基于 Neo4j 的中医妇科知识图谱构建研究

周 涛[1] 常 凯[1]
ZHOU Tao CHANG Kai

摘 要 通过详细阐述中医妇科知识图谱的构建流程，包括本体层的设计、实体与关系的知识抽取、数据的知识融合，以及最终的知识存储和可视化展示，提供了一种新的视角来理解和应用中医妇科知识，为中医学的研究和临床诊疗提供参考和借鉴。在本体层设计阶段，采用自顶向下的方法，运用 Protégé 软件进行本体编辑和构建，确保了本体结构的科学性和系统性。在知识抽取阶段，利用 Bi-LSTM-CRF 模型结合人工定义规则，有效提取了中医妇科文本中的关键实体和关系。在知识融合阶段，针对中医术语的复杂性和多样性，采用人工方法进行数据的整合和对齐，增强了数据的准确性和一致性。最后，在知识存储和展示阶段，通过 Neo4j 图数据库实现了知识的有效存储和直观展示。

关键词 知识图谱；中医妇科；知识抽取；Neo4j

doi：10.3969/j.issn.1672-9528.2024.06.046

0 引言

中医妇科是中医学的重要组成部分，其理论体系包括脏腑经络理论、阴阳五行理论等。在长期的实践中，中医妇科领域积累了丰富的经验，形成了独特的辨证论治方法和方药应用体系，如中医妇科强调整体观念和辨证论治，针对女性的生理特点和疾病特点，提出了补肾益精、调理冲任、疏肝解郁等治疗原则，其理论体系和实践经验丰富，对于女性的健康和疾病的防治具有重要意义。

知识图谱技术是一种强大的知识表示方法，能把一些复杂知识领域挖掘出来的数据和信息，通过可视化的图形象地展示，能较好地揭示事物的内在本质关系和发展趋势，为科学研究提供有价值的参考。在现代中医药领域，知识图谱能够为中医临床诊治提供方向，其应用领域越来越广[1]。本文以中医妇科疾病为研究对象，使用 Neo4j 图数据库软件可视化显示中医妇科中疾病、症状、方剂、药物、治法之间的联系，构建了中医妇科知识图谱，为中医妇科的研究与临床诊疗提供一定的借鉴和参考。

1 中医妇科知识图谱构建

1.1 知识图谱构建整体流程

本研究构建中医妇科知识图谱的流程如下：首先编写爬虫代码，获取中医妇科相关数据，使用人工标注进行数据预处理得到语料库，在中医临床专家和医生的指导下，参考中医妇科诊疗相关文献资料，设计中医妇科知识本体层；其次将预处理后的中医妇科数据进行实体和关系抽取，本研究是基于 Bi-LSTM-CRF 算法模型和人工定义规则进行实体和关系抽取；然后按照相关的规则，进行知识融合和实体对齐操作，完成对实体名称的规范统一；最后利用 Neo4j 图数据库软件对中医妇科知识进行存储和可视化显示，构建流程如图 1 所示。

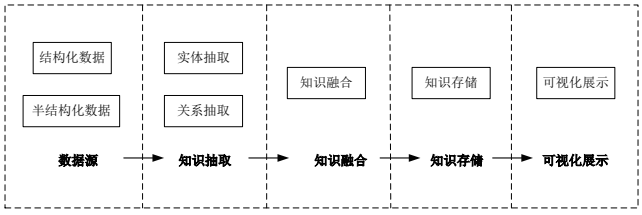

图 1 中医妇科知识图谱构建流程

1.2 数据来源和预处理

本研究使用的中医妇科数据来源于寻医问药（https://www.xywy.com/）、中国医药信息查询平台（https://www.dayi.org.cn/）和三九医药网（http://www.39yao.cn/）三个百科类知识网站。获取到中医妇科的数据后，对数据的预处理，流程如下：首先使用 Python 爬虫程序获取中医方剂数据集和中医药物数据集，由于获取的数据集大都是非结构化或半结构化的文本数据，爬取的数据大部分都是文本数据，由于中医文本具有特殊性，数据爬取完成后需要按照一定的规则进行数据清洗，去掉无关、冗余的信息。疾病和方剂部分的数据，主要是通过人工标注的方法进行，然后把人工整理后的数据按照方剂和药物进行分类，把对应的序号、方名、主治、功用等记录到 Excel 中，整合所有人工标注的数据，形成中医妇科文本语料库，相关数据示例如图 2 所示。

