水文数据库数据质量控制与管理应用研究

余宇峰[1],张建新[2],朱跃龙[1],万定生[1]

(1.河海大学计算机与信息学院,江苏 南京 210098;2.水利部水文局,北京 100053)

摘 要:数据质量问题已经成为水文信息化过程成败的重要影响因素。本文以水文数据库为背景,分析了水文数据质量评估关键指标维度、数据质量问题的来源及其分类,建立了水文数据库数据质量评估模型和控制模型,并从实践角度探讨了数据质量控制和改进的若干方法。

关键词:数据质量;维度;质量评估;质量改进

中图分类号:TP309.2 **文献标识码**:A **文章编号**:1000-0852(2013)03-0065-04

1 引言

当今社会是一个高速发展的信息社会,信息社会的核心基础是海量的、并不断快速增加的数据。庞大的数据是否具有存在价值和应用价值则完全取决于数据质量。

众所周知,数据也是企业重要的资产。随着信息化程度越来越高,数据量越来越大,低质量的数据会导致业务流程阻塞、成本增加甚至决策困难等一系列的严重问题。Card & Payments 的一项分析报告表明,每年仅仅因为错误或重复的客户信息就使企业付出了6 000 多亿美元的成本[1]。低质量的数据往往导致错误的决策,因此,越来越多的部门在进行信息系统建设时开始重视数据质量问题。

我国已经建成一系列的数据采集系统即水文站网,收集并积累了大量水文数据,水文数据库建设或水文数据信息化建设相继逐步展开。但由于原始水文数据量庞大,数据处理自动化程度不高,且水文数据处理存在人为差异,水文数据库数据质量参差不齐。

但是,水文数据是贯穿一切水文活动的主线,是水文行业的灵魂所在,水文数据的质量关乎着水文行业自身发展、关乎着涉水工程建设、关乎着国济民生,因此,水文数据的质量控制与管理越来越受到广泛重视。

以国家水文数据库建设规划需求分析,水文数据库数据质量管理对象主要是结构化数据、半结构化数据和非结构化数据,质量管理的内容主要是完整性、连续性、重复性、一致性、合理性和准确性等问题,质量管理的目标是对于数据库数据质量检测综合指标符合技术规范。

综上所述,水文数据库数据质量控制与管理的关键是如何检测到数据质量问题,依据什么样的理论基础和具体方法对问题数据进行清理,最终依据数据质量控制指标对水文数据库数据质量做出评价。

2 水文数据库质量分析

2.1 数据质量定义

数据质量(Data quality,DQ)通常被定义成"适合使用"的数据,即收集的数据能满足特定用户需求[2,3]。也有学者进一步把数据质量定义成一个信息系统中的数据模式和数据实例间相一致程度,并在此基础上实现数据的精确性(accuracy)、完整性(completeness)、有效性(validity)、一致性(consistency)等[4]。

根据数据质量的定义,结合水文数据应用领域的现状及数据收集、处理和分析的历史经验,水文数据库数据质量水平的衡量标准应围绕以下 3 个主要维度进行:

(1)准确性:描述数据是否与实体或属性定义相符合的程度,包括语法准确性(数据是否与域定义的值相对应)和语义准确性(数据是否满足业务逻辑要求)。

(2)完整性:描述为解决问题所获得的数据的广度、

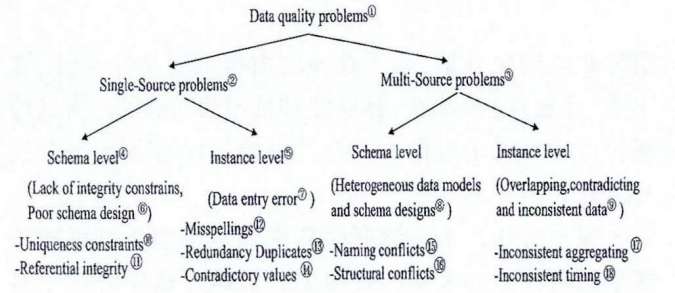
收稿日期:2012-11-02
基金项目:国家自然科学基金项目(51079040);水利部 948 项目(201016)
作者简介:余宇峰(1979-),男,湖北黄冈人,博士研究生,讲师,主要研究方向为数据挖掘、水利信息化。E-mail: yfyu@hhu.edu.cn

