大数据下的数据质量评价指标构建实践

刘金晶[1],王 梅[2]
(1. 北京锐安科技有限公司 大数据分析部,北京 100192;
2. 北京锐安科技有限公司 研究院,北京 100192)

摘 要:大数据下的数据特点决定了对其数据价值的萃取犹如沙里淘金,需要进行大量的数据处理、分析和挖掘才能获得其背后的价值。而进行数据分析与挖掘,且获得真正有价值的信息与知识,良好的数据质量得到保障是前提。因此,数据质量的量化评估成为这个过程中很重要的一环。通过综合国内外对数据质量评价体系的研究成果,结合所在行业和大数据系统的特点,提出了一个评价指标的框架,不仅包含数据本身的质量,而且包括数据处理过程与数据效能的质量。全面对大数据处理平台下的数据质量进行量化评估,是对数据质量评价体系在大数据生产系统进行实践的第一步,为大数据下的数据治理提供了新的研究和实践经验,也为后续进行持续的数据改进、数据治理、数据价值到信息价值的提炼提供借鉴。

关键词:大数据;数据质量;评价指标;量化评估
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2019)10-0046-05
doi:10.3969/j.issn.1673-629X.2019.10.010

Practice of Data Quality Evaluating Index Construction under Big Data

LIU Jin-jing[1], WANG Mei[2]
(1. Department of Big Data Analysis, Beijing Rui An Technology Co., Ltd., Beijing 100192, China;
2. Institute of Beijing Rui An Technology Co., Ltd., Beijing 100192, China)

Abstract: The characteristics of big data determine that the extraction of the underlying value in big data is like panning for gold in sand, which requires a lot of data processing, analysis and mining to obtain the value behind it. For data analysis and mining, and to obtain truly valuable information and knowledge, guaranteeing great data quality is premise. Therefore, quantitative assessment of data quality has become an important part of this process. By integrating the research results of data quality evaluation system at home and abroad, combined with the characteristics of the industry and big data system, a framework for evaluating indicators is put forward, including not only the quality of the data itself, but also the quality of the data processing and data performance. Comprehensive quantitative evaluation of data quality under the big data processing platform is the first step in the practice of the data quality evaluation system of the actual big data production system, which provides new research and practical experience for data governance under big data, and also provides reference for continuous data improvement, data governance, data value and information value extraction.

Key words: big data; data quality; evaluating index; quantitative assessment

0 引 言

人类历史上从未有哪个时代像现在一样,任何活动都带来了大量的数据[1],完全不受时间、地点的限制。由此进入的大数据时代,数据成为了一种基础资源、战略资源[2],已然在业界形成了共识。但大数据产生的背景,使得大数据有其自身的典型特点,其价值不是显性的可以被直接获取使用的,而是需要像沙里淘金一样,通过建立适当的分析模型,并运用相应的技术手段进行有效的深加工和挖掘分析[3],发现隐含在大数据中的价值并加以利用,进而指导决策,才能将大数据的真正效用发挥到极致。

而进行数据分析和挖掘,数据质量则是一个至关重要的因素。根据“垃圾进,垃圾出(garbage in, garbage out)”[4]的原理,如果数据质量存在问题,系统

收稿日期:2018-11-15 **修回日期**:2019-03-13 **网络出版时间**:2019-06-26
基金项目:国家高技术研究发展计划(863 计划)课题(2011AA010502)
作者简介:刘金晶(1984-),女,硕士,工程师,通信作者,研究方向为数据挖掘、元数据管理、数据质量管理、数据治理等;王 梅,高级研究员,研究方向为大数据生命周期管理、数据挖掘、数据治理等。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190626.0829.026.html

运算、分析的结果极有可能是错误的[5],甚至与真实世界南辕北辙。因此,数据质量是发挥大数据价值的必要条件。

1　数据质量评估体系的研究现状

数据质量是一种通过测量和改善数据综合特征来优化数据价值的过程[6]。提高和保障数据质量,首先要建立的是数据质量评估体系。虽然数据质量不是一个新事物,但在大数据环境下的数据质量相比传统行业,面临的问题更加突出和急迫[2]。数据质量的保障,需要多环节、全方位的一套治理体系。在这些环节中,数据质量评估是提高数据质量的基础和必要前提[7]。

