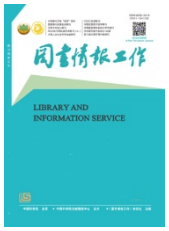


图书情报工作
Library and Information Service
ISSN 0252-3116,CN 11-1541/G2

《图书情报工作》网络首发论文

题目：基于本体的简帛医药知识组织研究
作者：李贺，祝琳琳，刘嘉宇，樊娇，侯力铁
DOI：10.13266/j.issn.0252-3116.2022.22.002
收稿日期：2022-05-08
网络首发日期：2022-12-22
引用格式：李贺，祝琳琳，刘嘉宇，樊娇，侯力铁．基于本体的简帛医药知识组织研究[J/OL]．图书情报工作. https://doi.org/10.13266/j.issn.0252-3116.2022.22.002

网络首发时间：2022-12-22 11:51:53
网络首发地址：https://kns.cnki.net/kcms/detail//11.1541.G2.20221219.1211.002.html
图书情报工作
第 66 卷 第 22 期　2022 年 11 月

基于本体的简帛医药知识组织研究*

■ 李贺[1]　祝琳琳[1]　刘嘉宇[1]　樊娇[1]　侯力铁[2]
[1] 吉林大学商学与管理学院　长春 130012　[2]长春中医药大学　长春 130117

摘　要：[目的/意义]中医药是中华文化的传承，对现代医学发展具有重要作用，应受到高度重视。利用数字技术组织中医古籍，对其进行保护与重用具有重要意义。[方法/过程]以简帛医药文献为研究对象，分别构建简帛医药书目本体和内容本体，并通过书目本体和内容本体的连接形成简帛医药文献本体模型；利用命名实体识别模型抽取实体概念，利用依存句法分析和规则的方法抽取实体关系，最终将简帛医药文献数据存入图数据库，实现简帛医药文献知识图谱可视化与检索查询结果呈现。[结果/结论]利用本体和知识图谱，提出简帛医药知识组织方法，建立具有逻辑关系的简帛医药知识链，形成中医药领域知识的语义网络，在中医古籍的智能化方面做出有益的探索，有助于中药知识的普及与传承。

关键词：简帛医药文献　知识组织　本体　知识图谱　数字人文
分类号：G203
DOI：10.13266/j.issn.0252-3116.2022.22.002

1　引言

中医药包含了中华民族深邃的哲学思想，是极具原创性的中华民族智慧结晶。获得诺贝尔医学奖的屠呦呦团队正是结合了《肘后备急方》中的中医知识发现青蒿素，开创了疟疾治疗新方法。中医古籍资源规模较大，根据《中国中医古籍总目》的记载，目前我国馆藏的中医文献有 13 455 种。自 20 世纪 70、80 年代以来，随着大量的简帛医药文献出土，各种简帛医籍陆续被整理公布，数量十分丰富。简帛医药文献是以竹简、帛书等形式记载的中医药内容，出土涉医简帛文献是传世中医古籍的文献源头[1]，它是研究古代医药知识的珍贵资料，是部分中医药书籍的祖本。例如湖北张家山出土汉简《脉书》和马王堆汉墓出土帛书《阴阳十一脉灸经》甲乙本可以被看作是《黄帝内经·灵枢·经脉》的祖本。中医药既是中华文化的传承，也对现代医学发展具有重要作用，中医古籍的保护与重用需要得到高度重视。然而，古籍原始载体的物理保存方式会随着时间而受到侵蚀，会导致原始载体的老化和破损。传统的数字化转化和存储保护了古籍载体原件，为古籍数据化及智能化奠定了良好基础，但很多中国古代医籍卷帙浩繁、晦涩难懂，存在知识散落难寻、数字化资源使用效率较低的问题。而对于出土的简帛医药文献，更是很少从本体和知识图谱角度组织文献书目和内容，知识以及知识之间的相互联系尚未得到有效的挖掘与利用。

鉴于此，本研究将以简帛医药文献为对象进行研究，以期利用计算机相关技术更好地利用中医药古籍中丰富的知识，推动中医古籍的深度开发和利用。本研究主要解决以下问题：①利用本体技术，全面构建简帛医药知识的书目本体和内容本体，进而完成简帛医药文献本体的构建，对简帛医药文献进行知识组织；②基于上述本体，利用命名实体识别和实体关系抽取等相关知识图谱技术，构建简帛医药文献知识图谱，凸显该领域的核心概念和知识关联，以图形的方式可视化展示中医药知识及其相互关系。

2　研究现状

2.1　古籍数字人文相关研究

数字人文领域的不断发展为未来计算科学、社交网络理论与分析、机器和深度学习等领域增加了需求[2]，L. Kaplan 提出了专注于大型文化数据集的处理

* 本文系国家社会科学基金项目“数据驱动的档案文献资源知识构建与知识服务研究”（项目编号：21BTQ109）研究成果之一。
作者简介：李贺，教授，博士，博士生导师；祝琳琳，副教授，博士，通信作者，E-mail：zhulinlinjlu@163.com；刘嘉宇，博士研究生；樊娇，本科生；侯力铁，馆员，博士。
收稿日期：2022-05-08　修回日期：2022-09-26　本文起止页码：16-27　本文责任编辑：王传清

和分析、关注整个数字文化、处理大数据3个具有挑战性的数字人文领域研究问题[3]。目前，关于古籍数字人文研究主要集中在古籍文献的组织与检索、古籍文献导读以及古籍文献的分析与利用3个方面。在古籍文献的组织与检索方面，夏翠娟等[4]从已经存在的元数据中进行知识挖掘，实现面向知识发现的数字人文服务；在古籍文献导读方面，面向普通大众普及中华传统文化知识，提供文献导读相关服务，例如，南通大学楚辞研究中心利用文本概念挖掘技术和语义分析技术构建语义检索模型，降低楚辞的阅读理解门槛；在古籍文献的分析与利用方面，王大学等[5]利用GIS技术构建开放的大型中国古籍资源库，该系统具备编绘专题地图和连接用户数据等功能模块，表达古籍基础地理信息的分布及演变知识。

2.2 古籍知识组织相关研究

目前，关于古籍的知识组织研究主要集中在古籍知识抽取、古籍知识库构建、本体构建以及知识图谱的构建等方面。在古籍知识抽取方面，高晶晶等[6]针对中医古籍生僻字处理过程中存在的缺字现象，提出了使用私用区造字法等进行集外字处理。朱玲等[7]以正则表达式作为提取规则进行中医古籍疾病相关知识的抽取。付璐等[8]抽样10种不同类型的清代医学书籍，探讨了古代中医书籍的分词规范。在古籍知识库构建方面，王国玺等[9]通过对医案古籍知识的提取和对中医古籍后控词表的完善与应用，开发了医案古籍知识库。M. C. Herrera-hernandez等[10]设计了一个基于网络的知识管理系统，作为探索中西医之间关键关系的培训和研究工具，促进整合主流治疗方式的相关医学诊断。在本体构建方面，汤伟平等[11]利用自动语义标注和本体构建工具建立了岭南温病古籍知识的本体框架。L. Abouenour等[12]建立了词汇信息本体库，用于阿拉伯语的问答应用，并在此基础上构建Arabic Verb-Net框架。J. Chen和S. Ou[13]基于建筑叙事理论提出了中国古建筑本体模型，利用语义Web技术重组中国古建筑相关信息，提出的本体为中国古建筑非结构化信息的语义标注提供了解决方案。在知识图谱构建方面，S. Ferre等[14]进行链接预测研究，通过链路推断知识图谱实体之间缺少的边。

