doi:10.3772/j.issn.2095-915x.2024.04.003

考虑数据维度与质量维度二重性的数据质量评估研究

开放科学（资源服务）标识码（OSID）

刘枏[1]　刘润[1]　李卓城[2]

1. 重庆交通大学经济与管理学院　重庆　400074；
2. 北京邮电大学玛丽女王海南学院　北京　572400

摘要：[目的 / 意义] 针对数据开发中关联性差、完整性不足所引发的高成本问题，本文构建了考虑数据维度与质量维度的数据质量评估模型，旨在评估数据的各维度质量并进行价值分析。[方法 / 过程] 首先通过数据质量相关的文献分析，利用修正德尔菲法构建了数据质量维度指标体系；之后使用层次分析法选取指标，采用蒙特卡洛采样计算 Shapley 值进而构建了针对数据不同维度的质量评估模型。在房地产数据质量评估案例中，对其 12 个数据维度运用质量评估模型进行评估，结果显示其中三个维度质量较差，建议删除，同时针对其他维度提出了修改意见。[局限] 未解决数据质量维度指标难以量化问题，也未对数据进行分类，仅提出笼统的评估模型。[结果 / 结论] 将数据维度和数据质量维度进行融合，并基于数据质量维度指标体系及 Shapley 构建了一个较为全面的数据质量评估模型，能更为详细地对数据库进行质量评估。

关键词：质量评估模型；数据维度；数据质量维度

中图分类号：G35；F49；TP391.1

Research on Data Quality Evaluation Considering the Duality of Data Dimension and Quality Dimension

LIU Nan[1]　LIU Run[1]　LI Zhuocheng[2]

1. School of Economics and Management, Chongqing Jiaotong University, Chongqing 400074, China;
2. Queen Mary Hainan College, Beijing University of Posts and Telecommunications, Beijing 572400, China

Abstract: [Objective/Significance] This article constructs a dimensional evaluation model based on Shapley value to address the high cost issues caused by poor relevance and insufficient integrity in data development. The purpose is to evaluate the quality of

基金项目　教育部基金项目“大数据资产的定价机理、方法和规范研究”（20XJAZH007）；重庆市社会科学规划基金项目“西部陆海新通道跨境数据互联互通机制研究”（2023ZDLH11）。

作者简介　刘枏（1965-），博士，教授，主要研究方向为数据分析、工程管理信息化，E-mail：liunan@cqjtu.edu.cn；刘润（2001-），硕士研究生，主要研究方向为数据分析；李卓城（2004-），本科生，主要研究方向为数据挖掘。

引用格式　刘枏 , 刘润 , 李卓城 . 考虑数据维度与质量维度二重性的数据质量评估研究 [J]. 情报工程 , 2024, 10(4): 25-35.

doi:10.3772/j.issn.2095-915x.2024.04.003

each dimension of data and conduct value analysis. [Methods/Processes] The research first analyzes the literature related to data quality, and uses the modified Delphi method to construct a data quality evaluation index system. Then, the analytic hierarchy process is used to select indicators, and the Monte Carlo sampling method is used to calculate the Shapley value and obtain the dimensional evaluation results of data. In the case study of real estate data quality evaluation, the quality evaluation model is applied to evaluate 12 data dimensions. The results show that three dimensions have poor quality and are recommended for deletion. At the same time, suggestions for modification are proposed for other dimensions. [Limitations] This article does not address the difficulty of quantifying data quality dimensions, nor does it classify data, only proposing a general evaluation model. [Results/Conclusions] The research integrates data dimensions and data quality dimensions, and constructs a relatively complete data quality evaluation model based on the data quality evaluation index system and Shapley value. It can conduct more detailed quality evaluation of the database.

Keywords: Quality Evaluation Model; Data Dimension; Data Quality Dimension

引言

随着人工智能、机械学习、深度学习等技术的快速发展，各种数据共享、交换、重复利用活动被广泛应用[1]，数据已成为驱动经济发展的重要资源和新型生产要素[2]；在大数据时代的背景下，充分发掘潜在数据价值，提高数据利用效率的需求愈发迫切。然而，关联数据质量低、数据结构差、完整性不足等数据质量问题导致的开发成本较高影响着数据资源的可持续发展与利用[3]。

