计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330791
61(12): 3154−3167, 2024

基于劳资博弈模型的实用查询定价新算法

王会举　黄玮煊　岳　晓
（中南财经政法大学信息工程学院　武汉　430073）
（wanghj@zuel.edu.cn）

Novel Practical Query Pricing Algorithm Based on Labor Game Model

Wang Huiju, Huang Weixuan, and Yue Xiao
（*School of Information Engineering*, *Zhongnan University of Economics and Law*, *Wuhan* 430073）

Abstract　With the promotion of data as a production factor, traditional query pricing methods face tremendous challenges in practical applications due to their overly strict premise assumptions, limited support for flexibility and dynamics, and inadequate consideration of key factors. To address these issues, we innovatively design a query pricing algorithm based on the labor game model. This algorithm models the participants in data transactions as labor unions and employers, and treats the data trading platform and data buyers as the labor union and employers, respectively. The data trading platform (labor union) is responsible for the fair and transparent calculation of the value of each traded dataset (wages), aiming to facilitate transactions as much as possible. Data buyers determine the purchase quantities of datasets based on their estimated value, personal needs, and budgets, thereby achieving a pricing strategy that balances the interests of all three parties. Experimental results demonstrate that compared with the popular Stackelberg game model, our algorithm better accommodates the interests of all parties and ensures greater fairness. Compared with traditional query-based data pricing methods, our pricing algorithm is more practical, offers greater flexibility and dynamics, and can dynamically adjust prices in response to changes in query results. The time complexity of our pricing algorithm is O(N), where N is the number of datasets related to the query, and it also guarantees no arbitrage.

Key words　query pricing; labor game model; data pricing; data trading; fair pricing

摘　要　在数据要素化的推动下，传统查询定价方法因其前提假设要求过高、灵活动态性支持有限、关键因素考虑不足等问题，面临落地难的巨大挑战. 为解决以上问题，创新设计了基于劳资博弈模型的查询定价算法，该算法利用劳资博弈模型对数据交易中参与方进行建模，将数据交易平台和数据买方分别视作工会和用人单位；数据交易平台（工会）负责各交易数据集价值（劳动者工资）公平透明计算，以尽可能促成交易为目标；数据买方根据各数据集估量价值、自身需求和自身预算，决定各数据集购买数量，藉此实现兼顾三方利益的交易数据集定价. 实验表明，该算法相比于流行的斯塔克伯格博弈模型，更能兼顾各方利益，更加公平；相比于传统的基于查询的数据定价方法，该定价算法更易落地应用、更具动态灵活性，可以跟随查询结果的变化实现价格的动态调整. 该定价算法时间复杂度为 $O(N)$（N 为查询相关数据集个数），且具有无套利性.

关键词　查询定价；劳资博弈模型；数据定价；数据交易；公平定价

收稿日期：2023-10-10；**修回日期**：2024-05-16

基金项目：国家社科基金后期资助项目（22FJYB022）；数字技术与现代金融学科创新引智基地项目（B21038）
This work was supported by the China National Social Science Fund Post-Funding Project (22FJYB022) and the Innovation and Talent Base for Digital Technology and Finance Project (B21038).

通信作者：岳晓（yuexiao@zuel.edu.cn）

中图法分类号 TP391

数据的爆发增长、海量集聚蕴藏了巨大的价值，为智能化发展带来了新的机遇[1]. 数据的流通汇聚是从数据“量变”到知识“质变”的关键步骤. 数据定价是数据要素流通的核心问题之一. 基于查询的数据定价(简称查询定价)是对数据库查询结果进行价格计算的算法[2]. 在数据要素化的背景下，传统查询定价算法面临落地难的巨大挑战：

1)用户关键需求考虑不足，难以促成交易. 传统查询定价方法较少考虑用户的关键购买需求，尤其是用户预算、多数据集组合购买及数据集倾向性等需求，容易导致交易失败.

2)前提假设要求过高，难以落地. 传统查询定价方法往往需要用户预设细粒度价格点如〈基础视图，价格〉点，查询基于预设基础视图进行，查询结果价格基于预设的基础视图价格点推导而出. 由于数据商品的非标性，由普通用户来预设价格点，在实际操作中并不现实.

3)用户查询的动态灵活性支持不足，难以满足用户对交易数据集的灵活定制化需求. 传统查询定价方法中查询的灵活性往往取决于基础视图的定义，导致查询的动态灵活性不足. 虽然已有研究关注到此问题，提出基于元组的定价方法[3]，但也存在空间或时间成本过高的问题，而且粒度只到元组，未到列级，动态灵活性仍然有限，也严重影响了其推广应用.

为解决传统查询定价中存在的以上遗留问题，本文研究创新性提出了一种兼顾数据交易参与双方利益的新查询定价方法——基于劳资博弈的查询定价算法. 该算法采用更加适合数据交易市场的新博弈模型——劳资博弈模型——来实现双方利益的博弈计算，最终实现查询公平价格的计算，避免目前常用的动态博弈模型——斯塔克伯格博弈——在数据定价中存在的不公平问题. 该算法无需数据提供者预设价格点，系统可以自动利用现有数据集价值量计算算法，通过劳资博弈模型计算出数据交易价格. 大量实验结果表明，基于劳资博弈的查询定价算法可以兼顾双方利益，实现多数据集动态公平定价，能更直接地指导交易，有效促成交易的达成.

本文研究贡献有 3 点：

1)首次提出多数据集组合购买定价问题.

2)创新设计了基于劳资博弈模型的查询定价算法. 该算法无需预设价格点，可以根据买方需求，实现多数据集的组合定价，兼具动态灵活性和公平合理性，更易促成数据交易的达成.

3)进行了大量实验验证分析，验证了基于劳资博弈的查询定价算法可以兼顾双方利益，实现多数据集动态公平定价，能更直接地指导交易，有效促成交易的达成.

1 相关工作

从定价原理角度来看，同本文工作相关的研究主要是基于查询的定价方法和基于博弈的定价方法，现综述如下.

1)基于查询的定价方法

Balazinska 等人[4]有远见地指出了查询定价对数据库社区的重要性，随后文献 [2–10] 进一步发展了这个观点，并提出了一系列查询定价方法. Koutris 等人[2]提出的查询定价模型允许卖家对部分视图指定明确的价格，模型会根据这些视图自动导出其他买家查询的视图集价格，且针对该模型提出了定价函数，并证明该函数满足定价模型的无套利和无折扣要求. Lin 等人[5]考虑了查询定价的套利情况，并提出了免套利定价函数，以解决基于查询定价所导致的套利问题. Upadhyaya 等人[6]提出了基于退款的历史感知定价. 文献 [7] 描述了不同套利条件下定价函数的可能空间，其理论框架在 Qirana 系统中得到了应用[8-9]. Tang 等人[3]提出了类似的基于元组的定价方法，即只有对结果集有贡献的元组才会被计费. 然而，此类查询定价需要预设一些细粒度的价格点如〈基础视图，价格〉，因此此类方法计算复杂度较高，并且很难根据结果集自身价值动态地调整价格. Li 等人[10]提出了一种基于量化噪声的定价模型，加入的噪声越多，获取的数据集的敏感度越低，数据集的价格越便宜. 该工作结合差分隐私和查询，以扰动的程度作为价值衡量的参考，此举一定程度上实现了动态定价，使得数据集价格可以随着扰动参数变化，但是其研究没有给出无扰动的原始数据集是如何定价的，如果还是通过给定价格点进行推导，那么该研究的定价方法还是无法完全实现动态定价. Chen 等人[11]也采用文献 [2] 预设价格点的方式，对图数据定价问题进行了研究.

基于查询的定价方法能够很好地满足交易数据集的定制化需求，且在实时系统中有着良好的发挥空间. 但现有查询定价研究主要关注单个查询结果

的购买定价问题,较少考虑多查询集的组合购买定价问题.此外,其预设价格点的前提要求在现实应用中存在落地难的问题.数据卖家如何合理地预设价格点才能满足双方收益最大化,实现公平定价,也是基于查询的定价方法需要解决的难题.