1. 湖北中医药大学 湖北武汉 430065

序号	方名	规范名	经典	出处	主治	功用	功用大类	功用小类	处方
10000002	桂枝汤	桂枝汤	经典	《伤寒论》	外感风寒。	解肌发表，	解表方	辛温解表	桂枝（9克） 芍药（9克） 生姜（9克） 大枣（3枚） 甘
10000007	正柴胡饮	正柴胡饮	经典	《景岳全书	外感风寒轻	解表散寒。	解表方	辛温解表	柴胡（9克） 防风（3克） 陈皮（4.5克） 芍药（6克）
10000050	加减葳蕤汤	加减葳蕤汤	经典	《重订通俗	素体阴虚，	滋阴解表。	解表方	扶正解表	生葳蕤（9克） 生葱白（6克） 桔梗（4.5克） 东白薇（
10000082	麻子仁丸（又名	麻子仁丸	经典	《伤寒论》	肠胃燥热，	润肠泄热，	泻下方	润下	麻子仁（500克） 芍药（250克） 枳实（250克） 大黄
10000083	济川煎	济川煎	经典	《景岳全书	老年肾虚。	温肾益精、	泻下方	润下	当归（9～15克） 牛膝（6克） 肉苁蓉（6～9克） 泽泻
10000086	五仁丸	五仁丸	经典	《世医得效	津枯肠燥证	润肠通便。	泻下方	润下	桃仁(30克） 杏仁(麸炒，去皮尖，30克） 松子仁（5克）
10000211	当归建中汤	当归建中汤	经典	《千金翼方	产后虚羸	温补气血，	温里方	温中祛寒	当归四两，桂心三两，甘草（炙）二两，芍药六两，生
10000220	当归生姜羊肉汤	当归生姜羊肉	经典	《金匮要略	产后血虚，	补气养血，	温里方	温中祛寒	当归 生姜 羊肉
10000243	补中益气汤	补中益气汤	经典	《内外伤辨	1．脾虚气	补中益气，	补益方	补气	黄芪（18克） 炙甘草（9克） 人参（6克） 当归（3克）
10000279	四物汤	四物汤	经典	《太平惠民	冲任虚损。	补血调血。	补益方	补血	当归（10克） 川芎（8克） 白芍（12克） 熟地（12克）
10000280	当归补血汤	当归补血汤	经典	《内外伤辨	血虚阳浮发	补气生血。	补益方	补血	黄芪（30克） 当归（6克）
10000293	补血定痛汤	补血定痛汤	经典	《万病回春	小产后瘀血	活血， 止痛	补益方	补血	当归 川芎 熟地黄 白芍 延胡索 桃仁 红花 香附 青皮
10000295	神应养真丹	神应养真丹	经典	《三因极一	四气侵袭肝	滋肝补肾、	补益方	补血	当归 天麻 川芎 羌活 白芍 熟地黄
10000306	通乳丹	通乳丹	经典	《傅青主女	产后气血两	补气养血，	补益方	气血双补	人参 黄芪 当归 麦冬 桔梗 川木通
10000309	加味圣愈汤	加味圣愈汤	经典	《医宗金鉴	产后血虚，	补气养血，	补益方	气血双补	人参 黄芪 当归 川芎 熟地黄 白芍 盐杜仲 续断 砂仁
10000316	当归芍药汤	当归芍药汤	经典	《千金要方	产后虚损，	调和气血，	补益方	气血双补	当归 白芍 人参 肉桂 生姜 甘草 大枣 地黄
10000319	肠宁汤	肠宁汤	经典	《傅青主女	妇人产后亡	补气补血	补益方	气血双补	当归 熟地黄 人参 麦冬 阿胶 山药 续断 甘草 肉桂
10000336	固阴煎	固阴煎	经典	《景岳全书	肝肾亏虚之	补益肝肾，	补益方	补阴	人参 熟地黄 麸炒山药 山萸肉 制远志 炙甘草 五味子
10000366	右归饮	右归饮	经典	《景岳全书	肾阳不足，	温补肾阳	补益方	补阳	熟地黄 山药 枸杞子 山萸肉 炙甘草 茯苓
10000380	养荣壮肾汤	养荣壮肾汤	经典	《傅青主女	产后感受风	nan	补益方	补阳	当归 防风 独活 肉桂 杜仲 续断 桑寄生
10000419	茯神散	茯神散	经典	《医宗金鉴	产后血虚，	健脾养血、	安神方	滋养安神	茯神 人参 黄芪 赤芍 牛膝 琥珀 龙齿 地黄 肉桂 当
10000423	至宝丹	至宝丹	经典	《灵苑方》	卒中急风不	化痰开窍，	开窍方	凉开	水牛角 玳瑁 琥珀 朱砂粉 雄黄粉 牛黄 冰片 麝香 安
10000464	趁痛丸	趁痛丸	经典	《经效产宝	治产后遍身	益气补血，	理气方	行气	牛膝 当归 肉桂 白术 黄芪 薤白 独活 生姜 炙甘草
10000473	桃核承气汤	桃核承气汤	经典	《伤寒论》	瘀热蓄于下	逐瘀泻热	理血方	活血祛瘀	火单 桃仁 桂枝 大黄 芒硝 炙甘草
10000478	生化汤	生化汤	经典	《傅青主女	产后血瘀留	养血祛瘀，	理血方	活血祛瘀	当归 川芎 火单 桃仁 炮姜 炙甘草
10000479	失笑散	失笑散	经典	《太平惠民	小肠气及心	活血化瘀，	理血方	活血祛瘀	酒五灵脂 生蒲黄
10000480	桂枝茯苓丸	桂枝茯苓丸	经典	《金匮要略	妇人宿有症	活血化瘀，	理血方	活血祛瘀	桂枝 茯苓 牡丹皮 火单 桃仁 白芍