数据价值开发的基础主题之一是数据质量评估[4]，相关研究也得到了国内外相关学者的广泛关注。刘枏等[5]、张小伟等[6]从数据质量效用的角度出发，分别基于 Stackelberg 模型及 shapley 值构建了新的数据交易定价模型。Taleb 等[7-8]创建了一个涵盖质量纲要、质量需求、质量维度及其计分规则的大数据质量框架，并将其用于大数据全生命周期的质量评估工作。Woodall 等[9]研究了可以动态配置的质量评估模型，并提出了针对不同数据集特征将评价指标进行移除和关键性排序的评估方法。Martín 等[10]研究了物联网数据流质量的定义和评估机制以及数据管理解决方案，并将评估模块集成到操作数据富集工具链中，通过智能算法提高数据质量。Buelvasj 等[11]针对空气质量检测系统提出了一个多维模型，该模型集成了相关质量维度的研究和领域专家的意见，可以生成数据质量综合评分指数。Guerra-garcía 等[12]提出了在信息系统开发项目中数据质量的标准化管理方法，并将数据质量需求与软件数据质量需求关联。Mirzaie 等[13]将数据流质量控制领域的研究进行了总结，将相关的质量研究方法分为质量评估、质量控制方法、质量控制技术、质量应用四个维度，并对相关方法进行了批判性总结。

数据集由众多样本构成，每个样本又包含多个特征，这些特征各自代表着数据的一个维度。上述研究综合考虑了数据质量和数据容量对于数据价值的影响，但缺乏对多维数据下单一维度数据的数据质量对数据库价值影响的思考。在多元化应用背景下，数据各维度具有差

异性价值。考虑维度的数据质量评估工作能够识别各维度在业务中的重要程度，从而明确数据质量关注重点；评估结果也能为数据的存储和管理方式提供针对性的优化建议，进而提升数据利用效率。

1 研究概述

数据库不同维度数据之间的质量问题交互性强，单一维度的质量问题常常在多维度数据的组合中显现，因此必须选择合适的方法对单维度数据质量进行评估。本研究试图通过对过往文献中数据质量维度的概念及计算规则进行总结，构建一个较为齐备的数据质量维度指标体系。在此基础上，针对不同类型的数据及其特性，使用专家评议法从评估体系选择适宜的评估指标和评分方法，并运用层次分析法对质量维度指标进行权重划分。针对不适用于单一维度的评分方法，通过对数据各种维度之间的所有任意组合进行评分，利用 shapley 值计算出单维数据的评分，并将其作为质量评估的依据（研究框架见图 1）。研究的主要贡献在于提出了一个综合考虑数据维度和质量维度的质量评估体系，并借助 shapley 值的方式对单维数据的质量进行判断，进而帮助研究人员更好地评估数据质量及价值，为数据的有效应用提供支持。

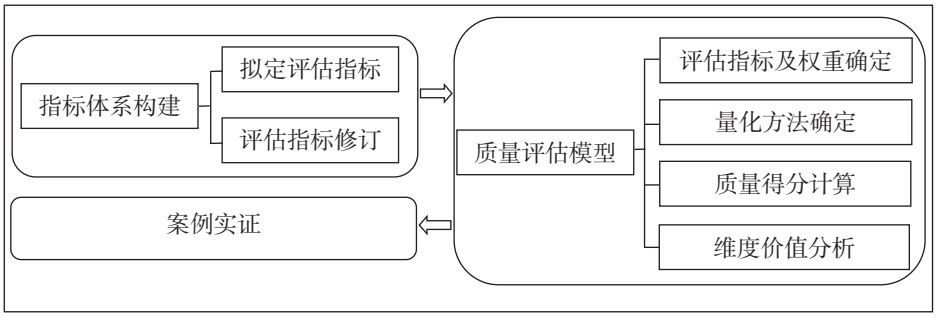

图 1 研究框架图

2 数据质量维度指标体系

数据质量是数据的固有特性，被定义为“在特定条件下满足使用者需求的能力”[8]。是衡量数据适合使用的指标。而数据质量维度用于衡量、量化和管理数据质量，每个质量维度代表数据质量的一个属性。为了方便管理，我们将质量维度指标分为四类，分别是固有质量、情景质量、表征质量、可达质量。其中，固有质量是指数据本身所固有的、与使用条件和使用环境无关的质量；情景质量用于衡量数据使用条件和使用环境的匹配程度；表征质量是衡量数据被有效表示的难易程度；可达质量用于衡量获取数据难易程度。本文选取 30 篇有关数据价值影响因素研究的标准与文献，并从中对数据质量维度指标进行提炼。使用修正德尔菲法征询专家意见，提出数据质量维度指标。

