计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(11):165−169 [doi: 10.15888/j.cnki.csa.006046]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于规则库的数据质量评估方法①

刘　芳[1], 李　敏[2], 任洪敏[1], 周兆明[3]

[1](上海海事大学 信息工程学院, 上海 201306)
[2](青岛西海岸新区管委, 青岛 266555)
[3](上海产业研究院, 上海 201306)

摘　要: 在当今大数据时代下, 数据质量的保证是大数据价值得以发挥的前提, 数据质量的评估是其中一个重要的研究课题. 本文基于规则库的数据质量评估方法, 提出了数据质量评估整体模型, 包括规则、规则库、数据质量评估指标、评估模板、评估报告. 设计了规则评估模板, 组合规则库中的规则, 根据数据质量评估指标的重要性设置规则的权重, 采用简单比率法和加权平均法相结合的评估方法, 计算评估结果并确定数据质量的等级, 利用了数据可视化技术来展现数据质量的评估结果. 本文既考虑了单个规则的执行合格率, 又考虑了各规则在数据质量评估模板中的比重, 公正地准确地评估数据质量, 并且简洁、直观地呈现评估结果.

关键词: 规则库; 数据质量; 评估模板; 数据可视化

引用格式: 刘芳,李敏,任洪敏,周兆明.基于规则库的数据质量评估方法.计算机系统应用,2017,26(11):165–169. http://www.c-s-a.org.cn/1003-3254/6046.html

Data Quality Evaluation Method Based on Rule Base

LIU Fang[1], LI Min[2], REN Hong-Min[1], ZHOU Zhao-Ming[3]

[1](College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)
[2](Qingdao West Coast New District Administrative Committee, Qingdao 266035, China)
[3](Shanghai Industrial Research Institute, Shanghai 201306, China)

Abstract: In today's era of big data, data quality is the premise of the significance of big data. The evaluation of data quality is one of the most important research topics. In this paper, the data quality assessment method based on rule base is put forward, and the overall model of data quality assessment is presented, which includes rules, rule base, data quality evaluation index, evaluation model and evaluation report. This paper designs the rule evaluation template, combines rules in the rule base, sets rule weight according to the importance of data quality evaluation index, adopts the evaluation method that combines the simple ratio method and the weighted average method, calculates the evaluation result, determines the grade of the data quality, and shows the evaluation result of data quality with the data visualization technology. In order to fairly and accurately assess the data quality, and concisely and intuitively present the evaluation results, the paper does not only consider the execution rate of a single rule, but also considers the proportion of each rule in the data quality evaluation template.

Key words: rule base; data quality; evaluation template; data visualization

随着网络和信息技术的不断发展, 各行各业都已经开始使用信息化技术, 并且在业务处理、交流中慢慢积累了大量的业务数据, 并且这些数据呈指数增长, 我们已进入到一个大数据时代. 在大数据时代下对于企业来说, 抓住大数据时代带来的机遇和优势, 是企业的核心竞争力. 但是保证数据的准确性、有效性, 即数

① 基金项目: 上海市科委重点项目 (SKY2015004)
收稿时间: 2017-02-23; 修改时间: 2017-03-09; 采用时间: 2017-03-13

据的质量, 是发挥大数据在商业决策中作用的前提. 因此如何评价、保证数据的质量, 已成为一个至关重要的问题. 目前关于数据质量的研究工作大致可以分为以下几类: 数据质量评估模型[1,2]、数据质量评价方法[3-6]、数据质量规则库模型[7]、数据质量评估算法[8,9]、数据质量评估在各个领域的应用[10-12].

文中提出了一套完整的基于规则库的数据质量评估方法, 由于规则库是通用的, 设计了数据质量评估模板, 针对具体的数据组合成不同的规则模板, 设置权重, 采用简单比率法和加权平均法计算评估结果, 并采用数据可视化技术, 简洁地、直观地呈现数据质量分析报告.

1 数据质量评估模型

1.1 数据质量评估框架及流程

设计的基于规则库的数据质量评估方法的框架如图 1 所示, 其组成部分包括: 规则库、数据质量评价指标、规则、评估模板、评估报告五个部分.