深度和规模等的完整程度，包括列完整性、表结构完整性和总体完整性等。

(3)一致性：描述关系表中的元组间或是文件的记录间，数据元素间是否存在的语义规则冲突，包括库表结构、字段逻辑、更新内容及异构数据库数据源间的一致性等。

2.2 质量问题分析

数据质量问题可以从数据来源(单源、多源)和数据模式(模式层、实例层)两个方面划分成 4 类[5]：单源模式层质量问题、单源实例层质量问题、多源模式层质量问题和多源实例层质量问题。图 1 表示了这种分类，并列出了每一类中典型的数据质量问题。



①数据质量问题,②单数据源问题③多数据源问题④模式层⑤实例层⑥缺少完整性约束糟糕的模式设计⑦数据⑧异质数据模型和模式设计⑨冗余、互相矛盾或不一致数据⑩唯一性约束⑪ 引用约束 ⑫拼写错误 ⑬ 相似重复记录 ⑭ 互柜矛后的字段⑮命各冲突 ⑯结构冲突 ⑰不一致的统计维度 ⑱不一致时间选择

图 1 数据质量问题分类

Fig. 1 Classification of data quality problems

数据质量评估是数据质量提高的基础和前提，它能对数据(整体或部分)的质量状况给出一个合理的度量，从而帮助数据用户了解系统的数据质量水平，并采取相应的处理过程来提高数据质量。数据质量评估分为定性和定量方法：定性方法主要依靠评判者的主观判断；定量方法则采用百分制分别计算各质量维度的得分，为人们提供一个系统、客观的数量分析方法。

水文数据质量评估主要采用定量评估方法，通过对数据质量评价框架、评价模型相关文献的研究，结合水文数据质量评价的需求，建立水文数据质量评价模型(见图 2)如下：

$$M=\{P,D,R,W\} \tag{1}$$

式中：P(Perspective)是数据质量评估的角度，水文数据质量评估角度包含两部分：历史数据和系统新产生数据；D(Dimension)是数据质量评估维度，包括准确性，完整性和一致性等三方面；R(Rule)是数据质量评估规则；W(Weight)是不同评估维度和评估规则的权重。

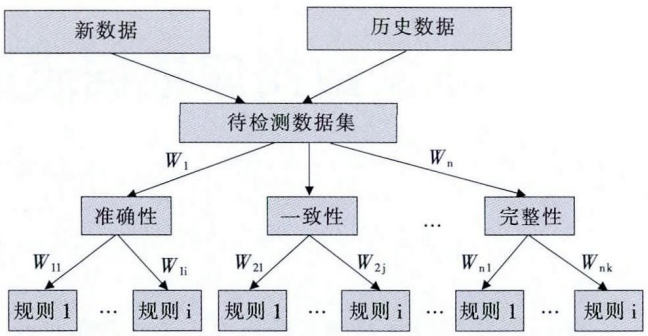


图 2 数据质量评估模型

Fig.2 The model for data quality assessment

定量评估通常采用属性加权算法，以评估分数(即无质量问题的记录数和总记录数之间的比值) 衡量数据质量的高低。通常，第 i 个评估维度的评估分数 S_i 可通过如下公式得到：

$$S_i=\sum_{j=1}^{n}(W_j*\frac{N_j}{M_j})*100 \tag{2}$$

式中：N_j为符合第 i 个评估维度第 j 条评估规则的记录总数；M_j为数据集总记录数；W_j为第 i 个维度第 j 个规则的权重且满足 $\sum_{j=1}^{n}W_j=1$；n 为第 i 个维度的总规则数。

综合单个维度评价结果和各维度的权重，可以计算出数据质量评价综合得分：

$$S=\sum_{i=1}^{k}W_i*S_i \tag{3}$$

式中：k 为待评估数据集的维度总数；W_i 为第 i 个维度的权重，且 $\sum_{i=1}^{k}W_i=1$。

3 水文数据质量控制与方法

3.1 质量控制模型

水文数据质量控制包含两个方面：数据生产过程中的质量保证和质量评估后的质量提高。本文着重讨论质量评估后的质量提高。

数据质量提高主要涉及模式层和实例层两个方面。模式层数据质量提高技术主要侧重理解数据模式及根据已有的数据实例重新设计数据模式；实例层面数据质量提高通常采用数据清洗，通过缺失数据处理、重复对象检测、异常数据检测、逻辑错误检测、不一致数据处理等实现数据质量的改进与提高。水文数据质量控制包含自动和手动两种方式，质量提高流程如图 3 所示。