拟定数据质量维度指标并对其进行提取分析，构建数据质量维度指标体系（见表 1）。

doi:10.3772/j.issn.2095-915x.2024.04.003

表 1　数据质量维度指标体系

质量维度		指标说明	指标来源
固有质量	数据容量	数据集中单位数据的数量	[4]、[7]、[19]、[22]
	数据完整性	数据集中各单位数据的完整程度	[3]、[13]、[14]、[15]
	数据一致性	数据集中各单位数据无矛盾且与其他数据保持一致的程度	[13]、[14]、[15]、[17]
	数据准确性	数据集所记录内容的正确性和精确度	[10]、[21]、[22]、[25]
	数据规范性	数据集符合数据标准、业务规则或权威参考数据的程度	[12]、[18]、[24]、[25]
	数据真实性	数据集反映其蕴含属性和相关实例的真实性和可靠程度	[19]、[25]、[28]、[29]
	数据全面性	数据集反映其预期属性和相关实例的完备程度	[12]、[16]
情景质量	数据相关性	数据集与业务任务的关联匹配程度	[11]、[23]、[25]、[19]
	数据及时性	数据集实时更新及传输同步的效率	[19]、[22]、[23]、[26]
	数据时效性	数据集与当前状态的关联匹配程度	[14]、[15]、[17]、[21]
	重复使用率	数据集在设置、替换、传输等重复使用中保持现有质量的程度	[13]、[20]、[24]
	独特性	数据集在特定背景下的独特价值或独特性质	[11]、[16]、[29]
表征质量	可理解性	使用者读取和解释数据集所蕴含属性的难易程度	[10]、[23]、[26]、[29]
	可机读性	数据集被计算机程序自动读取、解析和处理的难易程度	[8]、[10]、[16]、[21]
	数据代表性	数据集对其预期属性和相关实例的代表程度	[11]、[16]、[24]、[25]
	易操作性	对数据集进行替换、移动、分析等操作并保持其现有质量的难易程度	[10]、[12]、[18]、[25]
可达质量	可获取性	数据集查找、检索和获取的难易程度	[11]、[12]、[13]、[16]
	非歧视性	针对他人访问、获取、使用和分享数据集的潜在约束程度	[4]、[20]
	数据安全性	保护数据集不被窃取、盗用或破坏的难易程度	[13]、[25]、[26]、[27]

其次，使用修正德尔菲法对评估指标进行修正。共发放了《“数据质量维度指标体系”专家咨询问卷》35 份，回收有效问卷 20 份，问卷的回收率为 57.1%，问卷的有效率为 100%。研究量表主体部分采用内部一致性较高的五级李克特量表方式进行填答，分数越高，代表被试者认为该项指标对数据质量影响越大。对回收的问卷进行数据处理和分析，计算每个指标专家意见的平均值、标准差和专家意见的 Kendall 协调系数。平均值用于衡量指标的取舍，标准差和 Kendall 协调系数用于衡量专家意见的一致性。平均值低于 1.5 分的，说明专家意见处于影响程度较小，可以将其从指标体系中剔除。数据处理并分析后发现：“数据代表性”和“非歧视性”平均分均为 1.38902，说明专家认为数据代表性和非歧视性对数据质量影响较小。专家对所有指标评价的标准差均小于 1，这意味着专家的意见基本趋于一致和稳定。Kendall 协调系数检验呈现出显著性（$P \approx 0.000 < 0.01$），意味着 20 位专家的评价具有关联性，即说明评价具有一致性。同时 Kendall 协调系数为 0.875，大于 0.5，进一步说明评价一致性较强，因此不再进行第二轮的专家调查。

3　质量评估模型

质量评估是根据确定的质量评估对象、质量范围、测量及其实现方法实现质量评测的活

动过程[30]。基于此，本文的质量评估模型由评估指标选定与权重确定、评分方法确定、质量指标得分计算、维度价值分析四部分组成。

本节所用的符号描述如表 2 所示：

表 2　符号描述

符号	描述
Z	需进行质量评估的数据集
V_n	数据质量指标
D_i	数据集 Z 的第 i 维度所组成的数据集
M_n	可直接计算的 V_n 评分方法
S	评估模型所需的由多个 D_i 所组成的数据集
SV_i^n	由 Shapley 值计算的指标得分
L_n	需借助 Shapley 值计算的评分方法
SC_i^n	需借助 Shapley 值计算的指标得分

3.1　数据预处理

鉴于半结构化数据及非结构化数据质量评估工作所固有的复杂性，在启动评估工作之前，应根据具体的数据集和任务需求对数据进行定制化的处理，以便于为后续的数据评估工作奠定坚实的基础。

