中国科学 : 信息科学 2025 年 第 55 卷 第 3 期: 654–680
SCIENTIA SINICA Informationis
论文

基于信息熵的数据元件信息计量 —— 以数据定价及其电能统计应用分析为例

陶晓明[1,2], 彭劼扬[1], 王钺[1], 王有政[1], 胡成盛[3], 陆志鹏[3*]

1. 清华大学电子工程系, 北京 100084
2. 新疆大学计算机科学与技术学院, 乌鲁木齐 830046
3. 中国电子信息产业集团有限公司, 北京 100190
* 通信作者. E-mail: luzhipeng@cecdt.com.cn

收稿日期: 2024–06–05; 修回日期: 2024–09–09; 接受日期: 2025–01–06; 网络出版日期: 2025–02–19

欧盟 “地平线 2020” 研究与创新计划下玛丽 · 居里行动资助协议 (批准号: 101109045)、清华大学与中国移动通信集团有限公司联合研究所以及中央高校基本科研业务费专项资金 (批准号: FRF-NP-20-03)、国家杰出青年科学基金项目 (批准号: 61925105)、国家重大科研仪器研制项目 (批准号: 62227801)、国家自然科学基金青年科学基金项目 (批准号: 62401334)、国家自然科学基金科技活动项目 (批准号: 62442106) 和德国联邦教研部 (BMBF) 针对 “AITT – 人工智能辅助技术” 项目 (批准号: 03LB3058B) 资助

摘要 在数字经济时代, 数据作为关键生产要素, 其价值评估与定价问题一直是研究的热点. 当前, 信息价值评估与定价的研究主要集中在数据质量、数据需求和数据市场政策等方面, 但缺乏数据驱动的定量研究. 本文旨在探讨基于信息熵的数据元件数据定价, 并提出相应的定价机制. 随着数据在数字经济中的关键性地位日益凸显, 数据元件作为连接数据供给和需求的 “中间态” 具有重要意义. 首先, 本文回顾了信息价值评估与定价的研究现状, 指出了信息价值评估既是经济挑战也是数据科学挑战, 并阐述了建立最优定价结构的重要性. 其次, 本文定义了数据元件并构建了相应的数学模型, 强调了数据元件在数据资源开发、计量、定价等方面的作用. 进而, 通过基于数据元件的数据计量与定价模型, 研究了数据元件的内蕴信息及其与传统方法在信息量上的差异. 最后, 本文以电力数据定价为实证研究对象, 验证了所提出方法的有效性. 本文还强调了数据元件在实现数据要素高效配置中的重要性, 并指出了其未来发展方向. 本文的研究为实现数据要素的高效配置提供了新思路, 为数字经济的发展和数据要素的流通与共享提供了理论支撑和实践路径.

关键词 数据元件, 数据科学, 信息论, 数据要素, 数据治理, 电力定价

1 引言

随着全球范围内大数据热潮的兴起, 数字化正在以前所未有的程度推动着生产方式、生活方式和治理方式的深刻变革, 世界经济数字化转型已经成为大势所趋 [1]. 数字经济成为继农业经济、工业经济之后的新经济形态, 发展数字经济已成为把握新一轮科技革命和产业变革新机遇的战略选择, 也已

引用格式: 陶晓明, 彭劼扬, 王钺, 等. 基于信息熵的数据元件信息计量 —— 以数据定价及其电能统计应用分析为例. 中国科学: 信息科学, 2025, 55: 654–680, doi: 10.1360/SSI-2024-0169
Tao X M, Peng J Y, Wang Y, et al. Information metrics for data components based on information entropy: data pricing and its application analysis for electric energy statistics. Sci Sin Inform, 2025, 55: 654–680, doi: 10.1360/SSI-2024-0169

成为满足人民美好生活需要的重要途径 [2]. 数据作为新型生产要素, 是数字经济深化发展的核心引擎. 数据已快速融入生产、分配、流通、消费和社会服务管理等各环节, 对提高生产效率的乘数作用不断凸显. 同时, 数据的爆发增长、海量集聚蕴藏了巨大的价值, 为科技创新带来了新的机遇.

数据成为生产要素是数字经济发展的客观规律和内在要求. 2017 年, 习近平总书记主持中共中央政治局第二次集体学习时指出 “要构建以数据为关键要素的数字经济”; 2019 年 10 月, 党的十九届四中全会首次将数据确立为生产要素; 2020 年 4 月, 《中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见》提出要加快培育数据要素市场, 并在 2021 年 12 月 21 日颁布的《要素市场化配置综合改革试点总体方案》中提出建立健全数据流通交易规则; 2021 年 12 月, 国务院发布的《“十四五” 数字经济发展规划》指出, 数据要素是数字经济深化发展的核心引擎, 目标是在 2025 年初步建立数据要素市场体系, 到 2035 年力争形成统一公平、竞争有序、成熟完备的数字经济现代市场体系 [3]; 2022 年 6 月 22 日, 习近平总书记主持召开中央全面深化改革委员会第二十六次会议, 审议通过《关于构建数据基础制度更好发挥数据要素作用的意见》, 该文件于 2022 年 12 月 2 日正式发布, 指出数据正深刻改变着生产方式、生活方式和社会治理方式, 强调数据基础制度建设事关国家发展和安全大局, 明确从数据产权、流通交易、收益分配、安全治理等方面开展数据基础制度建设. 当前, 数据已成为我国重要的基础性战略资源.

数据要素概念的提出和不断深化, 标志着数字经济从数据资源化利用阶段逐渐转向数据要素的市场化配置阶段 [4]. 通过数据要素的市场化配置, 可实现规模化的数据开发、数据流通和数据应用, 进而实现经济社会的效率倍增、安全倍增以及财富倍增. 切实用好数据要素, 协同推进技术、模式、业态和制度创新, 将为经济社会数字化发展带来强劲动力 [5]. 首先有别于资本数据的、价值土地释放等传统过程生产要素, 数据要素投入生产过程并释放价值的方式更多是在复杂众多主体参与下对数据进行持续加工的过程, 在加工过程中借助各种技术工具不断改变着数据的形态; 其次, 数据具有很强的流动性, 在数据跨越经济主体的流动过程中, 一方面受安全和效率等技术条件的限制, 另一方面又必须处理好收益分配、风险界定等商业问题; 最后, 数据是一种非同质化的资源, 其价值释放过程与数据类型、应用场景密切相关, 不同场景下数据发挥的作用不同、价值实现的路径不同, 对其进行组织、加工、流通、应用的技术要求也存在显著差异.

数据成为生产要素是数字经济发展的客观规律和内在要求. 然而, 在探讨数据要素的市场化配置时, 我们需要明确区分信息价值定价和数据定价两个概念. 信息价值定价主要关注信息本身的经济价值, 即信息在特定应用场景下的实际效用和价值. 这涉及到信息的准确性、完整性、及时性以及其对决策过程的支持程度等因素. 信息价值定价更侧重于评估信息在特定业务或研究中的直接价值, 以及信息提供者为获取信息所付出的成本. 相比之下, 数据定价则更加广泛和复杂. 数据定价不仅涉及数据的直接价值, 还涉及数据的潜在价值和未来收益. 数据定价需要考虑数据的规模、质量、多样性、时效性等多个维度, 以及数据的采集、处理、存储、传输等成本. 此外, 数据定价还需要考虑数据的隐私、安全、合规性等因素, 以及数据在不同应用场景下的价值差异. 数据定价的目标是确定数据在市场上的合理价格, 以促进数据的流通和应用, 推动数字经济的发展.

然而, 数据要素化和市场化的现状无法满足快速增长的数据应用需求, 其中的一个关键问题是数据定价问题尚未解决. 在涉及数据共享、交换和再利用的经济活动中, 准确评估数据的价值至关重要 [6]. 虽然评估和表达数据价值的方法有很多, 但一种可广泛适用于大规模应用并为多方接受的方法是确定数据买卖的价格, 即数据定价 (data pricing). 如果没有一个标准化的、广为接受的数据估值方法, 要促进数据在经济活动中的广泛应用和流通就变得很困难. 缺乏确定数据价格的综合框架阻碍了数据与市场交易的无缝结合, 阻碍了数据作为可交易资产在经济中发挥作用 [7].

本文旨在深入探讨数据要素的市场化配置问题, 特别是聚焦于数据定价这一核心难题. 通过梳理数据要素概念的发展历程、分析数据要素与传统生产要素的差异性, 以及明确信息价值定价与数据定价的区别, 本文旨在构建一个更加清晰、全面的数据定价框架. 这一框架将综合考虑数据的多维度特

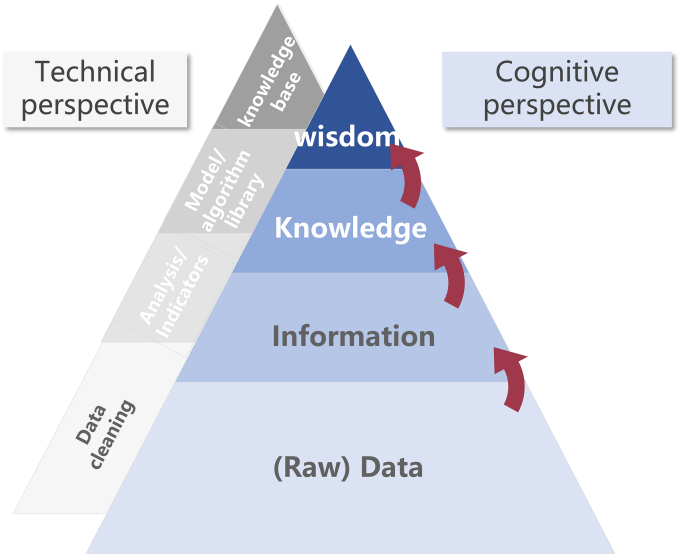

图 1　(网络版彩图) 知识金字塔 DIKW 体系.

Figure 1　(Color online) Knowledge pyramid DIKW model.

征、采集处理成本、隐私安全合规性以及应用场景下的价值差异, 以期为解决当前数据要素化和市场化过程中面临的定价难题提供新思路、新方法.

本文的贡献在于: 首先, 明确界定了数据定价与信息价值定价的区别, 为后续研究奠定了理论基础; 其次, 提出了一个综合性的数据定价框架, 该框架不仅考虑了数据的直接价值, 还兼顾了其潜在价值和未来收益; 最后, 通过深入分析数据要素的市场化配置机制, 为数据在经济活动中的广泛应用和流通提供了实践指导, 有助于推动数字经济的高质量发展.

2　文献综述

2.1　数据与生产要素

数据是数字经济时代的关键生产要素, 是国家基础性、战略性资源, 是推动经济社会高质量发展的重要引擎. 数据作为信息载体, 人们能够根据实际需求, 从海量数据中挖掘有用的信息、知识甚至智慧, 既实现了数据价值的利用, 也通过使用数据创造了新的信息价值. 组织理论家罗素 · 艾可夫 (Russell Ackoff) 在 1988 年提出的著名知识金字塔 DIKW 体系 [8], 即 “数据 – 信息 – 知识 – 智慧” (data-information-knowledge-wisdom, DIKW), 获得广泛认可, 如图 1 所示. DIKW 体系的概念辨析旨在实现数据的信息化, 进而提高信息、知识的管理和交流效率 [9].

在 DIKW 体系的基础上, 国内外众多学者和权威机构对数据进行了定义, 例如: 国际标准化组织认为数据是信息的一种形式化方式的体现, 以达到适合交流、解释或处理的目的. 国际数据管理协会认为数据是以文字、数字、图形、声音和视频等格式表征事实的信息 [10]. 中国信息通信研究院认为数据是对客观事物进行数字化记录或描述, 是无序的、未经加工处理的原始素材. 赛迪智库认为数据是用来记录客观事物或事件的符号, 包含任何以电子或者非电子形式对信息的记录. 欧盟《数字市场法 (提案)》(Digital Markets Act) 认为, 数据是行为、事实或信息的数字表现, 及其任何此类行为、事实或信息的汇编, 包括以声音、视觉、视听记录的形式. 美国《开放政府数据法案》(Open Government Data Act) 认为数据是任何形式或媒介所记录的信息 [11]. 《中华人民共和国数据安全法》认为数据是任何以电子或者其他方式对信息的记录. 此后, 相关法案、指南、报告等相继对网络数据、公共数据、

组织数据等概念做出界定, 但都未超出《中华人民共和国数据安全法》对数据的概念界定范围[1)].

结合上述所有定义来看, 信息化时代下的数据应具有以下特点.

(1) 不同的数据概念建立了特定范畴下的话语体系. 法学侧重内容本身, 即数据传递了怎样的信息; 信息科学侧重数据编码组织方式, 即如何通过计算机语言组织信息; 经济学侧重推理决策, 即如何从数据中挖掘有价值的信息.

(2) 不同数据概念表示特定范畴内主体与客体使用数据的目的、内容与方式, 即为什么使用数据, 使用怎样的数据, 通过什么方式进行使用. 数据记录与传递信息的介质在数字经济时代更多是电子化、数字化的, 也可以是纸质等传统方式.

(3) 不同的数据概念关注点各有侧重, 但均揭示了海量数据的内在联系和潜在价值. 数据挖掘、大规模利用是数据价值开发、催生新模式新业态、产生经济社会效益的主要途径. 因此, 急需通过一种工程化路径, 从海量数据中挖掘有用的信息、知识甚至智慧, 既实现数据价值的利用, 也通过人们使用数据创造新的价值.

在数据赋能实体经济的背景下, 数据可以通过归集、存储、传输、加工, 衍生出更多更有价值的信息, 参与生产、流通、分配、消费等经济循环, 支撑经济社会的高质量发展, 因此也是数字经济时代的核心生产要素 [12]. 广义的生产要素是进行生产经营活动所必需的一切资源及其环境条件. 今天所说的土地、劳动力、资本、技术、数据 5 种生产要素, 范畴更窄一点, 更像是生产经营活动所涉及的那些通用的、标准的、可规模化利用的资源 [13].

生产要素首先应该是一种商品, 但与普通商品不同, 生产要素不是用于消费的, 而是要投入生产过程, 与生产过程相结合形成产品或服务, 创造价值并获得收入. 生产过程中对要素的使用也不太像是消费, 用过就被消耗掉, 就没有价值了, 而是持续使用要素所提供的生产服务 [14]. 生产者购买生产要素, 购买的往往不是要素本身, 而是其在生产过程中的贡献, 即其服务. 生产者以一定的价格购买各种生产要素, 获得生产服务的过程, 其实就是收入在生产者和要素所有者之间进行分配的过程. 相同的生产要素在不同的生产者手中所产出的收入不同, 往往需要引入市场化的机制让要素流向更合适的主体, 通过适当的要素配置提升经济运行效率 [15]. 因此, 商品属性、服务属性是生产要素的基本属性; 而大规模参与经济循环, 参与市场化配置, 参与收入分配是生产要素的重要特征.