2.3 中医药知识图谱相关研究

中医药领域知识图谱的应用研究范围主要集中在中医临床病例医案[15]。相关研究学者全面总结了目前中医药各子领域的知识图谱应用情况[16－17]，除了有利用古代医案等作为数据源的中医医案领域知识图谱，还包括证候知识图谱、方剂知识图谱、中药知识图谱、中医核心知识图谱、中医药知识图谱、用方经验知识图谱等研究范畴。在中医药知识图谱技术应用方面，王蕊等[18]在中医特色疗法相关领域内使用知识地图技术，构建中医特色疗法知识地图。刘燕等[19]使用实体识别等技术构建医学百科知识图谱。张德政等[20]提出了基于本体的中医核心知识图谱构建方法，探讨了模型层和数据层的匹配映射机制，实现了中医基础理论本体和基于本体的中医核心知识图谱的构建。总体来说，与其他领域的知识图谱应用技术相似，中医药领域实体抽取方法主要包括基于规则的方法、基于统计的方法，近年来也有将机器学习模型与深度学习方法相结合进行实体抽取[17]。中医药领域实体关系抽取方法主要包括基于已知的实体类型和关系存在事实、基于关键词、模板、规则或依存句法、基于机器学习或深度学习[21]。综上所述，随着数字人文的发展，信息资源研究经历了数字化、数据化、关联化、智慧化阶段[22]，其知识组织深度不断提升。而关于古籍的研究，也由最初的古籍资源的数字化建设发展到当前古籍资源的数据化与关联化，结合本体和知识图谱等相关理论方法，逐渐形成对古籍资源的语义化研究阶段。但中医药古籍研究仍存在不足：①中医药古籍卷帙浩繁、晦涩难懂，如何理解古文中中医药术语是一个重要问题；②虽然研究者构建了养生本体、中药本体、证候本体等领域本体，但是面向整个中医药领域充分结合中医药文献书目和内容的本体模式层的联合构建研究相对较少；③出土涉医简帛文献是传世中医古籍的文献源头，但鲜有研究利用现代化信息技术手段对简帛医药知识进行组织。因此，本研究将利用简帛医书，从书目本体和内容本体两个方面构建简帛医药知识本体，通过实体抽取和关系抽取等技术构建简帛医药文献知识图谱。

3 简帛医药文献知识本体模型设计与构建

本体是知识语义组织应用最多的形式之一[23]，是一种高效的语义层面的知识建模工具。在中医药领域，作为知识描述框架，本体的构建对象包括证候、疾病、症状、方剂、针灸、医案、中药等[24－27]；本体的构建方法多采用骨架法、七步法及其结合或改进方法，既可以根据步骤构建新的本体，同时可以提取、复用或扩展已有成果[28]。在本体构建过程中，遵循客观性、完整性、可扩展性、最小编码偏差等构建原则。

中医药领域的本体构建对象、构建方法和构建原则等为简帛医药文献的本体构建提供了借鉴。本研究面向中医药领域，主要采用七步法流程，遵循本体构建原则，以简帛医药文献作为数据来源，以疾病、中药、制法、用法实体等为构建对象，提炼简帛医药文献知识，分别构建简帛医药书目本体和简帛医药内容本体，并最终形成简帛医药文献本体模型。

3.1 书目本体构建

3.1.1 确定领域和范围

简帛医药文献外部特征并不明显，通过描述简帛医药文献的题名、作者、载体、位置等各种外部特征，组织散落在各个图书馆或机构的文献资源，深入揭示不同版本和校注之间的关联关系，实现简帛医药文献资源的组织、整理和利用。

3.1.2 复用现有本体

在设计本体时，应该尽可能地复用已有本体的类和属性[29]，通过部分复用已有的成熟本体，可以减少类和属性的重复构建，有利于知识的共享和重用。BIBFRAME 书目描述框架是新型书目数据格式，可以对现有的书目资源进行整合，对解决图书馆资源在网络上的关联与识别具有重要作用[30]。BIBFRAME 主要通过创造性作品（Work）、实例（Instance）、规范（Authority）、注释（Annotation）4 个核心部分描述书籍结构特征，每个类下有很多规范的属性和子类。通过文献调研和对简帛医药书目的综合分析，本文将基于 BIBFRAME 书目框架，复用部分类和属性设计简帛医药文献书目本体。

3.1.3 列举重要术语

元数据规范地定义了某个领域的核心术语，而古籍元数据标准为中医古籍数据化奠定了良好基础，具有较高的权威性和一定的参考借鉴意义。本研究列举了题名、作者等多个重要术语。

3.1.4 定义类的属性和关系

借鉴中医文献元数据（Traditional Chinese Medicine Literature Metadata，TCMLM）和 BIBFRAME 书目本体，进一步自定义类和属性，使简帛医药书目本体更加完整。基于 BIBFRAME 基本框架，复用其他类和属性，并定义新增加的类和属性，修改已有模型。表 1 通过与 TCMLM 的对照，展示简帛医药书目本体的类和属性，其中前缀"bm"为自定义的类和对象属性。例如，为"bf:HeldItem"类定义了"bm:heldHistory""bm:storage"等多个属性。

表 1 中医药文献元数据与简帛医药书目本体的对应关系

中医药文献元数据		简帛医药书目本体	
元数据项	元数据元素	类	属性
标识	标识符	bf:Identifer	bf:identifierType
			bf:identifierValue
题名项	题名	bf:Title	bf:titleType
			bf:titleValue
责任者项	创建者	bf:Agent	bf:creator
	贡献者		bf:contributor
格式	格式	bf:Format	bf:formatType
			bf:formatValue
类型	类型	bf:Category	bf:categoryType
			bf:categoryValue
附注项	描述	bf:Summary	bf:summaryContent
	主题		bm:summarySubject
	覆盖范围		bm:coverage
出版项	出版者	bf:Provider	bf:providerName
	出版地点		bf:providerPlace
	印刷地点		bf:printPlace
	日期		bf:providerDate
版本项	版本	bf:Edition	bf:edition
来源	来源	bm:Source	bm:source
馆藏项	存储地点	bf:HeldItem	bf:heldPlace
	收藏历史		bm:heldHistory
	破损级别		bm:damage
	珍稀程度		bm:rarity
	权限		bm:jurisdiction
	保存方式		bm:storage

3.1.5 可视化展示

与作品本身相关的特征，如题名、作者、摘要等作为医药文献作品的一部分，不随作品呈现方式的改变而改变。而与作品载体相关的部分，如版本、出版者、单件存储等特征是医药文献实例的组成部分。一个简帛医药文献作品可以有多个实例，而一个实例只属于一个作品。最终构建的简帛医药书目本体模型见图 1，图 1 中展示类与类之间的关系，子类用 subClassOf 表示。箭头表示属性，箭头起点和终点分别代表属性的域和范围。

3.2 内容本体构建

3.2.1 确定领域和范围

病方、古经脉学等医药文献的内容蕴含大量中医药知识，知识之间的隐含关系不易被发现，很难为非中医药领域的相关人员带来直接价值，因此需要构建简帛医药内容本体模型，刻画简帛医药的内容特征，发现其中重要概念。

