

中医脾胃病临床文献大数据平台构建研究*

田野 于琦 朱玲 于彤 李宗友 张竹绿 姜威 王一萌 李菲
李敬华

(中国中医科学院中医药信息研究所 北京 100700)

〔摘要〕 提出运用计算机技术对中医脾胃病相关古今文献进行整合、存储、分析、加工与处理,形成中医临床文献大数据平台,详细阐述平台设计、资源整合方式、平台功能及应用价值。

〔关键词〕 中医;脾胃病;知识库;知识服务

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.01.014

Study on the Building of Big Data Platform for Clinical Literature of Spleen and Stomach Disease of Traditional Chinese Medicine TIAN Ye, YU Qi, ZHU Ling, YU Tong, LI Zongyou, ZHANG Zhulv, JIANG Wei, WANG Yimeng, LI Fei, LI Jinghua, Institute of Information on Traditional Chinese Medicine, CACMS, Beijing 100700, China

〔Abstract〕 The paper proposes to use computer technology to integrate, store, analyze and process ancient and modern literature related to spleen and stomach disease of Traditional Chinese Medicine (TCM) to form a big data platform for clinical literature of TCM, and expounds the design, resource integration mode, functions and application value of the platform in detail.

〔Keywords〕 Traditional Chinese Medicine (TCM); spleen and stomach disease; knowledge base; knowledge service

1 引言

中医学文献内容丰富,运用计算机技术对脾胃

病古今文献进行处理、加工,形成基于脾胃病的中医临床文献大数据平台专题知识库是脾胃病研究的一项基础内容。以中医脾胃病相关诊疗内容为核心,对中医药领域知识体系进行系统梳理,采集、加工高质量知识内容,建立各类知识点之间关联,基于古今文献、高质量证据、最新研究进展、专家论著建成中医临床文献大数据平台知识库系统。本系统整合名医经验、临床研究、临床指南、中医医案、诊疗技术、中医文献和方药知识等资源,有助于中医药知识共享和中医脾胃病诊疗技术研究,可与中医临床智能系统^[1]对接以辅助临床决策。本文对平台构建方法及知识内容进行详细介绍。

2 平台设计

〔收稿日期〕 2020-04-12

〔作者简介〕 田野,助理研究员,发表论文5篇;通讯作者:李敬华,副研究员。

〔基金项目〕 中国中医科学院基本科研业务费自主选题项目“中医脾胃病临床文献大数据平台构建及其示范应用研究”(项目编号:ZZ120305);国家自然科学基金面上项目“基于深度学习的脾胃病关键诊疗要素发现与临床优化决策研究”(项目编号:81873200);中国中医科学院基本科研业务费自主选题项目“中医传承工作站的系统集成与探索应用”(项目编号:ZZ130303)。

2.1 平台搭建

在充分调研相关文献^[2]、分析领域需求基础上设计基于脾胃病的中医临床文献大数据平台数据模型，确定元数据加工审核、人工语义标注、机器语义标注等一系列数据表字段类型、字段内容及表关系等，基于 Lucene、Mongodb 等开源技术搭建平台。收集脾胃病相关中医药领域纸书、期刊、古籍等文

献资源，经过光学字符识别（Optical Character Recognition, OCR）、文件格式转换（如将 PDF 格式文件转换为 TXT 格式文件）、元信息标注（如按古代、现代、主题内容、作者、来源等）等处理之后，将全文内容存入中医临床文献大数据平台。

