基于数据质量的公平数据定价

陈思莹，张丹，丁小欧，王宏志
哈尔滨工业大学计算学部，黑龙江 哈尔滨 150001

摘要
随着数据的爆炸式增长，以数据为关键要素的数字经济进一步发展。在数据市场中，建立公平高效的定价交易系统变得尤为重要。针对数据市场的公平性，提出了基于数据质量的数据市场模型。首先，以用户的需求为目标，制定综合数据质量的定价策略。其次，为防止用户的恶意欺诈行为，设计了保障公平数据交易的市场机制。最后，在原始数据交易的基础上，讨论了与数据质量相关的数据清洗服务，通过游戏理论中的机制设计多用户清洗价值分配机制。通过实验证明了按照模型构建系统的效率和有效性，可以保证数据市场的公平性。

关键词
数据市场；数据定价；公平性

中图分类号：TP311 文献标志码：A doi: 10.11959/j.issn.2096-0271.2024025

Fair data pricing based on data quality

CHEN Siying, ZHANG Dan, DING Xiaoou, WANG Hongzhi
Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China

Abstract
With the explosive growth of data, the digital economy, where data serves as a crucial element, continues to advance. In the context of data markets, establishing a fair and efficient pricing and trading system becomes paramount. Addressing fairness within data markets, this study introduces a data market model based on data quality. Firstly, aiming at user demands, a comprehensive pricing strategy based on data quality is formulated. Secondly, to mitigate malicious fraudulent behaviors from users, a market mechanism ensuring fair data transactions is designed. Lastly, building upon primary data transactions, data cleaning services related to data quality are discussed. A multi-user value allocation mechanism for cleaning is designed using principles from game theory. Experimental results demonstrate that constructing systems according to this model ensures both efficiency and fairness within data markets.

Key words
data market, data pricing, fairness

0 引言

数据市场，是所有潜在的数据要素供给方、需求方以及数据要素交易行为共同构成的系统[1]。随着数据量的快速增长，数据市场在经济活动中的作用越发重要。数据定价旨在量化数据的货币价值，是数据市场的关键功能，有助于识别和实现数据价值、确保公平交易以及推动数据创新等，促进数据市场高质量发展。然而，数据市场仍然没有形成统一的定价机制，不同的数据市场由于应用场景不同，有不同的定价机制[2-4]。众多学者围绕数据市场中的价值分析和定价问题提出了相应的方法。李然辉[5]认为数据资产的价值分为数据质量价值和数据应用价值，利用层次分析法综合上述两种价值进行分析。Yang等[6]定义了数据质量的效用函数，并提出了利益最大化问题，说明了数据市场可以实现利润最大化。Yu等[7]提出了同时考虑数据质量和数据版本控制的双层次数学规划模型解决数据定价问题。Heckman等[8]设计了一种综合定性分析和定量分析的方法，利用逻辑回归模型进行定价。Ding等[9]设计了基于公平性的数据市场。Deep等[10]设计了可以大规模实现基于查询的定价的轻量级框架。

数据质量对于数据拥有者分析问题、判断局势和做出决策至关重要，错误或者不准确的数据可能导致数据拥有者难以获取关键信息，甚至失去机会或者做出错误决策。同时，低质量的数据需要数据拥有者花费额外的时间和资源进行清洗和修复。对于数据质量高的数据，用户倾向性更高，愿意为其付出更多的金钱以弥补后续的决策损失[11-12]。因此，数据质量是数据市场定价的重要依据，**表1**中的例子展示了数据质量对用户的影响。首先，第2行和第3行数据

表1 低质量数据举例：大学申请信息

学校名称	位置	所属国家	国家代码	申请时间	录取分数
大学1	纽约	美国	001	–	90
大学2	伦敦	英国	0044	12/02/2013	100
大学3	纽约	美国	002	03/12/2013	3.5

的时间格式不统一，存在“月/日/年”和“日/月/年”两种记录方式。同时，虽然第一行和第三行数据有同样的“国家”属性，但是它们的“国家代码”属性却不一致。常识告诉我们，国家和其代码存在着依赖关系，表中的情况是不符合逻辑的。这样的错误可能要求用户购买额外的数据来解决冲突，增加了用户决策的成本。再看观察录取分数列，直觉上可以发现第三行的数据或者是错误数据，或者标度不同。因此，低质量数据给用户造成很大负面影响，需要更低的定价以弥补用户的损失。数据质量在数据库领域的相关研究已经比较成熟[13-15]，有较为完善的数据质量评价体系。

本文从数据质量的敏感性出发，研究基于公平性的数据定价，这里的公平性的保证以数据质量为基础。具体来说，对于数据市场的公平性，从3个方面进行维护。首先，对于用户层面，需要确保用户购买数据的价格与数据质量一致，从而保证对于用户的公平性。其次，对于交易平台，需要确保平台上所有用户交易的公平性，防止某用户恶意欺诈造成不公平，损害其他用户的利益。最后，针对数据交易后的清洗服务，保证服务提供者收益的同时，用户的付出也应该保证公平没有偏置。

从数据市场的公平性出发，本文设计了基于数据质量的公平数据市场中的定价策略。为了保证数据市场的公平性，首先，不同的买家对数据质量的要求不同，需要根据数据质量要求制定合适的数据定价策略。其次，交易平台可能会出现用户的恶意

欺诈行为，通过“欺骗”系统得到有利于自己而损害其他用户利益的结果，需要设计有保障的安全机制。

综上，实现公平数据市场的定价主要存在如下挑战。

- 需要制定合适的数据定价的策略，基于数据质量满足买家的定制化需求，实现对买家的公平性。
- 需要设计防止用户恶意查询的安全机制，保障这个交易平台所有用户的公平性。

针对上述挑战，本文做出了如下贡献。

- 在数据市场的定价过程中，考虑数据质量的因素，以用户的需求为基础将数据质量的各方面因素进行整合，得到公平的定价策略。
- 利用密码学原理和可信第三方，为数据市场提供公平的交易机制，保证用户和交易系统之间不能互相欺骗。
- 在公平定价的基础上，为用户提供数据清洗的增值服务，利用数据服务的合作模式，整合用户需求以负担数据清洗的昂贵代价。

