面向企业信息化的数据质量评估研究

黄武锋,郑 华
(广西财经学院 计算机与信息管理系,广西 南宁 530003)

摘 要:数据质量问题是企业信息化过程中面临的一项重要挑战,但针对数据质量评估的研究还缺乏足够的重视。文中从数据质量定义出发,阐述了数据质量的各个不同维度及其评估指标的确定,在对比分析已有成果的基础上给出了主观、客观两种评估方法,通过引入SOA上下文的可重用服务思想,设计了一种数据质量评估的服务框架,基于该框架对输入输出、流程管理、自动化评估等服务进行了阐述,并使用Web Services服务组件的形式实现了所有的功能需求。

关键词:企业信息化;数据质量;维度;数据质量评估;SOA
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2011)01-0185-04

Study of Data Quality Assessment for Enterprise Informationization

HUANG Wu-feng, ZHENG Hua
(Department of Computer and Information Management Guangxi University of Finance and Economics, Nanning 530003, China)

Abstract: The data quality problem is an important challenge in the process of enterprise informatization, but the study for data quality evaluation still lacks enough attention. Starting from the definition of data quality in this paper, the definitions of different dimensions of data quality are expounded, and the identification of subjective and objective evaluation methods are provided, then a kind of ideal data quality evaluation model is designed by embracing capabilities within the context of SOA, design a kind of data quality assessment service framework, based on the framework of input and output, process management, automated evaluation is discussed, and the use of Web Services have identified service component function of form all the needs. Finally, a further research direction is given.

Key words: enterprise informationization; data quality; dimensions; data quality assessment; SOA

0 引 言

随着信息社会的快速发展,企业各种管理信息系统的建设在经过不同的发展阶段后形成了复杂多变的数据环境。信息数据的大量产生,以及从各种渠道收集获取的不符合系统要求的数据造成了数据重复、不一致、格式混乱和缺陷等问题,从而对企业的数据分析、数据处理影响尤其严重。数据作为信息系统的基础和核心,对信息系统起着至关重要的作用,数据质量的高低对整个系统有直接的影响。好的数据质量是各种数据分析能够得到有意义结果的基本条件,而质量低劣的数据已经成为影响企业进行正确决策的重要因素。IDC公司(2008)在一项有关"中国数据集成与数据质量市场"调查的白皮书指出:"中国特殊的软件建设背景形成了复杂的数据环境;43%的接受调查的中国企业有10个以上的业务支撑软件系统,81%的接受调查的中国企业在使用2种以上的数据库产品,42%的接受调查的中国企业在以非结构化方式存储重要数据。在接受调查的100家大中型企业中,超过70%的接受调查的中国企业已经建设或正在建设数据集成项目,并重点关注数据质量热点问题。"可见,数据质量问题已经开始受到各行业的重视,但一直以来学术界和工业界更多的致力于数据质量的提高技术,如数据清洗的研究,而缺乏对数据质量问题的预防,即评估的研究。

本课题主要针对后者进行了研究,分析其内涵并重点关注评估指标的确定以及评估方法的选择,最后设计了一种理想的解决方案。

1 数据质量的内涵和基本指标

1.1 数据质量的内涵

目前的研究从不同的角度对数据质量进行定义:基于消费者的角度,基于制造的角度,基于产品的角度,基于价值的角度,先验的角度等。在许多文献中,

收稿日期:2010-04-08;修回日期:2010-07-05
基金项目:广西"十一五"规划2008年度课题(08FTQ001)
作者简介:黄武锋(1975-),男,广西容县人,讲师,研究方向为计算机应用;郑 华,副教授,博士研究生,研究方向为网络管理信息系统,电子商务。

数据质量 DQ(Data Quality)与信息质量 IQ(Information Quality)两个术语通用[1],定义多种多样。文献[2,3]认为数据质量是数据满足特定用户期望的程度;而文献[4]认为数据质量是数据适合使用的程度。现代数据质量概念主要包括以下几个方面:一是注重从用户角度来衡量数据质量,强调用户对数据的满意程度;二是数据质量需要建立一套有效的数据质量管理体系,从多个角度来评价数据的好坏;三是数据质量可以用人们对数据期待的特性以及所获得的数据的特性之间的差距来表示;四是数据质量最重要的一个特性即它是一个复杂的、多维度的概念。