数据预处理的两个重点是样本选取及维度提取。由于目前数据库容量大、格式多样，对数据整体进行评估工作可能存在困难，因此可使用随机抽样、等距抽样、分层抽样等抽样方法从数据源中选择少量样本Z来表示整个数据，但应保证样本与预评估数据库水平保持一致；针对数据维度特征不明显的数据，应根据数据应用场景及应用需求通过文本分析、图像分析、音频分析等技术结合数据特征对数据进行维度提取。质量评估模型需要基于维度将样本数据 Z 垂直拆分为多个 D_i，针对半结构化及非结构化数据，还需综合考虑数据特性和技术实现的复杂性设计合理的拆分策略与方法。

3.2　数据质量评估指标选定与权重确定

评估模型采用专家评议法确定 n 个数据质量指标 V_n（$1 \le n \le n'$）。在评标过程中，评估小组应预先收集确定数据特征、质量需求、技术方案及技术成本等评审内容，并对质量评估指标进行共同的分项分析、比较，然后进行评议，最终确定指标。

层次分析法（AHP）是用于指标权重确定的常用方法。质量评估模型中，在评估小组内部发放调查问卷，通过成员对评估指标两两间重要程度的判断确定评估指标权重，最后求取权重分配的平均值作为最终权重。

质量评估指标是数据质量评价的关键，它反映了数据的多方面属性。在确定评估指标时要将组织战略、数据需求及业务需求和数据质量指标关联起来，明确数据质量与数据需求之间的关系。要根据具体业务需求选择适当的数据质量维度评价指标并进行合适的权重分配。此外，在数据全生命周期中，同一质量维度在不同时间维度上的含义和内容可能有所不同，需要根据评价阶段和使用范畴来确定需要评定的质量评估指标。

3.3　指标评分方法确定

在确定指标评分方法时，数据质量指标量化方法与数据业务需求的关系亟需进一步明确。指标评分方法应具有一定的前瞻性，即结合已经确定或预估的业务发展方向，将未来的数据预期融入测量框架中。对于备选的评分方法，还应考虑成本及可用性，主要包括是否能

doi:10.3772/j.issn.2095-915x.2024.04.003

够量化、技术上能否实现、初步的资源消耗评估等。

3.4 基于shapley值的质量指标得分计算

3.4.1 shapley值的定义

shapley 值法是由 shapley 提出的为解决多人合作对策及合作受益分配的一种数学方法[31]。shapley 值的优势是按照成员对最终结果的边际贡献率将收益进行分配，即成员 i 所分得的收益等于其所参加合作所创造的边际利益的平均值。将 shapley 值法应用于质量评估模型之中可描述为：对于需进行质量评估的 i 维数据库 Z 的第 i 维数据单独组成的数据库 D_i，若 V_n 的评分方法 L_n 不能计算只包含单一维度的 D_i，则令 S 代表数据库 Z 除 D_i 外其他维度数据集组合的所有可能的任意子集（$S \subseteq \{\{D_{1,\dots,}D_{i'}\} \setminus D_i\}$），再通过计算所有 S、$S \cup \{D_i\}$ 的 V_n 指标得分 $L_n(S)$、$L_n(S \cup \{D_i\})$ 及其两者差值 $L_n(S \cup \{D_i\}) - L_n(S)$，进而得到 D_i 的边际贡献即 D_i 加入时各 S 指标得分的增量，并经过加权平均最终得到 D_i 的 V_n 指标得分 SV_i^n，如式（1）所示。

$$SV_i^n = \frac{1}{i'} \sum_{s \subseteq (\{D_{1,\dots,}D_{i'}\} \setminus D_i} \frac{L_n(S \cup \{D_i\}) - L_n(S)}{\binom{i'-1}{|s|}} \quad (1)$$

3.4.2 质量指标得分计算

对于需进行质量评估的 i 维数据库 Z 及其第 i 维数据单独组成的数据库 D_i（$1 \le n \le n'$），若 V_n 的评分方法可直接对 D_i 进行单独评估，则以 M_n 代指该评分方法。即 M_n 可直接对 D_i 进行评估并得到 D_i 的 V_n 指标得分 SC_i^n。若 V_n 的评分方法不能计算只包含单一维度的 D_i，则以 L_n 代指该评分方法。即通过计算 D_i 对所有 S 的边际贡献（$L_n(S \cup \{D_i\}) - L_n(S)$），再通过加权平均得到 D_i 的 V_n 指标得分 SV_i^n。

如式（1）所示，若要直接计算所有 D_i 的 shapley 值，则需计算出所有可能的边际贡献，这随着数据的增长呈指数级增长。因此我们采用蒙特卡洛采样[32]的方法来计算近似的 shapley 值。通过对不同维度的数据集的随机排列进行采样，然后扫描所有排列并计算所有数据集的边缘得分贡献，如图 2 所示。

