文章编号：2095-2163(2019)04-0071-05 中图分类号：TP311.13 文献标志码：A

数据质量量化评价研究与实现

庄计龙[1,2]，陈敏刚[2]
（1 东华大学 计算机科学与技术学院，上海 201620；2 上海市计算机软件评测重点实验室，上海 201112）

摘 要：近年来，随着科学技术的飞速发展，信息化、数字化社会正在形成。伴随而来的是数据质量问题越来越凸显。本文在分析了当前数据质量评价标准的基础上，确定以 GB/T 25000.24 为基础构建数据质量评价模型，并对指标权重进行研究。相比使用单个权重计算方法，本文综合 Delphi 法、层次分析法和基于信息熵的熵权系数法计算综合权重，使得权重进一步客观。针对当层次分析法的判断矩阵经计算不满足一致性时，重新构造判断矩阵成本高的问题，文章引入了诱导矩阵修正法来修正判断矩阵以尽可能避免重新构造判断矩阵。最后本文开发了相应的数据质量评价系统，有效地提高了数据质量评价工作的质量和效率。

关键词：数据质量；评价模型；层次分析法；熵权系数法

Research and implementation of quantitative evaluation of data quality

ZHUANG Jilong[1,2], CHEN Mingang[2]
(1 School of Computer Science and Technology, Donghua University, Shanghai 201620, China;
2 Shanghai Key Laboratory of Computer Software Evaluating and Testing, Shanghai 201112, China)

【Abstract】 In recent years, with the rapid development of science and technology, informationization and digital features of the society have attracted more and more attention. Simultaneously, the data quality issues are becoming more and more prominent. Based on the analysis of current data quality evaluation standards, this paper determines the data quality evaluation model based on GB/T 25000.24 and studies the index weights. Compared with the single weight calculation method, the Delphi method, the analytic hierarchy process and the entropy weight coefficient method based on information entropy are integratedly used to calculate the comprehensive weight, which makes the weight more objective. Aiming at the problem that the judgment matrix of the analytic hierarchy process does not satisfy the consistency and the cost of reconstructing the judgment matrix is expensive, the paper introduces the induction matrix correction method to modify the judgment matrix to avoid reconstructing the judgment matrix as much as possible. Furtherly, the paper develops a corresponding data quality evaluation system, which effectively improves the quality and efficiency of data quality evaluation.

【Key words】 data quality; evaluation model; analytic hierarchy process; entropy weight coefficient method

0 引 言

近年来，随着科学技术的飞速发展，信息化、数字化社会正在形成。计算机系统软件已经渗透到生活的各个方面，这些软件不断地产生新的海量数据。此外，不仅仅是 IT 行业，越来越多的行业涉及到了数据的处理，如银行、保险、零售业、等等，数据已经成为新时代最重要的资产之一[1]。

但这些数据可能由于人为录入的错误、人为篡改、机械故障等原因，往往会存在数据属性缺失、数据相似重复、数据属性值异常等问题。这些错误可能会造成数据冗余，浪费存储的空间，甚至可能导致数据分析挖掘时产生严重的偏差[2]。在对数据进行分析挖掘之前，数据质量的好坏对于人们能否准确利用数据获得决策信息非常重要，甚至决定着数据应用的成败[3]。虽然目前关于数据质量的研究已经蓬勃兴起，但工作主要集中在数据的存储、管理、挖掘分析等方面，数据质量问题没有得到足够的重视[4]。这些缺失数据或错误数据等原因导致了数据不能很好地利用，甚至造成很大的决策失误。因此已有越来越多的专家、学者意识到数据质量对数据分析挖掘的重要性并投身于相应的数据质量研究中。

1 构建数据质量评价模型

1.1 GB/T 数据质量模型

数据质量研究的诞生和发展主要是在国外，因

基金项目：上海市科学技术委员会项目(18DZ2203700)。

作者简介：庄计龙(1992-)，男，硕士研究生，主要研究方向：数据质量、大数据、数据可视化；陈敏刚(1978-)，男，博士，副研究员，主要研究方向：大数据与人工智能的测试与应用。

通讯作者：陈敏刚 Email：cmg@ ssc.stn.sh.cn

收稿日期：2019-04-11

此早期国内相关研究中的主要理论依据都是根据ISO/IEC发布的一系列标准。随着国内对数据质量的关注度逐渐提高,中国对数据质量测量的标准化也有了实质性的进展。在GB/T 25000.12-2017和GB/T 25000.24-2017(2018年5月1日开始实施)这2个国家标准中,为计算机系统中以某种结构化形式保存的数据定义了一种通用的数据质量模型,从固有的以及依赖系统的角度划分了质量特性以及对应的属性。其中包括15个特性,63个属性。

1.2 裁剪构建数据质量评价模型

裁剪指标的依据来源主要有:

