文章编号:1007-5321(2019)01-0120-06 **DOI**:10.13190/j.jbupt.2018-108

数据定价机制现状及发展趋势

彭慧波, 周亚建
(北京邮电大学 网络空间安全学院, 北京 100876)

摘要: 探讨了以合理定价为核心的数据交易机制. 介绍了国内外知名的数据交易平台; 将当前的数据交易定价机制归纳为 4 种模型:基于博弈论的协议定价模型、基于数据特征的第三方定价模型、基于元组的定价模型和基于查询的定价模型. 通过比较各模型的优缺点,结合数据定价的相关理论,分析了存在的问题, 讨论了一种适用于复杂真实交易环境,能准确衡量隐私和正确评估数据价值的方法,为相关研究提供参考.

关 键 词: 数据定价; 定价机制; 博弈论; 隐私度量
中图分类号: TP399 **文献标志码**: A

Data Pricing Mechanism Status and Development Trends

PENG Hui-bo, ZHOU Ya-jian
(School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: The construction of a data transaction mechanism centered on reasonable pricing has been discussed. Firstly, the well-known data transaction platforms at home and abroad has been introduced and the current data transaction pricing mechanisms can be summarized into four models: a protocol pricing model based on game theory, a third-party pricing model based on data characteristics, a tuple-based pricing model, and the pricing model of the query. Based on comparing the advantages and disadvantages of each model, combined with the relevant theory of data pricing, analyze the existing problems and discuss a method that is suitable for complex real-time trading environments which can accurately measure privacy and correctly evaluate the value of data.

Key words: data pricing; pricing mechanism; game theory; privacy measurement

大数据作为一种在一定程度上不可共享的资源,也逐渐演变成为一种可进行交易的商品,这就导致了数据资源成为人类社会一种必不可少的生产要素与战略资产[1]. 据《2016 年中国大数据交易产业白皮书》(以下简称"《白皮书》")的不完全统计, 2016 年中国大数据交易市场规模达到 62.12 亿元, 预计将在 2020 年达到 545 亿元[2]. 鉴于大数据交易在重要领域的巨大价值,中国各地掀起了建设数据交易中心的热潮,目前在建或者已经建成的包括贵阳大数据交易所、上海数据交易中心、武汉东湖大数据交易中心、华东江苏大数据交易平台等在内的国家级数据交易中心有十余家,数据堂、优易数据、数读、京东万象等在内的企业级数据交易平台有近 20 家. 但是由于缺乏相应的流通管理政策及机制,这些数据中心多数处于各自为政的"信息孤岛"状态. 消灭"信息孤岛"、引导大数据交易产业发展、让数据流动起来的一个有效办法,就是建立公平、合理的大数据交易机制,用市场的力量来优化数据资源

收稿日期: 2018-05-25
作者简介: 彭慧波(1994—), 男, 硕士研究生.
通信作者: 周亚建(1971—), 男, 副教授, E-mail:yajian@bupt.edu.cn.

的配置，使之能服务于国家的大数据战略和国民经济建设[3].

刘朝阳、Moiso、Balazinska 等[4-6]对数据交易问题进行了全面、深入的研究，发现数据定价机制是一个关键性的、共性的基础问题. 现行的数据定价方法存在以下几个问题：首先，由于缺乏统一且符合标准的交易渠道和交易规范，加之当前的大数据交易平台组成来源十分复杂（分为交易所平台、产业联盟性质的交易平台和专注于互联网综合数据交易和服务的平台），导致大数据交易总体看来集中度还不高，交易规模也并不大[4]，据《白皮书》调查，到 2016 年底交易市场规模仅占大数据市场总规模的 40% 左右；其次，数据交易仍然采用协议定价、拍卖定价、集合定价等定价方式，数据定价方法无法实现人工智能时代自动化定价的目标；再者，由于信息不对称，存在第三方非法套利的情况，严重影响数据交易市场的秩序[5]；此外，由于用户隐私保护的限制，再加上大数据本身极易复制的特性，使得大数据在价值上具有极高且不可恢复的固定成本和极低的可变成本[6]，导致了传统定价方法在大数据领域不再生效. 因此，数据定价方法成为了国内外科研工作者的一个研究重点.