1 基于数据质量的数据定价策略

本文通过把质量差异映射到价格差异中来提高数据市场对于用户的公平性，即质量高的数据将在数据市场中取得更好的价格。值得注意的是，在数据市场定价的过程中，影响数据价格的因素有很多，本文着眼于数据质量对数据定价的影响，即保证在数据质量背景下的数据市场定价的公平性。本节首先讨论数据质量的几个因素，然后讨论如何根据用户的需求将分散的数据质量因素综合起来，研究综合质量情况对价格的影响，最后讨论如何通过基于质量定价函数给出数据最终价格。

1.1 数据质量因素分析

1.1.1 数据质量选择

在讨论在线数据市场中的数据质量问题时，首先要进行数据质量的定量描述。考虑两个因素：第一个是评价数据质量的效率，数据市场上供应者提供的数据规模大，且数据集合更新时数据质量的评估需要重新进行，需要高效的量化方法；第二个是数据质量因素的多样化，为了将各个数据质量因素进行整合，要求不同质量因素的结果有相似的格式。

笔者选择了数据质量的4个方面：准确性、完整性、时效性、一致性。原因如下：首先这几个数据质量因素是研究界普遍认同的数据质量考量对象[12]；其次，这几个因素与数据的价值紧密相关，因而直接影响着数据价格。每个数据质量因素对应一些限制和要求，笔者认为违反了这些要求的数据是低质量的，着重考察对数据质量因素要求的违反数K_s，它可以反映该数据质量因素在整体上被违反的程度。

在下面的讨论中，假设数据库的模式有m个属性，$R=(R_1,\cdots,R_m)$。数据库实例D是R的一个实例，D共有n个元组。

1.1.2 数据准确性

准确性[16]指数据没有错误的程度。本文用数据的类型、格式和预定义的模式3种方式来检测对数据准确性准则的破坏。首先，用模式分析、域分析和数据类型分析对数据进行分析。其中模式分析发现数据记录的通用模式，域分析识别出数据常常落入的区间范围或高频值，数据类型分析使系统归纳数据集合每个属性的数据类型。然后通过3个准则识别不准确的数

据：违反属性模式、超出合理范围、数据类型错误。

用不准确数据的数量占数据集合的比例来度量整个数据集合的不准确程度。为了防止比例值趋向极端，采用比例的负对数形式，数据集合的准确性如式（1）所示：

$$K_1 = K_{\text{acc}} = -\log\frac{n_{\text{ac}}}{mn} \quad (1)$$

其中，n_{ac}表示不准确数据的数量。更准确的数据对应更大的K_1值。

1.1.3 数据完整性

完整性[17]是数据集合中数据未缺失的程度，以及数据集合是否有足够的深度和宽度。为了衡量数据的完整性，需要考察数据集合是否满足以下3个原则：①有合适的数据量；②足够的属性；③较少的缺失值。

首先，数据库的元组容量应该达到一个有意义的最小值。用$n_{\min}$表示最小需要的元组数量。对于给定的数据库实例，$n_{\min}$是一个常数，当$\frac{n_{\min}}{n}>1$时，数据库实例元组数不足。

其次，数据表应该含有足够的属性，即数据表应有足够的宽度。完整性用实际数据表属性集合对最小属性集合的覆盖程度来度量。其中最小属性集合$R_{\text{nec}}=\{R_1,R_2,\cdots,R_p\}$代表了某种类型数据在语义环境下的必需属性的集合。实际数据的属性出现在最小集中的个数记为$\tilde{p}$，数据属性的不完整程度用比值$\frac{p-\tilde{p}}{p}$表示。

最后，缺失值的存在可能也导致数据质量下降。本文用缺失值占总数据量的比例$\frac{n_{\text{mis}}}{mn}$来度量缺失程度，其中$n_{\text{mis}}$是缺失值的数量。

把这3个方面综合起来，数据集合的不完整程度如式（2）所示。这3个方面的权重由w_{com1}、w_{com2}、w_{com3}表示。权值的分布由用户根据不同的应用场景来决定。

$$K_2 = K_{\text{com}} = -\log\left(w_{\text{com1}}\times\frac{n_{\min}}{n}+w_{\text{com2}}\times\frac{p-\tilde{p}}{p}+w_{\text{com3}}\times\frac{n_{\text{mis}}}{mn}\right) \quad (2)$$

1.1.4 数据时效性

时效性[18]指的是数据对于某一个任务的没有过期、时间上的可用程度。把时效性引入数据质量的度量中保证了过期的或者陈旧的数据通过价格修正得到较低的价格。

为了衡量数据的时效性，笔者根据数据种类带来的语义信息得到数据的估计有效期，并根据此有效期来侦测过期数据的记录数n_{exp}。数据的低时效性程度如式（3）所示：

$$K_3 = K_{\text{tim}} = -\log\frac{n_{\text{exp}}}{m_t n} \quad (3)$$

其中，m_t指的是有有效时间戳的数据属性。

1.1.5 数据一致性

数据的一致性是数据符合函数依赖和条件函数依赖的程度。首先查找违背函数依赖的元组n_{vio}，然后统计它们在整个数据集合中的比例，如式（4）所示：

$$K_4 = K_{\text{con}} = -\log\frac{n_{\text{vio}}}{n} \quad (4)$$

1.2 数据质量因素的整合

在考量数据质量对数据价值的影响时，本文选择用清洗成本作为量化的标

准。因为数据质量对数据价值的影响是通过如下方式来完成的：低质量的数据倾向于花费消费者更多的时间和金钱来进行数据清洗，因此要消费者花费较高的价格来购买高质量的数据以避免后续的高强度清洗工作。

