实用的数据收集与清理模型的研究与应用①

Practical Data Collection and Clean-Up Model

张志宽　罗晓沛　(中国科学院研究生院 北京 100049)

摘　要： 从提高软件项目中数据收集工作效率的角度出发，在汇总当前主要数据清理方法的基础上，提出了一套完整实用的数据收集与清理模型。

关键词： 业务规则 数据收集 数据整合 数据库规则 数据清理模型

1　引言

在软件项目实施过程中，数据收集与数据清理工作是和软件设计、程序开发同等重要的一个关键环节。对于系统开发商来说，这项工作的成败直接决定着系统能否获得客户的认可，能否顺利通过系统验收，因为，对系统进行验收的人员往往是企业业务上的专家，他们不太了解计算机软件技术，但是对业务数据十分敏感，一旦业务数据出问题，小则导致验收推迟，大则导致系统被否决。另外，数据的收集与清理工作的效率，也直接影响到系统的进展。对于使用系统的企业，业务数据的质量直接关系到企业效益的提高。因此，采用科学的数据清理方法，建设高效的数据清理系统是十分重要的。

经查阅，现有资料主要是关于数据清理方面的文献，注重于对不符合条件的数据如何进行清理的研究，而对清理出数据后，如何进行高效处理讨论的比较少。本文在总结常用的数据清理的方法的基础上，提出一个基于三层模式的实用的数据收集与清理模型。

2　数据清理的方法

2.1 基于业务规则[1]的数据清理

业务规则是指符合业务的某一数值范围、一个有效值的集合，或者是指某一种数据模式，如地址或日期。业务规则根据其规则内容，可分成通用业务规则和特定业务规则。

通用业务规则是指对多数信息系统都适用的业务规则，如年龄不能为负，如果在数据表中年龄为负值，则表示无效的年龄。

特定业务规则是指针对某一种特定行业信息系统的业务规则，如在 ERP 的库存管理系统中，物料的出库日期要比入库日期早，这种规则就属于特定业务规则。

2.2 基于数据库规则[2]的数据清理

一般来说，数据清理是将数据库精简以除去重复记录，并使剩余部分转换成标准可接收格式的过程。数据清理标准模型是将数据输入到数据清理处理器，通过一系列步骤“清理”数据，然后以期望的格式输出清理过的数据。数据清理从数据的准确性、完整性、一致性、惟一性、适时性、有效性几个方面来处理数据的丢失值、越界值、不一致代码、重复数据等问题。

数据清理一般针对具体应用，因而难以归纳统一的方法和步骤，但是根据数据不同可以给出相应的数据清理方法。解决不完整数据(即值缺失)的方法大多数情况下，缺失的值必须手工填入(即手工清理)。当然，某些缺失值可以从本数据源或其它数据源推导出来，这就可以用平均值、最大值、最小值或更为复杂的概率估计代替缺失的值，从而达到清理的目的。

2.2.1 错误值的检测及解决方法

用统计分析[3]的方法识别可能的错误值或异常值，如偏差分析、识别不遵守分布或回归方程的值，也可以用简单规则库(常识性规则、业务特定规则等)检查数据值，或使用不同属性间的约束、外部的数据来检测和清理数据。

① 收稿时间:2009-03-10

2.2.2 重复记录的检测及消除方法

数据库中属性值相同的记录被认为是重复记录，通过判断记录间的属性值是否相等来检测记录是否相等，相等的记录合并为一条记录。

2.2.3 不一致性(数据源内部及数据源之间)的检测及解决方法

从多数据源集成的数据可能有语义冲突，可定义完整性约束用于检测不一致性，也可通过分析数据发现联系，从而使得数据保持一致。目前开发的数据清理工具大致可分为三类。

(1) 数据迁移(Data migration)工具允许指定简单的转换规则，如：将字符串 gender 替换成 sex。Prism 公司的 Warehouse Manager 是一个流行的工具，就属于这类。

