一种基于规则的数据质量评价模型

袁 满,张 雪
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:在对国际与国内关于数据质量定义及评价方面研究成果的分析发现,到目前为止,对这些问题的研究仍然存在许多缺陷,如数据质量的定义不统一,数据质量的评价指标描述不全面,数据质量评价体系不系统等。针对这些问题,提出了以七项指标为基础的全面的数据质量定义,并定义了基于七项指标的十五类数据质量约束规则,给出了它们之间的关系。定义了五元组来形式化描述数据质量评价指标算法,并以完整性评价指标为例详细描述了该算法及其实现过程。为使这些指标与约束规则精准描述及存储,最后基于元数据构建了系列支撑元模型。上述研究成果已在大型企业数据中心数据质量检测与评价中得到了初步应用,并且效果良好。

关键词:数据质量;数据质量评价指标;数据质量约束规则;数据质量评价指标算法;元数据;数据质量评价模型
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2013)03-0081-04
doi:10.3969/j.issn.1673-629X.2013.03.021

A Data Quality Assessment Model Based on Rules

YUAN Man, ZHANG Xue
(School of Computer Science & Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: From the finding of the research on the definition and assessment of data quality abroad and at home, research on these issues still exist many defects, such as the non-uniform definition of data quality, the incomprehensive description of data quality assessment, the unsystematic system of data quality assessment, etc. Aiming at these issues, give a comprehensive definition of data quality from the seven kinds of assessment indicators. Then define fifteen kinds of constraint rules of data quality based on the seven kinds of assessment indicators. And describe the relationship between them. Quintuple form is defined to formally describe the algorithm of the data quality assessment indicator. And the integrity assessment indicator is taken as an example to specifically describe the algorithm and its implementation. As for the accurate description and store of these indicators and constraint rules, a series of supporting meta-models are constructed based on meta-data. Research above has preliminarily been applied in the data quality testing and assessment of data center in large enterprises with good results.

Key words: data quality; data quality assessment indicator; data quality constraint rule; algorithm of data quality assessment indicator; meta-data; data quality assessment model

0 引 言

对于一个企业,只有高质量的数据才具有实际的应用价值,如何保证企业数据的质量问题一直是人们研究的热点。随着信息化水平的提高,数据在企业决策中的重要性也已经突显出来,低质量的数据将导致业务流程阻塞、成本增加、甚至决策困难等一系列的严重问题,直接影响着企业的生存和效益[1]。某油田数据中心存放了上千万条的数据,并且还以每天数万条的速度增加,若要使这些海量数据在生产管理、科学研究、企业决策中发挥应有的作用,使其真正为企业服务,那么对数据质量的管理就成为了企业发展中至关重要的一环。

在处理数据质量问题的各个过程中,数据质量评价是提高数据质量的基础和关键,它能用科学、客观的方法评价出数据的真实情况,使用户清晰、明确地了解到当前数据的状态,以便采取适当的手段提高数据的质量。文献[2]针对数据对用户是否可信与可用两个方面阐述了数据质量的评价指标,并提出了一种基于该评价指标的数据质量评价方法,但是它并没有给出具体评价指标所对应的规则集合。文献[3]主要研究了数据质量评价指标在选择、投影等运算中出现的问

收稿日期:2012-06-10;**修回日期**:2012-09-17
基金项目:黑龙江省教育基金项目(11541008)
作者简介:袁 满(1965-),男,黑龙江农安人,博士,教授,硕士生导师,CCF 高级会员,主要研究方向为信息集成、高级数据管理等;张雪(1986-),女,河北昌黎人,硕士研究生,主要研究方向为软件工程与集成技术。