(1)根据最新的国家相关数据质量标准;

(2)咨询相关领域的专业人士的意见;

(3)上海软件中心实习期间的见闻;

(4)统计相关信息系统的指标要素构成。

通过裁剪所得到的数据质量评价模型完备性、一致性、依从性、准确性、唯一性、现时性和保密性等7个一级指标构成。

2 改进数据质量评价指标权重分配方法

2.1 改进的层次分析法

处理数据质量评价过程中的权重分配需要使用层次分析法[5]。这里使用的层次分析法与传统意义上的层次分析法有区别,因而要做相应的改变。重新定义层次分析法的层次结构为目标层、指标维度层。因此新的层次分析法使用步骤如下:

(1)构建层次结构模型。层次分析法是确定权重的基础。首先需要通过对数据的理解和分析去设定顶层也即目标层,其次需要确定指标维度层;

(2)判断矩阵的建立与计算。通过所有指标维度的两两比较,然后按照某一尺度建立。这里通过邀请专家根据Santy提出的1-9标度方法作为评价尺度来建立判断矩阵;

(3)求解权重向量。在层次分析法中,最根本的计算任务就是求解判断矩阵的最大特征根及其所对应的特征向量。本文通过比较,在不损失精度的前提下,选择计算相对比较简单快捷的和积法作为求解权重向量的方法。设判断矩阵为 $\boldsymbol{D} = (d_{ij})_{n\times n}$。其具体计算步骤如下:

Step 1 $\boldsymbol{D}$ 中的元素按列归一化,即求:

$$\bar{d}_{ij} = d_{ij} / \sum_{k=1}^{n} d_{kj}, \quad i,j = 1,2,\cdots,n, \tag{1}$$

Step 2 将归一化后的矩阵的同一行的各列相加,即求:

$$\widetilde{w}_i = \sum_{j=1}^{n} \bar{d}_{ij}, \quad i = 1,2,\cdots,n, \tag{2}$$

Step 3 将相加后的向量除以 n 即得权重向量,即求:

$$w_i = \frac{\widetilde{w}_i}{n}, \quad i = 1,2,\cdots,n, \tag{3}$$

Step 4 计算最大特征根为:

$$\lambda_{max} = \frac{1}{n}\sum_{i=1}^{n} \frac{(Dw)_i}{w_i}, \quad i = 1,2,\cdots,n. \tag{4}$$

其中,$(Dw)_i$ 表示向量 Dw 的第 i 个分量。通过上述计算流程即可求得权重向量 W,$W = \{w_1, w_2, \cdots, w_n\}$。

(4)一致性校验。定义判断矩阵 $\boldsymbol{D}$ 的一致性指标(Consistence Index,C.I.)为:

$$C.I. = \frac{(\lambda_{max} - n)}{n-1}, \tag{5}$$

引入了一致性比例(Consistence Ratio,C.R.)这个一致性评价 ,即:

$$C.R. = \frac{C.I.}{R.I.}. \tag{6}$$

其中,$R.I.$ 为随机一致性指标(Random Consistency Index)。对于一致性比例,当 $C.R. < 0.1$ 时,认为该判断矩阵通过一致性校验,说明该判断矩阵的不一致性程度在容许范围内,则由其导出的特征向量即可作为子特性的权重向量。当 $C.R. > 0.1$ 时,称 D 不具有一致性。一般需要再次构造判断矩阵重复上述过程。为解决重新构造判断矩阵成本高的问题,文章引入了诱导矩阵修正法来修正判断矩阵以尽可能避免重新构造判断矩阵。具体说来:当阀值 $0.1 < C.R. < 0.2$ 时引入诱导矩阵修正法[6],记为AHPIM(Analytic Hierarchy Process with Induced Matrix)。

诱导矩阵修正法的计算步骤如下:

Step 1 计算判断矩阵 $\boldsymbol{D}$ 的每一列向量的归一化向量 $b_j = (b_{1j}, b_{2j}, \cdots, b_{nj},)^{\mathrm{T}}$,和排序向量 $w_i = (w_1, w_2, \cdots, w_n)^{\mathrm{T}}$。

Step 2 求得诱导矩阵 $\boldsymbol{C} = (c_{ij})_{n\times n}$。

Step 3 找出令 $|c_{ij} - 1|$ 的结果最大的一组 (i,j) 记为 (e,f)。

Step 4 若 $c_{ef} > 1$,转 Step4.1;否则,转到 Step4.2。

Step 4.1 若 $d_{ef} > 1, d'_{ef} = d_{ef} - 1$,转 Step5;若 $d_{ef} < 1, d'_{ef} = d_{ef}/(1 + d_{ef})$,转 Step5。

