

基于语义表达的中药概念数据模型构建

刘丽红 朱彦

(中国中医科学院中医药信息研究所,北京,100700)

摘要 目的:中药领域大量内容丰富、异质异构数据为领域信息交流与共享带来困难。建立基于语义表达的中药数据模型,一方面,能在语义层面上提供知识共享与重用;另一方面,能促进数据资源的集成、交换和应用的互操作。方法:采用本体与底层数据相结合的方法,基于现有本体中国中医药学语言系统(Traditional Chinese Medicine Language System, TCMLS),参照已有概念数据模型,结合基于基础数据进行概念抽取构建中药概念数据模型。结果:基于中医药一体化数据模型(Traditional Chinese Medicine Integrated Model, TCMIM)框架,分析目前中药领域数据资源,抽取中药及相关业务领域概念,构建中药概念数据模型框架。结论:运用本体与数据相结合的方法,在概念层面建立中药数据模型,在数据层面有利于数据建模的准确性,促进信息资源的集成与交换;在语义层面上有利于丰富补充领域本体。

关键词 中药;概念数据模型;语义表达

Building Chinese Medicine Conceptual Data Model Based on Semantic Representation

Liu Lihong, Zhu Yan

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

Abstract Objective: A large amount of rich content and heterogeneous data in Chinese medicine field bring difficulties for information exchange and sharing. On the one hand, it can provide knowledge sharing and reusing on semantic level, on the other hand, it can promote the integration, exchange and application of data resources. **Methods:** Combining ontology and data, Chinese medicine conceptual data model was established based on the existing ontology Traditional Chinese Medicine Language System (TCMLS), with reference to the existing conceptual data model to extract concepts. **Results:** Based on the Traditional Chinese Medicine integrated data model framework, the current data resources in the Chinese medicine field were analyzed, Chinese medicine and related business concepts were extracted to build the framework of Chinese medicine concept data model. **Conclusion:** With combination of ontology and data, Chinese medicine data model is built at the conceptual level, for the accuracy of data modeling on data level and it could promote the integration and exchange of information resources and supplement the domain ontology at semantic level.

Key Words Chinese medicine; Conceptual data model; Semantic representation

中图分类号:R-331;R286 文献标识码:A doi:10.3969/j.issn.1673-7202.2017.04.058

1 概述

数据模型分概念层、逻辑层、物理层的3层架构。概念数据模型(Conceptual Data Model)描述计算机系统将要处理的问题领域中的事物,它本身常常采用实体联系模型(E-R Model)。概念/语义数据模型可以被看作是“语义模型”或“信息模型”,它与“本体”基本上处于同样的相对位置。这些模型都可以从本体研究那里找到重叠或借鉴的东西^[1](如图1)。

本体自20世纪90年代引入计算机人工智能领域,目前主要有两层含义:一是哲学领域的存在,是本体论的研究对象;二是延伸到特定领域之中,指某套概念及其相互之间关系的形式化表达,包括概念

化、规范化、形式化和共享4个特征^[2]。信息科学中的本体论的目标是确定领域内共同认可的词汇,并从不同层次的形式化模型上给出这些词汇和词汇间相互关系的明确定义,从而获取相关领域的知识,提供对该领域知识的共同理解^[3]。谢琪等提出应用领域本体方法建立中医概念信息模型的思路与构想,解决中医药领域顶层设计理论指导不足的问题,是从顶层设计本体,发现语义关系,构建中医药概念信息模型。郭磊^[4]根据方证相关性构建中医药关系型数据库中数据概念模型,张丽颖^[5]等基于本体建立医学科研信息模型,提出本体建模的依据与方式方法。李明^[6]等采用本体方法,探讨中医证候本体表达模式的构建方法及其在语义检索、智能诊断、数据