数据质量维度是一组表达数据质量构成或者数据质量单一方面的数据质量属性,不同的研究对数据质量维度的分析是有区别的,尽管基本上都是围绕着数据质量的特性展开,但其所采用方法和描述的角度和出发点不尽相同。文献[5]采取二阶段调查方法研究了四类数据质量维度:固有质量、可访问性质量、语境质量、表达质量;文献[6,7]则从功能性特征、可靠性特征、效率特征、合用性特征、维护特征、可移植性特六个方面进行了描述;而文献[8,9]以符号学为基础,建立语义层次的维度、语用层次的维度、社会层次的维度等四个符号学层次共十一个质量维度。从上述数据质量维度方案可以看出,由于数据质量维度的复杂性和多样性,有些维度的重要性得到广泛认同,虽然每个维度都有具体的改进策略,但由于偏重于数据使用的环境和用户的差异,也不可能建立一套能被广泛接受的完整的数据质量维度。因此,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。

1.2 数据质量评估的基本指标

对数据质量进行评估能帮助企业准确地了解数据的内容、质量和结构。主管人员参与数据质量评估以及分析在数据检查过程中发现的问题对于数据质量评估来说都很重要。数据质量的评估过程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,难点在于数据质量的含义、内容、分类、分级、评价指标的确定等,其核心在于如何具体地评估各个维度[10]。在最有效的数据质量评估中,所有问题都将按照对业务影响从大到小的顺序列出,这将帮助 IT 机构节省项目成本。许多研究把评估各个数据质量维度的指标作为数据质量评估的基本指标,但在进行不同行业的数据质量评估时,从不同的角度对数据质量衡量的指标是不一样的,现有的一些分析评估指标的侧重点、覆盖范围、针对的领域也各不相同,因此无法形成一个标准。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[11]。但是,数据质量评估至少应该包含以下七个方面的基本指标(见表 1)。

表 1 数据质量评估的基本指标

评估指标	指标含义
完整性	主要包括实体缺失、属性缺失、记录缺失和字段值缺失四个方面
及时性	指数据提取、传送、处理、装载、展现的及时和快速性
合法性	主要包括格式、类型、值域和业务规则的有效性
唯一性	指主键唯一和候选键唯一两个方面
一致性	指不同系统之间的数据差异和相互矛盾的一致性
准确性	一个数据值与设定为准确的值之间的一致程度,或与可接受程度之间的差异
时效性	描述数据的时间特性对应用的满足程度。数据从产生、加工处理,到消亡,有一个相对的有效期。该特性描述了数据是当前数据还是历史的数据

同时,可以考虑定义一个指标:数据质量问题频率。

指标定义:数据质量问题频率 = 数据质量问题发生次数/存储的总数据量

指标单位:次/GB

根据数据质量评估指标将最终的评估结果划分为三个等级(见表 2):

表 2 数据质量的评估等级

数据质量等级	描述	统计口径
一级	数据质量差,需要重点监控	数据质量问题频率大于等于 1 次/GB
二级	数据质量一般	数据质量问题频率大于等于 0.5 次/GB,小于 1 次/GB
三级	数据质量好	数据质量问题频率小于 0.5 次/GB

通过对数据质量问题频率的考评和等级划分,就可以从企业数据中心众多的数据中解放出来,集中精力把有限的资源投入到需要重点关注的主题数据。因此数据质量可信等级是数据质量提高的有效途径。

2 数据质量的评估方法

进行数据质量研究的目的是提高各企业单位的数据质量,以求更好地利用数据制定正确的信息决策,获取更多的社会、经济效益。目前,数据质量的研究主要围绕两个方面展开:(1)数据质量评估;(2)数据质量提高。数据质量评估是解决数据质量问题的一个源头性问题。通过数据质量评估,及时掌握各类数据的可靠程度或差错率的大小,系统查找影响数据质量的因素,并有针对性地采取措施,提高数据质量[12]。文献[13]描述了用一个控制矩阵来显示数据集的质量高低,通过它来反映优化处理后数据集质量方面发生的变化,但对于数据质量的计算和数据的优化策略都没有涉及到。文献[3]提出的数据质量评估模型描述了将客观评价和主观评价结合体系来解决如何将客户反馈信息纳入到数据质量评估中来的问题。文献[14]将按照由选择模块、质量评估模块和简表模块构成的