Step 4.2　若 $d_{ef} > 1, d'_{ef} = d_{ef} + 1$，转 Step5；若 $d_{ef} < 1, d'_{ef} = d_{ef}/(1 - d_{ef})$，转 Step5。

Step 5　令 $d'_{fe} = 1/d'_{ef}, d'_{ij} = d_{ij}, i,j = 1,2,\cdots,n$，其中 $(i,j) \neq (e,f)$。

Step 6　判断修改后的判断矩阵是否满足一致性。若满足则停止修正，否则转 Step1。

2.2　面向权重的熵权系数法

这里引入基于信息熵[7]的熵权系数法[8]。如果某评价指标的熵越小，说明该指标提供的信息量就越大，在综合评价中所起的作用就越大，权重就越高。反之，若评价指标的熵越大，说明该指标提供的信息量就越小，在综合评价中所起的作用就越小，权重就越低[9]。应用熵权系数法可以尽可能消除人为因素对计算各指标权重的影响，使评价结果更为准确。

在本文实际的数据质量评价中，使用熵权系数法进行权重值求取的步骤如下。

2.2.1　评语集和指标集的确立

原始的熵权系数法所考虑的评估问题，一般是设有 n 个评价对象（方案），m 个评估指标，这样的设定方法并不适合本文数据质量评价的需求。因而本文对其所表述的含义进行如下修改，并将其记为 WEWCM（Weight-Oriented Entropy Weight Coefficient Method）。

设数据质量评语的集合记为 $C = \{c_1, c_2, \cdots, c_n\}$，$c_i \cap c_j = \varnothing$。设数据质量评价维度指标的集合记为 $D = \{d_1, d_2, \cdots, d_m\}$，$d_i \cap d_j = \varnothing$。

2.2.2　评价矩阵的建立

评价矩阵主要是通过采取专家评判和统计的知识，设专家人数为 t，即是对指标 d_i 评价的次数，t_{ij} 为指标 d_i 被评为评语集合中评语 c_j 的次数，则可得到初始评价矩阵 $\boldsymbol{T}$ 如下所示。

$$\boldsymbol{T} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix}.$$

2.2.3　指标权重的求解

上文通过专家打分和统计已经构建了初始评价矩阵。在上述步骤的基础上，就可结合信息熵的知识进行指标权重的求解。其具体计算步骤如下：

Step 1　初始评价矩阵 $\boldsymbol{T}$ 中，评价指标 d_i 被评为等级 c_j 的频率：

$$f_{ij} = \frac{t_{ij}}{\sum_{j=1}^{n} t_{ij}}, \quad i, = 1,2,\cdots,m, \tag{7}$$

Step 2　评价指标 d_i 的熵：

$$h_i = -\frac{1}{\ln n}\sum_{j=1}^{n} f_{ij}\ln f_{ij}, \quad i = 1,2,\cdots,m, \tag{8}$$

Step 3　评价指标 d_i 的熵权：

$$w_i = \frac{1 - h_i}{\sum_{i=1}^{m}(1 - h_i)} = \frac{1 - h_i}{m - \sum_{i=1}^{m} h_i}. \tag{9}$$

2.3　综合 AHPIM 与 WEWCM 权重

若把改进的层次分析法得到的权重结果记为 $W = \{W_1, W_2, \cdots, W_n\}$，把由面向权重的熵权系数法得到的权重结果记为 $w = \{w_1, w_2, \cdots, w_n\}$，则融合后的综合权重为：

$$\hat{w}_i = (W_i \times w_i) / \sum_{i=1}^{n} (W_i \times w_i). \tag{10}$$

3　数据质量量化评价设计与实现

3.1　功能性设计

功能性设计如图 1 所示。

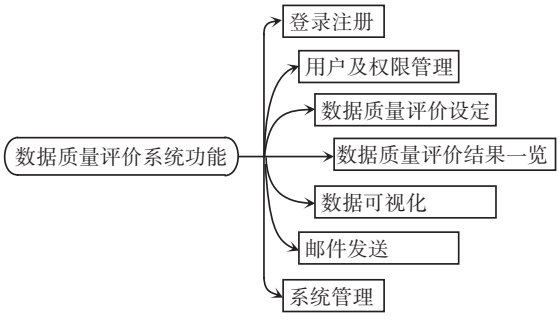

图 1　数据质量评价系统功能结构

Fig. 1　Data quality evaluation system functional structure

3.2　技术架构设计

结合最新的前后端分离技术，以及对各类技术应用研究和分析，设计系统的技术采用 B/S 架构[10]，如图 2 所示。

在前后端分离总体架构的基础上，逻辑上将技术架构分为 4 个层次，分别是视图层、业务逻辑层、数据访问层和数据层。前后端分离后，难以避免跨域问题。解决跨域问题核心代码如下：

```
public void addCorsMappings(CorsRegistry registry) {
    registry.addMapping("/**")
        .allowedOrigins("*")
```