2)基于博弈的定价方法

博弈定价是商品定价的常用方法.在数据定价领域,因为目前缺乏有效的实际交易价格确定方法,使得博弈定价在定价领域取得了非常好的效果.Haddadi等人[12]提出了一个基于斯塔克伯格博弈的定价模型,该模型是一个典型的双寡头完全信息动态博弈模型,其特点是不要求博弈双方同时做出决策,而是运行一方在对方决策后才进行决策,因此后决策的一方总是会取得博弈的有利地位.Xu等人[13]针对汽车共享设计了数据卖方、交易平台及数据买方三方参与的斯塔克伯格博弈数据定价方法.苑迎等人[14]面向云资源提出了一种基于非完全信息的动态博弈定价模型.Berz[15]利用讨价还价博弈来作为复杂谈判情况下提升拍卖性能的有效举措.Riederer等人[16]提出了利用拍卖会来出售个人数据.Susanto等人[17]提出了一种基于McAfee拍卖模型的双向拍卖机制,且证明了该机制能够达到纳什均衡.Ghosh等人[18]利用差分隐私和拍卖会来最大程度实现保护隐私的同时以最低价格买到最准确的个人数据.

基于博弈的定价解决了静态定价的问题,在理性博弈时可以真实体现出商品的价格,并且能够满足双边收益最大化,实现公平定价.但是其缺点也非常明显.首先是拍卖、讨价还价等博弈定价方法效率低下;其次是交易的数据包是已经打包好再进行出售的,难以满足交易中的定制化需求;再次是交易者们不总是理性的,特别是在拍卖、讨价还价时,商品成交价格可能会极大地偏离商品价值.基于斯塔克伯格的定价模型解决了以往拍卖讨价还价的低效率问题,尽量规避了在定价时交易者们容易出现的不理性行为,但是由于斯塔克伯格博弈模型是一个大小寡头的博弈模型,其博弈双方地位的不对等性易导致定价不公平性问题.

2 基于劳资博弈的查询定价算法

本节介绍如何利用劳资博弈模型实现查询定价算法,包括算法流程、建模及模型求解3部分内容.

2.1 问题描述

目标问题可建模描述为函数 $P=f(ds, Q, W, S, B)$,其中函数 f 即是本文研究所提的劳资博弈定价函数,其输入参数 $ds=\{D_1, D_2, \cdots, D_n\}$ 为待售数据集集合,$Q=\{Q_1, Q_2, \cdots, Q_n\}$ 为买方查询束,W 为买方对查询结果集的倾向排序,S 为买方购买策略,B 为买家预算;其输出为买家数据集最优购买组合及各数据集价格集合 $P=\{\langle Q(D_i), P_i\rangle, \langle Q(D_j), P_j\rangle, \cdots, \langle Q(D_k), P_k\rangle \mid 1<i, j, k<n\}$, $Q(D_i)$ 为从数据集 D_i 购买的查询结果集,P_i 为买方购买 $Q(D_i)$ 查询结果集应支付的价格.目标问题求解即是以参数 ds, Q, W, S, B 为输入的定价函数 f 的结果计算.

2.2 定价算法整体流程

查询定价的思想源于基于版本的数据定价.基于版本的数据定价思想是平台通过预设多个数据版本,并为每个数据版本预设价格,来满足不同数据买家的需求.基于查询的定价方法考虑了结构化数据的多版本,让买家提出任意的结构化查询语句,结构化数据定价系统根据查询语句生成指定的数据集,并自动生成价格.在Koutris等人[2-9]的一系列工作中,数据集的价格是基于卖家预设的价格点,借助定价函数来自动生成的.预设价格点方式存在落地难、动态灵活性不足等问题,本文研究对其思路进行了改进,将预设价格点操作替换为数据价值量的计算,并利用劳资博弈模型实现定价函数.改进后的定价步骤有:

1)交易平台对上架数据集价值进行衡量①;

2)买家输入查询语句②;

3)交易平台执行查询语句,获得每一个数据源 i 的查询结果集 D_i,计算结果集单位价值量 (D_i, θ_i);

4)交易平台利用劳资博弈定价函数 f 计算结果集 D_i 的价格.

举例说明定价算法流程如下:某服装公司拟购买一些男装交易数据用于构建男装销售预测模型.该公司登录某数据交易平台,选择电商类数据,发现有2个数据包可用,分别叫作“京东男鞋”和“京东男装_女装_童装”,分别记为数据集 D_1 和数据集 D_2.数据集 D_1 售价101 855元,大小290 MB;数据集 D_2 售价9 150元,大小22 MB.该公司建模人员在每个数据集上查询选择部分或者全部数据,查询结果分别记为 $R_1=Q(D_1)$ 和 $R_2=Q(D_2)$,数据交易平台根据该公司的

① 数据集价值衡量是数据定价的基础工作.附录A介绍了本文研究采取的用于后续实验验证的价值衡量方法.

② 如果数据源结构不同,则需要输入不同的查询语句.

预算 B、数据集倾向性(数据集购买优先级排序)及购买策略(均衡购买策略、部分优先购买策略等，详见 2.5 节算法 1)、基于价值量计算函数(详见附录 A)计算出 R_1 和 R_2 所含价值量，并计算出各自数据价值量单价，利用劳资博弈定价函数(详见 2.3~2.5 节后续内容)给出每个数据集的购买量及其价格.

2.3 劳资博弈模型构建

1946 年，里昂惕夫提出劳资博弈模型，用以描述代表劳资双方的工会与企业之间的博弈问题. 该模型是一个完全信息动态博弈模型. 模型假设工人工资完全由工会决定，雇工数量则由企业根据工会提出的工资要求自行调整；工会在促进企业健康发展的前提下权衡工人工资，以使工人接受工资、企业雇佣更多工人.

根据上述博弈情形，本文研究基于一般的数据交易情况，提出 2 个博弈变量，分别称作交易价值量 V和单位价格 P，并设定数据买家的收益函数为 R 和数据交易平台的收益函数为 U. 价值量是数据集价值的衡量指标，假设数据集 D的成本为 z，其含有的价值量为 θ，针对数据集 D的交易价值量 $V\in[0,\theta]$. 单位价格 P为数据买家愿为单位价值量 Δ支付的价格. 单位价值量 Δ是数据集交易的基本单位，$\Delta=1/\theta$价值量在实际的交易情形中是多样的，可以是数据集的质量量化值[19]，也可以是数据集的信息熵[20]等价值衡量因素. 在针对一些特殊数据类别进行定价的时候，还会有更加有针对性的价值衡量因素，比如隐私价值[10].

基于 V 和 P 这 2 个博弈变量和交易双方的收益函数，本次交易场景的收益矩阵如表 1 所示. 接下来将对本次博弈的子精炼纳什均衡解进行具体分析. 数据交易平台是本次博弈的先行动方，在劳资博弈模型中，充当着工会的角色. 工会在劳资博弈中决定工人的工资，数据交易平台在本文研究所涉及的交易情景中决定单位价格. 工会的收益是一个关于工资和就业人数双变量的函数，因此，数据交易平台的收益也是一个关于单位价格 P与交易价值量 V双变量的函数.

根据效用理论，数据买家的资源是有限的. 设数据买家预算为 B，输入一组查询 $Q=\{Q_1, Q_2, \cdots, Q_n\}$对应输出结果集 $R=\{R_1, R_2, \cdots, R_n\}$，则数据交易平台的收益函数为

$$U=\sum_{i=1}^{n}U(P_i,V_i)=\sum_{i=1}^{n}P_i\times V_i-z, \tag{1}$$

$$\text{s.t. } P_i>0,$$

$$0<V_i\leqslant V_{\max_i},$$

$$\sum_{i=1}^{n}U(P_i,V_i)\leqslant B.$$

式(1)中，P_i是结果集 R_i 的单位价格，V_i是买家针对结果集 R_i 成交的交易价值量，$V_{\max_i}$是结果集 R_i 可获得的最大价值量，z 是数据成本.

数据买家作为本次博弈的后动方，在劳资博弈模型中充当着企业的角色. 企业在劳资博弈中决定雇工的数量，对应地，数据买家决定交易价值量的多少. 企业的利润等于企业的雇工收益减去雇工成本，因此企业的收益是一个关于工资和雇工人数双变量的函数. 同理，数据买家的收益是一个关于单位价格与交易价值量的双变量函数. 设

$$R=\sum_{i=1}^{n}R(P_i,V_i)=\sum_{i=1}^{n}(G_i(V_i)-P_i\times V_i), \tag{2}$$

$$\text{s.t. } V_{\min_i}\leqslant V_i\leqslant V_{\max_i}.$$

式(2)中，$G_i(V_i)$ 是数据买家有关于结果集 R_i 交易价值量的收益函数，$V_{\min_i}$是数据买家所能接受的交易价值量的最小值，低于该值的结果集 R_i 对数据买家没有收益，即收益为 0.

2.4 劳资博弈模型均衡求解

本节讨论如何基于构建的博弈模型确定最优博弈解，也称为子精炼纳什均衡解. 买卖双方的交易目的均是获取各自的最大利润，可以公式化表示为

$$\begin{cases}\max\sum_{i=1}^{n}U(P_i,V_i),\\ \max\sum_{i=1}^{n}R(P_i,V_i).\end{cases} \tag{3}$$

求解此不定方程即可获得本次交易的最优 (P_i^*,V_i^*).