评估过程的数据质量评估架构进行,判断得到的数据质量值是否符合用户的期望值。从现有文献分析可知,由于数据质量与背景和用户密切相关,现有的数据质量测量和评估标准主要采取主观和客观相结合的方法[15]。

主观数据质量的评估采用传统的软件测量方法,即问卷调查方式对各个数据质量维度进行评估。调查对象可以是数据的使用者或数据的维护者,其评估结果基本依赖于数据的使用或维护对象对于数据使用要求的认识与理解以及他们对所使用数据的期望的满意程度,但这类对象往往对系统的数据质量要求认识不足或存在偏差,评估的结果中用户的主观意识占据绝大部分,但基本上能反映出该类数据质量维度的高低。

客观评估根据数据的评估标准或具体的任务来进行评估。客观评估任务可以是无关(指脱离具体的应用在上下文无关的环境中进行数据质量评估,评估的标准可以应用到任何数据集上。)或者相关(评估是在具体的应用环境中,评估标准必须是事先规定的行业标准或者数据库管理员提出的数据约束)。

数据质量客观评估可使用以下三种算法来实现:

(1)简单比率。

简单比率是期望值与数据重要性的比率值。通常期望值越高数据越重要,系统的期望值可与数据质量改进阶段相结合成为可变的值。在不同的时期和需求下,可以提高或降低期望值。如在数据质量控制初期,可取一个较低的值,而在系统数据质量慢慢提高后则可逐步提高期望值,对系统提出更加格的数据质量要求,以进一步提高系统数据质量。

(2)最大-最小运算。

最大-最小运算值适用于需要多个数据质量变量衡量的维度。最小值是多个规格标准化数据质量变量值的最小取值,而最大值是多个规格标准化数据质量变量值的最大取值,两者的值都介于0和1之间。

(3)加权平均。

使用加权平均值的前提就是必须对各个数据质量变量的重要性有一个清晰的认识。对于数据质量评估的多变量维度,可以使用加权平均值来代替最大-最小值进行评估更为合理。

据此综合利用主观评估和客观评估方法,数据质量所期望的值为主客观评估都达到最高的状态,如主观评估结果偏低或客观评估结果偏低,则必须对评估的数据进行深入调查,找出数据质量偏低的原因,然后采取正确的措施来提高数据质量。

3　一种理想的数据质量评估模式

从企业的角度来看,常常缺少一个权威性的数据源,所以无法为组织的核心数据提供一个完整且准确的视图。相反,对于不同的业务线、通道或产品类型,常常使用不同的技术以不同方式存储和处理数据。许多大型组织的核心企业信息分散在多个垂直系统上,而且存在重复的数据,每个系统根据自己的上下文维护信息,而不考虑整个企业的上下文,这进一步加重了业务过程中的不一致性——业务过程本身在企业的不同部分中常常很不一致。数据即使满足它原来的存储库和应用程序的规则和约束,也不一定能够满足企业级的需求。例如,一个标识符在某个系统中可能是唯一的,但是在整个企业范围内不一定真的唯一。在将数据通过业务系统向企业范围公开时,在原来的应用程序中无关紧要的质量问题可能会变成很严重的问题。例如,缺少值、冗余条目和不一致的数据格式在原来的应用程序中可能不会造成什么问题,但是在向新消费者公开时就可能出现问题。本课题的解决方案是在服务分析和设计期间进行数据质量评估。在对支持服务的源系统进行分类之后,就可以开始研究它们的数据质量问题。例如,应该检查数据是否符合相关的完整性规则。应该检查是否存在重复的数据,研究在数据匹配和聚合期间如何解决重复的数据。以这些分析为基础,可以采取适当的措施来确保服务的实现能够满足潜在服务消费者对数据精确性和含义的要求。