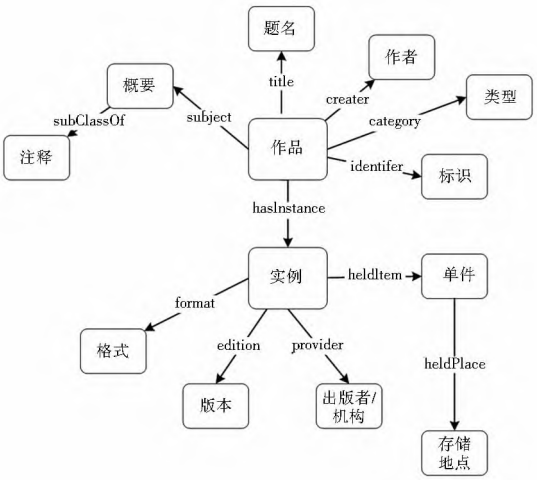


图 1　简帛医药书目本体模型

3.2.2　复用现有本体

遵循尽量复用已有本体模型和术语词表的本体构建原则，内容本体的设计将进一步借鉴中医药学语言系统(TCM Language System，TCMLS)。TCMLS 是面向中医药领域较为成熟的规范化顶层本体[31]，为中医药学语言系统中的所有概念提供了一体化的概念框架，建立了规范化的中医药术语体系[32]。简帛医药内容本体模型选择部分复用 TCMLS 本体中已有的概念和语义关系，通过语境和中医药领域的专业资料，结合简帛医药文献本身的特殊性进行相应扩充。

3.2.3　列举重要术语

依据本体构建数据源《简帛医药文献校释》以及《中医中药主题词表》和《中医药常用名词术语辞典》等中医药领域权威资料，列出重要术语。对于有争议的术语，则通过咨询领域专家进行重要性评估，选取认可度最广的术语。最终列出的重要术语包括疾病、证候、病因、病机、脏腑、经络、穴位、精、气、血、津液、中药、剂量、方剂、针灸等。

3.2.4　定义类和类的等级体系

类的等级关系主要是横向创建语义关系，描述概念的等级体系，本研究主要依据中医药领域的基本理论确定。以核心类“人体基础”为例，说明简帛医药核心概念的部分等级关系，见图 2。其中，脏腑、穴位、经络和基本物质为“人体基础”的子类。精、气、血和津液为“基本物质”的子类。

3.2.5　定义类的属性和关系

类的属性包括数据属性和对象属性。由于目前的中医药领域本体中对数据属性的定义较少，因此数据属性的定义主要参考《简帛医药文献校释》。定义类的对象属性即定义类之间的语义关系，包括等级关系

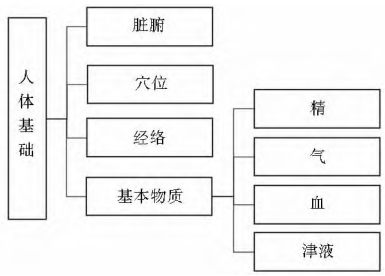


图 2　核心类“人体基础”中的等级关系

与非等级关系。上文已经构建类之间的等级关系，此处不再赘述。对于非等级关系，其定义首先考虑复用中医药领域核心概念间的语义关系，而后基于中医药基本理论，依据对文献内容的逻辑分析，确定概念之间的语义关系。简帛医药内容本体的类及其属性如表 2 所示：

表 2　内容本体的类和属性

类	属性/关系	属性/关系描述	类别
疾病	名称	规范的名称	数据属性
	别名	别名	数据属性
	疾病种类	属于的类别	数据属性
	并发症	疾病带来的并发症	数据属性
	预防	疾病的预防	数据属性
	现象表达	证候是疾病的现象表达	对象属性
	病机关系	疾病的作用机理	对象属性
治疗方法	名称	规范的名称	数据属性
	别名	别名	数据属性
	出处	治疗方法的来源	对象属性
	作用于	治疗方法作用于人体基础	对象属性
	治疗	治疗/被治疗的关系	对象属性
中药疗法	名称	规范的名称	数据属性
	别名	别名	数据属性
	类别	所属的类别	数据属性
	中药	组成方剂的中药	对象属性
	功效	方剂具有的功效	数据属性
	禁忌	方剂的禁忌症	数据属性
	不良反应	方剂的不良反应	数据属性
中药	名称	规范的名称	数据属性
	别名	别名	数据属性
	气味	四气和五味	数据属性
	毒性	中药的毒副作用	数据属性
	药物类别	药物所属的类别	数据属性
	对应	药物所对应的用量	对象属性
人体基础	名称	规范的名称	数据属性
	别名	别名	数据属性
	所属类别	所属的上级类别	数据属性
	分布规律	在人体内分布的规律	数据属性
	主	某部位对应的疾病或对应的证候	对象属性
病机	指导	病机对治疗方法的指导作用	对象属性
病因	导致	导致/引起/被引起的关系	对象属性
	影响	病因影响治疗方法	对象属性
制法	组成	是……的组成部分	对象属性
	使用	使用/被使用的关系	对象属性
用法	组成	是……的组成部分	对象属性

3.2.6　可视化展示

简帛医药内容本体模型见图 3，其中，虚线代表相关关系（语义关系），实线代表等级关系。例如，病因是疾病发生的原因，与疾病具有“导致”关系；证候是疾病的现象表达，与疾病之间具有“现象表达”的语义关系。

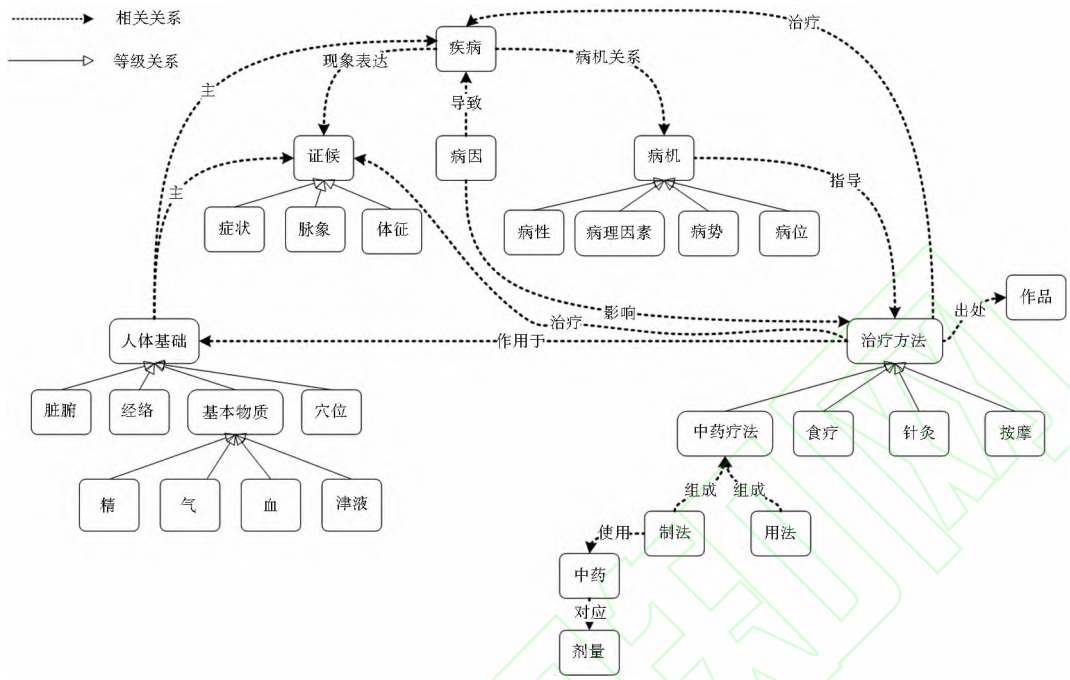


图 3　简帛医药内容本体模型

3.3　简帛医药文献本体构建

上文构建的简帛医药书目本体和简帛医药内容本体分别描述简帛医药文献资源的外部特征和内容特征。简帛医药内容本体中“治疗方法”类的一个对象属性是“出处”，通过该语义关系可与简帛医药书目本体中的“作品”类相关联，从而使简帛医药文献的内部特征与外部特征相互连接，形成简帛医药文献书目本体与内容本体之间的连接，实现简帛医药文献知识组织的统一，即完成简帛医药文献本体的构建。