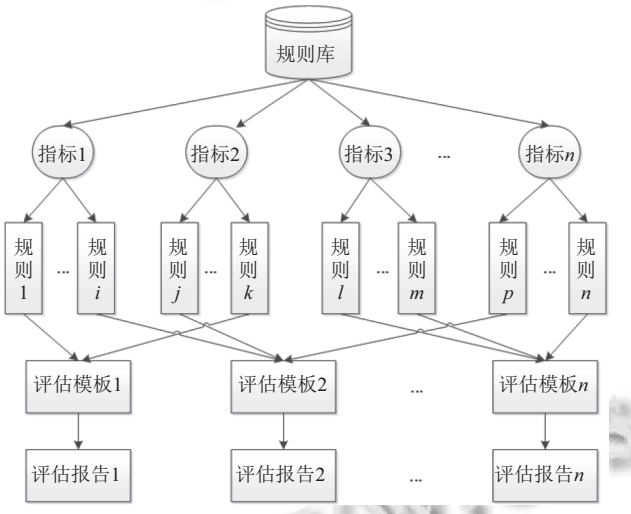

图 1 数据质量评估框架图

图 1 的评估框架图清楚的地展示了数据质量的评估流程, 首先设计数据质量的规则库, 定义数质量评价指标, 设计规则并赋予该规则所依据的数据质量评价指标, 针对具体的数据和规则库中的规则组合数据质量评估模板, 并且设置评估模板中规则的权重, 最终计算评估结果, 生成评估报告.

该框架面向数据仓库全体数据, 保障数据质量评估的准确和完整, 周期性的对仓库的增量数据实施评估.

1.2 数据质量评估指标

数据资源不同与产品, 具有用途个体化、多样化、不稳定等特点. 数据质量评价指标受行业领域、数据类型和应用目的等因素的影响极大, 较难制定面向所有学科领域的普适性数据质量指标体系. 为了对数据质量进行更加深入的分析和评估, 常常将数据质量划分为若干个更具体的数据质量评价维度. 不同的研究者有不同的划分方法. Diane M. Strong 等提出了一个目前被广泛引用的数据质量评估框架, 这个框架将数据质量划分为内在质量、可访问性质量、上下文质量和表达质量四个大的质量类, 每个质量类又可以再细分为若干更具体的质量维度[13].

因此将从准确性、完整性、一致性、可信性、时效性、易访问性、依从性、保密性、效率性、精准性、回溯性、易理解性、可用性、可移植性和易恢复性 15 个维度来评价数据质量, 如图 2 所示.

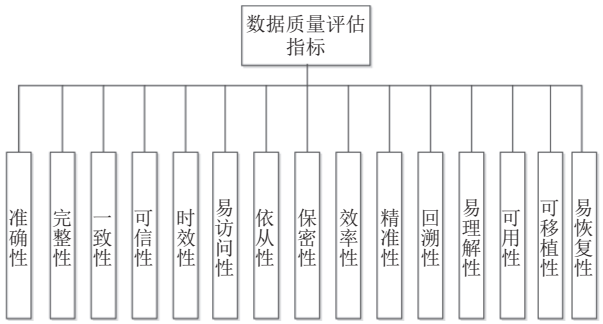

图 2 数据质量评估指标图

(1) 准确性, 用于描述一个值与它所描述的客观事物的真实值之间的接近程度.

(2) 完整性, 用于描述为解决问题所获得的数据的广度、深度和规模足够充分.

(3) 一致性, 用于描述数据项遵循预定的语法规则的程度. 主要包括: 参照完整性、数据格式一致性、数据不一致的风险度、结构的一致性、数据值一致性覆盖程度、语义一致性.

(4) 可信性, 是一个集合性术语. 它用来表示可用性及其影响因素: 可靠性、维修性、保障性, 它常用于非定量条款中的一般性描述.

(5) 时效性, 是指信息仅在一定时间段内对决策具有价值的属性.

(6) 易访问性, 主要包括: 用户易访问性、设备易访问性、数据格式易访问性.

(7) 依从性, 主要包括: 数据值依从性、数据格式依从性和技术依从性.

(8) 保密性, 又称机密性, 其与 Integrity(完整性) 和 Availability(可用性) 并称为信息安全的 CIA 三要素.

(9) 效率性, 是指数据处理过程中投入时间与得到成果之间的对比关系.

(10) 精准性, 是指数据的准确性高和精度高. 主要包括数据值的精确性、数据格式的精确性.

(11) 回溯性, 是指数据值本身、用户访问和系统依赖的数据值的可回溯性.

(12) 易理解性, 主要是指符号、语义、主数据、数据值、数据模型、数据呈现、和链接主数据的易理解性.

(13) 可用性, 是指数据对用户来说有效、易学、高效、好记、少错和令人满意的程度.

(14) 可移植性, 指将数据从某一种存储方式转换到另一种存储方式的难易程度.

(15) 易恢复性, 是指数据丢失、缺失、被改动之后的恢复程度, 即数据的备份.