为了整合数据质量因素，笔者选择了评估的数据质量值的线性和作为基本骨架。根据奥克姆剃刀理论，在没有黄金守则的情况下，我们选择最简单的方式来进行整合。数据质量不同因素的相对重要程度是根据用户基于应用的需求得到的。用户会被要求提供权重向量 $\boldsymbol{W}=[w_1,w_2,w_3,w_4]$，$w_1+w_2+w_3+w_4=1$，代表上面提到的数据质量因素的一方面在用户最终质量需求中的权重，该向量的分布体现了用户对数据质量的关注倾向。如果用户不能在完全没有指导的情况下将自己的应用需求准确地转换为权值并给出权重向量，系统会通过给出建议来辅助用户给出合适的权值。

例如，用户十分关注数据的完整性，他在购买数据之后会愿意花很多时间和金钱来提高数据的完整性。当他要进行针对完整性的数据清洗的时候，有3种主要的方式以供选择：①忽略所有缺失的数据记录；②用特殊值填补所有缺失值；③根据数据整体捕获填充缺失值。其中，第三种方式虽然时空代价高，但是结果最准确，用户会倾向于选择该方式。

假设清洗效果越好，清洗方法越昂贵。系统衡量并记录典型的数据清洗方法的消耗，并按照消耗和效果分成级别。然后，针对每一个清洗等级给出某一数据质量因素的权值范围，即给用户$\boldsymbol{W}$值的引导和建议。用户根据某一质量因素的清洗需求找到相应的清洗等级，得到对应的权重范围。在此范围内，用户有针对不同应用场景自主调节的自由。

对于上述完整性的例子，系统给出的建议范围可以为：①忽略缺失值及同级别的清洗：[0,0.1]；②特殊值填充及同级别清洗：[0.1,0.2]；③使用统计方法捕获缺失值：[0.2,0.3]；④使用机器学习方法捕获缺失值：[0.3,0.4]。对完整性要求很高的用户会使用最高等级的方法，此时可以把完整性的相对权重设置为0.35。

对于第i个数据质量因素的第j级别，系统有一个离线生成的清洗代价函数 $f_{ij}(K,D,V)$。函数反映的是不同清洗方法的预估时间消耗。用户给定第i个数据质量因素的权重w_i，系统根据w_i落入范围r相应选择f_{ij}。$F_i((K_i,D_i,V_i),w_i)=f_{is}(K_i,D_i,V_i)$，其中$w_i\in r_s$。最终的数据质量值如式（5）所示。

$$\mathrm{FQ}=\sum_1^4 F_i((K_i,D_i,V_i),w_i)w_i \quad (5)$$

1.3 数据质量影响的价格

本节将用前文方法得到的每个数据质量因素的最终量化值施加到数据价格上，在笔者的语境里就是用户查询的价格，笔者把这种价格影响叫作浮动。这种在基础查询价格上的浮动需要把数据质量的各个方面即不同的质量因素整合到一起，并反映出它们的相对重要程度。为了维护在线数据市场的公平性，还需要知道整个市场的一些标准质量参数 $S=(S_1,S_2,S_3,S_4)$，它们代表了在每个质量方面整个市场的平均水平。S可在数据市场的交易数据中学习得到。

已知当前数据库实例的质量评估结果K，用户的权值向量$\boldsymbol{W}$。然后通过两个步骤在原始价格上完成价格浮动。首先，用第1.2节中介绍的方法计算数据库实例的质量值FQ，并用S代替W，使用相似的方法得到数据市场标准质

量的质量值FQ_S。然后，根据FQ和FQ_S计算出最终基于质量的数据价格。第二步，为了平衡浮动对价格的影响，采取一种综合的浮动方式。最终的价格通过式(6)计算：

$$p_{\text{final}} = p + \frac{\text{FQ} - \text{FQ}_s}{\text{FQ}_s} pC \qquad (6)$$

其中，C是数据市场的参数，用来表示数据质量对最终价格的影响。假设，$\text{FQ}=1.5,\text{FQ}_S=1$，$C=0.1$，对于价格数值为2的数据，会增长到2.1，对价格为2 000的查询，则会增长到2 100。总体波动在比较合理的范围内。定价策略中的参数，系统通过统计或机器学习的方式给出用户指导。

原始数据的价格受其他各方面因素影响，利用本节中的方法，笔者在其他影响因素的基础上，将数据质量的影响作用于价格，使数据市场定价对于用户来说更具公平性。

2 数据市场公平交易机制

对于第1节中介绍的定价机制，如果用户恶意构造权重向量$\boldsymbol{W}$欺诈系统，可能得到更低的价格破坏市场的公平性。为了防止这种情况，笔者设计了防止欺诈的公平数据市场交易机制。本节首先定义数据市场的公平准则，然后根据准则设计工作函数，实现数据市场的整体公平交易机制。

2.1 数据市场公平准则

首先，讨论如何评价基于质量的数据定价函数的公平性。如果用户可以通过提供虚假信息欺骗系统得到更低的价格，那么数据市场就失去了公平性。对市场的公平性进行形式化定义，如果数据市场满足用户在真实需求向量为$\boldsymbol{W}$时无法构造向量$\boldsymbol{W'}$，确保：

$$\sum_1^4 \boldsymbol{F}_i\left((\boldsymbol{K}_i,\boldsymbol{D}_i,\boldsymbol{V}_i)\boldsymbol{w}_i'\right)\boldsymbol{w}_i' < \sum_1^4 \boldsymbol{F}_i\left((\boldsymbol{K}_i,\boldsymbol{D}_i,\boldsymbol{V}_i)\boldsymbol{w}_i\right)\boldsymbol{w}_i \qquad (7)$$

那么，这样的数据市场被视为不可欺骗的，即公平的。这个准则要求用户无法有意识地欺骗数据市场得到更低的价格，而非基于实际质量需求。同时，为了保证个人用户不会被数据市场系统欺骗，本文引入可信的第三方来进一步保证公平。

