软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(3):1396−1424 [doi: 10.13328/j.cnki.jos.006751]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

数据定价与交易研究综述*

江 东[1], 袁 野[2], 张小伟[1], 王国仁[2]

[1](东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

[2](北京理工大学 计算机学院, 北京 100081)

通信作者: 袁野, E-mail: yuanye@mail.neu.edu.cn

摘 要: 在大数据时代, 随着信息技术的发展, 各行各业都在收集海量数据. 数据是数字经济的基础, 蕴含有巨大价值. 但是由于缺乏高效可行的共享机制, 数据拥有方彼此之间缺乏沟通, 形成了一个个数据孤岛. 这不利于大数据产业的健康发展. 因此, 给数据分配一个合适的价格, 设计高效的数据交易市场平台成为消除数据孤岛、使数据充分流动的重要途径. 系统梳理进行数据定价与交易时涉及的技术性问题. 具体来说, 介绍数据定价与交易的难点和相关准则; 将大数据在市场中的生命周期分为数据收集与集成、数据管理与分析、数据定价和数据交易 4 个环节; 在大数据管理研究的基础上介绍适用于前两个环节的相关方法; 然后对数据定价思路和方法进行分类, 分析各类方法的适用场景以及优势和短板; 介绍数据市场的分类, 以博弈论和拍卖为例研究了数据交易中市场类型和参与人行为对交易过程及价格的影响. 最后, 对数据定价与交易的未来研究方向进行展望.

关键词: 数据定价; 数据交易; 数据市场; 定价模型; 数据管理

中图法分类号: TP311

中文引用格式: 江东, 袁野, 张小伟, 王国仁. 数据定价与交易研究综述. 软件学报, 2023, 34(3): 1396–1424. http://www.jos.org.cn/1000-9825/6751.htm

英文引用格式: Jiang D, Yuan Y, Zhang XW, Wang GR. Survey on Data Pricing and Trading Research. Ruan Jian Xue Bao/Journal of Software, 2023, 34(3): 1396–1424 (in Chinese). http://www.jos.org.cn/1000-9825/6751.htm

Survey on Data Pricing and Trading Research

JIANG Dong[1], YUAN Ye[2], ZHANG Xiao-Wei[1], WANG Guo-Ren[2]

[1](School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China)

[2](School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: In the big data era, an enormous amount of data is collected in every industry with the development of information technology. Data is the foundation of the digital economy, containing great value. However, for the lack of efficient and feasible data-sharing mechanisms, data owners seldom communicate with each other, which leads to the formation of data islands and is unfavorable to the healthy development of the big data industry. Hence, allocating a proper price to data and designing an efficient data market platform have become important ways to eliminate data islands and secure sufficient data flow. This study systematically sorts out the technical issues regarding data pricing and trading. Specifically, the difficulties and related principles of data pricing and trading are introduced. The life cycle of data in the data market is divided into four stages: data collection and integration, data management and analysis, data pricing, and data trading. Upon the research on big data management, related methods applicable to the first two stages are elaborated. After that, data pricing methods are categorized, and usage scenarios, advantages, and shortcomings of these methods are analyzed. Moreover, the classification of data markets is introduced, and the impact of market types and participants' behavior in data trading on the trading process and prices is studied with game theory and auctions as examples. Finally, future research directions of data pricing and trading are discussed.

* 基金项目: 国家自然科学基金 (61932004, 61732003, 62072087, U2001211); 中央高校基本科研基金 (N181605012)
收稿时间: 2021-11-01; 修改时间: 2022-05-09; 采用时间: 2022-07-26; jos 在线出版时间: 2022-10-27
CNKI 网络首发时间: 2022-12-29

Key words: data pricing; data trading; data market; pricing model; data management

随着互联网、移动设备、工业传感器等技术的日益发展, 全球数据规模日益增大. 根据中国信息通信研究院发布的《大数据白皮书 (2020 年)》[1]的预测, 2030 年全世界数据产生量将达到 612 ZB, 2035 年将达到 2 142 ZB. 各种各样不同来源的数据被收集整理, 存储在各个数据中心或厂商的终端设备中, 便形成了如今学界和业界的重点研究方向——大数据. 关于大数据的定义, 不同文章从各自角度出发给出了不同定义方式. 通常来说, 各种不同来源的数据汇集在一起形成体量巨大、可以创造出巨额商业价值的数据, 称为大数据. 大数据有着 4V 特性[2]: 第一, 数据体量巨大 (volume), 从最开始的 TB 级别, 跃升到后来的 PB, 再到现在的 ZB; 第二, 数据变化快速 (velocity), 这个变化速度不仅仅指数据生成速度, 还包括实时在线的要求, 其次, 数据处理、传送和存储的速度也是极高的; 第三, 数据种类多样 (variety), 由于数据来源不同, 各行各业收集到的数据也是多种多样的, 除了传统的结构化数据, 还有图、XML、视频等半结构化和非结构化数据; 第四, 也是大数据最重要的一个特性, 巨大的价值 (value), 即商业价值高、价值密度低, 在数据的海洋里往往需要通过不断寻找才能发现其中巨大的商业价值.

数据是新石油[3]. 因此, 像石油一样, 大数据对于国家和组织来说具有巨大价值. 但只有具有专业技能和积累的公司和组织才具备“开采”出“新石油”的资格. 想要发掘出大数据的价值, 不仅需要拥有数据分析能力, 还要具备数据发现和采集、数据整合的能力. 随着数据不断产生, 互联网公司、移动设备运营商、智慧工厂等一些为数据产生源头提供服务的公司或组织囤积了大量数据, 但是他们其中很大一部分并不具备发掘出数据全部价值的能力. 因此数据共享成为打破数据孤岛、创造良好数据流通生态的重要手段. 数据共享可以分为开放共享和有偿共享两种方式. 开放共享的难点在于拥有数据的公司或组织没有足够动力去主动分享数据, 特别是涉及核心用户习惯、蕴含巨大商业价值的数据. 即使是通过开放数据接口[4,5]、数据湖[6]等一些方式得到的共享数据, 往往也需要消费者进行二次整理才能变成其想要的数据形式. 另一种共享方式是有偿共享, 即把数据推向市场, 通过数据交易的方式共享数据. 数据交易可以建立起一套数据共享规则, 使买卖数据双方都能够在该规则运行下得到自己想要的结果. 在数据交易市场中, 拥有数据的组织和公司可以通过交易得到相应回报, 称为激励; 而数据购买者也在其数据分析任务的驱使下, 到数据交易市场购买数据. 随着近些年来大数据产业的发展, 我国已经成立了包括贵阳大数据交易所[7]、武汉东湖大数据交易中心[8]、北京国际大数据交易所[9]在内的多家大数据交易中心. 同时, 在国际上也出现了一大批数据交易平台, 如 Dawex[10]、Xignite[11]、WorldQuant[12]等. 因此, 通过数据在有买卖意愿的各方进行交易, 不仅打破了数据孤岛, 还可以让数据的价值被充分挖掘, 帮助企业、政府做出决策, 助力下一轮科技革命.

一般情况下, 数据交易市场的参与方包括数据拥有者、数据消费者和数据平台[13]. 数据拥有者通常是能够收集到数据的公司或组织, 也包括少数可以合法出售自己隐私数据的个人; 数据消费者是出现在数据市场上, 希望购买数据以解决自己使用需求的买家; 数据平台是数据交易的中间人, 也称为数据中介, 经纪人等, 负责对数据进行收集整合、设定收购和出售价格、为数据拥有者和数据消费者提供相关服务等.

虽然数据交易如今已经获得了足够多的重视, 但由于是新兴交易类型, 交易的商品又是与传统商品和数字信息商品性质存在较多不同的数据商品, 因此现在数据交易市场仍处在起步阶段. 为了建立高效的数据交易市场, 需要解决如下的问题. 第 1 个问题是数据交易前的准备工作, 包括数据收集、整合、分析等操作以体现出数据的价值. 由于消费者需要的数据类型是千变万化的, 因此在数据交易市场中流通的数据类型也是多种多样的. 这导致数据平台必须考虑如何从不同数据源头收集到不同类型的数据, 如何对这些数据进行整合以方便存储和出售, 如何对数据进行简单分析以确定其基本价值, 从而为定价和交易奠定基础. 第 2 个问题是如何确定数据的价格. 这是在进行数据交易前首先要考虑的问题. 由于数据消费者希望数据价格可以反映该数据对其任务的价值, 而数据拥有者和数据平台大多希望以数据收集、管理成本作为数据价格. 这种割裂导致进行交易的各方很难达成一致. 因此, 如何为交易的数据设计一个合适的价格, 使得拥有数据的公司和组织有着更高的意愿卖出数据, 同时又保证了消费者的经济利益, 满足其完成相应任务的需求, 是一个极具挑战性的问题. 最后, 是如何设计数据交易机制的问题. 数据交易和数据定价是互补的关系. 数据定价关注数据出售价格的设定, 而数据交易则更关注市场. 由于每个市场

参与人的行为、市场结构都会对数据交易过程产生影响, 因此如何设计交易机制, 构建真实可信的交易平台, 以保证参与交易的双方收益可以最大化, 确保数据交易能够公平、高效地进行, 也是一个值得深入研究的问题.

对于上述问题, 已经存在对数据定价与交易相关工作的综述, 但角度各有不同. Pei 等人[14]从经济学的角度对数据定价进行了完整的叙述, 并总结了数据定价时需要考虑的基本内容以及应该遵循的准则, 基于这些准则, 介绍了相应的数据定价方法. 类似地, 张小伟等人[13]对经济学中适用于数据定价的理论和方法进行了综述. 刘枌等人[15]以社会科学的视角介绍了大数据定价方法, 将其分为成本导向、市场导向、需求导向、利润导向以及基于生命周期的定价五种类型. 蔡莉等人[16]也对数据定价模型进行了综述, 将其分为基于数据质量的定价、基于信息熵的定价、基于查询的定价、基于博弈论的定价和基于机器学习的定价, 并对上述几种定价方法的优劣进行了分析. 上述文献各自存在不足之处. 文献 [13,14] 重点关注定价过程中涉及的经济学准则和方法, 没能对现存定价方法进行完整分类. 刘枌等人[15]虽然对数据定价方法进行了分类, 但是更多侧重于制度性和框架性的叙述, 没能够对定价方法的具体细节进行研究. 蔡莉等人[16]弥补了上述不足, 对数据定价策略和方法进行了详细分类, 对数据定价过程的介绍也较为全面. 但是忽略了与数据定价密不可分的数据交易部分, 因此, 除了对数据定价过程中需要遵循的准则以及数据定价方法进行全面综述外, 本文还将数据交易市场作为重点, 根据大数据在数据交易市场上的流通过程, 将其生命周期分为数据收集与集成、数据管理与分析、数据定价和数据交易 4 个环节, 详细介绍了每个环节需要进行的工作、存在的挑战以及相关解决方案.

在本文中, 我们对大数据定价与交易进行了全面综述, 以帮助大家对这一方向有完整的了解. 本文贡献如下.

1) 本文首先回顾了大数据定价与交易的相关研究工作, 在考虑大数据特点的基础上, 总结了每篇数据定价与交易相关文章中的侧重点, 并基于此介绍了数据定价与交易中存在的一些挑战和难点, 描述了需要遵循的一般性准则, 解释了这些准则的重要性.

2) 本文将大数据在数据交易市场中的生命周期分为了 4 个环节, 分别为数据收集与集成、数据管理与分析、数据定价、数据交易. 由于前两个环节在数据定价与交易相关文章中未受到足够的重视, 因此借鉴了大数据管理方向的相关文章, 总结了其中适用于数据交易市场的方法.

3) 本文对以往的数据定价相关研究工作进行了总结, 对流行的数据定价思路和方法进行了分类, 比较了每种方法的优势和局限性.

4) 本文研究了数据交易过程, 对数据市场结构进行了分类. 以博弈论和拍卖为例, 研究了数据定价方法中没有涉及的市场类型和数据市场参与人行为对数据交易过程产生的影响, 并详细介绍了博弈论和拍卖方法在数据交易场景中的分类及其应用, 总结了每个类别的优劣.

本文第 1 节介绍数据定价与交易的挑战和需要遵循的相关准则, 并简单介绍了大数据在数据交易市场中的生命周期. 第 2 节介绍数据交易平台需要完成的数据收集与集成工作. 在此之后, 第 3 节介绍数据交易市场涉及的数据管理与分析任务关键性问题. 第 4 节为数据定价方法进行了分类, 介绍了技术细节, 总结了每个方法的适用场景以及优劣之处. 第 5 节介绍了数据市场分类, 以博弈论和拍卖为例研究了市场结构和数据交易中参与人行为对数据价格的影响. 第 6 节对相关工作进行了介绍. 第 7 节总结全文, 并对未来的研究进行了展望.

1 数据定价与交易的难点和准则

1.1 数据定价与交易的难点

大数据具有 4V 特性, 即数据体量巨大 (volume)、数据变化快速 (velocity)、数据类型多样 (variety) 和数据价值巨大 (value)[2]. 数据体量巨大是指数据规模通常可以达到 PB 甚至 EB 的级别; 数据变化快速是指大数据的生成和更新有着极高的时效性, 对数据的存储、处理也有高速的要求; 数据类型多样是指大数据来源丰富, 各行各业所产生的数据都以多种形式汇集, 成为大数据的组成部分; 数据价值巨大是指数据商业价值高, 但是价值密度较低, 需要经过仔细筛选挖掘才能找到有价值的内容.

基于大数据以上的特性, 本文认为, 同传统资产相比, 在数据市场上交易的数据有着以下特性.

1) 多样性. 在大数据时代, 数据来源是多样的, 除了互联网作为我们最为熟知的数据产生源, 医疗设备、视频监控设备、物联网设备、工厂自动化设备、移动通信设备等都是数据产品的重要来源. 数据产品来源的多样化导致了数据产品形式的多样化. 从文本、音频、视频等各种非结构化数据, 再到半结构化、结构化数据.

2) 时效性. 随着云计算技术的发展和线上应用的推广, 大数据的产生和更新往往具有极高的时效性. 各种应用、设备随时随地产生数据, 这些数据也需要及时进行处理以投入使用. 大数据的时效性同时也是很衡量数据产品价值的重要指标, 过时数据对于买家往往有着较低的吸引力.

3) 可重复性. 同数字资产一样, 由于存储方式的特殊, 导致数据产品有着极低的复制代价, 同时, 被复制的数据还能保证其原有属性不变. 从另一方面来说, 数据产品在使用过程中也不会产生损耗和折旧, 可以重复利用.

4) 价值稀疏性. 在数据产品的价值方面, 刘朝阳等人[17]认为大数据价值具有稀疏性, 具体表现在价值的不确定性、价值的稀缺性和价值的多样性 3 个方面.

对于数据产品的交易, 难点来自数据产品本身的特性. 通常来说, 数据定价的难点在于数据来源的多样性以及自身结构的复杂. 数据产品的多样性导致需要对不同类型的数据设计不同的定价方法, 这些方法有着各自的出发点, 这导致很难保证定价结果的客观性, 从而影响交易. 其次, 数据的多样性增大了数据平台存储数据的花费, 不利于统一管理.

数据产品的时效性也是导致大数据定价与交易存在难点的问题之一. 文献 [18] 分析认为, 与传统产品不同的是, 数据产品的定价是具有时间依赖性的, 实时产生的数据在一段时间之后对于购买者就不再重要, 因此带有时间贴现 (time discounting) 的数据定价模型是连续时间动态规划问题, 有着极大挑战. 同时, 维持数据产品更新需要大量时间和处理代价, 这就要求一部分定价策略需要随着数据产品的更新进行实时调整, 这给数据定价方法提出了更高的设计要求.

数据产品的可重复性要求在数据定价时考虑隐私泄露问题, 同时, 由于复制代价极低, 数据买家在获得数据后可以将数据重新打包卖出, 这对数据售出的公平性产生了影响, 同时会降低卖家出售数据的积极性, 因此数据定价方法应该要考虑到隐私保护和版权保护的机制.

在数据价值方面, 价值是定价的基础. 刘朝阳等人[17]认为大数据的价值具有双向不确定性, 即在数据产品交易中, 买卖双方对产品价值很难达成一致[19]. 由于数据消费者希望数据价格可以反映该数据对其任务的价值, 而数据拥有者和数据平台大多希望以数据收集、管理成本作为数据价格. 因此, 难以实现广泛的大数据价值认同是当前数据定价面临的最突出的问题[20].

1.2 数据定价与交易的准则

除了来自数据本身特性所造成的一些难点, 在进行数据定价和交易时, 为了让交易顺利进行, 确保参与人能得到更高收益等目的, 参与数据交易的三方需要遵循一些准则, 包括真实性、公平性、无套利、收入最大化和个人理性.

1.2.1 真实性

在数据交易进行时, 往往要求买卖双方是真实的. 即买卖双方都是利己的, 并且仅提供能够使得自己利益最大化的价格. 如果一个数据市场满足上述条件, 则称该数据市场是真实的 (truthful). 换句话说, 在一个真实的市场中, 如果买家认定了某个产品的购买价格, 他就不会再支付多于该价格的金钱去购买此产品[14]. Cai 等人[21]认为真实性的定义应该为: 无论其他人如何操作, 对于任意的供应商和消费者来说, 都不能通过虚报真实价值来增加其收益. 简单来说, 真实性保证了每个人在参与交易时都不进行虚假操作. 许多交易和定价方法的设计都是在真实性的前提下进行的. An 等人[22]提出了基于真实拍卖的大数据交易模型, 认为一个真实的拍卖模型必须具有激励能力, 即竞标者必须真实地报告自己的价格, 才能获得最大收益. 并给出了真实性的形式化定义: 数据市场中的任意买家 i, 如果存在 $U_i \geqslant U_i'$, 那么该定价模型满足真实性. 其中 U_i 表示买家 i 通过真实报告获得的收益, U_i' 表示买家 i 通过虚假报告获得的收益. 类似地, Jiao 等人[23]认为真实性可以防止数据交易时的猜测行为, 并减少市场竞标策略中不必要的开销.