4　简帛医药文献知识图谱构建方法

本研究采用自顶向下的方法构建简帛医药文献知识图谱，知识图谱的构建将遵循预先设计好的简帛医药内容本体模型进行实例化。运用本体构建方法，以《简帛医药文献校释》中的周家台秦简《病方》《五十二病方》《足臂十一脉灸经》和《阴阳十一脉灸经》为切入点，利用深度学习方法开展简帛医药文献领域命名实体识别和实体关系抽取任务，构建结构化知识库，本研究在简帛医药文献组织的研究方法上属于新的探索与尝试。

4.1　数据获取

简帛医药文献以竹简、帛书等形式承载，大多从各地出土并被收藏保存至相关机构，非中医药领域的专业人员很难直接对此类文献开展相关研究。本研究选择的数据源是《简帛医药文献校释》，该书基于文献整理的角度全面收集了目前已经公开的简帛医药文献资料，在充分参考学术各界对简帛医药文献整理的研究成果基础上，对每篇医药文献进行校核和注释，详细解释说明其内容，并对文章中的关键字词进行解释。此外，该书整理了存在破损残缺的简帛医药文献资源，并通过原文语义推断等方法补充了缺失的部分文献。在《简帛医药文献校释》中，1993 年湖北省荆州市出土的周家台秦简《病方》主要以疾病为核心论述对应的疾病、证候以及治疗方法；《五十二病方》是一本方剂学著作，记述了 52 种疾病的治疗方法，是迄今所见最早、最完整的古医方专著；1973 年在湖南长沙马王堆汉墓中出土的《足臂十一脉灸经》和《阴阳十一脉灸经》是现存最古老的灸疗文献，这些都是简帛文献中非常重要的部分。因此，将选取周家台秦简《病方》《五十二病方》《足臂十一脉灸经》和《阴阳十一脉灸经》作为分析数据构建知识图谱。

首先，通过百度 AI 开放平台对数据源图片进行 OCR 文本识别，将中文繁体转化为中文简体，并通过人工方式校对；其次，在古文与注释等共同存在的文本

中，仅选取“释译”中的文本，即分析已完成文言文翻译的文本；最后，由于出土简帛存在部分残缺无法完全恢复，仅选择非残缺的“释译”文本。最终获得可编辑的文本数据，其中《病方》19 条、《足臂十一脉灸经》12 条、《阴阳十一脉灸经》11 条、《五十二病方》73 条，共两万余字，部分数据如表 3 所示：

表 3 《简帛医药文献校释》“释译”文本(部分)

来源	“释译”文本
《病方》	将黑豆放入肥牛胆内，盛满之后再聚结起来，悬挂在避光阴暗的地方，让它干燥。需要用药的时候，从牛胆内取十来颗黑豆放在粥中，饮服，可以治疗痢疾。如果疾病不愈，再加量饮服。饮粥之量愈足以送黑豆入肠胃
《五十二病方》	一方：外伤，取黄芩末和……猪油……放进布袋中加压，滤出药汁……用药汁冲洗伤口处。一方：陈旧外伤，可将杏仁研碎，以黏油脂调和，涂敷在伤口上，可把伤口内的虫子驱出。此方经过应用疗效好。一方：把芒硝溶解于温水中，用以冲洗伤口。一方：金刃外伤止痛方：取鼷鼠，杀死晾干研末，取鲇鱼焙烤成炭，研末，再取长石、辛夷、甘草三药分别和鼷鼠等量，将以上诸药混合搅拌。取三指撮药末，放入一杯温酒里，饮服。如果服药效果不明显，再适足增加药量，直到不再疼痛，即停止服药。此方灵验
《足臂十一脉灸经》	臂阳明脉，从手中指中部出来，沿着肘骨外侧前缘，行经肱部外侧前缘，向上到达枕骨，止于口部。臂阳明脉所主的病症是：牙痛。凡是出现这一类症状，都可以灸臂阳明脉来治疗
《阴阳十一脉灸经》	臂少阴脉，起始于臂部尺骨、桡骨之间，沿着尺骨上侧，顺着臂筋下侧，从肱部内侧出来，再进入心脏。该经脉被外邪侵挠，会出现下列症状：心痛，咽喉干渴而想喝水，这就是臂厥病，以上各种病症都要以臂少阴脉为主来治疗。本脉自生的病变有：胸侧部疼痛，只有一种病症

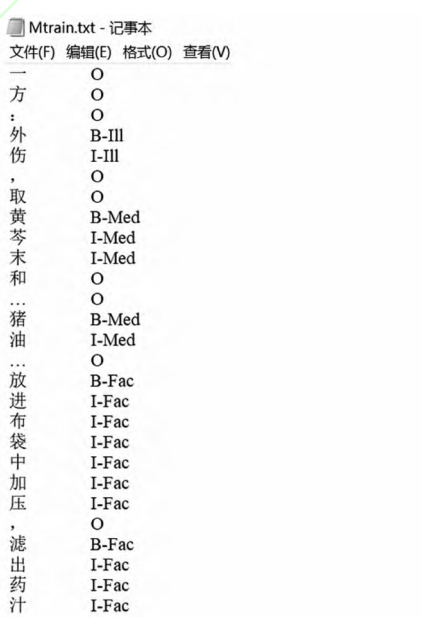
4.2 知识抽取

知识抽取包含实体识别、实体属性抽取以及实体关系抽取。由于所选简帛医药文献数据中关于实体属性的描述很少，因此关于实体属性以及数量很少的关系抽取主要采用人工方式筛选，本研究仅探讨存在数量较多的命名实体与实体关系的自动抽取。经过知识抽取后形成的数据集，可能存在表达冗余和语义歧义问题，而《简帛医药文献校释》充分参考了学术界的研究成果，是作者对公开简帛医药文献专业的校核和注释，术语规范[33]，不必进行额外的实体消歧等知识融合任务。

4.2.1 命名实体识别

中医药领域专业知识较强，需要辅以《简帛医药文献校释》校注及《中国中医药主题词表》等领域权威资料作为参考。基于这些参考资料，构建包含中药、疾病和经络词语的词典，便于后续实体标注和关系抽取的使用。

《足臂十一脉灸经》和《阴阳十一脉灸经》的文本表达较为统一规范，主要利用上述构建的词典进行分词，提取经络实体。对于《病方》和《五十二病方》的命名实体识别，主要采用条件随机场模型(Conditional Random Field，CRF)，CRF 模型是一种用于标注和切分有序数据的条件概率模型[34]，本文采用 CRF + + 进行系统实现。与本体中的相应类对应，主要标注中药、剂量、疾病、制法、用法 5 类实体。应用 BIO 三位标注法，随机选择《病方》和《五十二病方》中 70% 数据标注 5 类实体，实体标注规则为：中药类第一个字标为 B-Med、剂量类第一个字标为 B-Dos、疾病类第一个字标为 B-Ill、制法类第一个字标为 B-Fac、用法类第一个字标为 B-Use；中药类其他字标为 I-Med、剂量类其他字标为 I-Dos、疾病类其他字标为 I-Ill、制法类其他字标为 I-Fac、用法类其他字标为 I-Use，其他全部标为 O，具体标注情况见图 4。然后，利用标注的语料训练 CRF 模型。将剩余的 30% 数据作为测试集，利用 CRF 模型标注。最后，根据上述构建词典并结合人工审核，对 CRF 模型所标注的结果进行校正。


Mtrain.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V)