为了进一步演示博弈的求解流程，对买家的收益函数和卖家的成本做出以下假设. 根据文献 [21]以及数据交易场景，买家的收益取决于数据对买家的效用，故而假设数据商品 i的效用函数为

$$G_i(V_i)=S_i\times log_{a_i}\left(a_i-1+\frac{V_i}{V_{\max_i}}\right), \tag{4}$$

$$\text{s.t. } V_{\min_i}\leqslant V_i\leqslant V_{\max_i}.$$

式(4)中，$G_i(V_i)$ 是一个以交易价值量为自变量的增函数，$G_i(V_i)$ 在数值上表现为愿意为产品支付的

Table 1 Income Matrix

表 1 收益矩阵

交易平台策略	买家策略	
	接受单位价格	不接受单位价格
交易平台接受交易价值量	$(R(P,V)-z,U(P,V))$	(0,0)
交易平台不接受交易价值量	(0,0)	(0,0)

预期价格. 数值越大, 商品对消费者的效用就越大, 商品对消费者需求的满足程度越高, 消费者的支付意愿越大. S_i 是用户对商品 i 的最大支付意愿, α_i 为拟合参数, 用于表现不同的消费者对商品量的偏好度. 面对理智的消费者, 一个商品的最大效用不会超过消费者的最大支付意愿.

鉴于数据成本为固定值, 为便于讨论, 假设数据卖家的数据成本 $z=0$. 求解的目标问题变为

$$\begin{cases}\max\sum_{i=1}^{n}\left(S_i\times\log_{a_i}\left(a_i-1+\dfrac{V_i}{V_{\max_i}}\right)-P_i\times V_i\right),\\ \max\sum_{i=1}^{n}P_i\times V_i.\end{cases}\tag{5}$$

根据动态博弈的逆向归纳法, 首先分析数据买家的博弈最优解. 令 $\dfrac{\partial R(P_i,V_i)}{\partial V_i}=0$, 可得

$$\frac{S_i}{\ln 2\times(V_{\max_i}+V_i)}=P_i,\tag{6}$$

$$V_i^*(P_i)=\frac{S_i}{P_i\times\ln 2}-V_{\max_i}.\tag{7}$$

式(7)中,

$$\frac{S_i}{2\ln 2\times V_{\max_i}}<P_i<\frac{S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})}.$$

综上, 就获得了针对数据交易平台提出的不同单位价格 P_i, 买家的最优交易价值量 $V_i^*(P_i)$ 的函数表达式为

$$V_i^*(P_i)=\begin{cases}V_{\max_i},\ P_i\leqslant\dfrac{S_i}{2\ln 2\times V_{\max_i}}.\\ \dfrac{S_i}{\ln 2\times P_i}-V_{\max_i},\ \dfrac{S_i}{2\ln 2\times V_{\max_i}}<P_i<\dfrac{S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})}.\\ 0,\ P_i\geqslant\dfrac{S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})}.\end{cases}\tag{8}$$

此种情况下, 无论数据交易平台如何选择单位价格 P_i, 买家都会购买 V_i^* 数量的价值量. 因此, 交易平台的目标就是选出最优单位价格 P_i^*, 使自己收益最大化. 令

$$\max\sum_{i=1}^{n}U(P_i,V_i)=\max\sum_{i=1}^{n}P_i\times V_i.\tag{9}$$

将式(9)代入式(8)得

$$\max\sum_{i=1}^{n}U(P_i,V_i)=\sum_{i=1}^{n}\frac{S_i}{2\ln 2},\tag{10}$$

博弈的子精炼纳什均衡解为 $(S_i/V_{\max_i},V_{\max_i})$.

但是, 该收益只在买家预算充足, 即 $\sum_{i=1}^{n}\dfrac{S_i}{2\ln 2}\leqslant B$ 时成立. 在实际的交易中, 我们往往要考虑买家预算有限, 不支持每个结果集都购买 $V_{\max_i}$ 的价值量, 因此, 存在以下 3 种特殊情况.

1)情况 1. 买家选择每个结果集都要购买, 追求整体数据集均衡满足.

这种情况可以退化为一个大数据集的购买问题, 大数据集的最大支付意愿为 $\sum_{i=1}^{n}S_i$, 买家能够交易的最大价值量为 $\sum_{i=1}^{n}V_{\max_i}$, 能够交易的最小价值量为 $\sum_{i=1}^{n}V_{\min_i}$, 买家此时购买 $\sum_{i=1}^{n}S_i/(P_i\times\ln 2)-\sum_{i=1}^{n}V_{\max_i}$ 数量的价值量. 此时, 交易平台的定价区间为 $P\in\left(\sum_{i=1}^{n}S_i/\left(2\ln 2\times\sum_{i=1}^{n}V_{\max_i}\right),\sum_{i=1}^{n}S_i/\left(\ln 2\times\sum_{i=1}^{n}(V_{\max_i}+V_{\min_i})\right)\right)$ 问题的求解变为

$$\max\sum_{i=1}^{n}U(P,V)=\max\left(\frac{\sum_{i=1}^{n}S_i}{\ln 2}-\sum_{i=1}^{n}V_{\max_i}\times P\right),\tag{11}$$
$$\text{s.t.}\ \sum_{i=1}^{n}S_i\Big/\ln 2-\sum_{i=1}^{n}V_{\max_i}\times P\leqslant B,$$

解得

$$P^*=\left(\frac{\sum_{i=1}^{n}S_i}{2\ln 2}-B\right)\Big/\sum_{i=1}^{n}V_{\max_i}.\tag{12}$$

各数据集交易价值量为:

$$V_i^*=\frac{V_{\max_i}\times\left(\dfrac{\sum_{j=1}^{n}S_j}{P_i\times\ln 2}-\sum_{j=1}^{n}V_{\max_j}\right)}{\sum_{j=1}^{n}V_{\max_j}}.\tag{13}$$

此时, 买家对各个数据集的满足程度 $V_i^*/V_{\max_i}$ 都是一个相同的常数, 符合买家需求.

2)情况 2. 买家优先选择偏好数据集, 其他数据集均衡满足. 即对于其中 k 个结果集, 交易其 $V_{\max_i}$ 的价值量按照买家偏好度 $S_i/V_{\max_i}$ 决定, 其他的 $n-k$ 结果集交易 $S_i/(P_i\times\ln 2)-V_{\max_i}$ 的价值量, 且追求这 $n-k$ 个结果集整体根据偏好程度的均衡满足.

此时, 对于这 k 个结果集, 其价格 $P_i^*=S_i/2\ln 2\times V_{\max_i}$, 对于其他 $n-k$ 个数据集, 求解同情况 1. 问题的求解变为

$$\max\sum_{i=1}^{n}U(P_i,V_i)=\max\sum_{i=1}^{k}\frac{S_i}{2\ln 2}+\frac{\sum_{i=k+1}^{n}S_i}{\ln 2}-\sum_{i=k+1}^{n}V_{\max_i}P,$$
$$\text{s.t.}\ \sum_{i=1}^{k}\frac{S_i}{2\ln 2}+\sum_{i=k+1}^{n}\left(\frac{S_i}{\ln 2}-V_{\max_i}\times P_i\right)\leqslant B. \tag{14}$$

同情况 1, 解得

$$P^*=\frac{\sum_{i=1}^{n}S_i+\sum_{i=k+1}^{n}S_i}{2\ln 2}-B\Big/\sum_{i=k+1}^{n}V_{\max_i}.$$

3)情况 3. 买家追求偏好数据集得到最大程度满足, 且每个结果集都要参与交易.