1.2.2 公平性

在当前数据交易模式下, 数据通常来自不同的数据卖家. 为了确保卖家出售数据的积极性, 数据交易平台需要保证总收入在所有数据贡献者中按其贡献公平分布, 称为公平性 (fairness). Xiong 等人[24]给出了公平性的一般定义: 从用户角度来看, 公平性表示所有用户的收入在一段时间以内以公平的方式分配, 是衡量数据市场价格问题的重要方法. 数据定价的公平性在多种定价模型中均有涉及. Khokhar 等人[25]提出了基于信息熵的定价算法, 在数据交易的验证阶段, 通过限制不诚实的数据拥有者再次进入市场参与交易从而保证了收入分配的公平性. 但是文章没能明确给出在诚实的数据拥有者之间分配收入的方法. Delgado-Segura 等人[26]提出了一个公平交易市场, 提供了基于比特币的公平交易协议, 交易过程可以随时结束或终止, 以确保供应商和消费者都没有损失. 该方法缺点在于, 主要依靠比特币来保证公平交易, 并不能完全实现收入的公平分配. 上述两个实现公平性的方法有着应用范围窄, 普适性差的缺点. Koutris 等人[27]提出了 QueryMarket, 一个基于查询的数据定价系统, 并在其中引入了 FairShare 策略, 保证收入在数据卖家中公平分配. 在 FairShare 策略下的 k 个卖家中, 单个卖家 s_i 的收入计算方法如下:

$$rev(s_i, Q) = \frac{share(s_i, Q)}{\sum_{j=1,k} share(s_j, Q)} \cdot p(Q) \tag{1}$$

其中, $share(s_i, Q)$ 表示卖家在所有最低成本解决方案 Q 中可以获得的最大收入, $p(Q)$ 表示 Q 的价格. 该策略是关系数据库数据交易中较为常用的卖家收入计算方法. 一个适用范围更加广泛的方法是基于博弈论中著名的沙普利值法 (Shapley value)[28], 可以满足数据市场中收入公平分配的要求, 其计算方法如下:

$$s_i = \frac{1}{N} \sum_{S \subseteq \Delta \setminus z_i} \frac{1}{\left(\frac{N-1}{|S|} \right)} [v(S \cup \{z_i\}) - v(S)] \tag{2}$$

其中, N 代表博弈参与者, S 表示参与者组成的任意联盟, $v(S)$ 表示联盟 S 的收益函数, z_i 表示联盟 S 中的某个参与者. 沙普利的实现满足如下要求.

(1) 集体理性 (group rationality): 交易获得的收入必须全部分配给所有卖家.

(2) 公平性 (fairness): 对于一个卖家联盟 S 和另外两个卖家 s 和 s', $s, s' \notin S$, 若 $S \cup \{s\}$ 和 $S \cup \{s'\}$ 获得了相同的资金, 那么 s 和 s' 也应该收到相同的回报. 即, 对于效用的贡献度相同的卖家, 他们所收到的回报也应该相同; 对于一个卖家联盟 S 和一个额外的卖家 $s \notin S$, 若 $S \cup \{s\}$ 和 S 获得了相同的资金, 则 s 收到的回报为 0. 即, 没有贡献就没有回报.

(3) 可加性 (additivity): 如果分别为两个任务 T_1 和 T_2 回报 v_1 和 v_2, 那么完成两个任务 T_1+T_2 的回报是 v_1+v_2.

Jia 等人[29]就融合了基于模型定价和基于查询定价的特点, 提出了针对 kNN 模型的定价方法, 并在其中使用了沙普利值法来保证收入是公平分配的. 但是, 由于精确计算沙普利值的时间复杂度是指数级的, 因此通常情况下人们愿意使用近似算法来计算该值. 同时, 由于上文提到的数据产品复制代价极低、对于大部分数据定价场景来说效用函数难以计算等问题, 使用沙普利值思想实现收入分配的公平性依旧具有极大挑战.

1.2.3 无套利

套利是大数据定价中最需要关注的问题之一, 是指买家通过某种手段, 按照低于卖家规定价格获取数据产品的行为. 套利机会的存在会导致数据定价的不一致性, 并使得信息泄露的风险大大增加. Balazinska 等人[30]列举了构建云数据市场面临的挑战, 其中最重要的挑战之一就是要求定价函数必须满足无套利 (arbitrage free). Koutris 等人在文献 [31] 中提出了基于查询的定价, 研究了其中套利的问题. 在买家希望购买查询束 $Q = Q_1 + Q_2$ 的情况下, 如果买家分别创建两个账号去购买查询 Q_1 和 Q_2, 为了防止套利, 平台应该保证查询 Q 的价格至多不能大于查询 Q_1 和 Q_2 价格之和. Li 等人[32]提出了数据市场中聚合查询的定价方法, 认为由于数据产品具有可重复性的特点, 所以买家可以将购买到的查询进行结合或再处理, 从而可能得到未购买的查询结果. 文章将此类型的无套利定义为: 如果查询束 Q_1 得到的信息是查询束 Q_2 得到信息的子集, 那么 Q_1 的价格必须低于 Q_2 的价格. Li 等人[33]则解决了

有噪声的线性查询问题. 卖家将数据贡献给平台, 平台会给买家返回加有噪声的查询结果, 以此保护个人隐私. 但是文章同时提出了该方法的难点: 当平台向查询结果内添加的噪声方差较大时, 返回结果的精确度也会存在接近 1 的情况. 这就导致如果平台给具有较大噪声方差的结果设置一个较低价格, 会出现套利情况. Lin 等人[34]将套利行为进行了总结, 共分为 5 类.

① 基于价格的套利: 买家可能会通过重复咨询价格来推测除了查询价格以外的信息, 例如元组是否包含在两个表的连接中, 甚至可以得到一个表中的所有内容;

② 多账户套利: 对于查询束 $Q=Q_1+Q_2$, 买家分别创建两个账号去购买查询 Q_1 和 Q_2, 为了防止多账户套利, 平台应该保证查询 Q 的价格至多不能大于查询 Q_1 和 Q_2 价格之和, 即文献 [31] 中所提到的套利形式;

③ 后处理套利: 如果查询束 Q_1 得到信息是查询束 Q_2 得到信息的子集, 那么 Q_1 价格必须低于 Q_2 价格, 即文献 [32] 中所提到的套利形式;

④ 偶然套利: 如果买家希望随机购买符合要求的任意一条数据, 如果随机查询的价格低于确定查询某条记录的价格时, 会发生偶然套利;

⑤ 确定套利: 在平台给买家返回有噪声的查询结果并以噪声程度确定查询价格的模式下, 当平台向查询结果内添加的噪声具有较大方差时, 返回结果的精确度也会存在接近 1 的情况, 如果平台给具有较大方差的结果设置较低价格, 则会出现套利情况; 即文献 [33] 中提到的套利形式.

文献 [34] 对上述 5 种套利形式, 分别给出了相应的解决方案, 并将其整合到一个框架中. 该框架允许查询随机化, 并提出了两个可以解决上述所有套利形式的潜在定价函数. Deep 等人[35]介绍了两种套利形式, 称为信息套利和捆绑套利, 分别对应文献 [34] 中的②和③, 描述了无套利定价函数的机制. 并在文献 [36] 中设计了可扩展的、更适用于关系数据的定价框架.

1.2.4 收入最大化

收入最大化 (revenue maximization), 也称利润最大化 (profit maximization) 是在传统商品中已经被充分研究的问题. 对于卖家或经纪人来说, 较低的出售价格可以吸引更多的买家, 而较高的出售价格可以使自己得到更多的收入. 那么如何在二者之间进行平衡? 瓦尔拉斯均衡 (Walrasian equilibrium) 就解决了在完全竞争市场中如何保证卖家收入最大化问题[37]. Myerson[38]在 1981 年也提出了经典的单物品拍卖中的收入最大化方法. 传统产品与数据产品在收入最大化的计算上具有较大不同. 在竞争市场中, 传统产品在边际成本等于边际收益时达到卖家收入最大化. 然而由于数据产品边际成本几乎为 0, 因此该规则不适用于数据产品. 此外, 由于现代数据交易模式下, 与买家直接接触的往往是经纪人, 因此收购数据的经纪人并不一定清楚数据的具体用途, 因此很难给数据产品标定一个可以使自己收入最大化的价格[39]. 因此, 在数据产品上实现收入最大化需要进行专门研究.

通常情况下, 将数据交易市场中收入最大化看作最优化问题. 由于精确解决该最优化问题需要花费的时间复杂度过高, 研究者们都普遍在寻找简单精准的近似算法. 近来研究结果表明, 使用贝叶斯优化机制可以在单买家和多买家的拍卖中实现常数近似[40]和对数近似[41]. 为了解决经纪人收入最大化问题, 文献 [39] 将收入最大化问题转化为遗憾最小化 (regret minimization) 问题, 提出了基于上下文的动态定价算法, 实现了经纪人收入最大化. 类似地, Chawla 等人在文献 [42] 中将收入最大化问题限定在专一买家、无限供给条件下, 基于此提出了启发式方法, 满足单调性和次可加性. 文章对比了 3 种精简的定价函数, 并研究了其相应的收入最大化问题, 在保证无套利的同时最大化了经纪人的收益. 上述文章对收入最大化的研究都局限在基于查询的定价模式上, 交易数据仅限于关系数据库中的数据集. 随着人工智能技术的发展, 机器学习模型的训练数据也逐渐成为炙手可热的资源. Agarwal 等人[43]设计了一个交易机器学习训练数据的市场, 基于 Myerson 拍卖理论[38]提出了组合数据产品的拍卖方法, 除了满足上文所介绍的真实性、无套利之外, 还实现了卖家收入的最大化. 可以看出, 收入最大化思想最初是由传统交易方式如拍卖引入, 随着数据定价技术的发展, 逐渐在基于查询和模型的定价中也有所应用, 并结合相应特点提出了不同的实现方式[31,44].

1.2.5 个人理性

无论是在传统产品的交易还是数据产品的交易中, 个人理性都是机制或算法设计者需要假设的前提之一.

Ghosh 等人在文献 [45] 中提出了基于拍卖的隐私数据交易方法, 设计了满足真实性和个人理性的定价机制, 并将个人理性定义为: 买卖双方通过积极参与并向该机制真实地报告自己产品的价值, 他们就可以得到非负的收益. 另一篇由 Cai 等人[21]提出的双边拍卖数据定价机制也给出了类似的个人理性的定义: 买卖双方通过数据交易能够得到的收益都是非负的. 与真实性类似, 个人理性的保证也大多出现在拍卖中.

1.3 数据交易的生命周期

与传统商品交易需要市场类似, 要进行数据定价和交易, 就离不开数据交易市场. 但是, 数据作为一种特殊的虚拟商品, 有其自身的特别之处. 因此, 想要数据交易过程顺利进行, 设计公平合理的数据交易市场至关重要. 迄今为止尚未有数据交易市场的明确定义及统一模型, 因此本文以数据市场的角度出发, 结合大数据在数据交易市场中的生命周期, 对市场所需满足的条件及各个组成部分进行详尽介绍.

通常情况下, 数据市场的结构如图 1 所示.

图 1 数据交易市场结构

数据交易市场包括 3 个组成部分: 数据拥有者, 或者称为数据卖家、数据提供者, 负责向数据平台提供数据, 并接受数据平台给予的相应补偿; 数据平台也称为数据中间商、数据中介、经纪人等, 负责对收购到的数据进行集成整合, 设定数据收购价格并补偿数据拥有者, 设定数据出售价格, 为数据消费者提供查询其希望购买数据的接口和服务, 给数据消费者提供数据并对出售的数据提供隐私、版权保护等任务; 数据消费者又称为数据买家, 在数据交易中需要完成的任务是向数据平台提出需求, 并支付一定金钱从数据平台购买到自己所需的数据.

在进行数据交易时, 数据拥有者负责向数据交易平台提交数据以及与隐私等方面的要求; 数据消费者则需要向数据交易平台提交自己的数据购买需求; Fernandez 等人[46]认为, 除了执行数据定价与交易外, 数据交易平台相关的准备工作包括: 数据发现、数据集成、数据融合和事实发现. 其中, 数据发现是为了从现有数据源收集到的数据中挖掘出其具有的商业价值; 数据集成则是为了将收集到具有商业价值的数据进行整理、清洗和验证, 以便其满足数据平台的存储需求; 数据融合和事实发现是为了对多个来源的数据集进行融合, 为数据消费者提供最终的查询结果. 结合上述内容, 本文将数据在数据交易市场中的生命周期分为: 数据收集与集成、数据管理与分析、数据定价和数据交易 4 个部分. 数据收集与集成解决数据“从无到有”的问题, 并对源数据执行整合、清洗和验证等操作, 以便满足后续数据管理要求以及数据消费者的数据查询要求; 数据管理与分析是为了解决数据组织存储形式的问题, 同时对数据进行分析以得到其适用范围、出售模式和近似商业价值; 数据定价关注各种确定数据价格的方法; 而数据交易则重点考虑了数据市场类型、参与交易各方行为等对数据出售价格的影响. 接下来将对其进行逐个介绍.

2 数据收集与集成

2.1 数据收集

数据收集是数据定价与交易中最为基础的一个阶段. 随着信息技术的日益发展, 各行各业的设备无时无刻不在产生大量不同种类的数据. 使用合适的方法收集这些数据是消灭数据孤岛, 促进数据共享和交流, 实现数据交易常态化的重要保证. 本文所述的数据收集将从数据平台角度出发, 研究数据平台是如何获取大量数据以便售卖的. 从数据源头看, 数据拥有者主要包括个人数据拥有者和产生数据的公司、企业或者团体, 我们将其简称为集体数据拥有者. 个人数据拥有者出售数据的场景主要为个人隐私数据的出售[33,45,47]. 在该场景下, 数据拥有者通过出售

隐私数据获取相应补偿, 并通过控制隐私暴露程度来决定他们可以得到的补偿价格是多少. 集体数据拥有者由于其给用户提供服务或拥有数据产生设备的便利 (如互联网公司和通信供应商、大规模使用物联网设备的工厂等), 通常可以很方便地获取到大量数据. 上述两类数据拥有者都有着将数据出售给数据平台的动力. 个人数据拥有者一般将数据出售给平台以换取相应收入; 而对于集体数据拥有者来说, 大数据是下一代生产力解决方案的基础, 向数据平台出售数据, 除了可以获取收益以外, 还可以使得整个行业形成良好的数据交易氛围, 因而在提高服务、提高生产力和最大化数据价值方面, 集体数据拥有者有着强烈愿望将数据出售给数据平台.

由于上述两类数据拥有者有出售数据的愿望, 因此对于数据平台来说, 获取数据拥有者所贡献的数据是较为简单的. 但是, 当下的许多数据交易平台不仅仅依靠数据拥有者贡献数据, 也通过各种技术手段自行收集数据. 对于数据交易平台来说, 最容易获取到数据的地方便是互联网. 互联网上存在着大量结构化、半结构化、非结构化数据. 网络爬虫是获取上述数据较为简便和基础的方法, 即通过一定规则, 自动抓取互联网中的信息. 同时, 也存在着一些数据收集平台, 比如 Apache Flume[48]、Fluentd[49]、Logstash[50]和 Splunk Forwarder[51]等. 此外, 已经存在很多工作研究如何从互联网上抽取结构化数据[52,53]. WebTables[54,55]是其中最为成功的方法. WebTables 可以自动抽取以 HTML 表格形式发布在互联网上的数据, 并将其转化为关系数据库中的表. 比如, WebTables 可以抽取所有百度百科的信息框, 首先通过网络爬虫收集百科中的所有 HTML 表格, 然后应用分类器确定哪些表可以被视为关系数据库的表. 每个关系表由一个描述列和一组元组模式组成. 同时, 由于互联网数据的多样性, 上述的表抽取技术又被扩展到了更多的用途中. 其中一种便是通过以垂直表格和列表的形式提取关系数据, 并利用知识库将表格抽取扩展到识别 HTML 标签之外的地方[56,57].

数据收集作为数据交易市场中的第一环, 解决了数据交易中“从无到有”的问题. 上文总结了数据收集中常用的方法, 除了向数据拥有者购买数据以外, 数据平台自行通过网络爬虫等方法进行收集也是交易数据的重要来源. 但是, 由于数据来源的复杂性, 经过上述步骤收集到的源数据大多是异构的, 并且其组织过于松散, 不能直接出售, 因此需要对数据执行集成验证等操作, 我们将在下一小节中进行介绍.

2.2 数据集成

由于大数据来源和种类的多样性, 数据平台收集到的数据往往以各种各样的形式存在. 为了达到出售数据的目的, 就要求平台对数据进行一系列整理、去重、分析和验证等操作, 即数据集成. 数据集成的目的是使用相应策略, 将收集来的数据进行整合以满足出售需求[46]. 通常来说, 在数据交易平台上所需要进行的数据集成任务分为以下 3 个步骤[58]: 模式匹配、实体解析和数据融合.