题,并给出了相应的质量评价模型,但是该文献对数据质量评价指标的定义并不全面,只针对正确性和完整性进行了阐述。文献[4]给出了一个属性粒度的质量评价模型,定义了正确性评价指标,但是该模型只是针对正确性评价指标而建立,有很大的局限性。文献[5]给出了正确性的评价指标定义,及其数据质量评价算法,而并没有给出一个相应的数据质量评价模型。文献[6,7]提出的数据质量的评价指标较为全面,但是并没有说明每个指标的具体计算过程。文献[8]就用户使用的角度探讨了数据质量评价的方法,但是并没有给出定量的描述过程。文献[9]则认为在特定的数据使用环境中研究数据质量维度才是有意义的。可以看到,针对不同的领域,目前国内外对于数据质量评价问题的研究仍然存在很多不足之处。由于在进行某个具体的数据质量评估时,要根据具体的数据质量评估需求对数据质量评估指标进行相应的取舍[10]。因此,文中给出了一种基于规则的数据质量指标定义与算法,并建立了基于这些评价指标和算法的数据质量评价体系模型。

1 数据质量相关知识

由于数据可能出现的质量问题多种多样,数据质量评价以需求为导向[11],用数据质量评价指标描述这些问题成为定义数据质量的一个重要依据。由于传统文献在定义评价指标方面的不足,文中提出了七项数据质量评价指标 AI(Assessment Indicator),具体定义如下:

定义 1 完整性(Integrity):体现为实际的数据与所期望的数据在数量上的满足程度。

定义 2 历史性(Historic):也称为深度性,体现为该数据在其产生之日起到消亡时的固定周期内的时间满足程度。

定义 3 准确性(Accuracy):体现为实际的数据与所期望的数据间的一致程度。

定义 4 及时性(Timeliness):体现为数据在固定的或规定的时间内的完成程度。

定义 5 冗余性(Redundancy):体现为数据集内各个数据间的重复程度。

定义 6 一致性(Consistency):体现为数据与数据之间在某一特定条件下满足某一相同的条件或状态。

定义 7 关联性(Relevance):体现为不同或相同数据集之间数据的依赖程度。

以上七项数据质量评价指标所对应的十五类数据质量约束规则 CR(Constraint Rule)定义如下:

定义 8 非空约束规则:描述为该数据项上的数据不允许出现空值。

定义 9 连续性约束规则:描述为该数据项上的数据必须满足一定的连续取值。

定义 10 完整性约束规则:描述为该数据项上的数据必须满足完整性要求。

定义 11 历史性约束规则:描述为该数据项上的数据必须满足历史性要求。

定义 12 代码约束规则:描述为该数据项上的数据必须满足特定的代码规范。

定义 13 词法约束规则:描述为该数据项上的数据必须满足特定的词法要求。

定义 14 值域约束规则:描述为该数据项上的数据必须满足特定的取值规则。

定义 15 逻辑依赖约束规则:描述为该数据项上的数据必须与其它数据项上的数据满足某种逻辑关系(如大于、小于等)。

定义 16 及时性约束规则:描述为该数据项上的数据必须满足及时性要求。

定义 17 冗余性约束规则:描述为该数据项上的数据必须满足冗余性要求。

定义 18 等值函数依赖约束规则:描述为该数据项上的数据必须通过同一表中其它数据项上的数据计算得来。

定义 19 等值一致性依赖约束规则:描述为该数据项上的数据必须通过不同表中其它数据项上的数据计算得来。

定义 20 存在一致性依赖约束规则:描述为该数据项上的数据必须出现在其它表中的某一数据项中。

定义 21 逻辑一致性依赖约束规则:描述为该数据项上的数据与其它表中的数据项满足某种逻辑关系(如大于、小于等)。

定义 22 关联性约束规则:描述为该数据项上的数据必须满足关联性要求。

2 系统模型

2.1 评价体系结构模型

由于数据质量是一个相对的概念,在不同的时期、不同领域,数据质量有着不同的定义和评价标准[12],因此,国内外对于数据质量的评价系统很少,也没有形成一个完整的评价体系。数据质量问题的类型很多,因此要建立起一套完整的、全方面的数据质量检查和评价体系,离不开数据质量约束规则库的支持[13]。因此,文中提出了一种以规则元数据为基础的数据质量评价体系。由于以规则元数据为基础的质量评价模型能良好地涵盖数据质量的种类繁多、形式繁杂等问题,因此,基于规则元数据而建立起的一套数据质量评价体系拥有非常强大的完整性、可扩展性和灵活性等。