分析等方面的应用。周扬^[7]等提出了构建中药本体的思路,期望全面描述中药本质,揭示中药本身及中药间复杂的功效与物质关系,澄清中药的知识结构。娄苗苗^[8]等提出基于领域信息构建国家卫生信息概念数据模型的方法,目的是为元数据项目开发和管理提供一致、透明的语义背景。

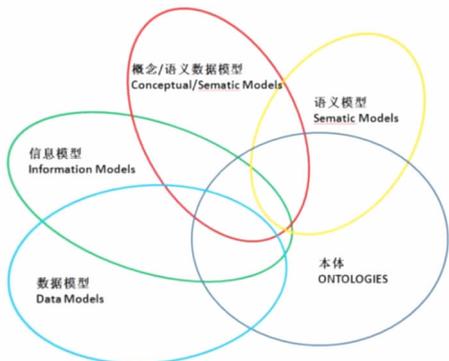


图 1 各模型与本体关系图

在一些行业领域,对于数据模型的研究较为成熟的如交通地理信息领域,其系统的交通地理信息系统数据模型^[9]基于应用出发,从传统数据模型、线性定位参照体系、导航数据模型至 GIS—T 时空数据模型,用来表达地理信息系统中数据之间的联系与逻辑组织形式,满足交通地理信息系统导航需要、多维化与时态化。在石油工业领域,熊方平等^[10]以实体为主线,将各专业数据有机、完整结合,形成一体化勘探开发数据库,满足各个层面应用研究需求的跨专业勘探开发一体化数据模型 EPDM。这些数据模型成熟且完善,覆盖了该行业各个业务领域过程环节数据,方便一体化管理,为跨专业的综合研究提供了基础。国内对于信息交换研究进行了一些探索,张玉海^[11]等对国家卫生信息概念数据模型与 HL7RIM 之间进行了映射关系的对照,确定了它们之间的关系,检验概念数据模型的范围。唐忠^[12]分析军用电子设备试验数据,研究设计了试验数据库概念模型,采用时间、设备和试验题三维关联数据模型,满足试验数据库管理系统应用需求。

在中医药领域,数据模型研究基本处于探索阶段,还未见系统完整的概念数据模型,语义 web 技术得到了一些应用^[13]。目前,由于没有统一语义的数据模型标准,没有参照,重复建设,低质量的数据模型直接影响应用系统的质量,更谈不上应用互操作,阻碍了信息资源的集成、交换。

中药概念数据模型属数据模型范畴,一方面通过梳理中药知识体系,分析中药类数据资源,基于基础数据进行中药概念信息抽取,另一方面利用中医

药领域内本体及现有模型指导约束概念及概念间关系的确定。利用本体与数据结合构建概念数据模型,明确概念含义,理清概念间关系,有利于数据建模的准确性,促进信息资源的集成与交换;在语义层面上有利于补充领域本体。

2 研究方法与步骤

2.1 中药及相关业务领域概念特征分析

2.1.1 中药业务领域数据资源分析 中药概念数据模型中的概念来源主要基于中药领域数据资源进行抽取。中药领域数据来源需根据概念数据模型构建需求设定。

一方面,中药领域核心概念来源要考虑反映中药核心知识体系,需分析中药领域权威辞典、工具书、教材,如《中华人民共和国药典》《中华本草》《中药大辞典》《中药学》《临床中药学》等;另一方面,中药概念数据模型是针对数据构建的模型,要考虑反映中药领域数据库信息特点。

中国中医科学院中医药信息研究所的中药相关数据库包括中国中药数据库、中药科技基础信息数据库、中药药理实验数据库、中药化学实验数据库、中药化学成分数据库、中国方剂数据库等;相关数据库包括中医临床疾病数据库、中医医案数据库、有毒中药数据库等。另外,中国科学院的中国中草药数据库、中药有效成分数据库、中药基本信息数据库、中国中成药数据库、中国英文中药数据库,上海中医中药数据中心的 SIRC/TCM 中医药信息系统,中国科学院武汉植物园的神农架药用植物数据库与华中药用植物数据库,复旦大学生物多样性科学研究所的上海市药用植物数据库查询系统等。通过分析这些数据库系统的数据集与核心数据元,利用权威辞典对数据元进行规范抽取,可以作为中药核心概念的可靠来源。