图 2 中医妇科相关数据示例

1.3 本体层设计

本体是知识图谱中的概念模型，类似于面向对象程序设计的类，用于描述领域中实体、关系和属性等因素，是知识图谱的核心架构和基础，将本体整理成相应的结构表，有助于对知识图谱的理解[2]。本体中类与类的层次结构构建主要有 3 种方式，分别为自顶向下法、自底向上法、自顶向下法和自底向上法相结合[3]。本研究采用自顶向下的方法构建中医妇科本体层，在相关领域专家的指导下，结合中医妇科学领域的特点，使用 Protégé 软件进行本体编辑，采用 Protégé 的“七步法”步骤进行构建，主要包括确定领域范畴、考虑复用本体可能性、列出本体中的重要术语、定义类与类之间的等级、定义相关的对象属性、数值属性、添加实例几个步骤进行[4]。首先构建本体的基本模型，在中医妇科本体的顶层类以下，分为“药物”“方剂”“治法”“症状”“疾病”等 5 个大类，然后根据中医妇科临床经验知识细化下一层结构，其中在疾病大类下，分为“带下病”“妊娠病”“月经病”“产后病”几个类，这几类下细分为产后发热、产后腹痛等，如表 1 所示，由于分类较多，仅展示部分。本研究使用 Protégé 软件构建的中医妇科本体层次结构如图 3 所示。

表 1 部分层级结构示例

一级类目	二级类目	三级类目	实例
疾病	产后病	产后发热	外感风寒型产后发热
		产后大便难	脾胃气虚型产后大便难
		产后腹痛	寒凝血瘀型产后腹痛
		产后血晕	瘀阻气闭型产后血晕

