基于规则的土壤数据校验模型研究与实现①

张　仁[1]　沈志宏[1]　黎建辉[1]　施建平[2]　(1 中国科学院计算机网络信息中心 北京 100190;
2 中国科学院南京土壤所 江苏 南京 210008)

摘　要： 数据校验是数据挖掘与知识发现中的重要一环。我国土壤观测数据由于台站观测人员上网条件、观测地记录不便以及需要适当的数据预处理等原因，无法实行在线入库，一般借助于 Excel 等软件来记录中间结果，再提交土壤分中心，这样的记录过程经常引入不必要的错误。提出了一个基于可定制规则库的土壤数据校验模型。模型主要包括数据格式转换模块、权限管理模块、元数据管理模块、重复记录去除模块、数据校验模块及规则定制与解析模块。低侵入式的轻量级设计，使得在大大减轻数据校验人员工作量的情况下，原有的数据填报流程不需要改变。可定制规则使得模型易于扩展。

关键词： 数据校验;基于规则;知识发现;数据挖掘;科学数据

Research and Implementation of Rule-Based Data Cleaning Model for Soil Data

ZHANG Ren[1], SHEN Zhi-Hong[1], LI Jian-Hui[1], SHI Jian-Ping[2]
(1.Computer Network & Information Center, Chinese Academy of Sciences, Beijing 100190, China;
2.The Institute of Soil Science, Chinese Academy of Sciences, Nanking 210008, China)

Abstract: Data validation is one of the most important phases in KDD (Knowledge Discovery and Data Mining). Since Internet and computer are unavailable in some observation station and data preprocessing is necessary, most soil observation data in our country could not be included in database online. Most of the data are stored and preprocessed by software like Microsoft Excel before they are reported to Soil Sub-Center. These steps often lead to some uncxpected errors. We present a customizable rule based model in this paper. The model consists of several modules: Data format transformation module, Privilege management module, Metadata management module, Record De-duplication module, Data Cleansing module and Rule customization & parser module. Low-invasive and light-weight design make the model validatc data successfully while without affecting the old data entry system. At the same time, Customizable Rule makes the model much easier to extend.

Keywords: real-time VBR video; smoothing algorithm; funnel; the shortest path; sliding window

在信息化的今天，数据仓库、数据挖掘和知识发现等在科研、商业、军事以及教育等等领域都逐渐地发挥着越来越大的作用。在知识发现的过程中，根据“进去是垃圾,出来也是垃圾(garbage in, garbage out)”这条原理，使得保障数据质量成为知识发现中的一个重要方面。而数据校验作为数据进行数据仓库的大门把关环节，正是为了保证在知识发现中所用到数据的质量的重要一环。 研究表明，超过 80%的研究人员在进行数据清洗的相关项目研究时，超过 40%的时间花在了数据预处理上。在研究数据清洗方面[1,2]，研究人员的主要研究重点，也都放在了模式层的转换集成上，而对于实例层次上的数据质量问题，因其数据的具体领域相关性，所受到的关注并不多。然而，通过分析，仍然能够找出一些实例层次上的数据的共性。有些

① 基金项目:中国科学院“十一五”专项项目;中国科学院知识创新工程重要方向项目 (KZCX2-YW-433-03)
收稿时间:2009-11-24;收到修改稿时间:2009-12-29

研究人员针对实例层的数据质量问题，也提出了相应的算法与模型。如文献[3]提出了一种可交互的数据清洗系统，文献[4]描述了一种基于本体的清洗方法等等。我国土壤观测数据由于台站观测人员上网条件、观测地记录不便以及需要适当的数据预处理等原因，无法实行在线入库，一般借助于 Excel 等软件来记录中间结果，再提交网络、存储条件完好的数据中心。对于观测台站上报的土壤观测数据，中国生态系统研究网络(CERN)土壤分中心目前仅依靠人工凭经验完成大批量数据的检验，这样的操作难免存在着遗漏，且发现问题常常滞后一年。此外，由于缺乏对数据有效性检验的系统理论和方法研究，同样存在着人员变动带来的数据校验不一致性。为解决分中心层次的土壤监测数据上报和入库过程中的质量控制问题，需研究数据校验模型并开发快速数据校验计算机辅助工具。建立与质量控制相关的背景数据库和参考值数据库，应用多年人工数据校验积累的经验进行规则定制，解决异常数据的判别和标识异常数据原因等关键技术问题。根据这种情况，本文结合土壤观测数据，提出了一种简单易用，基于可定制规则库的数据校验模型。该模型能很好地代替数据校验人员的手工查错工作，并且不影响原有的数据填报流程。

