

数字经济背景下数据质量评估模型研究

代 婕

贵州大学省部共建公共大数据国家重点实验室, 贵州 贵阳

收稿日期: 2025年4月30日; 录用日期: 2025年5月23日; 发布日期: 2025年6月25日

摘 要

在数字经济蓬勃发展的背景下, 数据不仅成为新型生产要素, 也在产业升级、智能决策和公共治理中发挥着重要作用。数据质量作为影响数据价值释放的关键因素, 正受到广泛关注。本文基于层次分析法与熵权法, 构建了涵盖准确性、完整性、冗余性和时效性的多维度数据质量评估模型, 实现了数据质量的量化分析。研究表明, 准确性和时效性是提升数据整体价值的关键因素。基于此, 本文提出应在数据管理过程中优先关注关键维度的质量保障, 从而提升数据利用的有效性与效率。

关键词

数字经济, 数据质量, 层次分析法, 熵权法

Research on Data Quality Assessment Model in the Context of Digital Economy

Jie Dai

State Key Laboratory of Public Big Data, Guizhou University, Guizhou Guiyang

Received: Apr. 30th, 2025; accepted: May 23rd, 2025; published: Jun. 25th, 2025

Abstract

Against the backdrop of a booming digital economy, data has not only become a new type of production factor, but also plays an important role in industrial upgrading, intelligent decision-making and public governance. Data quality, as a key factor affecting the release of data value, is receiving widespread attention. Based on hierarchical analysis and entropy weight method, this paper constructs a multi-dimensional data quality assessment model covering accuracy, completeness, redundancy and timeliness, and realizes the quantitative analysis of data quality. The research results show that accuracy and timeliness are the key factors to enhance the overall value of data. Based on this, this paper proposes that priority attention should be given to the quality assurance of key dimensions

in the data management process, so as to enhance the effectiveness and efficiency of data utilization.

Keywords

Digital Economy, Data Quality, Hierarchical Analysis, Entropy Weight Method

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

如今,数据已成为与土地、劳动力、资本、技术等并列的关键生产要素,被誉为“21 世纪的石油”,在各行各业的发展变革中发挥着引领作用。然而,在数字化转型的大背景下,尽管数据流通领域受到了空前的关注,各方也在积极探索数据的开发与流通,以实现充分释放数据要素潜力的目标,但我国数据要素市场尚未充分激活,其中一个关键障碍在于当前缺乏统一、科学的数据质量评价体系,导致数据在流通过程中“质”的问题难以量化,进而影响其应用价值的有效释放。数据质量是衡量数据有效性和可用性的核心标准。高质量的数据能够支撑更精准、更复杂的分析决策过程,助力人工智能、智能制造、精准营销等关键应用场景落地。因此,构建合理、可操作性强的数据质量评估方法,已成为当前数据要素市场亟需突破的难题。在此背景下,数据质量评估的重要性愈发凸显,成为推动数字经济高质量发展的关键一环。

2. 文献综述

随着数字经济的迅速发展,数据质量评估已成为数据管理与应用中的关键环节,其核心在于通过科学的方法或模型,对数据的优劣进行系统判断。早期研究多将数据质量等同于“准确性”,但随着数据应用场景的不断丰富,单一的维度已难以全面反映数据的价值与使用效果。因此,学界逐渐从多元视角出发,拓展了数据质量的内涵与评估方法。

从评估方法的分类来看,目前数据质量评估主要包括定性评价、定量评价和综合评价三种类型。定性方法主要从数据使用者的主观视角出发,通过观察、分析与判断构建或优化评价指标体系[1],常用方法包括专家评价法、用户反馈法及第三方评测法[2]。定量方法则以客观数据为基础,利用统计分析 with 指标评分等手段对数据质量进行量化描述,适用于结构化数据的质量评估[3]。综合评价方法融合了定性与定量的优势,在保留主观判断灵活性的同时,提升了评估结果的科学性与稳定性,已成为当前研究中应用最为广泛的方法之一,常见方法包括模糊综合评价法、层次分析法、德尔菲法与扎根理论等[4]。