在这种解决方案中,通过开发可重用的服务,同时利用已有的应用程序和数据,并谨慎地公开已有系统的功能,企业就可以发展它的IT功能,以便更好地适应变化,SOA服务架构是一种很好的选择。整个解决方案是建立在典型SOA解决方案框架上的,所有的服务都将通过集中的ESB(Enterprise Service Bus,企业服务总线)进行整合。对于不同的服务,有不同的企业中间件/产品支持其运行。图1是基于SOA的数据质量评估模式的服务框架视图。系统输入可来自不同的渠道,包括传统的纸质文件渠道、web渠道、邮件/传真/电邮渠道、电子表单渠道等,具体形式可以是各种类型的文件(数据库、XML等),通过独立的"输入管理服务"进行动态处理。"输出管理服务"主要与核心管理系统提供的服务相交互,以保证流程内输出数据的生成、存档及分发。"流程管理服务"将提供/整合监控服务以提供对服务、流程、工作量状态的可视化。"工作管理服务"负责捕捉、组织工作任务,并根据不同的工作类型、技能、工作量将其分配给用户。"交互处理服务"主要负责手工操作任务的实现。"自动化评估服务"主要基于预定义的指令及服务目录中的业务服务,提供流程编排及流程映射以处理自动化工作任务并跟踪手工操作任务。该方案使得数据源识别、工作任务自动化映射以及相关SOA服务将使"无人值

守”的自动化处理成为可能。对于不同的流程，人工操作可以被划分为独立的“交互处理服务”，“工作管理服务”则可以协助进行智能化工作任务分配。这样，人力资源将得到最大程度的重用，相应的人工操作成本将降到最低，整体操作流程的费用也可以得到很好的控制。

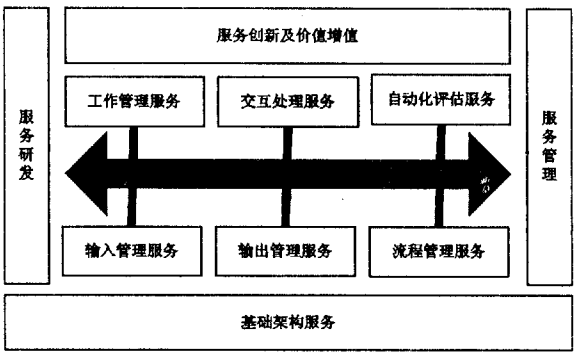

图 1 数据质量评估模式的服务框架规图

该方案的具体流程可划分为三个部分：

1)分析源系统。将数据质量度量指标应用到来自已识别的数据存储的数据上，以确定数据质量级别；解释数据度量指标结果，并将这些结果翻译为业务术语。创建详细的报告、图表和总结，以描绘数据质量级别和提供建议。

2)分析目标系统。发现所分析的源系统与目标系统之间的差异；创建用于消除差异的建议。

3)评估每个相关数据元素的校准(alignment)和协调(harmonization)需求。将数据质量度量应用到识别出的数据元素上，以评估当前的标准化和相匹配的支持，并将这些结果翻译为业务术语；创建详细的报告、图表和总结，以描绘标准化和匹配的级别，并提供建议。

整个评价流程需要通过软件系统来实现。为此，笔者基于 Web Services 技术，采用. NET 架构，把应用系统的功能及业务流程逻辑封装成标准的服务，通过服务的描述、发布与发现机制实现服务间的调用与组合，最终实现了数据质量评价的整个业务流程，该系统的结构如图 2 所示。