数据作为一种重要的资源很早就被认可, 但将其从资源上升成为生产要素则是一种革命性的跨越, 大多数资源都没有完成这种跨越. 这中间的原因部分源于数据的内在禀赋, 数据是服务性资源, 而非消耗性资源, 这与石油、煤炭等重要资源有所不同; 另一部分原因则源于数据对于数字经济的基础性作用, 就如同土地之于农业经济, 劳动力和资本之于工业经济一样. 将数据列为新的生产要素, 一方面是承认数据对于数字经济的基础性作用; 另一方面则是对数据参与经济的方式提出了新的、更高的要求. 旧的数据利用方式更多是以数据的持有方为中心的, 拥有大量数据资源的经济主体才有条件开发并释放数据的价值. 数据并没有在不同的主体之间真正流动起来, 流动到最高效的利用者手里 [16]. 数据在技术上天然就具有着很强的流动潜力, 有着高效配置的可能性. 为了真正释放数据的流动性, 使其大规模地参与经济循环, 才产生了数据要素的理论突破.

首先, 数据上升成为生产要素是要强化数据的商品属性, 数据从资源到商品的转化, 是在经济意义上使数据流转起来的前提; 之后, 在商品化的基础之上进一步引入市场机制和价格机制, 以实现数据资源的市场化配置, 促使数据真正向高价值的场景和用户处汇集; 同时, 这种基于市场机制的数据流动过程也是收益分配的过程, 数据的所有者基于所拥有的高价值数据分享了数据应用创造的收入, 这种基于要素的分配可以激励数据供给, 并促进数据产业的进一步分工 [17].

综上所述, 数据要素的提出, 旨在推动数据的规模化利用与市场化流通配置, 解决当前数据资源化利用阶段效率不高、受益面不大的问题, 促进数据大规模参与经济循环, 实现数据的社会化大生产.

1) https://fgw.sh.gov.cn/cmsres/7e/7e405f41fe814fe98ed4a0922f2866f6/31262692f1a9d65c7cdcbe566f0574bf.pdf.

2.2　信息价值评估的研究现状

信息价值评估既是一种经济挑战, 也是一种数据科学挑战 [18]. 从经济角度看, 信息价值评估涉及到满足消费者需求的资源分配, 需要深入理解市场动态和政策策略 [19]. 另一方面, 从数据科学的角度看, 建立最优的定价结构需要分析大量的数据集, 以识别能够指导定价决策的模式和趋势.

在信息时代的早期阶段, 关于信息价值评估和所有权权利的法律框架并不清晰. 数据传输主要通过政府协议在有限范围内进行, 其中著名的例子包括 Smart Disclosure [20]. 此时期的显著举措包括美国的 “Blue Button” 和英国的 “Midata Project” [21]. 2010 年, 美国推出了 “Blue Button” 应用, 该应用被美国退伍军人事务部的个人平台所采用, 使退伍军人能够访问他们的个人健康数据 [22]. 此外, 由美国能源部和环保署在 2010 年发起的 “Green Button” 项目, 涉及了超过 50 家公共事业和电力供应商, 为超过 6000 万户家庭和企业提供了以标准格式安全访问能源数据的途径. 2011 年, 英国在银行、移动和能源行业启动了 “Midata” 项目, 最初是由政府实体组织的一项自愿倡议, 通过 4 个阶段进行: “透明 – 访问 – 控制 – 转移” [23]. 该项目旨在帮助企业将数据转移给消费者, 并培育个人数据管理服务的新市场. 随后, 英国政府发布了《企业和监管改革法案 (2013)》(The Enterprise and Regulatory Reform Act 2013), 明确了开放机构、开放数据和开放主题的内容, 从而加速了 “Midata” 项目的实施 [24].

随着全球对个人数据相关法规的日益实施, 人们开始寻求新的由数据主体驱动的数据流通渠道. 2018 年 6 月, 包括 Google, Facebook, Microsoft 和 Twitter 在内的主要美国科技公司迅速推出了数据传输项目 [25]. DTP 利用 API 技术, 协同建立了数据传输标准的框架, 简化了所有参与公司之间的个人数据的高效交换. 2018 年, 韩国在金融领域引入了 MyData 模型. 韩国的 MyData 模型代表了以人为本的个人数据管理方法. 到 2021 年 6 月, 韩国有 34 家机构获得了 MyData 运营商的许可, 定制金融服务的用户基数已经增长到 670000.

在学术方面, 信息论在衡量各领域信息价值方面发挥着至关重要的作用. Omran 等 [26] 深入探讨了非金融公司发布会计信息的相关价值, 强调了经验关系和统计技术在评估该价值中的重要性. 在人工智能领域, Yager [27] 探讨了模糊度量作为表示不确定变量知识通用结构的应用, 并将其与概率论进行了类比. Giri 等 [28] 则基于组织信息处理理论和技术接受模型, 研究了供应链从业人员对基于区块链技术的协作的接受程度.

因此, 信息价值评估需要综合市场动态、政策策略以及大数据分析技术来优化资源配置和定价决策. 从早期依赖政府协议推动的数据传输项目, 到近年来全球范围内个人数据法规的实施和由数据主体驱动的数据流通渠道的出现, 信息价值的实现路径不断演变. 学术领域的研究也在不断深入, 信息论、人工智能及模糊度量等技术为评估信息价值提供了新视角和方法. 这些努力共同推动了信息价值评估体系的完善, 促进了数据经济的健康发展.

2.3　数据产品定价的研究现状

数据定价是一个跨学科的关键领域, 涵盖经济学到数据科学等多个方面. 在当前数据交易中的数据定价政策中, 衡量数据商品价值的指标有很多, 如数据生成日期、数据量、数据完整性等. Li 等 [29] 提出了一种基于熵的数据定价指标, 这有助于在数据交易中做出合理的定价决策. Liang 等 [30] 综述了数据定价模型, 对这些模型的优缺点进行了全面的比较, 并设计了数据交易平台及其解决方案, 该平台支持高效、安全、保护隐私的数据交易, 最后, 概述了数据交易生命周期中数据保护面临的挑战. 为了解决数据交易中的潜在问题 (例如, 数据丢失和泄漏), 加强数据交易的安全性, Xu 等 [31] 提出了一种基于区块链的联盟数据交易框架. 在该架构中, 服务提供商对原始数据进行处理, 并提供具有不同数据精度和隐私级别的分层质量数据给买方, 买方决定具体的数据购买策略. 为了充分利用物联网中的数据资产, 发挥其经济价值, Chuang 等 [32] 提出了一种基于物联网的数据定价、交易和保护功能的物联网数据经济系统. 该系统采用以客户为中心的数据价值评估模型和基于博弈论的定价模型, 促进

需求方以可接受的价格获得更高质量的数据, 供方获得更高利润的双赢交易. Luo 等 [33] 提出了一种基于拍卖机制的定价模型, 该模型为实现高效、公平的数据交易提供了一种新的途径. 此外, 该模型保护拍卖过程不受一种称为假名竞标攻击的新型欺诈行为的操纵, 攻击者可以使用多个匿名身份来改进他们的竞标. Tang 等 [34] 则专注于机器学习领域的数据定价, 强调了其在该领域的重要性. Ye 等 [35] 提出了一种基于度量的数据产品定价框架, 强调了为数据集和数据产品定义度量空间的必要性. Shen 等 [36] 基于差分隐私提出了一种个人大数据定价方法, 展示了数据定价方法的演进. Cong 等 [37] 研究了机器学习流程中的数据定价, 强调了该领域的原则和最新研究动态. Allouah 等 [38] 提出了一种量化数据信息内容以进行定价的新方法, 侧重于数据在定价策略中的价值. Miao 等 [39] 对现代数据定价模型进行了全面综述, 确定了数据定价领域的研究方向和新兴主题.

现行的数据定价策略通常受到卖方的影响, 导致买方对数据收集、清洗和打包的费用了解有限. 这种信息不对称导致了数据定价透明度的不足, 不利于卖方进行最优的市场定价, 也阻碍了买方在不同的数据服务提供商之间策略性地评估定价选项. 因此, 数据定价正在逐渐向更先进的定价结构转变. Huang 等 [40] 认为, 云计算服务的供应商采用混合定价策略是明智的, 这种策略将固定价格的预留服务和现货价格的按需服务结合在一起. Mei 等 [41] 构建了一个基于 Stackelberg 博弈的定价模型, 并主张在设备价格高和消费者评价差异大时, 采用纯捆绑策略, 而不是纯组件策略. Li 等 [42] 提出了一个理论框架, 用于根据其准确度为噪声查询答案分配价格, 并将价格分配给应该得到隐私损失补偿的数据所有者. Yu 等 [43] 提出了一个考虑了数据质量和数据版本策略的数据定价问题的双层数学规划模型. 结果表明, 当考虑多维数据质量时, 多版本策略实现了更好的市场细分, 更具利润, 更可行.

综上所述, 在数据定价领域, 不同的研究方法各有千秋. 传统成本法基于数据生成和处理的成本来定价, 其优点在于直观易懂, 但忽略了数据的潜在价值和市场需求, 可能导致定价偏低. 相比之下, 市场定价法通过市场供求关系来确定价格, 能够更灵活地反映数据的实际价值, 但其难点在于数据市场的信息不对称和交易不透明, 可能影响定价的准确性和公正性. 价值评估法则试图通过评估数据的内在价值和使用价值来定价, 具有更高的科学性和系统性, 但评估过程复杂且主观性较强. 而本文所探讨的创新性方法, 结合了上述方法的优点, 同时考虑了数据隐私、安全性及合规性等多维度因素, 旨在构建一个更加全面、公正且高效的数据定价机制, 从而凸显其独特的创新性和实际应用优势.

3　数据元件的定义与数学模型

3.1　数据产品和数据商品

在数据科学和数据交易领域, 对数据及其衍生物的深入理解是构建有效数据生态的基石. 这些衍生物包括数据产品、数据商品、数据元件以及数据集等多个概念, 每个都在数据价值链中发挥着不可或缺的作用 [44].

首先, 数据产品是基于原始数据, 通过加工、处理、分析和可视化等手段精心打造的有价值数据解决方案. 它不仅包含数据本身, 还融合了围绕数据的一系列服务, 如数据查询、分析和可视化, 旨在满足特定业务或研究需求, 从而提高决策效率和准确性. 数据产品的出现, 使得数据能够更直接地服务于业务, 成为推动业务发展的重要力量.

而数据商品, 则是数据市场上进行交易的对象, 具有明确的产权归属和交易价值. 无论是原始数据、数据产品还是数据服务 (如数据分析服务、数据 API 等), 只要它们能在市场上进行交易, 就都可以被视为数据商品. 数据商品化是数据经济发展的重要趋势, 它不仅促进了数据的价值实现和流通, 更使得数据成为了一种可交易的资产, 为数据经济的繁荣奠定了基础.

然而, 在数据产品的构建和数据商品的交易过程中, 我们不得不面对一个基础而重要的概念 —— 数据集. 数据集是数据的集合, 包含了多个数据项或数据记录, 它们可以是结构化的、半结构化的或非

结构化的. 作为数据分析和处理的基础, 对数据集的分析和处理通常需要结合特定的业务或应用场景进行. 为了将数据集转化为具有通用应用价值的数据单元, 我们引入了数据元件的概念. 数据元件是在数据集基础上, 经过数据抽取、转换、加载等处理形成的, 具有特定含义、格式和用途的数据单元. 它是数据产品的重要组成部分, 具有明确的业务价值和应用场景. 将数据集转化为数据元件, 不仅强调了数据集在特定业务或应用场景下的应用价值, 也凸显了数据从原始状态到实际应用过程中的价值转化过程.

综上所述, 数据产品和数据商品、数据元件和数据集在数据科学和数据交易领域都具有明确的区分. 数据产品是数据价值的直接体现, 数据商品是数据价值的流通媒介, 而数据元件则是数据从原始状态到实际应用过程中的关键转化单元. 理解这些概念, 不仅有助于我们更好地把握数据经济的发展趋势, 更有助于我们推动数据在业务中的应用和价值实现.

3.2　数据元件的定义

数据元件是连接数据供给和需求两端的 “中间态”, 它扮演着将原始数据转化为数据初级产品和交易标的物的角色, 同时也连接着数据资源与数据应用之间的关系. 通过将数据资源开发为数据初级产品, 可以实现数据的确权、计量、定价、监管和安全流通, 从而推动数据要素市场化的高效配置.

定义1 (数据元件)　作为近源数据的信息载体, 数据元件是具有一定主题, 经过对数据资源脱敏处理后, 根据需求由若干关联字段形成的数据集, 或由数据资源的关联字段通过建模形成的数据特征.

相比于传统数据集, 数据元件是针对特定业务需求和应用场景开发的, 具有明确的业务价值和应用场景. 相比之下, 数据集通常只包含原始数据, 不具备特定的业务价值. 在数据交易中, 买方通常关注的是数据能否满足其业务需求, 而非数据本身的格式或结构. 将数据集封装成数据元件, 可以简化数据交易过程, 降低买方的选择成本. 此外, 数据元件具有明确的业务含义和格式, 可以方便地在不同系统、平台和业务之间共享和复用. 这有助于提高数据利用效率, 推动数据资源的最大化价值发挥. 结合数据元件的上述定义, 数据元件的理论模型应具备如下特征.

(1) 内蕴信息可量化. 数据元件是数据产品开发的基础, 是近源数据的信息载体, 同时也是数据交易市场中的交易标的物. 而数据元件中内蕴的信息价值的可计量性则是其实现上述功能的前提条件. 因此, 数据元件模型需要建立一套统一的数据计量与定价框架 [45].

(2) 统一的数学表征. 数据要素规模化要求提高面向非特定应用的数据可用性, 即要求能够基于一套统一、标准化的基础数据服务于多种差异化的数据应用, 实现跨应用的数据复用. 因此, 数据元件统一的数学表征必不可少. 同时, 由于不同的数据元件对应不同的现实应用功能, 数据元件的数学表征应清晰刻画数据元件与实际功能之间的关联 [46].

(3) 原始数据间隐藏的逻辑关联的显式表达. 原始数据间往往隐藏着逻辑关联, 这些逻辑关联蕴含着重要的信息价值. 传统的模型往往仅是对原始数据的简单堆砌, 逻辑关联未被显式表达或直接忽略. 数据元件理论模型的核心思想就是重新构建并显式表达这些被忽略的原始数据间的逻辑关联. 从而能够明确计算数据元件理论模型与传统模型之间的信息差异 [47].