在这种情况中, 将有k个结果集交易其$V_{\max_i}$的价值量, m个结果集交易其$V_{\min_i}$的价值量, 剩余结果集交易$S_i/(P_i\times\ln 2)-V_{\max_i}$的价值量. 设该结果集为$t$, 问题的求解变为式(15)和式(16).

$$\max\sum_{i=1}^{n}U(P,V)=\max\left(\sum_{i=1}^{k}\frac{S_i}{2\ln 2}+\frac{S_t}{\ln 2}-V_{\max_t}\times P_t+\sum_{i=n-m+1}^{n}\frac{V_{\min_i}\times S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})}\right), \tag{15}$$

$$\sum_{i=1}^{k}\frac{S_i}{2\ln 2}+\frac{S_t}{\ln 2}-V_{\max_t}\times P_t+\sum_{i=n-m+1}^{n}\frac{V_{\min_i}\times S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})}\leqslant B. \tag{16}$$

由式(15)和式(16)解得

$$P_t=\frac{\sum_{i=1}^{n}\frac{S_i}{2\ln 2}+\frac{S_t}{\ln 2}+\sum_{i=n-m+1}^{n}\frac{V_{\min_i}\times S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})}-B}{V_{\max_i}}.$$

综上, 对于交易平台来说, 根据数据买家的最优定价策略, 平台的最优定价策略见式(17), 本次博弈的子精炼纳什均衡解见式(18).

$$P_i^*(V)=\begin{cases}\dfrac{S_i}{2\ln 2\times V_{\max_i}}, & \sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}\leqslant B\vee\left(\sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}>B\wedge 0<i<k\right),\\ \dfrac{\sum_{j=1}^{n}S_j\big/\ln 2-B}{\sum_{j=1}^{n}V_{\max_j}}, & \sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}>B\wedge k=0,\\ \dfrac{\sum_{j=1}^{k}\frac{S_j}{2\ln 2}+\frac{S_t}{\ln 2}+\sum_{j=n-m+1}^{n}\frac{V_{\min_j}\times S_j}{\ln 2\times(V_{\max_j}+V_{\min_j})}-B}{V_{\max_i}}, & \sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}>B\wedge 0<k<i<m,\\ \dfrac{S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})}, & \sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}>B\wedge 0<m\leqslant i.\end{cases} \tag{17}$$

$$(P_i^*,V_i^*)=\begin{cases}\left(\dfrac{S_i}{2\ln 2\times V_{\max_i}},V_{\max_i}\right), & \sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}\leqslant B\vee\left(\sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}>B,0<i\leqslant k\right),\\ \left(\dfrac{\sum_{j=1}^{n}S_j\big/\ln 2-B}{\sum_{j=1}^{n}V_{\max_j}},\dfrac{V_{\max_i}\times\left(\sum_{j=1}^{n}S_j\big/(P_i\times\ln 2)-\sum_{j=1}^{n}V_{\max_j}\right)}{\sum_{j=1}^{n}V_{\max_j}}\right), & \sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}>B\wedge k=0\wedge m=0,\\ \left(\dfrac{\sum_{j=1}^{k}\frac{S_j}{2\ln 2}+\frac{S_t}{\ln 2}+\sum_{j=n-m+1}^{n}\frac{V_{\min_j}\times S_j}{\ln 2\times(V_{\max_j}+V_{\min_j})}-B}{V_{\max_i}},\dfrac{\sum_{j=1}^{n}V_{\max_j}\times\left(B-\frac{\sum_{j=1}^{k}S_j}{2\ln 2}-\frac{V_{\min_i}\times S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})}\right)}{\left(\sum_{j=1}^{n}S_j+\sum_{j=k+1}^{n}S_j\right)\big/2\ln 2-B}\right), & \sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}>B\wedge 0<m\wedge 0<k<i<n-m+1,\\ \left(\dfrac{S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})},V_{\min_i}\right), & \sum_{j=1}^{n}\dfrac{S_j}{2\ln 2}>B\wedge 0<i\leqslant m.\end{cases} \tag{18}$$

2.5 劳资博弈定价算法

劳资博弈定价具体实现如算法 1 所示. 算法 1 中行①~③确定边界值, 行④~⑦计算预算充足情况下每个数据集的交易量及价格, 行⑧~㉕实现在预算不充足、买家不同偏好度及购买策略情况下每个数据集交易量及价格的计算. 其中行⑩~⑫, ⑬~⑱, ⑲~㉓分别对应 2.4 节中情况 1、情况 2 和情况 3 买家 3 种不同需求下交易量及价格的计算.

算法 1. 劳资博弈定价算法.

输入: 数据源集合 *Set*, 买家总预算 B, 效用函数参数 α, 买家的购买策略 *STRATEGY*;

输出: 最优定价组合 $priceSet=\left\{\left(P_i^*,V_i^*\right)\middle|1\leqslant i\leqslant n\right\}$.

① for each(*ds*: *Set*)/*根据式(10)确定纳什均衡解*/
② bv+=*ds.s*/2lnα; /**bv* 为边界值变量*/
③ end for
④ if($B\geqslant bv$){ /*买家预算充足*/
⑤ for each(*ds*: *Set*)
⑥ *priceSet.put*(*getBestP*(*ds*), *ds*.$V_{\max}$); /*根据式(10)计算数据集 *ds* 交易最大价值量时的价格*/
⑦ end for
⑧ }else if($B<bv$){/*买家预算不足*/
⑨ switch(*STRATEGY*);
⑩ case: *BALANCE*{/*情况 1: 据式(12)(13)计算得到各结果集的均衡价格与交易价值量*/
⑪ for each(*ds*: *Set*)
⑫ *priceSet.put*(*getBalanceP*(*B*, *Set*), *getBalanceV*(*ds*));
⑬ }case: *PREBALANCE*{/*情况 2*/
⑭ *BuyAllVMin*(*Set*, *B*);
⑮ *BuyPreferences*(*Set*, *B*);
⑯ for(*ds*: *Set*)/*计算剩余数据集的均衡交易价值量及价格*/
⑰ *priceSet.update*(*getBalanceP*(*B*, *Set*), *getBalanceV*(*ds*));
⑱ end for
⑲ }case: *PREFERENCE*{/* 情况 3*/
⑳ *BuyAllVMin*(*Set*, *B*);
㉑ *BuyPreferences*(*Set*, *B*);
㉒ *priceSet.update*(*getBetterP*(*B*, *ds*), *getBetterV*(*P*, *ds*)); /*根据式(15)计算预算不足时的购买价值量及价格*/
㉓ }
㉔ end switch
㉕ }
㉖ end if
㉗ return(*priceSet*).

算法 2 完成每个数据集最低价值量及对应价格的博弈计算, 实现所有数据集全部购买情况下的最低购买需求. 其中行①~③计算每个数据集的最低价值量及对应价格, 行④~⑥计算在所有数据集都最低价值量交易时交易后的剩余预算.

算法 2. 以最低价值量购买所有数据 *BuyAllVMin*.

输入: 数据源集合 *Set*, 买家总预算 B;

输出: *priceSet*, *B*.

① for each(*ds*: *Set*)
② *priceSet.put*(*getLowestP*(*ds*),*ds*.$V_{\min}$); /*依式(15)计算数据集 *ds* 的最低购买价值量及对应价格*/
③ end for
④ for each(int *p*: *priceSet*)
⑤ *B*– =*p*; /*剩余预算*/
⑥ end for
⑦ return *priceSet*, *B*.

算法 3 完成以数据集偏好为优先购买顺序, 对每个数据集购买的价值量和应支付的价格进行计算. 行①~⑧按数据集购买的偏好优先级遍历所有数据集, 并扣除购买该数据集最大价值量的价格(行⑤⑥), 直至预算用尽(行②③), 返回结果.

算法 3. 按偏好购买数据集 *BuyPreferences*.

输入: 数据源集合 *Set*, 买家总预算 B;

输出: *priceSet*, *B*.

① for each(*ds*: *Set*){/*按偏好遍历数据集*/
② if(*getBestP*(*ds*)>*B*+*priceSet.getP*(*ds*))
③ break; /*剩余预算不足以购买 *ds* 的最大价值量*/
④ end if
⑤ *B*– = *getBestP*(*ds*)–*priceSet.getP*(*ds*); /*计算剩余预算*/
⑥ *priceSet.update*(*getBestP*(*ds*), *ds*.$V_{\max}$); /*据式(10)计算每个数据集最优购买价值量及对应价格*/
⑦ }
⑧ end for
⑨ return *priceSet*, *B*.

不难看出，劳资博弈定价算法的时间复杂度为$O(N)$，其中N为候选数据集的个数. 考虑到数据交易现实情况，N远远小于数据量，因此理论上讲，该算法执行时间几乎可忽略不计.

2.6 无套利性验证

无套利原则的定义源于信息理论确定性的概念. 假如V是数据集D的一个视图集，当一个数据集D的查询$Q(D)$可以通过$Q(V)$来获取结果时，那么由视图集V确定查询结构Q可表示为$D \vdash V \twoheadrightarrow Q$.