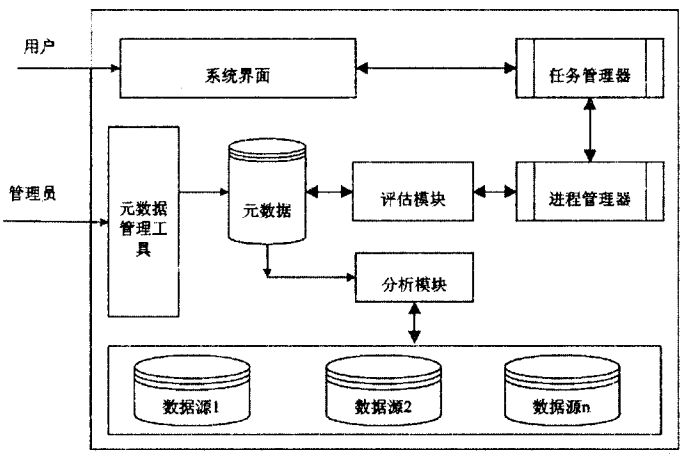

图 2 数据质量评估系统结构

4 结束语

随着人们对信息系统的依赖越来越强，数据质量的研究成果必将在人们生产和生活的各个方面发挥巨大的作用。

尽管对数据质量的研究取得了比较系统的研究成果，但在许多方面尚待进一步探索：

(1)数据评估标准信息库的建立；

(2)评估标准方法体系的建立和完善；

(3)数据质量评估方法；

(4)开发适用于信息系统从分析设计到运行维护全过程的、适合数据质量管理的信息系统模型也值得进一步研究。

数据质量评估的研究必将成为企业信息化发展的战略重点，成为企业新一轮发展的主要支撑。最后，需要强调的是：数据质量问题不外乎两方面原因，管理上(人)的因素和技术上的因素，建立健全科学、规范的数据质量管理机制，从组织、制度、技术等层面保障对数据的有效监控，是破解如何保证数据质量难题的关键。

参考文献：

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

[2] Kahn B K, Strong D M. Product and Service Performance Model for Info-rmation Quality[J]. An Update, 1998(4)：102-115.

[3] Cappiello C, Francalanci C, Pernici B. Data quality assessment from user's pespective [C]//IQIS. [s. l.]：[s. n.], 2004：35-43.

[4] Huang K-T, Lee Y W, Wang R Y. Quality information and knowledge management[M]. New Jersey：Prentice Hall, 1998：46-53.

[5] Wang R Y, Strong D M. Beyond Accuracy：What Data Quality Means to Data Consumers[J]. Journal of Management Information Systems, 1996(4)：80-86.

[6] Orr K. Data Quality and System Theory[J]. Communications of the ACM, 1998(2)：72-78.

[7] Zeist R H J, Hendriks P R H. Specifying Software Quality with the Extended ISO Model[J]. Software Quality Journal, 1996(4)：64-70.

[8] Shanks G, Corbitt B. Understanding Data Quality：Social and Cultural Aspects[C]//In Proceedings

(下转第 192 页)

一阶谓词表示成 Time(K,P,Ts,Tl)。类似地,公交班次在中途站点的时序时段逻辑可以表示为 Time(K,P,Ts,Tl,Td)。其中 Td 代表车辆进站后的停留时长[12]。

略微区别于上面两层模块,虽本级模块内仍旧能够划分成不同细分层次,但它们之间不可完全依序顺次进行,而必须多次往复递归,经搜索、检验与矛盾、冲突消除,方可获取最后旅行咨询结论。在该级模块中,表面看好像问题极其错综复杂,可深入探究本质时却不难发现:公交班次与中转接续地点才是关键。因为前项决定着换乘次数和评价准则主要成份保障,后项关系到咨询结论实用化程度。

公路旅行面临两种中转接续模式,即单点换乘和区段换乘。由于二者比较起来,区段换乘更加富有空间选择随意性与接续冲突、矛盾消解策略多样性,所以应当优先给予采纳。事实上,换乘地点选定既取决于公交班次交汇密度、接续时间裕量、食宿方便、错过末班车风险和应对时变、不确定性及突发事件等的有效对策搜寻的容易性,同时,又应当为因临时变故造成原有方案无法实施而选取次优替代留有余地。荒野区域绝非属于换乘禁地,只是在公交班次稀少、近邻夜间与行车规律性波动异常严重时,尤其应该谨慎。