2.2 数据市场工作流程

2.2.1 主要机制

为了满足前文提到的公平准则，提出一种数据市场工作机制，保证在本文的定价策略下，用户无法欺骗系统。数据市场共有三方参与：提供数据集合的数据供应商、在选定数据库上发起查询并购买查询结果的数据消费者、作为平台完成查询操作和质量相关计算的市场管理系统。为了保证各方利益并监督市场运营，引入一个可信无偏见的第三方(trusted third party，TTP，即可信第三方)进行管理和审计，并提供验证服务。

保证公平的主要机制是对用户隐藏质量权重向量$\boldsymbol{W}$和最终查询价格之间的映射关系，防止用户通过遍历权值组合来探测数据集合的质量分布情况。在实际操作中表现为，消费者不会得到每次消费的具体金额，而是加密后的价格值和明文的价格范围。如果用户想要验证费用的正确性，系统需要保证用户在不解密的情况下验证。

下面简述数据市场中用户的操作流

程，假设所有通信发生在授权的安全通道。首先，每个用户在市场申请一个账户并且存储一定的金额。用户发起查询请求，市场管理系统执行查询，计算原始查询价格，并根据用户权值向量计算出最终价格，处理后连同区间一起返回给用户。用户如果接受该范围，则返回查询结果，并从账户中扣除相应费用。如果用户想要验证市场管理系统是否正确地扣除了相应金额，可以向TTP发起查询请求。

2.2.2 系统工作函数

在详细介绍工作流程之前，首先了解系统工作的基本函数。其中，G_n代表了一个模n的整数乘法群。

$(g_i, s_{ki}, \mathbf{pub}_i) \leftarrow \mathbf{Reg}(1^s)$：市场管理系统提供的注册函数，输入安全参数$s$，系统生成唯一的用户辨识码，用户公私钥信息$g_i \in G$。对于安全参数$s$，$p$有$s$位，是$G_n$的阶。

$\boldsymbol{E}_{\text{value}} \leftarrow \mathbf{Enc}(\boldsymbol{i}, \text{value})$：为用户$i$加密价格值或者余额值的算法，返回$E_{\text{value}} = g_i^{\text{value}}$。

$(\text{range}) \leftarrow \mathbf{Range}(\boldsymbol{i}, \boldsymbol{p})$：市场管理系统确定向用户输出的价格范围的概率性函数，给定用户i和价格p，输出$p \in \text{range}$。

$(\boldsymbol{res}, \boldsymbol{p}) \leftarrow \boldsymbol{Query}(\boldsymbol{q}, \boldsymbol{W})$：市场管理系统查询函数。输入用户查询$q$和用户权值向量$\boldsymbol{W}$，得到查询结果res和查询原始价格$p$的确定性算法。

$(\text{consumptionID}, \boldsymbol{E}_p, \text{res}) \leftarrow \mathbf{Consume}(\boldsymbol{i}, \mathbf{Confirm}_i, \boldsymbol{q}, \boldsymbol{W})$：这是一个确定性算法，输入为用户$i$的确认购买信息和用户同意的查询信息$(q, W)$。系统为这次交易生成一个唯一的消费ID，加密$E_p = \text{Enc}(p, i)$，然后在数据集上运行Query执行查询。最后返回消费ID，加密后的价格和查询结果。同时保存此次交易的信息，并返回给TTP以供存档。

$\text{YES} \mid \text{NO} \leftarrow \mathbf{Verify}(\boldsymbol{E}_{B1}, \boldsymbol{E}_P, \boldsymbol{E}_{B2})$：这是一个确定性的算法，由用户端执行。算法输入为加密形式的购买前余额、购买后余额和此次购买价格。如果$E_{B1} \times (E_{B2})^{-1} = E_p$，则返回“YES”。$E_{B2}$可以用扩展欧几里得算法计算。

$\text{YES} \mid \text{NO} \leftarrow \mathbf{CheckBalance}(\boldsymbol{E}_B, \boldsymbol{i})$：TTP执行的确定性算法，用来检查用户提供的一对值是不是用用户i的信息加密的。如果$E_B = \text{Enc}(B_i, i)$，则返回“YES”，其中B_i是用户i的当前余额，因为TTP记录了每次消费，所以总是有用户最新的余额信息。

$\text{YES} \mid \text{NO} \leftarrow \mathbf{VerifyRange}\ (\text{ConsumptionID}, \boldsymbol{E}_p, \boldsymbol{i}, \text{range})$：TTP执行的确定性算法，输入某次消费的记录号码，用户ID、加密形式的价格和系统声称的价格区间。如果价格的真实值确实落在给定区间内，则输出“YES”。为了避免用户恶意利用TTP的范围查询功能来缩小范围，TTP会在收到查询申请时首先检索消费记录号，只有记录中的信息与用户给定的一致时才会给出回应。也就是说用户只能验证自己的真实消费。在用户记录和查询记录一致时，若有$p \in \text{range}$且$E_p = \text{Enc}(p, i)$，则返回“YES”值。

2.2.3 系统工作阶段

下面具体介绍如何利用上述工作函数实现数据市场的公平交易流程。

初始化：假设数据供应商、市场管理系统、TTP三者拥有自己的密钥对和对方的可信公钥。首先，数据供应商向市场提供出售的数据集合，并标明数据集中的独立价格点以用来计算查询价格。管理系统考察并存储数据库的质量信息。交易过程从注册开始，管理系统运行Reg算法，生

成并记录用户信息再将必要的信息告知用户，通知TTP以做记录。

查询：消费者构造一个查询q，然后根据系统建议给出权重向量$\boldsymbol{W}$，将$(q,\boldsymbol{W})$发送给市场管理系统，然后等待系统返回的价格范围。系统执行查询，运行$\mathrm{Query}(q,\boldsymbol{W})$。系统返回价格范围和加密价格$\left(\mathrm{Range}(p),\mathrm{Enc}(p,i)\right)$。如果用户同意价格区间，发回信息通知系统，管理系统之后从用户余额中减去相应的值，更新余额，并将加密的余额返回给用户。系统把成功交易的记录信息发给TTP待查。