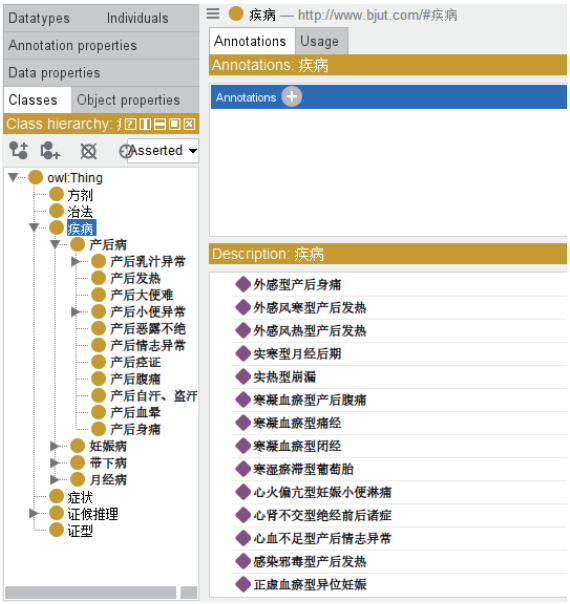

图 3 中医妇科本体层构建

在本体中类主要分为两种类型的属性，即数据属性和对象属性，对象属性即是类与类之间的关系[4]，如“方剂”治疗“疾病”、“症状”出现“疾病”、“治法”治疗“疾病”、“疾病”有…“症状”等，本研究构建的对象属性如表 2 所示。

表 2 类的对象属性

对象属性	反向属性	定义域	值域
用药为	被用药为	疾病	方剂
有…症状（has Symptom）	出现（appear_in）	疾病	症状
表现出（show）	是…的表现（reflect）	证型	症状
治疗（treat）	被治疗（be_treated）	治法	疾病
由…组成	组成	方剂	药物

1.4 知识抽取

中医妇科知识抽取主要分为实体抽取和关系抽取两部分，目的是把清洗后的中医妇科文本结合知识抽取技术，构造＜实体，关系，实体＞结构化的三元组信息。中医药领域实体抽取的主流方法是将传统的机器学习模型与深度学习方法相结合，如用于序列标记的长短期记忆 LSTM-CRF 模型 [5]。Bi-LSTM-CRF 模型是医学领域实体识别比较主流的深度学习模型，能有效提高模型的准确率 [6]。本研究使用 Bi-LSTM-CRF 模型和人为定义的规则，对中医妇科文本进行实体和关系抽取，抽取的实体主要包括疾病、症状、方剂、治法等。实体抽取的具体流程如下，首先对中医妇科文本信息通过 jieba 工具进行分词处理，其次采用 BIO 标注方法，对分词后的数据进行序列标注，标注实体起始位置，中间和非实体部分，然后使用 Word2vec 工具把实体进行词向量化生成字向量矩阵，并作为 Bi-LSTM 层的输入，最后通过 CRF 层计算概率，确定实体所属的类型。总共抽取到 7201 个实体，其中疾病实体 169 个，药物数 3402 个，方剂实体 2689 个，症状实体 740 个，治法实体 201 个。

实体关系抽取是构建中医妇科知识图谱的重要步骤，通过查阅相关资料，本研究基于人工定义的规则对中医妇科知识进行实体关系抽取，定义了 9 类实体关系，分别是“用药为”“治疗”“表现出”“有…症状”“出现”等。总共抽取出 28 427 个关系，其中“有…症状”关系有 2096 个，“表现出”关系有 904 个，“用药为”关系有 594 个，“治疗”关系有 611 个，“出现”关系有 2096 个，“由…组成”关系有 23 466 个。综上，基于 Bi-LSTM-CRF 模型和人工定义规则抽取的中医妇科数据的实体关系。

1.5 知识融合

本文通过 Bi-LSTM-CRF 模型和人工定义规则抽取中医妇科知识的实体和关系，但由于中医经典理论丰富、历史发展悠久、术语描述差异大等特点，抽取的数据中可能存在数据冗余和一些错误信息，如：出现一词多义。在不同的年代，对相同疾病的症状描述不同等情况，还需要对中医妇科数据进行知识融合和实体对齐等操作，提高知识图谱整体的准确性和规范性。本研究采用人工方式进行知识融合，参考《中医临床诊疗术语》[7]《中医临床常见症状术语规范》[8] 文献和标准术语概念，对实体数据采用字符串匹配的方法，对抽取的中医妇科实体数据进行归类。例如：“和气血”和“养气血”等含义相同的词语，把这类词分类记录并构建为同义词典，并按照术语规范，以“和气血”这个词语作为其同义词的标准描述方法，构建的同义词字典部分示例如表 3 所示。