对于数据质量评估,虽然业界已进行了大量的学术研究和应用探索,但在目前还没有完全统一的定义和体系化的标准。

文献[8-10]从不同的方面提出了数据质量的评估方法,文献[11]介绍了数据质量的评估过程,文献[4]介绍了统计学界的一些公认指标,主要包括准确性、时效性、相关性、客观性、可衔接性、完整性、可理解性、透明性、可操作性、可取性、可解释性、效益性、安全性等,以及 UN 下属的经济委员会提出的包含 11 个指标变量的数据质量评价体系。在国内,蔡莉等主导的研究中提出了包含 5 个指标的大数据质量评价体系,它们分别是可获得性、可靠性、可用性、相关性、可表达性。文献[12]则结合所在的石油行业的需求提出了完整性、准确性、一致性、深度性、及时性、冗余性等 6 个关键特性。

可以看到,众多的研究都集中在对数据质量关键特性的评价指标定义上面。而关于如何将概念定义落实到量化的、可采集、可计算的评价指标的行业实践经验,均较少涉及。

笔者通过参考这些公认的质量评价指标,结合行业领域、数据类型、应用目的、信息系统使用的技术等多方面的相关影响因素,构建了一套在行业领域内适用的质量评价指标并用于实践,取得了一定的效果。

2　大数据质量评价指标构建实践

构建一套质量评价体系,首先需要对质量评价的模型进行确定。笔者参考了国内外的众多研究成果,评估了质量评估模型与所在行业、信息系统特点的相关程度之后,最终以文献[7]所提出的模型作为基础,结合数据采集、数据集成、数据整合与清洗、数据处理与加工、数据持久化等数据流转环节的特点,建立了一个简单且有效可行的数据质量评估指标框架。

2.1　数据质量评价模型与评价方式

文献[7]提出数据质量评价体系需至少包含以下两个方面的基本评估指标:

(1)数据对用户必须是可信的,其中包括精确性、完整性、一致性、有效性、唯一性等指标。这些指标的具体含义如下:

精确性:描述数据是否与其对应的客观实体的特征相一致。

完整性:描述数据是否存在缺失记录或缺失字段。

一致性:描述同一实体的同一属性的值在不同的系统或数据集中是否一致。

有效性:描述数据是否满足用户定义的条件或在一定的值域范围内。

唯一性:描述数据是否存在重复记录。

(2)数据对用户必须是可用的,其中包括时间性、稳定性等指标。这些指标的具体含义:

时间性:描述数据是当前数据还是历史数据。

稳定性:描述数据是否是稳定的,是否在其有效期内。

文献[8,13]总结了数据质量的评价方法,有以下几种方式:

(1)简单比率法:指期望的结果(E)占总值(T)的比率即 E/T,反映数据质量某些方面的好坏程度。当结果等于或接近于 1 时,表明数据质量情况好,否则质量情况差。该计算方式还能用来进行纵向比较,反映数据质量的改进情况。

(2)最小/最大值法: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小值或最大值。最小值和最大值分别代表了最保守和最激进的评价方法,一般适用于比较复杂的度量体系。

(3)加权平均法:对于复杂的多指标的评价,如果评价者对每个指标在总体评价中的重要程度很容易量化,则可以使用加权平均法。为每个单独的指标设置权重 λ_i,取值在 0 和 1 之间,且 λ_i 的和等于 1,即 $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$,则最终的总体评价指标为 $X = \lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_n X_n$, X_i 代表不同的基础指标。

根据实际情况,笔者扩展了最小/最大值法,增加了平均值的评估方法。如果说最小值和最大值分别代表了最保守和最激进的评估方法,那么对这些指标求平均值,相对而言则是一个更稳妥、适中的评价方式。

2.2　指标构建与实践

笔者综合考虑了所在公安大数据领域的大数据处理平台的特点以及数据处理流程、数据来源、用户使用数据以及数据模型等多方面影响因素,分别从数据自身的质量、数据处理过程的质量和数据效能三个方面,提出了评价指标框架,对全生命周期的数据质量进行评估和度量。

根据指标是否具有对所有处理环节的数据质量进行评价的共通性,分为通用指标和特定指标两大类。通用指标指的是与数据的具体形态、处理的具体环节无关的评价指标,评价的是数据和数据处理过程本身的质量。而特定指标则和数据的形态格式与数据处理的具体环节紧密相关,在不同的实际环境中,会根据使用的数据接口、数据处理技术和功能的不同有不同的评价指标。

