

基于《中国药典》的中药知识图谱构建与应用^{*}

付涛涛¹ 陈艳梅¹ 李庆娜² 邵义明³ 苏国彬⁴ 弓孟春⁵

(¹ 神州医疗科技股份有限公司 北京 100080 ² 中国中医科学院西苑医院 北京 100091

³ 广东医科大学附属东莞第一医院 东莞 523808 ⁴ 广东省中医院 广州 510000

⁵ 广东医科大学多模态数据融合创新应用实验室 东莞 523808)

〔摘要〕 目的/意义 将电子信息领域的知识图谱技术引入中药术语研究，以生动的方式展示中药治疗作用与性味归经和中医疾病之间的辨证关系。方法/过程 采用自上而下的本体构建方式在 Protégé 平台搭建中药本体顶层结构。以 2020 版《中国药典》一部为数据来源，抽取其中的中药信息，并整理各轴位信息、消歧和归一。借助 Protégé 平台创建实体和关系，以语义网络资源描述框架 RDF 数据格式输出中药三元组数据，并以 Neo4j 图数据库存储、展示，形成系统化的中药知识图谱。结果/结论 完成基于《中国药典》的中药知识图谱构建，可通过数据统计分析、知识计量以及可视化图形绘制，充分揭示中药领域复杂的知识系统结构。

〔关键词〕 中药；《中国药典》；知识图谱；Protégé；Neo4j

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.10.007

Construction and Application of Traditional Chinese Medicine Knowledge Graph Based on the Chinese Pharmacopoeia

FU Taotao¹, CHEN Yanmei¹, LI Qingna², SHAO Yiming³, SU Guobin⁴, GONG Mengchun⁵

¹Digital Health China Technologies Co. Ltd., Beijing 100080, China; ²Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing 100091, China; ³The First Dongguan Affiliated Hospital of Guangdong Medical University, Dongguan 523808, China; ⁴Guangdong Provincial Hospital of Chinese Medicine, Guangzhou 510000, China; ⁵Innovative Application Laboratory of Multimodal Data Fusion, Guangdong Medical University, Dongguan 523808, China

〔Abstract〕 **Purpose/Significance** To introduce the knowledge graph technology in the field of electronic information into the study of traditional Chinese medicine (TCM) terminology, and to demonstrate the dialectical relationship between the therapeutic effects of TCM and its nature, flavor, meridians and TCM diseases in a vivid way. **Method/Process** The top-down ontology construction method is adopted to build the top-level structure of TCM ontology on the Protégé platform. Taking the 2020 edition of the Chinese Pharmacopoeia as the data source, the TCM information in the pharmacopoeia is split and extracted, and the information of each axis is sorted, disambiguated and normalized. With the help of the Protégé platform, entities and relationships are created, and the TCM triple data is output in RDF data format. Finally, the Neo4j graph database is used to store and display the RDF data to form a systematic TCM knowledge graph. **Result/Conclusion** It mainly realizes the construction of the knowledge graph of TCM in the pharmacopoeia, and fully reveals the complex knowledge

〔修回日期〕 2024-07-08

〔作者简介〕 付涛涛，硕士，发表论文 2 篇；通信作者：弓孟春，博士。

〔基金项目〕 国家重点研发计划项目（项目编号：2023YFC2706305）；国家中医药管理局监测统计中心自选课题（项目编号：YGZXKT2024013）。

system structure in the field of TCM through data statistical analysis, knowledge measurement and drawing of visual graphics.

[Keywords] traditional Chinese medicine (TCM) ; Chinese Pharmacopoeia ; knowledge graph ; Protégé ; Neo4j

1 引言

本体理论于 20 世纪 90 年代被引入计算机人工智能领域，是一种用于定义概念及其相互关系的结构化信息模型，是对真实世界中事物属性特征及其关系的抽象，是知识图谱的构建基础^[1-3]。知识图谱通过结构化的模式描述客观世界中概念与实体及其之间的关联信息，将互联网的海量信息转化成更便于人类认识和理解的形式，提供了一种更好地管理互联网海量碎片数据的方式。

中医经历了几千年的发展，积累了大量经验，但是知识体系并不完善。中医药领域信息孤岛现象明显，概念较为抽象，知识体系庞大，阻碍了中医药知识的互操作性和深度分析应用。中医药语言存在同物异名、信息碎片化、缺少系统结构等问题。知识图谱则可以使中医药知识统一化，促进中医药行业的发展。