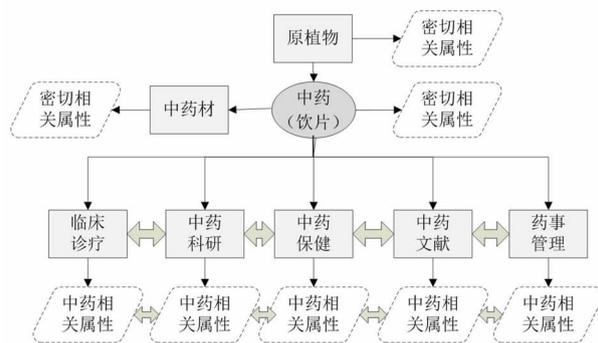


图 2 中药及相关业务领域概念

2.1.2 中药相关业务领域范围设定 TCMIDM 的构建研究提出了数据模型集成框架^[14]。基于此集

表 1 中药类数据元

二级分类	数据元名称	同义	定义
基本信息	中药名称	药名、中药名、规范名词	指中医理论指导下应用的药物的名称,包括中药饮片或药材
	中药英文名称	英文名、英文等价术语	指中药饮片或中药材的英文名称
	中药拼音名称	中药汉语拼音、拼音名	指中药的汉语拼音名称
中药,药材	中药材拉丁名称	拉丁文、拉丁名	指中药材的拉丁名称
	中药基原	基原、药材基原	指确定中药材来源的特征描述
	药用部位		指药用植物、动物和矿物的作为药用部分的描述
	中药材性状	性状、药材	指中药材的物理特征或形态描述
	中药炮制方法	药材炮制、炮制	指中药材在使用前根据病情和实际需要,采用进行炮制处理的方法描述
中药,饮片	药材贮藏	中药材贮藏	指经过加工处理后的药材,按一定规范要求包装、贮存过程的描述
	药材等级		指根据中药材的加工等级的分类描述
	饮片性状	性状	指中药饮片的物理特征或形态描述
	中药性味	药性、性味	治中药所具有的与治疗作用有关的性能,包括四气、五味等
	中药归经	归经	指中药对机体的选择性作用
	中药功能	功能、功效、功能与主治	指中药所发挥的有利的作用或效能的描述
	中药主治	应用、功能与主治	指中药治疗的主要病症的描述
	中药用法	用法	指中药的使用方法的描述
	中药用量	用量	指中药的使用剂量范围描述
	药理作用	药理、现代药理	指中药具有的现代药理学作用,中药对机体各种功能的影响及其作用原理的描述
	化学成分	中药化学成分、成分	指中药包含的化学成分
	用药禁忌	服药禁忌、使用注意、禁忌	指药物不适宜应用于某些疾病、情况或特定的人群等
	不良反应描述	不良反应	指服用中药后,对患者出现药物不良反应表现的描述
	中药配伍禁忌	配伍禁忌	指为防止不同中药合用产生剧烈的不良反应或降低和破坏药效,应该避免配合应用的某些中药的名称
	药用植物	中药应用鉴别	鉴别用药、应用鉴别
中药鉴定			指对中药进行鉴定的方法的描述
中药毒性		毒性反应、毒性	指药物对用药者靶组织(器官)发生的危害性反应的描述
植物中文名称			指药用植物的中文名称
植物拉丁名称			指药用植物的拉丁文名称
药用动物	植物科属	科属	指药用植物所属的科和属
	植物形态	原植(动、矿)物、动植物形态	指药用植物的形态特征描述
	动物中文名称		指药用动物的中文名称
	动物种类		指药用动物所属的科和属
药用矿物	动物拉丁名称		指药用动物的拉丁名称
	动物形态	原植(动、矿)物、动植物形态	指药用动物的形态特征描述
	矿物种类		指药用矿物所属的类和族
	矿物中文名称		指药用矿物的中文名称
	矿物拉丁名		指药用矿物的拉丁名称
	矿物形态	原植(动、矿)物、动植物形态	指药用矿物的形态特征描述