2 规则库与规则模板

2.1 规则及规则库

如何有效的评估数据质量, 关键在于数据质量规则的制定. 数据规则, 又称数据约束, 是客观世界的数据所应遵循的语义限制, 包括领域知识和业务规则[14]. 将所有的规则组织在一起, 又叫做规则库. 数据质量的分析, 是基于规则的定义, 对于不同的数据集, 不同的业务数据, 规则的制定是不同的, 因此本文基于“规则库”[15]数据质量评估方法, 建立一种可适用于大多数数据集的数据质量评估方法, 使得数据质量评估工具有通用性.

设计的数据质量规则, 包括序列标识、条件类型、源数据、操作符、参考数据类型、参考数据、规则名称、评价指标和操作九个元素. 如图 3 所示.

(1) 序列标识, 标识是第几条规则, 如果几个语句的序列标识相同, 说明这几个语句属于同一条规则.

(2) 条件类型, 主要定义了 IF、AND、OR, 用来表示同一条规则中的几条语句之间的关系.

(3) 源数据, 指待评估的数据.

(4) 运算符, 指源数据和参考数据之间的关系, 主要定义了 is、is not、is within、is not within、contain、<、<=、>、>=, 在将来的数据质量评估过程中对于具体的数据评估, 会增加运算符, 运算符体系将会越来越完善.

(5) 参考数据类型, 表示参考数据的数据类型.

(6) 参考数据, 指将要与源数据进行比较的数据, 可以是用户自己定义, 也可以是系统内定的.

(7) 规则名称, 简要说明该规则的功能.

(8) 评价指标, 指评价数据质量的维度.

(9) 操作, 指源数据符合或者不符合一条规则后根据需要进行数据统计、清洗等.

图 3 规则管理界面图

规则库的设计, 如图 4 所示.

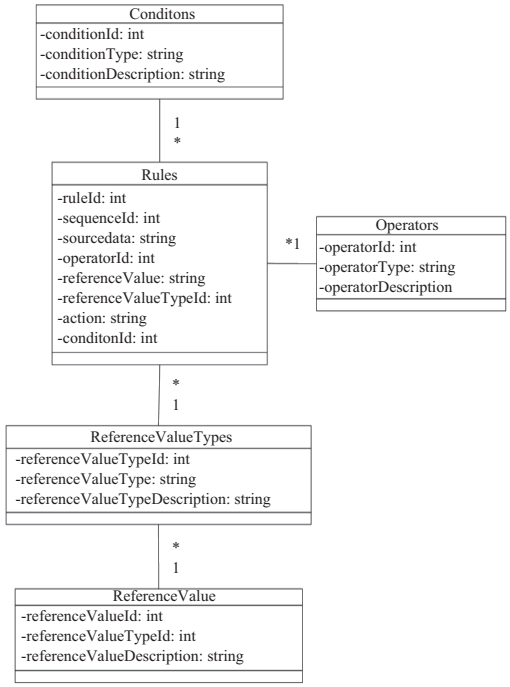

图 4 规则库设计图

2.2 数据质量评估模板

由于规则库是为了适用于大多数的数据质量评估, 而不是具体的、特定的数据, 因此评估不同的数据需要不同的规则, 就需要不停的删除规则、创建规则. 针

对此问题, 本文提出了使用数据质量评估模板, 用若干规则组成一个模板, 针对某具体数据进行评估, 并且生成评估报告. 设计的规则模板包括: 模板 Id、序列标识、规则名称、规则类型、是否已经执行和权重六个元素, 如图 5 所示.

添加 删除 修改 保存 模板ID Please Input Value

模板Id	序列标识	规则名称	规则类型	是否已经执行	权重
1	1	用户名或密码不能为空	完整性	是	0.1
1	2	用户ID不能为空且为数字	完整性	是	0.2
1	3		可信性	是	0.1
1	4		时效性	是	0.1
1	5		易访问性	是	0.2
1	6		完整性	是	0.2
1	7		精准性	是	0.05
1	8		完整性	是	0.05
2	2	用户ID不能为空且为数字	完整性	是	0.2
2	4		时效性	是	0.3
2	7		精准性	是	0.1
2	8		完整性	是	0.4
3	1	用户名或密码不能为空	完整性	是	0.3
3	3		可信性	是	0.3
3	5		易访问性	是	0.2
3	8		完整性	是	0.2

图 5 评估模板界面图

(1) 模板 Id, 用来表示那些规则属于哪一个模板.

(2) 序列标识, 同规则中的序列标识, 表示一条规则.

(3) 规则名称, 同规则中的规则名称, 简要描述规则的功能.