笔者对数据交易中的重点即数据定价方法进行了系统的综述，包括各主要数据交易中心使用的数据定价方法，结合国内外科研人员关于数据定价的相关理论，分析其存在的问题，为下一步的研究奠定了坚实的基础.

1　国内外数据交易平台

从 2008 年开始，全球大数据交易市场已经初见端倪. 目前，随着各国抢抓战略布局，不断加持发展力度，加之资本的青睐及投资，大数据交易市场呈现不断发展趋势. 但是，根据贵阳大数据交易所得到的数据，全球大数据产业发展还是以美国、欧洲和亚洲为主. 笔者将从国内和国外分别选取一家数据交易公司进行介绍.

1.1　Factual

开放位置数据库服务商 Factual 于 2008 年成立于美国加州洛杉矶. 作为一家提供开放位置数据的大数据公司，其致力于开发位置相关数据集，并与那些没有条件或能力拥有相关数据信息的公司进行分享，将信息共享推向大众化，实现一个以互联网为基础的实时数据交易市场.

根据公司创始人吉尔·埃尔巴兹所述：“透过位置，你就可以了解人们的生活模式，判断他们喜欢的东西，他们在那里，他们在做什么，他们要去做什么等. ”

Factual 工作流程可分为如下 5 个部分.

1）汇总：通过 Facebook、Apple、Nike 等公司提供的位置信息，收集各个方面的地理位置信息，并以原始格式存储，通过不断添加新的数据源，始终保留原始数据，以便稍后通过改进的规则和学习提取更多数据.

2）使用包含超过 1 万个潜在规则的引擎来提取核心属性，应用每个国家/地区的属性将混乱的原始数据转换为结构化、规范化数据.

3）解决方案：通过数十亿经过清理和验证的输入记录及经过确认的代表相同位置的地理特征，将记录归入最佳群集. 以 1 亿个地方实体作为基础构成集群，并由数以万计的集群最终实现解决引擎.

4）总结和实现：通过为每个数据源分配一个属性特定的信任权重来确定真实值，从一组聚类输入中导出一个地点的各个属性，并为每个属性确定最真实的正确值，随着深入地分析数据源，这些信任权重将通过启发式算法和机器学习不断更新，同时汇总后的数据可以通过下载提供数据发布.

5）质量保证：内部数据质量指标包括 22 个衡量数据质量的关键指标，这些指标将衡量数据的准确性和全面性.

通过以上方式，Factual 拥有的数据覆盖 7 500 万个位置，涵盖 50 个国家的商户、公园和其他景点. Facebook、CitySearch、AT&T 在内的一些大公司都会使用 Factual 来获取相关信息. 其不仅提供数据买卖，还提供数据托管、数据评分、买卖双方评分等服务. 目前，Factual 已经推出了 Infochimps Platform 流式数据（streaming data）处理平台，真正实现了实时的数据交易.

1.2　贵阳大数据交易所

贵阳大数据交易所于 2014 年 12 月 31 日注册成立，2015 年 4 月 14 日正式挂牌运营，成为 2017 年 4 月 25 日国家大数据（贵州）综合试验区首批重点企业，是中国乃至全球第一家大数据交易所. 2015 年 5 月 8 日，国务院总理李克强亲自批示贵阳大数据交易所，希望“利用大数据与传统行业的融合，形成‘互联网 + ’的战略支撑”.

截至 2017 年 10 月，贵阳大数据交易所发展会

员超1 500家,接入225家优质数据源,可交易数据产品达到近4 000个,可交易的数据总量超150 PB[2],并连续4年(2015—2018年)承办“数博会”核心论坛——中国(贵阳)大数据交易高峰论坛,发布《中国大数据交易产业白皮书(报告)》、《贵阳大数据交易观山湖公约》等成果,引领了大数据交易产业的发展.