在具体应用中,研究者逐步由单一评价指标向多维度建模转变,并引入多种权重分配方法,以增强评估体系的系统性与适用性。郭亚军[5]结合层次分析法与扎根理论,构建了适用于 AIGC 场景的数据质量评估模型;宋俊典[6]则基于层次分析法方法设计了包含权重集合、隶属函数与特征值集合的多维综合评价框架,实现了指标的系统判定。此外,针对电力行业的数据质量评估,王杨琛[7]采用熵权法计算各维度指标的客观权重,并引入灰色评估法确定数据质量等级;杨栋枢[8]则将层次分析法与熵权法相结合,构建了面向电力企业的数据质量综合评价模型;计蓉[9]通过模糊层次分析法与熵权法联合确定权重,再借助模糊综合评价法依据隶属函数评定数据质量等级。相关研究表明,层次分析法因其结构严谨、层次清晰,在数据质量权重确定中应用最为广泛;而熵权法则多用于电力等特定行业,尚未在更广泛的数据

应用场景中普遍推广。

值得注意的是，当前将层次分析法与熵权法进行组合的研究仍较为有限。作为一种融合主观判断与客观数据的组合赋权方法，该方法能够整合不同权重模型的计算结果，有效降低单一方法带来的偏差与不确定性。在实际应用中，层次分析法主要依赖专家打分，易受主观因素影响，导致评价结果波动较大。而熵权法则通过指标在数据集中所包含的信息量及其分布特征，客观地确定各指标的权重。为克服传统方法的局限，提升评价结果的科学性与稳健性，本文拟综合运用层次分析法与熵权法，对数据质量评价指标体系进行组合赋权。

3. 研究设计与方法

3.1. 评价方法

在构建数据质量评价体系的过程中，科学合理地确定各评价指标的权重，是确保评价结果客观性与有效性的重要前提。为此，采用层次分析法(Alytic Hierarchy Process, 简称 AHP)与熵权法(Entropy Weight Method, 简称 EWM)相结合的方式评价，以实现主观判断与客观数据的有机融合，从而提高评价体系的全面性与可靠性。

层次分析法是一种结合定性定量分析的多准则决策方法，由美国运筹学家托马斯·塞蒂(T. L. Saaty)提出。该方法将复杂问题分解为多个层次结构，先进行定性分析以明确各因素间的逻辑关系，再利用特征向量的数学计算方法，获取各因素在层次结构中的相对权重。具体而言，AHP 通过构建层次结构模型，将目标分解为多个子目标，并通过矩阵特征向量计算各层次的权重，最终通过加权求和递归计算出目标权重。

熵权法则是一种通过客观赋权进行综合评价的方法，能够根据评价指标在样本数据中的分布特征来调整其权重。该方法以信息熵为基础，衡量各指标值的离散程度。当某评价指标的熵值偏低，意味着该指标在样本之间存在较大差异，信息贡献较高，因此在评价中应分配更大的权重；反之，若某指标的熵值较高，则说明其分布较为均匀，所含信息有限，对最终结果的影响较小，应给予较低的权重。熵权值常被称为修正系数，可用于修正指标的初始权重，从而增强权重设置的客观性。

通过 AHP 法引入专家经验与结构性判断，结合 EWM 对样本数据的分布特征进行客观分析，构建了一套更具理论依据与实际适应性的权重确定方法，为后续数据质量的综合评价提供坚实基础。

3.2. 数据质量评价指标体系的构建

目前，数据质量评估维度相关研究成果颇为丰硕。众多学者基于各自的研究视角，针对不同的数据质量维度提出了多样化的评估指标框架，如表 1 所示。

Table 1. Data quality evaluation indicators
表 1. 数据质量评价指标

研究者	具体评价指标	年份
Xu T [10]	完整性、一致性、唯一性、准确性、相关性、及时性、可操作性、可理解性、可访问性、成本、盈利能力、风险性	2024
Wang J [11]	完整性、准确性、及时性、一致性、相关性	2024
计蓉[9]	完整性、准确性、唯一性、一致性、及时性	2024
王梦婷[12]	完整性、时效性、准确性、规范性、一致性	2024
Gong Y [13]	完整性、自我一致性、及时性、保密性、准确性、标准化、无偏性、易用性	2023

