中医药科学数据的数据质量研究*

□胡雪琴** 崔 蒙** (中国中医科学院中医药信息研究所 北京 100700)
陈 兵 (厦门大学信息科学与技术学院 厦门 361005)

摘 要:在国家科技部基础性工作-建设"中医药科学数据管理与共享服务中心"取得成果的基础上,以中国中医科学院中医药信息研究所目前存在的 110 个中医药数据库 260G 数据为基础数据源,结合网络和其他来源数据,讨论中医药数据仓库的建设及数据整合。技术上主要从元数据模型入手,研究中医药数据质量评价标准和关键维度评估算法,对构建的数据仓库进行数据质量的评估,从而完成高数据质量的中医药数据仓库。

关键词:中医药科学数据 数据质量 数据仓库 元模型

在医学领域,医学信息化水平的提升带来了大量的数据,同时也暴露了不少的问题,如"军卫一号"等 HIS 系统现在就深受数据质量低下的困扰[1]。自 2004 年以来,中国中医科学院中医药信息研究所组织全国 24 家中医院校与科研院所 200 余名科研人员共同构建了中医药科学数据汇交平台、共建平台和共享平台,为中医药行业提供 7×24h 不间断的数据共享服务[2],其数据准确的重要性可见一斑。低下的数据质量对数据仓库的决策支持影响极大,危害极深,造成的损失相当严重。因此,为了确保中医药数据研究的可信性,构建一个高质量的中医药数据仓库,推出一套可信的中医药数据质量评估标准就成为一项刻不容缓的工作。

一、数据质量已成为提高"中医药科学数据"利用率的瓶颈

中医药科学数据管理与共享平台作为国家科技基础条件平台,高质量的数据服务是非常必要的[2]。根据"垃圾进,垃圾出"(garbage in,garbage out)这条原理,在中医药科学数据管理中,要求数据仓库管理的数据必须可靠,减少错误,能反映中医药数据的实际情况。但是,由于系统集成和历史数据造成的原因,中医药科学数据共享平台中的数据仍存在较严重的数据质量问题。

前期我们对中医药多库集成平台的基础数据进行了调查分析,发现造成中医药科学数据质量低下的现象除了管理手段、数据仓库设计不完整、不规范,数据清洗策略不利外,也和不同时期开发的业务

收稿日期:2009-07-31
修回日期:2009-08-09
* 科学数据共享工程医药卫生科学数据共享网(2005DKA32405):中医药学科科学数据中心,负责人:刘保延,崔蒙。
** 联系人:胡雪琴,在站博士后,主要研究方向:中医药信息学,中国中医药语言系统研究,E-mail:huxueqin@gmail.com;崔蒙,本刊编委,博士生导师,主要研究方向:中医药信息学的学科建设,中医药信息数据库与网络建设,中医药信息数据的挖掘与利用研究,以及中医药软科学战略研究,Tel:010-64013395,E-mail:cui@mail.cintcm.ac.cn。

数据库之间数据结构不一致，不同源头系统最原子信息的粒度不一致、数据库各录入模块缺少输入验证过程，不能屏蔽非法格式的数据入库；验证程序不能发现格式正确但内容不正确的错误；以及数据库中一个数据字段存在多个数据元素、中医药概念中存在大量的同物异名、异物同名等现象有关，具体的质量问题情况见表1所示。

二、中医药科学数据的数据质量特点

通过对中医药科学数据质量问题的研究，我们发现中医药科学数据的数据质量具有以下一些特点：

1. 数据质量问题是多方面的

有数据采集方面的质量问题、数据传输方面的质量问题，还有数据存储方面的质量问题。前期的数据仓库设计不完整、不规范，数据清洗策略不利等多种质量问题。

2. 数据质量约束的对象是多样的，而且是一种多层次的划分方式，需要对不同的对象进行质量约束定义

对于中医药科学数据而言，数据质量约束至少应该针对数据集、属性、数据等对象。

3. 数据质量的元素是相对稳定的

数据质量的元素指数据约束的类型，如完整性、一致性、准确性等。数据质量元素目前没有统一的说法。但许多系统，包括一些专业领域都在试图定义局部的数据质量元素规范，以形成对数据质量的定量或非定量的衡量标准。

4. 数据质量指标计算是复杂的

数据质量元素定义的是一种概念，每一个数据元素都是需要用定量或定性的指标进行说明，这些指标的计算需要用一定的算法描述。关系数据库理论对模式上的约束定义已经非常完善，但其它方面的算法则需要进行更深入的研究。

可见，中医药数据仓库中的数据质量非常复杂，质量的多方面性使得单一质量模型满足不了中医药行业建立质量体系的需求，必须从更加抽象的层次来描述数据质量。质量内容的变化性要求数据质量模型具有较大的灵活性和可个性化定制，针对专门的数据质量模型进行计算的质量评估软件不能适应这种动态性的需求。