假设R和T是数据集D的2个视图(结果集)，且满足$D=R \bowtie T$，3个数据集的价格分别是$(D,P_1),(R,P_2),(T,P_3)$. 考虑一种情况，若$P_1 \geqslant P_2+P_3$，那么聪明的数据买家会发现购买数据集R和T可以用更低的价格获得数据集D，这将会损害数据交易平台的利益. 因此，有效的定价函数应遵循无套利原则，即当数据集D的查询结果Q可由查询束$Q_1,Q_2,\cdots,Q_k$得到时，定价函数$P=G(Q)$须满足

$$\sum_{i=1}^{k} G(Q_i) \geqslant G(Q). \tag{19}$$

定义 1. 无套利性. 当数据集D的查询结果Q可由查询束$Q_1,Q_2,\cdots,Q_k$得到时，买家获取到的单位价值量V必须满足

$$\sum_{i=1}^{k} R(V_i) \leqslant R(V). \tag{20}$$

式(20)的意义为多个子查询结果对买家的总收益不得大于单查询结果对买家的总收益，否则就有可能发生套利行为.

其次，为了确保本定价算法的合理性，在进行多个子查询时，交易的输入参数必须满足：

$$S \leqslant \sum_{i=1}^{k} S_i. \tag{21}$$

式(21)中，S为每个查询结果集的最大支付意愿. 该公式的意义为一个查询束获取到的各个子结果集的最大支付意愿之和不得小于等价查询Q获取的结果集的最大支付意愿.

由式(8)的劳资博弈定价结果可以得知，根据买家预算的不同，买家可获取到的单位价值量如式(22)所示.

$$V=\begin{cases}\displaystyle\sum_{i=1}^{n} V_{\max_i}, & \displaystyle\sum_{i=1}^{n}\frac{S_i}{2\ln a_i} \leqslant B.\\[2ex] \dfrac{V_{\max_i}\times\left(\displaystyle\sum_{i=1}^{n} S_i \Big/ (P_i\times\ln 2)-\sum_{i=1}^{n} V_{\max_i}\right)}{\displaystyle\sum_{i=1}^{n} V_{\max_i}}, & \displaystyle\sum_{i=1}^{n}\frac{S_i}{2\ln a_i} > B \wedge k=0 \wedge m=0,\\[2ex] \displaystyle\sum_{i=1}^{k} V_{\max_i}+\frac{\displaystyle\sum_{i=k+1}^{n} V_{\max_i}\left(B-\sum_{i=1}^{k} S_i \Big/ 2\ln 2\right)}{\left(\displaystyle\sum_{i=1}^{n} S_i+\sum_{i=k+1}^{n} S_i\right)\Big/ 2\ln 2-B}-\sum_{i=k+1}^{n} V_{\max_i}, & \displaystyle\sum_{i=1}^{n}\frac{S_i}{2\ln a_i} > B \wedge k\neq 0 \wedge m=0,\\[2ex] \displaystyle\sum_{i=1}^{k} V_{\max_i}+\sum_{i=1}^{m} V_{\min_i}+\frac{V_{\max_i}\left(B-\dfrac{\displaystyle\sum_{i=1}^{n} S_i}{2\ln 2}-\dfrac{V_{\min_i}\times S_i}{\ln 2\times(V_{\max_i}+V_{\min_i})}\right)}{\left(\displaystyle\sum_{i=1}^{n} S_i+\sum_{i=k+1}^{n} S_i\right)\Big/ 2\ln 2-B}, & \displaystyle\sum_{i=1}^{n}\frac{S_i}{2\ln a_i} > B \wedge k\neq 0 \wedge m\neq 0.\end{cases} \tag{22}$$

根据式(22)可知，买家单位价值量获取的影响因素有3个，分别是预算B、各数据集的最大支付意愿S_i和各数据集的最大单位价值量$V_{\max_i}$. 接下来我们将对这3个因素逐个分析.

1)预算B的分析. 预算B在买家一开始进入交易市场中就已经确定，其只与买家的最终目标有关，理智的买家在交易市场中无论采取何种交易行为，其交易目标都是确定的，因此预算B不随交易行为的改变而改变.

2)数据集的最大支付意愿S_i的分析. 在式(21)中，我们可知$S \leqslant \sum_{i=1}^{k} S_i$，根据式(22)，当买家预算充足时，买家获取的单位价值量与S_i无关，当买家预算不足时，由式(22)可知，S_i的增大将导致获取到的单位价值量减少，此时买家拆分成子查询束的行为是不明智的，故明智的买家不会拆分成子查询束.

3)数据集的最大单位价值含量$V_{\max_i}$的分析. 当查询结果Q可由查询束$Q_1, Q_2, \cdots, Q_k$得到时，即查询结果集$R = R_1 \bowtie R_2 \bowtie \cdots \bowtie R_k$时，各结果集的自然连接结果依然是原数据集$R$，也就是说，买家将1个查询分解为多个子查询，其最终获取到的数据集是相同的，拆分之后各数据集的单位价值量之和$\sum_{i=1}^{n} V_{\max_i}$发生变化，不会影响其最终获取到的单位价值量.

综上所述，本文定价算法满足$\sum_{i=1}^{k} R(V_i) \leqslant R(V)$，满足查询定价的无套利性要求.

3 实验结果分析

1)实验内容. 本节对基于劳资博弈的查询定价算法(以下简称查询-劳资博弈)进行实验验证分析. 鉴于其无套利性已在2.6节分析，本节主要验证其定价动态灵活性、公平合理性和高效性. 为表明该定价算法具备无需预设价格点的优势，本节实验采用附录A中介绍的隐私含量作为数据集的价值量.

2)基准测试. 基于查询的定价方法侧重于定价的灵活性，而基于斯塔克伯格博弈的定价方法侧重于定价的合理性. 因此，本文研究使用查询定价方法作为对比算法，来验证查询-劳资博弈定价算法的动态灵活性，同时使用斯塔克伯格博弈定价方法来分析其公平合理性. 为说明查询-劳资博弈定价算法的有效性，本文研究分别采用最新的查询定价方法[8]和斯塔克伯格博弈定价方法[13]作为基准测试算法.

虽然关系数据库具备高性能的特性，但由于本文研究是第1个提出多数据集组合购买定价问题的研究，同基于关系数据库的传统查询定价方法研究问题不同. 因此，在分析性能时，本文研究所提方法同传统查询定价方法不具备直接对比性. 基于此，在分析查询-劳资博弈定价算法性能时，本文研究采用相对分析手段，将从劳资博弈定价算法的执行导致的查询-劳资博弈完整执行时间的增加比例入手，对算法性能进行分析.

3)数据集. 实验数据集包含1个真实数据集和1个标准模拟数据集. 真实数据集用于对算法合理性和灵活性进行分析，标准模拟数据集用于对算法性能进行分析. 真实数据集采用从Kaggle网站上获取的名为personal data的test_data公开数据集[22]. 为确保该真实数据集的代表性，本文研究调研了京东万象等数据交易平台，从该数据集中抽取出10列交易频率高的个人基本信息项，构造出一个8 756行、40列的实验数据集，并基于该真实数据集，派生出4个对比数据集. 另外，由于涉及到隐私，本文研究对测试集中具有高度指向性的数据进行了脱敏处理. 模拟数据集选用TPC-H标准测试集[23].

4)实现及实验平台. 所有算法用Java语言实现. 由于查询-劳资博弈定价算法是对数据库查询的定价，理论上来说，可以对任意关系数据库查询结果进行定价，考虑到代表性，数据的存储和查询基于流行的开源关系数据库MySQL 5.17进行. 实验平台配置为：CPU 2.5 GHz 4核 Intel Core i7，内存16 GB 1 600 MHz DDR3，磁盘500 GB SSD.

3.1 定价动态灵活性验证

本实验利用SQL语句查询得到2个数据集. 为了实验的严谨性，本实验复制了这2个数据集，采取控制变量原则，仅调整了2个数据集的隐私含量，生成了4个数据集，分别命名为数据集D_1、数据集D'_1、数据集D_2、数据集D'_2. 其中，数据集D_1和数据集D'_1、数据集D_2和数据集D'_2的数据完全相同，但是其隐私含量不同. 这样的实验设置可以避免因数据变化造成的查询定价结果变化. 数据集详情如表2所示.

Table 2 The Table of Dataset Detailed Information

表2 数据集详情表

数据集	数据条数	隐私含量
D_1	20 000	26 756
D'_1	20 000	18 679
D_2	30 000	47 839
D'_2	30 000	29 375

对算法动态灵活性的验证分析，本文研究选取经典的基于查询的定价方法(查询定价)作为基准测试. 2种定价方法的比较结果如图1所示.

