Y, Deng S, et al. Text analytics to support sense-making in social media: A language-action perspective [J]. MIS Quarterly, 2018, 42(2): 427 – 464.

[38]姚加权, 冯　绪, 王赞钧, 等. 语调、情绪及市场影响: 基于金融情绪词典[J]. 管理科学学报, 2021, 24(5): 26 – 46.
Yao Jiaquan, Feng Xu, Wang Zanjun, et al. Tone, sentiment and market impacts: The construction of Chinese sentiment dictionary in finance[J]. Journal of Management Sciences in China, 2021, 24(5): 26 – 46. (in Chinese)

[39]Liu L, Dzyabura D, Mizik N. Visual listening in: Extracting brand image portrayed on social media[J]. Marketing Science, 2020, 39(4): 669 – 686.

[40]张　维, 曾大军, 李一军, 等. 混合智能管理系统理论与方法研究[J]. 管理科学学报, 2021, 24(8): 10 – 17.
Zhang Wei, Zeng Dajun, Li Yijun, et al. Hybrid intelligence management system research: Theory and methods[J]. Journal of Management Sciences in China, 2021, 24(8): 10 – 17. (in Chinese)

[41]Sturm T, Gerlach J P, Pumplun L, et al. Coordinating human and machine learning for effective organizational learning [J]. MIS Quarterly, 2021, 45(3): 1581 – 1602.

[42]邝云娟, 傅　科. 考虑消费者后悔的库存及退货策略研究[J]. 管理科学学报, 2021, 24(4): 69 – 85.
Kuang Yunjuan, Fu Ke. Inventory and consumer returns policies under consumers' anticipated regret[J]. Journal of Management Sciences in China, 2021, 24(4): 69 – 85. (in Chinese)

[43]Xia F, Chatterjee R, May J H. Using conditional restricted Boltzmann machines to model complex consumer shopping patterns[J]. Marketing Science, 2019, 38(4): 711 – 727.

[44]Zhang Y, Li B, Luo X, et al. Personalized mobile targeting with user engagement stages: Combining a structural hidden Markov model and field experiment[J]. Information Systems Research, 2019, 30(3): 787 – 804.

[45]Ye M, Cui S, Wang Y, et al. MedPath: Augmenting health risk prediction via medical knowledge paths[C]. Proceedings of the Web Conference 2021, 2021: 1397 – 1409.

[46]Zhang J T, Guo J, Zhou B, et al. A simple two-sample test in high dimensions based on L 2-norm[J]. Journal of the American Statistical Association, 2020, 115(530): 1011 – 1027.

[47]Ordozgoiti B, Pai S, Kołczyńska M. Insightful dimensionality reduction with very low rank variable subsets[C]. Proceedings of the Web Conference 2021, 2021: 3066 – 3075.

[48]Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423 – 443.

[49]Shi C, Hu B, Zhao W X, et al. Heterogeneous information network embedding for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(2): 357 – 370.

[50]Liu W, Zhou P, Zhao Z, et al. K-BERT: Enabling language representation with knowledge graph[C]. Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020: 2901 – 2908.

[51]Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests[J]. Journal of the American Statistical Association, 2018, 113(523): 1228 – 1242.

[52]Janssen M, Brous P, Estevez E, et al. Data governance: Organizing data for trustworthy Artificial Intelligence[J]. Government Information Quarterly, 2020, 37(3): 101493.

[53]Margetts H, Dorobantu C. Rethink government with AI[J]. Nature, 2019, 568: 163 – 165.

[54]Wang T, Zhang X, Feng J, et al. A comprehensive survey on local differential privacy toward data statistics and analysis [J]. Sensors, 2020, 20(24): 7030.

[55]Abowd J M, Schmutte I M. An economic analysis of privacy protection and statistical accuracy as social choices[J]. American Economic Review, 2019, 109(1): 171 – 202.

[56]Lambrecht A, Tucker C. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads[J]. Management Science, 2019, 65(7): 2966 – 2981.