续表

王景鸿[14]	准确性、一致性、冗余性、完整性、时效性	2023
黄倩倩[15]	规范性、完整性、准确性、一致性、时效性、可访问性	2022
Taleb I [16]	完整性、一致性、准确性、及时性	2021
Zhang H [17]	准确性、可及性、完整性、及时性、一致性、可理解性	2020
Yang J [18]	准确性、完整性、冗余性	2019
Cichy C [19]	完整性、及时性、准确性、一致性、可访问性	2019
El Alaoui I [20]	可靠性、可用性、相关性	2019
王明政[21]	完整性、一致性、精确性、规范性、时效性、丰富性、可恢复性、可交换性	2019
李晶晶[22]	依从性、保密性、准确性、完整性、一致性、时效性、易访问性	2018
张晓冉[23]	正确性、完整性、一致性	2018
Stahl F [24]	准确性、数据量、可用性、完整性、延迟、响应时间、及时性	2016

通过对过往研究梳理发现，不同学者在数据质量评估中采用的维度标准存在较大差异，这种差异性导致评估结果在方法和结论上呈现多样化趋势。尽管维度设置越丰富，理论上越有助于更全面地反映数据质量，但维度过多也可能引发指标交叉、评估重复等问题。比如，时效性和及时性在定义上高度重合，准确性、规范性、精确度等指标的评估目标是判断数据是否真实可靠。

因此，本研究首先对现有研究中提出的数据质量维度进行了系统梳理与归类，在统计分析基础上发现，完整性、准确性、时效性是出现频率较高的维度，而冗余性虽不常作为独立指标，但在部分研究中以数据的唯一性或规范性体现其评估功能，具备独立使用的基础。

基于上述分析，本文选取完整性、准确性、冗余度和时效性四个维度构建数据质量评估体系，如表 2 所示。这些指标在前人研究中具有较强的代表性，内涵区分明确，能够减少冗余，提高评价模型的简洁性与逻辑清晰度。此外，这四个维度分别从数据的结构完整性、内容正确性、重复程度和时间相关性等不同层面展开，关注点相互独立，能够避免指标间的交叉与重复，为后续的量化分析提供良好基础。

Table 2. Data quality metrics and their definitions

表 2. 数据质量指标及其定义

评价指标	指标解释及说明
完整性 Q_1	数据缺失或空白的程度
准确性 Q_2	数据与真实数据的接近程度
冗余性 Q_3	相同字段重复出现的程度
时效性 Q_4	及时、有效地获取数据的水平

1) 数据完整性计算

数据完整性反映了数据集或数据库中信息项的完备程度，是衡量数据质量的重要指标之一。保障数据的完整性有助于避免因信息缺失而引发的误操作或错误输出，从而提高数据应用的有效性。其计算公式如式(1)所示。

$$Q_1 = \left(1 - \frac{ic}{tnc}\right) \times 100 \tag{1}$$

其中, ic 表示缺失数据的单元格数量, tnc 为单元格总数。值越接近 1, 表示完整性越高。

2) 数据准确性计算

准确性是数据质量的核心组成部分, 用于衡量样本信息在多大程度上能够真实反映客观现象。准确性的计算如式(2)所示。

$$Q_2 = \left(1 - \frac{ec}{tnc}\right) \times 100 \quad (2)$$

其中, ec 代表存在错误记录的单元格数量。该指标越高, 说明数据的真实程度越强, 可信度越高。

3) 数据冗余性计算

冗余性反映的是数据库中同类信息出现重复的程度, 特别是字段或数据行的重复。在实际应用中, 冗余过多会引发数据异常、冲突或不一致等问题, 从而降低整体质量。其计算公式如式(3)所示。

$$Q_3 = \left(1 - \frac{dr}{tnr}\right) \times 100 \quad (3)$$

其中, dr 为重复记录的行数, tnr 是数据的总行数。该值越大, 冗余程度越低, 数据质量越优。

4) 数据时效性计算

数据的时效性评估反映数据产品的产生或提供是否及时, 高时效性的数据能更好地支持决策分析, 尤其在时间敏感场景中尤为关键。滞后产生或延迟获取的数据可能导致结果失真, 降低数据价值。以交付时间 dt 为参考, 若数据实际采集时间为 it , 其计算公式如式(4)所示。