一	O
方	O
：	O
外	B-Ill
伤	I-Ill
，	O
取	O
黄	B-Med
芩	I-Med
末	I-Med
和	O
…	O
猪	B-Med
油	I-Med
…	O
放	B-Fac
进	I-Fac
布	I-Fac
袋	I-Fac
中	I-Fac
加	I-Fac
压	I-Fac
，	O
滤	B-Fac
出	I-Fac
药	I-Fac
汁	I-Fac

图 4 简帛医书实体识别语料标注示例

4.2.2 实体关系抽取

不同实体关系抽取方法的结合可以获得更好的效果[35]。由于《病方》《五十二病方》《足臂十一脉灸经》和《阴阳十一脉灸经》四部简帛文献的文字表达规律并不相同，因此分别采用基于词性标注与依存句法的方法以及基于规则的方法抽取三元组。

（1）基于词性标注与依存句法的实体关系抽取。《病方》和《五十二病方》主要是各种疾病的治疗方法。本研究将基于词性标注与依存句法抽取实体关系，主要包括中药疗法与制法的组成关系、中药疗法与用法的组成关系、制法与中药的使用关系、中药疗法与疾病的治疗关系、中药与剂量的对应关系。依照上文标注和抽取的实体结果，分别更新中药和疾病词典，并构建剂量、制法和用法词典，载入词典并依据词典进行分词处理。利用哈工大社会计算与信息检索研究中心研制的语言技术平台（Language Technology Platform，LTP）对已分词的结果进行词性标注和依存句法分析[36]。中药和疾病实体的词性应被标注为名词，制法和用法实体应被标注为动词，剂量实体应被标注为数量词或形容词。抽取实体关系三元组的流程见图 5，抽取规则如下：①被标注为名词的实体通过分别与中药词典和疾病词典进行匹配可确定其为中药实体或疾病实体。若为疾病实体，则可抽取出 < 中药疗法，治疗，疾病 > 三元组；若为中药实体，后面紧接剂量实体（词性标注为数量词或形容词）且依存句法关系为定中关系（ATT），则可抽取出 < 中药，对应，剂量 > 三元组。②被标注为动词的实体视为制法实体或用法实体，若该实体的依存句法关系为核心关系（HED），则可抽取出 < 中药疗法，组成，制法 > 三元组；若该实体的依存句法关系为并列关系（COO），则可抽取出 < 中药疗法，组成，用法 > 三元组。③在分别确定中药实体和制法实体之后，若中药实体的依存句法关系为主谓关系（SBV），则可抽取出 < 制法，使用，中药 > 三元组。

为保证准确定性，根据以上抽取规则得出的三元组仍需要结合进一步人工审核，最终确定中药疗法与制法的组成关系、中药疗法与用法的组成关系、制法与中药的使用关系、中药疗法与疾病的治疗关系、中药与剂量的对应关系。

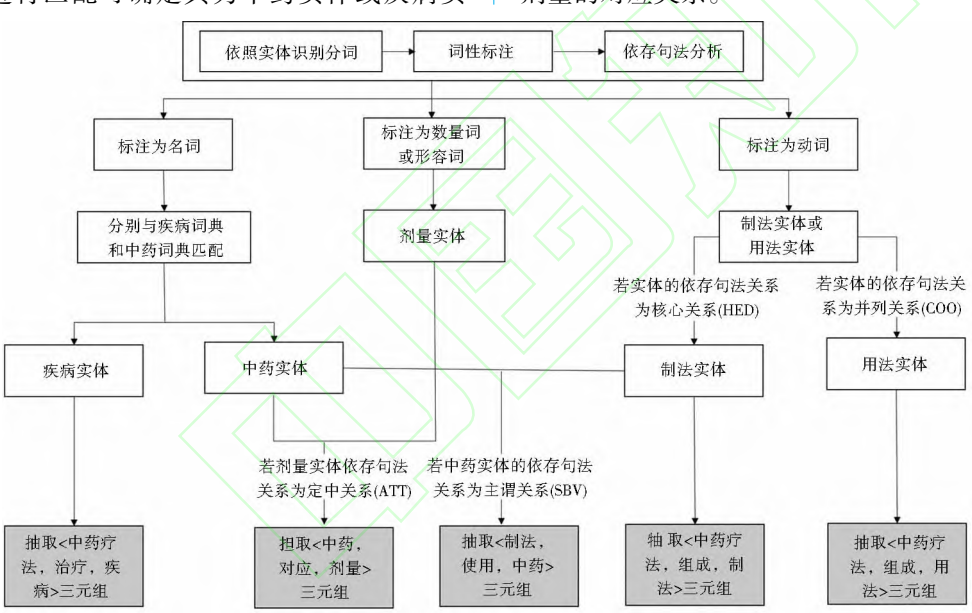


图 5　利用依存句法分析抽取实体关系三元组流程

（2）基于规则的实体关系抽取。由于《足臂十一脉灸经》和《阴阳十一脉灸经》的文本表达较为规范，如“某某脉所主的病症是：”“会出现下列症状：”“本脉自生的病变有：”，因此主要采用基于规则的实体关系抽取方法。首先，定位到句子中的上述表达和冒号，从冒号之前确定经络实体和关系（主）；然后，在冒号之后的语句中，通过逗号分隔，确定经络实体所主的证候实体或疾病实体，即抽取 < 经络，主，证候 > 或 < 经络，主，疾病 > 三元组。例如，图 6 展示了基于规则所抽取的“足少阳脉”与相应证候实体或疾病实体的关系。

足少阳脉	主	足小指和次指麻痹
足少阳脉	主	小腿外侧痛
足少阳脉	主	小腿感到寒冷
足少阳脉	主	膝关节外侧痛
足少阳脉	主	大腿外侧痛
足少阳脉	主	股骨大转子部外侧痛
足少阳脉	主	胸侧痛
足少阳脉	主	肩痛
足少阳脉	主	腋下发出臭气
足少阳脉	主	锁骨上窝处疼痛
足少阳脉	主	颈项瘫肿
足少阳脉	主	耳朵聋
足少阳脉	主	枕骨痛
足少阳脉	主	耳前痛
足少阳脉	主	外眼角痛
足少阳脉	主	前胸之侧痛

图 6　足少阳脉所主疾病/证候的关系抽取结果

4.2.3 实验结果与分析

（1）评价指标。命名实体识别实验和实体关系抽取实验的评价指标包括准确率（Precision）、召回率（Recall）、F1 值（F1-score）。准确率和召回率的取值在 0－1 之间，数值越大表明准确率或召回率越高。但是二者有时会出现矛盾的情况，需要综合考虑 F1 值，该值较高说明实验方法比较有效。相关计算公式分别为：

$$\text{Precision} = \frac{\text{CORRECT}}{\text{CORRECT} + \text{OVERPRED}} \times 100\% \quad \text{公式（1）}$$

$$\text{Recall} = \frac{\text{CORRECT}}{\text{CORRECT} + \text{MISSING}} \times 100\% \quad \text{公式（2）}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \quad \text{公式（3）}$$

其中，CORRECT 表示识别出的正确实体数或者正确抽取的三元组个数；MISSING 表示本该被正确抽取但未被抽取的实体个数或者本该被正确抽取但未被抽取的三元组个数；OVREPRED 表示本不该被抽取但被模型抽取出的实体个数或者本不该被抽取但被模型抽取出的三元组个数[37]。