(4) 信息的动态性. 传统的模型除了不考虑原始数据信息的逻辑关联外, 也甚少考虑信息的动态变化. 然而, 实际数据所内蕴的信息经常发生变化. 因此, 数据元件理论模型的目标之一就是在计量信息时一并考虑信息的动态性 [48].

引入数据元件这种 “中间态” 后, 以数据元件为中心的数据要素化治理过程被明确分割为元件开发和元件应用两个关键环节. 首先是基于原始数据, 通过特征选择、特征抽取、聚合分析、统计分析等方法开发数据元件; 随后, 将数据元件作为安全流通、公允定价的数据 “中间态”, 以此作为流通要素, 赋能于应用, 并建立相关的定价审核机制.

4　数据元件的治理与数学模型

数据元件的数学模型构建如下: 数据空间是一个包含不同原始数据的集合, 表示为 $\widetilde{\boldsymbol{V}}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n)$. 集合中的元素 $\boldsymbol{v}_n=(v_n^{(1)},v_n^{(2)},\ldots,v_n^{(i)})$ 为不同的向量, 代表不同的原始数据. 向量中的每个元素 $v_n^{(i)}$ 代表一个信息计量单元. 集合 $\widetilde{\boldsymbol{V}}$ 中的不同元素间存在各种逻辑关联, 这些逻辑关联可以定义为一个集合 $\widetilde{\boldsymbol{E}}\subseteq\{(\boldsymbol{v}_n,\boldsymbol{v}_m):(\boldsymbol{v}_n,\boldsymbol{v}_m)\in\widetilde{\boldsymbol{V}}^2,\boldsymbol{v}_n\neq\boldsymbol{v}_m\}$. 基于上述定义, 数据空间中的成对关系可以表征为一个图 $\widetilde{\boldsymbol{G}}=(\widetilde{\boldsymbol{V}},\widetilde{\boldsymbol{E}})$, 其中 $\widetilde{\boldsymbol{V}}$ 和 $\widetilde{\boldsymbol{E}}$ 分别是顶点和边的集合 (根据附录 A 的介绍).

接下来, 考虑 $\widetilde{\boldsymbol{G}}$ 的一个子图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$, 其中 $\boldsymbol{E}\subseteq\widetilde{\boldsymbol{E}}$ 并且 $\boldsymbol{V}\subseteq\widetilde{\boldsymbol{V}}$. 对于子图 $\boldsymbol{G}$, 存在一个邻接矩阵 $\boldsymbol{A}$ 刻画图的连接关系. 根据邻接矩阵 $\boldsymbol{A}$, 可以定义图的拉普拉斯 (Laplace) 矩阵 $\boldsymbol{L}$, 满足 $L_{ij}=[(\delta_{ik}\sum_k A_{ik})-A_{ij}]$. 考虑到原始数据的信息会随着时间 $\boldsymbol{t}$ 演化, 则关于子图 $\boldsymbol{G}$ 的信息演化算子 (information propagator) 可以表示为 $\boldsymbol{K}=\boldsymbol{e}^{-\boldsymbol{tL}}$.

根据数据元件的概念定义, 数据元件是联系原始数据和特定功能数据产品的桥梁. 那么, 对于特定的应用功能, 存在一个操作 $\boldsymbol{M}$, 可以定义满足此种特定功能的数据元件为子图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$, 且子图的信息演化算子 $\boldsymbol{K}$ 与 $\boldsymbol{M}$ 可交换, 即 $\boldsymbol{KM}=\boldsymbol{MK}$.

综上, 数据元件的数学表征定义可归纳如下.

(1) 数据元件是原始数据空间的一个子图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$, 它同时包含了原始数据 $\boldsymbol{V}$ 的信息和原始数据间逻辑关联 $\boldsymbol{E}$ 的信息. 此定义对应数据元件与原始数据资源的联系.

(2) 对于旨在实现某种特定数据应用功能的数据元件 $\boldsymbol{G}$, 需满足以下关系: $\boldsymbol{KM}=\boldsymbol{MK}$. 其中, $\boldsymbol{M}$ 是实现应用功能的操作算子, $\boldsymbol{K}$ 是图 $\boldsymbol{G}$ 的信息演化算子. 上述交换关系表明, 数据元件并非原始数据空间中原始数据的随机组合, 只有满足上述交换关系的子图才能被视为数据元件. 此定义对应数据元件与数据产品的联系.

我们建立的数据元件理论模型与过往传统方法中直接使用的原始数据存在显著差异. 传统模型仅考虑原始数据集合 $\boldsymbol{V}$, 而不考虑原始数据间的逻辑关联 $\boldsymbol{E}$. 同时, 传统模型也忽略了实际原始数据源中信息的动态变化. 由于原始数据间存在关联, 一旦某个原始数据发生变化, 其他关联的数据也会有相应的变动. 而由于传统方法不考虑数据间的动态性和关联性, 使得它几乎无法刻画局部数据变化所造成的整体数据产品的全局变化, 从而导致数据应用层面出现错误和信息冗余等问题.

5　基于数据元件的信息计量与定价

5.1　基于数据元件的数据信息量计量

5.1.1　数据元件信息计量理论模型

本小节将详细介绍数据元件的信息计量模型, 其目标在于: (1) 精确计量数据元件的内在信息量; (2) 评估数据元件模型相较于传统方法在信息量上的差异.

在实际应用中, 针对同一描述对象, 往往存在多个不同的数据源 (参考 5.1.3 小节), 如图 2 所示. 这些数据源中的数据不仅描述了对象本身, 还揭示了对象之间的复杂关系, 同时数据中也可能包含重复和冗余的信息. 为了有效处理这些信息, 我们利用第 4.2 小节中构建的数据元件数学模型来表征这些多源数据.

考虑数据元件的图表示 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$, 其中每个数据元件的原始数据元 $\boldsymbol{v}_n=(v_n^{(1)},v_n^{(2)},\ldots,v_n^{(i)})$ 随时间变化的主方程可以表示为

$$\mathrm{d}\boldsymbol{v}_n=\sum_m\boldsymbol{A}_{nm}\varphi\left(\boldsymbol{v}_n,\boldsymbol{v}_m\right)\mathrm{d}t+\sigma\left(\boldsymbol{v}_n\right)\mathrm{d}W_t. \tag{1}$$

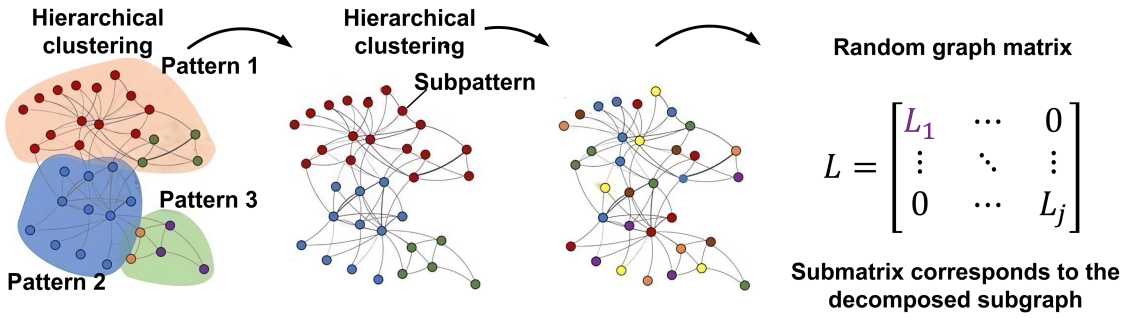

图 2　(网络版彩图) 数据元件网络的构成示意图.

Figure 2　(Color online) Schematic diagram of data component network.

在此方程中, $\varphi(\boldsymbol{v}_n, \boldsymbol{v}_m)$ 是数据元 $\boldsymbol{v}_n$ 和 $\boldsymbol{v}_m$ 之间的关联函数, 它量化了不同数据源或数据元之间信息的相互依赖和影响程度. 这种关联性可以通过多种数学方法定义, 如相似度、相关性、因果关系或互信息等.

一个具体的关联函数示例是基于皮尔逊相关系数 (Pearson correlation coefficient) 的变体, 它衡量了两个数据元之间线性关系的强度和方向. 对于数据元向量 $\boldsymbol{v}_n$ 和 $\boldsymbol{v}_m$, 它们之间的皮尔逊相关系数可以定义为

$$\rho_{nm} = \frac{\sum_{i=1}^{d}(v_n^{(i)} - \bar{v}_n)(v_m^{(i)} - \bar{v}_m)}{\sqrt{\sum_{i=1}^{d}(v_n^{(i)} - \bar{v}_n)^2}\sqrt{\sum_{i=1}^{d}(v_m^{(i)} - \bar{v}_m)^2}}, \tag{2}$$

其中, d 是数据元的维度, $\bar{v}_n$ 和 $\bar{v}_m$ 分别是 $\boldsymbol{v}_n$ 和 $\boldsymbol{v}_m$ 的均值. 在实际应用中, 这个相关系数可以转换为一个非线性的关联函数 φ, 用于上述主方程中, 以更准确地反映数据元之间的相互作用.

此外, $\sigma(\boldsymbol{v}_n)$ 建模了所有其他可能影响数据元 $\boldsymbol{v}_n$ 产生动态变化的因素, 这些因素被视为随机白噪声, 符合维纳过程 W_t. 当同一原始数据元中的信息计量单元 $v_n^{(i)}$ 相互独立时, 上述方程可以导出数据元件中的福克 – 普朗克方程 (Fokker-Planck equation), 用于进一步分析数据元的动态变化.

最后, 由信息熵的定义, 可以得到数据源 n 的信息熵为

$$H(v_n) = H\left(P\left(v_n^{(1)}\right), \ldots, P\left(v_n^{(i)}\right)\right) = -\sum_i P\left(v_n^{(i)}\right)\log_2 P\left(v_n^{(i)}\right), \tag{3}$$

其中, $P(v_n^{(i)})$ 是信息计量单元 $v_n^{(i)}$ 的概率分布. 通过计算信息熵, 可以量化数据源 n 的信息量, 进而为数据定价和信息管理提供重要依据.

$\sigma(\boldsymbol{v}_n)$ 将所有可能使数据元 $\boldsymbol{v}_n$ 产生动态变化的因素建模为随机白噪声, 该噪声符合维纳过程 W_t. 当同一原始数据元中的信息计量单元 $v_n^{(i)}$ 相互独立时, 上式可导出数据元件中的福克 – 普朗克方程

$$\frac{\partial P(v_n^{(i)} \mid t)}{\partial t} = -\frac{\partial P(v_n^{(i)} \mid t)\sum_m \boldsymbol{A}_{nm}\boldsymbol{\varphi}(v_n^{(i)}, \boldsymbol{v}_m)}{\partial v_n^{(i)}} + \frac{1}{2}\frac{\partial^2 \sigma(v_n^{(i)})^2 P(v_n^{(i)} \mid t)}{\partial {v_n^{(i)}}^2}, \tag{4}$$

其中 $P(v_n^{(i)} \mid t)$ 是时刻 t 数据源 n 中信息计量单元 i 为 $v_n^{(i)}$ 的概率. 当系统趋近于稳态时, 有 $P(v_n^{(i)} \mid t) = P(v_n^{(i)})$. 由信息熵的定义, 可得数据源 n 的信息熵为

$$H(v_n) = H\left(P\left(v_n^{(1)}\right), \ldots, P\left(v_n^{(i)}\right)\right) = -\sum_i P\left(v_n^{(i)}\right)\log_2 P\left(v_n^{(i)}\right). \tag{5}$$

根据信息熵的可加性原理, 信息元件的总信息熵值 $H(\boldsymbol{G})$ 可由其子系统的熵动态加和所得. 假设数据元件模型 $\boldsymbol{G}$ 由 n 个长度分别为 $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_n$, 边数为 $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_n$ 的原始数据源构成, $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$, 其中 $\boldsymbol{L} = \boldsymbol{i}_1 + \boldsymbol{i}_2 + \cdots + \boldsymbol{i}_n$, $\boldsymbol{K} = \boldsymbol{k}_1 + \boldsymbol{k}_2 + \cdots + \boldsymbol{k}_n$, 则数据元件的信息熵 $H(\boldsymbol{G})$ 可由下式计算:

$$H(\boldsymbol{G}) = \sum_{\alpha=1}^{n}\frac{\boldsymbol{i}_\alpha \boldsymbol{k}_\alpha}{\boldsymbol{L}\boldsymbol{K}} H\left(P\left(v_\alpha^{(1)}\right), \ldots, P\left(v_\alpha^{(\boldsymbol{i}_\alpha)}\right)\right) + H\left(\frac{\boldsymbol{i}_1\boldsymbol{k}_1}{\boldsymbol{L}\boldsymbol{K}}, \ldots, \frac{\boldsymbol{i}_n\boldsymbol{k}_n}{\boldsymbol{L}\boldsymbol{K}}\right). \tag{6}$$

表 1　实验参数.

Table 1　Experimental parameters.

Parameter symbol	Parameter meaning
$\mathrm{d}t$	Time step
n	Number of data sources in data component (nodes)
k	Number of logical connections between data sources (edges)
σ	Noise and other uncertainties
$\boldsymbol{A}_{nm}$	Logical connection between data sources
i	Total number of information units in each data source
T	Total iteration time

接下来, 我们将计算传统模型的信息熵. 传统模型可以认为是原始数据的简单收集堆砌. 假设传统模型中有 n 个长度分别为 $\boldsymbol{i}_1,\ldots,\boldsymbol{i}_n$, 边数为 $\boldsymbol{k}_1,\ldots,\boldsymbol{k}_n$ 的原始数据源 $\boldsymbol{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n)$, 其中 $\boldsymbol{L}=\boldsymbol{i}_1+\boldsymbol{i}_2+\cdots+\boldsymbol{i}_n$, $\boldsymbol{K}=\boldsymbol{k}_1+\boldsymbol{k}_2+\cdots+\boldsymbol{k}_n$, 则传统模型的信息熵为

$$H'(\boldsymbol{V})=\sum_{\alpha=1}^{n}\frac{\boldsymbol{i}_\alpha\boldsymbol{k}_\alpha}{\boldsymbol{LK}}H\left(P'\left(v_\alpha^{(1)}\right),\ldots,P'\left(v_\alpha^{(\boldsymbol{i}_\alpha)}\right)\right)+H\left(\frac{\boldsymbol{i}_1\boldsymbol{k}_1}{\boldsymbol{LK}},\ldots,\frac{\boldsymbol{i}_n\boldsymbol{k}_n}{\boldsymbol{LK}}\right). \tag{7}$$

假设原始数据信息计量单元的变化速率很慢, 则有 $P'(v_\alpha^{(\boldsymbol{i}_\alpha)})\approx P(v_\alpha^{(\boldsymbol{i}_\alpha)}\mid t=0)$. 定义 $\Delta H\equiv H(\boldsymbol{G})-H'(\boldsymbol{V})$, 则

$$\Delta H=\sum_{\alpha=1}^{n}\frac{\boldsymbol{i}_\alpha\boldsymbol{k}_\alpha}{\boldsymbol{LK}}\left[H\left(P\left(v_\alpha^{(1)}\right),\ldots,P\left(v_\alpha^{(\boldsymbol{i}_\alpha)}\right)\right)-H\left(P'\left(v_\alpha^{(1)}\right),\ldots,P'\left(v_\alpha^{(\boldsymbol{i}_\alpha)}\right)\right)\right]\leqslant 0. \tag{8}$$

这个不等式表明, 数据元件模型通过动态分析和数据整合, 相比传统模型, 在信息表达上更为高效, 具有更低的总信息熵.