$$Q_4 = \left[1 - \frac{dt - it}{dt}\right] \times 100 \quad (4)$$

其中, Q_4 的得分越高, 说明数据生成或获取越及时, 满足时效性需求的程度越高。

4. 数据质量评价指标权重计算

上一节已针对数据质量评估维度开展量化剖析。但在实际应用中, 各个指标的重要程度存在差异, 若在评估时对各维度一概而论地赋予相等权重, 极有可能降低数据质量评估结果的科学性与准确性。就准确性来说, 其作为数据质量的关键属性, 体现了数据与客观实际的契合度, 这几乎是所有数据使用者对数据的基础要求, 理应在权重分配中占据较大比例; 而时效性则在部分特定场景中才具有显著价值, 其相对重要性较低。基于此, 有必要对各指标的重要程度进行合理赋权, 以提高评估模型的可靠性与适用性。本节将进一步阐述 AHP 和 EWM 这两种方法的具体应用过程, 并介绍主客观权重的融合方法, 以实现数据质量综合权重的科学计算。

4.1. 层次分析法

层次分析法通过将整体问题按逻辑关系分解为多个层次结构, 从而简化分析过程、增强问题的透明度与可理解性。其具体实施流程如下:

1) 构建层次结构图, 见图 1

2) 建立判断矩阵

判断矩阵的元素通过两两比较各维度指标的重要性来赋值。通常采用九级标度法, 依据专家意见对指标相对重要性进行量化。具体标度方法如表 3 所示。

为了确定各数据质量维度的相对权重, 本文设计了《专家意见征询打分表》, 向 12 位专家发放问卷, 其中包括 7 位高校教授和 5 位企业管理者。问卷采用 1~9 标度的成对比较方式, 邀请专家根据前述四个

数据质量维度之间的重要性关系进行打分。根据专家评分结果，构建了数据质量维度重要性比较的判断矩阵，具体数值如表 4 所示。

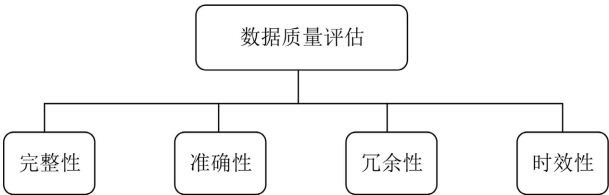


Figure 1. Data quality evaluation index system
图 1. 数据质量评价指标体系

Table 3. Judgment matrix of data quality dimensions
表 3. 数据质量维度判断矩阵

标度	含义
1	表示两个指标相比较，具有同等重要性
3	表示两个指标相比较，前者相较于后者略为重要
5	表示两个指标相比较，前者比后者显著重要
7	表示两个指标相比较，前者比后者极为重要
9	表示两个指标相比较，前者比后者绝对重要
2、4、6、8	表示上述相邻判断之间的中间取值

Table 4. Judgment matrix of data quality dimensions
表 4. 数据质量维度判断矩阵

指标	完整性	准确性	冗余性	时效性
完整性	1	1/2	4	1/3
准确性	2	1	7	2
冗余性	1/4	1/7	1	1/8
时效性	3	1/2	8	1

根据上表的数值可以构造矩阵 A。

$$A = \begin{bmatrix} 1 & 1/2 & 4 & 1/3 \\ 2 & 1 & 7 & 2 \\ 1/4 & 1/7 & 1 & 1/8 \\ 3 & 1/2 & 8 & 1 \end{bmatrix} \tag{5}$$

3) 一致性检验

一致性检验用于评估多准则评判中的逻辑一致性，通过量化专家在比较不同要素权重时的协调性，确保评估结果的内在一致性，并能有效检测打分过程中的逻辑矛盾。通过 SPSS 软件计算出判断矩阵的最大特征值为 $\lambda_{\max} = 4.015$ ，再根据式(6)计算一致性指标 CI。

$$CI = \frac{\lambda_{\max} - n}{n - 1} \tag{6}$$

其中, n 为判断矩阵的阶数。

最后计算一致性比率 CR , CR 的计算如式(7)所示。

$$CR = \frac{CI}{RI} \quad (7)$$

其中, RI 表示平均随机一致性, 通过查表 5 获取。

Table 5. RI section value

表 5. RI 部分取值

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.52	0.89	1.11	1.25	1.35	1.4	1.45