近年来国家层面对于医药学，特别是中医药领域高度重视，提高了对中医医院标准化建设、中医药信息化人才培养和创新发展的要求^[4]，也推动促进了中医药领域数字化、结构化、智能化的研究。医学领域有许多成熟的知识库，例如基因本体^[5-6]、医学系统命名法 - 临床术语（systematized nomenclature of medicine - clinical terms, SNOMED CT）知识库^[7]、encyclopedia^[8]等。本研究参考 SNOMED - CT 的顶层框架逻辑，借助 Protégé 和 Neo4j 工具创建并存储中药知识三元组数据，构建的中药知识图谱系统可以直观地展示中药饮片、中成药和适应证之间的关系，便于深层次地了解药典中隐藏的知识，使更多用户可以根据自身症状获取药典推荐的相关用药知识信息。

2 构建方法

采用自上而下的本体构建方式，即先为知识图

谱定义好本体与数据模式，再将实体加入到知识库，主要包括知识获取、知识抽取、知识融合、知识表示以及知识存储^[9-10]等环节，见图 1。

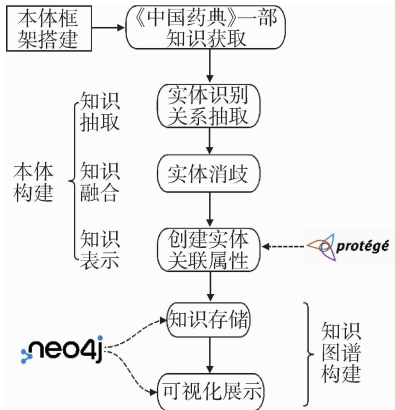


图 1 中药知识图谱构建流程

2.1 本体框架搭建

中医药领域顶层本体是在语义层面总结中医药知识的基本概念框架，是中医药领域本体与通用顶层本体连接的基础^[11-12]。本研究在 SNOMED CT 19 个顶层概念的基础上^[13-14]，结合《中药学语言系统语义网络框架》的中药性能定义进行概念分类，搭建中药本体层级结构模式，见图 2。

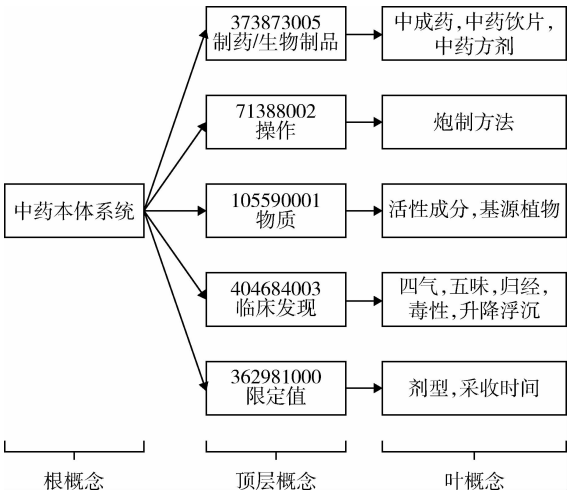


图 2 中药本体层级结构模式

2.2 知识获取

以《中国药典》一部作为中药知识的唯一数据来源，基于文本规律性的分布特性，通过 Excel 文本分列技术对药典文本进行属性类别拆分，获取中药关联的属性信息和实体信息。主要以药典中的下列轴位信息为研究要点：品名、来源、处方、制法、性状、含量测定、炮制、性味与归经、功能与主治、用法与用量、注意、规格、贮藏等^[15]。

2.3 知识抽取

2.3.1 实体识别 主要任务是从文本中识别出实体，并标明其类别和属性。由于药典中的句式具有明显的标志符号，采用基于规则的方法识别和抽取

实体。如语句开头具有“【】”“。”“、”等符号，句中包含“用于”“归... 经”等关键词，均可作为属性名称的位置切割点。通过 Excel 文本分列功能拆分中药信息包含的实体。例如：中药中包含“功能主治”的部分原文为：“宣肺，利咽，祛痰，排脓。用于咳嗽痰多，胸闷不畅，咽痛音哑，肺痈吐脓”，句号前面的内容属于中医治法，按照逗号分割可得到“宣肺/利咽/祛痰/排脓”4个中医治法关系对应的实体信息；句号后面的内容属于中医疾病，按照逗号和关键词“用于”分割可得到“咳嗽痰多/胸闷不畅/咽痛音哑/肺痈吐脓”4组中医疾病关系对应的实体信息。此外，通过参考《中医药学语言系统语义网络框架》，梳理并抽取《中国药典》中中药本体知识的重要术语，见表1^[16]。