模式匹配是数据集成的第一步, 也是最重要的一步. 大致过程为: 数据交易平台首先对收集到的数据生成统一视图, 称为中间模式, 将不同数据属性和中间模式属性进行匹配, 然后阐明数据源内容和中间模式之间的语义关系. 对于数据交易平台来说, 其收到的源数据是多种多样的, 进行模式匹配最重要的就是选择这些数据的启发式信息, 这些信息以 3 种形式存在: 文本、结构和约束. 首先, 利用文本信息进行匹配, 使用信息检索中常用的本文处理方法, 对数据属性名称、数据实体等启发式信息进行处理, 计算其相似度; 其次, 可以利用结构信息, Madhavan 等人[59]将需要进行模式匹配的数据构建为树结构, 通过计算两棵树之间的相似性来反应数据模式之间的相似度. 最后, 由于进行交易的数据许多自身就带有约束信息, 可以用来计算模式的相似度, 如在物联网数据交易中使用值域来区分不同来源的数据[47].

由于数据交易平台所收集到的数据源自不同的贡献者, 因此对同一个数据实体会产生不同的描述, 实体解析, 就是指将不同描述的实体进行解析并映射到现实世界中实体的过程[60]. 传统实体解析方法大多基于成对比较, 在有大批量数据需要处理的数据交易场景中并不适用. 目前主流方法是使用分块技术. 即首先对收集到的数据进行预处理, 将其分为更小的数据块, 在每个小块内进行实体解析. 数据融合表示结合多个源头的数据信息, 改进最终结果质量的过程[46], 其最重要的任务就是解决多个数据源冲突的问题. 随着互联网上数据规模的增大, 其中存在的错误数据也越来越多, 为了高效地提升融合后数据的质量, Yin 等人[61]首先提出了真值发现 (truth discovery) 问题, 即通过可靠的数据源找出冲突数据中真实值的方法, 并逐渐成为与数据融合享有同等地位的研究课题. 在数据交易市场中, 由于数据来源的多样性, 会存在多个源头的数据适用于同一个购买需求的可能性, 实体解析和数

据融合就是解决这种冲突的办法. Fernandez 等人[46]为解决数据交易中产生的数据集成问题, 提出了 DoD 引擎 (dataset-on-demand engine), 负责出售数据前的实体解析和数据融合任务. DoD 引擎的输入是消费者的数据购买需求, 输出是满足其购买需求数据组合 (Mashups). 考虑如下的场景: 当一个数据消费者希望购买天气数据, 但是有许多不同的数据拥有者都可以提供该数据. 文献 [46] 认为可行的解决方案是, 将表示统一实体的元素和其不同的取值列在一起, 并提供了多种数据融合算子 (比如基于多数投票的方法选择真值), 让数据消费者手动选择想要使用的算子.

数据集成效果的好坏直接影响到数据消费者获取数据质量和数据交易平台出售数据效率的高低. 数据定价与交易的相关工作大都注重于交易和定价机制的设计, 忽视了数据集成的相关工作. 但是在传统数据管理领域, 已经存在许多解决方案[60,62−67]可供参考. 设计选用高效可行的数据集成方法是数据平台在出售数据之前要进行的必须操作, 是提高数据出售效率、价格和售出数据可用性的重要手段. 经过集成后的数据应该按照格式或使用途径具有明确分组, 各个分组之内、不同分组之间应该具有高内聚、低耦合的特性.

3 数据管理与分析

数据管理和分析是进行数据交易前的必要过程: 为不同数据选择适当的组织形式和存储方式, 对数据进行分析以得到不同数据应该用于何种用途、适用于何种交易模式, 同时初步分析出其商业价值, 以便设定价格进行交易. 本节将在考虑数据来源、形式多样性等前提下, 对数据平台在数据管理与分析上应该做出的工作进行介绍. 现存的数据市场并不支持多种类型数据进行同时交易, 并且尚未有文章对数据平台中数据的组织和存储有深入研究. 因此本节将在统一交易平台的视角下对数据平台所应当考虑的数据组织形式和存储方式以及数据分析任务进行探讨.

随着分布式的兴起, Hadoop 分布式文件系统 (Hadoop distributed file system, HDFS) 是大数据系统中最常见的一种异构存储方式. HDFS 支持多种文件[68]. 除了支持 CSV、XML、JSON 等类型文件, 还支持图片等二进制文件格式, 为了解决数据体积过大的问题, HDFS 还支持压缩后的文件格式, 如 Snappy、Gzip 等. HDFS 支持列式存储格式如 Parquet 和基于行的存储格式如 Avro, 可以实现轻松的架构管理. 但是, 由于数据定价和交易的环境下, 平台需要对不同来源、不同类型的异构数据进行统一管理并进行后续查询、出售等操作, 因此单独的 HDFS 并不能很好支持数据平台所拥有全部数据的存储.

上述需求在大数据的存储和分析中也极为常见. Stonebraker[69]提到了多存储 (polystore) 的概念, 支持对异构数据多种存储方式的集成访问. Hai 等人[70]提出了多存储的方式, 适用于数据交易平台的存储模式. 文章按照数据源格式将其分别存储到关系数据库 (如 MySQL)、基于文件的数据库 (如 MongoDB) 或图数据库 (如 Neo4j) 中. 对于不能直接存储到关系数据库或非关系数据库中的文件, 则存储在 HDFS 中. 如果上述默认存储方式都不能满足用户需求的话, 数据交易平台还可以根据其数据类型自定义选择存储方式. 多存储概念的提出为数据交易平台的数据管理提供了存储上的解决方案. 但是由于大数据的时效性极强, 很容易产生数据集过时的问题. 因此, 数据交易平台还要负责对存储的数据进行更新、维护等操作. Liu 等人[71]提出了动态数据市场框架, 来解决上述问题. 该框架包含了一个在线共享计划选择算法 ManagedRisk, 可以保持数据视图的生成效率, 同时支持对数据视图的动态更新.

除了传统的元数据建模等方便数据存储和管理的分析任务, 以数据交易平台的视角来看数据分析, 还包括对数据适用的范围和出售模式以及商业价值进行分析, 以提高数据交易效率, 确定数据出售价格基准.

由于数据交易平台从各个源头收集到和数据拥有者主动贡献的数据具有类型多样、形式多样的特点, 为了满足数据消费者针对不同任务提出的不同数据需求, 数据交易平台有必要对收集到的数据进行适用范围和出售模式的分析, 满足买家对于数据的期望, 让数据投放更加精准, 从而提高数据的出售效率. 例如收集到的传感器数据、图像音频数据或者带有标签的数据更倾向于出售给有构建机器学习模型需求的数据消费者, 若上述数据是包含隐私的个人数据, 数据交易平台则要考虑出售这些数据会导致的隐私暴露问题, 以及对数据拥有者给予相应的补偿; 对于存储在关系数据库中的数据, 则需要为其中的某些条目或者视图设置相应价格, 以便数据消费者查询.

除了上述的适用范围和出售模式的分析, 对于大型数据交易平台来说, 还需要对数据的商业价值进行初步分析, 从而设定相应的价格基准, 为后续的定价提供依据. 需要注意的是, 在该阶段对数据的分析并不是要计算出数据价格, 而是根据数据能够给购买者提供的效用、数据的市场价值等因素估算出大致价格范围, 为后续定价过程提供参考. 但是, 现有的文献如 [31,45] 等大都假设在数据出售前数据交易平台已经了解了数据的基础价格, 忽略了确定价格基准的过程. 通常来说, 可以通过数据挖掘的方法获取数据中的商业价值. 但是, 这类方法也存在相应的问题[72]. 第一个挑战侧重于数据访问和计算过程. 由于数据交易平台倾向于使用分布式存储系统, 系统中数据量在不断增长, 数据交易平台必须具备处理分布式和大规模数据存储的能力. 大多数数据挖掘算法需要将所有必要的数据加载到主内存中, 这在数据交易的场景下显然是一个技术挑战, 因为从分布式存储系统移动数据的代价是极其昂贵的. 第二, 由于不同数据对于不同数据消费者来说有着不同的意义, 其对数据的不同用途也会导致数据商业价值发生变化, 因此采用数据挖掘方法也不能得出让买卖双方都满意的价值结果. 对于数据交易平台来说, 它在本阶段的任务是给需要出售的数据设定一个价格基准, 因此 Chen 等人[44]采用了市场调查的思路, 即让数据拥有者或者数据交易平台在出售数据前先进行市场调查, 以确定代表潜在数据消费者对机器学习模型实例的需求和价值分别对应误差的关系曲线. 该曲线将需求和价值表示为训练后机器学习模型误差的函数. 数据交易平台则通过市场调查得来的曲线, 构建呈现给数据消费者的价格-误差曲线, 即确定每个误差值所对应数据的价格基准.

上文总结了数据交易平台在数据管理和分析阶段需要完成的相关任务, 需要注意的是, 虽然这些任务是设计一个统一数据交易平台所必须要考虑的, 但是很少有文章研究上述内容在数据定价与交易领域的应用. 然而, 这些内容在大数据管理方向中已被广泛的研究, 因此本文挑选了部分适合于数据定价与交易的内容, 并对其做了介绍.

4 数据定价

在数据交易时, 数据价值通常用交易时价格来体现, 因此数据定价是数据交易中最为重要的任务之一. 本节将介绍在数据定价时各方所使用的衡量数据价格的各种机制, 根据其思路不同, 将其分为 3 种类型, 如表 1 所示. 需要注意的是, 3 种定价思路并不是互斥的, 由于其侧重点各不相同, 因此在数据交易过程中可能同时存在. 基于任务的定价更侧重于根据数据对于消费者而言能产生效用的大小而定价, 比如: 基于查询的定价是根据该查询任务以及每个条目的组合方式计算出整体价值; 基于模型的定价是根据该机器学习模型或预测任务对数据消费者能够产生的效用来定价. 基于价值的定价则关注数据的内在价值, 比如: 基于隐私补偿的定价, 其补偿值大小就是依靠数据中包含隐私的多少而确定的; 基于数据质量的定价则是依据收集到的数据质量的高低来划分版本, 从而确定价格. 而与上述两种思路完全不同的是, 以博弈论和拍卖为代表的基于经济学定价思路则主要依靠市场类型、机制设计和参与人之间的关系来确定价格. 该定价思路既可以看作是定价模型, 因为其包含了确定价格的功能, 同时还侧重于在确定了数据基础价格之后, 市场类型、机制设计和参与人行为等与数据本身无关的因素对价格产生的影响. 因此基于经济学的定价不仅仅是一类定价方法, 该研究内容与数据交易机制的设计息息相关, 因此在本节中仅对其进行简单介绍, 具体内容在数据交易部分进行更深入的讨论.

表 1 数据定价方法分类

定价思路	思路介绍	定价方法	参考文章
基于任务的定价	依据该数据产品对于数据消费者执行某项任务所能产生的价值来确定价格	基于查询的定价 基于模型的定价	[27,31−33,73−77] [29,43,44,78]
基于价值的定价	依据该数据产品的内在价值如隐私包含程度、数据质量优劣来确定其价格	基于隐私补偿的定价 基于数据质量的定价	[33,45,47,75,79] [80−84]
基于经济学的定价	在确定基础价值的前提下, 依靠市场如供需关系、市场类型等经济学方法来确定数据价格, 主要考虑市场类型和参与人行为对价格产生的影响	基于花费的定价、基于供需关系的定价、基于博弈论的定价、基于拍卖的定价	[23,43,85−95]

4.1 基于查询的定价

由于现有的数据市场所采用的数据定价策略大多数只允许买家选择固定的某些视图, 不支持个性化 SQL 查询操作, 因此, Koutris 等人在文献 [31] 中首次正式提出了基于查询的数据定价 (query-based data pricing) 概念, 并提出了能够给任意查询分配价格的定价框架, 首先, 卖家需要给一定数量的数据视图指定价格点 (price points), 当查询来临时, 将其价格设定为与查询结果相关的所有视图价格和的最小值. 提出的框架除了满足无套利公理外, 文章定义了无折扣 (discount-free) 公理: 要求定价函数计算查询价格时使用某个视图的价格不能低于卖家预先定义的价格点. 文章证明了当存在连接查询时, 在大规模数据库上计算任意查询的价格是 NP 难问题, 并描述了一种可以在多项式数据复杂度上计算出价格的连接查询. 该模型的灵活性在于, 它可以给任意查询分配价格, 而不仅仅限制买家购买特定的视图. 但是本文仅给出了理论框架部分, 没有进行进一步实验研究. 且文章提出的方法仅支持简单查询语句, 不能满足数据市场中进行复杂查询的需求. Koutris 等人[27]则基于文献 [31] 的理论设计了查询定价系统 QueryMarket, 改进了文献 [31] 中只能对一部分简单的查询进行定价的缺点, 将无套利定价问题转化为整数线性规划问题, 大大降低了算法执行大规模 SQL 查询的时间复杂度. 文章还研究了收入在查询结果贡献者之间公平分配的问题. 此外, 由于数据消费者在购买数据时可能会进行多次查询, 这容易产生对同一数据进行多次收费的问题. 因此, 文章引入了记录查询历史的方法, 解决了买家多次查询可能包含重复数据, 从而导致重复收费的问题. 除了文献 [31] 外, 在解决重复收费问题上, Upadhyaya 等人[73]提出了退款 (refunds) 的概念, 将支付过程分为了两个步骤, 买家在收到数据时按原价进行正常的支付, 发现有重复购买的数据时, 则可以向平台提出退款申请, 并提交重复购买的证明, 支持多个买家进行分组退款.

由于上述文章设计的定价方法仅考虑较为基础的 SQL 查询语句, 对复杂查询操作支持度较低. 因此, Li 等人[32]基于文献 [31] 中的理论, 对独立于数据库实例的线性聚合查询定价方法进行了初步讨论, 证明了在一些情况下精确计算聚合查询价格的开销是巨大的. 因此文献 [74] 针对该问题, 提出了支持近似聚合查询的定价框架. 文章采用了 Sampling 技术, 可以在误差范围内提供查询的近似结果, 并提供了将现存的定价模式转化为精确和近似聚合查询定价模式的框架. 在 Nget 等人[75]提出的个人数据定价框架中, 支持对含有噪声的数据进行聚合查询, 并提出给每个数据卖家应得的隐私补偿, 文章采用差分隐私作为衡量隐私补偿的依据. 同样, Li 等人[33]也结合了差分隐私和基于查询的定价理论, 允许消费者进行带有噪声的查询, 并对查询造成的隐私损失进行了量化, 根据隐私损失的多少对数据进行定价.

上文中所提到的定价方式都是基于视图进行定价. 买家所购买到的数据大都以视图为单位, 但是视图粒度的定价对于很多应用场景来说过粗, 虽然在很多情况下视图粒度可以转化成元组粒度, 但会引起严重的可扩展性问题. 因此 Tang 等人[76]提出了基于最小来源元组的定价: 对构成最终查询结果的元组进行追踪, 将其看成是一个整体, 取可以构成查询结果的最小元组集作为定价依据. 文章采用 P-Norms 作为价格聚合方法, 提出了精确算法和一系列的近似算法. 类似地, Shen 等人[77]也提出了基于元组粒度的个人数据定价平台, 在平台中对个人数据进行正面评级和反向定价, 保证了定价模型的透明性, 减少了个人数据交易市场中存在不对称的可能性.

基于查询的定价算法最初仅支持较为简单的查询语句, 经过改进, 现如今已经可以支持大批量复杂查询, 同时还可以进行近似匹配. 基于查询的定价方法本质属于基于任务的定价方法, 需要先设定元组价格或条目价格, 然后依据此价格, 通过算法生成数据消费者所需购买任意视图的价格. 该方法一般适用于存储在结构化数据库或非结构化数据库中的易于查询的数据, 有着定价灵活、设置完基础价格后不需要进行更多维护的优点. 但是, 由于基于查询的定价方法出售的数据是多个条目的集合, 单个数据条目并无特殊价值, 因此其价格的可解释性较低[15], 同时, 生成价格的时间复杂度普遍较高. 再者, 大数据较强的时效性导致离线定价算法存在不能实时更新价格的问题. 因此, 如何解决以上问题也是研究者需要着重考虑的.

4.2 基于模型的定价

随着大数据产业的不断进步, 使用机器学习模型进行大数据分析已经成为行业最通用的准则之一. 目前已有相当多的工作集中在研究机器学习模型性能和准确率上, 但在如何以高性价比获取数据上的工作研究较少[44]. 数据定价思想的兴起为上述问题提供了较好的解决思路. Jia 等人[29]专门为 kNN 模型设计了定价机制. 采用沙普利

值法来衡量每个数据点对模型的贡献度, 以此为依据对其进行定价. 而 Chen 等人在文献 [44,78] 中提出了基于模型的定价理论, 卖家直接出售训练好的机器学习模型实例, 而不是训练数据, 使用模型精准度的不同来划分不同价格水平. 框架如图 2 所示.

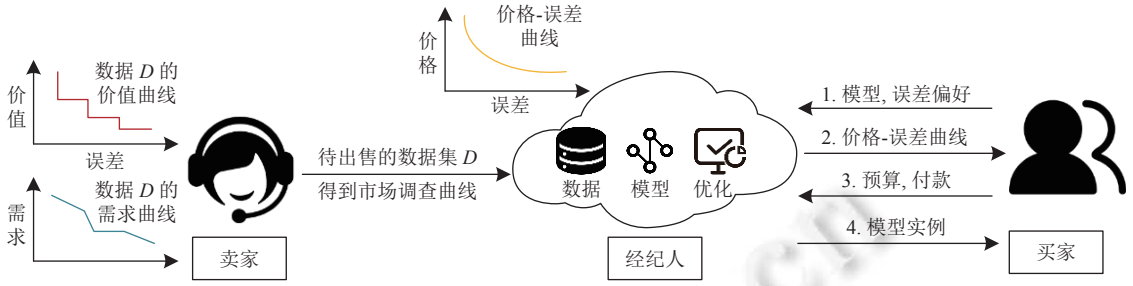


图 2 基于模型的定价方法框架[44]

框架包含 3 种参与者, 卖家提供数据集, 买家希望从市场中购买机器学习模型, 经纪人负责在中间协调.