三、中医药数据质量研究的主要内容

1. 数据质量元模型的构建

本文设计的元数据管理系统的逻辑架构主要由5部分构成：元数据获取部分、元数据存储部分、元数据管理部分、元数据应用部分、元数据服务接口部分；元数据管理系统的物理结构主要由3部分构成：数据库服务器、应用服务器、用户平台，具体设计的元数据管理系统的逻辑架构见图1所示。

2. 数据质量控制标准模型的构建

数据仓库质量本身是一个主观性的问题，而要使之发挥相应的作用就必须量化质量，使其具有多个指标和决定因素，

表1 中医药数据仓库建设所涉及到的业务表的数据质量分析

数据库	主要内容	相关业务表	质量分析	解决办法
中药化学实验数据库	全面介绍中药化学成分，共收录中药化学成分14032种，对每一种化学成分的理化性质、化学结构、临床应用等进行研究。	化学实验表 实验室条件表 化学成分纯化表 化学成份分离表	一个字段存在多个数据元素 实验室名称重复 化学成分鉴定不完整 含量测定不完整 测定方法不完整 实验结果不一致	可通过字段拆分等手段可以较好的解决。
中药基础数据库	数据库是以中药标准化与中医药科技期刊文献科学实验数据为依据，搭建可供中药数据挖掘与知识发现的应用平台。	品种表 一般药理表 一般临床药理表 化学成分表	动植物资源分布不完整 中药材基源分类不完整 观察指标不完整 指标检测方不完整 对症治疗不完整 功效不完整 治疗疾病不完整	可通过文本相似度、最大序列匹配、潜在匹配等方法解决。
针灸数据库	主要围绕针灸所临床就诊患者的临床病例所建立的一个加工界面，是以提取针灸临床病历中的有效信息为目的的一个加工程序。	患者基本信息表 针灸治疗表 按摩治疗表 针灸穴位表	病案号不完整 患者转归不一致 针刺手法不完整 针刺频率不完整 按摩手法存在同物异名、异物同名的现象 穴位类型不完整	可用我们自主研制的“词雀”软件，再通过中医药一体化语言系统和标准词表来干预，对所需要的数据进行清洗和整理。

以便于综合评判。本部分为了能够实现这一自动分析评估过程，以实体对象和实体对象的数据质量控制标准为依据，把相关的数据质量控制标准定义到数据库中作为数据质量控制标准元数据，为实现对所评估对象的数据质量的自动分析和评估服务。该元数据不仅承担数据质量控制标准的映射任务，同时也是数据质量评测的核心，它包含了分析结果信息、评估结果信息、系统管理信息、系统应用信息等内容。中医药科学数据的数据质量评测关系模型见图 2 所示。

3. 数据质量的分析任务

数据质量分析任务是按照数据质量控制标准分类进行的，目前设计的数据质量控制标准分 4 方面进行：

①针对逻辑结构的逻辑结构标准执行情况分析；

②针对整体的完整性分析、冗余性分析、一致性分析、深度性分析、及时性分析、放射性分析、活动度分析；

③针对数据的非空约束分析、值域约束分析、代码标准分析、取值标准分析、词法约束标准分析、偏差分析、等值函数依赖分析、逻辑依赖关系分析、一致性分析（等值一致性、存在一致性、逻辑一致性）、连续性分析；

④将进行基础数据描述分析。

数据质量分析任务的核心是通过对元数据的标准定义进行处理，生成可执行的分析查询语句或存储过程，并在数据质量分析对象数据库内执行，形成查询结果，并经由分析程序进行记录、汇总，形成分析结果信息，存储到分析结果元数据表中。

4. 数据质量评测标准的建立

本部分依据数据质量分析结果，按照数据质量定义模型，由数据质量分析评估程序定义的算法计算出数据质量各种评估指标和关键特性指标。每一关键特性指标在不同的评估范围中有不同的计算方法，因而在评估算法中要分别描述但这些算法共同遵循了如下原则：

①反映数据实际情况的各项指标，如问题数据个数、问题记录数、问题数据项个数、问题分类数目等指标，是基于数据质量分析结果得出的统计数据。

②反映关键特性的汇总指标，如完整性、一致性、问题数据发生率等指标。

③反映关键特性的加权汇总指标，如加权完整性、加权一致性、加权问题数据发生率等指标。

数据质量模型根据以上数据质量分析结果经过总体评估操作给出数据质量评估报表。

5. 解决数据质量问题的步骤

目前中医药数据仓库并没有统一标准，所建的数据仓库质量低下，我们可以先借助其他行业数据质量标准和数据库、数据仓库理论，构建可信的中医药数据仓库元模型，制定中医药数据仓库数据质量评价标准，设计数据质量评估算法。具体的实现步骤可按下面图 3 的技术路线来执行。