（2）结果分析。《足臂十一脉灸经》和《阴阳十一脉灸经》的文本表达较为规范，实验结果的评价指标极高。在命名实体识别实验中，对比使用 CRF 模型所标注的测试集数据与经过人工审核校对的数据，分别计算准确率、召回率和 F1 值，计算结果均为 83.33%。在实体关系抽取实验中，由于所用简帛医药文献数据是自建数据集，缺少可以对比的基准模型与已经标注好的测试集。因此，在未经过人工审核校对的三元组数据中随机人工抽检 30% 数据集开展结果评测，分别计算准确率、召回率和 F1 值。基于词性标注与依存句法的实体关系抽取实验中计算结果分别为 100%、66.7% 和 80%。未经过人工审核校对的三元组数据仍存在一定误差，原因主要在于：①中医古籍文献中蕴含的思维很难让机器通过文字表象去学习[38]；②词性标注可能存在误差，由于制法和用法实体多为词组，词性标注不准确，例如“混合研磨”被标注为“时间名词”；③依存句法分析可能存在误差，一些比较复杂的句子无法很好地匹配，不利于三元组关系的抽取。

总体来看，对于本研究的简帛医药文献数据，当前知识图谱构建各环节所选择方法具有较好适用性，规范文本表达的知识抽取可以全自动识别且准确率极高，然而现有自动化方法仍然无法同时保证复杂句子知识抽取的高准确率、高召回率和高 F1 值。因此，为了提高知识图谱质量，仍要人工审核予以辅助。

4.3 知识图谱可视化与检索查询

4.3.1 知识图谱可视化

基于上述过程，利用 Neo4j 图数据库导入节点、属性和关系形成知识图谱可视化展示。

（1）《病方》知识图谱结果。图 7 展示从该文献中抽取的部分疾病、证候、中药疗法以及组成的中药、剂量等实体，并以知识图谱的方式组织部分实体属性和实体关系。其中，类“中药”匹配映射的实例为“乌头”、类“剂量”匹配映射的实例为“二七”、类“疾病”匹配映射实例为“弊病”；“乌头”与“二七”具有对应关系；制法实例“反复煮沸清洗”组成“中药疗法”，“中药疗法”与“弊病”具有治疗关系。因此，采用中药疗法治疗弊病，具体方法为将羊矢三斗、乌头二七、大如手的牛脂反复煮沸清洗然后服用。

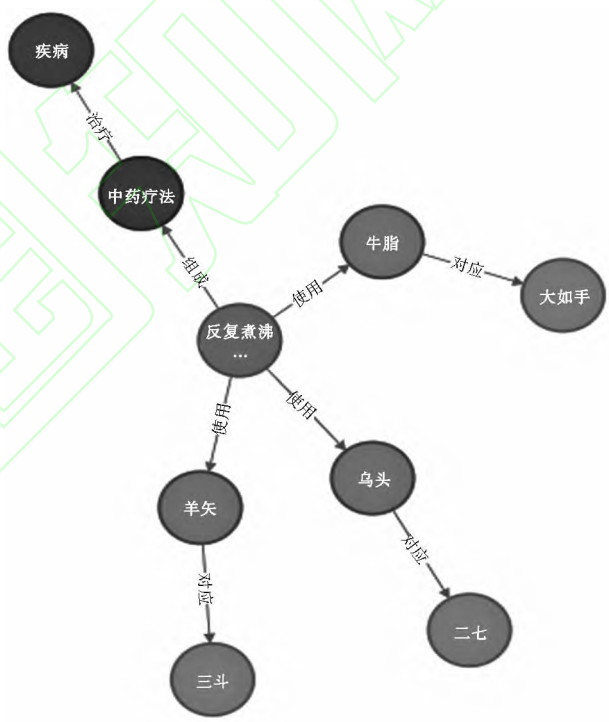


图 7 弊病治疗知识图谱

（2）《五十二病方》知识图谱结果。外科：《五十二病方》中记载的外伤主要采用中药疗法进行治疗，有三种疗法，其中一个治疗方法是将油脂、甘草、肉桂、干姜、蜀椒、茱萸等中药揉碎放到酒里饮服，见图 8。治疗金刃外伤同样有三种疗法，其中一个治疗方法是将羊粪焙烤成炭，外敷伤处，见图 9。

儿科：除外科疾病外，婴儿疾病也是《五十二病方》中非常重要的一部分。全身发热、腹部膨大等证候均是婴儿痫病的现象表达，该疾病可采用中药疗法治疗，见图 10。

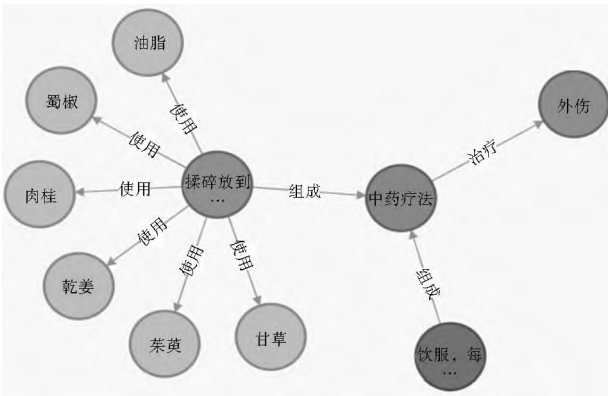


图 8　外伤治疗知识图谱

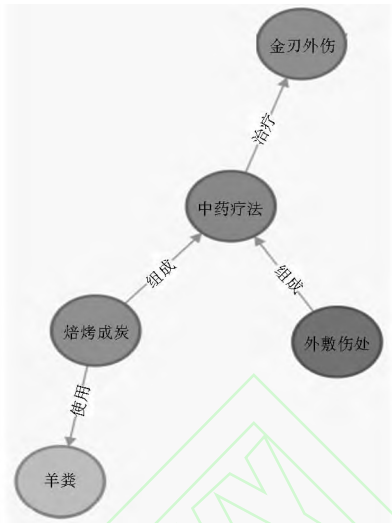


图 9　金刃外伤治疗知识图谱

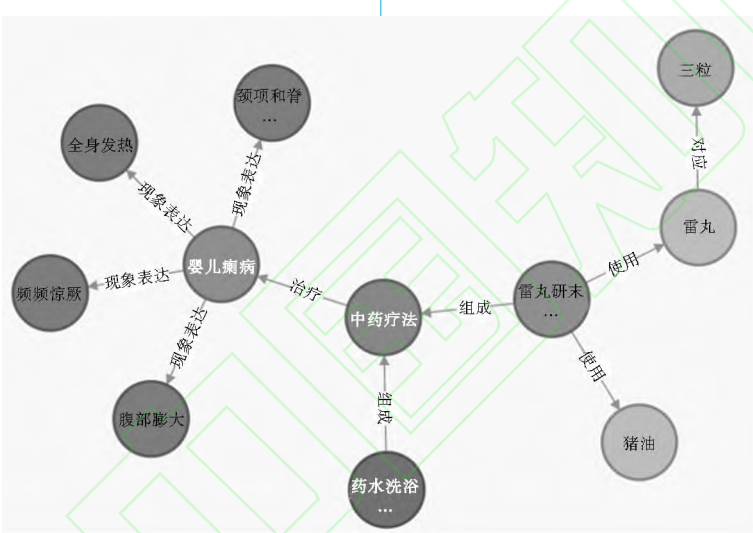


图 10　婴儿痫病治疗知识图谱

内科：《五十二病方》还包含一部分对内科疾病的阐述。痉病主要是由于在受到各种外伤后被风邪侵入体内等原因导致，该病的治疗方法有四种，其中一个治疗方法是将一把葱白放入半斗白酒中煮沸，然后饮服，见图 11。

（3）《足臂十一灸经》和《阴阳十一脉灸经》知识图谱结果。《足臂十一灸经》和《阴阳十一脉灸经》知识图谱主要展示经脉及其所主疾病或证候的中医知识，包括经络类的实例、疾病类的实例、证候类实例。以足少阴脉为例（见图 12），从知识图谱中可以看出足少阴脉所主证候主要包含肝痛、心痛、足部发热等十二项，所主疾病为哮喘。