根据能否通过采集到的信息直接进行计算,又可以分为基础指标和综合指标两大类。基础指标是通过采集信息就可以通过简单的计算得出,而综合指标则需要结合对数据的使用需求、数据重要程度考量、指标计算的可行性等多方面因素之后形成规则,依据规则再进行计算得到的评价指标,一般使用的评价方法为最小/最大值法、平均值法或加权平均法。

最后,构建质量评价指标框架,如图 1 所示。

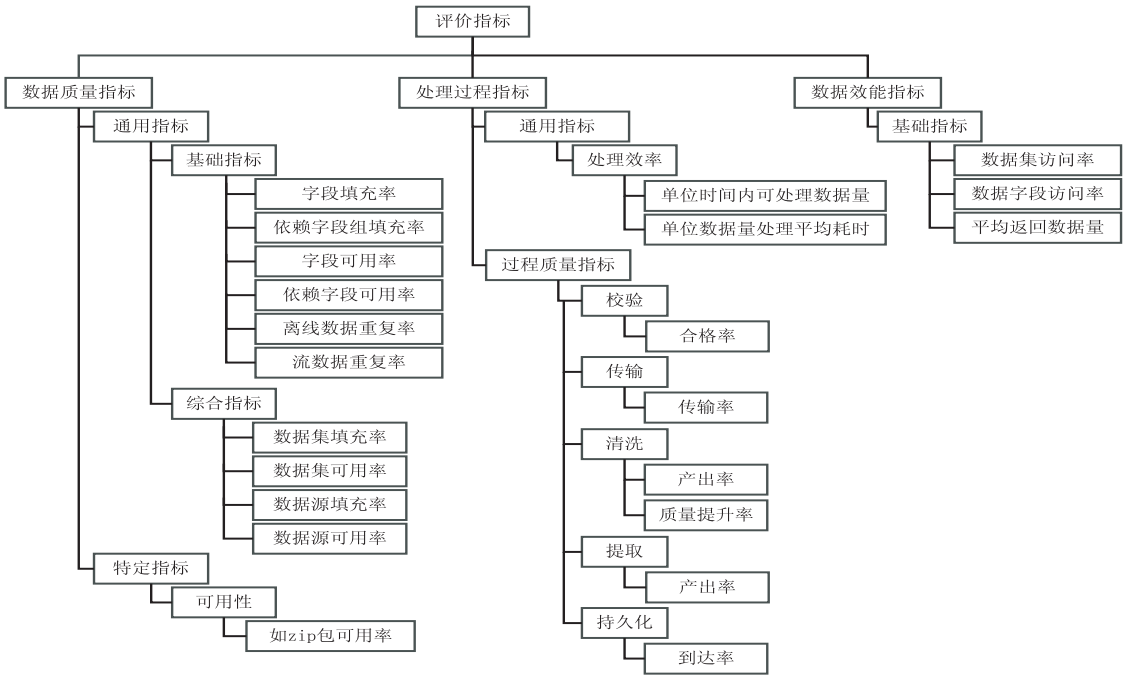

图 1 数据质量评价指标结构

2.2.1 评价数据质量的指标

结合文献[7]提出的评估指标和现有系统的数据处理与使用的特性,最终选取了以下几类指标:

完整性:数据的记录和信息是否完整,是否存在缺失的情况;

可用性:数据对使用者来说是否是可用的、有效的,合并了一致性、有效性和准确性;

重复性:根据指定的判重规则计算重复率。

详细的评价指标与评价方法如表 1 所示。

表 1 数据质量评价指标

是否通用	是否综合	质量维度	指标名称	描述	评价方法
通用指标	基础指标	完整性	字段填充率	字段非空的比例	简单比率法
			依赖字段填充率	存在依赖关系的字段共同非空的比例	简单比率法
		可用性	字段可用率	字段值有效、可用的比例	简单比率法
			依赖字段可用率	存在依赖关系的字段同时非空的比例	简单比率法
		重复性	离线数据重复率	全量数据中存在重复数据的比例	简单比率法
			实时数据重复率	实时流中存在重复数据的比例	简单比率法
	综合指标	完整性	数据集填充率	数据集由多个字段组成,根据字段的填充率计算整个数据集的填充率	加权平均法,不同的字段重要性不同,权重不同
			数据源填充率	数据源由多个数据集组成,根据数据集的填充率计算整个数据源的填充率	加权平均法,不同的数据集重要性不同,权重不同
		可用性	数据集可用率	数据集由多个字段组成,根据字段的可用率计算整个数据集的可用率	加权平均法,不同的字段重要性不同,权重不同
			数据源可用率	数据源由多个数据集组成,根据数据集的可用率计算整个数据源的可用率	加权平均法,不同的数据集重要性不同,权重不同

