计算机科学1998 Vol. 25 №. 1

基于主动规则对象的数据质量管理

Data Quality Management Based on Active Rule Object

胡 华[1,2]

(杭州商学院计算机与信息工程系 杭州310035)[1]

寿宇澄 高 济 何志均[2]

(浙江大学计算机科学与工程学系 杭州310027)[2]

摘 要 This paper analyzes the specialities in procedure of creation, processing and consumption of data information and provides an approach based on active rule object to realize total data quality management. The approach can efficiently manage and handle data information quality based on abstract rule sets.

关键词 Data Information, Data quality management, Rule, Active Object

1 前言

数据质量的控制与管理研究一般而言可从技术和管理两个方面进行。在技术上，不仅要保证数据的精确和高效率，而且还要保证系统中有关数据在逻辑上一致完整且合法，Pamela Cahn[7]和 Tim Quinlan[8,9]分别从技术的角度讨论了数据库的质量控制和数据质量评价问题；现有的数据库产品亦在支持多种数据类型的一定精度的数据操作基础上，还提供数据在逻辑上的一致性和完整性检验控制机制；在管理上，一般侧重于采用以 TQM(全面质量管理)为基础的数据质量控制管理的方法在数据的目标需求和过程控制方面进行质量管理与控制，如 G. J. van der Pijl[1], D. M. Strong[11] 和 Richard Y. Wang[2]分别所做的工作。在几乎所有的研究工作中，研究者们有一个共同的观念：由于质量管理经常要考虑人的因素，因此对企业中所有的信息处理和控制而言，很难实现一个完备的全自动化质量控制管理方案。虽然如此，很多研究者还是注意到：由于企业中计算机控制处理的信息具有：操作过程重复性高、数据特征的结构性强且比较适于过程式的自动化处理的特点。对这类信息的收集和处理，可以采用全面质量管理的方法实行自动化的质量控制处理。遗憾的是虽然有不少研究者提及这一点，但我们还没有看到这一方面的一个具体解决方案。本文根据企业智能信息系统中的信息数据产生和使用过程的特点，在面向对象和主动式数据库技术的基础上，提出了一种基于主动规则对象的全面数据质量管理方法。该方法在对象独立性和规则的知识性基础上，可有效地适应并处理信息系统中因操作、设计和企业环境变化所引起的数据质量问题，提高了相应系统中的数据信息质量。

2 数据质量控制管理指标

从全面数据质量管理的角度来看：数据信息的质量控制和管理可分为：信息系统生命周期和数据产品值链两个方面[2]。前者主要涉及信息系统的设计、开发和交付过程中系统的质量控制与管理；而后者则主要涉及数据信息在系统中的产生、处理和使用过程中的数据质量控制与管理。

2.1 信息系统生命周期

胡华 讲师，博士，目前的主攻方向为人工智能和 CIMS 集成；寿宇澄 副教授，博士，主要研究的领域为：人工智能、面向对象的数据库系统、PDM 和 CIMS 集成；高济 教授，博士导师，主要研究的领域为人工智能和 CIMS 集成；何志均 教授，博士导师，主要研究的领域为人工智能和 CIMS 集成。

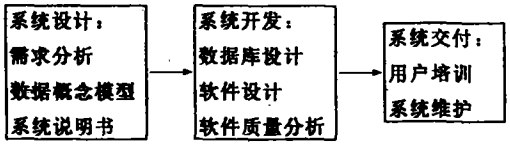

图1 信息系统生命周期

数据是以一定的方式组织存储在信息系统中的，因此信息系统性能的优劣在很大程度上直接影响着数据信息的质量指标。信息系统生命周期是通过将信息系统的生成过程划分为设计、开发和交付等几个阶段的来进行数据质量控制与管理。其中：