(2) 数据清洗(Data scrubbing) 工具使用领域特有的知识(如邮政地址)对数据作清洗。它们通常采用语法分析和模糊匹配技术完成对多数据源数据的清理。某些工具可以指明源的“相对清洁程度”。工具 Integrity 和 Trillum 属于这一类。

(3) 数据审计(Data auditing) 工具可以通过扫描数据发现规律和联系。因此，这类工具可以看作是数据挖掘工具的变形。

3 实用的数据清理模型

利用上述数据清理方法，可以有效的检测出不合格的数据。然而，数据收集与清理工作的目标不仅是检测出不合格的业务数据，还要对不合格的业务数据进行再收集，再清理，以确保这些数据达到业务要求。因此，从数据收集与清理工作的整体来看，清理数据只是数据收集工作中的一项关键任务，它的效率只能在局部影响整体工作的效率。那么，要从整体上提高数据收集与清理工作的质量与效率，必须创建一个合理的数据清理模型，这个模型应当覆盖从数据收集、数据录入、数据清理、不合格数据再清理到正确上传到目标系统的全过程。只有各个环节明确定义、相互协作才能从整体上提高数据收集与清理工作的效率。

图 1 是本文提出的一个数据收集与清理的模型。根据软件工程的定义，这个模型把软件系统与操作人员都看成系统的一个部分。它把数据收集与清理的整个任务合理分配给系统和人，人员主要分为规则定义人员和数据清理人员，系统中的人主要负责规则定义与执行清理任务上传清理数据；软件系统主要通过三层模式等方式，提供多人操作的平台，利用计算机的运算速度优势，执行既定的清理规则，快速反馈清理结果。这样系统的各部分相互配合，周而复始，直到所有数据都达到目标系统的数据要求为止。利用这个模型可以达到从整体上提高数据收集与清理工作效率的目的。

此模型中，中间数据库与数据清理平台是模型的核心内容。

3.1 中间数据库

中间数据库是目标系统所需数据的集合，它去掉了目标系统对数据的限制条件。建立中间数据库的主要目的是：

(1) 对不同来源的数据，以相同的格式存储；

(2) 加快数据收集的速度。

3.2 数据清理平台

数据清理平台是对中间数据库中的数据按业务规则、数据库规则进行清理，把符合条件的数据提供给目标系统，把不符合条件的数据及时反馈给数据清理人员进行再处理。因此，数据清理平台应当以 B/S 模式开发，这样有利于多个数据清理人员同时进行工作，从而提高数据收集与清理工作的效率。

3.3 数据提取平台

数据提取平台是一个功能相对简单的模块，主要完成对不同来源的数据进行整合，上传到中间数据库的功能。它除了对数据整合上传之外，不做任何效验工作，这也是为了提高数据清理工作的效率。