验证：用户查询自己的加密形式的余额$\boldsymbol{E}_B$，用户拥有购买前余额、购买后余额和购买价格的加密值$\left(\boldsymbol{E}_{B1},\boldsymbol{E}_P,\boldsymbol{E}_{B2}\right)$，可以运行Verify函数验证本次余额变化的正确性。

TTP上的记录调查：TTP作为可信第三方，根据用户提供的检索信息进行检查和审计。

- 用户可以查询初始余额的加密形式的正确性。用户已知自己账户的真实余额j，将其与自己的加密余额一同向TTP发起验证，因为TTP上存有用户的g^i，可完成验证。不仅可以确定初始值的正确，还可以验证每次消费的正确，保障用户整个流程中不被数据市场恶意收费。同样的验证还可以用在用户向账户充值时，以提高可用余额的情况。
- 用户查询自己的数据消费是否属于市场管理系统声称的范围。用户把消费ID、加密价格、收到的价格区间和自己的用户ID发给TTP。TTP运行VerifyRange，返回结果。

3 数据清洗增值服务价格分配

数据质量并非一成不变的，可以通过数据清洗服务提升数据质量，因此当数据市场可以提供数据清洗服务时，侧重于数据质量的数据定价机制需要进一步改进。数据清洗过程通常耗时而昂贵，巨大数据集合上的清洗费用常常在个人用户的负担范围之外，同时，数据清洗产生的数据质量的提升可以在多用户之间共享，因此笔者在多用户合作的设定下，设计数据市场上面向独立用户的数据清洗服务。需要确保多用户条件下的公平性[19]，对于某一清洗服务，购买服务的用户付出的代价需要保证一致性和公平性。

本节首先对给用户提供数据清洗增值服务的特点进行了分析，然后根据该特点设计了为数据清洗服务进行价值分配的方法。

3.1 数据清洗价值分配分析

对于数据市场上的某一数据集合，有购买了此数据集的查询使用权限用户集合$I=\{1,\cdots,m\}$。他们对数据质量有不同的要求，希望通过购买数据市场提供的增值清洗服务来获取更高质量的数据。

前文考察了数据质量的4个方面，每个方面都有不同的数据清洗方法。设$J_k=\{1,\cdots,n\},k\in1,2,3,4$是第$k$个质量方面的所有的数据清洗方法，得到所有数据市场可以提供给用户的清洗方法的集合$J=\{J_1\cdots,J_4\}$。对于同一质量方面的清洗方法，按照清洗效果和清洗消耗递减的方式进行排序。例如，如果考察数据的完整性，那么就可能有J_{11}为利用贝叶斯网络进行缺失值插补，J_{12}为利用统计规律进行缺失值预测，J_{13}为删除缺失值。

当数据市场系统决定进行某一项清洗操作之后，把清洗结果存储在一个数据库增量中，并决定哪些（为本次清洗付费的）

用户可以获取清洗结果。系统把信息记录在一个授权对(i,j)里面，代表用户i有权获取j的清洗结果。最终整个系统的授权信息将被记录到系统配置a中，a包含了所用实际实施的清洗方法j的集合，还记录了所有的授权对。同时，用$S_j=\{i|(i,j)\in a\}$来表示所有被授权j的用户的集合。整个分配机制的目标就是通过用户的出价和系统的消耗选出合适的配置，根据配置完成清洗，并计算出所有相关用户的价格。

从用户和清洗服务提供者来说，都需要保证各自的公平性。对用户来说，用v_{ij}表示用户i从清洗j中得到的收益。当用户购买了多种清洗方式时，用户得到的总价值就是$V_i(a)=\sum v_{ij}\geqslant 0$。用户$i$通过它的出价$b_{ij}$来向系统声明他从清洗$j$中获取$(i,j)\in a$的价值。如果用户可以通过声明假的价值$(b_{ij}\neq ij)$来取得更低的价格或使其他用户被收取更高的价格，那么系统就是不公平的。因此机制的设计要避免恶意报价对系统的愚弄。对于清洗服务提供者来说，每个清洗方法都需要消耗资源，把清洗j的消耗用$C_j\geqslant 0$来表示。数据市场系统要保证收取的总价格高于清洗消耗时，才会执行这次清洗操作。

3.2 运用改进的夏普利值机制进行价值分配

一方面，在价值分配的过程中，需要保证分配的公平性；另一方面，需要确定所有用户付出的代价的总和能够满足清洗消耗。因此，笔者利用游戏理论中的夏普利值机制（Shapley Value Mechanism）[20]进行价值分配。该机制的简要思想由算法1所示，算法的输入是清洗算法的费用和用户报价集合，输出是中标的用户和最终用户需要支付的价格。算法的基本思想是利用用户愿意出价的上限得到结果。首先假设所有用户都在购买集合内，此时按照集合内的人数把清洗价格均分，得到每个人的价格，把报价小于应付价格的用户从集合中删去。然后重复上述过程，直到集合不再变化或者为空，得到最终用户集合和价格集合。

算法1 Shapley Value Mechanism

输入：清洗费用C_j；用户报价$b_{1,j},b_{2,j},\cdots,b_{m,j}$

输出：中标用户S_j；每个用户应付的价格p_{ij}

$S_j\leftarrow 1,\cdots,m$

repeat

$$p\leftarrow\frac{C_j}{|S_j|}$$

$$S_j\leftarrow\{i\,|\,i\in S_j,p\leqslant b_{ij}\}$$

Until S_j不再变化，或者为空

$p_{ij}\leftarrow$ p if $i\in S_{ij}$

$p_{ij}\leftarrow 0$ if $i\notin S_{ij}$

return $(S_j,(p_{ij})_{i=1,m})$

上述算法针对单一清洗方法设计，但是用户在某一个质量方面能接受的清洗方法为一个集合，因此设计针对多种清洗方法的价值分配策略。如算法2所示，依旧借助夏普利值机制，用$\theta_{i,k}=(J_{pi},b_{i,k})$来表示用户的报价，其中$J_{pi}$是用户$i$对于此质量方面所有可以接受的清洗方法的集合，$b_{i,k}$则是集合中的任意清洗愿意付的最高价格。首先，计算每一个质量方面的所有清洗中被执行的集合，受益用户集合及其相应应付价格，即配置a。算法执行过程中，按照J_k中的原有顺序执行，保证数据的最终质量提升最大。分别对4个数据质量方面执行完成后，将每个用户应付价格加和即可。