首先, 卖家或者经纪人进行市场调查, 以确定代表潜在买家对机器学习模型实例的需求和价值分别对应误差的关系曲线. 该曲线将需求和价值表示为训练后机器学习模型误差的函数. 经纪人则使用市场调查的信息来构建呈现给买家的价格-误差曲线. 买家指定所需的价格或误差预算, 经纪人根据误差和价格预算计算出合适的机器学习模型, 将模型返还给买家. 在模型训练方面, 文章提出了一个噪声注入机制, 允许经纪人对每组数据只训练一次, 得到一个最优模型, 当某个买家提出购买需求时, 经纪人向模型内注入随机高斯噪声, 并将结果返回给买家. 定价机制根据注入到模型实例中的噪声多少来确定价格: 加入的噪声方差越低, 训练出的模型效果越好, 价格就越高. 该框架可以为买家提供不同版本的机器学习模型, 以迎合买家不同需求.

该文章还证明了提出的定价方法是单调并且是次可加的, 即保证了定价函数是无套利的. 同时, 文章认为收入最大化问题的核心是通过给定点插入单调和次可加函数的问题, 即使在简单的收入模型下, 收入最大化问题也是难以解决的, 因此设计了一个优化框架, 对次可加性的约束进行了松弛操作, 使得在多项式时间内实现收入最大化的近似保证.

上述文章虽然同时考虑到了数据拥有者, 数据平台和数据消费者, 但是却没能将参与数据交易三方进行形式化描述. 而 Agarwal 等人[43]以一种更加全面的角度, 分别对三者进行了参数化操作, 并用数学模型规范描述了对其在数据市场中的行为. 与文献 [44,78] 不同的是, 在文献 [43] 所提出的定价模型中, 数据消费者需要向数据平台发送其预测任务和支付意愿, 其中支付意愿是由该预测任务精度的边际上升大小确定的. 数据平台根据提供的信息为其选择相应的数据条目执行预测任务, 将预测结果返回给数据消费者. 支付价格是由上一轮数据消费者的支付意愿和支付价格依据收益最大化原则计算出来的. 数据消费者在完成支付后拿到的是预测任务的输出值而非完整的模型. 此外, 为了应对数据的可复制性, 文章将数据相似度引入补偿计算阶段, 为提供相似数据的数据拥有者重新计算补偿值, 以得到一个相对公平的补偿价格. 但是文章也存在相应的问题: 数据消费者产生了新的预测需求, 仅提供给其预测值, 则会导致数据消费者需要再次执行交易流程. 数据消费者需要将预测任务提供给平台, 可能产生隐私泄露的风险. 同时, 该模型对于参与数据交易三方的建模较为理想化, 不能满足复杂状况下的数据交易需求.

基于模型的定价是随着机器学习、人工智能技术的发展而演化出的一种具有高度针对性的定价方式, 属于基于任务的定价模式. 这种定价方法与机器学习模型高度契合, 方便了数据交易市场中具有特定目标的买卖双方进行沟通和交流, 有助于机器学习领域数据的充分流通. 但是就目前的研究现状来看, 由于针对性过强, 该方法的适用范围较差. 而且, 如果想设计一个统一、普遍适用的模型定价平台, 则会面临着繁重的训练任务, 同时, 基于模型的数据定价方法只能大致预估模型的效果, 而实际应用效果可能与数据消费者的效用预期产生出入等缺点也是需要深入思考的问题.

4.3 基于隐私补偿的定价

自数据分析、深度学习技术开始蓬勃发展以来, 个人数据就被看作是互联网世界中的新石油[3]. 每一分钟都

有新的个人数据产生并被收集. 因此, 隐私保护也就成为业界关注的重大课题. 而数据交易过程中, 无疑会涉及个人数据的交易, 个人数据交易在雅虎、谷歌等互联网巨头中也早已屡见不鲜. 其中包含的个人隐私可以作为衡量数据价格的重要指标. 为了应对卖家在数据交易中产生的隐私损失问题, 也为了激励更多人出售个人数据, 文献 [33,45,47] 等都提出需要给数据卖家一定的隐私补偿. 因此如何衡量隐私损失以对卖家进行补偿是基于隐私补偿的定价需要研究的问题之一. 通常情况下, 使用 ε 差分隐私来衡量隐私保护的水平, 其定义如下.

假设数据集 T 经随机算法 M 处理后的输出结果集合为 Y, Y 的任意子集为 D, 对于任意邻近数据集 T 和 T', 若算法满足不等式:

$$\frac{\Pr(M(T)=D)}{\Pr(M(T')=D)} \leqslant e^{\varepsilon} \tag{3}$$

则称算法 M 提供了 ε 差分隐私保护. 差分隐私的出现, 解决了数据定价中衡量隐私损失的问题, 现阶段基于隐私补偿的定价方法大都采用了差分隐私的 ε 值作为确定价格的参数.

Ghosh 等人[45]首先提出个人隐私数据交易, 并给卖家提供隐私补偿. 该文章考虑了数据拥有者对于其出售的数据持有不同的隐私态度. 为了揭示该数据拥有者的隐私态度, 数据平台或经纪人使用拍卖的方法, 让每个数据拥有者提交能反映其隐私态度的出价, 并根据收到的出价, 决定从数据拥有者那里购买的隐私水平, 然后生成一个带有噪声的查询输出, 以确保该隐私水平得到保证. 文章选择了差分隐私作为评价隐私泄露水平的方法, 基于差分隐私 ε 值对数据拥有者进行补偿. 但是文章设计的隐私补偿机制问题在于, 即便同一批出售数据的数据拥有者对于隐私有着不同的估值, 该机制也会对所有数据被使用的拥有者计算一个相同的 ε 值, 并基于此对其进行隐私补偿. 这就导致该机制会对某些数据拥有者过多的隐私保护, 缺少对隐私数据补偿机制进行个性化定制的能力. Zhang 等人[79]将这种机制称为“伪个性化”, 因为其不能体现出数据拥有者对于隐私保护水平的差异. 文献 [79] 则针对上述缺点提出了可以保证卖家个性化隐私需求的定价机制. 同样采用差分隐私衡量隐私损失, 并且在支持查询以高精准度输出结果的情况下, 确保数据拥有者自定义的差分隐私参数能够得到满足. 文章使用了反向拍卖机制决定购买哪个卖家的数据以及应该支付多少隐私补偿. 同样地, Li 等人[33]所采用的隐私损失衡量方法也是基于差分隐私的, 并定义了微支付函数, 用以确定给某次查询所设计的框架实现了查询价格和隐私补偿之间的平衡. 但是文章采用线性隐私衡量机制, 并允许用户自定义隐私损失系数 c_i, 所以在 ε 值相同的情况下, 用户会倾向于定义过高的隐私系数 c_i, 从而获取不当的过高利润, Zhang 等人[79]的方法也存在类似问题. 因此如何在补偿价格和隐私损失之间取得适当的平衡, 是个人数据市场中重要的挑战. Nget 等人[75]设计了可以在上述两个方面取得高效平衡的定价机制. 首先, 给基于隐私补偿定价中的支付模式 (payment schemes) 做出了形式化的定义: 支付模式是一个非减函数 $w:\varepsilon \to R^{+}$, 代表中介和卖家之间就卖家的实际隐私损失 ε_i 应该获得补偿的量. 文章设计了两种支付模式: 对数支付模式和次线性支付模式, 分别对应低风险低回报和高风险高回报. 卖家可以根据自己的隐私损失或风险倾向选择对应的支付模式.

通常情况下, 基于隐私补偿的定价是以数据平台的视角, 考虑在其收集数据时, 遇到包含个人隐私数据的情况. 本质上属于基于数据内在价值的定价思路, 依据数据拥有者对于隐私损失风险的承担能力和对收益的渴望程度来确定补偿价格, 同时还要考虑到过低的隐私暴露可能降低数据消费者效用的问题. 基于隐私补偿的定价重点关注数据拥有者和数据平台之间的交互, 对于数据的存储形式则无太高要求. 在隐私损失的衡量方法上, 除了差分隐私外, 信息熵也可以作为数据中包含隐私水平的衡量机制, 应受到更多关注.

4.4 基于数据质量的定价

数据质量也是确定大数据价值的一个重要属性. 依据数据质量确定数据价值, 重点分为两个方面: 确定数据质量维度和版本控制 (versioning). 早在 20 世纪初, Wang 等人[96]对数据质量特征进行了两阶段的分类研究, 制定了相关的分层框架, 将数据质量特征分为了 15 个维度. 在数据交易的热门领域——互联网数据中, Naumann 等人[97]将数据质量准则分为了 4 个类别: 内容相关 (content-related)、技术相关 (technical)、知识相关 (intellectual) 和实例相关 (instantiation-related), 从这 4 种类别中详细研究了 22 个衡量互联网数据质量的维度. 版本控制是依据数据在每个质量维度下的得分, 给数据分为不同的版本, 以满足不同消费者对数据的需求, 并以此设定不同价格.

Stahl 等人[80]借用了文献 [97] 的 22 个维度, 依据是否可以自动获得这些维度, 将其划分为自动、手动和混合 3 类, 并设计了一个适用于数据定价的数据质量打分系统. 该系统的主要目的是为来自不同数据拥有者的类似数据提供比较依据. 系统设计了一个线性加权打分机制, 允许数据买家根据自己偏好为不同的数据质量维度设计不同的权重. Yu 等人[81]则研究了在垄断平台下基于多数据质量维度的定价问题. 该平台依据数据质量的多个维度, 并且将不同维度之间的相互作用也考虑在内, 设计了版本控制策略, 并且建立了一个两层的编程模型, 包括一个领导者 (数据平台) 和多个追随者 (数据消费者). 第 1 层, 领导者根据多个数据质量维度和其中的相互作用, 决定不同的版本和其出售价格, 以最大化自己的收益; 第 2 层, 潜在的消费者根据自己对不同质量的偏好需求做出自主选择, 决定购买的版本.

上述方法在进行给数据进行定价时, 虽然考虑了综合的基于数据质量的打分系统, 但是由于综合的数据质量维度存在形式和尺度不统一、消费者的效用函数难以计算从而影响定价等问题, 难以适用于高效的数据定价应用场景[82]. Yang 等人[83]在总结了数据质量的不同衡量维度之后, 选取了精准度 (accuracy)、完整度 (completeness) 和冗余度 (redundancy) 作为衡量数据质量的方式, 分别表示数据源中具有正确值的数据比例、数据集中完整数据的比例和数据源中重复记录的比例. 并使用 Stahl 等人[84]的打分方法, 允许依据上述 3 种维度对数据进行连续的版本划分, 以产生不同质量水平的数据. 最后计算出的质量分数是在 (0, 1) 之间的. 随后, 文章以机器学习的分类算法为例, 提出了基于质量水平的效用函数, 并基于从经济学角度考虑的消费者支付意愿函数, 共同计算出某个质量水平数据的出售价格. 除此之外, 在 Zhang 等人[82]设计的以质量为导向的数据定价策略中, 考虑了精准度、完整度、及时性和一致性 4 个维度, 采用与文献 [83] 相同的线性加和方式将其整合在一起. 对于不同质量水平的效用函数, 提出了 Floating 方法. 首先根据所有数据库实例的数据质量计算出质量标准 FQ_S, 再针对某个数据库实例计算出其质量 FQ, 最终使用如下方法计算出价格:

$$p_{\text{final}} = p + \frac{(FQ - FQ_S)}{FQ_S} \times pC \tag{4}$$

其中, p 是初始价格, C 是常量参数, 反映了质量对最终价格的影响程度.

基于数据质量的定价方法依靠效用价值理论, 以数据质量作为数据价值或者消费者效用的决定性因素, 通常情况下, 适用于质量维度较少或者容易量化的定价任务中. 基于数据质量的定价方法由于从自身角度考虑数据价值以及从消费者角度考虑数据效用, 通常具有较好的透明度以及较高的可解释性. 但是, 对于数据来说, 质量并非决定其价值的唯一因素, 同时在大数据时代各种数据纷繁复杂, 很难找到统一的、令各方满意的数据质量衡量维度, 此外, 数据质量和消费者效用之间的关系本身也不容易量化. 因此, 在设计数据定价方法时, 将基于数据质量的定价作为确定数据价值的一个参考维度则更为妥当.

4.5 基于经济学的定价

基于经济学的数据定价是依据如供需关系、博弈论等经济学中的基本理论为数据确定价格的方法. 其中最简单也最为基础的是基于花费的定价方法. 该方法考虑商品的所有成本, 并将总成本的一个比率设定为利润, 以此确定价格[72]. 一般来说, 数据产品的成本可以分为收集成本即收集数据所产生的花费、存储成本即数据长时期存储在本地数据库或云端数据库产生的花费、复制成本即数据在被出售或传播时所产生的花费等. 该定价方法的优点是模型简单便捷, 但是仅考虑了数据的内在属性来决定数据价格[98], 而没有顾及市场的供需关系等外在属性[99]. 同时, 由于每个阶段的花费很难具体量化到每一个数据条目上, 当卖出部分数据时很难为其设定科学的价格. 此外, 上文提到由于大数据复制代价极低, 因此随着数据在市场上的传播, 价格会变得越来越低, 同时竞争对手容易将数据复制为己用, 导致数据出售者不再有出售数据的欲望, 影响数据市场的健康发展.

供需关系模型是经济学中决定商品价格的模型之一, 其关系用供需函数来描述. 在市场中, 用 P 来表达数据产品的单位价格, Q 表达数据产品交易数量, 那么需求曲线表示在其他因素不变的情况下, 数据消费者愿意购买的数据量随着数据单位价格的变动而变动, 公式表示为 $Q_D = Q_D(P)$; 供给曲线是指在其他因素不变的情况下, 数据拥有者愿意提供的数据量随着数据单位价格变动而变动, 公式表示为 $Q_S = Q_S(P)$. 基于上述关系, 我们可以构建出供给和需求的关系[72]. 如图 3 所示.

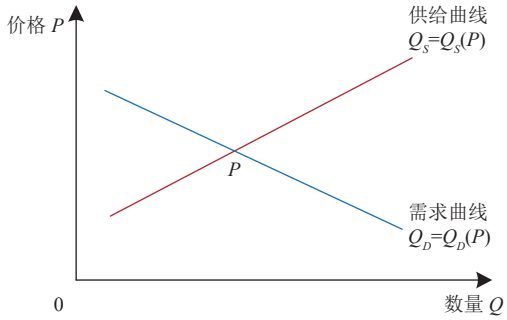


图 3 数据市场中供给需求关系图

由图 3 可以看出, 供给曲线和需求曲线必定相交. 交点即为买卖双方的平衡条件. 在此均衡点处, 有 $Q_D = Q_S$, 价格 P 称为清算价格, 此时市场中商品没有短缺或过剩, 不存在使得价格变动的外在压力. 该模型有两个基本特点[100]: 第一, 该模型所描述的是买卖双方在进入市场之后一系列持续性的一致行为; 第二, 买卖双方不能随便改变决定价格的进程, 所有决策过程都是由市场决定的. 因此, 该模型可以使得市场是公平的[72]. 但是, 由于价格完全交由市场决定, 难免会导致寡头垄断的情况出现, 这也是供需关系模型所存在的问题.

上文描述的两种方法的定价思想较为简单, 因此并未被数据定价文章广泛采用. 经济学中的另一种重要研究内容, 博弈论, 则广受数据定价研究者青睐. 近年来出现了许多使用博弈论中方法研究数据定价[85−91]的文献. 博弈论主要研究决策主体的行为发生直接相互作用时的决策以及这种决策的均衡问题. 在数据定价中的应用主要包括 3 个方面: 非合作博弈、Stackelberg 博弈和讨价还价. 非合作博弈的前提是数据交易的参与者之前不可能达成具有约束力的共识, 即都处于冲突状态, 以竞争的方式参与交易. 该博弈模型要求参与者在进入市场时即公布自己的价格策略, 同时在知道对手价格策略的前提下, 以自身收益最大化为目标计算出该博弈的纳什均衡, 即可得到成交价格. Stackelberg 博弈模型要求参与者中有领导者和追随者. 领导者首先发布自己的价格策略, 追随者观察到该策略后, 再决定自己的价格策略并发布, 双方都根据对方策略来决定自己策略以达到收益最大化, 如此往复以达到最终交易价格. 讨价还价是交易的各方经过一轮或多轮谈判就达成交易价格的过程. 而作为不完全信息博弈的重要应用, 拍卖也是最流行的数据定价机制之一[23,43,92−95]. 拍卖是通过市场驱动参与拍卖的双方在规则框架内进行自主竞价, 从而对商品进行分配, 并赋予对应的价格[101]. 在数据定价中的应用主要分为密封拍卖, 组合拍卖, 双边拍卖等 3 种方式.