从图1可得，由于数据集D_1和数据集D'_1数据相同、隐私含量不同，传统查询定价方法给这2个数据

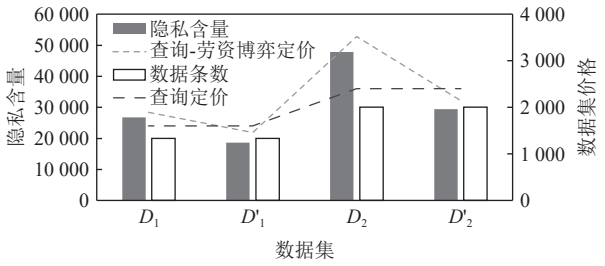

Fig. 1 Comparison results of query-labor game pricing and traditional query pricing

图1 查询-劳资博弈定价与传统查询定价的结果对比

集定价相同，查询-劳资博弈定价算法给出的定价随数据集隐私含量的变化呈正比例变化. 数据集 D_2 和数据集 D_2' 数据相同、隐私含量不同，现有的查询定价方法给这 2 个数据集定价相同，本文研究提出的查询-劳资博弈定价随着数据集隐私含量的变化呈正比例变化. 结果表明，相比于现有的查询定价方法，查询-劳资博弈定价算法可以跟随数据价值的变化动态调整价格，因此更具优越性.

3.2 定价公平合理性验证

公平合理性验证选择斯塔克伯格博弈模型作为基准对比算法. 假设斯塔克伯格博弈模型和劳资博弈模型双方采用相同的收益函数. 买家需求如表 3 所示.

Table 3 Experimental Datasets

表 3 实验数据集

数据集	V_{max}	V_{min}	S	S/V_{max}
D_1	10 000	7 000	700	0.080
D_2	20 000	4 000	400	0.020
D_3	30 000	5 000	500	0.017

1)设买家预算 B=1 500. 此时买家预算充足，基于斯塔克伯格博弈定价方法的定价结果如表 4 所示. 本文研究提出的查询-劳资博弈定价算法的定价结果如表 5 所示.

Table 4 Pricing Results of Stackelberg Game Pricing Algorithm Under Sufficient Budget

表 4 预算充足时斯塔克伯格博弈定价算法定价结果

数据集	买家花费或平台收益	买家所获隐私含量	买家需求满足程度/%	买家额外收益
D_1	656.25	10 000	100	43.75
D_2	375.00	20 000	100	25.00
D_3	468.75	30 000	100	31.25

Table 5 Pricing Results of Query-Labor Game Pricing Algorithm Under Sufficient Budget

表 5 预算充足时查询-劳资博弈定价算法定价结果

数据集	买家花费或平台收益	买家获得的隐私含量	买家需求满足程度/%	买家额外收益
D_1	505	10 000	100	195
D_2	288	20 000	100	112
D_3	360	30 000	100	140

图 2 和图 3 分别显示了斯塔克伯格博弈模型和劳资博弈模型在平台收益和买家收益 2 个关键指标的对比.由结果可知，2 个定价方法对于买家的需求满足程度都达到了 100%，但是基于斯塔克伯格博弈

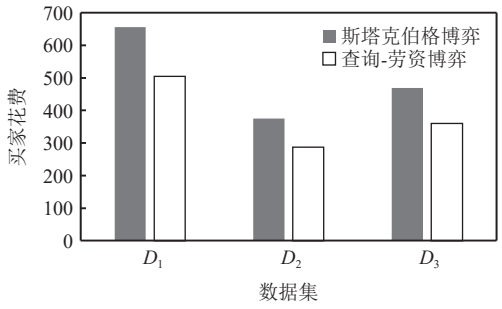

Fig. 2 Platform profits comparison

图 2 平台收益对比

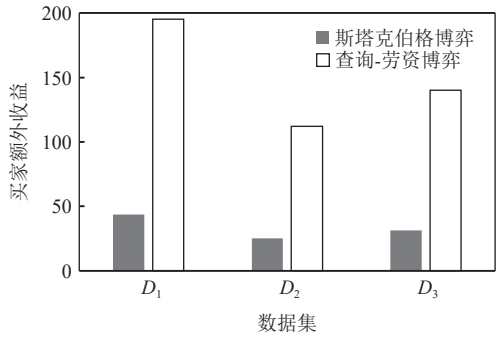

Fig. 3 Buyer extra profits comparison

图 3 买家额外收益对比

的定价方法最后买家收益为 100，基于查询-劳资博弈的定价算法最后买家收益为 447. 基于斯塔克伯格博弈的定价方法由于结果集单位价格 P 及交易价值量 V 都由卖家设置，因此，在预算充足的情况下，相比于斯塔克伯格博弈定价算法，查询-劳资博弈定价算法对买家更为有利，更容易促成交易.

2)设买家预算 B=1 000. 买家购买策略为优先购买策略即选择每个数据集都要购买且优先满足偏好度高的数据集.

此时买家预算不足，基于斯塔克伯格博弈定价算法的定价结果如表 6 所示. 基于劳资博弈的查询定价算法的定价结果如表 7 所示.

图 4 和图 5 分别显示了斯塔克伯格博弈模型和劳资博弈模型在买家所获得的隐私含量(价值量)和

Table 6 Pricing Results of Stackelberg Game Pricing Algorithm with Priority Strategy Under Insufficient Budget

表 6 预算不足且采取优先策略时斯塔克伯格博弈定价算法的定价结果

数据集	买家花费或平台收益	买家获得的隐私含量	买家需求满足程度/%	买家额外收益
D_1	437.5	7 000	70.00	0
D_2	250	4 000	20.00	0
D_3	312.5	5 000	16.67	0

Table 7 Pricing Results of Query-Labor Game Pricing Algorithm with Priority Strategy Under Insufficient Budget

表 7 预算不足且采取优先策略时查询-劳资博弈定价算法的定价结果

数据集	买家花费或平台收益	买家获得的隐私含量	买家需求满足程度/%	买家额外收益
D_1	505	10 000	100	195
D_2	288	20 000	100	112
D_3	207	16 829	56.10	0

需求满足程度的对比. 由表 6 和表 7 中结果可知, 2 个博弈方法交易平台的收益相同, 都为 1 000, 但是对比之下, 2 个方法的买家需求满足程度不同, 基于斯塔克伯格博弈的定价方法中买家的需求满足程度为 70%, 20%, 16.67%, 整体小于查询-劳资博弈定价算法中买家的需求满足程度 100%, 100%, 56.10%. 此外, 斯塔克伯格博弈方法并不能按照用户的偏好去优先满足偏好度高的数据集. 因此, 在预算不足的情况下, 查询-劳资博弈定价算法比斯塔克伯格博弈定价算法对买家更有利, 更易促成交易的达成.

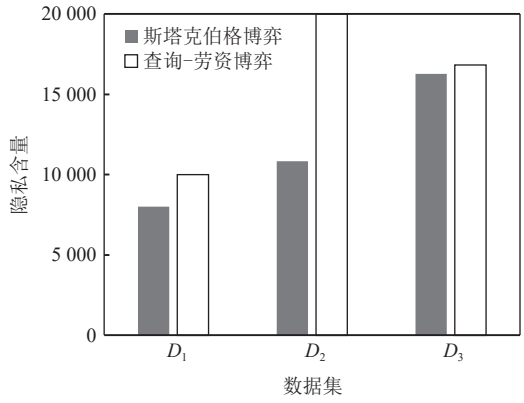

Fig. 4 Comparison of privacy content obtained by buyers

图 4 买家获得的隐私含量对比

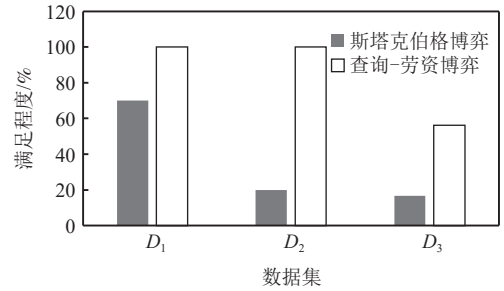

Fig. 5 Comparison of the satisfaction degree of buyers' needs

图 5 买家需求满足程度对比

综上, 本文研究提出的查询-劳资博弈定价算法更加兼顾买卖双方的合理利益, 定价更加公平、更易促成交易的达成.

3.3 性能分析

本实验基于 TPC-H 数据集派生出 5 份价值量不同的对比数据集, 并假定用户需求场景为预算不足且购买策略为各数据集均衡购买. 为准确控制实验参数, 每个元组价值量参数设置为相同值. 如前文所述, 因难以找到合适的直接对比基准算法, 本节采取相对分析方法, 以查询-劳资博弈定价算法在查询总执行时间的占比作为分析指标. 以 TPC-H 每组查询的第 2 个查询为测试查询, 不同扩展因子(scale factor, SF)参数对应的实验结果如图 6 所示.