表 3 同义词典示例

同义词字典	示例	标准描述
治法词典	和血气、和血、养血气、祛寒和血、和气血	和气血
症状词典	半身不遂 半身不随 半身不收 半身不举 半身手足不遂	半身不遂

1.6 知识存储与展示

Neo4j 是一个高性能的 NoSQL 图形数据库，与传统的关系型数据库相比，Neo4j 的数据不是存储在二维数据表中，而是以图的形式进行存储。本研究使用 Neo4j 图数据库软件进行数据存储和可视化的展示。知识图谱的可视化是指将知识单元之间的关系转化为能更好理解的图的形式，用以表现抽象的事物。Neo4j 图数据库可以批量进行数据导入和存储功能，使用 Cypher 语言进行数据的增删改查和图谱展示操作。本研究的数据存储过程如下：在数据存储之前，需要在 Protégé 软件中把编辑好的本体数据导出为 Rdf 格式文件，然后在 Neo4j 图数据库软件中，使用 Cypher 语言将 Rdf 文件导入，最后编写相应的查询语句，实现图谱中对应节点的查询。图 4 是中医妇科知识图谱的可视化展示，在图谱中能看到有很多的实体节点，节点之间相互关联较多，不利于查看和分析，为了解决这个问题，把图谱按照本体层进行分类，并使用不同的颜色标注用以区分。图谱中包含很多种疾病实体节点，每种疾病与相关症状关联，每个症状与对应的治法连接，治法与相关的方剂连接，方剂与组成的药物连接，形成了从疾病与症状、症状与治法、治法与方剂、方剂与药物的逻辑关系，从而构建出了基于“疾病 - 症状 - 治法 - 方剂 - 药物”的中医妇科知识图谱。与常规的检索方式相比，知识图谱能快速找到某个实体相关联的知识，发现知识之间的内在联系，构造出更加丰富、庞大的中医妇科知识库 [9]，该图谱的构建对于中医妇科疾病的诊疗和用药有一定的辅助作用。

2 结语

本研究通过构建中医妇科知识图谱，不仅可以更好地组织和展示复杂的中医妇科知识，而且有助于揭示不同实体之间的内在联系，为中医妇科的研究和临床诊疗提供了新的视角和工具。虽然当前的研究已经取得了一定的成果，但在中医知识的深入挖掘、数据的精准性和系统的完善性方面仍有提升空间。未来，随着技术的进步和数据的积累，中医妇科知识图谱的构建和应用将更加完善，对于推动中医学的现代化和国际化具有重要意义。