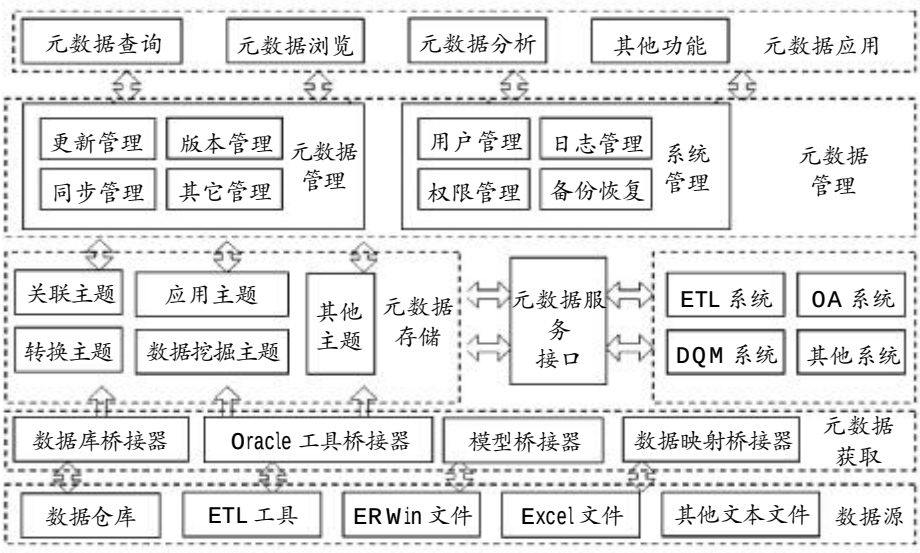

图 1　数据仓库元数据管理系统逻辑架构

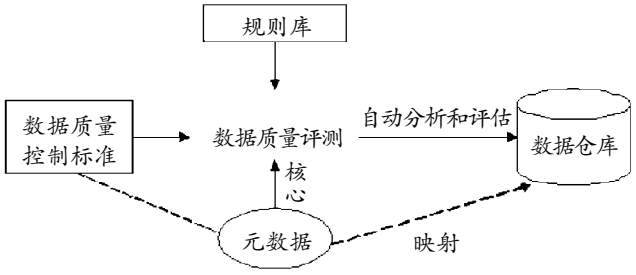

图 2　中医药数据质量评测关系模型图

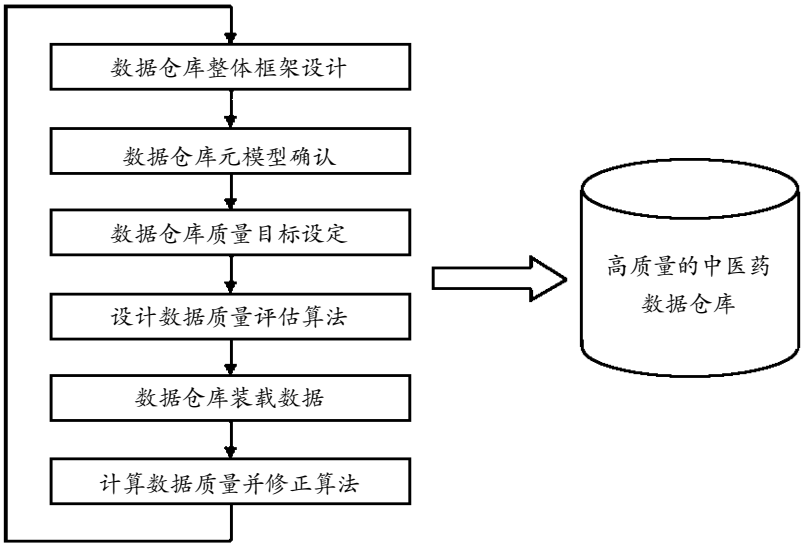

图 3 解决数据质量问题的 7 个步骤

四、问题与展望

高质量的中医药数据仓库是实现深层次挖掘与利用的工具，具有广泛应用前景和庞大应用市场。本项研究对于改进中医药数据质量评估方法，改善现有中医药数据质量，完善中医药科学数据共享平台有着很大的提高和促进作用，研究成果可以应用于我国的中医药数据管理具体实践中去。今后，通过中西医数据的语义对照、疗法对应等研究可以延伸到整个医学数据质量评估，应用到临床实践中协助医生日常工作。在数据质量管理上所投入的资金和时间，将在现在和未来得以高额的回报。

参考文献

1 王志勇，吴骋，余志明.“军卫一号”数据仓库建设中数据质量问题的研究. 医学信息，2006，19(9)：1503~1505.

2 崔蒙，尹爱宁，范为宇，等.中医药科学数据建设研究进展.中国中医药信息杂志，2006，13(11)：104~105.

Research on data quality of Chinese medicine scientific data

Hu Xueqin[1], Cui Meng[1], Chen Bing[2]

(1. Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

2. Institute of Information Science and Technology, Xiamen University, Xiamen 361005, China)

Abstract: Based on the achievement of the MOST fundamental research project " Traditional Chinese Medicine (TCM) Scientific Data Management and Shared Services Center" and taking TCM 110 databases, totally 260G data as the basic data source, this paper mainly discusses the construction of the TCM data warehouse and data integration. Starting with the Meta data model, it analyzes the assessment standards and key dimensions evaluation algorithm of the TCM data quality, aiming to establish a TCM data warehouse of high quality.

Keywords: Chinese medicine scientific data; data quality; data warehouse; Meta data

(责任编辑：李沙沙，责任译审：张立崴)