文中将该模型划分为数据层、业务逻辑层和用户层三部分,如图 1 所示。

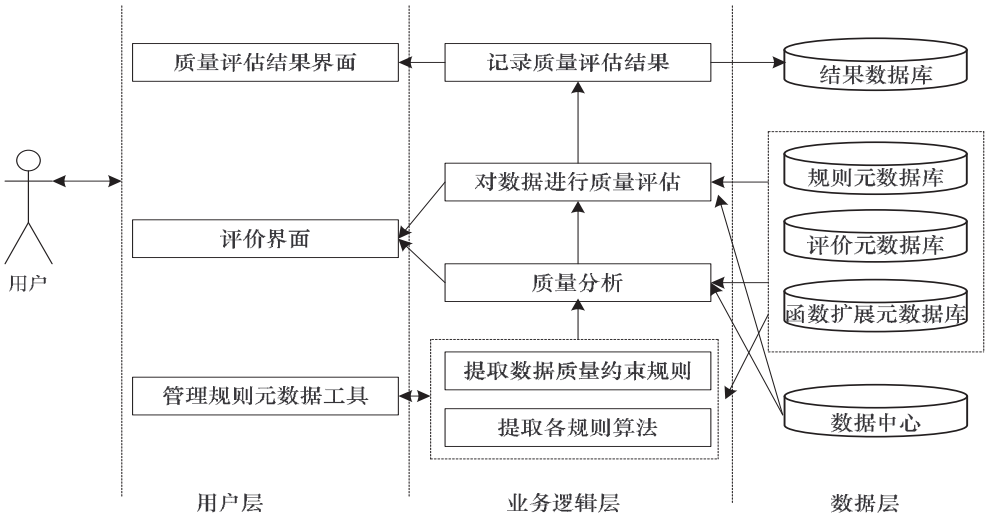

图 1　系统整体构架

首先,数据层上各个元数据模型为质量评价系统提供了良好的数据基础及模型支撑。其次,业务逻辑层通过提取、分析、计算、评价等过程,实现对数据质量的评价。最后,用户层为用户提供良好的界面接口。

2.2　数据质量支撑元数据模型构建

元数据定义为描述数据的数据,组织良好的元数据在实现数据的存取、转换分析、管理过程中,成为系统元模型可扩展的关键[14]。因此,构建基于规则元数据的数据质量评价体系,包括五个支撑元模型,分别是数据字典元模型、约束规则元模型、函数扩展元模型、评价元模型以及评价结果元模型。

2.2.1　数据字典元模型

数据字典元模型中存储了描述实体的元模型数据,包括描述数据库信息的数据源、描述数据表所属专业信息的专业、描述数据源中表信息的数据表以及描述表中字段信息的数据项,它们之间的关系定义如下:

DATASET{DSID, DSNAME, URL, UNAME, PWD, TSPACE, DSDESC, CREMAN, CREDATE}

SPECIAL{SID, SNAME, SDES}

TABLE{TID, SID, DSID, TNAME, SNAME, TDES, CREMAN, CREDATE }

DATAITEM{DIID, TID, DINAME, DIDES, DSTYPE, PRINTYPE, ISCANDIKEY, ISREL, RELFIE, PARID}

2.2.2　约束规则元模型

约束规则元模型中存储了所有的数据质量约束规则,以及规则与实体数据间的关系,它们之间的关系定义如下:

CON_INFOR{DBID, TID, FIEID, CRID, FILCAND}

LOC_CON_INFOR{DBID, TID, FIEID, CRID, DCRNUM}

INTE_CON_INFOR{DBID, TID, FIEID, CRID, INTEINFOR}

在进行数据质量评价时,通过函数扩展元模型在规则元模型中抽取相应的数据质量约束规则,然后计算对应的数据质量各个指标的信息。规则元模型中引用的数据质量对象信息都来自数据字典元模型。

2.2.3　函数扩展元模型

函数扩展元模型主要包括两部分扩展元模型:质量指标的扩展元模型和约束规则的扩展元模型,它们之间的关系定义如下:

ELE_CON_MAPPING{ASSECATECODE, CATEID}

ASSESS_ASSINGN_INFOR{ASSINGNID, ASSESTADATE}

ASSESS_ASSINGN_TAB_INFOR {ASSINGNID, ASSECATECODE, TID, REMARK}

ELE_CATE{ASSECATECODE, ASSECATENAME}

通过对这些函数扩展元模型的定义,为以后系统的扩展性提供了必要的元数据支持。

2.2.4　评价元模型

评价元模型中存储了用于进行数据质量评价的各个函数信息,数据质量指标、数据质量约束规则和函数之间的映射关系以及评价的各个流程信息,它们之间的关系定义如下:

ANAL_FUN_INFOR{FUNID, FUNINFOR, NSPACODE}

ASSESS_CATE_TAB{ASSECATECODE, FUNID, ASSECATENAME }

CON_RULE_ASSINGN _TAB{CRID, FUNID, CRCODE, CRNAME}

评价元模型是评价过程的基础,在数据质量每次评价过程中,为相应的操作调用对应的处理函数、分析对应的数据质量约束规则并实现对应数据质量指标的评价。

2.2.5　结果元模型

结果元模型中存储了包括数据质量评价指标信息和数据质量评价结果信息两部分,它们之间的关系定义如下:

ASSESS_ASSINGN_INFOR{ASSINGNID, ASSESTADATE}

ASSESS_PRO_INFOR{ASSIGNID, GROID, NUM, ASSECATECODE ,CRID, PROINFOR }

ASSESS_ASSINGN_TAB_INFOR {ASSINGNID, ASSECATECODE, TID, REMARK}

CON_LOG_INFOR{ASSINGNID, TID, ASSECATECODE, CRID, PRINKEYINFOR}

结果元模型为用户提供一个良好的系统展示结果,在每次评价过程中,结果元模型都会记录相应数据质量指标信息与记录评价结果的日志信息,包括评价过程中产生的过程信息、违反约束规则的数据信息、运行错误的数据信息等。

2.3 基于系列元模型的数据质量评价算法

依据上面定义的评价指标算法,以五个元模型为基础,采用关系代数语言对数据质量评价算法的流程描述如下:

步骤 1:$R=\pi(\sigma(INTE\_CON\_INFOR)\bowtie\sigma(CON\_RULE\_ASSINGN))\bowtie\sigma(CON\_RULE\_META\_MAPPING)$;//首先以完整性为例,定义完整性约束规则以及该约束规则与元模型间的映射关系。

步骤 2:$I=\pi_{FILCAND,FIEID,INTERDATE}(\sigma(CON\_INFOR)\bowtie\sigma(INTE\_CON\_INFOR)\bowtie\sigma(R))$;//利用步骤 1 的结果找到与完整性有关的所有表间的关系,找到过滤条件信息。

步骤 3:$P=\pi(\sigma(DATASET\cap SPECIAL\cap TABLE\cap DATAITEM)\bowtie\sigma(I))$//通过步骤 2 得到的表间关系,遍历整个实体数据库,得到不满足条件的问题数据。

步骤 4:$A=\pi(\sigma(\sigma(ANAL\_FUN\_INFOR)\bowtie\sigma(ASSESS\_CATE\_TAB))\bowtie\sigma(CON\_RULE\_ASSINGN\_TAB))\bowtie\sigma(P))$//在评价元数据库中获得评价的所需要的函数信息、约束规则信息以及之间的关系和评价的流程,结合步骤 3 中所得到的问题数据,通过公式,计算评价指标的结果,并将其记录到评价结果数据库中。

步骤 5:$\pi(\sigma(ASSESS\_PRO\_INFOR)\cup\sigma(CON\_LOG\_INFOR)\cup\sigma(CON\_ERR\_INFOR)\cup\sigma(INTE\_ASSESS\_INFOR))$//查看评价结果信息,包括评价进程信息、约束规则日志信息、约束规则错误信息和完整性评价信息,最后利用图形化界面将结果展示出来。