由上文的描述可以看出, 以供需关系模型、博弈论和拍卖为代表的基于经济学的定价思路与基于价值和任务的定价思路不同, 后者更加侧重于根据数据实际效用或价值以设定相应价格, 而前者更加侧重于市场类型, 机制设计和参与人行为等与数据本身无关的因素对价格的影响. 基于经济学定价的思路既可以看作是定价模型, 包含了确定价格的功能, 可以在数据分析后直接作为确定数据价格的重要依据, 另一方面, 又能在其他思路确定的价格基础上通过市场类型, 机制设计和参与人行为进一步对成交价格和数据出售方式产生影响. 而在整个数据的生命周期中, 市场类型的区分和参与人行为的规范都与数据交易机制的设计息息相关. 因此不能够将其简单的看作数据定价方法. 有关于经济学定价在数据交易方面的体现, 我们将在下一节做更深入的讨论.

5 数据交易

数据交易作为大数据生命周期的重要组成部分, 与数据定价有着互补的关系: 数据定价侧重于设定数据价格, 而数据交易则需要研究数据市场类型、参与交易各方行为等对市场和数据出售价格的影响. 因此在第 4 节介绍数据定价的基础上, 本节对数据交易过程中涉及的市场类型、机制设计和参与人行为规则进行了研究. 本节首先介绍了市场结构, 并对其进行了简单分类, 介绍了相应分类下的文章, 然后以博弈论和拍卖为例, 研究了数据市场环境下的交易机制设计问题, 并总结了各种方法的优劣.

5.1 数据市场结构

在经济学中, 通常将市场分为如下 4 种类型: 完全竞争 (perfect competition)、垄断竞争 (monopolistic

competition)、寡头垄断 (oligopoly) 和完全垄断 (monopoly). 市场结构可以在一定程度上决定产品的交易价格, 在数据市场中也不例外. 因此, 在分析数据交易前要首先确定数据市场的类型.

在完全竞争市场下, 边际成本即每增加一单位产品所需要的成本基本等于出售价格. 这无疑极大增加了市场透明度. 使消费者可以享受到更好的服务和更低的价格. 但是对于卖家来说, 在激烈竞争下会导致利润减少, 产品同质化严重. 因此卖家会选择降低数据质量的方式来压缩成本, 导致低质量的数据产品充斥市场, 从而引起恶性竞争, 缩小市场规模. 完全竞争市场不存在垄断因素, 因此在日常生活中很难见到类似的市场.

垄断竞争市场属于垄断和竞争因素并存, 但竞争因素更多一点的市场. 在实际生活中的零售行业接近该市场模型. 在垄断竞争市场中, 厂商数量往往有很多, 其所生产的产品往往存在有一定的差异, 新进卖家在市场中立足门槛不高, 市场有着较高容忍度. 这类市场的竞争手段属于非价格竞争, 利用价格以外的因素如广告等形式实现. 但是由于数据收集和分析有着较高的门槛, 不会有过于多的卖家参与到市场中, 因此属于此类别的数据市场较少.

寡头垄断是在市场竞争后, 为数不多存活下来的厂商组成的市场. 在寡头垄断下的市场中, 垄断寡头有足够大的权利增加自己产品的利润. 这些厂商对市场上流通的数据产品具有极强的控制能力, 包括其存储方式、分析过程和产品价格. 博弈论中的伯兰德模型、古诺模型和 Stackelberg 模型均是研究寡头垄断下的市场.

在完全垄断结构中, 只有唯一的厂商垄断整个行业. 厂商通常采用价格歧视的方法作为竞争手段, 针对消费者提供同样的商品或服务, 但对于消费者不同的需求来设定不同的价格[102]. 在这种市场结构下, 数据平台可以从数据拥有者和数据消费者身上攫取最大化的利润, 从而获得最大收益. 但是这是以降低市场繁荣度为代价的, 由于市场有着较高的准入门槛, 缺乏竞争也使得数据市场活力变得更低.

文献 [102] 将数据市场分为了寡头垄断、完全垄断和强竞争 3 种结构, 其中提到的强竞争结构类似于上文所述的垄断竞争市场. 由于现存的数据交易市场文章在市场分类方面定位较为模糊, 因此本文将其简单的分类为垄断和竞争两类, 并对相应的文章总结如表 2.

表 2 数据交易市场分类

参考文章	市场类型	主要贡献
[31]	垄断	设计了一个定价框架, 支持在给定特定几个视图价格的情况下, 可以自动推导出任意查询的价格. 证明了价格满足无套利和无折扣两个属性. 提出了连接查询的有限形式, 称为泛化链查询(GCHQ), 可以在PTIME时间复杂度内计算出GCHQ的价格
[81]		在数据定价时, 考虑数据的多维性和维度之间的相互作用, 用以衡量数据质量. 设计了版本控制策略, 提出了基于数据质量的双层规划模型, 并使用遗传算法来求解该模型
[87]		设计了一个基于联盟区块链的数据交易框架, 在框架中, 数据拥有者决定原始数据的定价策略, 服务提供者即数据平台对数据进行处理, 划分成为不同等级出售给数据消费者. 将该问题表述为3层的Stackelberg博弈模型, 并证明了该均衡的存在性
[103]		提出了一个由传感器、数据源、服务提供者和消费者(订阅者)组成的市场模型, 根据数据效用函数得出最优的服务订阅费, 根据Stackelberg博弈得出最优定价, 以最大化数据源的利润
[77]		文章借助对现有数据定价模型和策略的比较分析, 提出了一种基于元组粒度的个人大数据定价模型. 该模型通过调查影响数据价值的数据属性, 分析数据元组的价值如何随信息熵、权重值、数据参考指标、成本等因素变化, 对个人大数据实施正评级和反向定价
[23]		提出了一种基于拍卖的大数据市场模型. 服务提供者在市场中占垄断地位. 模型首先根据数据集大小对数据分析性能的影响来定义数据成本和效用. 并根据大数据“无限供应”的特性, 提出了贝叶斯利润最大化拍卖. 通过求解利润最大化拍卖得到最优的服务价格和数据量
[104]	竞争	为物联网服务提供商提出了一种新的定价方案, 以确定分别提供给传感器所有者和服务用户的传感数据购买价格和物联网服务订阅费. 此外, 采用了捆绑策略, 允许多个提供商组成联盟并捆绑提供他们的服务, 从而吸引更多用户并获得更高收入. 联盟之中的多个供应商处于完全竞争关系
[33]		文章提出了一个理论框架, 将有噪声的查询答案作为衡量其准确度的函数, 并作为在数据拥有者之间分配价格的函数. 对数据市场中差分隐私和基于查询的定价的关键原则进行了扩展. 由于存在多个数据拥有者和数据消费者向平台出售、购买数据, 因此属于完全竞争市场

上文对基于经济学的定价方法进行过简单的介绍, 其中最广泛使用的是基于博弈论和拍卖的定价方法. 由于博弈论和拍卖不仅仅是定价方式, 还涉及市场结构和数据拥有者、数据消费者与数据平台在进行交易时做出的决

策和行为对数据市场以及数据价格产生的影响, 因此我们在下文中对二者进行介绍.

5.2 基于博弈论的数据交易

博弈论是经济学中重要的研究方法, 也称为对策论, 是研究决策主体的行为发生直接相互作用时的决策以及这种决策的均衡问题. 博弈是指两个或者两个以上理性的个体或组织, 在一定规则的约束下, 参加一系列的竞争性行为, 并且综合考虑对手可能实施的行为, 在其基础上做出最有益于自己的决策. 为了方便下文的叙述, 将博弈论中用到的基本概念介绍如下[13].

1) 参与人: 参与人是指一个博弈中的决策主体. 该主体可以是人, 也可以是一个团体组织. 参与人通过选择合适的决策使得自己的收益能够达到最大.

2) 行动和策略: 行动是指参与人在博弈的某个时间点采取的决策变量; 策略是一种规则, 决定了参与人在某种情况下应该采取何种行动, 即参与人将按照这种规则来采取行动; (如“敌进我退, 敌驻我扰, 敌疲我打, 敌退我追”是一种策略, 这里, “敌”与“我”是参与博弈的双方, “进”“退”“驻”“扰”“疲”“打”“退”“追”是 8 种不同的行动, 由策略规定于何时采取何种行动)[13,85].

3) 效用函数: 效用函数是指在一个特定的策略组合下参与人能够得从此次博弈得到的确定效用水平, 反映了参与人对此次博弈结果的期望. 效用函数可以是连续的, 也可以是离散的, 取值正负均可. 在博弈中, 每一方都要有自己的效用函数, 但并不要求一定了解另一方的效用函数.

上文已经提到, 数据市场的参与者分为了数据拥有者、数据消费者和数据平台三方. 这三者也是博弈的参与人. 博弈要求参与人均为理性人, 即采取的所有策略和行为都是利己的, 并且尽可能以最小的成本使其利益最大化.

常见的用于数据市场的博弈论模型有 3 种: 非合作博弈模型、Stackelberg 模型和讨价还价模型. 接下来将对其进行逐个介绍.

5.2.1 基于非合作博弈的数据交易

非合作博弈是指一种参与人之间不可能组成联盟或者达成一种具有约束力协议的博弈[105].

对于非合作博弈的定义, Luong 等人[106]设计了一个物联网数据交易模型, 在该模型下, 所有数据拥有者以竞争的方式参与交易, 这是非合作博弈的典型应用场景. 在博弈中, 数据拥有者是参与人, 可以自主进行决策. 用 (V,π) 表示某个博弈, 拥有 n 个参与人, 其中 V_i 表示第 i 个参与人选择的定价策略空间. V 是每个参与人策略空间的笛卡尔积: $V=(V_1\cdot V_2\cdot V_3\cdot\ldots\cdot V_n)$, π_i 表示每个参与人 i 得到的支付向量. 令 $v_i\in V_i$ 表示参与人 i 的定价策略, 可以得到 n 个参与人的策略向量 $v=(v_1,v_2,v_3,\ldots,v_n)$, 同时由于博弈参与人 i 的策略受其他参与人的影响, 用向量 $\overline{v_i}=(v_1,\ldots v_{i-1},v_{i+1},\ldots,v_n)$ 表示除参与人 i 以外的其他所有参与人策略所构成的策略集合. 那么就有参与人 i 采用策略 v_i 以得到 π_i 的支付. 纳什均衡即表述一个给定的策略向量 $v^*=(v_1^*,v_2^*,v_3^*,\ldots,v_n^*)\in V$, 在其他参与人不改变自己策略的情况下, 没有任何一个参与人希望通过改变自己的策略来提高收益[13,72], 即:

$$\forall i,v_i\in V_i:\pi_i(v_i^*,\bar{v}_i^*)\geqslant\pi_i(v_i,\bar{v}_i^*) \tag{5}$$

上式表示在这种状态下, 每个人做出的都是最优的选择. 但是纳什均衡并不存在于所有博弈中, 此外, 某些场景下单个博弈也可能存在不止一个纳什均衡. 因此, 想要将非合作博弈应用于数据交易, 就要求该场景下的博弈有且仅有一个纳什均衡.

在“理性人”的前提下, 参与博弈的各方都会将对手致力于使其收入最大化作为预设. 因此上文总结的竞争市场[33,104]中的均衡就是非合作博弈均衡. 在这种市场中, 每个参与人的价格策略都是在其他参与人价格策略公布之后确定的, 以期使得自己收入最大化. 但是由于数据产品的特殊性, 传统商品边际成本等于边际收益的最大化收益方法无法用于数据市场. Li 等人[107]提出了一种交易方法, 参与交易的双方是数据消费者和数据拥有者, 该方法既能得到合适的数据交易价格, 同时还能够避免数据拥有者利益遭受损失. 非合作博弈的纳什均衡要求参与博弈双方公布自身策略, 并且在已知对方策略的前提下采取行动, 由于在实际生活中满足上述要求的情况较少, 因此非合作博弈纳什均衡是很难计算的. 同时, 由于竞争市场在数据交易时并不是普遍存在的, 因此基于非合作博弈的数据交易模式并未广泛使用.

5.2.2 基于 Stackelberg 博弈的数据交易

由于非合作博弈存在上文所述的种种缺陷, 因此本文考虑一个更加实用的情形: 一个参与人 (领导者) 先行发布自己的价格策略, 另一个参与人 (追随者) 依据领导者的策略做出相应策略选择, 并进行优化, 以得到最优的价格策略, 这种模式被称为 Stackelberg 博弈[13,108]. 在 Stackelberg 博弈中, 参与人 1 (领导者) 首先确定自己的价格策略 v_1, 参与人 2 (追随者) 在观察到 v_1 后, 确定自己的价格策略 v_2, 该博弈属于完全信息动态博弈. 由于参与人 1 (领导者) 先于参与人 2 (追随者) 行动, 不能掌握 v_2 的信息, 所以对于参与人 2 (追随者) 来说, 其价格策略是一个从 $V_1 \to V_2$ 的映射 T. Haddadi 等人[109]和 Lv 等人[110]证明了使用 Stackelberg 模型可以让参与博弈的各方实现收入最大化, 同时可以使领导者获得相比于追随者更大的收益.

在涉及具体交易市场之前, Mei 等人[85]分别针对捆绑销售和独立销售的情形, 在数据拥有者和数据平台之间设计了基于 Stackelberg 博弈的交易模式, 将数据拥有者看作领导者, 数据平台看作追随者. 讨论了在保证数据拥有者和数据平台收入最大化的情况下, 数据拥有者应该如何采取策略最大化自己的收益. Liu 等人[86]设计了一个分为两阶段的 Stackelberg 博弈用以解决数据平台和数据消费者之间的数据交易问题. 文章假定市场包含多个数据拥有者提供数据、一个数据平台和一个数据消费者购买数据, 数据平台可以获得所有交易参与人的相关信息. 在第 1 阶段, 数据拥有者按照自己的估值为数据设置初始出售价格, 同时, 数据消费者可以得到数据平台作为领导者所公布的价格策略. 第 2 阶段中, 数据消费者根据领导者所公布的策略, 选择适当策略作为自己的购买决策. 两阶段完成后, 数据平台根据数据拥有者的服务质量和数据消费者的购买意愿来决定哪位数据拥有者胜出, 并由该名数据拥有者与数据消费者进行交易.

Stackelberg 博弈模型作为在传统商品交易中广泛应用的博弈模型, 在数据市场中也有着很强的实用性. 但是需要注意的是, 由于在数据市场中数据拥有者的主体往往不甚明确, 因此需要谨慎选择数据拥有者作为领导者.

5.2.3 基于讨价还价的数据交易

讨价还价是指参与博弈的各方经过一次或多次谈判就某种物品的分配达成协议的过程. 在 Mao 等人[111]提出的定价模型中, 用 r_o 表示数据消费者为此次交易准备的最高价格, 即保留价格, 类似的, 数据拥有者的保留价格用 r_c 表示; 数据拥有者和数据消费者分别对数据报出自己的价格策略 p_o 和 p_c, 数据拥有者希望采取可以使自己期望收入 $\pi_o(p_o, r_o)$ 最大化的最优策略 p_o^*, 即: $\pi_o^*(p_o^*, r_o) \geqslant \pi_o(p_o, r_o), \forall p_o$, 同样地, 对于数据消费者来说也有最优策略 p_c^* 可以使自己期望收入 $\pi_c(p_c, r_c)$ 最大化, 即: $\pi_c^*(p_c^*, r_c) \geqslant \pi_c(p_c, r_c), \forall p_c$. 在该模型中, 如果有 $p_c^* \geqslant p_o^*$ 则以价格 $p = kp_c^* + (1-k)p_o^*, 0 \leqslant k \leqslant 1$ 成交. 最终得到了该博弈的纳什均衡解为 (p_c^*, p_o^*)[72].

在上文中已经提到, 差分隐私在数据定价中占有举足轻重的地位. Jung 等人[91]在个人数据市场提出了一个公平协商的框架, 参与交易的各方可以通过该框架使用差分隐私来确定隐私的暴露程度 ε 以及每暴露一单位的隐私所对应的价格. 框架允许数据拥有者根据自己对隐私暴露的容忍程度、数据消费者根据自己对数据需求的迫切程度和数据精准度预期以及预算分别和数据平台进行讨价还价博弈, 以确定最终成交价格, 同时保证了交易的公平性.

讨价还价适用于复杂市场环境下确定数据产品的最终价格, 其最后得到的结果是合作博弈的最终均衡状态, 因此也经常被用于资源分配等领域, 如传感器网络[111]、无线体域网[112]、频谱分配[113]. 值得注意的是, 在讨价还价博弈中, 参与交易各方的谈判往往需要耗费较长时间和较大资源, 因此讨价还价的针对具体问题的停止条件需要谨慎设计.

基于博弈论的数据交易方法着重考虑市场中参与人的决策以及互动行为对成交价格的影响, 由于非合作博弈存在纳什均衡难以计算等缺点, 因此上述 3 种方法中的 Stackelberg 博弈和讨价还价博弈是数据交易中最常用的方法. 同时, 基于博弈论的数据交易方法有着广泛的适用性, 能够不受交易数据类型和数据消费者想要进行任务的限制, 这也是其在数据市场中常用的原因之一. 但是由于基于博弈论的交易方法整体上侧重于对市场的宏观分析, 缺少对数据内在价值的考量, 所以在一些情况下并不能精确体现出数据自身的具体价值, 因此交易方法应该与定价方法进行结合. 同时, 有些复杂的交易场景难以完美建模、纳什均衡解难以计算等问题也是在设计基于博弈论的交易方法时应该考虑的现实问题.

5.3 基于拍卖的数据交易

拍卖可以看作是博弈论中不完全信息博弈的一种具体应用形式, 在传统商品和数据商品交易[114−117]中的应用十分广泛. 拍卖是一种经济驱动的方案, 其目的是通过买卖双方的竞价过程分配商品并赋予相应的价格[101]. 由于拍卖更多应用于不完全信息的环境中, 并且形式较为简单, 同时又对市场的公平性、高效性有着很好保证, 因此很适合用来解决大数据交易问题. 在拍卖中用到的一些概念总结如下[13].