[57]姜富伟, 马　甜, 张宏伟. 高风险低收益? 基于机器学习的动态 CAPM 模型解释[J]. 管理科学学报, 2021, 24(1): 109 – 126.

Jiang Fuwei, Ma Tian, Zhang Hongwei. High risk low return? Explanation from machine learning based conditional CAPM model[J]. Journal of Management Sciences in China, 2021, 24(1): 109 – 126. (in Chinese)

[58]洪 亮, 欧阳晓凤. 金融股权知识大图的知识关联发现与风险分析[J]. 管理科学学报, 2022, 25(4): 44 – 66.
Hong Liang, Ouyang Xiaofeng. Knowledge association discovery and risk analysis based on financial equity knowledge graph [J]. Journal of Management Sciences in China, 2022, 25(4): 44 – 66. (in Chinese)

[59]Timoshenko A, Hauser J R. Identifying customer needs from user-generated content[J]. Marketing Science, 2019, 38 (1): 1 – 20.

[60]Mejia J, Mankad S, Gopal A. A for effort? Using the crowd to identify moral hazard in new york city restaurant hygiene inspections[J]. Information Systems Research, 2019, 30(4): 1363 – 1386.

[61]Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596 (7873): 583 – 589.

[62]Wahl B, Cossy-Gantner A, Germann S, et al. Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? [J]. BMJ Global Health, 2018, 3(4): e000798.

[63]Androutsopoulou A, Karacapilidis N, Loukis E, et al. Transforming the communication between citizens and government through AI-guided chatbots[J]. Government Information Quarterly, 2019, 36(2): 358 – 367.

[64]袁 韵, 徐 戈, 陈晓红, 等. 城市交通拥堵与空气污染的交互影响机制研究——基于滴滴出行的大数据分析[J]. 管理科学学报, 2020, 23(2): 54 – 73.
Yuan Yun, Xu Ge, Chen Xiaohong, et al. Study on the interactive mechanism of urban traffic congestion and air pollution: A big data analysis based on DiDi Chuxing[J]. Journal of Management Sciences in China, 2020, 23(2): 54 – 73. (in Chinese)

[65]陈国青, 任 明, 卫 强, 等. 数智赋能: 信息系统研究的新跃迁[J]. 管理世界, 2022: 180 – 195.
Chen Guoqing, Ren Ming, Wei Qiang, et al. Data-intelligence empowerment: A new leap of information systems research [J]. Management World, 2022: 180 – 195. (in Chinese)

Panoramic PAGE framework for big data research on managerial decision-making

CHEN Guo-qing[1], *ZHANG Wei*[2], *REN Zhi-guang*[3], *GUAN Yue*[4*], *WEI Qiang*[1]

1. School of Economics and Management, Tsinghua University, Beijing 100084, China;
2. College of Management and Economics, Tianjin University, Tianjin 300072, China;
3. Department of Management Science, National Natural Science Foundation of China, Beijing 100085, China;
4. School of Economics and Management, Communication University of China, Beijing 100024, China

Abstract: The rapid progress of big data and artificial intelligence increases the complexity and dynamics of managerial decision-making, and also provides considerable space for research and applications. As a directional blueprint, the proposed panoramic PAGE framework features a 4 × 3 matrix based on the nature of big data issues, playing an important role in explorations on theoretical paradigm (P), analytics technologies (A), resource governance (G) and innovation enabling (E). In light of the NSFC's grand research plan on big data, this article elaborates the core concepts and links of the panoramic PAGE framework, as well as the key scientific issues of related contexts, problem-solving route-maps and novel contributions. The aim of this article is twofold: one is to highlight important directions and innovative breakthroughs regarding research on applications of big data-related disruptive technologies, from the perspective of big data research on managerial decision-making; the other is to shed light on large-scale, multidisciplinary and multi-contextual research initiatives in management science and related domains.

Key words: big data; managerial decision-making; panoramic PAGE framework; decision-making paradigm; research paradigm