(1)系统设计阶段，对数据质量的管理控制主要体现在对系统需求分析、数据概念模型和系统数据说明书等工作的质量控制管理；

(2)系统开发阶段，对数据质量的管理控制体现在数据库的实现、软件工程开发方法和软件质量分析方法；

(3)系统交付与维护阶段，对数据质量的管理控制则体现在进行用户培训和系统维护，加深用户对系统功能的理解，减少误操作，提高系统可靠度。

2.2 数据产品值链

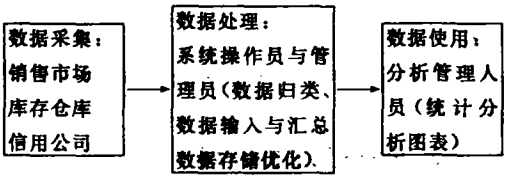

图2 数据产品值链

数据产品值链描述的是系统中数据产生、处理和使用时的质量控制与管理问题。映射到具体的处理上就是防止或减少因操作错误、数据设计理解错误和企业工作方式临时改变带来的数据存取效率以及逻辑不一致和完整性问题。如根据对当前库存数据来检验生产调度和销售的合理性或根据借贷方是否相等来检验记帐凭证的合法性等。

3 基于规则的主动对象

使用面向对象技术的系统，根据对象的行为特点，可以将对象划分为：被动对象(Passive Object)、主动对象(Active Object)和软件代理(Software Agent)三类[13]。其中主动对象不仅响应发送给它的消息，还可主动监控有关的事件并执行有关的操作。主动对象的控制一般采用 ECA(Event-Condition-Action)规则来进行。ECA 规则的执行是原子的；当其监控的事件发生时，规则就被激活并检验有关的条件，若条件满足将执行其动作。

此外为了更好地描述和控制主动对象，可以用一个具体的规则系统来描述分析有关对象的规则。规则系统是近年来主动式数据库领域内极为活跃的一个研究方向。由于引入了规则处理，主动数据库将原本需要写入应用程序的大量数据管理和处理策略抽象出来并用规则实现，从而达到了一定程度的知识独立。对主动数据库的研究表明：有效使用规则的两个关键是：解决规则库容量增大时所带来的规则集管理和语义冲突问题，以及保证规则执行结果一致且可终结。我们采用 E. Baralis 等[12]提出的规则模块化分层方法来解决上述问题。

定义1 事件是基于 ECA 的主动对象外部发生的，主动对象感兴趣的所有活动、变化和事情。

定义2 ECA 规则集是一个三元组〈E,C,A〉，其中：C 为系统当前状态的断言；A 为进行数据处理操作的序列集。

定义3 一个基于 ECA 规则的主动数据库是一个二元组〈E,R〉，其中：E 为规则所作用的对象实体；R 为 ECA 规则集。

定义4 一个基于 ECA 规则的主动状态是一个二元组〈D,T〉，其中：D 为系统状态；T 为触发的 ECA 规则集。

定义5 一个基于 ECA 规则的静止状态是一个二元组〈$D,\varnothing$〉，其中：触发的 ECA 规则集为空集。

定义6 〈E,R〉为基于 ECA 规则的主动数据库，$S\subseteq R$ 是 R 中的一个规则子集，则：S 的输入事件集(IE_s)中的事件为触发 S 中规则的事件；S 的输出事件集(OE_s)中的事件为执行 S 中规则而引发的事件。

定义7 如果主动数据库〈E,S〉中的规则被准确执行，经过一系列状态变化后可得到主动状态〈D_1,T_1〉($T_1\subseteq S$)，且规则处理最后终止于静止状态〈$D_2,\varnothing$〉，则规则子集 S 是一个局部聚合层。

定义8 对于主动数据库〈E,R〉中的规则层 S_i 和 S_j，如果 $OE_j\cap IE_i=\varnothing$，则我们称 S_j 不触发 S_i。

定义9 对于主动数据库〈E,R〉，$S=\{S_1,\cdots,S_n\}$；$S_i\subseteq R$，$i=1,\cdots,n$〉满足条件：1)S_i($i=1,\cdots,n$)是局部聚合的；2)可以在 S 上定义顺序$<$，即如果 $S_i<$

