

建立中医药数据服务与利用平台*

□刘 静** (中国中医科学院中医药信息研究所 北京 100700)

摘 要:随着中医药数据库群的建设与应用,以往的数据库建设方式形成了中医药信息孤岛现象越来越严重,难以适应中医药信息领域发展的需求。本文探讨了中医药信息数据研究与服务的发展趋势,强调了建立中医药数据服务与利用平台的重要性,在建设完善中医药数据服务与利用平台的基础上,将开展面向用户需求的、灵活多变的专题服务,提供挖掘算法工具与优秀案例,从而推动中医药事业的发展。

关键词:中医药信息 数据库建设 数据服务与利用 数据挖掘

中国中医科学院中医药信息研究所自 1984 年开始建设中医药期刊文献数据库,经过 25 年的建设,目前已累积的数据量达 150G,这些数据包括了大量关联结构型的数据,如:中药化学实验数据库、中药药理实验数据库、临床疾病数据库、临床个案数据库、针灸临床医案数据库等等。这些数据库在建设的过程中,采集了大量结构化的数据,如疾病的名称、疾病的证候、疾病的症状、治疗疾病的方剂、方剂中所包单味药物、单味药的剂量与炮制方法、针灸治疗的穴位及主次穴的分布、疾病的治则等等。目前这些数据库已经能够在共享网上进行检索应用,提供了大量的关联检索结果,但这些结果均以单个数据库为检索个体,缺少各个数据库之间的关联检索,也难以进行数据的综合利用和数据挖掘,迫切需要建立起中医药数据服务与利用平台。

一、中医药信息将走向开方式服务

中医药数据库的建设已经历了 25 年的岁月,这是一个将数据文件系统走向数据库的面向单个应用的过程,已经形成了中医药的数据库群,但是这些数据库均

以单个数据库为检索个体,缺少各个数据库之间的关联检索,形成的仍然是信息的孤岛,也难以进行数据的综合利用和数据挖掘,为了今后的数据利用与数据服务,应将数据层进行分离,数据层不再属于任何一个单独的应用系统,利用应用信息层来管控共同的核心数据,解决多个应用之间数据的一致性、实时性、共享与隐私的保护等问题,并以更为松耦合的方式对各类应用提供信息服务。此外,数据的组织也从为其具体业务逻辑服务转换为独立于业务逻辑的服务。推动中医药信息走向开放架构是一种必然的趋势,

二、中医药信息管理将以服务的形式存在

以往的中医药信息数据库建设模式是先有明确的需求,然后再进行软件工程 IT 建设,然后再推动中医药行业对数据库的应用,在应用建设完成后再考虑中医药数据的整合,尚未进行中医药数据利用与服务的构架研究。以往的过程,是中医药数据库发展的必然要走的路,也是中医药信息数据库建设的捷径。随着中医药数据库建设的规模日趋扩大,形成的信息孤岛现象也就越来越严重,迫使从事中医药信息建设领域的人们,开始考虑进行中医药数据利用与服务形式的构架研究。

收稿日期: 2009-07-24

修回日期: 2009-08-09

* 科技部基础性工作医药卫生科学数据共享网(2005DKA32405): 中医药学科学数据中心, 负责人: 刘保延, 崔蒙。

** 联系人: 刘静, 副研究员, 主要研究方向: 中医药信息数据研究, Tel: 010-64048243, E-mail: liuj@mail.cintcm.ac.cn。

随着计算机技术在中医药信息领域的深入应用,中医药行业对中医药信息的服务需求也具有了多样化、灵活的应用模式等特点,以往的针对需求而建设的应用模式越来越难以适应数据服务的需求。把已建好的各类中医药数据,从以往的应用中剥离出来,变成中医药数据平台的整体数据层,把对整个中医药数据层的信息管理应用与单个专向数据库的应用脱离,构架整体中医药数据应用层,数据应用层不属于任何一个单独的数据库,而是以服务的形式存在,逐步形成一种可管理控制的中医药数据开放式服务数据。

这种开放式的服务必须要考虑到中医药各领域数据的差异性和复杂性,对中医药数据建模提出了更高的要求,中医药数据的建模必须依据中医药本体论,但又要具有更好的灵活性和可扩展性。数据的存储与访问将进行隔离,今后越来越多的服务不是通过 SQL 直接访问数据库底层,而是通过 XML 服务接口来访问与获取信息。

三、中医药数据服务与利用平台的建设

中医药数据服务与利用平台的构架是整个平台的核心部分,应基于对中医药行业发展与技术走向的深刻理解,在归纳中医药数据业务要素的同时,处理好数据和功能的配合,并具有合理的灵活性。

整体中医药数据服务与利用平台可分为 4 个层次,从上到下依次为:

1. 中医药数据业务应用层

中医药数据系统的使用者,包括 Web 门户和通过 API 访问系统的应用。为利于今后中医药科学数据的挖掘与利用,应进行分布式计算平台层与数据挖掘平台层的建设。

可提供分布式计算平台层,包括 3 部分功能:①分布式文件系统:提供分布式数据文件存储功能,提供具备高可靠性、高稳定性的存储平台;②并行编程环境:提供基于 Mapreduce 的变成模型,及任务调度、任务执行、结果反馈等功能;及向平台提交作业功能;③分布式系统管理:实现对平台的分布式系统管理。

数据挖掘平台层:包括 5 部分功能:① workflow 模块:实现对各个数据挖掘步骤及模块总控、调度功能;②数据加载模块:将源数据从其他外设中计算平台系统中;③并行 ETL 模块:对原始数据进行预处理以得到挖掘数据;并行数据挖掘工具向计算平台提交待执行的 ETL 任务,由云计算平台执行并反馈结果,存放于 DSF;④并行数据挖掘算法模块:实现满足业务需要