算法2 改进的Shapley Value Mechanism

输入：本方面可选的清洗方法集合J_k；清洗费用$(C_j)_{j=1,n}$；用户报价$b_{1,j},b_{2,j},\cdots,b_{m,j}$

输出：配置a；每个用户应付的价格$(p_{ij})_{i=1,m;j\in J_k}$

$a\leftarrow\Phi, p_{ij}\leftarrow 0, \forall i=1,m\forall j\in J_k$

$l=|J_k|$

for j=1 to l do

$(S_j,(p'_{ij})_{i=1,m})\leftarrow \text{Shapley}-\text{Mech}(C_j,(b_{ij})_{i=1,m})$

if $S_j\neq\Phi$ then

$a\leftarrow a\cap\{j\}$

for each $i\in S_j$ do

$a\leftarrow a\cap\{(i,j)\}$

$p_{ij}\leftarrow C_j/|S_j|$

$\forall j\in J_k, b_{ij}\leftarrow 0$

end for

$C_j\leftarrow\infty$

end if

end for

return $(a,(p_{ij})_{i=1,m;j=1,l})$

4 系统实验及评价

原型系统搭建在现有数据库管理系统上，并和数据供应商、数据购买者以及可信第三方之间进行交互。

4.1 实验环境设置

系统运行在2.5 GHz Core i3CPU和8 GB内存的笔记本电脑上。系统使用Python实现，运行在MySQL数据库管理系统之上。使用现实世界中在真正使用中的数据市场AggData和Windows Azure Marketplace中销售的数据来衡量系统性能。共挑选了5个数据集合，包括英国大学地理位置信息（University）、历史气候信息（Weather）、国家代码及信息（Country Code）、美国1997—2011年各州各产业GDP信息（GDP）、2004—2010年400个慈善机构的完整列表（Philahthropy）。

系统中的所有参数和数据库其他信息存储在单独的配置文件中。这些信息包括特定属性的模式信息、函数依赖信息等。在本文的实验中，参数是通过分析数据手动得出的，在实际的数据市场中有了数据积累，参数通过统计或机器学习方法从原有数据中获得。

4.2 实验结果及分析

4.2.1 效率性

利用不同元组数和属性数的数据库实例衡量系统的效率表现。实验中，笔者向系统提交不同的数据集合并且发起不同的查询。衡量数据市场对提交的反应时间，即系统完成数据集合质量信息评价的时间，然后发起查询进行测试，衡量系统对查询的反应时间。

数据规模及测试结果见**表2**。从**表2**可以看出，在5个数据集合中，有3个可以在1 s之内完成提交处理，余下的两个处

表2 系统效率测试

项目	University	Weather	Country Code	GDP	Philanthropy
行数	590	20 750	206	72 900	2 798
列数	12	17	27	8	9
平均处理时间	0.075	1.248	0.100	4.230	0.268
平均查询时间	0.037	0.024	0.010	0.054	0.015