相对熵值 $\Delta H\leqslant 0$ 可表明数据元件相比传统模型具有更强的确定性, 而这也是数据产品所希望达到的效果 —— 完全展现数据本身内在的规律性. 因此, ΔH 事实上量化了数据元件的相对信息价值. 从上述可知, 数据元件的相对信息价值来源于: (1) 将原始数据间的关联性考虑在内; (2) 考虑数据中信息的动态性.

上述观点对于数据元件的现实应用有两点重要意义: (1) ΔH 可为数据元件的最终定价提供重要的参考意义; (2) 上式可拆分为每个原始数据源 n 中每个信息计量单元 $\boldsymbol{i}_n$ 在数据元件中和在过去方法中的信息熵的差值的加和. 每一项差值对应的是现实中考虑将信息计量单元 $\boldsymbol{i}_n$ 纳入数据元件框架的工序所带来的附加信息价值, 因此可为数据元件产业链条的经济利益分配提供理论指引.

接下来, 本文将基于上述理论, 通过计算仿真, 利用两个在一定程度上基于真实场景的算例来证明数据元件的信息熵必然不大于传统模型的信息熵. 同时, 我们将会基于算例来展示如何计算相对信息熵, 以及通过考虑信息随时间的更新和不同的关联性来分配建构数据元件所得经济利益.

5.1.2　参数介绍

本小节介绍实验中涉及的参数以及参数所代表的意义. 对于每个步长 $\mathrm{d}t$, 算法步骤如下.

(1) 初始化以下参数. 数据元件所包含数据源数目 n、数据源间逻辑关联 $\boldsymbol{A}_{nm}$ 及其总数 k、噪声 σ (在下述算例中, σ 视为常数), 如表 1 所示.

(2) 初始化数据元件中每个数据源的每个信息计量单元 $v_n^{(i)}$. 由于每个信息计量单元的取值所服从的分布可能性众多, 因此我们将会根据实际情况按照两种方法初始化 $v_n^{(i)}$. 第 1 种情况, 该信息单元的取值随机性很强, 这种情况下我们不失一般性地认为 $v_n^{(i)}$ 服从一个近似均匀的分布. 第 2 种情况,

表 2　逻辑关联函数 $\varphi(\boldsymbol{v}_n, \boldsymbol{v}_m)$.

Table 2　Logical correlation function $\varphi(\boldsymbol{v}_n, \boldsymbol{v}_m)$.

Correlation type	$\varphi(\boldsymbol{v}_n, \boldsymbol{v}_m)$
Positive correlation	$av_m^{(i)} - v_n^{(i)} \quad (a > 0)$
No correlation	$av_m^{(i)} - v_n^{(i)} \quad (a = 0)$
Negative correlation	$av_m^{(i)} - v_n^{(i)} \quad (a < 0)$

该信息单元的取值确定性很高, 这种情况下我们不失一般性地认为 $v_n^{(i)}$ 服从狄拉克 (Dirac) δ 函数, 在仿真中我们采用方差很小的高斯 (Gauss) 分布代替狄拉克 δ 函数.

(3) 确定关联函数 $\varphi(\boldsymbol{v}_n, \boldsymbol{v}_m)$. 在现实场景中, 两个数据源之间的信息计量单元总数目不一定相同. 但通过填补缺失值或插入 "0", 可以令不同数据源具有相同的信息计量长度. 因此, 不失一般性地可以认为 $\boldsymbol{v}_n$ 和 $\boldsymbol{v}_m$ 是具有相同长度的向量. 同时, 数据源的信息单元 $v_n^{(i)}$ 有可能和 $\boldsymbol{v}_m$ 中多个信息单元存在逻辑关联. 这种情况可以通过数据预处理的手段使得所有数据源中的信息计量单元都互相独立. 因此, 可不失一般性地认为 $v_n^{(i)}$ 只和 $\boldsymbol{v}_m$ 中某一信息单元存在逻辑关联. 信息单元之间的逻辑关联可粗略地划分为 3 种关系: 正关联、无关联、负关联. 在实际场景中, 我们不失一般性地用表 2 中函数量化这 3 种关联性. 可以证明所有可解析的关联函数的泰勒展开的一阶项为表 2 所示函数.

(4) 根据模型主方程动态更新信息单元 $v_n^{(i)}$, 并储存每个时刻 t 的信息单元状态. 模型主方程的积分计算使用四阶龙格 – 库塔法, 步长 $\mathrm{d}t = 0.1$.

(5) 重复步骤 (4), 直到主方程收敛.

对于上述算法步骤, 我们对每个场景重复模拟 100 次. 因此, 我们可以统计得到概率分布 $P(v_n^{(i)})$. 此分布等价于上述理论模型中福克 – 普朗克方程的稳态解 $P(v_n^{(i)})$. 基于所得概率分布, 可计算信息熵 $H(\boldsymbol{G} \mid t)$. 同时, 根据传统模型信息熵的定义, 可通过仿真计算 $v_n^{(i)}$ 的初始分布来计算 $H'(\boldsymbol{V})$, 从而得到相对信息熵 ΔH.

5.1.3　案例分析

电能数据元件具有如下应用功能: (1) 面向政府部门, 支撑政府及时掌握经济运行动态、预测经济形势、科学制定宏观决策分析; (2) 面向企事业单位, 进行用能结构分析、行业对比、节能降费空间分析和实施建议. 电能的数据资源由大数据局和德阳电力提供, 共 16 个可供引用的数据源.

对这 16 个数据源进行预处理和不同的组合封装, 现已开发出 41 个数据元件. 这 41 个数据元件可大致分类为以下 4 类: 综合指标类、能耗类、经济产值类、用电量统计. 本文接下来的数值仿真计算将会针对以上几类典型元件进行. 由于缺乏相关元件的具体组装信息, 本文将会从原始数据源入手进行仿真. 具体思路是: 对数据源的潜在用途进行分类, 并考察特定用途下不同组合封装的数据元件模型的信息熵, 并进一步将仿真计算所得信息熵与传统模型的信息熵作对比.

算例 1.　首先以用电量统计为典型类别, 原始数据中涉及此类别的有表 3 中的 1, 3∼6, 8, 9, 12∼16 共计 12 个数据源. 其中, 这 12 个数据源又可分为两类, 第 1 子类为绝对用电量, 第 2 子类为用电占比. 接下来, 我们将会首先分别对这两子类进行计算仿真分析.

涉及绝对用电量的数据源有 1, 4, 12∼14 共 5 个, 我们考虑将这些数据源组合封装成数据元件. 数据元件的节点数 n 可以从 1∼5 不等, 而逻辑关联的最大数目 $k_{\max}$ 则根据组合数学计算得出, 即 $k_{\max} = \frac{n(n-1)}{2}$, 这个公式用于计算 n 个不同项中取两个进行组合的总数, 代表了不同数据源之间可能存在的最大关联数目. 在本例中, 由于有 5 个数据源, 所以 $k_{\max}$ 的值为 10, 表示在 5 个节点完全关联的情况下, 最多可以有 10 种不同的逻辑关联.

对于实际仿真, 我们考虑 k 作为实际存在的逻辑关联数目, 其取值范围从 0 (毫无关联) 到 $k_{\max}$ (完全关联). 因此, 在本例中, k 的取值范围为 0∼10.

表 3　电能数据资源概览.

Table 3　Overview of electric energy data resources.

Serial number	Data provider	Table name
1	State Grid Dezhou Power Supply Center	Industrial electricity consumption of city power company
2	State Grid Dezhou Power Supply Center	Enterprise load rate of city power company
3	State Grid Dezhou Power Supply Center	Electricity consumption proportion of each region in city power company in 2020
4	State Grid Dezhou Power Supply Center	Top ten electricity consumers of city power company
5	State Grid Dezhou Power Supply Center	YoY and MoM growth rate of electricity consumption by industry in city power company in 2020
6	State Grid Dezhou Power Supply Center	Electricity consumption proportion of city power company
7	State Grid Dezhou Power Supply Center	Patent information of city power company
8	State Grid Dezhou Power Supply Center	Electricity consumption proportion and YoY comparison of tertiary industry and residents in each region of city power company
9	State Grid Dezhou Power Supply Center	Electricity consumption proportion of industries in city power company (1)
10	Municipal Data	Power factor of city power company (monthly)
11	Municipal Data	Power factor of city power company (daily)
12	State Grid Power Company	Electricity consumption of industrial enterprises in each time period of city power company (1)
13	State Grid Power Company	Electricity consumption of industrial enterprises in each time period of city power company (2)
14	State Grid Power Company	Monthly electricity consumption of industrial enterprises of city power company
15	State Grid Dezhou Power Supply Center	Electricity consumption proportion of each industry in each region of city power company in 2018
16	State Grid Dezhou Power Supply Center	Electricity consumption proportion of industries in city power company (2)

当 $k = k_{\max} = 10$ 时, 我们假设数据源之间具有完全关联; 相反, 当 $k = 0$ 时, 我们假设数据源之间没有任何关联. 对于绝对用电量数据源, 我们假设其关联为正关联. 在这里, 为了简化模型, 假设数据源中的噪声很小并且是一个常数, 即 $\sigma = 0.0001$.

基于实际情况和计算时长的考虑, 我们在每组实验中设定数据源的信息计量单元长度为 $i = 20$. 同时, 我们设定主方程的数值积分步长为 $\mathrm{d}t = 0.1$. 为了探究不同初始化条件对结果的影响, 我们采用了两种不同的初始化方式: (1) 均匀分布; (2) 方差为 0.01 的高斯分布. 对于每种初始化方式, 我们都进行了 100 次独立的仿真计算, 以统计得到概率分布 $P(v_n^{(i)})$, 并基于这些概率分布计算了信息熵值. 表 4 展示了这些仿真实验的结果, 包括不同 k 值、不同初始化方式下的信息熵值统计.

从表 4 中, 可以得到如下重要结果. 根据表 4 中的数据, 可以看到数据元件的信息熵 $H(\boldsymbol{G})$ 确实低于传统模型下数据产品的信息熵 $H'(\boldsymbol{V}_1)$ 和 $H'(\boldsymbol{V}_2)$. 这验证了理论模型的预测, 即数据元件通过考

表 4　算例 1 的结果.

Table 4　Results of case 1.

n	k	$H(\boldsymbol{G})$	$H'(\boldsymbol{V}_1)$	$H'(\boldsymbol{V}_2)$	ΔH_1	ΔH_2
5	10	0.67	4.56	3.37	−3.89	−2.70
5	8	1.36	4.59	3.42	−3.23	−2.06
5	6	1.70	4.61	3.48	−2.91	−1.78
5	4	1.36	4.61	3.45	−3.25	−2.09
5	2	0.69	4.60	3.45	−3.91	−2.76
5	0	0.68	4.56	3.38	−3.88	−2.70

虑不同信息源之间的关联和信息的动态更新, 能够消除冗余、减少噪声、提高信息的时效性和准确性. 当数据源之间存在关联时, 数据元件能够捕捉到这些关联所带来的额外信息, 从而更全面地反映用电量数据的实际情况. 同时, 由于数据元件考虑了信息的动态更新, 它能够及时反映数据的变化趋势, 减少信息的不确定性. 这些因素共同作用, 使得数据元件的信息熵低于传统模型下的数据产品.

信息熵的降低意味着信息价值的提升. 因此, $|\Delta H|$ (即传统模型下数据产品的信息熵与数据元件信息熵之差) 可以作为定价的重要基准. 它反映了数据元件相比传统数据产品在信息价值上的提升程度, 为数据产品的定价提供了依据.

从表 4 中还可以发现, 当数据源之间毫无关联 ($k = 0$) 或完全关联 ($k = k_{\max}$) 时, 信息价值的提升程度 (即 $|\Delta H|$) 相对较大. 这是因为在这两种极端情况下, 数据源之间的逻辑关联和动态更新几乎被完全确定或消除, 使得用电量数据的规律性能够更好地体现在数据元件中. 具体来说, 当数据源之间毫无关联时, 虽然数据元件无法捕捉到关联所带来的额外信息, 但由于它消除了冗余和噪声、提高了信息的时效性, 因此仍然能够显著提升信息价值. 而当数据源之间完全关联时, 数据元件能够充分利用这些关联信息, 更全面地反映用电量数据的实际情况, 从而进一步提升信息价值.

相反地, 当数据源之间的逻辑关联介于毫无关联和完全关联之间时, 数据元件的信息价值提升程度虽然相对较小, 但仍然具有明显的信息增值. 这是因为数据元件能够捕捉到一定程度的关联信息, 虽然不如完全关联时全面, 但已经足够反映用电量数据的主要规律性. 此外, 数据元件的动态更新特性也使得它能够及时反映数据的变化趋势, 进一步减少信息的不确定性. 因此, 在定价时应该充分考虑数据源之间的逻辑关联程度. 完全关联的数据元件应该具有最高的价格, 因为它们能够最全面地反映用电量数据的实际情况. 而逻辑关联程度较低的数据元件虽然价格相对较低, 但仍然具有一定的信息价值.

算例 2.　继续以用电量统计为例子作为第 2 个算例. 我们考虑 12~14 这 3 个关于工业企业用电量的不同细粒度的统计数据集. 这 3 个数据源可以组成一个关于工业企业用电量的数据元件. 由于数据源中信息计量单元数目的不同, 我们将探究这一因素对数据元件相对信息价值的影响. 在这个算例中, 由于 3 个数据集都是关于同一事件的统计信息, 我们认为它们之间完全关联, 因此节点数目 $n = 3$, 逻辑关联的最大数目 $k = 3$. 我们改变数据源的总长度 i, 并保持与算例 1 中相同的其他模拟仿真参数, 结果如表 5 所示. 表 5 的结果可视化如图 3 所示.

首先, 我们需要明确数据模型中的符号与实际数据的对应关系. 在这里, 每个信息计量单元可以对应于原始数据集中的一段时间或一段区域内的用电量数据. 我们假设信息计量单元的概率分布定义域为用电量数据的可能取值范围, 而数据向量则是由多个信息计量单元组成的序列, 代表整个数据集的用电量情况.