根据上述一致性检验的计算步骤, 该判断矩阵的一致性指标为 CI 为 0.005, 结合表 5 中提供的随机一致性指标, 当指标维度为 $n=4$ 时, RI 的取值为 0.89, 依据公式(7)计算一致性比率 CR 为 0.0056。由于 CR 值明显小于 0.1, 说明该判断矩阵具有良好的一致性, 满足层次分析法对一致性检验的基本要求。

4) 计算权重

在通过一致性检验之后, 就可以求解权重向量。根据式(8)判断矩阵 A 的每个元素按列归一化处理。

$$b_{ij} = \frac{a_{ij}}{\sum_{i=1}^n a_{ij}}, i, j = 1, 2, 3, \dots, n \quad (8)$$

计算得到归一化后的判断矩阵如表 6 所示。

Table 6. Normalized judgment matrix

表 6. 归一化后的判断矩阵

质量维度	完整性	准确性	冗余性	时效性
完整性	0.160	0.233	0.200	0.096
准确性	0.320	0.467	0.350	0.578
冗余性	0.040	0.067	0.050	0.036
时效性	0.480	0.233	0.400	0.289

得到归一化后的判断矩阵, 按式(9)对矩阵 B 的每一行元素求和。

$$c_i = \sum_j b_{ij}, i, j = 1, 2, 3, \dots, n \quad (9)$$

最后, 将每行元素除以列和, 如式(10)所示, 得到各维度指标 i 的权重。

$$w_i^A = \frac{c_i}{\sum_{i=1}^n c_i}, i = 1, 2, 3, \dots, n \quad (10)$$

结合式(9)和式(10)计算得到数据质量各指标权重 w_i^A 。最终, 基于层次分析法计算得到的数据质量指标的权重向量为 $W^A = [0.172, 0.429, 0.048, 0.351]$ 。

4.2. 熵权法

采用熵权法对数据质量维度的各项指标权重进行计算, 能够在一定程度上克服人为主观因素带来的

偏差，提升权重分配的客观性和科学性，从而使评价结果更加合理准确。熵权法的具体实施步骤如下：

1) 评语集和指标集的确定

传统的熵权系数法在处理评估问题时，通常假设存在 m 个评价对象以及 n 个评估指标，然而，这种设定方式并不完全符合本研究对数据质量评价的具体需求。因而本研究设定针对数据质量评价的集合为 $D=\{d_1, d_2, \cdots, d_m\}$ ， $d_i \cap d_j = \emptyset$ 。将数据质量评价分为 5 个等级，具体的等级划分规则如表 7 所示。

Table 7. Importance level of data quality evaluation indicators
表 7. 数据质量评价指标重要性等级

等级	评分	描述
非常不重要	1 分	该指标对整体评价几乎没有影响
不重要	2 分	该指标对整体评价影响较小
中等重要	3 分	该指标对整体评价有一定影响
重要	4 分	该指标对整体评价影响较大
非常重要	5 分	该指标对整体评价极为重要

数据质量评价维度指标的集合记为 $Q=\{q_1, q_2, \cdots, q_n\}$ ， $q_i \cap q_j = \emptyset$ 。

2) 评价矩阵的建立

采用德尔菲法，征集 12 位专家对数据质量评估指标重要程度的打分情况，进而构建起相应的评估矩阵 $T=(t_{ij})_{n \times m}$ ($i=1, 2, \cdots, n; j=1, 2, \cdots, m$)，其中 n 为数据质量维度数，即 $n=4$ ， m 为评价等级总数，即 $m=5$ 。评价矩阵的元素 t_{ij} 表示数据质量维度 q_i 被评为等级 d_j 的次数，初始评估矩阵 T 为 4×5 的矩阵，每一行对应的评价维度分别是完整性、准确性、冗余性和时效性，如式(11)所示。

$$T = \begin{bmatrix} 1 & 2 & 3 & 3 & 3 \\ 0 & 1 & 1 & 3 & 7 \\ 6 & 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 4 & 5 \end{bmatrix} \quad (11)$$