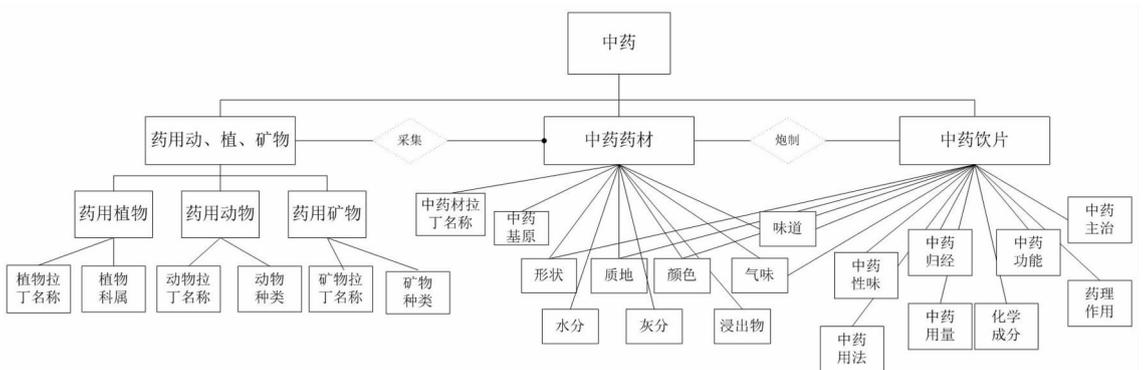


图 3 中药核心概念示例

成框架,中药概念数据模型包含中药及相关业务领域术语资源、数据资源、信息系统等方面概念的抽取,进行中药概念数据模型构建,以中药为核心进行相关业务领域数据资源概念抽取如图2。

2.2 中药及相关业务领域概念及概念间相互关系

杨喆^[15]等基于国家卫生信息概念数据模型为基础,提炼和描述卫生领域数据元,表达数据元及其语义关联。该方法与中药概念数据模型中基础概念抽取一致。在中药及相关业务领域概念抽取中,以中药知识体系为核心概念体系,相关业务领域概念为补充完善。前期研究^[16]对中药知识资源体系进行了比较,结合“中医药数据元标准研究”中药类数据源^[17],进行中药类数据元整理修订,结果如表1。基于数据元进行核心概念的抽取与描述,示例如图3。

在进行中药概念数据涉及的业务领域分析研究中,需结合调研咨询、业务梳理、专家讨论等,辨析中药及相关业务领域概念含义及其在理论中的地位、性质、作用,确定核心概念、非核心概念,比较概念含义及其重要性,确定核心概念、非核心概念,对中药领域概念与相关业务领域概念进行梳理。

2.3 中药概念数据模型构建 参考中药工具书、教材、相关专著及中药规范数据元研究成果,根据其含义及其在理论中的地位、性质、作用,确定核心概念、非核心概念,对核心概念与非核心概念之间的层次、类属、逻辑关系等进行深入辨析,确定其相互关系;结合中药相关业务领域,分析中药域与其他域之间的关系,进行中药概念及关系的考量。以本体论方法为指导,全面分析提取概念间的语义关系,以TCMLS的56种语义关系为参照,结合其语义关系适用情况,进行适当的增删调整,梳理中药概念间的语义关系,基于统一语义表达,构建中药概念数据模型。

中药概念数据模型的建立,一方面希望可以作为数据物理模型的上层,指导数据库规范建设,方便数据交换与共享,另一方面作为领域概念信息的抽取,可以丰富和补充TCMLS,完善中医药领域顶层本体建设。