3 应用案例

井下作业是油田实现原油增产的主要措施,涉及的业务比较庞杂,具体包括开发井压裂、探井压裂、修井、酸化、补孔完井、堵水、生产管理、特种工艺作业、机械制造以及机加机修、环保项目等。在进行井下作业数据中心构建过程中,为了保障进入到数据中心的数据质量,就必须结合井下作业的具体业务规则,对用于存储这些业务信息的数据模型中的各种属性进行数据质量规则的定义。规则约束定义完成之后,就可以基于这些规则按照不同的质量指标评价算法对用户所感兴趣的数据集进行质量评价。在该应用案例中,对数据中心 368 张表中的 5341 个字段进行了各种数据质量约束规则的逐一定义。以完整性约束规则为例,对钻井地质信息(CD_WELLBORE_T)的完整性约束规则进行定义之后,使用该表中数据项井 ID 的不重复数据个数作为数据表的完整性键值定义信息,定义完毕之后,对其进行完整性数据质量指标的评价。

4 结束语

随着领域数据中心的建立,实现了领域数据的集成存储、集中管理,为领域应用的信息共享提供了保障。但先决条件是这些数据中心中的数据必须是符合高质量要求的,所以数据质量问题是领域构建数据中心必定会面临的问题。

文中针对国际与国内在这方面的研究成果,提出了基于规则的数据质量评价系统,在其中定义了七项数据质量评价指标和十五类数据质量约束规则,并构建了实现功能的元数据模型。这些成果在油田井下作业数据中心数据质量管理中得到了应用的检验,并且应用效果良好。因此,这些成果对于数据中心中的数据质量管理具有重要的参考与借鉴意义。

参考文献:

[1] 余宇峰,万定生. Benford 法则在水文数据质量挖掘中的应用研究[J]. 微电子学与计算机,2011,28(8):180-186.

[2] 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(9):3-4.

[3] Parssian A, Sumit S, Varghese J S. Assessing data quality for information products: impact of selection, projection, and cartesian product[J]. Management Science, 2004, 50(7): 967-982.

[4] 陈卫东,张维明. 数据质量模型及选择运算中的质量传播研究[J]. 计算机工程与应用,2007,43(27):1-3.

[5] Yang W L, Wang R Y, Ziad M. Data quality[M]. New York: Kluwer Academic Publishers, 2001.

[6] Pipino L L, Lee Y W, Wang R Y. Data Assessment[J]. Communications of the ACM, 2002, 45(4): 211-218.

[7] 刘 慧,刘 敏,韩 兵. 基于维度的信息系统数据质量评估指标体系研究[J]. 信息系统工程,2010,20(6):102-105.

[8] Even A, Shankaranarayanan G. Utility-driven assessment of data quality[J]. The Data Base for Advances in Information

(下转第 89 页)

位分配结果及获得的总收益 R 与 MER 一致,γ 风险规避参数越大,其舱位分配结果及获得的总收益 R 与 FCFS 策略一致,即 RAMDP 模型为风险规避决策者提供了更多决策选择。

表 3　各等级货物非齐次泊松到达过程及货物重量

等级	非齐次到达的阶段及到达的重量
1	27(3.0),37(7.0),56(3.0),77(5.0),93(6.0),94(4.0),119(3.0),130(2.0),134(6.0),197(5.0),230(4.0)
2	15(15.0),21(12.0),30(16.0),50(25.0),59(20.0),60(21.0),81(21.0),90(29.0),92(13.0),102(16.0),115(21.0),120(18.0),127(30.0),137(26.0),169(16.0),175(24.0),247(16.0)
3	17(67.0),18(47.0),38(54.0),62(45.0),69(51.0),79(69.0),81(73.0),87(58.0),101(43.0),118(59.0),149(46.0),160(58.0),162(38.0),181(69.0),211(61.0),234(69.0),239(61.0),260(69.0)
4	52(272.0),101(167.0),125(129.0),144(234.0),179(166.0),198(247.0),199(129.0),227(189.0),228(189.0),238(198.0),267(122.0),284(94.0),287(138.0),297(223.0),298(171.0)
5	275(340.0),301(415.0),313(456.0),317(374.0),322(391.0),363(330.0)
6	304(482.0),371(542.0),375(672.0),391(621.0)