贵阳大数据交易所以电子交易为主要交易形式,面向全国提供数据交易服务. 数据的价格由交易的数据买卖双方协商制定,数据内容和交易价格在平台网站挂出,如果买家想要购买数据,在平台拍下就算交易成功. 当然,大数据交易的是数据分析结果而不是数据本身,该分析结果是通过数据的清洗、分析、建模、可视化后的结果,保障了普通人的隐私.

2 主要的数据定价方法

目前,国内外对大数据交易的研究可分为2种类型:一种是从学术研究的角度,综合考虑各类影响因素建立定价模型,以模型作为定价的依据,典型的方法包括基于查询的定价模型、基于机器学习的数据定价模型等[7];另一种是大数据交易从业人员从操作实践经验出发,以市场交易的实际过程和实际定价作为依据,建立数据定价的经验公式来指导交易过程,典型的实例包括基于博弈论的协议定价模型、基于数据特征的第三方定价模型等.

2.1 基于博弈论的协议定价模型

协议定价就是数据拥有者和数据购买者通过协商,对价格达成统一,这也是目前应用最为广泛的数据定价方法[8]. 首先,数据拥有者根据自身对数据的认识,率先为打算出售的数据定价;其次,数据购买者如果认可数据拥有者提出的价格,则二者交易成功,否则,可通过反复磋商的方式进行议价;最后,二者如能达成一致则交易成功,若不能则交易失败[9]. 若存在多名数据购买者,并且数据购买者有独占数据需要的时候,可采用拍卖的方式对数据进行定价,出价最高者获得数据的购买权. 根据此种情况,如图1所示,张晓玉[10]根据博弈论方法运用“一对一”和“一对多”的讨价还价模型,对大数据这种特殊商品的价格确定过程进行了详细分析,对大数据定价协商过程进行了建模.

如图2所示, Riederer等[11]同样也提出了一种数据拍卖模型,用户根据自身隐私信息含量提出交易底价,数据购买者在数据交易平台上通过拍卖的方式购买经过脱敏处理的数据.

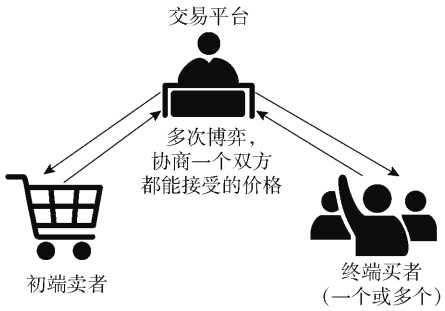

图1 基于博弈论的讨价还价模型

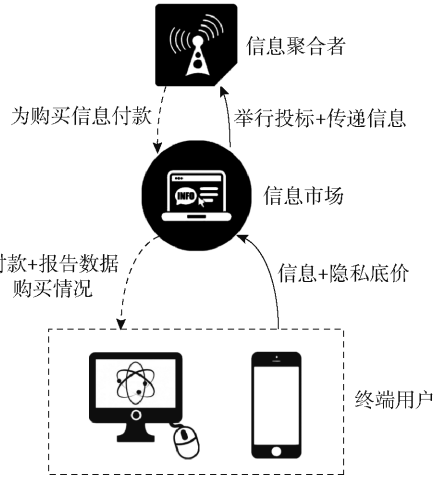

图2 基于隐私的数据拍卖模型

协议定价是当前实践中应用最为广泛的数据定价方法,其能够较为便捷地对数据进行定价. 但是,由于数据买卖双方信息的不对称,对数据价格认识的不一致[12],往往不能准确地评估数据价值,致使数据价格出现偏差,可能会出现非法套利的情况.

2.2 基于数据特征的第三方定价模型

当前,国内外大数据交易平台普遍采取的一种方法是可信第三方定价. 在数据拥有者无法准确针对数据进行定价的情况下,可委托可信第三方进行交易. 例如,Azure、Datamarket、上海数据交易中心、贵阳大数据交易所等大数据交易平台均可根据平台自有的包括数据量、数据完整性、数据时间跨度、数据稀缺性等在内的数据质量评价指标,对数据进行定价. 通过第三方定价方法,每个数据集的价格都将根据数据属性因素和数据集的数据量进行计算[13].