S_j 则有：①S_i 中的规则比 S_j 中的规则有更高的优先级；②S_j 不触发 S_i；则我们称 S 为 R 的一个事件分层。

根据以上定义，容易证明：

引理 规则集 R 若存在一个定义于其上的事件分层 $S=\{S_1,\cdots,S_n;S_i\subseteq R,i=1,\cdots,n\}$，则 R 上的任意规则处理序列将终止。

最后，随着系统处理环境和状态在时间轴上的变化，主动对象的状态亦在不断地进化。综上所述，我们有：

定理 基于 ECA 的主动规则对象状态可以由三元组〈T,R,D〉描述，其中：T 为离散的时间序列；R 为对象基于的 ECA 规则；D 为对象中的当前数据和作用于数据之上的操作。

4 基于主动规则对象的数据质量控制系统实现结构

基于 ECA 的主动规则对象的数据质量控制系统的实现结构如图3所示。

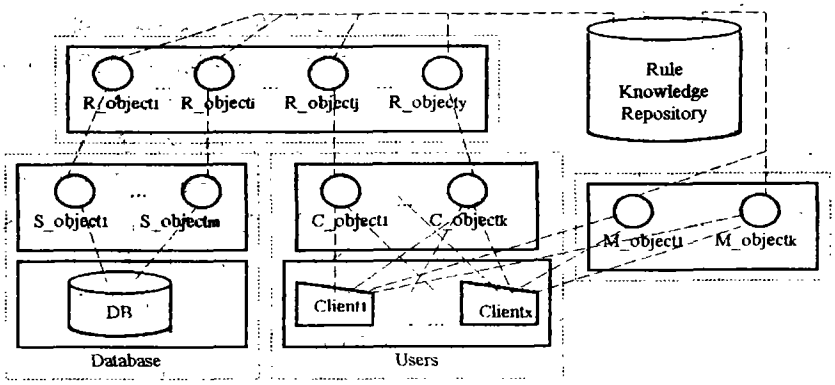

图3

在图3中，我们把基于 ECA 的主动对象分为：界面操作服务类、数据库操纵服务类、流程协调控制类和规则知识字典维护类等四大类。前三类对象基于处理的规则知识，参照质量指标进行抽象后存放于规则知识库(Rule Knowledge Repository)中并对质量控制管理规则知识进行重载并进化；第四类对象则负责对规则知识库进行维护管理。

4.1 界面操作服务类对象

界面操作类对象在系统操作环境的前端，它们的作用包括：提供信息系统操作的界面；接受并检验用户数据操作的合法性；向其它对象转发用户数据操作的请求；响应系统事件进行状态进化达到操作环境自动适应系统环境变化的要求。

4.2 数据库操纵服务类对象

数据库操纵服务类对象在数据库服务器端。它们根据其操作的数据库不同，有关的操作部分将被重载，以满足对不同数据库的操作请求。该类对象的一个最重要作用是利用规则知识，保证数据库中的数据信息一致和完整。此外，该类对象还能根据数据库状态事件激活改变系统状态的事件以便系统通过自动调节全局知识字典和操作界面表现来适应有关的变化。

4.3 流程协调控制类对象

流程协调控制类对象起到对系统状态不同阶段的各界面类对象和数据库操纵服务类对象的中介协调作用。根据系统状态的不同阶段特点，这类对象的操作部分也将被重载，以满足对不同对象间的行为进行权衡协调。此外，该类对象还能根据其他类的对象激活的系统事件激活改变描述系统状态的全局规则知识字典并通知有关的对象进行进化。

4.4 规则知识字典维护类对象

规则知识字典维护类对象主要负责响应用户对规则知识字典的维护和控制。

结语 我们采用 C++ 语言对本文的方法进行了实现，实践证明：在面向对象和主动式数据库技术的基础上，本文的全面数据质量管理方法能够较好地适应并处理信息系统中因操作失误、理解设计错误和企业环境变化所引起的数据信息质量问题，提高了相应系统中的数据信息质量。(参考文献共15篇略)