续表 1

是否通用	是否综合	质量维度	指标名称	描述	评价方法
特定指标		可用性	如 zip 包可用率	某个环节使用的输入数据形态为 zip 包,评估输入 zip 包中可用的 zip 包占总体数量的比例	简单比率法

其中,zip 包可用率就是一个典型的特定指标。某个数据流转环节中,定义的数据接口是遵循行业规范对数据文件和数据描述文件进行压缩后的 zip 包,其中数据文件的命名、数据分隔符、数据描述文件的格式、里面包含的数据项内容、数据项的值等都需要遵循相应的行业标准规范。如果输入的数据不符合定义的格式和要求,那么数据将无法被解析,等同于无效数据。因此,在这个环节,zip 数据包的可用率就是一个非常重要且必要的特定监测指标。

2.2.2　评价数据处理过程的指标

数据处理的基本过程一般包括校验、传输、清洗、提取、持久化等类型。每一个处理过程都有可能带来数据处理前后的数量变化、质量变化。不同的数据处理过程不同,衡量其处理质量的指标也存在差别。

同时,质量高的处理过程应该在处理时效有保证的前提下,提升输出数据相对输入数据的质量。因此,处理过程的质量也不能孤立的使用过程指标就能判定,还需要配合处理前后的数据质量才进行综合判定。

因此,对数据处理过程[14]的质量可以提出以下评价指标,如表 2 所示。

表 2　数据处理过程评价指标

指标分类	处理环节	指标名称	指标描述	评价方法
通用指标	处理效率	平均处理速度	单位时间可处理的数据量	平均值法,一段时间的处理数据量除以耗时
		平均处理耗时	单位数据量的处理平均耗时	平均值法,多次单位数据量的处理时间除以次数
不同环节的质量指标	校验	合格率	校验后符合数据定义和规则的数据与总体数据的比例	简单比率法
	传输	传输率	传输后的数据占传输前数据里的比例	简单比率法
	清洗	产出率	清洗后的数据量占清洗前数据总量的比例	简单比率法
		质量提升率	清洗后的数据质量相对清洗前的数据质量的提升百分比	简单比率法
	提取	产出率	提取后的数据量占来源数据总量的比例	简单比率法
	持久化	到达率	持久化成功的数据量占待持久化的数据总量的比例	简单比率法

图中不同环节的质量指标虽然评价的处理环节不同,但却也与具体的处理技术和细节无关。因此,如果在实际系统中,对监控更细节的处理质量存在需求,则可以根据实际情况添加更具体的评价指标。

2.2.3　评估数据效能的指标

数据最终需要为应用、为终端用户所用才能展现价值,其质量的好坏才有意义。前面数据采集的再好、质量保证的再高、处理的再快,如果用户不用或极少使用这些数据,或者使用的效果不如用户所期望,那么这些数据的价值也不算得到了体现,需要根据用户的需求进行调整。

考虑到应用系统对数据的访问、使用情况能在一定程度上反映数据的利用价值,因此提出如表 3 所示的指标,作为评估数据最终价值也即数据效能的指标。同时也可以作为数据的使用情况反馈,为数据分析和数据处理的优化、调整提供参考依据。

表 3　数据效能指标

指标名称	指标描述	评价方法
数据字段访问率	字段在所有被访问字段中的比率	简单比率法
数据集访问率	数据集在所有被访问数据集中的比率	简单比率法
平均返回数据量	应用访问数据集时平均返回的数据量	平均值法,一段时间内应用获取该数据集时的平均返回数据量

简单比率法按其定义,其指标反映的是相对期望值(一般为 1)的符合程度,其值越是接近 1,表明质量越高,否则反之。但对于评价数据的使用效能而言,数据字段和数据集的访问率是不可能以 1 为期望值的。

所以这两个指标更多用来做排名,查看访问率排名靠前的数据集和字段是否如需求所期望的,如果不是,那么就可以指导设计人员或开发人员进行相应的调整。同时,这个指标也可以用来做纵向对比,即调整之后的访问率相比调整之前的访问率,是不是有相应的提升,提升的幅度是否达到了调整的期望。