为体现不同评价等级的区分度和重要性，通过将每个评价等级对应的评分与评价频次相乘，得到评价矩阵的元素，如式(12)所示。

$$s_{ij} = t_{ij} \times d_j, \quad i, j = 1, 2, \cdots, n \quad (12)$$

由此，获得加权评价矩阵 $S=(s_{ij})_{n \times m}$ ，如式(13)所示。

$$S = \begin{bmatrix} 1 & 4 & 3 & 12 & 15 \\ 0 & 2 & 3 & 12 & 35 \\ 6 & 6 & 6 & 4 & 0 \\ 0 & 2 & 6 & 16 & 25 \end{bmatrix} \quad (13)$$

3) 指标权重的计算

在前文通过专家打分和统计分析构建了评价矩阵的基础上，进一步引入信息熵理论，以此来求解各指标的权重。

加权评估矩阵 S 用于反映各维度在不同评分等级下的加权得分。矩阵元素由评分值与相应被选次数相乘得出。由于加权后数值差异较大，若直接用于计算，可能造成权重失衡，因此有必要进行标准化处

理。为消除各指标间量纲不一致带来的影响,并保持评分之间的相对差异,本文采用极差标准化(Min-Max Normalization)方法对矩阵数据进行归一化处理,标准化公式如式(14)所示的标准化公式。

$$y_{ij} = \frac{s_{ij} - \min \{s_{i1}, s_{i2}, \dots, s_{im}\}}{\max \{s_{i1}, s_{i2}, \dots, s_{im}\} - \min \{s_{i1}, s_{i2}, \dots, s_{im}\}} \quad (14)$$

其中, s_{ij} 表示第 i 个维度在第 j 个评分等级下的加权得分, y_{ij} 为标准化后的值。该方法可将所有评分归一至[0,1]区间,保证各维度数据在同一数量级下进行信息熵的计算。由此得到如式(15)所示的标准化评估矩阵 $Y = (y_{ij})_{n \times m}$ 。

$$Y = \begin{bmatrix} 0.000 & 0.214 & 0.571 & 0.786 & 1.000 \\ 0.000 & 0.057 & 0.086 & 0.343 & 1.000 \\ 1.000 & 1.000 & 1.000 & 0.667 & 0.000 \\ 0.000 & 0.080 & 0.240 & 0.640 & 1.000 \end{bmatrix} \quad (15)$$

在完成矩阵的标准化处理后,我们能够更准确地计算评价指标 q_i 被评为等级 d_j 的相对频率 p_{ij} , 如式(16)所示。

$$p_{ij} = \frac{y_{ij}}{\sum_{j=1}^m y_{ij}}, \quad i=1, 2, \dots, n \quad (16)$$

则评价指标 q_i 的熵值 e_i 的计算如式(17)所示。

$$e_i = -\frac{1}{\ln m} \sum_{j=1}^m p_{ij} \ln p_{ij}, \quad i=1, 2, \dots, n \quad (17)$$

可求得数据质量指标的信息熵如表 8 所示。

Table 8. Entropy values of quality indicators information

表 8. 质量指标信息熵值

质量维度	完整性	准确性	冗余性	时效性
e_i	0.807	0.593	0.865	0.707

最后,可根据式(18)求解得到评价指标 q_i 的权重值 w_i^E 。

$$w_i^E = \frac{1 - e_i}{\sum_{i=1}^n (1 - e_i)} \quad (18)$$

则数据质量指标的客观权重向量为 $W^E = [0.188, 0.396, 0.131, 0.285]$ 。

4.3. 组合权重

上文详细地介绍了使用层次分析法和熵权法计算数据质量维度权重的过程,然而,这两种方法都存在一定的局限性,难以单独准确地反映评价指标的真实权重。其中,主观权重依赖于专家的专业知识和经验积累,能够体现领域专家的判断,但在一定程度上缺乏客观性和稳定性;而客观权重则基于数据的客观表现,能够较为真实地反映数据特征,但缺乏对特定领域专业知识的考量。因此,结合两种方法计算综合权重,有助于提高权重分配的客观性和准确性。