无独有偶,Niyato等[14]结合了经济学中Stackelberg模型和机器学习中的分类算法,将数据交易

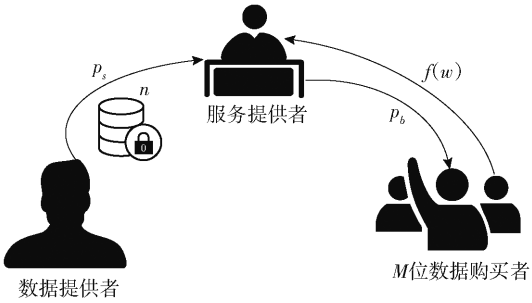

图 3　第三方数据市场模型

分为数据提供者、服务提供者和数据消费者 3 个部分，如图 3 所示. 当数据量为 n 的时候，数据提供者向数据平台提供的使用数据的价格为 p_s；M 位数据消费者愿意为数据付款的概率密度为 $f(w)$，其中 $t\in[0,T]$；服务提供者设定数据提供者提供的“生数据”的单元价格为 p_b，通过最大化 p_b，使 p_b 尽可能逼近 p_s 来实现效益最大化.

$$p_b=\arg\max_{p_b} p_b n(p_b)$$

$$\text{s. t. } n(p_b)=\arg\max_{n,p_s}\left(p_s M\int_{p_s}^{T} f(t)\,\mathrm{d}t-p_b n\right) \tag{1}$$

然而，采用第三方辅助定价的方法，首先必须保证第三方数据交易平台是完全可靠的. 但是，当前国内外数据交易平台背景复杂，政府、企业、个人都参与其中，没有形成一个统一、规范的数据交易平台. 对于用户而言，平台缺乏透明度，会导致信息误传和非对称信息的产生[15]. 其次，采用第三方定价模式，交易的数据往往是整个数据集，并没有针对每个数据元组进行定价，如果用户需要的是部分数据，则必然要购买整个数据集，造成了一定程度上的浪费. 此外，采用人工标注的方式进行第三方定价，无法针对海量数据集自动生成数据的价格，缺乏时效性.

为了实现数据的自动定价，可针对数据的特性设计某种动态计价模型. 自动定价的目的是为了解决海量数据交易的效率问题，但是目前尚存在无法全面找出影响数据定价的特征，以及无法准确评估数据价格等理论问题和必须通过离线方式进行计算等技术问题.

2.3　基于元组的定价模型

所谓元组指的是待交易数据集中的每条交易数据. 而人工定价领域中，被定价的对象往往是一个数据集合，并不能针对数据集合中的每个元组进行定价. 目前的定价模型往往假设集合中所有的元组在价格上都是完全一致的. 当然在实践过程中由于不同元组中包含隐私信息量的不同，这一假设是完全不成立的. 为了解决这个问题，Balazinska 等[16]提出在元组这一结构粒度上设定数据的价格，并基于公共数据库中数据的顺序建立定价函数. 定义价格函数 f 为

$$f:\ D\to R^{+} \tag{2}$$

其中：D 为数据集合，R^{+} 为数据价格. 包含多个元组的数据集价格，就是各个元组单一价格之和.

在以上研究的基础上，Shen 等[17]提出了积极分级和反转定价机制. 数据属性根据其影响数据质量的程度被分为不同等级，每个数据元组的准确价格都是根据数据属性因素和数据集的整体价格进行计算. 将数据元组作为最基本的数据度量组合，结合信息熵、权重、数据引用指数、花费等影响数据价值的因素对每个元组进行定价. 但是，没有对选取信息熵、权重、数据引用指数、花费作为影响数据定价的特征的原因进行说明. Shen 等[17]的具体计算过程如下：