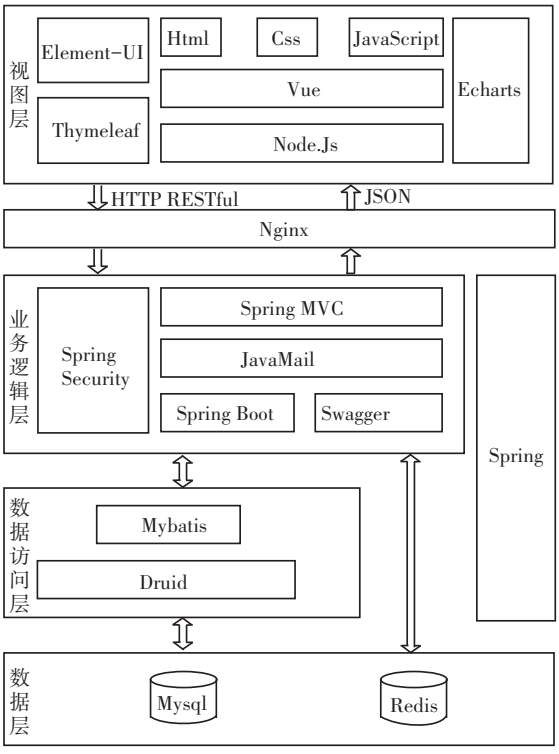

图 2 系统技术架构

Fig. 2 System technology architecture

```
        .allowCredentials(true)
        .allowedMethods("GET","POST", "DELETE",
"PUT")
        .maxAge(3600);
    }
```

3.3 系统功能模块实现

首先明确系统的开发环境和开发工具,前端基于 Node 框架,所使用的开发工具为 WebStorm,后端基于 JDK1.8,所使用的开发工具为 IntelliJ IDEA。这里仅给出数据质量评价配置模块的实现说明。

数据质量评价配置模块是本文所开发系统的核心功能模块,具体实现流程是:在前端系统的数据质量综合量化指标维度编辑界面,设置好相应规则约束等字段,然后把数据以 JSON 的形式发送给后端进行相应指标计算,并将结果保存到数据库和 Redis 缓存中,供后面计算总得分、可视化以及评价报告使用。这部分为了提高运算的速度,充分发挥 CPU 的性能,系统使用线程池技术。模块时序如图 3 所示。

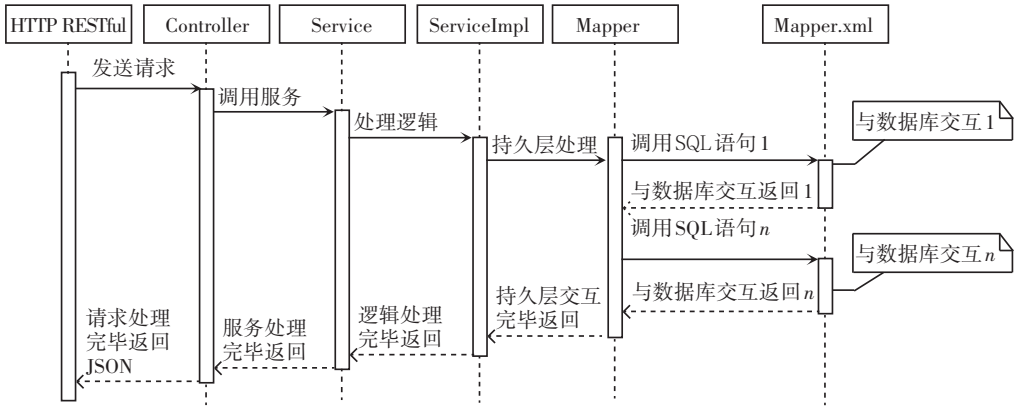

图 3 系统模块时序图

Fig. 3 System module timing diagram

4 实 验

文章使用真实电商领域的数据集进行数据质量评价实验。

(1)利用 AHPIM 计算权重。通过一系列步骤算出权重为:

$W = \{0.094, 0.054, 0.104, 0.037, 0.134, 0.292, 0.285\}$,

(2)利用 WEWCM 计算权重。通过一系列步骤算出权重为:

$w = \{0.149, 0.184, 0.149, 0.230, 0.184, 0.070, 0.035\}$,

(3)综合 AHPIM 与 WEWCM 计算综合权重。

根据公式(10)求得质量维度的综合权重为:

$\hat{w} = \{0.136, 0.097, 0.151, 0.083, 0.239, 0.198, 0.096\}$.

在确定了指标的综合权重后,权重也作为电商领域数据的默认权重保存到系统中。接下来在所设计并实现的数据质量评价平台上评价数据的数据质量。最后得到评价分数如图 4 所示。

(下转第 78 页)