运用几何平均法,将主观权重与客观权重进行综合,从而计算得出组合权重 w_i , 其计算如式(19)所示。

$$w_i = \frac{\sqrt{w_i^A w_i^E}}{\sum_{i=1}^n \sqrt{w_i^A w_i^E}} \quad (19)$$

其中, w_i^A 为主观权重, w_i^E 为客观权重。

通过计算可得, 数据质量评价的组合权重向量为: $W = [0.182, 0.417, 0.080, 0.332]$ 。结果表明, 在本研究构建的数据质量评估模型中, 准确性和时效性是最为关键的指标。该结果在逻辑上与现实数据使用需求高度契合: 准确性保障了数据本身的可靠性, 是各类数据分析与决策的基础; 而时效性则确保数据内容能够真实反映当前状态, 是动态环境下数据价值得以实现的前提。

5. 数据质量评估模型应用

结合前述所确定的数据质量维度及其组合权重, 本文构建了一个加权评分模型, 以量化衡量数据质量水平的优劣, 具体计算如式(20)所示。

$$F_q = \sum_{i=1}^4 w_i \cdot Q_i \quad (20)$$

其中, w_i 表示第 i 个数据质量维度的权重, Q_i 表示样本数据在该维度上的得分。

该评价函数 F_q 可用于对数据样本整体质量的定量评估, 其值越高, 表明该组数据在多个质量维度上的表现越优, 整体数据质量水平越高。能够更全面、系统地反映数据质量的实际状态, 兼顾了主客观权重与客观表现的结合。实现对不同数据样本间质量差异的横向比较, 也为数据资源在流通、利用等阶段的质量评估奠定了方法基础。

6. 结论

在数字化转型不断加速的背景下, 数据已成为重要的战略性资源, 其质量直接关系到数据要素价值的释放和数字经济的健康发展。当前我国数据要素市场尚处于起步阶段, 亟需建立科学有效的数据质量评价体系, 以促进数字经济的发展。

为此, 本文聚焦于数据质量评估问题, 构建了包含准确性、完整性、冗余性和时效性四个维度的评价指标体系, 并引入层次分析法与熵权法相结合的方式, 确定各维度权重, 最终形成加权综合评分函数 F_q , 实现了对数据质量水平的量化衡量。研究结果表明, 准确性和时效性是影响数据质量的核心因素。在模型中权重最高, 说明数据是否真实可靠、能否及时更新, 是决定其有效性的关键要素。因此, 建议在数据采集、处理与管理过程中, 优先保障数据的准确性与更新的及时性, 通过完善校验机制与优化更新流程, 切实提升整体数据质量水平。本研究所构建的评估模型兼顾主观判断与客观数据特征, 具备较强的适用性和可扩展性。不仅为数据资源的流通与利用提供了评估依据, 也为构建规范、有序的数据要素市场提供了方法支持。

不过, 需要指出的是, 不同应用场景下, 各质量维度的重要性可能存在显著差异。例如, 在金融实时风控系统中, 时效性的重要性可能优于完整性; 而在医疗记录管理或政务档案等数据场景中, 完整性与冗余控制可能更受重视。因此, 本文模型的评估结果具有一定的情境依赖性, 需结合实际应用场景进行灵活调整。未来可进一步完善不同行业场景下的评价指标体系, 提升模型适用性。

参考文献

- [1] 蔡莉, 梁宇, 朱扬勇, 等. 数据质量的历史沿革和发展趋势[J]. 计算机科学, 2018, 45(4): 1-10.
- [2] 刘桂锋, 聂云贝, 刘琼. 数据质量评价对象、体系、方法与技术研究进展[J]. 情报科学, 2021, 39(11): 13-20.
- [3] 张宁, 袁勤俭. 数据质量评价述评[J]. 情报理论与实践, 2017, 40(10): 135-139.