算法框架：质量得分计算

输入：数据库Z、$D_i(1 \le i \le i')$，评分方法M_n、$L_n(1 \le n \le n')$
输出：各维度指标得分SC_i^n、SV_i^n

```
if M_n(D_i) is ValueError then
    Initialize SV_i^n = 0、 t=0;
    While convergence criteria not met do
        Let π^k be a random permutation of {1,...,i'};
        t+=t;
        for j=1 to i' do
            SV^n_{π^k(i)} = L_n(D_{π^k(1)},...,D_{π^k(i)}) − L_n(D_{π^k(1)},...,D_{π^k(i-1)})
            SV^n_{π^k(i)} = ((t−1)/t) SV^n_{π^{t-1}(i)} + (1/t) SV^n_{π^t(i)}
        end for
    end for
else SC_i^n = M_n(D_i)
end if
return SV_i^n、SC_i^n
```

图 2 算法框架

3.5 维度价值分析

根据质量得分可以衡量各 D_i 的质量水平并支持决策的制定和调整，对于低质量的 D_i，若该维度存在使用必要性，则对 D_i 进行清洗和修正使其满足使用要求；若 D_i 为非必要维度或数据处理成本较高，则可以考虑舍弃该维度数据。同时，在数据生命周期较长的情况下，则质量得分可以作为依据辅助数据业务人员判断哪些

D_i 及质量维度需要重点关注，以便于更好地维护数据库，满足相关业务需求。

4 案例实证

本文选取二手房数据集验证本文所构建的数据资产维度化评估体系的适用性。案例背景为某房产中介平台拟开展企业数据库构建工作，现已初步收集了部分数据，但发现存在维度过多、部分维度质量较差等问题。为方便后续的数据集收集整理及数据维度管理，需要对已收集的部分数据进行维度化质量评估并给出质量管理意见。

4.1 质量评估需求及数据预处理

拟建的房产交易数据库旨在支持企业发展决策、房产定价、房产信息优化、数据交易等活动。但线上收集的数据在某些维度上可能存在较为严重质量问题。同时，企业也需要通过评估了解不同维度的关键程度以及它们的质量问题，以便于优化成本并更好地完成数据库构建工作。

案例数据为结构化数据，包含户型、建筑面积、朝向、楼层、装修、建筑年代、电梯、产权性质、住宅类别、建筑结构、建筑类别、学校、单价、成交周期、浏览量等十六个维度，无需进行维度提取及标准化处理工作。对于户型、装修档次、朝向等非数值维度，采用赋值法来实现指标的数值化。维度的量化情况见表3。目前已收集 8942 条数据，由于数据为结构化数据且各个体可视为等概率分布，为方便质量评估工作，采用简单随机抽样，随机抽取 300 条数据作为分析样本。

表 3 数据维度表

序号	指标	量化方法
1	房屋户型	一室一厅记为 1，两室一厅记为 2，两室两厅记为 3
		三室一厅记为 4，三室两厅记为 5，四室两厅记为 6
2	建筑面积	直接使用该套房产的建筑面积（单位：平方米）
3	朝向	房屋朝向为南、西南、东南均记为 1，其他朝向记为 0
4	楼层	低楼层、中楼层、高楼层分别记为 1、2、3
5	装修档次	毛坯、简装、中档、高档、豪华装修分别记为 1、2、3、4、5
6	建筑年代	直接使用该套房产所处楼栋的建成年代（单位：年）
7	电梯	有电梯记为 1，无电梯记为 2
8	产权性质	个人产权记为 1，商品房记为 2
9	住宅类别	普通住宅记为 1，非普通住宅记为 2
10	学校	非学区房记为 1，学区房记为 2
11	单价	房屋售出时每平方米单价（单位：元）
12	总价	房屋售出时交易总价（单位：元）

4.2 评估指标及权重选定

经数据质量评估专家与数据业务相关人员讨论与投票，在充分考虑数据库潜在应用价值的同时，结合房地产行业相关特点，围绕数据处理工作中的实际情况，根据现有评价能力和评价手段，从数据质量维度指标体系中选取并确定数据完整性、数据一致性、数据准确性、数据独特性四个数据质量评价指标。向专家发放《“二手房数据质量评估指标体系”AHP 专家咨询问卷》10 份，以确定一般情况下数据质量评估指标权重。根据专家反馈意见分别将专家对于评估指标权重的分配进行计算并进行一致性检验，最后求取权重分配的平均值作为数据质量评估指标的权重参考。

doi:10.3772/j.issn.2095-915x.2024.04.003

表 4　维度指标表

类别	质量维度	指标权重
固有质量	数据完整性	0.306860859
固有质量	数据一致性	0.251900705
固有质量	数据准确性	0.300769442
情景质量	数据独特性	0.140468993