3 结语

运用本体与数据相结合的方法,基于基础数据进行中药及相关业务领域概念及概念间抽取,概念及概念间关系的确定需结合利用领域本体TCMLS,重用概念及概念间关系,兼顾HL7 RIM模型,建立

概念间关系,在概念层面建立中药数据模型。这种构建概念数据模型的模式,概念来源于基础数据,能够反映数据结构特点,利用语义表达概念间关系,可在语义层面上提供知识共享与重用,为中药本体的构建提供来源。

概念数据模型一方面要有较强的表达能力,另一方面也需要简单、清晰,易于理解。在实际应用过程中,必须进一步转换成为逻辑数据模型、物理数据模型,才能在具体数据库中实现^[18]。

参考文献

- [1]余彤鹰. 札记:本体及数据、信息、领域、企业建模与模型. <http://www.ee-forum.org/wp/pub/ty/2011-02-p2491.html>, 2016-06-07.
- [2]Studer Rudi, Richard Benjamins, Dieter Fensel. Knowledge engineering: principles and methods [J]. Data and Knowledge Engineering, 1998, 25(1/2): 161-197.
- [3]苏里, 朱庆伟, 陈宜金, 等. 基于地理本体的空间数据库概念建模 [J]. 计算机工程, 2007, 33(12): 87-89.
- [4]郭磊. 根据方证相关理论构建中医药关系型数据库概念模型探讨 [J]. 山西中医, 2009, 25(3): 60-61.
- [5]张丽颖, 马鸣华, 金岩. 基于本体的医学科研信息模型的建立与分析 [J]. 中华中医药学刊, 2013, 31(10): 2259-2261.
- [6]李明, 张昌林, 包汉飞, 等. 中医证候本体表达模式的构建与应用研究 [J]. 上海中医药杂志, 2013, 4(1): 7-11.
- [7]周扬, 王振国. 中药 Ontology 概念关系体系的构建探析 [J]. 中国中医药信息杂志, 2009, 16(3): 96-97.
- [8]娄苗苗, 杨喆, 刘丹红, 等. 基于领域信息的卫生信息概念数据模型构建方法 [J]. 中国数字医学, 2015, 10(1): 74-77.
- [9]石建军, 许国华, 何民, 等. 交通地理信息系统数据模型的研究进展 [J]. 北京工业大学学报, 2004, 30(3): 311-322.
- [10]熊方平, 马进山, 陈新燕, 等. 中国石油一体化勘探开发数据模型研究与实践 [J]. 信息技术与信息化, 2011, (3): 49-55, 60.
- [11]张玉海, 徐勇勇, 刘丹红, 等. 国家卫生信息概念数据模型与HL7RIM的映射研究 [J]. 第四军医大学学报, 2006, 27(11): 999-1001.
- [12]唐忠, 周含冰, 何荣茂, 等. 军用电子设备试验数据库概念模型设计 [J]. 计算机与数字工程, 2013, 41(1): 148-150.
- [13]于彤, 崔蒙, 李敬华. 语义 Web 在中医药领域的应用研究综述 [J]. 世界中医药, 2013, 8(1): 107-109.
- [14]刘丽红, 朱彦, 李海燕, 等. 中医药一体化数据模型的构建 [J]. 中国数字医学, 2015, 10(10): 70-72.
- [15]杨喆, 刘丹红, 娄苗苗, 等. 基于信息建模的数据元标准化方法 [J]. 中国数字医学, 2016, 11(2): 58-60, 70.
- [16]刘丽红, 贾李蓉, 刘静, 等. 中药本体相关概念描述探讨 [J]. 中国数字医学, 2016, 1(2): 90-92.
- [17]董燕, 张竹绿, 李海燕, 等. 中医药数据元提取规则与分类探讨 [J]. 中国数字医学, 2013, 8(11): 79-80, 88.
- [18]梁礼方. 数据模型 [J]. 金融科技时代, 2014(1): 45-51.