1）假定 P_D 为数据集的需求价格，C 为收集、分析数据以及运行数据交易平台的花费，供给价格为 P_S，则

$$P_S=P_D-C \tag{3}$$

2）假定信息熵（q）、权重（w）、数据引用指数（r）的系数分别为 α、β、γ，其满足约束：

$$\alpha+\beta+\gamma=1 \tag{4}$$

3）P_i 为数据集中第 i 个数据元组的价格，其计算公式为

$$P_i=P_S\left(\frac{w_i}{w}\alpha+\frac{q_i}{q}\beta+\frac{r_j}{r}\gamma\right) \tag{5}$$

其中：q_i、w_i、r_i 分别为第 i 个数据元组的信息熵、权重和数据引用指数，其计算公式为

$$\sum_{i=1}^{n}\frac{q_i}{q}=1,\ \sum_{i=1}^{n}\frac{w_i}{w}=1$$

$$\sum_{i=1}^{n}\frac{r_i}{r}=1 \tag{6}$$

采用元组的方式对数据进行定价，虽然能够公式化地计算出数据价格，从技术上表示了元组在一定程度上具有的价值. 但是，在现实交易的过程中，用户需求偏好也从一定程度上决定数据的价格，而通过现有元组定价的方式无法表征数据获取难度、数据稀缺性、用户通过数据获得的盈利等数据价值. 此外，在大数据环境下，数据往往包含用户的隐私，因此影响数据价格的因素十分复杂，如果仅仅通过

几个简单的公式进行计算,不能完全表示数据的价格.

2.4 基于查询的定价模型

待交易的数据往往是存储在结构化或非结构化的数据库中,用户需要购买的数据往往需要从数据中查询获得,因此就诞生了基于查询的定价模型.该模型允许卖方指定一些视图的价格,允许买方根据自身需要进行任意查询来购买需要的数据,同时模型能够通过指定视图的价格生成其他任意视图的价格. 这样查询到的数据价格就是一系列能够组合出该查询的视图中最优的情况. 基于查询的定价模型应满足以下2个条件:

1) 抗套利(arbitrage-free):以购买全美国的商业数据为例,美国全国的数据价格应该比分别购买50个州的价格之和便宜;

2) 免贴现(discount-free):当确定每一个视图价格时,对于整体数据库而言,应在各视图之和的基础上有一个折扣.

这一模型存在一个问题,原始的基于查询的定价模型限制了用户只能以固定的数量或通过预定义的视图购买数据. Tang 等[18]对数据库中的每个元组分配价格,然后通过生成满足查询结果的最小视图,来定义任意查询的价格. 但是,用户提出的查询往往是重复且有冗余的. 为了保证高效生成查询结果, Tang 等[18]还提出使用了 MiniCon 算法对用户提出的查询进行修正,在查询结果一致的情况下,对查询过程过程进行优化,保证生成数据价格过程的时效性.

虽然研究人员已经对基于查询的数据定价模型做出了改进,但是,基于查询的定价模型仍然存在需要解决的问题. 首先,针对需要的数据生成查询是一个 NP-hard 问题,其算法复杂度较高. 其次,如何选择预先定价的视图,以及基本视图如何定价. 再者,基于查询的定价模型考虑的情境是数据通过离线方式进行交易,但是在大数据条件下往往很短的时间内就有大量的数据生成,预先设定的视图就不能覆盖新生成的数据. 因此,如何解决以上问题就是国内外科研工作者需要考虑的.

3 结束语

实现数据的合理定价,推动数据交易机制的不断完善和数据产业的成熟、壮大,将有利于改变数据资源利用率不高和盈利能力不强的现状,促进中国大数据产业健康、稳定、可持续发展. 中国大数据的起步时间不晚,起点不低,当前在数据交易、定价方面存在的问题对国内的学术圈和产业界可能是一个百年不遇的"弯道超车"的机遇.

为了实现有效的数据定价,下一步的研究重点是寻找一种有效的数据价值度量技术,目前比较理想的方法是以隐私度量作为核心的价值参考,并综合评估其他价值影响因素,构建普适的、可解释的数据定价模型. 当然,如果国家相关主管部门能在涉及交易和定价的关键技术领域加大投入,多给予优惠政策的支持和产业化引导,将有助于国内的大数据生态圈在全球大数据产业链的形成和博弈过程中占据有利的竞争态势,获取更大的话语权.