本文先对系统的总体架构及各个模块的功能进行了介绍，然后介绍了系统实现及其关键技术。

1 数据校验技术

1.1 数据校验基本概念

数据校验是数据清洗中的第一个环节，数据清洗处理检测和消除数据中的错误，以达到提高数据质量的目的[5]。在数据校验之后，根据规则进行数据修改，便可达到数据清洗的目的。

数据清洗按数据源有单数据源和多数据源两种情况：单数据源时，分为单字段的错误(包括拼写，日期类型错误，引用类型错误，邮政编码不存在，邮政编码与城市不对应等等)、字段之间关联错误(如年龄不等于现在的日期减去出生日期)和记录重复错误(由于姓名地名及专门名词等是否缩写导致)等，在数据入口时，可以对模式层添加约束，来减少这类错误；多数据源中的数据错误则在单源情形的基础上，增加了与数据集成相关的数据质量问题，比如模式的集成与转换、数据格式的表示不一致、重复实体检测与消除等。

按数据质量问题所处的层次可分为模式(Schema Level)层数据质量问题和实体层(Instance Level)数据质量问题。

1.2 现有数据校验工具

目前国内外，存在着很多形式的数据校验与清洗工具，如针对专有的数据格式的校验工具 MD5 值校验工具、Cmis30 数据校验工具、CRC 数据校验计算工具等；针对 web 表单数据的 java 开源包 validator 等；通用的商业软件数据清洗工具也一般都具有校验功能，如 WinPure Clean & Match、matchIT Data Cleansing Software Suite。

此外还有一些数据清洗研究人员提出的原型系统一般也包括相应的异常数据检测模块，如可交互的数据清洗系统 Potter’s Wheel[3]，它用 C++及 java 实现了这一系统，该系统的主要功能在处理数据格式的转换(字段的拆组等)上，与所需要的数据校验功能并不十分吻合。OntoClean[4]是一种基于本体的数据清洗框架，基于本体的方法可以在知识层而不仅仅是数据层发现数据质量问题，但这一方法要求有清洗的数据领域的本体描述。文献[6]描述了基于规则引擎的数据清洗方法，实验结果表明利用规则引擎比硬编码更加高效，对于本模型的设计有一定借鉴意义。同时，也有许多的商业软件，用于数据清洗的不同方面。如清洗邮编，姓名，地址等的错误和重复。

然而，这些系统与模型要么过于复杂，要么功能过于单一，有时，没有数据清洗相关知识的专业人员，无法使用这些系统，而熟悉数据清洗知识的编程人员，则没有领域知识。针对这些情况，本文提出了一种基于可定制规则的简单数据校验模型。

2 基于可定制规则的数据校验模型

本文针对土壤数据校验过程中存在的问题，提出了一种基于可定制规则的简单数据校验模型，该模型简单易用，适合于多个层次的用户使用及扩展。

对于普通用户，可用系统提供的图形界面的方式配置规则文件；而有 xml 基本知识的用户，则可借助于规则描述文档，仿照 xml 文件的格式进行直接编辑；有 java 编程知识的用户，还可以实现自己的规则类，扩展规则类型。

2.1 模型流程

该模型由数据格式转换，重复记录去除，规则定制与解析，数据校验，结果展示与交互等模块组成，系统总体流程图如图 1 所示。

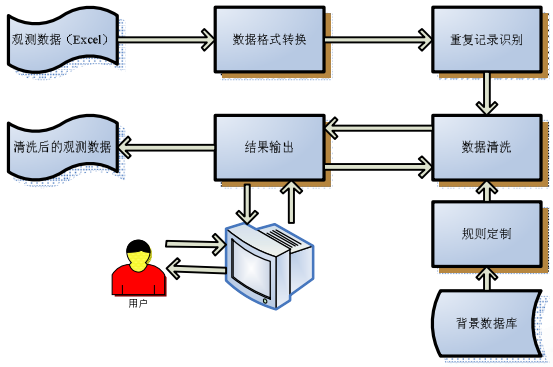

图 1　系统流程图

2.2 功能模块

(1)数据格式转换：用户选择需要校验的数据表项，然后载入对应的 Excel 数据表，由数据格式转换模块将其封装为 Java 的数据对象列表(DataObject List)，在数据转换的过程中，还可以发现数据类型的错误，转换过程如图 2 所示。