1) 投标方 (bidders): 又称为投标人, 是指在拍卖过程中提交标书并打算在市场上购买商品的人. 在大数据市场中, 投标方通常是数据消费者.

2) 拍卖人 (auctioneer): 拍卖人在拍卖中担任代理人的角色, 负责拍卖流程的正常运转, 获胜者的确定, 以及进行支付和收入的分配. 在数据市场中的角色类似于数据平台, 但是不负责数据的收集.

3) 卖家 (seller): 卖家是指进行招标出售商品的所有者. 在大数据市场中, 卖家是指数据拥有者;

4) 投标方的估价 (valuations): 拍卖时, 投标方和卖家都需要对他们请求和出售的商品进行估价. 该价格可以高于或低于最终的交易价格, 是由拍卖人在拍卖过程中决定的.

5) 清算价格 (cleaning price): 在拍卖过程中, 卖家和投标方分别提出要价和出价. 要价表示卖家提出的商品售价, 出价表示投标方提出的投标价格, 即他们期望为商品支付的价格. 清算价格是指拍卖人根据收入最大化等原则确定的商品最终交易价格.

根据拍卖的形式不同, 可以将常用的拍卖分为密封拍卖、组合拍卖和双边拍卖. 典型的基于拍卖的大数据交易市场框架如图 4 所示[22].

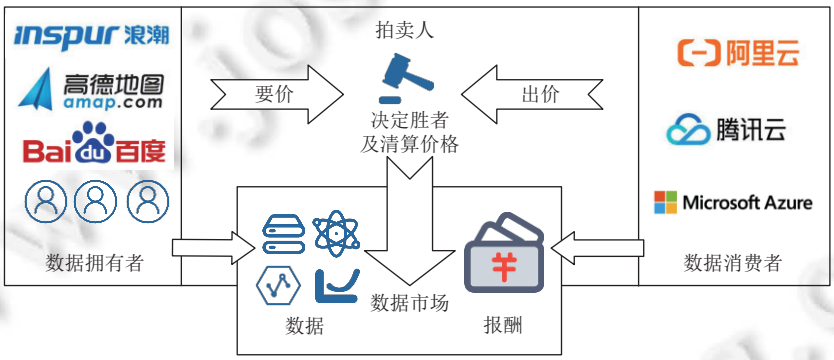


图 4 基于拍卖的大数据交易市场框架[22]

5.3.1 密封拍卖

密封拍卖主要包括第 k 价格密封拍卖和 VGC (Vickrey-Clarke-groves) 拍卖[118]. 其中第 k 价格密封拍卖主要分为第一价格密封拍卖[119]和第二价格密封拍卖[120].

在第一价格密封拍卖中, 多个投标方在密封投标信息的情况下, 以书面方式参与拍卖, 彼此之间不知道对方的出价. 拍卖的获胜者是出价最高的人, 并且以其出价水平获得该商品. 第一价格密封拍卖机制较为简单, 虽然能够保证卖家收入最大化, 但是存在着诸如投标方难以计算报价、难以保证投标方之间不结盟从而导致不公平等现象的发生等问题. 因此, 引入第二价格密封拍卖, 又称为 Vickrey 拍卖, 与前一种拍卖形式类似, 投标方彼此之间同样不知道对方的出价, 获胜者是出价最高的人, 而其仅需支付第二高的价格就可以获取商品. 在第二价格密封拍卖中, 每个投标方的占优策略是使出价等于自己对这件商品的估价[13]. 即满足真实性. 但是同第一价格密封拍卖相同, 第二价格密封拍卖难以保证投标方之间形成联盟、拍卖人和投标方之间串通等行为发生, 从而影响拍卖的公平性.

为了考虑社会福利, Ausubel 等人[121]引入了 VCG 拍卖. VCG 拍卖是 Vickrey 拍卖的扩充, 其定义如下. 现市场中存在 M 个商品表示为 $T=\{t_1,t_2,\ldots,t_M\}$, 存在 N 个投标方表示为 $B=\{b_1,b_2,\ldots,b_M\}$. 在 VCG 拍卖中, 获胜者第 i 个投标方 b_i 需要补偿其他 $N-1$ 个投标方的社会价值损失. 在拍卖中, 若投标方 b_i 对商品 t_j 的出价是最高的,

为 $v_i(t_j)$, 则其需要支付的价格为:

$$P = V_{N\setminus\{b_i\}}^{M} - V_{N\setminus\{b_i\}}^{M\setminus\{t_j\}} \tag{6}$$

其中, V_N^M 表示由 M 件商品所创造的社会价值. 在 Vickrey 拍卖中, 该值等于第二高的出价.VCG 拍卖的结果是不完全信息静态博弈的纳什均衡, 即贝叶斯纳什均衡[122]. 虽然 VCG 拍卖能够在实现社会福利最大化的同时保证拍卖的真实性, 但是在实际应用中, 也存在社会福利最大化结果难以计算, 以及收益和激励机制表现不好等问题.

5.3.2 双边拍卖

双边拍卖是数据交易市场中常见的拍卖方式之一, 也被广泛的应用于证券交易[123]、智能电网[116,124]等场景. 与上述几种拍卖方式不同的是, 在双边拍卖中, 多个投标方和卖家同时向拍卖人提交自己的出价和要价. 其交易规则为, 清算价格 p 是由拍卖人最终决定的, 当且仅当卖家要价 p_o 小于等于投标方出价 p_c 时交易才成立. 由于在双边拍卖中, 投标方和卖家的出价和要价可以分多轮进行. 为了竞争成功, 投标方每轮的出价必须越来越高, 而卖家的要价必须越来越低. 可以得到产品的清算价格图如图 5 所示[124]. 在收集了相关信息后, 拍卖人根据清算价格以及投标方向卖家支付的价格来匹配这些出价和要价[72].

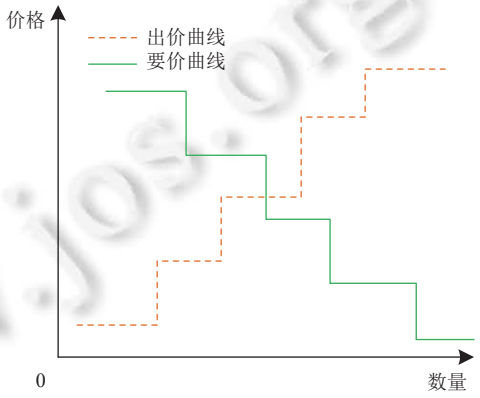


图 5 双边拍卖数据市场价格的形成[124]

5.3.3 组合拍卖

在大数据市场中, 单一类型的数据无法满足数据消费者对数据产品类型多种多样的需求, 因此一般情况下数据消费者希望购买大量组合在一起而非简单糅合的数据[72]. 因此提出了组合拍卖方法. 在组合拍卖中, 卖家提供灵活的可捆绑销售的多种数据产品的组合[125], 买家根据自身对数据的需求提出相应的出价, 而拍卖人根据出价中包含的约束条件, 并综合考虑卖家的物品分配能力, 找到最佳的组合方案, 确定最终的清算价格以及获胜的卖家[13]. 组合拍卖与密封拍卖等拍卖方式相比, 优点在于其经济效率要远高于上述的单物品拍卖方式, 同时, 可以达到买卖双方的收入最大化. 然而, 组合拍卖最终获胜卖家的计算是一个 NP 难问题, 因此想得出最优分配方案需要极大代价. Agarwal 等人[43]设计了一个交易机器学习训练数据的市场, 基于 Myerson 的拍卖理论提出了组合数据产品的拍卖方法, 满足数据消费者高效购买机器学习训练数据的需求.

Cao 等人[94]认为, 仅有一个数据收集者的数据市场是不符合实际情况的, 因此设计了具有多个数据拥有者、收集者和用户的数据市场, 数据市场参与人有着不同的效用函数. 使用迭代拍卖来协调参与人之间的数据交易, 该拍卖机制可以防止直接访问参与人的效用函数, 从而保障其隐私不被泄露, 同时实现了社会福利最大化. Cai 等人[21]认为, 现有的激励机制忽略了数据消费者之间偏好和利益冲突共存的情况, 因此提出了一种双边拍卖机制 DTPCI, 解决了数据消费者对市场偏好的多样性、数据消费者之间的利益冲突和交易各方的策略选择 3 大挑战. 该拍卖机制包含分组规则和数据交易规则, 可以实现非负的社会福利. 在去中心化交易市场的研究上, Li 等人[107]则提出了基于区块链的数据交易市场, 让拥有少量信息的中间商来对资源进行管理和分发. 提出的迭代双边拍卖交易方法满足对参与人的隐私保护, 可以实现社会福利最大化, 同时满足真实性和个人理性. 上述的基于拍卖的数据交易模

型中, 一般都是数据拥有者和数据消费者直接进行交易, 拍卖人只做中间协调, 效率较为低下. Jiao 等人[23]第一次在拍卖中引入服务提供商, 服务提供商可以收集数据拥有者的数据, 并对其进行隐私泄露的补偿, 同时自己对得到的批量数据进行整合分析, 为数据消费者提供整合后的服务, 通过贝叶斯利润最大化拍卖来交易数据. 基于拍卖的数据交易方法进一步总结见表 3.

表 3 基于拍卖的数据交易方法比较

参考文献	定价模型	市场结构			概述
		卖家	中间商	买家	
[23]	双边拍卖	多个	一个	多个	提出了基于双边拍卖的大数据市场模型. 引入服务提供商, 可以收集数据拥有者的数据, 补偿其隐私, 同时对数据进行处理, 以满足数据消费者的需求, 为其提供服务而不是原始数据. 将拍卖过程中的利润最大化问题定义为贝叶斯最优机制, 通过对其求解得到服务的最优价格和数据分配, 同时保证该方法是理性真实的
[43]	组合拍卖	多个	一个	多个	基于拍卖设计了一个双边的机器学习数据交易市场, 并给出了参数化的定义. 提出了真实的、零遗憾的机制, 用于拍卖基于 Myerson 支付函数和乘法权重算法的特定类别的组合商品. 该市场可以保证用户报价的真实性, 并对收益进行公平分配
[92]	VCG拍卖	一个或多个	一个	多个	基于VCG拍卖提出了多粒度服务组合的动态定价方案. 服务提供商对组合服务中不同粒度的服务进行投标, 并根据收到的投标, 用户决定在满足QoS 约束的同时最小化总成本的组合. 获胜者确定过程为整数规划模型, 以社会福利最大化方式确定获胜者. 该机制同时保证了真实性
[93]	适用多种拍卖规则	一个或多个	一个	多个	基于同态加密和安全网络协议设计来解决CPS中数据拍卖的隐私保护问题. 提出通用的隐私保护拍卖方案, 拍卖者和中间平台组成了不受信任的第三方交易平台. 使用同态加密和一次性填充技术对拍卖中投标信息进行伪装. 进一步通过签名验证机制增强了安全性
[94]	双边拍卖	多个	无	多个	针对具有多个数据拥有者、收集者和消费者的数据市场提出了迭代拍卖机制来对其中的数据交易问题进行协调, 可以在不访问用户隐私的同时实现社会福利最大化
[95]	第一价格密封拍卖	多个	一个	多个	开放和匿名的在线环境可能导致拍卖数据的价格无法达到公平真实的水平. 提出了第一个基于智能合约的反串通数据拍卖机制. 该机制使得互不信任、理性的买卖双方在没有可信第三方的情况下安全参与数据拍卖. 智能合约中设计的数据拍卖机制, 可以有效防止串通, 实现数据拍卖的公平性和真实性

拍卖方法作为博弈论中不完全信息博弈的一部分, 在传统商品市场和数据市场中都有广泛的应用. 上文中所提到的 3 种拍卖方式各有侧重: 以 VCG 拍卖为代表的密封拍卖侧重于保证交易的公平性和真实性, 但是可能存在减少卖家收益、多个数据产品需要多次拍卖等缺点; 组合拍卖则侧重于提供多种数据产品灵活捆绑销售的交易方式; 双边拍卖中的拍卖人则能够以中介的身份在买卖双方之间进行协调, 大大增加了数据交易时买卖双方的沟通效率. 但是由于拍卖时为了保证真实性, 往往需要投标方提交真实信息, 因此存在泄露投标方隐私的风险. 此外, 虽然拍卖的适用性较强, 但拍卖机制设计、如何设立可信的第三方拍卖平台等问题也是使用基于拍卖的数据交易方法时所必须要考虑的.

6 相关工作

数据定价与交易涉及数据管理、数据库、经济学、深度学习和人工智能等多个研究方向的内容. 已经存在一些文章从各自的角度对数据定价进行介绍. 张小伟等人[13]对经济学中适用于数据定价的理论和方法进行了综述. 类似地, Pei 等人[14]从经济学角度对数据定价进行了完整的叙述, 研究了数据定价的动机. 文章总结了数据定价时

需要考虑的基本内容, 分为版本控制、真实性、收入最大化、公平性、无套利、隐私保护和计算高效, 并基于上述内容, 分别叙述了现存的数据定价模型. 贯穿于整篇文章的是数据产品和数字产品的对比. 同时, 文章指出了数据定价现在存在的一些挑战, 包括数据供应链、数据价值评估等方面的问题. 刘枂等人[15]也对大数据定价方法进行了综述, 并将其分为成本导向、市场导向、需求导向、利润导向以及基于生命周期定价的 5 种定价类型, 对比了成本法、协议定价、市场法、收益法、基于质量以及基于查询的定价 6 种主流定价方法的优劣势, 并通过大数据定价流程展现了不同定价方法的各自特点. 同样, 蔡莉等人[16]也对数据定价模型进行了综述, 并将其分为基于数据质量的定价、基于信息熵的定价、基于查询的定价、基于博弈论的定价和基于机器学习的定价, 并对上述几种定价方法的优劣进行了分析. 文章还阐述了现阶段数据定价存在的挑战分别体现在: 价值评估、交易规则和隐私保护 3 处. 除此之外, Fricker 等人[126]对数据市场中的定价问题进行了介绍. 文章将数据市场分为了单卖家和多卖家两种类型, 分别总结了定价方法能够实现的目标, 如社会福利最大化、收入最大化、一致性和公平性等. 并且对定价时考虑的数据价值维度进行了阐述.

上述综述对数据定价领域相关内容进行了详细的介绍, 但都存在一些不足之处. 首先, 文献 [13,14] 对数据定价的介绍偏向于经济学领域, 强调数据定价中应该遵循的各种规则, 但是没能对现存定价方法进行完整分类. 刘枂等人[15]则更偏向于社会科学领域, 更多讨论了数据定价中存在的制度性和框架性问题, 并基于此对定价方法进行了分类, 但没能对定价方法的具体细节进行研究. 蔡莉等人[16]和 Fricker 等人[126]弥补了上述不足, 对数据定价策略和方法进行了详细分类, 同时对定价过程也进行了较为全面的介绍. 但是, 上述文章存在的共同问题是, 虽然对数据定价过程的涵盖较为全面, 但是对于和数据定价密不可分的数据交易部分却介绍甚少. 因此, 除了对数据定价过程中需要遵循的准则以及数据定价方法进行全面综述外, 本文将数据交易市场作为重点, 根据大数据在数据交易市场中的流通过程, 将其生命周期分为数据收集与集成、数据管理与分析、数据定价和数据交易四个环节, 详细介绍了每个环节需要进行的工作、存在的挑战以及相关解决方案.

7 总结与展望

近年来, 由于大数据产业的快速发展, 数据已经成为炙手可热的战略资源. 大数据对于个人和组织来说都具有巨大的价值. 但是由于拥有数据收集和分析的能力的公司相对较少, 而这些公司往往倾向于将数据保留在自己的数据中心, 这便形成了数据孤岛. 数据孤岛的存在严重妨碍了大数据产业的健康发展. 因此数据共享的呼声越来越高. 数据共享中研究最为广泛的方法便是数据交易. 本文对大数据在数据交易市场的流通环节进行了总结, 将其生命周期分为了数据收集与集成、数据管理与分析、数据定价和数据交易 4 部分, 介绍了每个部分存在的挑战, 为每个部分中的相关工作进行了分类和总结. 由于数据交易领域对前两个部分的研究相对较少, 因此本文借鉴了数据管理方向中的相关工作, 总结了其中适用于数据交易的方法. 对于数据定价, 本文对数据定价的相关方法进行了总结, 并依据其思路的不同, 将其分为了基于任务的定价、基于价值的定价和基于经济学的定价. 这 3 类定价方法并不是简单的互斥关系, 而是存在着相互交叉的领域: 比如基于价值的定价方法就作为确定价格基准的途径, 在基于任务的定价方法中时有出现. 与数据定价互补的任务是数据交易. 由于数据交易市场在确定了数据价格之后, 还要考虑交易参与人对交易过程和数据价格产生的影响. 这就是数据交易机制设计的问题. 本文首先介绍了数据交易市场的不同分类, 并以博弈论和拍卖为例, 介绍了数据市场中交易机制的设计方法, 总结了这些方法的使用场景和不足之处. 本文的目的是给数据定价与交易领域涉及的问题做全面的总结和综述, 希望本文可以为新进入该领域的学者提供一个完整的了解.

接下来基于本文对数据定价与交易方面研究进展的梳理以及当前工作的不足之处, 针对数据在大数据交易市场的生命周期, 给出未来可以研究的方面.