班次换乘合理性与换乘冲突不存在绝对标准,其核心便决定于中转换乘时序如何认定,8:00 发车 8:30 到站可视作换乘接续不合理和冲突,但若认为来日再搭乘又似乎不属于问题。为消除此类二义性与认识不准确,有必要设立普遍认同的换乘合理、矛盾、冲突相对化标准。本标准不可一成不变,应当根据出行急迫性、换乘地服务设施完善性、公交运营状态、班次密度、消费接受档次和旅行者身体健康状况等综合兼顾、妥当调整。另外,也不排除反复试探或逐渐寻优。

4 结束语

如何圆满协助每个人实现最佳化搭乘长途公交车旅行,不但富有非常明显的经济、资源有效利用、环保与推动社会快速发展价值,同时也将极大地扩充人工智能科学研究内涵,并促使其尽快走向完善。本文涉及研究就是针对公路交通领域面临难题开展的一项实质性探索,它在知识利用、特征与广义拓扑路网建模和公路交通建模(这些内容另外辟文论述)等基础之上,采用分级递阶思想借助区域穿越规划、路线、路段规划与班次、中转接续规划三大主体模块予以兑现。由当前实验结果看,该项工作成效已初步显现,并映射出巨大的生命力和实用化意义。当然,也不排除可能仍隐含某些局陷与不足,特诚请同行给予批评、指正。

参考文献:

[1] 杨新苗,王 炜,马文腾.基于GIS的公交乘客出行路径选择模型[J].东南大学学报,2000,30(6):87-91.

[2] 侯 刚,周宽久.基于换乘次数最少的公交网络最优路径模型研究[J].计算机技术与发展,2008,18(1):44-47.

[3] WANG Hong-jian, XIONG Wei. Research on global path planning based on ant colony optimization for AUV[J]. Journal of Marine Science and Applicaton,2009,16(8):58-64.

[4] 陈则王,袁 信.基于分层分解的一种实时车辆路径规划算法[J].南京航空航天大学学报,2003,35(2):193-197.

[5] 朱 达,佟 琼.基于旅客出行选择的旅行时间价值研究[J].北京交通大学学报,2007,12(10):42-85.

[6] 付梦印,李 杰.基于分层道路网络的新型路径规划算法[J].计算机辅助设计与图形学学报,2005,17(4):719-722.

[7] Frontzek T,Goerke N, Eckmiller R. A Hybrid Path Planning System Combining the A*-Method and RBF-Networks[J]. URISA Journal,2002,25(8):78-85.

[8] 金炳尧.一个用于多目标优化的进化规划算法[J].微机发展(现更名:计算机技术与发展),2001,11(5):25-28.

[9] 杨淮清,薛明昊,余冠华.一种面向中转换乘的铁路网建模方法研究[J].计算机技术与发展,2010,20(4):211-214.

[10] Pang G K H, Takabashi K, Yokota T. Adaptive route selection for dynamic route guidance system based on fuzzy neural approaches[J]. IEEE Transactions on Vehicular Technology, 1999,48(6):2028-2041.

[11] 冯 林,姜 浩.基于时间约束 Petri 网的工作流可调度性分析[J].计算机技术与发展,2006,16(11):34-37.

[12] 林 闯,刘 婷,曲 扬.一种不确定时段的扩展时段时序逻辑:时间 Petri 网表示和线性推理[J].计算机学报,2001,24(12):1299-1309.

(上接第 188 页)

of the 10th Australasian Conference on Information Systems. [s.l.]:[s.n.],1999:58-64.

[9] Pipino L, Y Lee, Wang R Y. Data Quality Assessment[J]. Communications of the ACM,2002(5):14-22.

[10] 韩京宇,徐立臻,董逸生.数据质量研究综述[J].计算机科学,2008,35(2):1-12.

[11] 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.

[12] 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):3-15.

[13] Scannapieco M, Catarci T. Data Quality under the Computer Science Perspective[J]. Archivi & Computer,2002(2):1-15.

[14] Crosby P B. Quality is Free:The Art of Making Quality Certain [M]. New York:McGraw-Hill,1988:33-36.

[15] 齐艳珂,李晓举,周 青.数据质量的研究[J].中小企业科技,2007(7):10-11.