(4) 规则类型, 同规则中的评价指标.

(5) 是否已经执行, 指该条规则是否已经执行.

(6) 权重, 表示该条规则在模板中的重要性, 一个模板中的所有规则的权重之和为 1.

3 数据质量评估方法

3.1 数据质量指标计算方法

文献[16]中, 提出了三种数据质量评价方法: 一是简单比率法, 指期望的结果 (E) 占总值 (T) 的比率 E/T, 反映数据质量某些方面的好坏程度; 二是最小/大值法, 适用于衡量数据质量中需要对多种指标进行加总的维度, 评价的关键是要找出各类指标中的最大值或最小值. 最小值法是一种保守的评估方法, 它赋给维度一个不超过它的最差数据质量指标的值. 最大值是一种不保守的评估方法, 一般适用于比较复杂的度量体系; 三是加权平均法, 为了确保评价值标准化, 每个指标的权重必须被限定在 0 和 1 之间, 并且他们的和等于 1, 即 $\lambda_1+\lambda_2+\ldots+\lambda_n=1$, $X=\lambda_1X_1+\lambda_2X_2+\ldots+\lambda_nX_n$, 其中 X_i 代表数据质量评价指标, λ_i 代表评价指标的权重, $i=1, 2, \ldots, n$.

考虑到待评估的大数据量和评估性能问题, 实施简单、快速的质量评估, 采用简单比率法和加权评平均法相结合的方法, 并且将文献[16]中提出的加权评价法融入的数据质量评估模板中. 方法描述如下:

Step1. 执行规则, 采用简单比率法, 所有符合规则的数据数 (F) 占所有的源数据数 (S) 的比率 F/S, 即每条规则执行合格率 R=F/S.

Step2. 应用某一评估模板对某一特定的数据进行评估, 并在评估模板中设置规则的权重, 并且一个模板中的所有规则的权重之和为 1, 即 $W_1+W_2+\ldots+W_n=1$, $(W_1, W_2, \ldots, W_n)$ 属于 M_i, 其中 M_i 代表某个模板.

Step3. 最后将每条规则的执行结果和每条规则的权重数之积相加, 就得出某一模板的评估结果, 即 $S=(R_1W_1+R_2W_2+\ldots+R_nW_n)*100$, R_i 代表某条规则的执行结果, W_i 代表某条规则在同一模板中的权重数, S 代表某一模板的评估结果.

3.2 评估等级计算

将每一条规则都转化为正则表达式, 匹配源数据与参考数据, 统计出合格数据所占比例, 然后再结合加权平均法计算出最终的评估分数. 根据分数将数据质量分为 A、B、C、D、E 五个等级: A 级为质量最优的数据, 分数在 90 到 100 分之间; B 级的数据质量为良, 分数在 80 到 89 之间; C 级的数据质量为中, 分数在 70 到 79 之间; D 级的数据质量为合格, 分数在 60 到 69 之间; E 即的数据质量为差, 即不能使用的数据, 需要进行数据清洗和数据转换, 分数在 60 以下. 表 1 将展现评估模板 1 对数据库用户表的评估结果, 其中该模板包含规则 1、规则 2、…、规则 8 共八条规则, 且八条规则的权重已经根据规则的评估指标的重要性给予赋值.

因此, 由评估等级可以看出来, 该模板评估处理的数据质量等级为 B. 本文提出的评估体系和评估方法, 即考虑了模板中各个规则的重要性, 又考虑了各个规则执行后数据的合格率, 精确地评估出数据的质量.

4 评估结果可视化

ECharts, Enterprise Charts 商业产品图表库, ECharts 开源来自百度商业前端数据可视化团队, 基于 html5 Canvas, 是一个纯 Javascript 图表库, 提供直观, 生动, 可交互, 可个性化定制的数据可视化图表.

表 1 数据质量评估结果

规则ID	源数据总数	合格数据数	合格率	权重
1	239	228	0.95	0.1
2	239	239	1	0.2
3	239	230	0.96	0.1
4	239	199	0.83	0.1
5	239	226	0.95	0.2
6	239	151	0.63	0.2
7	239	203	0.85	0.05
8	239	137	0.57	0.05
总结	(0.95*0.1+1*0.2+0.96*0.1+0.83*0.1+0.95*0.2+0.63*0.2+0.85*0.05+0.57*0.05)*100=86.1			

因此, 本文使用此技术来实现评估结果的可视化, 如图 6 将展现评估模板 1 对数据库用户表的评估结果, 其中该模板包含规则 1、规则 2、…、规则 8 共八条规则.