从表 5 中, 可以得到以下重要结论: 数据源中的信息计量单元的数量不会剧烈影响数据元件的绝对信息价值 $H(\boldsymbol{G})$, 但会增加数据元件的相对信息价值. 原因在于, 数据元件的绝对信息价值主要由数据源之间的逻辑关联决定, 而在这个算例中, 逻辑关联是固定的 ($k = 3$). 然而, 信息计量单元数量的

表 5　算例 2 的结果.

Table 5　Results of case 2.

n	k	i	$H(\boldsymbol{G})$	$H'(\boldsymbol{V_1})$	$H'(\boldsymbol{V_2})$	ΔH_1	ΔH_2
3	3	20	0.69	4.09	3.19	−3.40	−2.50
3	3	200	0.69	6.22	3.66	−5.53	−2.97
3	3	2000	0.69	7.42	3.71	−6.73	−3.02

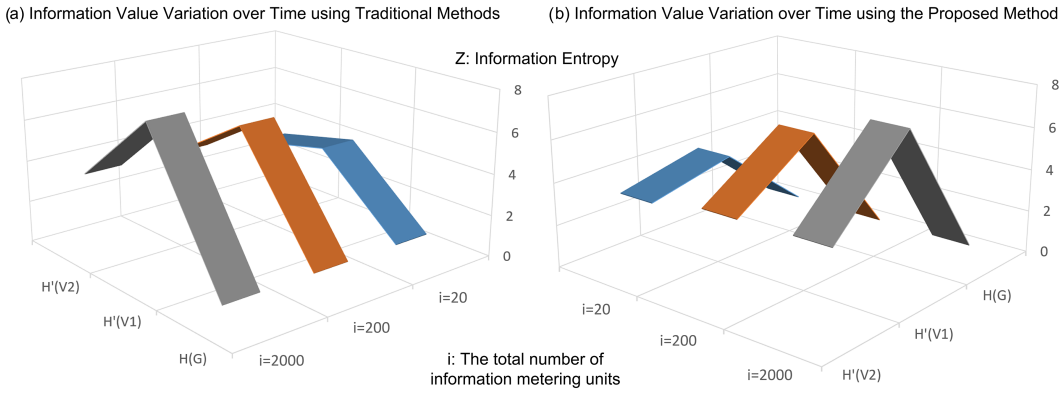

图 3　(网络版彩图) (a) 传统方法用于信息价值量化; (b) 所提出的方法用于信息价值量化.

Figure 3　(Color online) (a) Traditional method is used to quantify the information value; (b) the proposed method is used to quantify the information value.

多少直接影响了数据源中潜在的噪声水平. 随着信息计量单元的增加, 数据源中可能包含的噪声也随之增加, 这导致传统模型下数据产品的信息熵 $H'(\boldsymbol{V}_1)$ 和 $H'(\boldsymbol{V}_2)$ 也随之增加. 尤其是当初始化分布为均匀分布时 (代表信息单元历史分布的随机性很强, 即噪声很大), 这种增加趋势更为显著. 因此, 即使数据元件的绝对信息熵 $H(\boldsymbol{G})$ 保持不变, 其相对信息价值 (即 $|\Delta H|$) 也会因为原始数据所包含的处理噪声的成本的增加而增加.

这个算例不仅展示了信息计量单元数目对数据元件相对信息价值的影响, 还强调了在实际应用中考虑数据噪声和数据处理成本的重要性. 通过合理地选择信息计量单元的长度和数量, 我们可以更好地利用数据元件来提高数据的信息价值和实用价值.

算例 3.　实际场景中, 数据源之间的关联往往对数据的整体价值有着重要影响. 以用电量占比数据为例, 当多个数据集共同描述同一地区的产业用电情况时, 这些数据集之间的关联是显著的. 然而, 在实际操作中, 有时这些关联可能会被忽略, 导致数据价值的损失. 下面, 我们将通过用电量占比数据所对应的数据元件来分析这一问题, 并探讨逻辑关联对数据元件信息价值的影响.

我们考虑一个包含 7 个数据源的数据元件 a, 这些数据源分别刻画了德阳市不同产业的用电占比情况. 由于这些数据源共同描述了同一地区的用电情况, 我们可以合理假设它们之间存在完全关联. 现在, 我们假设在构建数据元件时, 忽略了其中某些数据源与其他数据源之间的关联, 得到新的数据元件 b. 为了量化这种关联被忽略后信息价值的损失, 我们进行了以下实验.

实验设定信息计量单元总长度 $i = 20$, 并逐次忽略一个数据源及其与其他数据源的关联. 例如, 当数据元件 a 的节点数 $n = 7$、逻辑关联数 $k = 21$ 时, 忽略其中一个节点的关联后, 其效用等同于节点数 $n = 6$、逻辑关联数 $k = 15$ 的数据元件 b. 我们计算了这两种情况下数据元件的信息熵以及传统模型下数据产品的信息熵, 并计算了它们之间的差值, 即相对信息价值的损失. 实验结果如表 6 和图 4 所示, 并得出以下结论.

(1) 当数据元件中存在的逻辑关联被忽略时, 其相对信息价值会随之降低. 这是因为逻辑关联是数据元件能够更全面地反映实际情况的关键因素之一. 当关联被忽略时, 数据元件无法充分利用这些

表 6　算例 3 的结果.

Table 6　Results of case 3.

n	k	$H(\boldsymbol{G})$	$H'(\boldsymbol{V_1})$	$H'(\boldsymbol{V_2})$	ΔH_1	ΔH_2
7	21	0.69	4.90	3.48	−4.21	−2.79
6	15	0.69	4.70	3.41	−4.01	−2.72
5	10	0.67	4.56	3.37	−3.89	−2.70

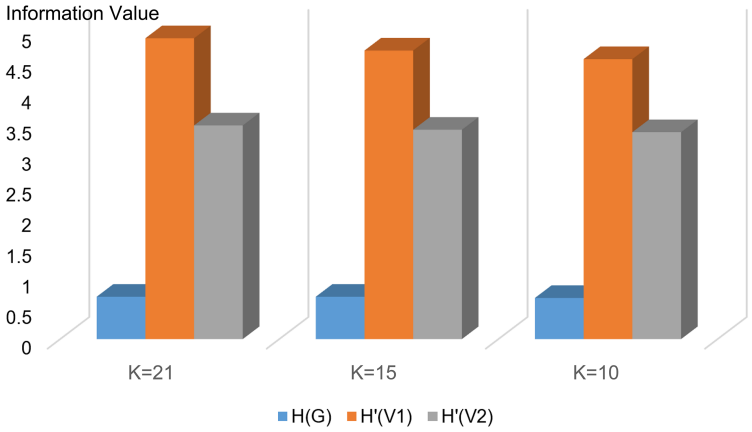

图 4　(网络版彩图) 数据间逻辑关联对信息价值的影响.

Figure 4　(Color online) Impact of logical associations between data on information value.

关联信息, 从而导致信息价值的减少.

(2) 逻辑关联建构错误对于由噪声较大的原始数据源组装而成的数据元件而言, 会带来更大的相对信息价值损失. 这是因为噪声较大的数据源本身就包含较多的不确定性, 当它们之间的关联被错误地忽略时, 这种不确定性会进一步放大, 从而导致信息价值的损失更为严重.

(3) 每个逻辑关联所带来的相对信息价值变化可以量化为 δH, 即相邻两行 ΔH 的差值. 这一指标反映了单个逻辑关联对数据元件信息价值的影响程度. 在实际应用中, 我们可以根据 δH 来评估数据元件加工链条上每个步骤的经济价值, 为数据元件获利后的经济利益分配提供参考基准.

综上所述, 逻辑关联对数据元件的信息价值具有重要影响. 在实际应用中, 我们应该充分重视数据源之间的关联, 确保在构建数据元件时能够充分利用这些关联信息, 以提高数据元件的信息价值和实用价值. 同时, 我们也需要关注数据源本身的噪声水平, 采取合适的方法来降低噪声对数据元件信息价值的影响.

5.2　基于数据元件的数据定价

5.2.1　数据元件信息价值评估

数据元件的信息价值与其信息附加价值和数据质量具有密切关系. 信息附加价值 (information additional value, IAV) 衡量的是数据元件的构建相比于过去一般数据处理方法所带来的额外熵减 ΔH. 这里的 “额外熵减” 指的是通过特定数据处理技术或方法, 使得数据的不确定性降低, 即信息的确定性增加. 数据质量则衡量的是数据元件构建过程中信息保持的完整性, 包括原始数据集 $\boldsymbol{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n)$ 的质量以及数据集间逻辑关联 $\boldsymbol{E}\subseteq\left\{\{\boldsymbol{v}_n,\boldsymbol{v}_m\}:(\boldsymbol{v}_n,\boldsymbol{v}_m)\in\boldsymbol{V}^2,\boldsymbol{v}_n\neq\boldsymbol{v}_m\right\}$ 的准确性.

在构建数据元件时, 若忽视了数据之间的逻辑关联或数据源本身存在质量问题, 都会导致信息损失. 此外, 如果逻辑关联构建错误, 也会对数据元件的信息价值产生负面影响. 因此, 数据元件信息价值的评估模型需要考虑这两部分因素.

第 1 部分是计算信息附加价值. 根据数据元件信息计量模型, 信息附加价值可由下式量化:

$$\mathrm{IAV} = \mathrm{IAV}(\Delta H), \tag{9}$$

其中, ΔH 表示额外熵减, 而 IAV 是 ΔH 的函数. 由于 $\Delta H \leqslant 0$ (表示熵减), IAV 是 ΔH 的减函数. 这意味着 ΔH 越小 (即信息确定性越大), 信息附加价值 IAV 就越大.

第 2 部分是量化数据质量. 数据质量包含两大类指标 $\boldsymbol{q}_V = (q_v^{(1)}, q_v^{(2)}, \ldots)$ 和 $\boldsymbol{q}_E = (q_e^{(1)}, q_e^{(2)}, \ldots)$, 分别对应原始数据源质量指标和关系结构的整合质量指标. 这些指标用于衡量数据源的准确性、完整性、一致性等, 以及逻辑关联的正确性和完整性.

数据质量的量化模型如下:

$$\mathrm{Quality}(\boldsymbol{Z}) = \left(\boldsymbol{Z}\left(\boldsymbol{q}_V, \boldsymbol{q}_E\right)^{\mathrm{T}} \boldsymbol{\beta}\right)^2, \tag{10}$$

其中, $\boldsymbol{Z}\left(\boldsymbol{q}_V, \boldsymbol{q}_E\right)$ 是数据质量指标矩阵, $\boldsymbol{\beta}$ 是指标权重矩阵. 这个模型通过加权求和的方式综合考虑了不同质量指标对数据质量的影响.

基于以上对数据体量、数据质量和信息附加价值的评估指标, 数据元件的信息价值 I 可表示为

$$I(\Delta H, \boldsymbol{Z}) = \mathrm{Quality}(\boldsymbol{Z}) \cdot \mathrm{IAV}(\Delta H). \tag{11}$$

这个模型将信息附加价值和数据质量两个因素相结合, 全面评估了数据元件的信息价值. 在实际应用中, 可以根据具体需求和数据特点调整指标和权重, 以获得更准确的评估结果.

5.2.2　数据定价模型

基于相关定价研究和上述对数据元件信息价值的评估, 本文提出两种不同交易场景中的价格生成机制: 议价交易机制和竞价交易机制. 以下将对这两种交易场景进行详细介绍.

第 1 种定价机制是议价交易机制. 这种机制通常适用于由数据生产者根据生产成本、数据元件的内在价值等因素与数据需求者共同协商议定价格的场景. 对于数据元件的协议价, 需综合考虑其开发和运维成本、信息价值等因素.

$$\mathrm{Price}_{\mathrm{Negotiated}}(I, C) = (1 + \gamma) \cdot f(I(\Delta H, \boldsymbol{Z}), C), \tag{12}$$

其中, $\mathrm{Price}_{\mathrm{Negotiated}}(I, C)$ 表示数据元件的协议价; γ 是预期收益率; f 是联合定价函数, 用于确定信息价值和成本对价格的影响; $I(\Delta H, \boldsymbol{Z})$ 代表数据元件的信息价值, 由额外熵减 ΔH 和数据质量 $\boldsymbol{Z}$ 决定; C 代表数据元件的成本, 可细分为

$$C = \mu \left(\sum_{n,i} C_{v_n^{(i)}} + C_{\mathrm{Integration}} \right), \tag{13}$$

其中, μ 是领域调节系数, $C_{v_n^{(i)}}$ 是数据源中某字段的开发和维护成本, $C_{\mathrm{Integration}}$ 是量化数据源之间逻辑关联的成本.

第 2 种定价机制是竞价交易机制. 这种机制通常适用于在数据交易市场组织下, 买卖双方根据数据的稀缺性、时效性等因素, 按照价格优先原则进行交易的场景. 对于数据元件的市场指导价, 可基于市场调节因子和市场评估价确定.

$$\mathrm{Price}_{\mathrm{Auction}} = \alpha \cdot \mathrm{Price}_{\mathrm{Market\ Estimate}}, \tag{14}$$

其中, $\mathrm{Price}_{\mathrm{Auction}}$ 是市场指导价; α 是市场调节因子, 根据市场供需关系等因素动态调整; $\mathrm{Price}_{\mathrm{Market\ Estimate}}$ 是市场评估价, 由数据的稀缺性 S、时效性 T、信息价值 ΔH、数据质量 $\boldsymbol{Z}$ 和成本 C 共同决定:

$$\mathrm{Price}_{\mathrm{Market\ Estimate}} = f(S, T, \Delta H, \boldsymbol{Z}, C). \tag{15}$$

这种定价机制允许市场力量在数据交易中发挥作用, 通过竞价过程实现数据的优化配置.

接下来, 我们模拟一个竞价过程, 其中多个买家 (数据需求者) 对同一数据元件进行竞买. 假设市场调节因子 α 初始设定为 1.0, 但会根据市场供需关系动态调整. 例如, 当某种数据元件的需求远超供应时, α 可能会上升, 以反映市场对该数据元件的渴求程度. 模拟竞价过程如下.

(1) 初始报价. 买家根据市场评估价和自己的需求紧迫性, 提交初始报价.

(2) 竞价轮次. 市场组织者设定多个竞价轮次, 每轮结束后, 买家根据当前最高报价和自身策略调整报价.

(3) 交易达成. 当某个轮次中没有新的更高报价出现时, 最高报价者获得数据元件, 交易达成.