参考文献：

[1] 羊艳玲，李燕，帅亚琦，等．基于中医医案的知识图谱构建

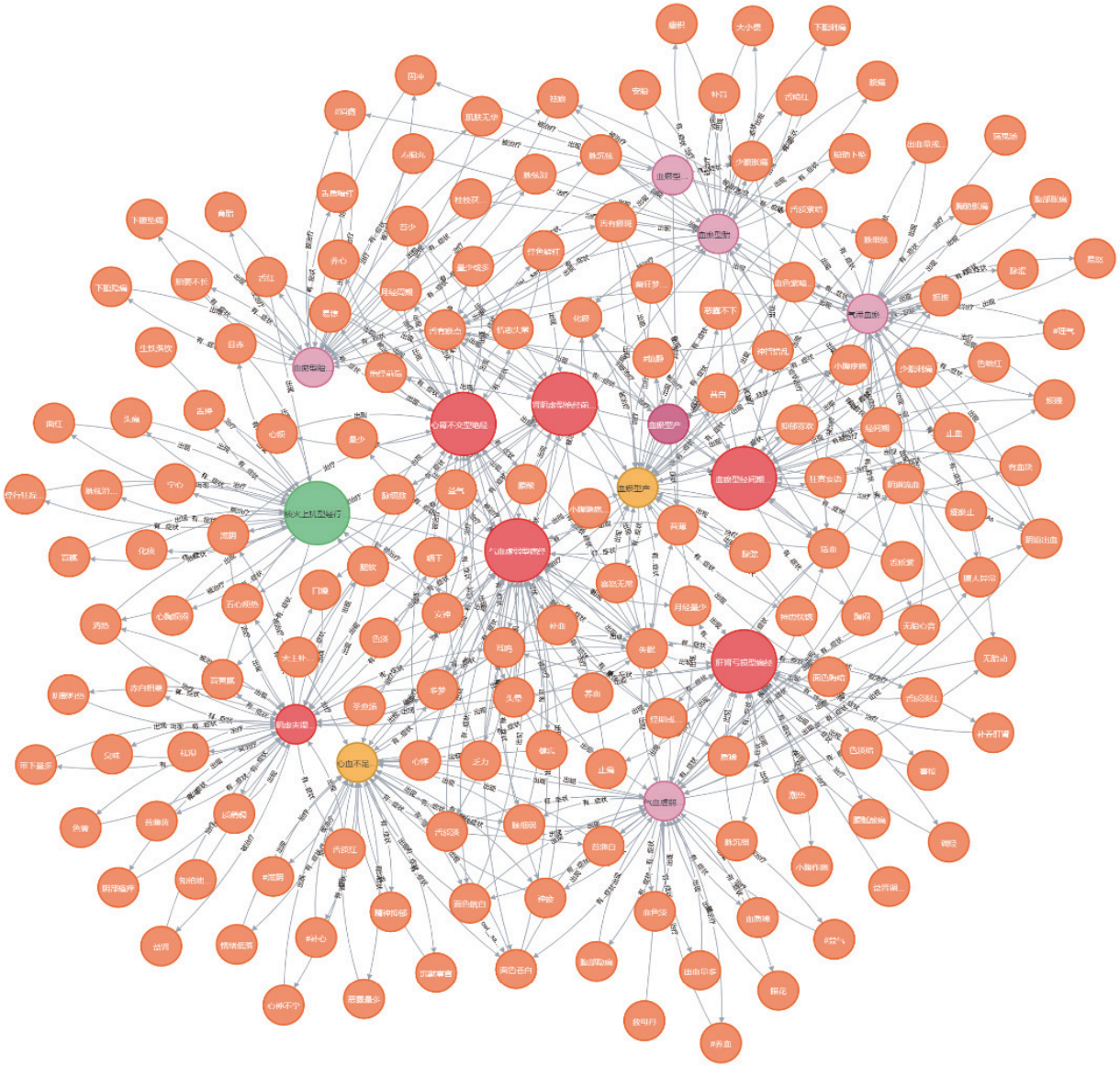

图 4 中医妇科知识图谱的可视化展示

[J]. 医学信息学杂志 , 2022, 43(10): 50-54.

[2] 刘金垒 , 惠小珊 , 张振鹏 , 等 . 基于中医诊疗指南的冠心病知识图谱构建 [J]. 中国实验方剂学杂志 , 2023, 29(07): 208-215.

[3] 贺雅洁 . 中医妇科肿瘤本体的构建研究 [D]. 长春 : 长春中医药大学 , 2023

[4] 刘丽 . 中医经方知识图谱的研究与构建 [D]. 济南 : 山东中医药大学 , 2022

[5] 王松 , 李正钧 , 杨涛 , 等 . 中医药知识图谱研究现状及发展趋势 [J]. 南京中医药大学学报 , 2022, 38(3): 272-278.

[6] 张坤丽 , 胡晨馨 , 宋玉 , 等 . 基于多源数据的中文产科知识图谱构建 [J]. 郑州大学学报 (理学版), 2023, 55(1): 8-14.

[7] 国家技术监督局 . 中医临床诊疗术语 [M]. 北京 : 中国标准出版社 , 1997.

[8] 中华中医药学会 . 中医临床常见症状术语规范 [M]. 北京 : 中国中医药出版社 , 2017.

[9] 徐安迎 , 胡孔法 , 杨涛 . 基于 Neo4j 的肺癌中医诊疗知识图谱构建研究 [J]. 世界科学技术 - 中医药现代化 , 2023, 25(4): 1456-1461.

【作者简介】

周涛（1996—），男，湖北黄冈人，硕士研究生，助理实验师，研究方向：中医药数据处理。

常凯（1987—），通信作者（changkai@hbtcm.edu.cn），男，湖北武汉人，硕士研究生，高级实验师，研究方向：中医药人工智能、中医药信息学。

（收稿日期：2024-03-21）