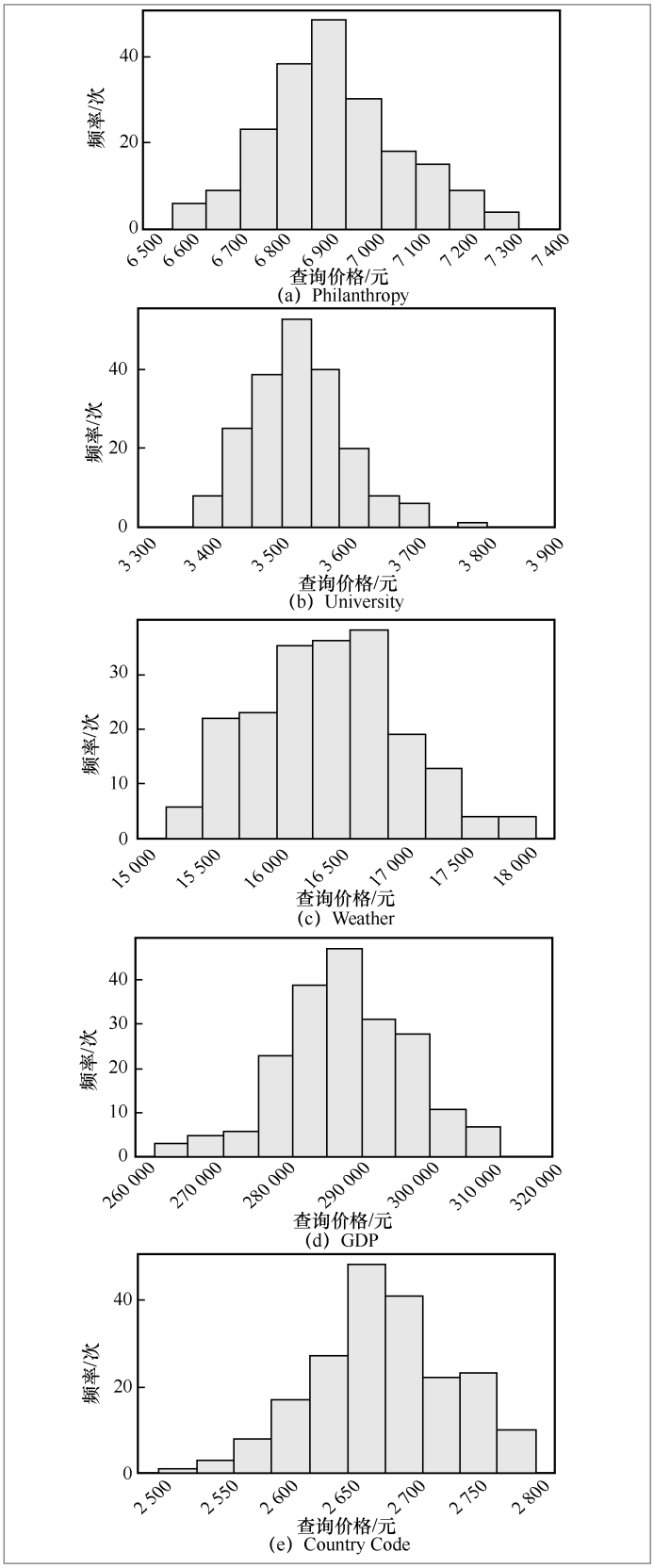


图1 随机权值向量下的价格分布

理时间也小于5 s。这样可以看出，系统可以应对较大的数据集合，也可以有效适应数据更新。对于数据查询时间，全部在0.1 s以下，保证了在线查询的用户交互性。

4.2.2 有效性

因为基于用户不同的查询权值向量，查询的最终价格不同，所以没有确定性的方法来测试算法的正确性。本文设计两种方式来评估定价系统。

第一个是对价格分布的验证，对于均匀的用户需求，价格的分布应该接近正态分布。只有那些有特殊需要的用户会导致价格落在距离期望值较远的地方。用200 个随机分布的权值向量W作为输入，测试同一个数据集合上的一个查询。结果如**图1**所示，5个结果的分布规律大多遵从正态分布的形式。也就是说，在用户需求均匀时，查询的价格呈正态分布。对于用户来说，用户以小概率获得极端价格，保证了公平性。对于系统来说，交易的价格稳定在某一区间。

第二个方面是测试数据质量和数据价格的关系。为了避免原始查询价格的影响，笔者用同一数据库上的同一个查询进行测试。通过手动随机增加错误数据，降低数据质量，并在此过程中考察系统给出的价格的变化，即考察数据价格随错误率的变化关系。

从**图2**可以看出，随着数据集合中数据错误率的增加，查询的价格呈递减的趋势。由此可见，本文的定价策略可以有效反应数据质量的变化。错误率高的数据，数据质量低，用户将以更低的价格获取。增加错误率，数据的准确性、一致性、完整性和时效性就会有所降低，那么数据质量值随之降低，最后导致价格浮动降低，尽管

浮动受整个市场平均水平的影响，但数据质量降低，价格整体呈递减趋势，与实验结果相合。

5 结束语

以数据为关键因素的数字经济快速发展，数据的重要性不断上升。合理评估数据的价值和质量，加强数据资产的管理，对数据市场的高质量发展有重大意义。本文从针对数据质量而言的市场公平性出发，提出一个基于数据质量的数据市场模型：包括面向不同用户质量需求的数据定价策略，利用可信第三方保障公平数据交易的市场机制和针对多用户增值服务的清洗价值分配机制。最后，通过实验证明了按照模型构建系统的效率和有效性，可以保证数据市场的公平性。

参考文献:

[1] 李兵兵. 我国数据市场发展的理论基础与路径[J]. 社会科学动态, 2022(11): 34-37.
LI B B. Theoretical foundation and path on the development of Chinese data market[J]. Dynamics of Social Sciences, 2022(11): 34-37.

[2] ZHANG M X, BELTRÁN F, LIU J M. A survey of data pricing for data marketplaces[J]. IEEE Transactions on Big Data, 2023, 9(4): 1038-1056.

[3] 尹传儒, 金涛, 张鹏, 等. 数据资产价值评估与定价：研究综述和展望[J]. 大数据, 2021, 7(4): 14-27.
YIN C R, JIN T, ZHANG P, et al. Assessment and pricing of data assets: research review and prospect[J]. Big Data Research, 2021, 7(4): 14-27.