4.3.2　知识图谱检索查询

通过构建 Cypher 语句，检索查询数据库中简帛医药文献中包含的全部中药信息，查询结果见图 13，包含鹿角、甘草、肉桂等多种中药实例。此外，点击节点会显示该节点的属性，如中药灌青的一个属性是“another_name”，该属性的值为灌，即中药灌青也可称作灌。再如使用 Cypher 中的 match 语句获取疾病与病因的语义关系，查询结果见图 14。根据查询结果可以获得婴儿索痉、痉病等疾病的病因。

5　结语

本研究以简帛医药文献为研究对象，通过构建简帛医药书目本体和内容本体进而构建简帛医药文献本体。应用自然语言处理等相关技术构建知识图谱，对简帛医药文献知识进行有效组织。

本研究在理论方面具有一定贡献。首先，扩展了知识组织在中医简帛医药文献方面的研究。传统的中医古籍知识组织模式较难适应数据化的发展需要，通

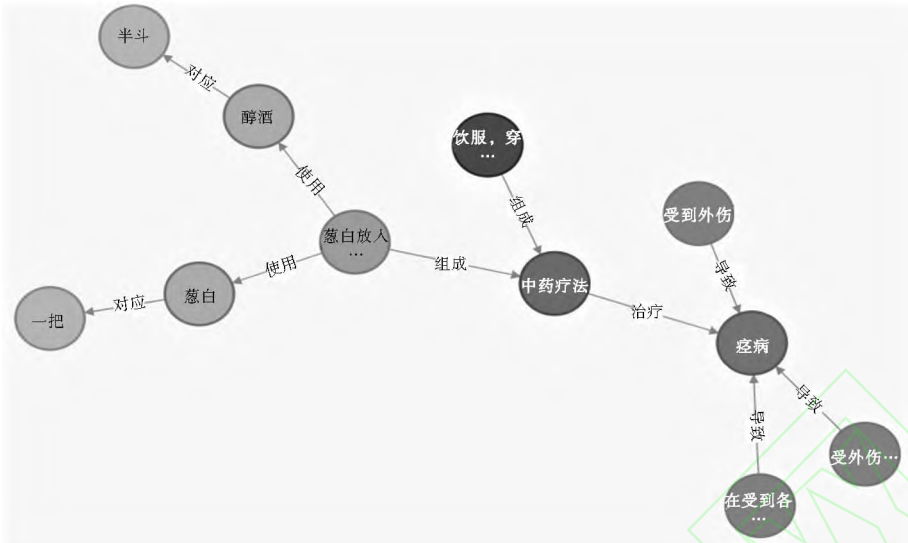


图 11　痉病治疗知识图谱

图 12　经络知识图谱

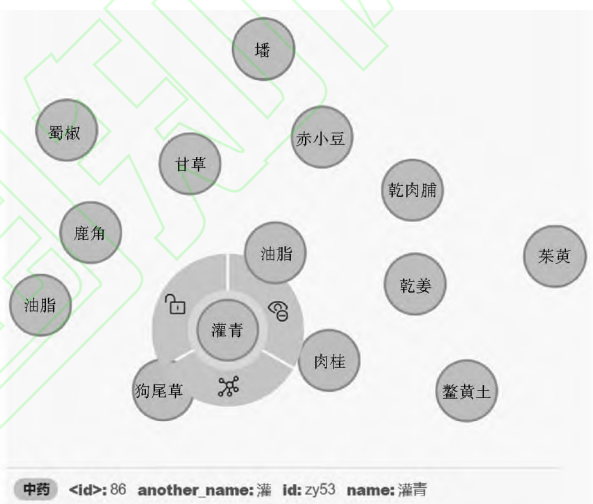


图 13　查询全部中药实体的查询结果

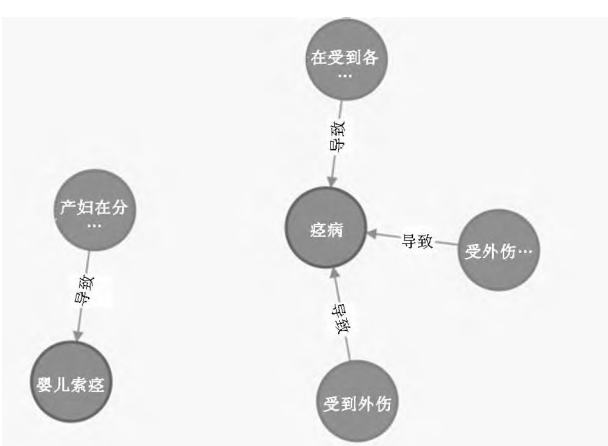


图 14　查询“导致”语义关系的查询结果

过引入信息组织中的本体和知识图谱的思想，为中医简帛医药研究与服务提供了一种新的视角。其次，结合书目本体和内容本体全面搭建了简帛医药文献本体模型。本研究分别构建简帛医药书目本体和内容本体，通过对书目本体和内容本体的连接，将散落在各个文献库中的异构简帛医药知识有机地融合起来，形成了完整且全面的本体框架。最后，有助于逐步建立起简帛医药文献领域的知识库。通过简帛医药文献知识图谱的构建，对简帛医药知识进行了有序组织，实现中医药资源的结构化表述和语义关联，在简帛医药知识图谱方面做出了有益的探索，有助于挖掘和发现隐含的知识。

本研究在实践方面具有一定贡献。首先，有利于

用户直观地掌握知识和知识之间的相互关系。知识并不是孤立存在,知识之间的相互作用具有更大的价值,通过具有逻辑关系的知识链条,满足用户全面获取中医知识的需求。其次,有利于中药知识的普及与传承。知识图谱在宏观上为用户展示了庞大的中医药知识体系,可以通过概念之间的跳转,获取、浏览、学习相关中医知识,降低了大众的认知负荷。通过现代化信息手段将中医药知识转化为图谱形式存储起来,有利于对中医药文化进行传播传承和创新发展。最后,有利于知识资源供应商或机构降低知识服务成本,提高服务的灵活性。相关机构可以基于构建的简帛医药文献本体模型,结合关联数据等技术构建更加大型的知识图谱,满足用户个性化搜索和深度获取资源的需求,提升用户体验。

本研究仍存在一定局限性,需要在未来研究中予以改进。其一,应进一步增加简帛医药文献和其他医药文献数据资源,丰富书目本体中的相关实体和内容本体中的相关实体属性,以帮助实验模型更准确地识别出实体和实体关系。其二,中医药知识的专业性较高,知识抽取的技术并不能完全实现自动化,仍要加入人工校正和专业知识的辅助,在未来研究中应进一步融入中医领域专家观点和知识,开展需求分析。最后,虽然本文在知识组织层面实现了简帛医药文献本体和知识图谱的构建,但对本体和知识图谱的应用研究方面仍存在不足,未来研究将以本体和知识图谱为基础,深入开展挖掘、推理、推荐等相关应用研究。

参考文献:

[1] 张如青. 出土简帛医书对澄清后世医学误解与疑义的作用[C]//中华中医药学会第二十二届医古文学术研讨会论文集. 北京:中华中医药学会,2013:26-31.

[2] UPADHYAY S, UPADHYAY N. Future directions and a roadmap in digital computational humanities for a data driven organization [J]. Procedia computer science, 2017,122:1055-1060.

[3] KAPLAN F. A map for big data research in digital humanities[J]. Frontiers in digital humanities,2015,2(1):1-7.

[4] 夏翠娟. 文化记忆资源的知识融通:从异构资源元数据应用纲要到一体化本体设计[J]. 图书情报知识,2021(1):53-65.