现在, 我们通过一个具体的例子来说明竞价交易模型下的价格形成. 假设某种数据元件的市场评估价为 1000 单位货币, 初始有 5 个买家参与竞价, 他们的初始报价分别为 900, 950, 1000, 1050 和 1100. 第 1 轮竞价后, 最高报价为 1100, 该买家暂时领先. 第 2 轮竞价中, 其他买家根据市场反馈调整报价, 其中一人将报价提高到 1150, 以尝试赢得交易. 第 3 轮竞价竞争进一步加剧, 最高报价达到 1200. 交易达成时, 在某一轮竞价中, 如果没有新的更高报价出现, 且所有买家均确认不再调整报价, 则交易以当前最高报价 1200 单位货币达成. 通过这个过程, 我们可以看出竞价交易机制能够有效地反映市场供需关系, 使得数据元件的价格更加贴近其真实价值. 同时, 竞价过程也促进了市场资源的优化配置, 使得数据能够在最需要它的地方发挥作用.

5.2.3　案例分析

本小节将会通过 3 个例子来论述, 主要集中在议价场景下讨论:

(1) 如何根据数据元件的信息价值对其进行定价, 以及信息价值与最终价格之间的关系;

(2) 数据元件的价格如何随信息计量单元的数量和数据资源之间的关联数变化;

(3) 如何利用信息计量模型来对数据元件加工链条上的经济个体分配经济利益.

在本小节中, 我们关心的重点为数据元件的 “相对价格”. 这是因为数据元件作为产品在市场流通, 会受到大量和其本身内蕴信息价值无关的因素的影响, 例如元件的稀缺性、市场整体景气程度、人力成本等诸多环境因素. 因此, 我们的主要焦点将会集中在由数据元件相对信息价值所决定的 $f(I(\Delta H, \boldsymbol{Z}), C)$ 函数上. 这个函数我们在 5.2.2 小节中已经进行了介绍, 它主要由数据元件的相对信息价值 $I(\Delta H)$ 决定, 代表了数据元件的信息基准价格.

案例 1. 首先, 我们以用电量的绝对数值统计的相关数据源 (表 3 中数据源 1, 4, 12∼14) 组装开发而成的数据元件进行第 1 类分析. 本案例中, 原始数据向量是指对应于真实数据集中的用电量绝对数值统计数据, 即表 3 中数据源 1, 4, 12∼14 的具体数值. 信息计量单元是指在本案例中, 信息计量单元可以是不同时间段 (如日、月、年) 的用电量数据. 概率分布定义域是指对于用电量数据, 概率分布定义域可能包括不同时间段的用电量范围、变化趋势等. 关于符号与真实数据对应关系, 本案例使用 H 和 ΔH 来量化数据的不确定性和信息附加价值. 具体来说, H 可以表示为基础数据产品的信息不确定性, 而 ΔH 则表示数据元件相对于基础数据产品的信息附加价值. 这些值可以通过对真实数据进行分析和计算得到.

从表 3 可得到两个重要结论: (1) 数据元件相对信息价值由 ΔH 量化; (2) 给定数据元件所包含的数据资源数目, 对于可以实现同一数据应用目标的不同数据资源组合方式, 选取相互关联性极小或极大的方式能带来最高的信息附加价值.

这两个结论直接可应用到数据元件的定价上. 首先, 假设忽略成本因素, 并认为 f 随信息不确定性 H 线性递减, 有

$$f(H) = \beta_0 + \beta_1 H, \tag{16}$$

表 7　数据定价案例 1.

Table 7　Data pricing case 1.

n	k	ΔH_1	ΔH_2	$f(\Delta H_1)$	$f(\Delta H_2)$
5	10	−3.89	−2.70	107.78	105.40
5	8	−3.23	−2.06	106.46	104.12
5	6	−2.91	−1.78	105.82	103.56
5	4	−3.25	−2.09	106.50	104.18
5	2	−3.91	−2.76	107.82	105.52
5	0	−3.88	−2.70	107.76	105.40

其中 $\beta_1 \leqslant 0$. 对于不考虑数据资源关联性和信息动态性的数据产品, 其信息价值可以表示为

$$f(H_0) = \beta_0 + \beta_1 H_0. \tag{17}$$

对于数据元件, 其信息附加价值 ΔH 决定了其相对于基础数据产品的信息价值增量. 因此有 $f(H) - f(H_0) = \beta_1 \Delta H$, 所以数据元件的信息价格函数为

$$f(\Delta H) = f_0 + \beta_1 \Delta H, \tag{18}$$

其中, f_0 可以看作是现有非数据元件的数据相关产品的参考价格, 我们假设它为常数. 需要注意的是, 上述讨论未考虑成本 C. 在实际场景中, 数据元件的成本可能会高于现有的数据产品, 因此其基础价格 (即 f_0) 可能会更低. 因此, 上述公式中的数据元件价格 $f(\Delta H)$ 应视为下限, 即

$$f(\Delta H) \leqslant f(\Delta H, C) \leqslant \text{price}(\Delta H, C) = (1+\gamma) f(\Delta H, C), \tag{19}$$

其中, $\gamma \geqslant 0$ 表示预期收益率.

接下来, 我们将表 7 中的 ΔH 的计算结果结合上述公式进行讨论. 我们假设 $f_0 = 100, \beta_1 = -2$. 表 7 的后两列展示了在给定固定数量的数据资源组装开发数据元件时, 数据资源之间几乎不关联和完全关联两种情形下数据元件的价格下限 $f(\Delta H)$ 的对比. 由于数据元件的成本 C 主要由两部分组成: 单个数据资源的成本 $C_{v_n^{(i)}}$ (其中 n 表示数据资源的类型, i 表示该类型的第 i 个实例) 和关联数据资源的成本 C_k. 在这里, 我们假设关联成本 C_k 与关联度 k 成正比, 即 $C_k \propto k$.

在数据资源几乎不关联的情况下, 关联度 k 较低, 因此关联成本 C_k 也相对较低. 而在数据资源完全关联的情况下, 关联度 k 较高, 导致关联成本 C_k 显著增加. 由于总成本 C 是单个数据资源成本和关联成本的总和, 即

$$C = \mu \left(\sum_{n,i} C_{v_n^{(i)}} + \sum_k C_k \right), \tag{20}$$

其中, μ 是一个常数因子. 完全关联情况下较高的总成本 C 会导致对应的数据元件价格下限 $f(\Delta H, C)$ 也更高. 这是因为价格下限通常是基于成本来设定的, 以确保数据元件的供应在经济上是可行的.

因此, 从成本分析的角度来看, 数据资源之间的关联程度越高, 数据元件的价格下限也越高.

案例 2. 我们继续以用电量的绝对数值统计的相关数据源 (12∼14) 组装开发而成的数据元件进行第 2 类分析. 从表 4 可得到重要结论: 数据元件的相对信息价值会随着数据源中的信息计量单元 (例如字节) 的增加而增多, 原因是数据元件能更好地清理原始数据中的信息噪声. 接下来, 我们将表 4 中的 ΔH 的计算结果结合上述公式进行讨论. 我们假设 $f_0 = 100$, $\beta_1 = -2$.

从表 8 中, 我们可以看到根据 ΔH 的不同值, 可以计算出不同信息计量单元数量下数据元件的价格下限. 由于 β_1 为负值, 随着 ΔH 的增加 (即信息附加价值的增加), 价格下限 $f(\Delta H)$ 也会增加, 这

表 8　数据定价案例 2.

Table 8　Data pricing case 2.

n	k	i	ΔH_1	ΔH_2	$f(\Delta H_1)$	$f(\Delta H_2)$
3	3	20	-3.40	-2.50	106.8	105.00
3	3	200	-5.53	-2.97	111.16	105.94
3	3	2000	-6.73	-3.02	113.46	106.04

表 9　数据定价案例 3.

Table 9　Data pricing case 3.

n	k	ΔH_1	ΔH_2	$f(\Delta H_1)$	$f(\Delta H_2)$
7	21	-4.21	-2.79	108.42	105.58
6	15	-4.01	-2.72	108.02	105.44
5	10	-3.89	-2.70	107.78	105.40

反映出信息价值的提升对价格下限的正向影响. 同时, 如果考虑到成本 C 的变化, 价格的实际值可能会更高, 特别是当信息计量单元数量增加导致成本上升时.

具体来说, 当数据元个数 n 与逻辑关联数 k 维持固定比例时, 传统模型中的信息不确定性 H_0 会随着数据计量单元 i 的增大而增大. 然而, 由于数据元件的设计和优化, 其能够更有效地处理这些信息, 导致数据元件的相对信息价值 $|\Delta H|$ 也随之增大. 这可以从表 8 中 ΔH_1 和 ΔH_2 的变化得到验证.

接下来, 我们根据给定的 $f_0 = 100$ 和 $\beta_1 = -2$, 计算了数据元件的价格基准 $f(\Delta H)$. 从表 8 的最后两列可以看出, 随着数据体量的增大 (即 i 的增大), 价格基准 $f(\Delta H)$ 也在增大. 这进一步验证了数据体量与数据元件相对信息价值之间的正相关关系.

此外, 根据成本计算公式 (20) 以及假设 $C_{v_n^{(i)}} \propto i$, 可以推断出数据元件的字段开发和维护成本也会随着数据体量 i 的增大而增大. 因此, 在考虑成本因素后, 实际的数据元件价格 $f(\Delta H, C)$ 随 i 的增长速率会高于仅考虑信息价值的 $f(\Delta H)$.

综上所述, 本案例分析表明, 在数据元件的设计和定价过程中, 数据体量是一个重要的考虑因素. 通过增大数据体量, 数据元件能够提高其相对信息价值, 但同时也需要承担更高的开发和维护成本. 因此, 在制定数据元件的定价策略时, 需要综合考虑其信息价值和成本因素.

案例 3.　我们以用电量占比数据 (表 3 中 3, 5, 6, 8, 9, 15, 16) 所对应的数据元件作为例子来分析如何对数据元件加工链条上的经济个体分配经济利益. 从表 5 可得到重要结论: 当数据元件中存在的逻辑关联建构错误或被忽略时, 其相对信息价值会随之损失, 而损失的相应的附加信息价值恰好就是对应此逻辑关联的价值. 首先, 我们明确数据元件中的两个关键参数: 数据元个数 n 和逻辑关联数 k. 这两个参数共同决定了数据元件的相对信息价值 ΔH. 我们假设信息价值函数 $f(\Delta H)$ 的形式已知, 其中 $f_0 = 100$ 和 $\beta_1 = -2$ 是给定的常数.

接下来, 我们利用表 9 中的数据来量化逻辑关联对应的信息价值. 表 9 中第 2 和 3 行展示了当数据元件中缺失了某个节点 (数据源) 时, 所对应的逻辑关联数从 21 减少到 15, 即减少了 6 条边 (逻辑关联). 相应地, 我们观察到损失的信息确定性为 $\delta H_1 = \Delta H_1\ (n = 7, k = 21) - \Delta H_1\ (n = 6, k = 15) = -0.2$. 这意味着这 6 个不同的逻辑关联共同对应了 0.2 单位的信息价值损失.

为了确定每个逻辑关联所对应的价值, 可以将总的信息价值损失平均分配到每个缺失的逻辑关联上. 因此, 每个逻辑关联的价值可以计算为 $\frac{\delta H_1}{\text{缺失的逻辑关联数}} = \frac{-0.2}{6} \approx -0.033$. 但请注意, 由于我们关注的是附加信息价值, 即正值, 所以实际上我们关心的是由于逻辑关联缺失导致的价格减少. 因此, 需要计算的是每个逻辑关联对应的价格损失.

价格损失可以通过比较缺失逻辑关联前后的价格基准来计算, 即 $f[\Delta H_1\ (n = 7, k = 21)]$

表 10　数据规模扩展与扰动实验表.

Table 10　Data scale expansion and perturbation experiment table.

Dataset size	Average price before perturbation $\bar{P}_{\text{before}}$	Standard deviation before perturbation $s_{P,\text{before}}$	Average price after perturbation $\bar{P}_{\text{after}}$	Standard deviation after perturbation $s_{P,\text{after}}$
D_1	5.00	1.20	5.02 ± 0.05	1.22 ± 0.03
D_2	5.05	1.18	5.07 ± 0.04	1.19 ± 0.02
D_3	5.03	1.15	5.05 ± 0.03	1.16 ± 0.02
D_4	5.02	1.10	5.03 ± 0.02	1.11 ± 0.01

$-f[\Delta H_1\ (n=6,k=15)]$. 假设这个差值等于 0.4 (这里是一个假设值, 实际值应基于表 9 中的具体数据计算), 则每个逻辑关联对应的价格损失约为 $\frac{0.4}{6} \approx 0.067$. 这样, 就得出了每个逻辑关联在数据元件中对应的经济价值大约为 0.067 单位.

5.2.4　模型鲁棒性分析

为了全面评估基于信息熵的数据定价模型的鲁棒性, 我们将通过具体的公式、数值表格以及额外的数据扰动实验来详细展示. 首先, 我们定义原始数据集 D, 其中每个数据点 x_i 具有对应的信息熵值 $H(x_i)$, 这些值共同构成了数据集的信息熵分布.

为了模拟数据收集过程中的随机误差或系统偏差, 我们设计了一个数据扰动实验. 在此实验中, 随机选择数据集 D 中的一部分数据点 (例如, 10%), 并对这些数据点的信息熵值进行扰动. 扰动量 ΔH 从正态分布 $N(0,\sigma^2)$ 中随机抽取, 其中 σ 控制扰动的强度. 通过比较扰动前后的数据定价结果, 可以评估模型对噪声的敏感性和鲁棒性.

基于信息熵的数据定价模型通常将数据点的价格 $P(x_i)$ 与其信息熵 $H(x_i)$ 相关联, 采用如下公式:

$$P(x_i) = \alpha \cdot \mathrm{e}^{\beta \cdot H(x_i)}, \tag{21}$$

其中, α 和 β 是模型参数, 需根据具体应用场景进行调整.

为了评估模型在不同规模数据集上的表现, 我们进行了数据规模扩展实验, 具体步骤如下:

(1) 数据集构建. 从一个小规模数据集 D_1 开始, 逐步增加数据量, 形成一系列规模递增的数据集 D_2, D_3, D_4;

(2) 模型定价. 对每个数据集使用基于信息熵的数据定价模型进行定价, 计算每个数据集的平均价格 $\bar{P}$ 和标准差 s_P;

(3) 稳定性分析. 分析不同规模数据集下平均价格和标准差的稳定性, 以评估模型的鲁棒性.

数据规模扩展实验及数据扰动实验的数值如表 10 所示. 在数据规模扩展实验中, 我们观察到几个关键趋势:

(1) 平均价格的稳定性. 随着数据集规模的增加, 平均价格 $\bar{P}$ 保持了相对稳定, 表明模型在不同规模的数据集上能够产生一致的价格评估;

(2) 标准差的减小. 标准差 s_P 的逐渐减小说明价格评估的离散程度在降低, 即模型对不同数据点的价格评估变得更加一致和稳定;

(3) 扰动实验的鲁棒性. 通过对比扰动前后的平均价格和标准差, 我们发现模型对适度的数据扰动具有一定的鲁棒性, 价格评估的波动在可接受范围内.