2.2 平台架构（图 1）

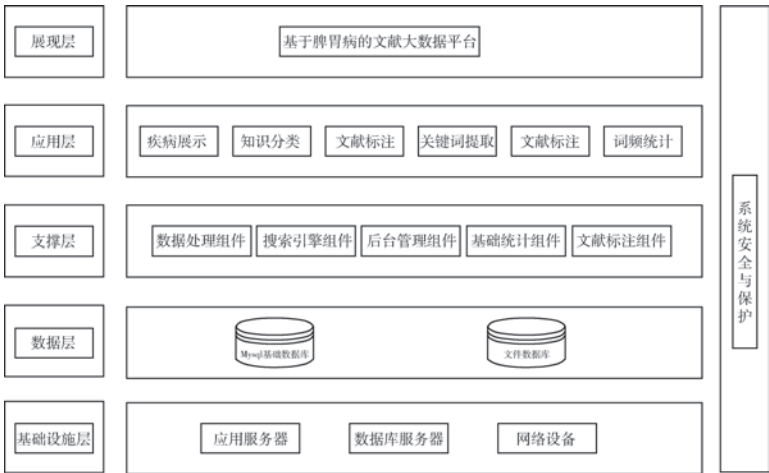


图 1 平台整体架构

2.2.1 基础设施层 位于最底层，是大数据服务平台最基础部分，用以实现对物理服务器以及相关网络与安全组件的虚拟化功能，对应系统包括服务器、网络系统、存储系统、安全系统、运维系统和其他基础系统。

2.2.2 数据资源层 位于基础设施层之上，将数据库（关系/非关系型数据库）、Web 资源、文件系统等异构数据源进行抽取及清洗、转换、加工后，按统一数据标准规范加载到各类数据库中，通过构建数据库、对象数据库和索引数据库对各类资源进行分布式存储与管理。

2.2.3 应用支撑层 位于数据资源层之上，数据资源层分散、异构的应用和信息资源在该层进行聚合，通过统一访问入口实现结构化数据资源、非结构化文档和互联网资源、各种应用系统跨数据库、跨系统平台无缝接入和集成，提供支持信息访问、传递以及协作的集成化环境，实现个性化业务应用的高效开发、集成、部署与管理。采用面向对象、组件式设计等多项技术提供的构件系统，是跨领域、与具体业务无关、通用的基础服务，能随着业

务系统发展变化而扩展、伸缩。应用系统一般通过应用开发接口或者声明性描述使用构件服务，通过组件对应用系统进行再构造，形成更强大的系统。

2.2.4 应用层 面向用户服务，该层使用应用支撑层提供的构件服务，向用户提供类型丰富的业务服务系统。

2.2.5 展现层 该层为平台提供对外服务的系统展示页面。

3 资源整合

3.1 资源获取

3.1.1 构建数据仓储系统 通过构建数据仓储系统对中医资源中的结构化、非结构化文档和互联网资源进行基于数据的整合、组织和存储，初步构建各类信息资源的数据仓储中心，为后续资源利用和服务提供基础。数据仓储系统将打破现有资源库的存储方式，通过本地数据中心实现对图书、期刊、论文、图片、音视频等多类型资源的数据收割，通过构建数据库、对象数据库和索引数据库，对各类

资源实现分布式存储与管理。

3.1.2 数据资源获取系统功能 可以对互联网网站内容进行采集,对期刊、论文等文献进行数据采集等。对采集信息实现长期保存,支持与数据仓储间的数据集成与整合,使采集信息无缝接入本地数据仓储,从而为资源检索与获取提供服务^[3]。

3.1.3 平台服务 可提供包括元数据仓储、数据资源获取(对互联网站内容采集、纸书数据采集等)、统一检索(平台提供数据检索)、中医文献大数据分析(语义关联检索、词表抽取、主题汇聚)在内的多种服务。

3.2. 资源利用

3.2.1 概述 通过人工标注(用于强监督学习)和机器标注(用于弱监督学习)相结合的方法对中医药文本进行概念实体、语义关系和内容标注,从而生成语料库并进行机器学习算法测试和模型训练^[4]。

3.2.2 概念标注 对脾胃病中医临床医案文本中的证候、症状、病因、病机、中药、方剂等概念实体进行人工标注,可通过语义标注软件将文本中的概念实体标注为不同颜色。使用命名实体识别算法进行测试,训练命名实体识别模型。