其中，数据完整性和数据独特性可通过直接对数据库各维度进行单独评估得到不同维度的指标得分，数据一致性及数据准确性的评分方法只针对数据集而不能只计算单一维度，因此需要通过算法对各维度之间任意组合的所有数据集进行评估，并利用 shapley 值得到各维度的指标得分。各指标度量方法如下。

数据完整性：完整性应衡量各维数据被完整赋值没有空缺的程度。

N= 单维数据被赋值元素个数 / 单维数据中元素个数（得分范围：0~1 分）

数据独特性：数据集在特定背景下的独特价值或独特性质。

M= 由评估专家对不同维度进行主观评分（得分范围：0~1 分）

数据一致性：一致性应衡量数据集根据相关规则要求保持一致的程度。

U= 各数据集 *Kendall W* 协调系数得分（得分范围：0~1 分）

数据准确性：准确性应衡量数据集真实、准确客观地反映现实情况的程度。

A= 数据集中异常值个数 / 数据集中元素个数（得分范围：0~1 分）

4.3　房地产数据质量评估

4.3.1　完整性

将全部数据导入 spss 软件，通过描述统计得到各维度缺失值及得分情况如下（见表 5）：

表 5　完整性得分表

	统　计						
个案数	数据维度	户型	建筑面积	朝向	楼层	装修	建筑年代
	有效	8468	8942	8933	8935	6946	6861
	缺失	474	0	9	7	1996	2081
	N	0.947	1	0.99899	0.999	0.777	0.767278
	数据维度	电梯	产权性质	住宅类别	学校	总价	单价
	有效	7083	7041	6087	8942	8942	8942
	缺失	1859	1901	2855	0	0	0
	N	0.7921	0.787407739	0.68072	1	1	1

由统计表可知，户型、建筑面积、朝向、楼层、学校、总价、单价的缺失值较少，拥有较高的完整性。而装修、建筑年代、电梯、产权性质、住宅类别的缺失值较多，其中住宅类别的完整性最差。

4.3.2　数据独特性

经评估人员与数据业务人员讨论，针对房地产数据集各维度独特性难以量化的特点，选取数据特征、数据类型、数据规模作为数据独特性的评估因素，采取专家打分法对各维度指

标得分进行量化，并求取各专家打分情况的平均值得到各维度得分情况如下（见表 6）：

表 6　独特性得分表

统　计						
数据维度	户型	建筑面积	朝向	楼层	装修	建筑年代
M	0.745	0.855	0.678	0.642	0.784	0.798
数据维度	电梯	产权性质	住宅类别	学校	总价	单价
M	0.722	0.678	0.633	0.542	0.874	0.912

4.3.3　数据一致性

本案例中数据一致性较侧重于各维度数据之间的关联性，因此选用 *Kendall W* 度量数据集一致性。*Kendall W* 系数范围为 [0,1]，当 $W \leq 0.4$ 时，说明数据一致性较差；当 $0.4 < W \leq 0.6$ 时，说明数据一致性一般；当 $0.6 < W \leq 0.8$ 时，说明数据一致性较好；当 $0.8 < W \leq 1$ 时，说明数据一致性为优。选用 matlab 软件计算 *Kendall W* 系数，将样本数据集以矩阵形式导入 matlab 软件中，并将 *Kendall W* 系数算法嵌套入质量评估模型中得到各维度得分情况如下（见表 7）：

表 7　一致性得分表

统　计						
数据维度	户型	建筑面积	朝向	楼层	装修	建筑年代
U	0.8208	0.933636364	0.65482	0.727	0.758	0.996909
数据维度	电梯	产权性质	住宅类别	学校	总价	单价
U	0.7205	0.741545455	0.73827	0.762	0.947	0.9645

4.3.4　数据准确性

本案例采用 *K* 最近邻算法识别数据集中的异常值，该模型是通过搜寻最近的 *K* 个已知类别样本对未知类别样本进行预判，基于此，我们只需依次计算每个样本点与它最近的 *K* 个样本的平均距离。再利用计算的距离与阈值进行比较，如果大于阈值，则认为是异常点。

