可扩展数据清理软件平台的研究

陈 伟[1]，丁秋林[2]

(1. 南京审计学院信息科学学院 南京 210029; 2. 南京航空航天大学计算机应用研究所 南京 210016)

【摘要】提出一种可扩展的数据清理软件平台，该软件平台具有开放的规则库和算法库，规则库用来存放清理规则，算法库用来存放清理算法，算法库中包含多种算法，并可对其扩展；通过在规则库中定义清理规则以及从算法库中选择合适的清理算法，可使该软件平台适用于不同的数据源，从而使其具有较强的通用性和适应性；通过多种算法的清理，提高了数据清理的综合效果。最后，通过实例验证了该平台的效果及可行性。

关 键 词 数据清理; 软件平台; 规则库; 算法库
中图分类号 TP311.52 文献标识码 A

Study on the Extensible Data Cleaning Software Platform

CHEN Wei[1]，DING Qiu-lin[1]

(1. School of Information Science, Nanjing Audit University Nanjing 210029;
2. Computer Application Institute, Nanjing University of Aeronautics and Astronautics Nanjing 210016)

Abstract An extensible data cleaning software platform is proposed, which has open rules library and algorithms library. Rules library is used to store rules and algorithms library is used to store algorithms. Algorithms library has many algorithms and can be extended. Through defining rules in rules library and choosing proper cleaning algorithms from algorithms library, the software platform can be used to various data sources, which makes it universal and adaptive. The synthetic result is improved through data cleaning with many algorithms. Finally, the effect and feasibility of this extensible data cleaning software platform is proved through an example.

Key words data cleaning; software platform; rules library; algorithms library

由于信息化技术的运用，企业内部积累了大量的电子数据。这些数据对企业来说非常重要，但由于各种原因，导致企业现有系统数据库中存在这样或那样的脏数据，主要表现为重复记录、错误数据、不完整数据等[1-2]，若不进行清理，脏数据会扭曲企业从电子数据中获得的信息，影响企业信息系统的运行效果，并对企业构建数据仓库、建立决策支持系统、应用商务智能带来隐患。为了使企业信息系统中的数据更准确、一致，数据清理就显得很重要。目前，国内外已提出了一些有效的清理算法，或根据某种算法开发出一些各具特点的应用系统，或开发一些针对特定应用领域的清理软件。但是，由于数据清理的复杂性，对不同的数据源，要求数据清理适应不同的数据类型、数据数量以及具体业务。一种数据清理算法无论采用多有效的措施，不可能在所有问题上都表现出好的清理效果，即不可能依靠一种或少数几种算法普遍良好地解决各种数据清理问题。有必要提供一种包含一系列数据清理算法以及辅助算法的数据清理平台，为不同背景下的数据清理提供清理方法和清理算法方面的支持。

本文提出一种可扩展的数据清理软件平台，该软件平台具有开放的规则库和算法库，并提供大量的数据清理以及其他辅助算法。当对数据源进行清理时，根据具体业务，通过预定义清理规则和选择合适的算法，清理数据源中的种种错误，具有较强的通用性和适应性，可提高数据清理的综合效果。

1 数据清理原理

简单地讲，数据清理就是从数据源中清除错误数值和重复记录等。即利用有关技术如数理统计、数据挖掘或预定义清理规则等，从数据源中检测和消除错误数据、不一致数据和重复数据，从而提高数据的质

收稿日期：2003 – 04 – 29
作者简介：陈 伟(1976 –)，男，博士，讲师，主要从事数据清理方面的研究.

量。一般来说，数据清理包括以下步骤[2]：

(1) 数据分析

数据分析是指从数据中发现控制数据的一般规则，比如字段域、业务规则等。通过对数据的分析，可定义出数据清理的规则，并选择合适的清理算法。

(2) 数据检测

数据检测是指根据预定义的清理规则及相关数据清理算法，检测数据是否正确，如是否满足字段域、业务规则，或检测记录是否重复等。

(3) 数据修正

数据修正是指手工或自动地修正检测到的错误数据或处理重复记录等。

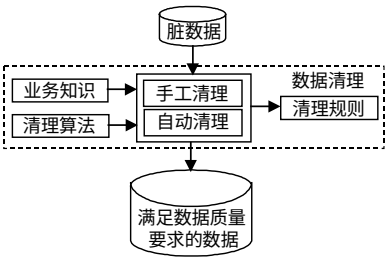

图1　数据清理原理

数据清理的原理如图1所示[3]。

2　可扩展数据清理软件平台的原理

2.1　软件平台的功能及清理方法

数据清理软件平台主要具有以下功能：

(1) 清理重复数据记录[4]