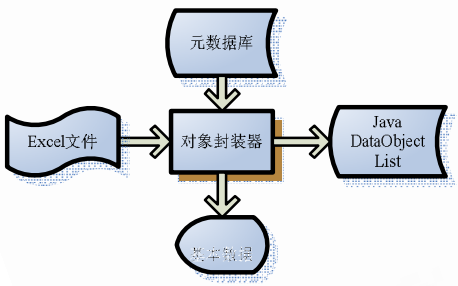

图 2　数据格式转换

(2)重复记录识别：根据编辑距离及基于距离算法，找出相似度高的记录，供用户判断是否存在重复录入的情况。

(3)规则定制：用户定制校验规则，规则定制中，有些值是来自于历年观测数据的均值、最大值最小值、年变异以及空间变异等统计值或者背景数据库中的台站标识、观测方法等其它知识，用户可能根据自身对计算机语言的理解能力，采取不同层次的定制，如：借助于规则编辑界面对已有规则进行可视化编辑，直接编辑 XML 文件进行规则配置，扩展 DataValidationTool 类进行校验规则中所使用到的功能以及实现 AbstractRule 类，实现用户所需的规则，扩充原有的规则库。

(4)数据校验：根据用户自定义的规则，对目标数据表进行校验，并且将校验结果记录到每一个 DataObject 对象的 errorList 中。

(5)数据输出：用户可根据校验结果，选择是否将结果输出，输出有两种方式，即以 Excel 表的方式导出加上了数据校验结果的 Excel 文件和将数据输出到数据库中长期保存。在数据输出之前，用户可以先对规则进行编辑之后，再进行校验，直至结果满意为止。

3　模型实现

规则引擎的框架设计如图 3 所示，主要包括以下几个关键技术环节。

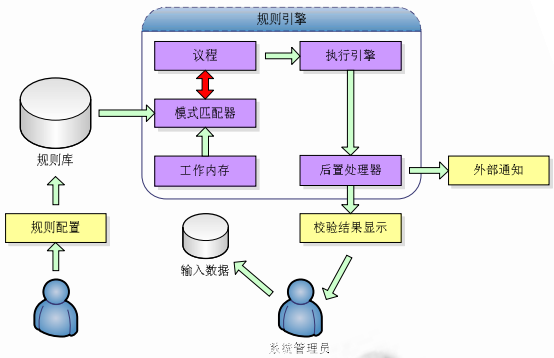

图 3　规则引擎框架图

(1)规则定义与描述：在设计采用上采用规则定义语言(RDL，Rule Definition Language)，规则定义语言 RDL 是在模板语言 VTL(Velocity Template Language)的基础上，进行二次开发改造的。VTL 是一个开放源吗的模版引擎，由 apache.org 小组负责开发，Velocity 现在最新的版本是 Velocity1.3.1，http://jakarta.apache.org/velocity/index.html?可以了解 Velocity 的最新信息。Velocity 允许我们在模版中设定变量，然后在运行时，动态的将数据插入到模版中，替换这些变量。规则采用 XML 进行规则库的形式化描述，以保证规则引擎能够适用于不同的应用环境。

规则定制采用了 xml 语言描述，使用开源框架 JDOM 进行解析，注入到程序中，这样便可实现规则的灵活定制。规则定义的 xml 示例代码如下：

```
//rules.xml
<?xml version="1.0" encoding = "gb2312" ?>
<ruleset>
    <rule>
        <title>表土养份标准差检验</title>

    <name>OM_SURFACE_TIME_STDEV</name>
        <description>标准差年变异</description>
        <suggestion>样地数据标准差>2 倍多年样地平均标准差，数据有问题</suggestion>
        <expression>!$record.getStDev($OM)  ||
$record.getStDev($OM)                    le
2*$tool.getTimeStDev($record,$OM)</expression>
        <params>
            <param>
                <title>OM</title>
                <name>OM</name>
                <value>OM</value>
                <type>TEXT</type>
            </param>
        </params>
    </rule>
    ……
</ruleset>
```

规则描述中的$为 Velocity 中的变量引用语法，record 及 tool 变量，在程序初始化时生成。

(2)规则解析：规则解析借助于 Velocity 引擎中的 Velocity 类，由模式串解析后返回的字符串 true 或 false 来决定记录是否符合模式串描述的规则。后续程序会将结果记录到每一个 DataObject 的 ErrorList 中。