由图 6 可以看出, 规则 2、5、6 的权重值比较高, 并且该三个规则执行时数据的合格率比较高, 因此得到的规则评估分数中, 这三个规则的评估分数所占的比率比较高, 即该模板所评估的数据质量的评估等级很大程度上取决于这三个规则.

5 结束语

随着信息科技的蓬勃发展, 数据已经成为一种无形的、全新的资源, 使人们对数据的质量要求越来越高. 然而大数据时代下, 数据种类繁多且数据量大的特征, 使得数据质量评估的难度加大. 本文提出一套整体的数据质量评估方法, 其中规则库和评估模板的应用使得该评估方法具有通用性, 针对不同的业务数据设计不同的规则存放于规则库中, 再使用规则评估模板针对某一具体的数据进行评估, 并且考虑数据质量各个方面的评估维度. 利用 echarts.js 实现评估结果的可视化, 使评估结果更加清晰、明了, 更有利于找出影响数据质量的因素, 对数据进行更改、恢复、清洗.

进一步的研究工作包括实时数据质量的评估、探索逻辑规则校验、多维评估延伸、统计分布验证等质量评价方法.

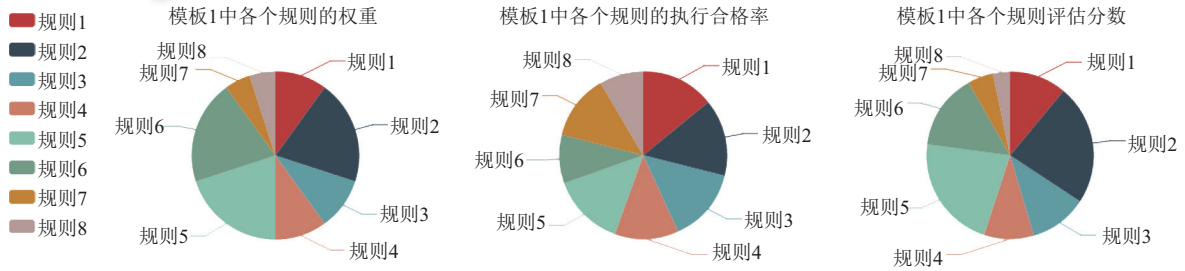

图 6 评估结果统计

参考文献

1 袁满, 张雪. 一种基于规则的数据质量评价模型. 计算机技术与发展, 2013, 23(3): 81–84, 89.
2 刘伟. 基于元数据的数据质量控制与评估模型研究[硕士学位论文]. 大庆: 东北石油大学, 2011.
3 邓丽华. 浅析统计数据质量评估方法. 中国市场, 2013, (38): 85–86. [doi: 10.3969/j.issn.1005-6432.2013.38.034]
4 祝君仪. 大数据时代背景下统计数据质量的评估方法及适用性分析. 中国市场, 2015, (29): 41–42.
5 陈苏, 柏文阳, 徐洁磐. 一种新的数据质量模型的研究. 计算机应用研究, 2005, 22(7): 48–50.
6 管尊友, 冯建华. 一个可扩展的数据质量元模型. 计算机工程, 2005, 31(8): 74–76, 226.
7 史峰. 基于规则库的数据质量分析. 武汉职业技术学院学报, 2010, 9(3): 79–83.
8 王慧锋, 段磊, 胡斌, 等. 带间隔约束的序列数据质量评价算法设计. 计算机科学与探索, 2015, 9(10): 1180–1194.
9 周青, 张乐坚, 李峰, 等. 自动站实时数据质量分析及质控算法改进. 气象科技, 2015, 43(5): 814–822.
10 朱巧玉. 基于质量规则矿政属性数据评价. 黑龙江工程学院学报, 2014, 28(6): 13–16.
11 宗威, 吴锋. 大数据时代下数据质量的挑战. 西安交通大学学报 (社会科学版), 2013, 33(5): 38–43.
12 刘军华. 大数据视野下统计数据质量演变的信息回归、分布与趋势. 统计与信息论坛, 2015, 30(9): 7–11.
13 陈卫东, 张维明. 属性粒度数据质量模型及其评价指标研究. 计算机科学, 2010, 37(5): 139–142.
14 杨青云, 赵培英, 杨冬青, 等. 数据质量评估方法研究. 计算机工程与应用, 2004, 40(9): 3–4, 15.
15 王树西, 白硕. 事实库、规则库的一体化全文索引算法. 计算机科学, 2006, 33(4): 174–176.
16 张胜. 数据质量评价指标和评价方法浅析. 科技信息, 2014, (2): 259.