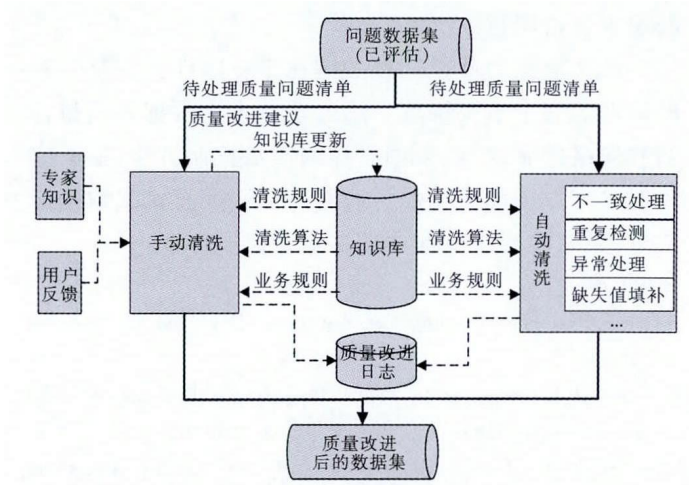


图 3　水文数据质量提高模型

Fig. 3 The model for hydrological data quality improvement

自动质量改进根据质量评估结果，自动从知识库中匹配质量改进处理规则和质量提高算法完成数据质量问题处理。自动质量改进方法能自动完成重复记录校核和清洗、缺失数据填补与清洗、逻辑错误修正、不一致数据处理等质量问题，并将数据质量改进信息记录进日志库。

手动质量改进结合质量评估报告,采用专家评议、用户反馈等人工干预模式，手动完成数据质量改进处理,并将数据质量改进信息记录进日志库。手动质量改进方法完成自动质量改进步骤无法处理的质量问题,根据手动质量改进过程、步骤及专家知识,生成新的质量改进规则和算法并更新质量知识库。

3.2　质量控制方法设计与实现

3.2.1　重复数据处理

现实世界中的同一实体存在不同数据模式下的表征为语法上相同或相似的不同记录，这些相同或相似重复记录检测与清洗是数据质量控制的一个重要环节。常用的重复记录处理方法有:排序 & 合并法、索引法、机器学习法、上下文相关法、专家系统法等。

水文数据中的重复记录主要体现在同一测站检测数据的重复录入上。由于测站的历史沿革和管理归属等问题,同一测站存在多个不同测站名称。采用排序 & 合并法,先对测站名称相似的记录排序,然后根据测站对应的时间属性进行判断。若名称相似的测站在时间序列上有重叠,则比较名称相似的测站在同一时间维度上的施测水文要素属性值。如果比较结果完全相同,则判定为重复记录并予以清洗,否则保留数据项。

3.2.2　缺失数据处理

实际的数据集中经常会存在缺失数据，常常会对分析结果有很大的影响。造成缺失的原因通常是数据采集过程中设备故障或人工遗漏。数据分析过程中忽略缺失数据或使用常量(如“0”)代替,会使分析结果准确性降低。水文行业常用的缺失数据处理方法有:

(1)人工法:根据上下文相关数据源,手工填补缺失数据,或由领域专家估计补全。

(2)均值法:用缺失数据前后数据点的时间序列均值代替。

(3)相关法:用其它相关序列的均值填补。如雨量站的降雨量时间序列的缺失，可以参考临近的多个雨量站在该时间雨量值的均值。

(4)最可能法:通过回归分析、贝叶斯形式化方法工具或判定树推导出可能数据值以填补。

3.2.3　异常数据检测

异常是指和其余数据集不一致的观测值，如水文分析中的极端流量值[6]。水文数据集中异常的存在会影响水文系统设计、运行和管理的决策过程。对异常的处理通常面临两难困境：保留异常值会使得水文序列统计分析变得非常复杂，而简单的处理又会降低预测模型的精度。异常一般是由两种原因造成,一是数据固有变异性造成，另外一种则是由于度量或人工错误导致,在数据清洗时这两者都应予以关注。在数据清洗领域主要采用统计学方法、聚类方法、最近邻法、关联规则法等对异常数据进行检测。

3.2.4　逻辑错误检测

现实数据中往往存在字段取值与实际情况不相符的逻辑错误,如年龄<16 岁 & 婚否=已婚。数据清洗过程中主要根据领域知识建立规则体系来自动处理数据中的逻辑错误。

(1) 通过逻辑约束规则检测单个字段取值的逻辑错误。如利用字段的数据类型、长度、取值范围等约束规则检测逻辑错误数据;