[4] 任洪润, 朱扬勇. 基于数据市场类型的数

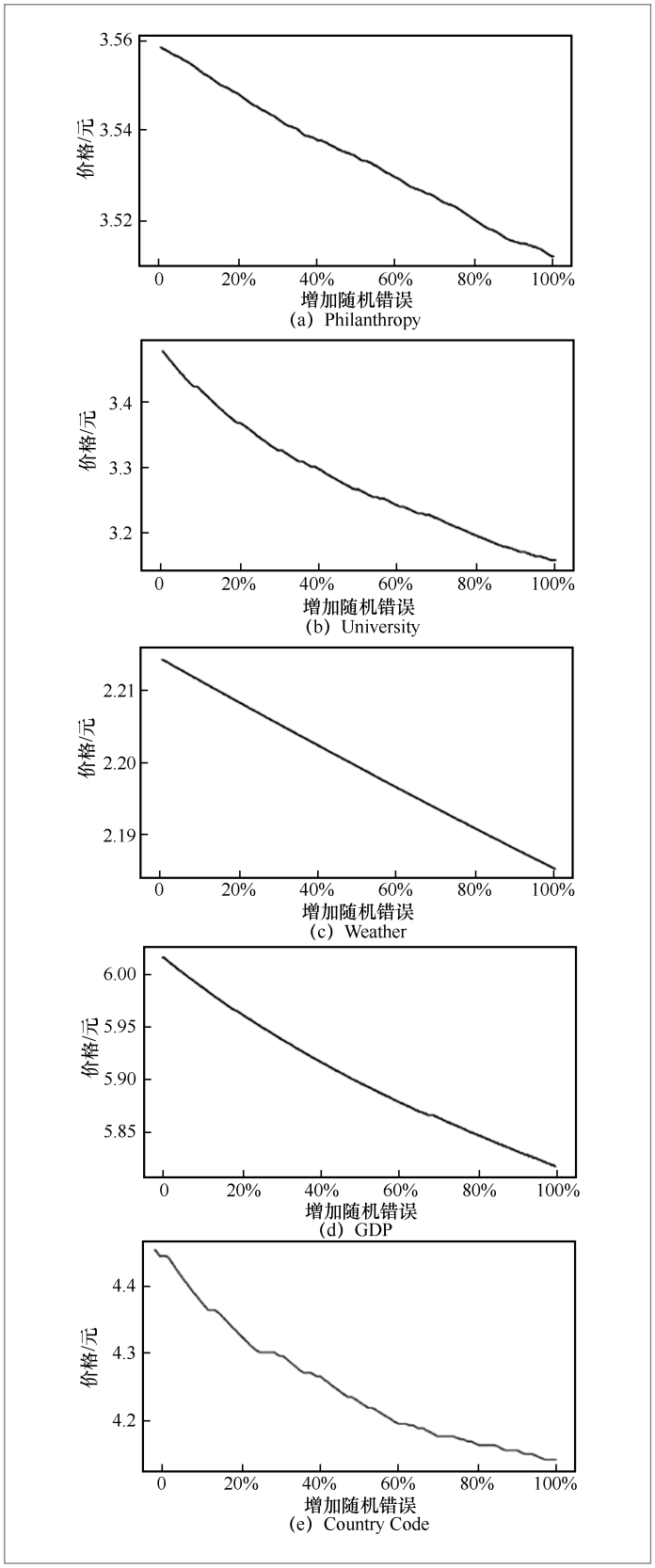


图 2 人工增加错误时价格变化

据定价模型研究[J]. 大数据, 2023, 9(4): 116-138.
REN H R, ZHU Y Y. Research on data pricing model based on data market type[J]. Big Data Research, 2023, 9(4): 116-138.

[5] 李然辉. 数据资产价值评估模型的理论研究与技术实现探讨[Z]. 2018.
LI R H. Theoretical research and technical realization of data asset value evaluation model[Z]. 2018.

[6] YANG J, ZHAO C C, XING C X. Big data market optimization pricing model based on data quality[J]. Complexity, 2019, 2019: 5964068.

[7] YU H F, ZHANG M X. Data pricing strategy based on data quality[J]. Computers and Industrial Engineering, 2017, 112(C): 1-10.

[8] HECKMAN J R, BOEHMER E, PETERS E H, et al. A pricing model for data markets[J]. iSchools, 2015.

[9] DING X O, WANG H Z, ZHANG D, et al. A fair data market system with data quality evaluation and repairing recommendation[C]//Proceedings of Asia-Pacific Web Conference. Cham: Springer, 2015: 855-858.

[10] DEEP S, KOUTRIS P, BIDASARIA Y. QIRANA demonstration[J]. Proceedings of the VLDB Endowment, 2017, 10(12): 1949-1952.

[11] PEI J. A survey on data pricing: from economics to data science[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(10): 4586-4608.

[12] CHEN L J, KOUTRIS P, KUMAR A. Model-based pricing for machine learning in a data marketplace[EB]. arXiv preprint, 2018, arXiv: 1805.11450.

[13] DING X O, WANG H Z, SU J X, et al. Cleanits[J]. Proceedings of the VLDB Endowment, 2019, 12(12): 1786-1789.

[14] 丁小欧, 王宏志, 于晟健. 工业时序大数据质量管理[J]. 大数据, 2019, 5(6): 1-11.
DING X O, WANG H Z, YU S J. Data quality management of industrial temporal big data[J]. Big Data Research, 2019, 5(6): 1-11.

[15] 丁小欧, 王宏志, 张笑影, 等. 数据质量多种性质的关联关系研究[J]. 软件学报, 2016, 27(7): 1626-1644.
DING X O, WANG H Z, ZHANG X Y, et al. Association relationships study of multi-dimensional data quality[J]. Journal of Software, 2016, 27(7): 1626-1644.

[16] SIDI F, SHARIAT PANAHY P H, AFFENDEY L S, et al. Data quality: a survey of data quality dimensions[C]//Proceedings of the 2012 International Conference on Information Retrieval & Knowledge Management. Piscataway: IEEE Press, 2012: 300-304.

[17] WANG R Y, STOREY V C, FIRTH C P. A framework for analysis of data quality research[J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(4): 623-640.

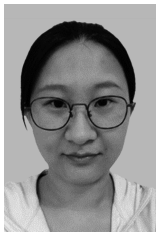
[18] STRONG D M, LEE Y W, WANG R Y. Data quality in context[J]. Communications of the ACM, 1997, 40(5): 103-110.

[19] 张小伟, 江东, 袁野. 基于博弈论和拍卖的数据定价综述[J]. 大数据, 2021, 7(4): 61-79.
ZHANG X W, JIANG D, YUAN Y. A survey of game theory and auction-based data pricing[J]. Big Data Research, 2021, 7(4): 61-79.

[20] SHAPLEY L S. A value for n-person games[M]. Santa Monica: RAND Corporation, 1952.

作者简介

陈思莹（2001– ），女，哈尔滨工业大学计算学部硕士生，主要研究方向为数据清洗。



张丹（1992– ），女，博士，哈尔滨工业大学计算学部研究员，主要研究方向为数据质量、以人为中心的人工智能。

丁小欧（1993– ），女，博士，哈尔滨工业大学计算学部助理教授，主要研究方向为数据清洗、时间数据质量管理、时间数据挖掘、工业数据清理和多元时间序列数据中的异常行为挖掘。

王宏志（1978– ），男，博士，哈尔滨工业大学计算学部教授，计算机科学与工程系主任，海量数据计算研究中心主任，黑龙江省大数据科学与工程重点实验室主任，主要研究方向为数据库和大数据。

收稿日期： 2023-12-31

通信作者： wangzh@hit.edu.cn

基金项目： 国家重点研发计划资助项目（No. 2021YFB3300502）；国家自然科学基金资助项目（No.62202126，No.62232005）；中国博士后科学基金项目（No.2022M720957）；黑龙江省博士后面上资助项目（No.LBH-Z21137）

Foundation Items: The National Key Research and Development Program of China (No. 2021YFB3300502), The National Natural Science Foundation of China (No.62202126, No.62232005), China Postdoctoral Science Foundation (No.2022M720957), Heilongjiang Postdoctoral Financial Assistance (No.LBH-Z21137)