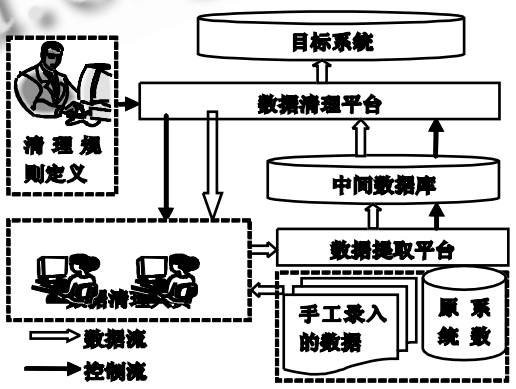

图 1 实用的数据清理模型

3.4 数据清理人员定义

按软件工程中的定义使用与维护系统的人员也是系统的重要组成部分。因此，对数据清理模型中人员

职责的定义也是很重要的。在数据清理工作中，人员的职责主要分为定义数据清理的规则、执行数据清理。

负责定义数据清理规则的人员主要对数据清理平台中的规则进行定义，从而清理出不符合条件的数据。

负责执行数据清理的人员主要完成通过数据提取平台，上传业务数据，并处理清理平台反馈回来的不合格的数据。

3.5 模型中的信息流

数据收集模型主要是通过信息的流转完成数据收集与清理工作的。模型中的信息流主要分为数据流与控制流。

数据流是收集来的原始数据在模型中各个环节流转的过程，最终被清理合格并上传到目标系统。其流转过程如图 2 所示。

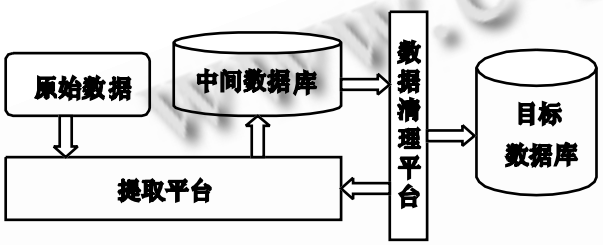

图 2　数据清理模型中的数据流

通过数据流的流转，数据从原始数据转换为合格的业务数据，数据质量得到了提高。这一流程的关键因素是定义中间数据库中数据项与数据格式，它是目标数据中所需数据的集中体现，是数据清理平台的工作的数据源。经过数据清理平台的严格筛选，一部分数据被提交给目标数据库，一部分数据转入再收集的流程。

控制流对数据清理工作进行控制，从而加速数据的再清理过程。它是通过在数据提取平台、中间数据库与数据清理平台三个环节，加入数据清理相关信息的方法实现的。其过程如图 3 所示。

控制流的关键因素是要详细定义数据清理的组织信息与人员信息，使系统快速定位到数据提供者。其原理主要是依据把大量信息收集的工作，快速准确分配到一线工作人员，从而提高数据收集与清理的速度。在实际工作中，数据收集与清理往往面对的是海量的数据，但是，一线工作人员的配置，能够覆盖到每一个数据点，因此，只有通过系统快速把数据收集与清理任务下发到一线人员，才能切实提高工作效率。

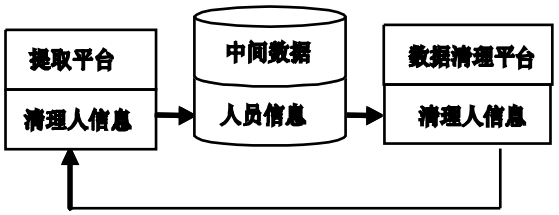

图 3　数据清理模型中的控制流

4　应用实例

供电公司为提高自身的管理水平，采用德国的 SAP 套装软件搭建企业的 ECM(客户关系管理)系统。套装软件对数据质量要求很高，现有系统中的 17 万用户的数据，只有小部分数据可以转换到 SAP 系统中，其中大部分数据需要从现场收集。这样加强了系统数据收集工作的难度。同时，任务重时间紧，根据总公司要求，项目一年零六个月的时间进度。这就要求整个 SAP 项目的实施，从需求确定、软件开发与部署、数据收集与清理各个阶段压缩工期。

在这样的背景下，经过数据清理组的认真分析，最终按上述数据收集与清理模型，搭建了数据清理系统，成立了数据清理的组织。系统中各部分紧密合作，从而使数据收集工作仅用了六个月的时间，就收集并核实了 17 万户用户数据，数据质量 99%以上合格，从而创造了 SAP 系统上线实施用时最短、数据最准确的记录。

5　结语

数据收集与清理是软件工程中关键的工作。此项工作的成败，对供需双方都有重大的意义。因此，本文在分析了主要的数据清理方式的基础上，提出了实用强的数据收集与清理模型。该方法具有简单、易用、清理准确度高等优点。这种方法的清理效果取决于对具体业务的分析以及定义规则的数目。这种方法具有普遍的适用性，尤其适合于大量数据的收集与清理工作。

参考文献

1 陈永府,杨小献,黄正东,陈立平.基于规则的数据收集研究.计算机工程与设计, 2007,(1):158－161.

2 陈伟,陈耿,朱文明,王昊.基于业务规则的错误数据清理方法.计算机工程与应用, 2005,(14):172－174.

3 薛小平,张思东,王小平,曹晓宁.RFID 网络的数据清理技术.计算机工程, 2008,(7):92－94.