3.2.3 关系标注 对脾胃病中医临床医案文本中的病证证相关关系、方证关系、药证关系、方剂组成关系等语义关系进行人工标注,可通过语义标注软件将文本中的两个概念实体连接起来,标注关系的种类(也可用不同的颜色表示)。使用语义关系抽取算法进行测试,训练命名实体识别模型。

3.2.4 内容标注 标出方剂以及医案中的各种内容,使用自动标引算法进行测试,训练出知识点抽取模型。其中重点测试医案结构化算法效果:拟利用原始医案(文本)及结构化医案(弱监督)生成训练语料,训练医案知识自动抽取模型,实现医案文本自动结构化。

3.2.5 知识图谱生成 基于命名实体识别模型和关系抽取模型,对整个语料集进行实体和语义关系自动标注,在此基础上自动构成知识图谱^[5]。平台搭建及服务流程,见图2。

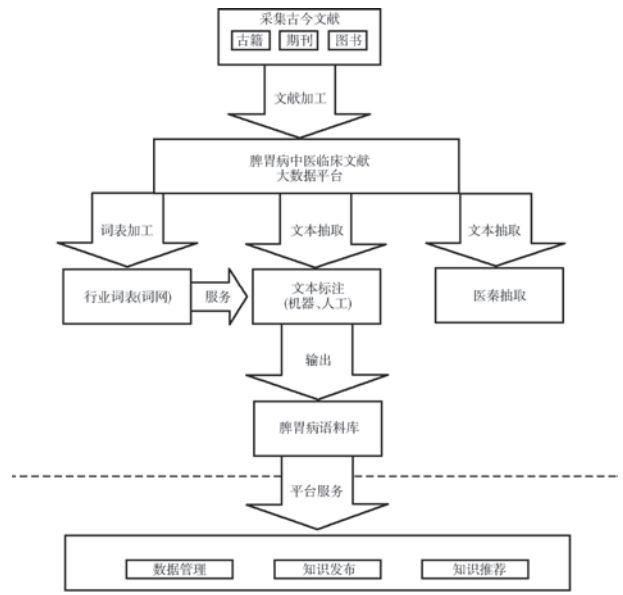


图2 平台搭建及服务流程

4 平台功能

4.1 知识检索与可视化

将包括胃痛、胃脘痛、痞满、噎膈、胃缓、痢疾、食积、胃肠功能障碍、胃肠神经官能症、功能性便秘、胃炎、胃溃疡、十二指肠溃疡、功能性消化不良、肠易激综合征、胃肠痉挛、消化道出血等疾病在内的脾胃病相关中医药期刊、书籍等文献数据整合为相同结构数据并上传平台,使用者可依据不同需求统一查询、检索、调用数据。可提供数据格式包括 EXCEL、PDF、XML。

4.2 语料库生成

通过人工标注(用于强监督学习)和机器标注(用于弱监督学习)相结合方法,对中医药文本中的中医疾病、西医疾病、证候、症状体征、方剂、中药等内容进行概念实体标注、语义关系标注和内容标注,从而生成脾胃病相关语料库,基于语料库进行机器学习算法测试和模型训练。

4.3 知识抽取

对期刊文献 PDF 文件进行格式转换,以便通过制定规则对期刊文献中的医案内容进行抽取。现代期刊中对于医案的描写方式基本规则化,仅有少许差别,对姓名、性别、年龄、诊次、舌质舌苔、脉

象、中医诊断、西医诊断、方剂、中药组成等内容进行规则制定,实现期刊中医案内容抽取^[6]。

4.4 平台服务示范

4.4.4.1 脾胃病相关知识挖掘 分别选取 10 个脾胃病中医及西医疾病，基于词网对全文献进行内容标注，采用机器自动标注，将标注后的内容进行人工审核并进行再标注，以知识图谱形式进行展示。展示内容主要包括中医疾病、西医疾病、症状、证候、方剂、中药。以列表形式列出文献中与某一脾胃病相关联的中西医疾病、症状、证候等内容段落章节，点击节选内容可链接原文进行内容查看。