- [4] 查先进, 陈明红. 信息资源质量评估研究[J]. 中国图书馆学报, 2010, 36(2): 46-55.
- [5] 郭亚军, 刘振阳, 冯思倩, 等. Aigc 质量评估指标体系构建研究——以 ChatGPT 为例[J]. 图书情报工作, 2024, 68(22): 93-104.
- [6] 宋俊典, 刘丰源. 一种支持数据质量评价的方法与应用研究[J]. 计算机应用与软件, 2018, 35(5): 328-333.
- [7] 王杨琛, 林佳能, 苏志勇. 基于灰色熵权法的大数据质量评估[J]. 微型电脑应用, 2022, 38(1): 110-113.
- [8] 杨德胜, 杨栋枢. 基于熵权和层次分析法的数据质量评估研究[J]. 现代电子技术, 2013, 36(22): 39-42.
- [9] 计蓉, 侯慧娟, 盛戈峰, 等. 基于组合赋权法和模糊综合评价的电力设备状态数据质量评估[J]. 高电压技术, 2024, 50(1): 274-281.
- [10] Xu, T., You, X., Yang, M., Shi, Y. and Mao, R. (2024) Data Asset Quality Evaluation Framework Based on a Hybrid Multi-Criteria Decision-Making Method. *Quality and Reliability Engineering International*, **41**, 801-820. <https://doi.org/10.1002/qre.3692>
- [11] Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S. and Aggarwal, S. (2023) Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. *Journal of the Knowledge Economy*, **15**, 1159-1178. <https://doi.org/10.1007/s13132-022-01096-6>
- [12] 王梦婷. 移动应急结构化大数据质量评价及提升方法研究[D]: [硕士学位论文]. 西安: 西安理工大学, 2024.
- [13] Gong, Y., Liu, G., Xue, Y., Li, R. and Meng, L. (2023) A Survey on Dataset Quality in Machine Learning. *Information and Software Technology*, **162**, Article 107268. <https://doi.org/10.1016/j.infsof.2023.107268>
- [14] 王景鸿. 面向数据市场交易量的两阶段数据定价研究[D]: [硕士学位论文]. 武汉: 中南财经政法大学, 2023.
- [15] 黄倩倩, 赵正, 刘钊因. 数据流通交易场景下数据质量综合管理体系与技术框架研究[J]. 数据分析与知识发现, 2022, 6(1): 22-34.
- [16] Taleb, I., Serhani, M.A., Bouhaddioui, C. and Dssouli, R. (2021) Big Data Quality Framework: A Holistic Approach to Continuous Quality Management. *Journal of Big Data*, **8**, Article No. 76. <https://doi.org/10.1186/s40537-021-00468-0>
- [17] Zhang, H. and Xiao, J. (2020) Quality Assessment Framework for Open Government Data. *The Electronic Library*, **38**, 209-222. <https://doi.org/10.1108/el-06-2019-0145>
- [18] Yang, J., Zhao, C. and Xing, C. (2019) Big Data Market Optimization Pricing Model Based on Data Quality. *Complexity*, **2019**, Article 5964068. <https://doi.org/10.1155/2019/5964068>
- [19] Cichy, C. and Rass, S. (2019) An Overview of Data Quality Frameworks. *IEEE Access*, **7**, 24634-24648. <https://doi.org/10.1109/access.2019.2899751>
- [20] El Alaoui, I., Gahi, Y. and Messoussi, R. (2019) Big Data Quality Metrics for Sentiment Analysis Approaches. *Proceedings of the 2019 International Conference on Big Data Engineering*, Hong Kong SAR, 11-13 June 2019, 36-43. <https://doi.org/10.1145/3341620.3341629>
- [21] 王明政, 孟伟强, 黄建华, 等. 基于层次分析法的教育网络安全数据质量管理研究[J]. 信息安全, 2019(12): 79-87.
- [22] 李晶晶. 数据质量评估模型及评估工具研究[D]: [硕士学位论文]. 上海: 东华大学, 2018.
- [23] 张晓冉, 袁满. 通用数据质量评估模型及本体实现[J]. 计算机研究与发展, 2018, 55(6): 1333-1344.
- [24] Stahl, F. and Vossen, G. (2016) Fair Knapsack Pricing for Data Marketplaces. In: *Lecture Notes in Computer Science*, Springer, 46-59. https://doi.org/10.1007/978-3-319-44039-2_4