而平均返回数据量,也是根据用户期望的需求不同而不同,因此,没有很统一的标准,需要根据实际情况制定参考标准。

2.3 实践与应用效果

基于上述指标框架,笔者所在单位开发了一套数据 KPI 监控的系统,数据质量的指标已经完全在系统中实现,数据处理环节的指标有部分已经实现,目前已用于对大数据平台的整个数据流的质量进行监控,取得了良好的效果。在没有进行质量监控之前,问题数据的发现往往都是在后端,通过倒推检查才能找到问题的源头,问题发现的晚,解决耗时长。而通过这套质量 KPI 系统,每一个环节的质量数据即时产生,即时评估,不符合质量指标及时告警,及时解决,大大提升了问题暴露的速度和解决效率,给系统运维人员和用户带来了很大的便利,也提升了整个平台的数据质量。

而数据效能指标,涉及到用户对数据的价值评估,根据数据-信息-知识-智慧[15]的金字塔体系,按照文献[16]的定义,归属于信息质量的范畴,目前业界大部分工作也还只处在研究阶段,笔者所做的尝试就是提出了一些可以进行采集与计算的量化指标,将概念上的信息质量变成了可以进行评价比较的数据,但实际应用效果还需要进行不断的调整与实践验证。

3 结束语

通过综合国内外对数据质量评价体系的研究成果,结合所在行业和大数据系统的特点,提出了一种评价指标框架,并在实际系统中进行了实践应用,取得了良好的效果,为当前大数据处理平台下的数据治理提供了重要的研究和实践经验。通过以上实践,实现了对现有系统的数据质量和数据处理过程的质量进行量化评估,是提升数据质量进而挖掘数据价值的第一步,让数据质量从理论研究到实践应用往前多走了一步,为后续进行持续的数据改进、数据治理、从数据价值到信息价值的提炼打下了基础。

参考文献:

[1] 孟小峰,慈 祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展,2013,50(1):146-169.

[2] 宋 敏,覃 正. 国外数据质量管理研究综述[J]. 情报杂志,2007,26(2):7-9.

[3] 宗 威,吴 锋. 大数据时代下数据质量的挑战[J]. 西安交通大学学报:社会科学版,2013,33(5):38-43.

[4] 马一鸣. 政府大数据质量评价体系构建研究[D]. 长春:吉林大学,2016.

[5] 刘金晶,曹文洁. 大数据环境下的数据质量管理策略[J]. 软件导刊,2017,16(3):176-179.

[6] 方幼林,杨冬青,唐世渭,等. 数据仓库中数据质量控制研究[J]. 计算机工程与应用,2003,39(13):1-4.

[7] 杨青云,赵培英,杨冬青,等. 数据质量评估方法研究[J]. 计算机工程与应用,2004,40(9):3-4.

[8] PIPINO L L,LEE Y W,WANG R Y. Data quality assessment[J]. Communications of the ACM,2002,45(4):211-218.

[9] WAND Y,WANG R Y. Anchoring data quality dimensions in ontological foundations[J]. Communications of the ACM,1996,39(11):86-95.

[10] WANG R Y,STOREY V C,FIRTH C P. A framework for analysis of data quality research[J]. IEEE Transactions on Knowledge and Data Engineering,1995,7(4):623-640.

[11] Firstlogic. Data quality assessment:a methodology for success[R]. [s. l.]:Firstlogic,2003.

[12] 刘学霞,曾昭虎,周之光. 基于元数据的数据质量分析评估系统模型及实现[C]//2005 石油数据管理与应用国际学术研讨会论文集(2005 石油数据管理与应用国际学术研讨会). 大庆:大庆油田,2009:36-41.

[13] 张 胜. 数据质量评价指标和评价方法浅析[J]. 科技信息,2014(2):259.

[14] 孔 钦,叶长青,孙 赟. 大数据下数据预处理方法研究[J]. 计算机技术与发展,2018, 28(5):1-4.

[15] BENSON P R. ISO 8000 data quality[EB/OL]. (2009-10-01). https://www. ewsolutions. com/iso-8000-data-quality/.

[16] 宋立荣,李思经. 从数据质量到信息质量的发展[J]. 情报科学,2010,28(2):182-186.