综上所述, 数据规模扩展实验和数据扰动实验的结果均表明, 我们提出的基于信息熵的数据定价模型在不同规模的数据集上均表现出良好的鲁棒性, 且对适度的数据扰动具有一定的抵抗能力. 这些结果增强了模型在实际应用中的可靠性和可预测性.

5.3　模型实用性分析

本文所提出的数据定价模型, 通过量化数据元件的相对信息价值 ΔH 及其与经济价值之间的关系 $f(\Delta H)$, 为数据定价提供了一种新颖的视角和方法. 模型的实用性体现在其通过引入信息熵概念有效量化数据元件中的信息价值, 为数据交易双方提供了清晰的价值评估基础, 促进了交易的公平性. 同时, 模型还充分考虑了数据元件的开发和维护成本 C, 确保了定价的合理性, 能够更准确地反映数据的实际价值. 此外, 模型中的参数 (如 f_0, β_1 等) 具备高度的灵活性, 可根据不同行业和领域的实际需求进行调整, 进一步拓宽了模型的应用范围, 使其能够广泛适用于多样化的数据定价场景. 本文所提出模型的潜在应用场景如下所述.

(1) **数据交易平台.** 在数据交易平台上, 买卖双方可以利用该模型对各类数据产品进行定价. 卖家可以基于模型评估其数据产品的信息价值, 并结合开发成本和市场需求设定合理的售价; 买家则可以根据模型分析数据产品的潜在价值, 从而做出更明智的购买决策. 这不仅提高了数据交易的透明度和效率, 还有助于促进数据市场的繁荣发展.

(2) **企业内部数据管理.** 企业在管理和利用自身数据资源时, 也可以运用该模型来评估不同数据集的价值. 通过比较不同数据元件的信息价值和成本, 企业可以优化数据资源配置, 提高数据利用效率, 并为企业决策提供有力的数据支持. 此外, 该模型还有助于企业识别高价值的数据资产, 进而制定有效的数据保护策略, 防止数据泄露和滥用.

(3) **跨行业数据合作.** 在跨行业数据合作中, 不同行业的企业往往拥有各自独特的数据资源. 通过运用该模型, 各参与方可以清晰地了解各自数据资源的价值, 并基于价值进行公平的利益分配. 这有助于促进跨行业数据共享和整合, 实现数据资源的最大化利用, 推动产业创新和升级.

(4) **政府数据开放与监管.** 政府在推动数据开放和共享的同时, 也需要对数据的使用进行监管, 以确保数据安全和个人隐私保护. 通过运用该模型, 政府可以评估不同数据集的公共价值和社会影响, 从而制定合理的开放策略和监管措施. 这有助于在保障数据安全和个人隐私的前提下, 最大限度地发挥数据的公共价值和社会效益.

6　数据元件的挑战及未来发展启示

本文深入探讨了数据定价问题在数字经济中的核心挑战与应对策略, 特别是在双碳节能和跨行业数据流通背景下的重要性. 研究成果揭示了数据价值量化的复杂性及其与多种因素 (如数量、质量、应用场景等) 的紧密联系, 强调了建立科学、合理、公平的数据定价机制的必要性和紧迫性. 通过案例分析和理论探讨, 本文提出了数据元件在数据价值实现中的关键作用, 以及数据定价策略在特定领域 (如通信网络赋能工业、车联网车路协同) 中的实际应用价值.

在实际应用上, 本文的研究为数据交易市场的规范化发展提供了理论支持和实践指导, 有助于推动数据资源的高效配置和价值最大化. 在双碳节能领域, 本文的研究促进了数据流通与能耗优化的结合, 为人工智能大模型等前沿技术的绿色应用提供了新思路.

6.1　数据元件面临的挑战

通过上述案例可以得知, 数据定价问题在宏观层面上主要体现为数据作为新型生产要素在数字经济中的价值评估与交易机制问题. 随着数据量的爆炸性增长和大数据技术的快速发展, 数据已成为推动经济增长、促进产业升级的重要力量. 然而, 由于数据的非标准化、非同质化以及价值释放过程的复杂性, 数据定价面临着诸多挑战.

首先, 数据的价值难以准确量化. 数据的价值不仅取决于其本身的数量和质量, 还与其应用场景、处理方式、使用方式等因素密切相关. 这使得数据价值的评估变得十分复杂, 难以形成统一的标准和方法. 目前, 数据交易市场仍处于起步阶段, 缺乏完善的市场机制和法律法规支持. 数据交易过程中存

在信息不对称、隐私泄露、权益纠纷等问题, 影响了数据交易的效率和公平性. 传统的定价机制往往基于成本或市场供需关系来确定价格, 但这种方法难以适用于数据这种新型生产要素. 数据定价需要综合考虑数据的价值、成本、市场需求、应用场景等多个因素, 建立科学、合理、公平的定价机制.

此外, 在双碳节能的背景下, 数据定价与跨行业数据流通对能源消耗的影响愈发凸显, 特别是在人工智能生成内容服务的快速发展中, 数据要素的角色愈发关键. 以人工智能大模型为例, 其在生成高质量内容的同时, 也伴随着巨大的能耗挑战. 数据准确性对于人工智能大模型的运行效率至关重要. 准确的数据输入能够降低模型的计算复杂度, 减少不必要的计算量, 从而降低能耗.

综上所述, 数据定价与跨行业数据流通在双碳节能的大背景下具有极其重要的意义. 通过优化数据定价机制, 促进数据的跨行业流通, 可以更有效地利用数据资源, 为节能减排提供有力支持. 同时, 我们也需要关注人工智能大模型的能耗问题, 通过降低模型的计算复杂度进而降低整体的能源消耗. 这些措施的实施, 将有助于更好地应对全球气候变化带来的挑战, 推动绿色低碳经济的发展.

因此, 数据元件在实现数据价值的价值链中扮演着至关重要的角色. 一方面, 数据元件作为信息的承载者, 包含着原始数据所携带的信息, 以满足业务场景的需求; 另一方面, 数据元件是交易的对象, 可作为数据资产计量和定价的基本单元, 解决数据资产化的问题. 在前沿技术中, 数据价值评估同样发挥了至关重要的作用, 例如:

(1) **通信网络赋能工业中的信息传递.** 在工业 4.0 时代, 远程控制和协同操控是实现智能制造的关键技术. 而通信网络的时延和稳定性直接影响到远程控制和协同操控的效率和准确性. 因此, 对于通信网络中传输的数据, 需要根据其时延和稳定性要求来制定合理的数据定价策略. 例如, 对于需要实时响应的数据, 可以采用高价位的优质数据传输服务, 以确保数据的及时性和准确性.

(2) **车联网中车路协同场景全局信息对于交通状况改善的价值.** 在车联网中, 汽车自主感知信息和车路协同场景全局信息对于改善交通状况具有重要价值. 这些数据可以帮助车辆实现更智能的驾驶决策, 减少交通事故和拥堵现象. 因此, 对于车联网中的数据, 需要根据其对于交通状况改善的贡献程度来制定数据定价策略. 例如, 对于能够显著提高交通流畅度的全局信息, 可以采用高价策略以激励更多的数据提供方参与数据共享和交易.

6.2　未来发展启示

展望未来, 数据元件的研究和应用将具有广阔的前景. 随着数字化、网络化和智能化的进一步发展, 数据的数量、类型和复杂性都将不断增加, 如何科学地评估和定价数据的价值将成为一个越来越重要的问题. 数据元件作为一种新的研究方法, 将为解决这一问题提供更加有效的途径. 此外, 数据元件的研究和应用也将对经济社会各领域产生深远的影响.

例如, 在智慧能源领域, 通过数据价值的准确评估和合理定价, 能促进能源的智能化管理和能源的高效利用. 电能数据不仅包含了电网的运行状态、供需关系等基本信息, 还蕴含着能源使用的模式、节能潜力等重要价值. 因此, 如何构建一套科学、合理的电能数据价值评估体系, 成为了亟待解决的问题. 这需要考虑数据的来源、质量、实时性、安全性等多个维度, 以及其在能源管理、节能减排、市场交易等方面的应用潜力. 其次, 电能数据作为一种新型生产要素, 其定价机制应当综合考虑数据的价值、成本、市场需求、应用场景等多个因素. 如何平衡数据提供方、使用方和平台方的利益, 确保数据的公平交易和有效利用, 是制定合理定价策略的关键. 最后, 需要关注电能数据的安全和隐私保护问题. 在能源互联网中, 电能数据的传输和共享涉及多个环节和多个主体, 如何确保数据的安全性和隐私性是一个重要的问题. 需要研究如何采用先进的加密技术、访问控制技术等手段, 保护电能数据的安全和隐私. 同时, 还需要关注数据泄露、非法获取等风险, 制定相应的应急预案和处置措施.

在金融领域, 通过对数据的科学评估和定价, 可以为金融产品的设计和风险管理提供更加科学和准确的依据; 在科技领域, 通过对数据的深入分析和挖掘, 可以发现新的规律和趋势, 推动科技创新和产业升级; 在社会服务管理领域, 通过对数据的合理利用和保护, 可以提高社会服务的质量和效率, 促

进社会的可持续发展.

综上所述, 数据要素及其市场化配置的需求将带来全新的应用场景. 在这种场景中, 数据需要在不同利益主体之间进行大规模的流通、共享和协同处理. 因此, 这种数据要素化过程需要一种便于流通、使用和管理的新的数据形态和信息组织方式. 这种数据形态需要具备两个重要特点: 一是有效性, 能够高效支持大规模的数据应用; 二是流通性, 方便实现跨利益主体的数据流动.

通过上述案例可以得知, 数据元件在实现数据价值的价值链中扮演着至关重要的角色. 一方面, 数据元件作为信息的承载者, 包含着原始数据所携带的信息, 以满足业务场景的需求; 另一方面, 数据元件是交易的对象, 可作为数据资产计量和定价的基本单元, 解决数据资产化的问题. 未来, 数据元件还将具备以下属性和功能.

6.2.1　数据元件将成为适应数据要素化需求的数据组织方式

数据具有多变的特征. 随着数据的加工和价值释放过程, 数据的形态不断演变, 经历了数字化、数据化、知识化的转变. 数字化将物理世界状态映射到信息空间中, 为后续加工处理提供基础, 并实现生产方式、分工形式、商业模式的改变; 数据化基于数字化内容, 将事实和观察结果转化为可量化分析的形式; 知识化是深度挖掘数据, 基于积累的经验和信息挖掘出数据背后的规律, 形成具有泛化和推广能力的知识和技术. 值得注意的是, 在这一演进过程中, 不仅数据的内容发生变化, 信息承载和组织方式也在不断变化. 在数字化阶段, 使用比特作为信息的载体, 重点关注物理世界状态变动对应的数据变动; 数据化阶段通常使用表格来组织相关信息, 表格中同一行的不同字段代表着一组存在关联的属性; 在知识化阶段, 信息组织方式更加复杂, 可用谓词表达、生成式表达、语义网络等方法.

数据要素及其市场化配置的需求将带来全新的应用场景. 在这种场景中, 数据需要在不同利益主体之间进行大规模的流通、共享和协同处理. 因此, 这种数据要素化过程需要一种便于流通、使用和管理的新的数据形态和信息组织方式. 这种数据形态需要具备两个重要特点: 一是有效性, 能够高效支持大规模的数据应用; 二是流通性, 方便实现跨利益主体的数据流动. 因此, 需要引入数据元件作为数据要素的承载形态.

6.2.2　数据元件将成为数据要素化治理的信息载体

数据的组织方式对数据的可用性产生影响, 决定了数据是否能够有效且便捷地支持上层应用. 随着信息化进程的推进, 数据的规模和复杂性不断增加. 为确保数据的可用性, 需要利用技术手段对复杂的数据进行整理和维护. 数据治理即是对数据组织方式进行规范化处理和维护的过程.

对数据组织方式的规范化处理在信息系统的各个层级都在不断进行. 在数据库最初生成时, 由大量表格组成的数据库系统通常存在大量的数据冗余, 同样的字段在不同表格中反复出现. 这种冗余不仅浪费存储空间, 更重要的是, 当相同的数据在不同表格中重复出现时, 可能因内容不一致而引发冲突和错误. 这种错误在频繁的插入、删除和修改操作中难以避免. 为解决这些问题, 需要对数据库设计进行规范化, 从而引入了数据库范式. 数据库范式规范了数据表的结构, 并约束了数据表之间的关系. 其根本目标是节省存储空间, 避免数据不一致性, 提高关系操作的效率, 并同时满足应用需求. 数据库范式的引入使得跨表的数据操作和流转变得更加顺畅. 这种数据库的规范化过程也可以视为数据治理的过程, 它发生在单一信息系统内部的数据治理.

面对数据要素市场化的新场景, 数据将在更广泛的范围内进行流通和共享, 更重要的是, 这种数据流动将跨越不同利益主体或治理主体的边界. 此时, 需要处理的不仅仅是数据内容的一致性和语义的一致性, 还需要关注伴随数据流动而来的收益和风险如何在不同利益主体和治理主体之间分摊. 因此, 数据要素市场化需要新的治理过程, 需要对市场化配置和流通的数据进行规范化, 以便管理收益和风险.

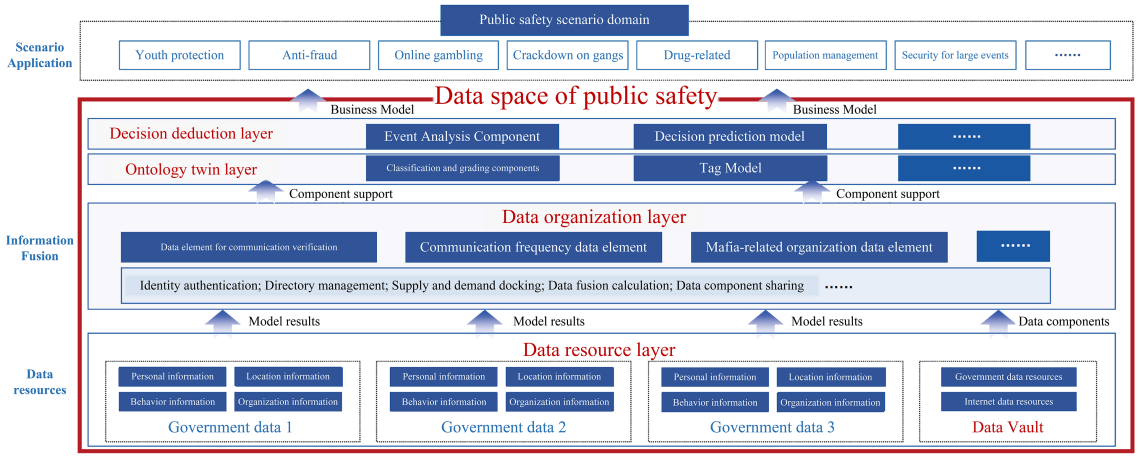

图 5　(网络版彩图) 政务数据要素化管理示意图.