将样本数据集以矩阵形式导入 matlab 软件中，并将 *K* 最近邻算法嵌套入质量评估模型中得到各维度得分情况如下（见表 8）：

表 8　准确性得分表

统　计						
数据维度	户型	建筑面积	朝向	楼层	装修	建筑年代
A	0.8924	0.9765	0.8454	0.7741	0.8324	0.8042
数据维度	电梯	产权性质	住宅类别	学校	总价	单价
A	0.7132	0.5585	0.6547	0.8433	0.9874	0.9872

4.4　评估结果分析

从数据维度层面的得分结果看，总得分在三分以下的维度是住宅类别、产权性质、电梯，分别为 2.706693、2.765453、2.947759；从质量维度层面对住宅类别、产权性质、电梯三个维度进行分析，它们的数据完整性与数据准确性存在严重的质量问题，建议业务人员在考虑数据获取与处理成本的同时，重点关注这三个维度的数据质量以提高数据库整体质量水平；同时其独特性和一致性评分也均处于较低标准，建议相关业务人员优化这三个维度数据的结构

doi:10.3772/j.issn.2095-915x.2024.04.003

及处理流程。

从质量维度的得分结果来看，数据库整体的一致性、独特性得分均处于较好水平，说明数据库整体潜在应用价值及与业务需求匹配程度均较好。针对数据完整性而言，除住宅类别、产权性质、电梯外，仅楼层维度评分为 0.7741，其他维度评分均在 0.8 以上，建议业务人员考察相关数据来源及收集流程，重点提高楼层的数据准确性。针对数据完整性而言，除住宅类别、产权性质、电梯外，装修和建筑年代的数据完整性评分也较低，同时装修和建筑年代的一致性、独特性、准确性质量维度评分较高，说明其数据准确性及潜在价值较好，因此在后续的数据库构建及维护工作中应重点关注装修和建筑年代的数据缺失问题。

5 结论

本文通过使用文献调查法和修正德尔菲法建立了兼具科学性、系统性的数据质量维度指标体系；并以此为基础，利用层次分析法、shapley 值法及蒙特卡洛采样构建了一个较为完备的数据质量评估模型；并以某二手房数据集质量评估案例作为实证应用样本，进一步探讨了所构建质量评估模型的有效性和可操作性。本文的创新体现在：将数据维度和数据质量维度进行融合，尤其是从维度层面考虑了数据的质量问题，以此来保证数据的可用性，在一定程度上弥补了当前数据质量研究的不足；针对部分数据质量维度评分方法不适用于单一维度评分的情况，通过对数据各种维度之间的所有任意组合进行评分，再利用蒙特卡洛采样计算 shapley 值得出单维数据的质量评分。

虽然本文为单维数据质量评估提供了新思路，但也还存在一些不足，例如，没有解决上文所提及的部分数据质量维度指标难以量化的问题，同时实证案例中选取的是结构化数据，针对非结构化数据的维度提取及基于维度的数据库拆分等预处理工作未进行详细说明。另外本文没有对数据进行分类，所以提出的是一种笼统的数据质量评估模型，不能为不同类型的数据结构提出定制化的评估方案等，这都是下一步研究需要解决的问题。

参 考 文 献

[1] JIAN P. A Survey on Data Pricing: From Economics to Data Science[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(10): 4586-4608.

[2] 中华人民共和国国务院 . 中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要 [EB/OL]. (2021-03-13) [2024-01-23]. https://www.gov.cn/xinwen/2021-03/13/content_5592681.htm

[3] 孙嘉睿 , 安小米 . 开放政府数据质量评估指标体系研究 [J]. 情报理论与实践 , 2023, 46(6): 94-100, 78.

[4] CABALLERO I, GUALO F, RODRíGUEZ M, et al. BR4DQ: A methodology for grouping business rules for data quality evaluation[J]. Information Systems, 2022, 109: 102058.

[5] 刘枏 , 徐程程 , 陈俞宏 . 基于效用理论的数据定价方法研究 [J]. 价格理论与实践 , 2022(11): 164-167, 211.

[6] 张小伟 , 江东 , 袁野 , 等 . MaSS: 基于单位数据贡献的模型定价框架 [J]. 计算机科学与探索 , 2023, 17(9): 2252-2264.

[7] TALEB I, SERHANI M A, BOUHADDIOUI C, et al. Big data quality framework: a holistic approach to continuous quality management [J]. Journal of Big Data, 2021, 8(1): 41.

[8] TALEB I, SERHANI M A, DSSOULI R, et al. Big Data Quality: A Survey [C]// proceedings of the IEEE International Congress on Big Data (IEEE BigData) Part of the IEEE World Congress on Services, San

Francisco, CA, F Jul 02-07, 2018. Ieee: NEW YORK, 2018: 166-173.