表 4　不同策略下的舱位分配情况与总收益状况对比

	6	5	4	3	2	1	R
FCFS	2317	2306	309	0	56	11	167788.5
0.01	2317	2306	309	0	56	11	167788.5
0.005	2317	2306	293	69	0	15	166977.1
0.003	2317	1891	309	260	180	41	169125.2
0.001	1835	2306	171	298	339	48	174081.2
0.0005	1835	1891	309	577	339	48	175945.3
0.0001	1835	1061	877	847	324	48	178300.2
0.00001	1293	1435	1075	796	339	48	180016.2
MER	1293	1435	1075	796	339	48	180016.2

6　结束语

文中以航协 TACT 发布的空运货物价格作为货物等级划分的依据,并且考虑货运订舱行为的不确定性,构建了基于风险规避的有限阶段 RAMDP 决策模型,该模型具有结构化的舱位保护策略,而且风险规避因子为决策者提供更多策略选择。实验分析验证了该模型适用于解决不确定性航空货运舱位优化控制问题。文中的研究只限于单航段情形,而且未考虑超售因素,未来航空货运舱位优化控制模型应该包含超售情形,并将模型扩展到多航段及整个航线网络中,以进一步提高航空货运收益。

参考文献:

[1] Popescu A, Keskinocak P, Johnson E, et al. Estimation air cargo overbooking based on a discrete show-up-rate distribution [J]. Interfaces, 2006(3): 248-258.

[2] Kasilingam R G. An economic model for air cargo overbooking under stochastic capacity[J]. Computer Industry Engineering, 1997, 32(1): 221-226.

[3] 张永莉. 应用收益管理方法的航空货运销售[J]. 中国民航学院学报, 2003(5): 24-26.

[4] 桂云苗, 朱金福. 航空货运动态舱位控制模型研究[J]. 预测, 2007(6): 53-56.

[5] Barz C. Risk-averse capacity control in revenue management [M]. Berlin: Springer-Verlag, 2007: 47-62.

[6] Barz C, Waldmann K H. Risk-sensitive capacity control in revenue management[J]. Mathematical Methods of Operations Research, 2007, 65(3): 565-579.

[7] 罗远浩, 何小通, 吴　立. 航空货运舱位重量等级需求统计分析(一)[J]. 空运商务, 2009(5): 29-30.

[8] Mongin P. Expected Utility Theory[M]//The Handbook of Economic Methodology. [s. l.]: [s. n.], 1998: 342-350.

[9] 郭文英. 期望效用理论的发展[J]. 首都经济贸易大学学报, 2005(5): 11-14.

[10] 周国梅, 傅小兰. 决策的期望效用理论的发展[J]. 心理科学, 2001, 24(2): 219-220.

[11] Joongwoo B. Capacity control in network revenue management: clustering and risk-aversion[D]. USA: MIT, 2010.

[12] 邓永录, 梁之顺. 随机点过程及其应用[M]. 北京: 科学出版社, 1992: 69-70.

[13] Kirkwood C W. Approximating risk aversion in decision analysis applications[J]. Decision Analysis, 2004(1): 55-72.

(上接第 84 页)

Systems, 2007, 38(2): 75-93.

[9] 黄武锋, 郑　华. 面向企业信息化的数据质量评估研究[J]. 计算机技术与发展, 2011, 21(1): 185-188.

[10] 丁海龙, 徐宏炳. 数据质量分析及应用[J]. 计算机技术与发展, 2007, 17(3): 236-238.

[11] 王晓华. 电信数据挖掘的数据质量评估技术[D]. 杭州: 浙江大学, 2010.

[12] Missier P, Embury S, Greenwood M. Quality views: Capturing and exploiting the user perspective on data quality[C]//Proc of 32th VLDB. Seoul, Korea: [s. n.], 2006: 977-988.

[13] Yuan Man, Liu Wei. A novel data quality controling and assessing model based on rules[C]//ISECS'10 Proceedings of the 2010 Third International Symposium on Electronic Commerce and Security. Guangzhou: Academy Publisher, 2010: 29-32.

[14] 李　聃. 元数据在数据仓库中的研究与应用[D]. 成都: 西南石油大学, 2007.

一种基于规则的数据质量评价模型

作者： 袁满， 张雪
作者单位： 东北石油大学 计算机与信息技术学院,黑龙江 大庆163318
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(3)