参考文献:

[1] Gkatzelis V, Aperjis C, Huberman B A. Pricing private data[J]. Electronic Markets, 2012, 25(2): 1-15.

[2] 贵阳大数据交易所. 2016年中国大数据交易产业白皮书[EB/OL]. (2016-05-25)[2018-05-30]. http://www.cbdio.com/BigData/2016-06/02/content_4965656_all.htm.

[3] 连玉明. 重新定义大数据[M]. 北京: 机械工业出版社, 2017: 164-175.

[4] 刘朝阳. 大数据定价问题分析[J]. 图书情报知识, 2016(1): 57-64.
Liu Chaoyang. Analysis on pricing of big data[J]. Intelligence, Information and Sharing, 2016(1): 57-64.

[5] Moiso C, Minerva R. Towards a user-centric personal data ecosystem the role of the bank of individuals' data[C]//16th International Conference on Intelligence in Next Generation Networks. New York: IEEE, 2012: 202-209.

[6] Balazinska M, Howe B, Koutris P, et al. A discussion on pricing relational data[C]// Tannen V, Wong L, Libkin L, et al. In Search of Elegance in the Theory and Practice of Computation. Berlin: Springer, 2013: 167-173.

[7] Tsai Y C, Cheng Y D, Wu C W, et al. Time-dependent smart data pricing based on machine learning[C]//Mouhoub M, Langlais P. Advances in Artificial Intelligence. Berlin: Springer, 2017: 103-108.

[8] 王文平. 大数据交易定价策略研究[J]. 软件, 2016, 37(10): 94-97.
Wang Wenping. Research on big data transaction pricing strategy[J]. Computer Engineering and Software, 2016, 37(10): 94-97.

[9] 陈筱贞. 大数据交易定价模式的选择[J]. 港澳经济,

2016(18)：3-4.
Chen Xiaozhen. The choice of big data transaction pricing model[J]. Hong Kong and Macao Economy, 2016(18): 3-4.

[10] 张晓玉. 基于讨价还价博弈的大数据商品交易价格研究[D]. 鞍山：辽宁科技大学, 2016.

[11] Riederer C, Erramilli V, Chaintreau A, et al. For sale: your data: by: you[C]//ACM Workshop on Hot Topics in Networks. New York: ACM, 2011: 13.

[12] Muschalle A, Stahl F, Löser A, et al. Pricing approaches for data markets[C]//Castellanos M, Dayal U, Rundensteiner E A. Enabling Real-Time Business Intelligence. Berlin: Springer, 2012: 129-144.

[13] 干春晖, 钮继新. 网络信息产品市场的定价模式[J]. 中国工业经济, 2003(5)：34-41.
Gan Chunhui, Niu Jixin. Pricing model for the network information product market[J]. China Industrial Economy, 2003(5): 34-41.

[14] Niyato D, Alsheikh M A, Wang P, et al. Market model and optimal pricing scheme of big data and internet of things (IoT)[C]//IEEE International Conference on Communications. New York: IEEE, 2016: 1614-1810.

[15] 杨琪, 龚南宁. 我国大数据交易的主要问题及建议[J]. 大数据, 2015, 1(2)：38-48.
Yang Qi, Gong Nanning. Reflections on big data exchange of China[J]. Big Data Research, 2015, 1(2): 38-48.

[16] Balazinska M, Howe B, Dan S. Data markets in the cloud: an opportunity for the database community[J]. Proceedings of the VLDB Endowment, 2011(4): 1482-1485.

[17] Shen Yuncheng, Guo Bing, Shen Yan, et al. A pricing model for big personal data[J]. Tsinghua Science and Technology, 2016, 21(5): 482-490.

[18] Tang Ruiming, Wu Huayu, Bao Zhifeng, et al. The price is right[C]//Decker H, Lhotska L, Link S, et al. Database and Expert Systems Applications. Berlin: Springer, 2013: 380-394.