(1) 构建完整的数据供应机制

数据供应是大数据交易能够进行的首要保证. 但是由于数据来源和种类丰富多样, 导致数据收集方式和集成方式难以形成统一标准, 加之数据高时效性和复制代价极低的特点, 对高效进行数据收集与集成提出了更大的挑战. 因此要发展生态可持续的数据交易市场, 就必须构建完整的数据供应机制[14]. 数据供应机制可以将数据拥有

者和数据平台连接起来, 为数据收集与集成提供标准化流程. 同时, 需要在数据供应机制中添加反馈功能, 使得数据收集者可以及时将数据出售情况反馈给数据拥有者, 使数据供应和消费能够得有效的协调与平衡.

(2) 建立高效的市场调查分析机制

在构建了完整的数据供应机制后, 虽然可以得到标准化格式的数据产品, 但为了给接下来的数据定价提供依据, 仍旧需要对数据产品适用场景和用户偏好进行分析, 建立高效的市场调查分析机制. 该机制应该满足如下两个方面的功能.

• 分析数据适用领域. 数据产品必须在特定的领域才能发挥作用, 特定领域中的机制, 法规和约束可能在某些方面对数据产品价值产生影响[14]. 因此必须分析该类型数据产品的适用领域, 得出数据产品对数据消费者的效用水平, 从而为定价提供依据.

• 调查市场偏好. 针对市场偏好进行差异化的定价是大数据市场化发展的必然趋势[15]. 文献 [44] 认为在数据定价前需要进行市场调查, 以确定特定领域的数据消费者对某种质量水平数据的偏好程度. 因此应该在数据定价前对数据交易市场中的潜在客户进行统一调查, 对其需求和购买意愿进行分析, 为版本控制和数据定价提供重要参考.

(3) 构建数据定价理论框架

由于不同的参与者对数据产品有着不同的预期和评价, 所以导致现有的单一指标数据定价方法都存在自身的局限性, 难以完全满足各方的需求. 因此应该结合数据的 4V 特性, 针对现有定价方法的不足, 提出完善的数据定价理论框架, 该框架应该包括如下内容.

• 统一的价值评价体系. 数据定价理论框架的首要内容就是统一价值评价体系, 寻找能让各方都满足的价值度量技术. 该价值评价体系应该以数据质量为基础, 结合数据消费者和数据平台在整合数据产品上的花费, 考虑数据消费者的效用指标, 还要以历史成交价格为参照, 综合评估其他能够影响数据价格的因素, 构建出统一的、可解释的、客观的数据价值评价体系.

• 动态定价机制. 现存的大多数定价方法都是静态定价, 但是由于数据有着极强的时效性, 数据消费者的需求也会随着时间的变化而变化. 因此数据价格也应该随之变动. 为了使数据价格更加贴合实际, 应该建立数据价格与时间的函数关系模型, 捕捉和监测数据内容和数据价格的变化方向, 探索动态定价机制.

• 定价模型联系实践. 已经提出的大部分数据定价方法都是假定了较为理想化的定价场景, 并在其上研究定价的理论模型, 很少有研究能将理论模型与现实生活中的数据定价实践相结合. 因此将模型推向市场时可能难以完全满足用户需求. 应该对实践中的定价场景和规则进行建模, 着力研究数据定价在实践中的效果, 以对其进行改进.

(4) 完善数据交易机制

数据交易是数据定价的互补过程, 重点关注数据在市场上流通时市场类型, 机制设计和参与人行为对数据价格的影响. 但是由于数据市场刚刚起步, 数据交易机制尚处于初级阶段, 仍然存在着很多不足. 数据交易机制的设计直接影响数据拥有者的出售意愿和数据消费者的购买意愿. 因此需要对当前的数据交易机制进行完善. 主要分为以下几个方面.

• 保护隐私和版权. 由于上文所叙述两种数据交易方法涉及公开策略 (博弈论) 或进行投标 (拍卖) 的过程, 同时数据本身就包含隐私成分, 因此容易导致隐私的泄露. 此外, 由于数据的可复制性, 出售的数据会以一个极低的代价传播出去, 损害数据拥有者对数据的版权. 上述问题都会导致过度的数据分享, 不仅会降低数据价格, 还会使得数据拥有者交易欲望变低, 影响数据市场发展. 因此在数据交易时必须研究相应的隐私和版权保护机制. 未来可以通过制定隐私版权保护规则如缴纳保证金、设置审查规范和惩罚措施等制度方法和数据脱敏、数据加密等技术方法对隐私和版权进行保护.

• 创造公平、真实的交易环境. 现存文章虽然对交易的公平性和真实性进行过一定的探索, 但是仍旧存在片面, 顾此失彼的缺点. 因此未来需要分析每次交易时所处的市场环境, 设计相应的交易机制, 确保交易价格对参与人来说都是公平的. 同时还应该规范交易参与人的行为, 以确保其交易时可以如实上报自己的花费或收入, 保证交易环境的真实性.

• 建立历史记录反馈机制. 为每次成交或交易失败的数据建立元数据库, 存储其数据类型、数据容量、质量水

平、成交价格和客户类型等交易元数据, 根据元数据对交易结果产生原因进行分析. 并设置完整的反馈通道, 保证前序步骤可以根据交易结果调整策略, 以促进数据市场的健康发展.

References:

[1] China Academy of Information and Communications Technology (CAICT). White paper of big data (2020). 2020 (in Chinese). http://www.caict.ac.cn/kxyj/qwfb/bps/202012/t20201228_367162.htm

[2] Huang XR. From complexity science to big data technique. Journal of Changsha University of Science & Technology (Social Science), 2014, 29(2): 5–9 (in Chinese with English abstract). [doi: 10.3969/j.issn.1672-934X.2014.02.001]

[3] The Economists. The world's most valuable resource is no longer oil, but data. 2017. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

[4] Data.gov. https://www.data.gov

[5] Ubaldi B. Open government data: Towards empirical analysis of open government data initiatives. 2013. https://www.oecd-ilibrary.org/governance/open-government-data_5k46bj4f03s7-en

[6] Delta lake. https://delta.io

[7] Global big data exchange. https://www.gzdex.com.cn/

[8] Donghu big data. http://www.chinadatatrading.com/

[9] Beijing International Data Exchange. https://www.bjidex.com/

[10] Dawex: Sell, buy and share data. https://www.dawex.com/en/

[11] Xignite. https://www.xignite.com/

[12] WorldQuant. https://www.worldquant.com/data-exchange/

[13] Zhang XW, Jiang D, Yuan Y. A survey of game theory and auction-based data pricing. Big Data Research, 2021, 7(4): 61–79 (in Chinese with English abstract).

[14] Pei J. A survey on data pricing: From economics to data science. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(10): 4586–4608. [doi: 10.1109/TKDE.2020.3045927]

[15] Liu N, Hao XJ, Chen YH. A review and comparative analysis of domestic and foreign research on big data pricing methods. Big Data Research, 2021, 7(6): 89–102 (in Chinese with English abstract). [doi: 10.11959/j.issn.2096-0271.2021063]

[16] Cai L, Huang ZH, Liang Y, Zhu YY. Survey of data pricing. Journal of Frontiers of Computer Science and Technology, 2021, 15(9): 1595–1606 (in Chinese with English abstract). [doi: 10.3778/j.issn.1673-9418.2103069]

[17] Liu ZY. Analysis on pricing of big data. Documentation, Information & Knowledge, 2016, (1): 57–64 (in Chinese with English abstract). [doi: 10.13366/j.dik.2016.01.057]

[18] Zhang M, Arafa A, Huang JW, Poor HV. Pricing fresh data. IEEE Journal on Selected Areas in Communications, 2021, 39(5): 1211–1225. [doi: 10.1109/JSAC.2021.3065088]

[19] Tang SS, Liu YT. China's big data transaction urgently needs a breakthrough. China Development Observation, 2016, (13): 19–21 (in Chinese with English abstract). [doi: 10.3969/j.issn.1673-033X.2016.13.007]

[20] Hu YL. Research on status quo and pricing issue of big data trade. Prices Monthly, 2017, (12): 16–19 (in Chinese with English abstract). [doi: 10.14076/j.issn.1006-2025.2017.12.04]

[21] Cai H, Zhu YM, Li J, Yu JD. Double auction for a data trading market with preferences and conflicts of interest. The Computer Journal, 2019, 62(10): 1490–1504. [doi: 10.1093/comjnl/bxz025]

[22] An D, Yang QY, Yu W, Li DH, Zhang Y, Zhao W. Towards truthful auction for big data trading. In: Proc. of the 36th IEEE Int'l Performance Computing and Communications Conf. San Diego: IEEE, 2017. 1–7. [doi: 10.1109/PCCC.2017.8280501]

[23] Jiao YT, Wang P, Niyato D, Abu Alsheikh M, Feng SH. Profit maximization auction and data management in big data markets. In: Proc. of the 2017 IEEE Wireless Communications and Networking Conf. San Francisco: IEEE, 2017. 1–6. [doi: 10.1109/WCNC.2017.7925760]

[24] Xiong ZH, Niyato D, Wang P, Han Z, Zhang Y. Dynamic pricing for revenue maximization in mobile social data market with network effects. IEEE Trans. on Wireless Communications, 2020, 19(3): 1722–1737. [doi: 10.1109/TWC.2019.2957092]

[25] Khokhar RH, Iqbal F, Fung BCM, Bentahar J. Enabling secure trustworthiness assessment and privacy protection in integrating data for trading person-specific information. IEEE Trans. on Engineering Management, 2021, 68(1): 149–169. [doi: 10.1109/TEM.2020.2974210]

[26] Delgado-Segura S, Pérez-Solà C, Navarro-Arribas G, Herrera-Joancomartí J. A fair protocol for data trading based on Bitcoin

transactions. Future Generation Computer Systems, 2020, 107: 832–840. [doi: 10.1016/j.future.2017.08.021]

[27] Koutris P, Upadhyaya P, Balazinska M, Howe B, Suciu D. Toward practical query pricing with QueryMarket. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2013. 613–624. [doi: 10.1145/2463676.2465335]

[28] Shapley LS. A value for n-person games. Classics in Game Theory, 1997: 69.

[29] Jia RX, Dao D, Wang BX, Hubis FA, Gürel NM, Li B, Zhang C, Spanos CJ, Song D. Efficient task-specific data valuation for nearest neighbor algorithms. Proc. of the VLDB Endowment, 2019, 12(11): 1610–1623. [doi: 10.14778/3342263.3342637]

[30] Balazinska M, Howe B, Suciu D. Data markets in the cloud: An opportunity for the database community. Proc. of the VLDB Endowment, 2011, 4(12): 1482–1485. [doi: 10.14778/3402755.3402801]

[31] Koutris P, Upadhyaya P, Balazinska M, Howe B, Suciu D. Query-based data pricing. In: Proc. of the 31st ACM SIGMOD-SIGACT-SIGAI Symp. on Principles of Database Systems. Scottsdale: ACM, 2012. 167–178. [doi: 10.1145/2213556.2213582]

[32] Li C, Miklau G. Pricing aggregate queries in a data marketplace. In: Proc. of the 15th Int'l Workshop on the Web and Databases. Scottsdale: ACM, 2012. 19–24.

[33] Li C, Li DY, Miklau G, Suciu D. A theory of pricing private data. ACM Trans. on Database Systems, 2014, 39(4): 34. [doi: 10.1145/2691190.2691191]

[34] Lin BR, Kifer D. On arbitrage-free pricing for general data queries. Proc. of the VLDB Endowment, 2014, 7(9): 757–768. [doi: 10.14778/2732939.2732948]

[35] Deep S, Koutris P. The design of arbitrage-free data pricing schemes. In: Proc. of the 20th Int'l Conf. on Database Theory (ICDT 2017). Berlin, Springer. 2017. 1–12, 18.

[36] Deep S, Koutris P. QIRANA: A framework for scalable query pricing. In: Proc. of the 2017 ACM Int'l Conf. on Management of Data. Chicago: ACM, 2017. 699–713. [doi: 10.1145/3035918.3064017]

[37] Roughgarden T. Twenty Lectures on Algorithmic Game Theory. Cambridge: Cambridge University Press, 2016. 111–112. [doi: 10.1017/CBO9781316779309]

[38] Myerson RB. Optimal auction design. Mathematics of Operations Research, 1981, 6(1): 58–73.

[39] Niu CY, Zheng ZZ, Wu F, Tang SJ, Chen GH. Online pricing with reserve price constraint for personal data markets. In: Proc. of the 36th IEEE Int'l Conf. on Data Engineering. Dallas: IEEE, 2020. 1978–1981. [doi: 10.1109/ICDE48307.2020.00218]

[40] Babaioff M, Immorlica N, Lucier B, Weinberg SM. A simple and approximately optimal mechanism for an additive buyer. Journal of the ACM, 2020, 67(4): 24. [doi: 10.1145/3398745]

[41] Cai Y, Zhao MF. Simple mechanisms for subadditive buyers via duality. In: Proc. of the 49th Annual ACM SIGACT Symp. on Theory of Computing. Montreal: ACM, 2017. 170–183. [doi: 10.1145/3055399.3055465]

[42] Chawla S, Deep S, Koutris P, Teng YF. Revenue maximization for query pricing. Proc. of the VLDB Endowment, 2019, 13(1): 1–14. [doi: 10.14778/3357377.3357378]

[43] Agarwal A, Dahleh M, Sarkar T. A marketplace for data: An algorithmic solution. In: Proc. of the 2019 ACM Conf. on Economics and Computation. Phoenix: ACM, 2019. 701–726. [doi: 10.1145/3328526.3329589]

[44] Chen LJ, Koutris P, Kumar A. Towards model-based pricing for machine learning in a data marketplace. In: Proc. of the 2019 Int'l Conf. on Management of Data. Amsterdam: ACM, 2019. 1535–1552. [doi: 10.1145/3299869.3300078]

[45] Ghosh A, Roth A. Selling privacy at auction. In: Proc. of the 12th ACM Conf. on Electronic Commerce. San Jose: ACM, 2011. 199–208. [doi: 10.1145/1993574.1993605]

[46] Fernandez RC, Subramaniam P, Franklin MJ. Data market platforms: Trading data assets to solve data problems. Proc. of the VLDB Endowment, 2020, 13(12): 1933–1947. [doi: 10.14778/3407790.3407800]

[47] Riederer C, Erramilli V, Chaintreau A, Krishnamurthy B, Rodriguez P. For sale : your data: By : you. In: Proc. of the 10th ACM Workshop on Hot Topics in Networks. Cambridge: ACM, 2011. 13. [doi: 10.1145/2070562.2070575]

[48] Apache flume. http://flume.apache.org/

[49] Fluentd. https://www.fluentd.org/

[50] Logstash. https://www.elastic.co/cn/logstash/

[51] Splunk. https://www.splunk.com/en_us/download/universal-forwarder.html

[52] Dalvi N, Kumar R, Soliman M. Automatic wrappers for large scale web extraction. Proc. of the VLDB Endowment, 2011, 4(4): 219–230. [doi: 10.14778/1938545.1938547]

[53] Bohannon P, Dalvi N, Filmus Y, Jacoby N, Keerthi S, Kirpal A. Automatic Web-scale information extraction. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. Scottsdale: ACM, 2012. 609–612. [doi: 10.1145/2213836.2213912]

[54] Cafarella MJ, Halevy A, Wang DZ, Wu E, Zhang Y. WebTables: Exploring the power of tables on the web. Proc. of the VLDB

Endowment, 2008, 1(1): 538–549. [doi: 10.14778/1453856.1453916]

[55] Cafarella M, Halevy A, Lee H, Madhavan J, Yu C, Wang DZ, Wu E. Ten years of webtables. Proc. of the VLDB Endowment, 2018, 11(12): 2140–2149. [doi: 10.14778/3229863.3240492]

[56] Elmeleegy H, Madhavan J, Halevy A. Harvesting relational tables from lists on the Web. Proc. of the VLDB Endowment, 2009, 2(1): 1078–1089. [doi: 10.14778/1687627.1687749]

[57] Chu X, He YY, Chakrabarti K, Ganjam K. TEGRA: Table extraction by global record alignment. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. Melbourne: ACM, 2015. 1713–1728. [doi: 10.1145/2723372.2723725]

[58] Hui JY, Li LL, Zhang ZG. Integration of big data: A survey. In: Proc. of the 4th Int'l Conf. of Pioneering Computer Scientists, Engineers and Educators. Zhengzhou: Springer, 2018. 101–121. [doi: 10.1007/978-981-13-2203-7_9]

[59] Madhavan J, Bernstein PA, Rahm E. Generic schema matching with cupid. In: Proc. of the 27th Int'l Conf. on Very Large Data Bases. Roma: Morgan Kaufmann Publishers Inc., 2001. 49–58.