的数据挖掘算法;并行数据挖掘工具平台向云计算平台提交待执行的聚类算法任务,由云计算平台执行并反馈结果,存放于 DFS;⑤并行结果展示模块:将并行数据挖掘算法的结果展示给用户。

2. 视图层

向中医药数据应用层提供统一的资源访问界面,并返回全局模式的数据视图。

3. 集成层

实现中医药信息资源的全局模式与局部模式之间的映射,执行用户和应用的访问请求。集成层为中医药数据服务与利用平台的核心,分布式、异构的数据资源在集成层被映射层单一的全局数据源,屏蔽了底层数据源的异构多样性,使外部应用为统一的数据源。采用描述性元数据对集成层的数据源进行统一管理。

4. 数据层

集中中医药科学数据中心目前可应用的数据资源。可从疾病、证候、症状、治测、方剂、单味药、人体指标、穴位、药理作用分类、药理实验对象、动物、药物化学名称、研究单位等等概念作为入口,根据每个概念的本体属性与关联关系进行数据的整合,将整合后的数据部署于数据层。

四、基于中医药数据服务与应用平台的数据挖掘

1. 数据挖掘的基本概念

数据挖掘(Data Mining)是指从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

数据挖掘与传统的数据分析(如,查询、报表、联机应用分析)的本质区别在于:数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识,所得到的信息应具有事先未知、有效和实用 3 个特征。数据挖掘通过预测未来趋势及行为,做出前瞻的、基于知识的决策。

2. 数据挖掘的基本步骤

(1) 收集数据。

可以从多个数据源获取数据,对数据进行集成。

(2) 数据选择。

分析所收集到的数据,并决定选择数据和排除数据的原则和标准。

(3) 数据清洗。

检查收集的数据是否满足数据挖掘的要求,目的是将数据质量提升到符合所选择的分析技术的需求,

以保证数据信息的正确性和一致性。

(4)数据转换。

将已有的数据根据挖掘分析的要求进行格式及逻辑转换。

(5)建立模型。

选择合适的算法进行挖掘验证。

(6)结果分析。

对挖掘结果进行解释及可视化的分析。

(7)知识同化。

将分析所得的知识集成到业务信息系统, 或为决策提供参考。

3. 中医药数据挖掘的方向

中医药数据的挖掘与利用应围绕着中医理论基本特色“理、法、方、药”, 这是中医学诊断与治疗操作规范的四大要素, 辨证论治是理法方药运用于临床的过程。以往的数据挖掘研究主要是 3 个方面。

(1)中药研究。

采用模糊数据、统计学、图形挖掘待方法, 研究中药化学成分药效基团、指纹图谱等。

(2)方剂研究。

采用关联规则、聚类分析, 对应分析和频繁集方法, 研究方剂与症状间的关系、组方配伍规律、方剂结构、核心药物及“方药证”的对应关系。

(3)证候研究。

采用聚类分析、人工神经网络、决策树等方法进行证候模型的研究。

今后中医药数据挖掘仍将把“理、法、方、药”的 4 大要素作为重点, 并适当增加中医基础理论研究的数

据建模与数据挖掘研究。

五、面向用户的信息获取方式服务

1. 面向用户需求的专题服务

在中医药数据服务与利用平台的基础性, 中医药数据已经过了清洗与整合, 可比较方便的为用户提供灵活的专题服务, 如针对疾病重点研究室的专病数据库、新药创制的中药先导化合物数据库、同类功效单味药的组方配伍专题数据库、某类证候与疾病症状体内指标关系数据库等, 并可采用聚类分析、关联规则分析等方法对其进行研究。

2. 提供挖掘算法工具及优秀案例

目前比较成熟的数据挖掘方法有 9 种, 分别为分类技术、统计学习、关联分析、链接分析、聚类分析、基层学习、时序分析、集成技术、图形挖掘。利用整合后的数据, 进行数据挖掘模型研究, 将成功的算法及案例部署于中医药数据服务与利用平台上, 让服户可以参照挖掘研究的实例, 进行各自所需求的研究, 以促进从事中医临床、科研工作的人员理解数据挖掘的含义, 学会使用数据挖掘的常用工具, 从而推动中医药事业的发展。

参考文献

- 1 刘晶炜. 信息管理经历三大变革. 计算机世界, 2009, 30:38.
- 2 熊玉庆, 唐新怀. X Couple 一个新型异构信息集成平台. 微电子学与计算机, 2007, 24(9):171~173.
- 3 王懿. 论数据挖掘技术在信息分析中的应用. 科技情报开发与经济, 2009, 19(10):79~80.
- 4 姚美村, 袁月梅, 艾路, 等. 数据挖掘及其在中医药现代化研究中的应用. 北京中医药大学学报, 2002, 25(5):20~23.

Establishment of a TCM data service and application platform

Liu Jing

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

Abstract: With the development and application of individual TCM databases, under the conventional data construction pattern, the "isolated island" phenomenon is becoming more serious, and the demand to develop TCM information cannot be satisfied. This paper discusses the development tendency of TCM information data research and service, and emphasizes the importance of establishing a TCM data service and application platform, based on which to provide flexible and user-oriented subject services and offer mining algorithms tools and selective cases, so as to put forward the TCM industry.

Keywords: TCM information; database development; data service and application; data mining

(责任编辑:李沙沙, 责任译审:张立巍)