Figure 5　(Color online) Schematic diagram of government data element management.

6.2.3　数据元件是连接数据供需两端的 “中间态”

数据的价值开发过程伴随着数据形态的演变. 最初, 数据以原始数据的形态出现, 是人类对客观事物的数字化记录或描述. 由于原始数据是无序的、未经加工处理的素材, 尚不具备使用价值, 难以直接投入生产. 当数据具备了使用价值并能直接投入社会生产经营活动中时, 原始数据就转化为数据资源. 进一步, 数据资源必须进入市场参与流通才能激活价值, 当数据在生产经营活动中实现流通, 并为使用者创造经济效益时, 它便转化为数据要素. 最后, 数据经生产形成产品或服务时, 数据便以产品和服务的形态呈现出来. 这一数据价值开发的过程经历了数据资源化、数据资源要素化、数据要素产品化 3 次重大的形态转变和价值增值.

以数据元件为中心, 在未来可实现数据价值链和数字资产链 “双链融合”. “数据资源 – 数据元件 – 数据产品” 的形态转变, 使得数据更有效地承载高价值信息, 推动由 “数据资源” 转化为 “数据资产”, 形成 “资产链条”. 从数据资源到数据元件的转化提升了数据品质, 提高了数据价值密度和标准化程度, 实现了第 1 层的数据增值. 从数据元件到数据产品的转化完成从标准化的数据元件到特定应用场景和专业化服务的适配, 实现了第 2 层的数据增值. 这样两层增值过程, 形成数据的 “价值链条”. 通过数据资源两次赋能, 打通数据资产链和数据价值链, 同步催生数据资源、数据元件和数据产品 3 类市场, 实现数据要素高效配置.

总的来说, 数据元件实现了数据资源与数据应用的解耦, 形成了数据有效保护层, 从而隔离了数据从资源端到应用端的泄漏风险以及从应用端到资源端的滥用风险. 这促进了数据的高效流通和安全配置, 解决了数据流通与安全之间的矛盾. 数据元件具有 “数据可用不可见、数据不动程序动” 的特点, 使得原始的数据资源在应用过程中不直接流向应用端, 从而隔离了数据泄露的风险; 同时, 数据应用端在使用过程中也不直接接触原始数据, 从而隔离了数据被滥用和被篡改的风险. 未来的政务数据要素化管理示意图如图 5 所示.

7　结论

在当前的数字经济时代, 数据已成为新的关键生产要素. 释放数据要素的价值、发挥数据在数字经济发展中的关键支撑作用, 已成为数字社会时代发展的必然规律, 也是我国新发展阶段的内在要求, 贯彻新发展理念的实践路径, 以及构建新发展格局的关键支撑. 为了实现数据要素的高效配置, 使数据 “供需两端” 真正贯通, 本文提出以数据元件作为连接数据供需两端的 “中间态”, 实现原始数据与

数据应用的 “解耦”, 解决 “安全与流通对立” 的难题. 作为支持数据要素流通和共享的关键概念, 数据元件需具备统一的内在信息和可计量的标准化模型. 本文的贡献可归纳如下.

(1) 本文明确给出了数据元件的理论定义, 并梳理了数据元件理论模型的四大特征. 通过与传统数据处理方法的对比, 为后续提出标准化的数学框架奠定了基础. 本文创新性地提出了数据元件的图论表征模型, 给出了数据元件的明确数学定义, 并从数学表征角度对比了数据元件与传统数据处理方法, 从数理理论角度对数据元件的优越性进行了解释和说明.

(2) 提出了数据元件的信息计量模型, 给出了数据元件一般性的关于信息熵的解析及计算方法. 同时, 推导了数据元件相比传统模型的额外熵减的一般表达式. 从理论上严格证明了数据元件具有更高的信息附加价值, 为数据元件的最终定价提供重要参考, 并为数据元件产业链条的经济利益分配提供理论指引.

(3) 基于信息计量模型, 评估了数据元件的信息价值. 并依据数据元件的信息价值提出了定价机制, 给出了数据元件基准价格的一般表达式. 同时, 通过案例分析展示了数据元件中各个信息计量相关因素如何影响数据元件的定价. 最后, 通过实例验证, 展示了如何确定数据元件加工每个步骤中所产生的经济利益增值.

由本文的研究可知, 数据元件的信息计量方法为数据交易双方提供了更为清晰的定价依据和谈判基础. 在传统的数据交易中, 由于数据价值的难以量化, 交易双方往往难以就价格达成共识, 导致交易效率低下. 而数据元件的信息计量方法则能够明确展示数据的价值构成和增值过程, 使得交易双方能够基于客观的数据信息进行价格谈判, 从而提高交易效率并降低交易成本.

补充材料　附录 A–C. 本文的补充材料见网络版 infocn.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

参考文献

1 Welch T F, Widita A. Big data in public transportation: a review of sources and methods. Transp Rev, 2019, 39: 795–818

2 Pan W, Xie T, Wang Z, et al. Digital economy: an innovation driver for total factor productivity. J Business Res, 2022, 139: 303–311

3 Gao L. Research on the development path of digital economy in China's central cities — from the perspective of policy. Dissertation for Master's Degree. Beijing: Central University of Finance and Economics, 2022 [高璐. 我国中心城市数字经济发展路径研究 —— 基于政策扩散视角. 硕士学位论文. 北京: 中央财经大学, 2022]

4 Cai Z, He Z. Trading private range counting over big IoT data. In: Proceedings of the 39th International Conference on Distributed Computing Systems (ICDCS), 2019. 144–153

5 Monino J L. Data value, big data analytics, and decision-making. J Knowl Econ, 2021, 12: 256–267

6 Zheng J H, Zhou N. Data-driven, digital transformation and the new development pattern. J Shandong Univ (Philos Soc Sci), 2023, 6: 93–105 [郑江淮, 周南. 数据要素驱动、数字化转型与新发展格局. 山东大学学报 (哲学社会科学版), 2023, 6: 93–105]

7 Lu Z P. Research on the engineering path of secure and trusted data space. Inform Commun Technol, 2023, 17: 49–55 [陆志鹏. 安全可信数据空间的工程化路径研究. 信息通信技术, 2023, 17: 49–55]

8 Frické M. The knowledge pyramid: the DIKW hierarchy. Ko Knowl Organ, 2019, 46: 33–46

9 McDowell K. Storytelling wisdom: story, information, and DIKW. Asso Info Sci Tech, 2021, 72: 1223–1233

10 Karkošková S. Data governance model to enhance data quality in financial institutions. Inform Syst Manag, 2023, 40: 90–110

11 Shkabatur J. Transparency with (out) accountability: open government in the United States. Yale Law Policy Rev, 2012, 31: 79

12 Stylos N, Zwiegelaar J, Buhalis D. Big data empowered agility for dynamic, volatile, and time-sensitive service industries: the case of tourism sector. Int J Contemp Hosp Manag, 2021, 33: 1015–1036

13 Wang D D, Dong J J. Quantitative evaluation of US and Chinese scientific data management policies based on the PMC-AE index model. J Modern Inform, 2023, 43: 111–121 [王丹丹, 董金金. 基于 PMC-AE 指数模型的中美科学

数据管理政策量化评价. 现代情报, 2023, 43: 111–121]
14 Price Ii W N, Cohen I G. Privacy in the age of medical big data. Nat Med, 2019, 25: 37–43
15 Jeske L, Placzek S, Schomburg I, et al. BRENDA in 2019: a european ELIXIR core data resource. Nucleic Acids Res, 2019, 47: D542–D549
16 Flewelling H A, Magnier E A, Chambers K C, et al. The pan-STARRS1 database and data products. Astrophys J Suppl Ser, 2020, 251: 7
17 Ivezić Ž, Kahn S M, Tyson J A, et al. LSST: from science drivers to reference design and anticipated data products. Astrophys J, 2019, 873: 111
18 Jones C I, Tonetti C. Nonrivalry and the economics of data. Am Economic Rev, 2020, 110: 2819–2858
19 Fan W, Geerts F. Foundations of Data Quality Management. Berlin: Springer, 2022
20 Bar-gill O. Smart disclosure: promise and perils. Behav Public Policy, 2021, 5: 238–251
21 Fisher H, Theodore K, Power P, et al. Routine evaluation in first episode psychosis services: feasibility and results from the MiData project. Soc Psychiat Epidemiol, 2008, 43: 960–967
22 Mandl K D, Gottlieb D, Mandel J C, et al. Push button population health: the SMART/HL7 FHIR bulk data access application programming interface. NPJ Digit Med, 2020, 3: 151
23 Erkmen A, Aydın M N. A comparison between right to data portability and United Kingdom's midata initiative. 2019
24 Brown I. The UK's Midata and Open Banking programmes: a case study of data portability and interoperability requirements. Technol Regul, 2022, 2022: 113–123
25 Frické M. The knowledge pyramid: the DIKW hierarchy. Ko Knowl Organ, 2019, 46: 33–46
26 Omran M, Tahat Y A. Does institutional ownership affect the value relevance of accounting information? Int J Account Inform Manag, 2020, 28: 323–342
27 Yager R R. Using fuzzy measures for modeling human perception of uncertainty in artificial intelligence. Eng Appl Artif Intell, 2020, 87: 103228
28 Giri G, Manohar H L. Factors influencing the acceptance of private and public blockchain-based collaboration among supply chain practitioners: a parallel mediation model. Supply Chain Manag, 2023, 28: 1–24
29 Li X, Yao J, Liu X, et al. A first look at information entropy-based data pricing. In: Proceedings of the 37th International Conference on Distributed Computing Systems (ICDCS), 2017. 2053–2060
30 Liang F, Yu W, An D, et al. A survey on big data market: pricing, trading and protection. IEEE Access, 2018, 6: 15132–15154
31 Xu C, Zhu K, Yi C, et al. Data pricing for blockchain-based car sharing: a stackelberg game approach. In: Proceedings of IEEE Global Communications Conference, 2020. 1–5
32 Chuang I H, Huang S H, Chao W C, et al. TIDES: a trust-aware IoT data economic system with blockchain-enabled multi-access edge computing. IEEE Access, 2020, 8: 85839–85855
33 Luo Z, Yang S, Chen Y Q, et al. An auction-based pricing model for big data trading. In: Proceedings of the 7th International Conference on Information Science and Control Engineering (ICISCE), 2020. 208–212
34 Tang Z, Lv Z, Wu C. A brief survey of data pricing for machine learning. In: Proceedings of the 8th International Conference on Signal, Image Processing and Pattern Recognition, 2020
35 Ye Y, Zhang Y, Liu G, et al. A measure based pricing framework for data products. Web Intell, 2020, 18: 249–260
36 Shen Y, Guo B, Shen Y, et al. Personal big data pricing method based on differential privacy. Comput Secur, 2022, 113: 102529
37 Cong Z, Luo X, Pei J, et al. Data pricing in machine learning pipelines. Knowl Inf Syst, 2022, 64: 1417–1455
38 Allouah A, Bahamou A, Besbes O. Pricing with samples. Oper Res, 2022, 70: 1088–1104
39 Miao X, Peng H, Huang X, et al. Modern data pricing models: taxonomy and comprehensive survey. 2023. ArXiv:2306.04945
40 Huang J, Kauffman R J, Ma D. Pricing strategy for cloud computing: a damaged services perspective. Decision Support Syst, 2015, 78: 80–92
41 Mei L, Li W, Nie K. Pricing decision analysis for information services of the Internet of Things based on Stackelberg game. In: Proceedings of the 2nd International Conference on Logistics, Informatics and Service Science, 2013. 1097–1104
42 Li C, Li D Y, Miklau G, et al. A theory of pricing private data. ACM Trans Database Syst, 2014, 39: 1–28
43 Yu H, Zhang M. Data pricing strategy based on data quality. Comput Indust Eng, 2017, 112: 1–10
44 Lu Z P, Meng Q G, Wang Y. Element-based Governance of Data. Beijing: Tsinghua University Press, 2024 [陆志鹏, 孟庆国, 王钺. 数据要素化治理. 北京: 清华大学出版社, 2024]

45 Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fusion, 2021, 76: 243–297
46 Septian A, Darhim A, Prabawanto S. Mathematical representation ability through geogebra-assisted project-based learning models. J Phys-Conf Ser, 2020, 1657: 012019
47 Gupta M K, Chandra P. A comprehensive survey of data mining. Int J Inf Tecnol, 2020, 12: 1243–1257
48 Zhou L, Fu A, Yang G, et al. Efficient certificateless multi-copy integrity auditing scheme supporting data dynamics. IEEE Trans Depend Secure Comput, 2020, 19: 1118–1132

Information metrics for data components based on information entropy: data pricing and its application analysis for electric energy statistics

Xiaoming TAO[1,2], Jieyang PENG[1], Yue WANG[1], Youzheng WANG[1], Chengsheng HU[3] & Zhipeng LU[3*]

1. *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
2. *School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China*
3. *China Electronics Corporation, Beijing 100190, China*
* Corresponding author. E-mail: luzhipeng@cecdt.com.cn

Abstract In the era of the digital economy, data has seen its value assessment and pricing remain a research hotspot as a critical factor of production. Currently, research on information value assessment and pricing primarily focuses on data quality, demand, and market policies, with a notable lack of data-driven quantitative studies. This paper explores pricing for data components based on information entropy and proposes a corresponding pricing mechanism. As data's pivotal role in the digital economy becomes increasingly evident, data components, which serve as an "intermediate state" connecting data supply and demand, are of great significance. First, the paper reviews the current research on information value assessment and pricing, finding that this assessment poses challenges to both economics and data science, and elaborates on the importance of establishing an improved pricing structure. Second, it defines corresponding mathematical models for data components and constructs, emphasizing their role in resource development, measurement, and pricing. Furthermore, through a measurement and pricing model based on data components, it investigates the intrinsic information within data components and the differences in their information content compared to traditional information communication methods. Finally, this paper includes an empirical study on the pricing of electricity data to validate the effectiveness of the proposed method. It underscores the significance of data components in achieving the efficient allocation of data elements and outlines future directions in their development. This paper will provide novel insights into achieving the efficient allocation of data elements, offering theoretical support and practical pathways for the development of the digital economy as well as the circulation and sharing of data elements.

Keywords data component, data science, information theory, data elements, data governance, electricity pricing