(3)背景知识库：由专家定义数据格式，阈值，可用于填充缺失值的某些值，重复记录度量规则，数据规则定制中引用的某些值也来源于背景知识库等。背景知识库除由土壤领域的专家制定的参考值外，也有一些来源于多年土壤观测数据的统计值，如均值，中位数，2 倍标准差，离群值检验，空间变异，年变异的统计值等。同时保存了土壤观测的领域背景知识及遍布我国的各个土壤数据观测台站基本信息，如台站标识，观测方法，分析方法，标准样品测定值等。

已积累近 10 年的观察数据、样地背景数据、初步建立了分析数据质量评估方法，为数据质量检验评估打下良好的基础。

(4)规则分发：由于规则是采用 XML 语言进行描述的，因此，可在各个校验系统中自由分发。在台站上报数据之前，也可在台站部署相应系统，规则在台站与综合中心及各分中心之间进行自由分发。系统部署与规则分发如图 4 所示：

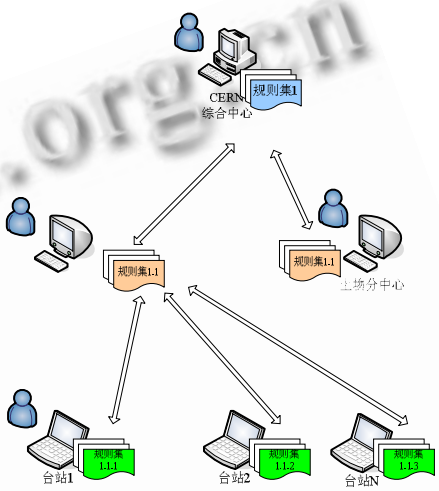

图 4　系统部署与规则分发

4　结语

应用表明，本文所提出的数据校验模型能很好地代替数据校验人员的手工查错工作，并且不影响原有的数据填报流程。简单易用，便于日后扩充。系统运行的界面如图 5 所示。

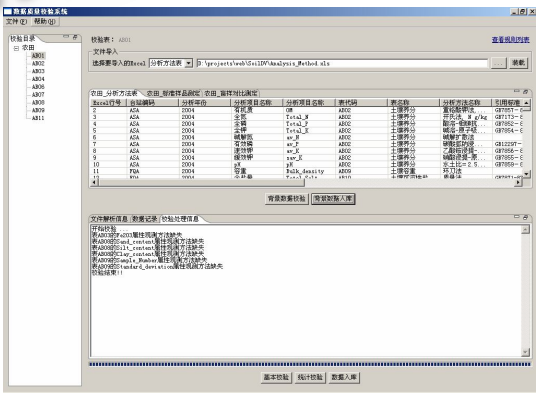

图 5　数据校验系统运行界面

本文所提出的数据校验模型虽然主要是针对土壤

(下转第 65 页)

(上接第 81 页)

数据的校验，但其灵活的规则定制方式，可以应用于任何关系型数据库管理系统(DBMS)中的任何关系表。对于其它领域的数据校验，仍有借鉴意义。该数据校验模型，将与中科院计算机网络信息中心科学数据中心拥有自主软件著作权的软件——可视化关系数据库管理发布系统 VisualDB (http://vdb.csdb.cn)进行整合，为 VisualDB 提供通用的数据清洗与数据校验功能，同时可定制规则还可作为其它系统构建复杂查询条件的组成部分。

参考文献

1 Maletic JI. A Marcus Data Cleansing: Beyond Integrity Analysis. Proc. of the Conference on Information Quality.2000. 200 − 209.

2 郭志懋,周傲英.数据质量和数据清洗研究综述.软件学报, 2002,13(11):2076 − 2082.

3 Raman V, Hellerstein J. Potter's wheel: an interactive data cleaning system. In: Proceedings of the 27th International Conference on Very Large Data Bases. 2001.381 − 390.

4 Wang X, Hamilton HJ, Yashu Bither. An Ontology-Based Approach to Data Cleaning. Technical Report. 2005.

5 Rahm E, Do HH. Data cleaning: problems and current approaches. IEEE Data Engineering Bulletin. 2000.23(4):3 − 13.

6 叶舟,王东.基于规则引擎的数据清洗.计算机工程, 2006,32(23):52 − 54.