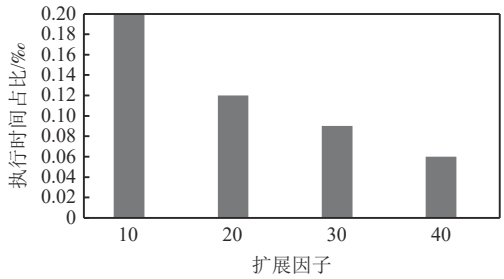

Fig. 6 Performance evaluation of the query-labor game pricing algorithm

图 6 查询-劳资博弈定价算法执行性能评价

从图 6 可以看出, 在各数据规模下, 查询-劳资博弈定价算法执行时间占整个查询执行时间的比例均在 0.2‰以下, 甚至更低, 且数据量越大, 其执行时间占比越低, 其执行均可在数毫秒内完成, 这同其算法复杂度 $O(N)$ 的理论值是一致的, 说明本文研究所提定价方法具备高效的优良特性.

4 结 论

针对现有查询定价算法落地难的挑战, 本文研究创新利用劳资博弈模型对数据交易市场进行建模, 据此设计了基于劳资博弈的查询定价算法. 该算法权衡考虑了各参与方的利益, 利用劳资博弈给出交易数据价格, 具有动态灵活、公平透明、易应用等优良特性. 对比传统查询定价和斯塔克伯格博弈模型, 大量实验验证了本文研究定价算法的高效性、优越性和实际落地性.

作者贡献声明: 王会举提出论文研究方向、思路及参与论文的修改; 黄玮煊负责实验及实现; 岳晓提出指导意见并修改论文.

参 考 文 献

[1] CPC Central Committee and the State Council. “14th five year plan” digital economic development plan[EB/OL]. (2022-03-25) [2023-05-12]. https://www.gov.cn/gongbao/content/2022/content_5671108.htm?eqid=8776104300000c760000000664564e72&eqid=cc7670d30007fc920000000664867f05 (in Chinese)
(中共中央国务院. “十四五”数字经济发展规划[EB/OL]. (2022-03-25)[2023-05-12]. https://www.gov.cn/gongbao/content/2022/content_5671108.htm?eqid=8776104300000c760000000664564e72&eqid=cc7670d30007fc920000000664867f05)

[2] Koutris P, Upadhyaya P, Balazinska M, et al. Query-based data pricing[C]//Proc of the 34th SIGMOD Symp on Principles of Database Systems (PODS). New York: ACM, 2012: 167−178

[3] Tang Ruiming, Wu Huayu, Bao Zhifeng, et al. The price is right: Models and algorithms for pricing data[C]//Proc of the 24th Int Conf on Database and Expert Systems Applications. Berlin: Springer, 2013: 380−394

[4] Balazinska M, Howe B, Suciu D. Data markets in the cloud: An opportunity for the database community[J]. Proceedings of the VLDB Endowment, 2011, 4(12): 1482−1485

[5] Lin Bingrong, Kifer D. On arbitrage-free pricing for general data queries[J]. Proceedings of the VLDB Endowment, 2014, 7(9): 757−768

[6] Upadhyaya P, Balazinska M, Suciu D. Price-optimal querying with data apis[J]. Proceedings of the VLDB Endowment, 2016, 9(14): 1695−1706

[7] Deep S, Koutris P. The design of arbitrage-free data pricing schemes[C]//Proc of the 20th Int Conf on Database Theory (ICDT). Wadern: Schloss Dagstuhl, 2017: 12: 1−12: 18

[8] Deep S, Koutris P. Qirana: A framework for scalable query pricing[C]//Proc of the 43rd ACM Int Conf on Management of Data. New York: ACM, 2017: 699−713

[9] Deep S, Koutris P, Bidasaria Y. Qirana demonstration: Real time scalable query pricing[J]. Proceedings of the VLDB Endowment, 2017, 10(12): 1949−1952

[10] Li Chao, Li Y D, Miklau G, et al. A theory of pricing private data[J]. ACM Transactions on Database System, 2014, 39(4): 1−28

[11] Chen Chen, Yuan Ye, Wen Zhenyu, et al. GQP: A framework for scalable and effective graph query-based pricing[C]//Proc of the 38th IEEE Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2022: 1573−1585

[12] Haddadi S, Ghasemi A. Pricing-based Stackelberg game for spectrum trading in self-organised heterogeneous networks[J]. IET Communications, 2016, 10(11): 1374−1383

[13] Xu Chengzhen, Zhu Kun, Yi Changyan, et al. Data pricing for blockchain- based car sharing: A stackelberg game approach[C/OL]//Proc of IEEE Global Communications Conf. Piscataway, NJ: IEEE, [2023-05-12]. https://ieeexplore.ieee.org/document/9322221

[14] Yuan Ying, Wang Cuirong, Wang Cong, et al. An uncompleted information game based resources allocation model for cloud computing[J]. Journal of Computer Research and Development, 2016, 53(6): 1342−1351(in Chinese)
(苑迎, 王翠荣, 王聪, 等. 基于非完全信息博弈的云资源分配模型[J]. 计算机研究与发展, 2016, 53(6): 1342−1351)

[15] Berz G. Game Theory Bargaining and Auction Strategies: Practical Examples from Internet Auctions to Unvestment Banking[M]. Berlin: Springer, 2016

[16] Riederer C, Erramilli V, Chaintreau A, et al. For sale: your data: by: you[C]//Proc of the 10th ACM Workshop on Hot Topics in Networks. New York: ACM, 2011: 13: 1−13: 6

[17] Susanto H, Zhang Honggang, Ho S Y, et al. Effective mobile data trading in secondary ad-hoc market with heterogeneous and dynamic environment[C]//Proc of the 37th IEEE Int Conf on Distributed Computing Systems (ICDCS). Los Alamitos, CA: IEEE Computer Society, 2017: 645−655

[18] Ghosh A, Roth A. Selling privacy at auction[C]//Proc of the 12th ACM Conf on Electronic Commerce. New York: ACM, 2011: 199−208

[19] Yang Jian, Zhao Chongchong, Xing Chunxiao. Big data market optimization pricing model based on data quality[J]. Complexity, 2019(2): 1−10

[20] Li Xijun, Yao Jianguo, Liu Xue, et al. A first look at information entropy-based data pricing[C]//Proc of the 37th IEEE Int Conf on Distributed Computing Systems (ICDCS). Piscataway, NJ: IEEE, 2017: 2053−2060

[21] Yu Haifei, Zhang Mengxiao. Data pricing strategy based on data quality[J]. Computers & Industrial Engineering, 2017(112): 1−10

[22] Kaggle. Personal dataset [EB/OL]. [2023-04-11]. https://www.kaggle.com/ datasets/wangtaobo/personal-data-datasets-for-big-data-research

[23] TPC. TPC-H benchmark[EB/OL]. [2023-10-11]. http://www.tpc.org/tpch/

Wang Huiju, born in 1979. PhD, associate professor, master supervisor. Member of CCF. His research interests include big data, high performance database, graph database, and AI.
王会举, 1979 年生. 博士, 副教授, 硕士生导师. CCF 会员. 主要研究方向为大数据、高性能数据库、图数据库、人工智能.

Huang Weixuan, born in 1999. Master. His main research interests include data pricing and big data.
黄玮煊, 1999 年生. 硕士. 主要研究方向为数据定价、大数据.

Yue Xiao, born in 1980. PhD, associate professor, master supervisor. Her main research interests include big data and data trading.
岳 晓, 1980 年生. 博士, 副教授, 硕士生导师. 主要研究方向为大数据、数据交易.

附录 A.

个体对于隐私价值的衡量是复杂的，在数据交易中，隐私价值取决于买家态度、卖家用途、数据时效性等多方面因素，这些方面往往是主观且复杂的，为了保证隐私衡量算法的普适性，学者们往往会根据研究问题的情景有所侧重地简化个体权衡过程. 在现有的研究中，Clauß等人[1]考虑到隐私信息是信息的一种特殊类别，开始研究利用信息熵计算隐私信息的信息量，从而量化隐私价值. 彭慧波等人[2]发现仅使用信息熵来衡量隐私价值是不合理的，因此引入了隐私主体分级修正信息熵算法量化的隐私价值. 然而，彭慧波等人[2]仅考虑量化数据主体的隐私价值，没有考虑到数据的隐私价值还与隐私信息和隐私信息之间的关联有关①. 为解决此不足，本文研究基于文献 [2] 的研究，进一步评估了数据关联隐私价值，以更充分衡量数据隐私价值.