[5] 王大学,陈熙,杨光辉. 基于 GIS 的中国古籍地理信息系统研究[J]. 复旦学报(自然科学版),2016,55(6):684-688,697.

[6] 高晶晶. 中医古籍数字化生僻字的处理[J]. 中国中医药图书情报杂志,2014(3):28-30.

[7] 朱玲,朱彦,杨峰. 基于中医疾病相关语义关系的正则表达式及知识抽取研究[J]. 世界科学技术-中医药现代化,2016,18(8):1241-1250.

[8] 付璐,李思,李明正,等. 以清代医籍为例探讨中医古籍分词规范标准[J]. 中华中医药杂志,2018,33(10):4700-4705.

[9] 王国玺,李兵,张华敏,等. 基于知识组织的医案古籍知识库的构建与思考[J]. 西部中医药,2019,32(9):49-52.

[10] HERRERA-HERNANDEZ M C, LAI-YUEN S K, PIEGL L A, et al. A Web-based knowledge management system integrating Western and Traditional Chinese Medicine for relational medical diagnosis[J]. Journal of engineering in medicine, 2016,230(12):1601-1073.

[11] 汤伟平,许坚,李兵,等. 岭南温病古籍知识库构建的思路探讨[J]. 中国医药导报,2020,17(11):136-139.

[12] ABOUENOUR L, NASRI M, BOUZOUBAA K, et al. Construction of an ontology for intelligent Arabic QA systems leveraging the conceptual graphs representation[J]. Journal of intelligent & fuzzy systems, 2014, 27(6):167-167.

[13] CHEN J, OU S. Research on the construction of the semantic model for Chinese ancient architectures based on architectural narratives [J]. The electronic library,2020,38(4):769-784.

[14] FERRE S. Link prediction in knowledge graphs with concepts of nearest neighbours [C]// European semantic Web conference. Berlin: Springer,2019:84-100.

[15] 王菁薇,肖莉,晏峻峰. 基于 Neo4j 的《伤寒论》知识图谱构建研究[J]. 计算机与数字工程,2021,49(2):264-267,396.

[16] 曾子玲,张华敏,于彤,等. 知识图谱及其关键技术在中医药领域的研究与应用综述[J]. 世界科学技术-中医药现代化,2022,24(2):780-778.

[17] CHENG M, XIONG S F, LI F, et al. Multi-task learning for Chinese clinical named entity recognition with external knowledge[J]. BMC medical informatics and decision making, 2021, 21(1):1-11.

[18] 王蕊,于彤,符永驰,等. 中医特色疗法知识地图的初步构建与研究[J]. 中国数字医学,2015,10(9):86-88.

[19] 刘燕,傅智杰,李姣,等. 医学百科知识图谱构建[J]. 中华医学图书情报杂志,2018,27(6):28-34.

[20] 张德政,谢永红,李曼,等. 基于本体的中医知识图谱构建[J]. 情报工程,2017,3(1):35-42.

[21] 罗计根, 杜建强, 聂斌, 等. 基于双向 LSTM 和 GBDT 的中医文本关系抽取模型[J]. 计算机应用研究, 2019, 36(12):3744-3747.

[22] 彭博. 数字人文视角下的网络文物信息资源知识图谱构建[J/OL]. 图书馆论坛:1-11[2022-04-25]. http://kns.cnki.net/kcms/detail/44.1306.g2.20220317.1011.002.html.

[23] GRUBER T R. Toward principles for the design of ontologies used for knowledge sharing[J]. International journal of human-computer studies,1995, 43(5/6): 907-928.

[24] 王明强,张磊,崔一迪,等. 利用 Neo4j 存储中医皮肤病"病-证-治"本体方法的研究[J]. 世界科学技术-中医药现代化,2020,22(8):2914-2921.

[25] 崔一迪,王明强,陈欣然,等. 痤疮的中医药本体构建研究[J]. 世界科学技术-中医药现代化,2019,21(12):2867-2872.

[26] 刘畅. 基于本体的岭南灸法古籍知识库构建研究[D]. 广州:广州中医药大学,2019.
[27] 乔幸潮. 中药材本体构建研究[D]. 北京:中国中医科学院,2020.
[28] 于彤,刘静,李海燕,等. 中医药子本体抽取方法研究进展[J]. 中国数字医学,2017,12(1):26－28.
[29] 夏翠娟,刘炜,张磊,等. 基于书目框架(BIBFRAME)的家谱本体设计[J]. 图书馆论坛,2014,34(11):5－19.
[30] 王兴兰. BIBFRAME 的发展及其在国内外图书馆的应用[J]. 图书馆杂志,2022,41(3):79－87.
[31] 尹爱宁,张汝恩. 建立《中医药一体化语言系统》[J]. 中国中医药信息杂志,2003,10(3):90－91.
[32] 于彤,崔蒙,李海燕,等. 中医药学语言系统的语义网络框架:一个面向中医药领域的规范化顶层本体[J]. 中国数字医学,2014,9(1):44－47.
[33] 周祖亮,方懿林. 简帛医药文献校释[M]. 北京:学苑出版社,2014.
[34] MIHALCEA R, TARAU P. TextRank: bringing order into texts [C]// Proceedings of the 2004 conference on empirical methods in natural language processing. Barcelona: Association for Computational Linguistics, 2004: 404－411.
[35] 张莹莹. 基于知识图谱的舌像诊疗系统研究与构建[D]. 绵阳:电子科技大学,2019.
[36] WANXIANG C, ZHENGHUA L, TING L. LTP: A Chinese language technology platform[C]// International conference on computational linguistics: demonstrations. Beijing: ACL, 2010: 13－16.
[37] 杨普全. 名人专题数据平台知识图谱构建方法研究与应用[D]. 上海:东华大学,2022.
[38] 杨雅斐. 基于本体的数据挖掘法对《金匮要略》黄疸病源流的古籍文献研究[D]. 成都:成都中医药大学,2020.

作者贡献说明:
李贺:负责写作指导与审阅,论文定稿;
祝琳琳:负责论文撰写;
刘嘉宇:负责论文修改;
樊娇:负责文献调研与论文撰写;
侯力铁:负责中医相关知识指导。

Research on the Organization of Bamboo and Silk Medical Knowledge Based on Ontology

Li He[1] Zhu Linlin[1] Liu Jiayu[1] Fan Jiao[1] Hou Litie[2]
[1] School of Business and Management, Jilin University, Changchun 130012
[2] Changchun University of Chinese Medicine, Changchun 130117

Abstract: [**Purpose/Significance**] Traditional Chinese medicine is the inheritance of Chinese culture and plays an important role in the development of modern medicine. It should be highly valued. It is of great significance to use digital technology to organize, protect and reuse ancient Chinese medicine books. [**Method/Process**] The article took bamboo and silk medical manuscripts as the research object, and constructed the bibliographic ontology and content ontology of bamboo and silk medical manuscripts, respectively. Through the connection of the bibliographic ontology and the content ontology, the ontology model of bamboo and silk medical manuscripts was formed; The named entity recognition model was used to extract entity concepts, and the method of dependency syntax parsing and rules was used to extract entity relationships. Finally, the data of the bamboo and silk medical manuscripts was stored in the graph database, and the visualization and retrieval query results of the knowledge graph for bamboo and silk medical manuscripts were presented. [**Result/Conclusion**] The proposed knowledge organization method using the ontology and knowledge graph establishes a logical knowledge chain of the bamboo and silk medicine, and forms a semantic network of domain knowledge in the Chinese medicine, which makes useful explorations in the intelligence of ancient Chinese medicine books, contributing to the popularization and inheritance of Chinese medicine knowledge.

Keywords: bamboo and silk medical manuscripts knowledge organization ontology knowledge graph digital humanities