对于完全重复记录，采用“排序→比较”的检测方法，先将数据库中的记录排序，然后通过比较邻近记录是否相同来检测完全重复记录；对于相似重复记录，清理方法是记录排序→记录相似性检测→合并相似重复记录。

(2) 清理错误数据

通过在规则库中预定义规则来检测数据是否满足属性域、业务规则等，从而检测出错误数据，清理效果取决于对业务的分析以及定义规则的数目。

(3) 清理不完整数据[5]

不完整数据的清理可以总结为首先采用记录可用性检测算法检测记录的可用性，其次删除不可用的记录，然后对可用记录采用回归、判定树归纳等算法预测可能值来填充，也可人工处理。

2.2　软件平台的工作原理

(1) 数据分析

分析所要清理的数据源，定义数据清理的规则，并选择合适的清理算法，使其能更好地适应所要清理的数据源。

(2) 数据清理

把数据源中需要清理的数据通过Java数据库连接(Java Data Base Connectivity，JDBC)接口调入软件平台，调用算法库中的相应算法对数据源进行预处理标准化数据记录格式，并根据预定义的规则，把数据记录中的相应字段转化成同一格式。然后，按照对数据源的分析，分步执行数据清理。清理过程一般为首先清理错误数据，然后清理相似重复记录，最后清理不完整数据。

(3) 清理结果检验

数据清理运行结束后，在系统窗口中显示出数据清理结果，根据清理结果和警告信息，手工清理不符合系统预定义规则的数据，处理未清理的数据，从而完成系统的数据清理。另外，通过查看数据清理日志，检验数据清理的正确性，对清理错误进行修正。

可扩展数据清理软件平台的工作原理如图2所示。

2.3　规则库与算法库

从图2可以看出，规则库与算法库是可扩展数据清理软件平台的核心。

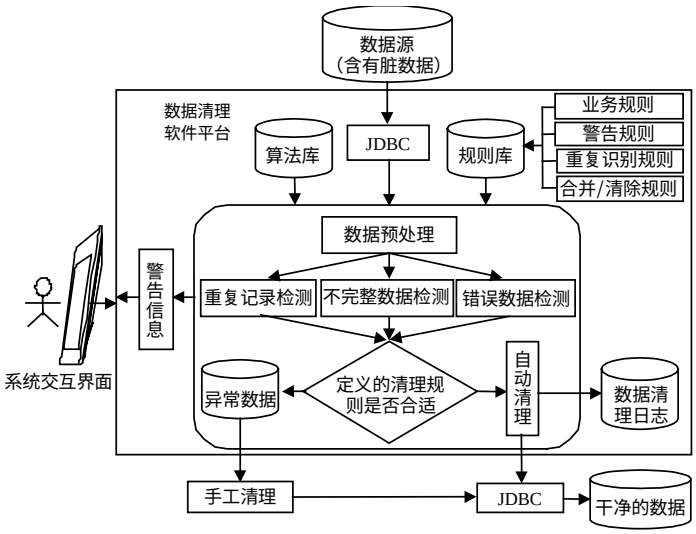

图2 可扩展数据清理软件平台的工作原理

规则库用来存放关于数据清理的如下规则：

(1) 业务规则

业务规则是指符合业务的某一数值范围或某一有效值的集合，或者是指某种模式，如地址或日期。业务规则能帮助检测数据中的例外情况，比如违反属性依赖的值、超出范围的值等。

(2) 重复识别规则

指定两条记录为相似记录的条件，比如距离的阈值δ。

(3) 合并/清除规则

指定对两条重复记录如何进行处理。对于一组所检测出的相似重复记录有两种处理方法：一种是把一条记录看成是正确的，其他记录则看成是含有错误信息的重复记录；另一种是把每一条重复记录看成是数据源的一部分，目的是合并重复记录，产生一条具有更完整信息的新记录。相似重复记录的处理由用户根据具体的业务分析在规则库中预定义合并/清除规则来完成。

(4) 警告规则

指定对特殊事件的处理规则及相应提示信息。

可根据具体的业务，在规则库中定义相应的规则，或者修改已有的规则，从而使可扩展数据清理软件平台适用于不同的数据源，具有较强的通用性和适应性。

算法库用来存放数据清理所需要的算法。多种数据清理算法通过Java程序实现后，以类的形式存放在算法库中，供数据清理时根据不同的情况来调用相应的合适算法。通过选择相应的清理算法多次对数据源进行清理，可提高数据清理的综合效果。另外，在算法库中可不断扩充新的数据清理算法，供数据清理时选用。