[9] WOODALL P, BOREK A, PARLIKAD A K. Data quality assessment: The Hybrid Approach [J]. Inf Manage, 2013, 50(7): 369-382.

[10] MARTÍN L, SáNCHEZ L, LANZA J, et al. Development and evaluation of Artificial Intelligence techniques for IoT data quality assessment and curation [J]. Internet of Things, 2023, 22: 100779.

[11] BUELVAS J H, MÚNERA D, GAVIRIA N. DQ-MAN: A tool for multi-dimensional data quality analysis in IoT-based air quality monitoring systems [J]. Internet of Things, 2023, 22: 100769.

[12] GUERRA-GARCÍA C, NIKIFOROVA A, JIMÉNEZ S, et al. ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design [J]. Data & Knowledge Engineering, 2023, 145: 102152.

[13] MIRZAIE M, BEHKAMAL B, ALLAHBAKHSH M, et al. State of the art on quality control for data streams: A systematic literature review [J]. Computer Science Review, 2023, 48:100554.

[14] VALENCIA-PARRA Á, PARODY L, VARELA-VACA Á J, et al. DMN4DQ: When data quality meets DMN [J]. Decision Support Systems, 2021, 141: 113450.

[15] 全国信息技术标准化技术委员会 . 信息技术数据质量评价指标 : GB/T 36344-2018[S]. 北京 : 中国标准出版社 , 2018: 13.

[16] VALVERDE C, MAROTTA A, PANACH J I, et al. Towards a model and methodology for evaluating data quality in software engineering experiments [J]. Information and Software Technology, 2022, 151: 107029.

[17] GONZÁLEZ-VIDAL A, RAMALLO-GONZÁLEZ A P, SKARMETA A F. Intrinsic and extrinsic quality of data for open data repositories [J]. ICT Express, 2022, 8(3): 328-333.

[18] WOOK M, HASBULLAH N A, ZAINUDIN N M, et al. Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling [J]. Journal of Big Data, 2021, 8(1): 49.

[19] LIU C, NITSCHKE P, WILLIAMS S P, et al. Data quality and the Internet of Things [J]. Computing, 2019, 102(2): 573-599.

[20] LIONO J, JAYARAMAN P P, QIN A K, et al. QDaS: Quality driven data summarisation for effective storage management in Internet of Things [J]. Journal of Parallel and Distributed Computing, 2019, 127: 196-208.

[21] 张夏子钰 , 周林兴 . 大数据时代档案数据质量 : 评估与优化 [J]. 北京档案 , 2023(5): 15-18.

[22] 吴小娥 . 物联网数据质量评估应用 [J]. 宝鸡文理学院学报 (自然科学版), 2022, 42(4): 50-54, 62.

[23] 赵文轩 , 李春旺 . 关联数据质量评价方法研究述评 [J]. 情报理论与实践 , 2016, 39(2): 134-138, 128.

[24] LIU J, LI J, LI W, et al. Rethinking Big Data: A Review on the Data Quality and Usage Issues [J]. ISPRS journal of photogrammetry and remote sensing, 2016, 115: 134-142.

[25] HARYADI A F. Requirements on and Antecedents of Big Data Quality: An Empirical Examination to Improve Big Data Quality in Financial Service Organizations [D]. Delft: Delft University of Technology, 2016

[26] 莫祖英 . 大数据质量测度模型构建 [J]. 情报理论与实践 , 2018, 41(3): 11-15.

[27] SHEN Y, GUO B, SHEN Y, et al. A Pricing Model for Big Personal Data [J]. Tsinghua science and technology, 2016, 21(5): 482-490.

[28] WANG R, STRONG D. Beyond Accuracy: What Data Quality Means to Data Consumers [J]. Journal of management information systems, 2016, 12(4): 5-33.

[29] LUGMAYR A, STOCKLEBEN B, SCHEIB C. A Comprehensive Survey on Big-Data Research and Its Implications–What Is Really 'New'in Big Data?-It's Cognitive Big Data! [C]//Proceedings of the 10th Pacific Australesian Conference on Information Systems Taiwan, 2016: 45.

[30] 李华锋 , 辜汝桐 , 胡海青 , 等 . 基于熵权法的航行通告数据质量评价方法 [J]. 民航学报 , 2022, 6(4): 1-5.

[31] 吴黎军 , 项海燕 . 基于信息熵的 n 人合作博弈效益分配模型 [J]. 数学建模及其应用 , 2013, 2(Z2): 50-54.

[32] CASTRO J, GÓMEZ D, TEJADA J. Polynomial calculation of the Shapley value based on sampling[J]. Computers & Operations Research, 2009, 36(5): 1726-1730.