4.4.4.2 词频统计 平台全面收录中医脾胃病诊疗相关古籍文献、科普文献、期刊文献、书籍资源,以10个中医定义的脾胃病及其对应的西医疾病为示例,收录相关病证文献题名、期刊来源、年、卷、期、页码、类型、作者、单位、省份、基金、人名、标签、关键词、主题词、分类号、摘要等信息。提供10种中医疾病及西医疾病的症状、名医、方药的词频统计结果,见图3。

4.4.4 医案抽取 通过制定规则分别对期刊医案中涉及的诊次、舌质舌苔、脉象、辨证、中医诊断、西医诊断、中药组成、方剂、按语等内容制定抽取规则，对期刊中医案内容进行抽取。如文中提及的 XXXX、XX、XXX 诊、X 诊（XXXX 年 XX 月 XX 日）、X 剂后 X 诊、X 日后 X 诊，则判定为诊次信息；XX 舌、XX 苔、舌 XX、苔 XX、舌质 XX，苔 XX 等，则判定为舌质舌苔信息进行内容抽取。

5 结语

本项目在对中医药防治脾胃病相关文献进行系统梳理和深度分析基础上,充分利用中医药领域术语系统、知识库和文献等数字资源,采用专家协作加工与人工智能方法相结合的方式构建中医临床文献大数据平台专题知识库,基于专题知识库搭建中医脾胃病临床文献大数据平台,实现知识发布、知识导航、知识浏览、知识检索、知识发现等知识服务,为临床、科研、教学提供技术支持。目前临床医生及科研人员可通过网站访问平台。后续将持续对后台知识库及前台服务平台内容进行扩容,以期针对更多脾胃系统疾病进行完整的内容补充及服务提供。

参考文献

- 1 朱玲, 李敬华, 于彤, 等. 基于本体的功能性胃肠病病知识库构建研究 [J]. 湖南中医药大学学报, 2018, 38 (10): 1154-1160.
- 2 于彤, 李敬华, 于琦, 等. 中医药防治哮喘病知识服务平台研发 [J]. 中国数字医学, 2017, 12 (6): 85-87, 90.
- 3 李敬华, 于彤, 朱玲, 等. 功能性胃肠病中医临床知识库构建研究 [J]. 世界科学技术-中医药现代化, 2018, 20 (3): 349-354.
- 4 于琦, 李敬华, 徐丽丽, 等. 中医医案与名医传承技术平台建设实践 [J]. 科技新时代, 2017 (4): 55-56.
- 5 于彤, 李敬华, 于琦, 等. 中医养生知识图谱的构建与应用 [J]. 中国数字医学, 2017, 12 (12): 64-66.
- 6 Yu Tong, Chen Huajun, Mi Jinhua, et al. DartWiki: a semantic wiki for ontology-based knowledge integration in the biomedical domain [J]. Current Bioinformatics, 2012, 7 (3): 278-288.

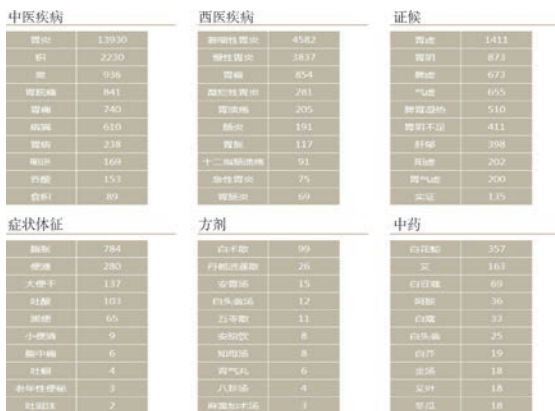


图3 基于脾胃病的中医临床文献大数据平台词频统计页面

4.4.3 关键词提取 将文献中的词汇按照频次进行提取,通过图片形式对与脾胃病相关联的关键词汇进行展示,见图4。



图4 关键词汇展示