2.4 可扩展数据清理软件平台的实现

可扩展数据清理软件平台采用Jbuilder8开发，通过JDBC接口可以从Oracle、Sybase等任何关系型数据库中获取数据。算法库就是存放数据清理算法的类库，算法采用Java编程实现，供数据清理时调用，所需的新算法可通过Java编程实现后扩充到算法库中。规则库的实现可分成以下两种方式：

(1) 通过规则语言实现，规则语言一般采用IF-THEN规则，这种方式主要用手业务规则、警告规则等清理规则。

(2) 通过在数据库中创建一个数据表实现，这种方式主要用于重复识别规则、不完整识别规则、错误识别规则等清理规则。

软件平台提供数据清理规则定义界面，供数据清理时根据具体的业务分析，定义或修改规则库中的数据清理规则。

2.5 可扩展数据清理软件平台的特点

(1) 该软件平台包含多种数据清理算法，在对数据源进行数据清理之前，通过预定义清理规则和选择合

适的算法，使该软件平台具有较强的通用性和适应性。

(2) 在数据清理过程中，人工的交互是必要的，因为很多错误是不可预料的，不可能所有错误都被自动清理。在数据清理过程中，如果出现异常错误，通过该软件平台所定义的警告规则，系统会给出相应的警告信息，提示用户手工处理，具有交互性。

(3) 该软件平台具有图形界面，可用来显示数据清理过程和清理结果，很方便地和数据清理过程进行交互，具有直观性。

(4) 由于该软件平台具有开放的规则库和算法库，可不断扩充新的数据清理算法，并可根据具体业务定义所需的清理规则，具有很好的可扩展性。

3 实　例

本文所提出的数据清理软件平台在某医疗保险信息系统工程项目中得到应用，其关键清理步骤简要说明如下：

(1) 错误数据清理

对于错误数据的清理，采取的方式是通过在规则库中定义规则来检测数据是否满足属性域、业务规则，从而发现错误数据。对该医疗保险信息系统业务的分析，在规则库中定义了近百条业务规则，很好地完成了错误数据的清理工作。

(2) 重复数据记录清理

重复数据记录清理的有效性主要有查全率R(Recall)和查准率P(Precision)两个度量标准，其中：

$$R=\frac{\text{正确识别出的重复记录数}}{\text{实际的重复记录数}}$$

$$P=\frac{\text{正确识别出的重复记录数}}{\text{识别出的重复记录数}}$$

以下以该医疗保险信息系统数据库的“参保单位”数据表中的数据为例，说明数据清理过程。

在进行相似重复记录清理之前，首先，要确定各参数的取值，并在规则库中进行定义；其次，运行相似重复记录的检测工作；最后，在完成相似重复记录的检测后，与人工检测结果进行比较，结果为：

$$R=98.4\%$$

$$P=98.1\%$$

对于从“参保单位”数据表中检测出的相似重复记录，采取“完整规则”来清除，即从一组相似重复记录中选择最完整的记录，删除其他记录。从检测结果可以看出，该数据清理软件平台能较好地完成对相似重复记录的清理工作。

4 总　结

数据清理是企业信息化建设中的一项重要任务，但由于数据清理的复杂性，数据清理具有较大的难度。本文根据数据清理的研究现状，提出了一种可扩展的数据清理软件平台。通过原理分析及实际应用，可以看出该软件平台具有较强的通用性和适应性，具有较好的清理效果。随着研究的进展，将进一步完善该软件平台，不断增加清理功能。

参　考　文　献

[1] Galhardas H, Florescu D, Shasha D. Declarative data cleaning: language, model, and algorithms[C]. In: Proceedings of the 27th VLDB Conference, Roma Morgan Kaufmann, 2001: 371-380

[2] Rahm E, Do H H. Data cleaning: problems and current approaches [J]. IEEE Data Engineer Bulletin, 2000, 23(4): 3-13

[3] Lee M L, Ling T W, Low W L. IntelliClean: a knowledge-based intelligent data cleaner[C]. In: Proceeding of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston: ACM Press, 2000: 290-294

[4] Monge A E. Matching algorithms within a duplicate detection system [J]. IEEE Data Engineer Bulletin, 2000, 23(4): 14-20

[5] 陈　伟, 丁秋林. 数据清理中不完整数据的清理方法[J]. 微型机与应用, 2005, 24(2): 44-45, 55

编　辑　熊思亮