根据文献 [2] 的研究，数据主体隐私价值可以用如下公式来量化：

$$\theta_{x_i} = O_{x_i} \times H(x_i), \tag{A1}$$

其中，θ_{x_i}是元组x_i的隐私主体含量，O_{x_i}是元组x_i的隐私主体等级，$H(x_i)$是元组x_i在以信源为数据集x时所求得的个体熵.

量化得到数据主体隐私价值后，下一步的工作是量化数据的关联隐私价值. Osothongs 等人[3]提供了一种通过贝叶斯公式量化信息间披露关系的方法.

贝叶斯公式是计算条件概率的一种常见方法，常用于计算给定观测值的后验概率. 本文研究将个人信息分为保密、公开 2 类，贝叶斯公式可以计算当一条信息公开后，另一条信息的后验概率，实现隐私关联影响的量化计算.

设x,y为某人的 2 项不同信息，则用户选择公开x后继续公开y的概率为

$$P(y|x) = \frac{P(x \cap y)}{P(x)}, \tag{A2}$$

其中，$P(x \cap y)$表示用户同时公开x,y这 2 项信息的概率，$P(x)$表示用户公开x的概率.

在得出数据集中所有信息项两两之间的隐私关联后，可以利用有向图来可视化信息项之间的隐私关联，设节点x,y为不同的 2 个信息项，$P(y|x) > P(x|y)$. 节点x与节点y形成的边$\langle x,y\rangle$为 2 个信息项之间的隐私关联，边的方向由高价值信息项节点x指向低价值信息项节点y②. 边的权重$w(x,y) = P(y|x)$.

假设数据集$D = \{x,y,z,k,j\}$共 5 个属性，5 个属性两类之间的后验概率关系如表 A1 所示.

Table A1 Posteriori Probability Relation Between Attributes of Dataset *D*

表 A1 数据集D的属性之间的后验概率关系

先披露属性	对后披露属性的影响				
	x	y	z	j	k
x		$P(y\|x)$		$P(j\|x)$	$P(k\|x)$
y			$P(z\|y)$		$P(k\|y)$
z				$P(j\|z)$	
j					
k				$P(j\|k)$	

将数据集中所有的属性都建立节点形成有向图G，如图 A1 所示.

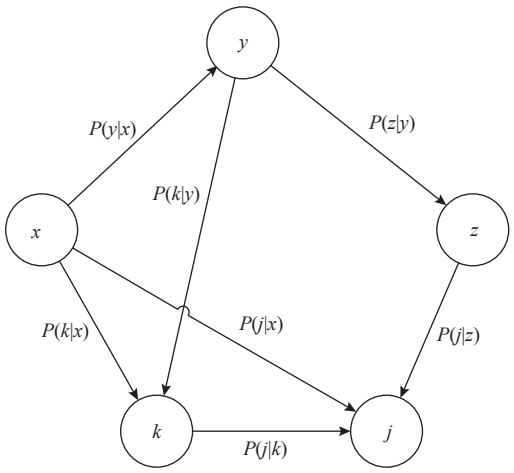

Fig. A1 Node digraph of dataset *D*

图 A1 数据集D的节点有向图

构建出图G后，为方便更好地显示信息间的披露影响，可以将图 G 转为为一个森林结构，如图 A2 所示.

$G_f = \langle V,E,W\rangle$是一个代表个人信息关系的图，其中$f$是所关注的个人信息项，$V = \{v_i, v_{i+1}, \cdots, v_{n-1}, v_n\}$是图$G_f$的节点集合，$E = \{e_{i,i+1}, e_{i+1,i+2}, \cdots, e_{n-2,n-1}, e_{n-1,n}\}$是图$G_f$的边集合，$W = \{w_e | e \in E\}$是图$G_f$中每一条边的权重集合.

假设v_n是根节点，v_i是其他节点. 若v_n通向v_i的通

① 举个例子，一位用户的电话号码和购物偏好被同时公开所带来的商业价值一定大于该用户单独公开 2 条信息的价值之和.

② 在现实中，低价值信息的公开一般不影响是否公开高价值信息的决策.

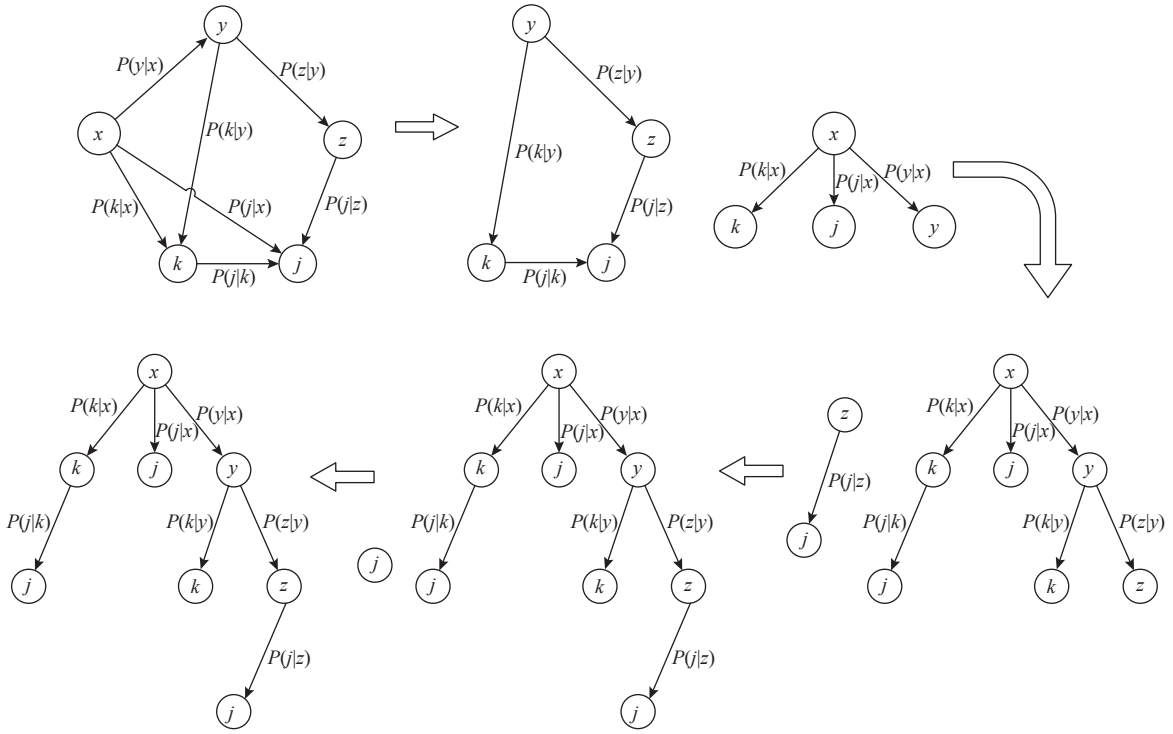

Fig. A2 Forest map formed after dataset *D* transformation

图 A2 数据集*D*转化后形成的森林图

路为$v_i, v_{i+1}\cdots, v_{n-1}, v_n$，则信息项$v_n$的公开对信息项$v_i$公开的影响为：

$$W_{n,i} = W_{e_{v_n,v_i}}. \tag{A3}$$

若个人数据集X中存在一个数据元组$x_i = \{x_{i1}, x_{i2}, \cdots, x_{ij}\}$，含有的信息关联隐私含量为

$$\sigma_{x_i} = \sum_{k=1}^{j} W_{x_i,x_k}, \tag{A4}$$

其中，σ_{x_i}为元组x_i的信息关联隐私含量，$x_i \neq x_j$.

综上，若存在一个个人数据集$X = \{x_1, x_2, \cdots, x_i\}$，则该数据集的隐私含量为

$$\theta = \alpha \times normal(\theta_{x_i}) + \beta \times normal(\sigma_{x_i}), \tag{A5}$$

其中，θ为数据集X的隐私含量，α和β为调整参数，*normal*()为量纲统一函数.

附录 A 参考文献

[1] Clauß S, Schiffner S. Structuring anonymity metrics[C]//Proc of the 2nd ACM Workshop on Digital Identity Management. New York:ACM, 2006: 55−62

[2] Peng Huibo, Zhou Yajian. A pricing model based on privacy measurement[J]. Software, 2019, 40(1): 57−62(in Chinese)
(彭慧波, 周亚建. 基于隐私度量的数据定价模型 [J].软件, 2019, 40(1): 57−62

[3] Osothongs A, Suppakitpaisarn V, Sonehara N. A proposed method for personal attributes disclosure valuation: A study on personal attributes disclosure in Thailand[C]//Proc of the 7th Int Conf on Information Technology and Electrical Engineering (ICITEE). Piscataway, NJ: IEEE, 2015: 408−413