[60] Zhu C, Cao J. Summary and prospect on entity resolution. Computer Science, 2015, 42(3): 8–12, 18 (in Chinese with English abstract). [doi: 10.11896/j.issn.1002-137X.2015.3.002]

[61] Yin XX, Han JW, Yu PS. Truth discovery with multiple conflicting information providers on the Web. IEEE Trans. on Knowledge and Data Engineering, 2008, 20(6): 796–808. [doi: 10.1109/TKDE.2007.190745]

[62] Domshlak C, Gal A, Roitman H. Rank aggregation for automatic schema matching. IEEE Trans. on Knowledge and Data Engineering, 2007, 19(4): 538–553. [doi: 10.1109/TKDE.2007.1010]

[63] Xie ZZ, Liu QZ, Bao ZF. Sifting truths from multiple low-quality data sources. In: Proc. of the 1st Asia-Pacific Web (APWeb) and Web-age Information Management (WAIM) Joint Conf. on Web and Big Data. Beijing: Springer, 2017. 74–81. [doi: 10.1007/978-3-319-63579-8_7]

[64] Rekatsinas T, Joglekar M, Garcia-Molina H, Parameswaran A, Ré C. SLiMFast: Guaranteed results for data fusion and source reliability. In: Proc. of the 2017 ACM Int'l Conf. on Management of Data. Chicago: ACM, 2017. 1399–1414. [doi: 10.1145/3035918.3035951]

[65] Li YL, Li Q, Gao J, Su L, Zhao B, Fan W, Han JW. Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery. IEEE Trans. on Knowledge and Data Engineering, 2016, 28(8): 1986–1999. [doi: 10.1109/TKDE.2016.2559481]

[66] Li Q, Li YL, Gao J, Zhao B, Fan W, Han JW. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. Snowbird: ACM, 2014. 1187–1198. [doi: 10.1145/2588555.2610509]

[67] Zhang HT, Li Q, Ma FL, Xiao HP, Li YL, Gao J, Su L. Influence-aware truth discovery. In: Proc. of the 25th ACM Int'l on Conf. on Information and Knowledge Management. Indianapolis: ACM, 2016. 851–860. [doi: 10.1145/2983323.2983785]

[68] Grover M, Malaska T, Seidman J, Shapira G. Hadoop Application Architectures: Designing Real-world Big Data Applications. Sebastopol: O'Reilly Media Inc., 2015.

[69] Stonebraker M. The case for polystores. 2015. http://wp.sigmod.org/?p=1629

[70] Hai RH, Geisler S, Quix C. Constance: An intelligent data lake system. In: Proc. of the 2016 Int'l Conf. on Management of Data. San Francisco: ACM, 2016. 2097–2100. [doi: 10.1145/2882903.2899389]

[71] Liu ZY, Hacigümüs H. Online optimization and fair costing for dynamic data sharing in a cloud data market. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. Snowbird: ACM, 2014. 1359–1370. [doi: 10.1145/2588555.2593679]

[72] Liang F, Yu W, An D, Yang QY, Fu XW, Zhao W. A survey on big data market: Pricing, trading and protection. IEEE Access, 2018, 6: 15132–15154. [doi: 10.1109/ACCESS.2018.2806881]

[73] Upadhyaya P, Balazinska M, Suciu D. Price-optimal querying with data APIs. Proc. of the VLDB Endowment, 2016, 9(14): 1695–1706. [doi: 10.14778/3007328.3007335]

[74] Wang XW, Wei XH, Liu YY, Gao S. On pricing approximate queries. Information Sciences, 2018, 453: 198–215. [doi: 10.1016/j.ins.2018.04.036]

[75] Nget R, Cao Y, Yoshikawa M. How to balance privacy and money through pricing mechanism in personal data market. In: Proc. of the SIGIR 2017 Workshop on eCommerce Co-located with the 40th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tokyo: CEUR-WS.org, 2019.

[76] Tang RM, Wu HY, Bao ZF, Bressan S, Valduriez P. The price is right-models and algorithms for pricing data. In: Proc. of the 24th Int'l Conf. on Database and Expert Systems Applications. Prague: Springer, 2013. 380–394. [doi: 10.1007/978-3-642-40173-2_31]

[77] Shen YC, Guo B, Shen Y, Duan XL, Dong XQ, Zhang H. A pricing model for big personal data. Tsinghua Science and Technology, 2016, 21(5): 482–490. [doi: 10.1109/TST.2016.7590317]

[78] Chen LJ, Koutris P, Kumar A. Model-based pricing: Do not pay for more than what you learn! In: Proc. of the 1st Workshop on Data Management for End-to-end Machine Learning. Chicago: ACM, 2017. 1. [doi: 10.1145/3076246.3076250]

[79] Zhang MX, Beltrán F, Liu JM. Selling data at an auction under privacy constraints. In: Proc. of the 36th Conf. on Uncertainty in Artificial Intelligence. San Diego: AUAI Press, 2020. 669–678.

[80] Stahl F, Vossen G. Data quality scores for pricing on data marketplaces. In: Proc. of the 8th Asian Conf. on Intelligent Information and Database Systems. Da Nang: Springer, 2016. 215–224. [doi: 10.1007/978-3-662-49381-6_21]

[81] Yu HF, Zhang MX. Data pricing strategy based on data quality. Computers & Industrial Engineering, 2017, 112: 1–10. [doi: 10.1016/j.cie.2017.08.008]

[82] Zhang D, Wang HZ, Ding XO, Zhang YC, Li JZ, Gao H. On the fairness of quality-based data markets. arXiv:1808.01624, 2018.

[83] Yang J, Zhao CC, Xing CX. Big data market optimization pricing model based on data quality. Complexity, 2019, 2019: 5964068. [doi: 10.1155/2019/5964068]

[84] Stahl F, Vossen G. Fair knapsack pricing for data marketplaces. In: Proc. of the 20th East European Conf. on Advances in Databases and Information Systems. Prague: Springer, 2016. 46–59. [doi: 10.1007/978-3-319-44039-2_4]

[85] Mei LJ, Li W, Nie K. Pricing decision analysis for information services of the Internet of Things based on Stackelberg game. In: Zhang ZJ, Zhang RT, Zhang JL, eds. LISS 2012. Berlin, Heidelberg: Springer, 2013. 1097–1104. [doi: 10.1007/978-3-642-32054-5_155]

[86] Liu K, Qiu XY, Chen WH, Chen X, Zheng ZB. Optimal pricing mechanism for data market in Blockchain-enhanced internet of things. IEEE Internet of Things Journal, 2019, 6(6): 9748–9761. [doi: 10.1109/JIOT.2019.2931370]

[87] Xu CZ, Zhu K, Yi CY, Wang R. Data pricing for Blockchain-based car sharing: A Stackelberg game approach. In: Proc. of 2020 IEEE Global Communications Conf. Taipei: IEEE, 2020. 1–5. [doi: 10.1109/GLOBECOM42002.2020.9322221]

[88] Kang X, Zhang R, Motani M. Price-based resource allocation for spectrum-sharing femtocell networks: A Stackelberg game approach. IEEE Journal on Selected Areas in Communications, 2012, 30(3): 538–549. [doi: 10.1109/JSAC.2012.120404]

[89] Yao HP, Mai TL, Wang JJ, Ji Z, Jiang CX, Qian Y. Resource trading in Blockchain-based industrial Internet of Things. IEEE Trans. on Industrial Informatics, 2019, 15(6): 3602–3609. [doi: 10.1109/TII.2019.2902563]

[90] Rawat DB, Shetty S, Xin CS. Stackelberg-game-based dynamic spectrum access in heterogeneous wireless systems. IEEE Systems Journal, 2016, 10(4): 1494–1504. [doi: 10.1109/JSYST.2014.2347048]

[91] Jung K, Park S. Privacy bargaining with fairness: Privacy-price negotiation system for applying differential privacy in data market environments. In: Proc. of the 2019 IEEE Int'l Conf. on Big Data. Los Angeles: IEEE, 2019. 1389–1394. [doi: 10.1109/BigData47090.2019.9006101]

[92] Wu QW, Zhou MC, Zhu QS, Xia YN. VCG auction-based dynamic pricing for multigranularity service composition. IEEE Trans. on Automation Science and Engineering, 2018, 15(2): 796–805. [doi: 10.1109/TASE.2017.2695123]

[93] Gao WC, Yu W, Liang F, Hatcher WG, Lu C. Privacy-preserving auction for big data trading using homomorphic encryption. IEEE Trans. on Network Science and Engineering, 2020, 7(2): 776–791. [doi: 10.1109/TNSE.2018.2846736]

[94] Cao XY, Chen Y, Ray Liu KJ. Data trading with multiple owners, collectors, and users: An iterative auction mechanism. IEEE Trans. on Signal and Information Processing over Networks, 2017, 3(2): 268–281. [doi: 10.1109/TSIPN.2017.2668144]

[95] Xiong W, Xiong L. Anti-collusion data auction mechanism based on smart contract. Information Sciences, 2021, 555: 386–409. [doi: 10.1016/j.ins.2020.10.053]

[96] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 1996, 12(4): 5–33. [doi: 10.1080/07421222.1996.11518099]

[97] Naumann F. Quality-Driven Query Answering for Integrated Information Systems. Berlin, Heidelberg: Springer, 2002. [doi: 10.1007/3-540-45921-9]

[98] Nagle TT, Hogan JE, Zale J. The Strategy and Tactics of Pricing. 5th ed., Upper Saddle River: Prentice Hall, 2010.

[99] Fama EF, French KR. Commodity futures prices: Some evidence on forecast power, premiums, and the theory of storage. In: Malliaris AG, Ziemba WT, eds. The World Scientific Handbook of Futures Markets. Singapore: World Scientific, 2015. 79–102.

[100] Sen A. Rational fools: A critique of the behavioral foundations of economic theory. Philosophy and Public Affairs, 1977, 6(4): 317–344.

[101] McAfee RP. A dominant strategy double auction. Journal of Economic Theory, 1992, 56(2): 434–450. [doi: 10.1016/0022-0531(92)90091-U]

[102] Muschalle A, Stahl F, Löser A, Vossen G. Pricing approaches for data markets. In: Proc. of the 6th Int'l Workshop on Business Intelligence for the Real-time Enterprise. Istanbul: Springer, 2012. 129–144. [doi: 10.1007/978-3-642-39872-8_10]

[103] Niyato D, Alsheikh MA, Wang P, Kim DI, Han Z. Market model and optimal pricing scheme of big data and Internet of Things (IoT). In: Proc. of the 2016 IEEE Int'l Conf. on Communications. Kuala Lumpur: IEEE, 2016. 31–36. [doi: 10.1109/ICC.2016.7510922]

[104] Niyato D, Hoang DT, Luong NC, Wang P, Kim DI, Han Z. Smart data pricing models for the Internet of Things: A bundling strategy approach. IEEE Network, 2016, 30(2): 18–25. [doi: 10.1109/MNET.2016.7437020]

[105] Nash J. Non-cooperative games. Annals of mathematics, 1951, 54(2): 286–295. [doi: 10.2307/1969529]

[106] Luong NC, Hoang DT, Wang P, Niyato D, Kim DI, Han Z. Data collection and wireless communication in internet of things (IoT) using economic analysis and pricing models: A survey. IEEE Communications Surveys & Tutorials, 2016, 18(4): 2546–2590. [doi: 10.1109/COMST.2016.2582841]

[107] Li ZN, Yang ZY, Xie SL. Computing resource trading for edge-cloud-assisted Internet of Things. IEEE Trans. on Industrial Informatics, 2019, 15(6): 3661–3669. [doi: 10.1109/TII.2019.2897364]

[108] Simaan M, Cruz JB. On the Stackelberg strategy in nonzero-sum games. Journal of Optimization Theory and Applications, 1973, 11(5): 533–555. [doi: 10.1007/BF00935665]

[109] Haddadi S, Ghasemi A. Pricing-based Stackelberg game for spectrum trading in self-organised heterogeneous networks. IET Communicatons, 2016, 10(11): 1374–1383. [doi: 10.1049/iet-com.2016.0033]

[110] Lv XY, Zhang RT, Yue JJ. Competition and cooperation between participants of the internet of things industry value chain. International Journal on Advances in Information Sciences and Service Sciences, 2012, 4(11): 406–412. [doi: 10.4156/aiss.vol4.issue11.50]

[111] Mao YX, Cheng T, Zhao HY, Shen N. A strategic bargaining game for a spectrum sharing scheme in cognitive radio-based heterogeneous wireless sensor networks. Sensors, 2017, 17(12): 2737. [doi: 10.3390/s17122737]

[112] Moulik S, Misra S, Gaurav A. Cost-effective mapping between wireless body area networks and cloud service providers based on multi-stage bargaining. IEEE Trans. on Mobile Computing, 2017, 16(6): 1573–1586. [doi: 10.1109/TMC.2016.2571286]

[113] Azimi SM, Manshaei MH, Hendessi F. Cooperative primary-secondary dynamic spectrum leasing game via decentralized bargaining. Wireless Networks, 2016, 22(3): 755–764. [doi: 10.1007/s11276-015-0999-8]

[114] An D, Yang QY, Yu W, Yang XY, Fu XW, Zhao W. Sto2Auc: A stochastic optimal bidding strategy for microgrids. IEEE Internet of Things Journal, 2017, 4(6): 2260–2274. [doi: 10.1109/JIOT.2017.2764879]

[115] Iosifidis G, Gao L, Huang JW, Tassiulas L. An iterative double auction for mobile data offloading. In: Proc. of the 11th Int'l Symp. and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks. Tsukuba: IEEE, 2013. 154–161.

[116] An D, Yang QY, Yu W, Yang XY, Fu XW, Zhao W. SODA: Strategy-proof online double auction scheme for multimicrogrids bidding. IEEE Trans. on Systems, Man, and Cybernetics:Systems, 2018, 48(7): 1177–1190. [doi: 10.1109/TSMC.2017.2651072]

[117] Das SR, Sundaram RK. Auction theory: A summary with applications to treasury markets. Working Paper 5873, Cambridge: National Bureau of Economic Research, 1997. [doi: 10.3386/w5873]

[118] Nisan N, Ronen A. Computationally feasible VCG mechanisms. Journal of Artificial Intelligence Research, 2007, 29: 19–47. [doi: 10.1613/jair.2046]

[119] Kirchkamp O, Poen E, Reiß JP. Outside options: Another reason to choose the first-price auction. European Economic Review, 2009, 53(2): 153–169. [doi: 10.1016/j.euroecorev.2008.03.005]

[120] Edelman B, Ostrovsky M, Schwarz M. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. American Economic Review, 2007, 97(1): 242–259.

[121] Ausubel LM, Milgrom P. The lovely but lonely Vickrey auction. In: Cramton P, Shoham Y, Steinberg R, eds. Combinatorial Auctions. Cambridge: The MIT Press, 2006. 17–40. [doi: 10.7551/mitpress/9780262033428.003.0002]

[122] Lucier B, Leme RP, Tardos E. On revenue in the generalized second price auction. In: Proc. of the 21st Int'l Conf. on World Wide Web. Lyon: ACM, 2012. 361–370. [doi: 10.1145/2187836.2187886]

[123] Yang J. The efficiency of an artificial double auction stock market with neural learning agents. In: Chen SH, ed. Evolutionary Computation in Economics and Finance. Berlin, Heidelberg: Springer, 2002. 85–105. [doi: 10.1007/978-3-7908-1784-3_5]

[124] Li DH, Yang QY, Yu W, An D, Yang XY. Towards double auction for assisting electric vehicles demand response in smart grid. In: Proc. of the 13th IEEE Conf. on Automation Science and Engineering. Xi'an: IEEE, 2017. 1604–1609. [doi: 10.1109/COASE.2017.8256333]

[125] Nisan N, Roughgarden T, Tardos É, Vazirani VV. Algorithmic Game Theory. Cambridge: Cambridge University Press, 2007. 300.

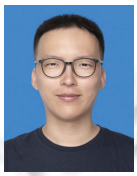
[126] Fricker SA, Maksimov YV. Pricing of data products in data marketplaces. In: Proc. of the 8th Int'l Conf. of Software Business. Essen: Springer, 2017. 49–66. [doi: 10.1007/978-3-319-69191-6_4]

附中文参考文献:

[1] 中国信息通信研究院. 大数据白皮书(2020年). 2020. http://www.caict.ac.cn/kxyj/qwfb/bps/202012/t20201228_367162.htm

[2] 黄欣荣. 从复杂性科学到大数据技术. 长沙理工大学学报(社会科学版), 2014, 29(2): 5–9. [doi: 10.3969/j.issn.1672-934X.2014.02.001]
[13] 张小伟, 江东, 袁野. 基于博弈论和拍卖的数据定价综述. 大数据, 2021, 7(4): 61–79.
[15] 刘枘, 郝雪镜, 陈俞宏. 大数据定价方法的国内外研究综述及对比分析. 大数据, 2021, 7(6): 89–102. [doi: 10.11959/j.issn.2096-0271.2021063]
[16] 蔡莉, 黄振弘, 梁宇, 朱扬勇. 数据定价研究综述. 计算机科学与探索, 2021, 15(9): 1595–1606. [doi: 10.3778/j.issn.1673-9418.2103069]
[17] 刘朝阳. 大数据定价问题分析. 图书情报知识, 2016, (1): 57–64. [doi: 10.13366/j.dik.2016.01.057]
[19] 唐斯斯, 刘叶婷. 我国大数据交易亟待突破. 中国发展观察, 2016, (13): 19–21. [doi: 10.3969/j.issn.1673-033X.2016.13.007]
[20] 胡燕玲. 大数据交易现状与定价问题研究. 价格月刊, 2017, (12): 16–19. [doi: 10.14076/j.issn.1006-2025.2017.12.04]
[60] 朱灿, 曹健. 实体解析技术综述与展望. 计算机科学, 2015, 42(3): 8–12, 18. [doi: 10.11896/j.issn.1002-137X.2015.3.002]

江东(1996－), 男, 博士生, 主要研究领域为大图数据分析, 数据定价.



张小伟(1996－), 男, 硕士生, 主要研究领域为数据定价.

袁野(1981－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为大数据管理, 数据库理论与系统.

王国仁(1966－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为不确定数据管理, 数据密集型计算, 可视媒体数据管理与分析, 非结构化数据管理, 分布式查询处理与优化技术, 生物信息学.