(2) 通过检测字段之间以及记录之间的关系来检测逻辑错误数。通过大量数据分析挖掘字段之间的完整性约束，然后采用函数依赖或特定应用的业务规则来检测以改正逻辑错误数据。

3.2.5　不一致数据处理

多源数据集成过程中，同一实体在不同源数据之间往往会存在重叠交差情况，从而导致不一致数据的产生。不一致数据处理是数据清洗面临的主要问题。水

文数据库质量控制主要采用规则依赖于约束实现不一致数据的处理：

(1)域一致性约束：通过关系的属性列的域一致性约束规则实现不一致性数据的检测与处理。如同一属性字段的值域、计量单位、数据有效位数等域属性在数据加工、传输过程中数据应一致。

(2)包含依赖约束：通过关系的属性列与其它属性列或其他关系的属性列之间的包含关系实现不一致数据检测与处理。如：测站一览表中测站所在行政区划代码必须包含在行政区代码表行政区划码属性中。

(3)函数依赖约束：通过一个关系的多个属性之间或一个关系的属性和其它关系的属性间存在的函数依赖关系实现不一致数据的检测与处理。如：月降水量表中月降水量属性等于日降水表中该月降水量的总和。

4 结论

水文数据库数据质量问题是水文信息化过程中一个关键问题。由于水文数据具有很强的专业知识和领域背景，对水文数据库中的数据进行质量评估和改进的需求日益明显。

水文数据库质量问题关键在于一致性，完整性和准确性等三个主要维度；通过建立水文数据库质量评估和质量控制模型，采用具体的质量控制方法，能够实现对水文数据库数据质量改善，从而改善水文数据库质量，更好地提供专业化的水文信息服务。

参考文献：

[1] Kate Fitzgerald. Weeding Out Bad Data [M]. Card & Payments, 2007.

[2] Kahn B K, Strong D M. Product and Service Performance Model for Information Quality :An Update [C].IQ, 1998,102-115.

[3] Cappiello C, Francalanci C, Pernici B. Data Quality Assessment from User's Perspective [C]. IQIS, 2004.

[4] Aebi D, Perrochon L. Towards improving data quality [A]. In: Proc. of the International Conference on Information Systems and Management of Data [C]. 1993, 273-281.

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

[6] W.W. Ng, U.S. Panu, W.C. Lennox. Chaos based analytical techniques for daily extreme hydrological observations [J]. Journal of Hydrology, 2007, 342, 17 - 41.

Data Quality Control and Management for Hydrological Database

YU Yufeng[1], ZHANG Jianxin[2], ZHU Yuelong[1], WAN Dingsheng[1]

(1. *College of Computer and Information Engineering, Hohai University, Nanjing* 210098, *China*;
2. *Bureau of Hydrology, MWR, Beijing* 100053, *China*)

Abstract: The data quality has an important influence on the informatization construction. This paper firstly analyzed the key dimensions such as completeness, consistency and accuracy of date quality, and then proposed an efficient data quality management framework based on those dimensions. The quality analysis and control model was built, and some methods and techniques to improve the data quality in hydrology database were discussed in practice view.

Key words: data quality; dimension; quality assessment; quality improvement

(上接第 64 页)

Analysis of Dam-break Flood Transfer in Yunnan Mountainous Area

LIU Xinyou[1,2], LI Zishun[1], ZHU Jun[1], YIN Binghuai[1]

(1. *Yunnan Bureau of Hydrology and Water Resources Management, Kunming* 650106, *China*;
2. *Asian International Rivers Center, Yunnan University, Kunming* 650091, *China*)

Abstract: Affected by the terrain, reservoirs are taken as the main water sources in the Yunnan mountainous area. Because the most reservoirs were built many years ago, there is a greater risk of dam break under the influence of under the influence of precipitation concentrated in the monsoon climate. Therefore, it is important to make dam-break flood routing. In this paper, a dam-break flood routing model was used for the actual conditions of the Yuannan mountainous area and the Hexi Reservoir in Changning County was taken as a study case to determine the relevant parameters. The results show that if dam break occurs, reservoir will collapse fully, with a great amount of flood flow and rapid transfer.

Key words: dam-break flood; flood transfer; Yunnan mountainous area