表1 中药本体的典型术语及实体

术语	定义	实体举例
毒性	中药对机体所产生的不良反应	有大毒
四气	中药4种基本属性：寒、热、温、凉	温，热
五味	中药的5种味道：酸、苦、辛、甘、咸	苦，酸
归经	中药作用归属、趋向于某脏腑、经络或特定部位等的定位、定向理论	肝经，肺经
中药	在中药理论和临床经验指导下用于防治和医疗保健的药物，包括中药材、饮片、中成药和方剂等	小青龙汤，三黄片，党参
中药饮片	以中药材为原材料，按中药理论，经过加工炮制后，可直接用于中医临床或制剂生产的处方药品	三七、白术
中成药	在中药理论指导下，为预防及治疗疾病的需要，按规定的处方和制剂工艺加工制成，为中药复方或单方使用的成品药剂	乌鸡白凤丸
证候	证的外候，疾病过程中一定阶段的病位、病因、病性、病势及机体抗病能力的强弱等本质有机联系的反应状态，表现为临床可被观察到的症状等	肺虚证、气阴两虚证

2.3.2 属性关系抽取 主要任务是从给定的结构化和非结构化文本中抽取两个实体或多个实体之间的关系。实体的属性关系分为两类：对象属性和数据属性，对象属性是一个对象作为实体的属性^[17]，表示属性值对应的概念或实体信息，用于描述两个实体信息之间的关系。以中药饮片“海金沙”为例，其对象属性关系有适应证、四气、五味、归

经、治法、采收季节、证候、活性成分、用药剂量等。数据属性指通过一个字符串为值描述实体的属性，表示属性值对应的具体参数，用于描述实体信息与参数之间的关系。中药饮片“海金沙”的数据属性关系有汉语拼音、拉丁名、药典页码。将清洗归一后的属性信息按照不同的特性划分到对象属性或数据属性的值域，见表2。

表2 中药本体系统属性关系的定义域、值域和属性类别

编号	属性关系 - 英文	属性关系 - 中文	定义域	值域	属性类别
1	harvest_time	采收时间	中药饮片	采收时间	对象属性
2	medicinal_part	药用部位	中药饮片	药用部位	对象属性
3	original_plant	基源植物	中药饮片	药用植物	对象属性

续表 2

编号	属性关系 - 英文	属性关系 - 中文	定义域	值域	属性类别
4	processed_method	炮制方法	中药饮片	炮制方法	对象属性
5	toxicity	毒性	药物	毒性	对象属性
6	channel_distribution	归经	中药饮片	归经	对象属性
7	chemical_formula	化学式	化学物质	化学式	对象属性
8	contraindication	禁忌	药物	用药禁忌	对象属性
9	drug_frequency	用药频次	药物	用药频次	对象属性
10	original_family	基源科目	中药饮片	基源科目	对象属性
11	five_tastes	五味	中药饮片	五味	对象属性
12	four_properties	四气	中药饮片	四气	对象属性
13	has_active_ingredient	活性成分	药物	化学物质	对象属性
14	has_dosage	用药剂量	药物	用药剂量	对象属性
15	has_dosage_form	剂型	药物	剂型	对象属性
16	therapeutic_effect	治疗作用	药物	中医治法	对象属性
17	indication_of	适应证	药物	中医疾病	对象属性
18	prescription	处方	药物	中药饮片	对象属性
19	processed_product	炮制品	中药饮片	中药饮片	对象属性
20	original_type	基源类别	药物	基源类别	对象属性
21	storage_method	贮藏方式	药物	贮藏方式	对象属性
22	TCM_syndrome	中医证候	药物	中医证候	对象属性
23	ChP_page	药典页码	药物	字符串	数据属性
24	ID	编码	药物	字符串	数据属性
25	latin_name	拉丁文名	中药饮片	字符串	数据属性
26	pinyin	拼音	药物	字符串	数据属性
27	synonyms	同义词	药物	字符串	数据属性

2.4 知识融合

主要解决实体命名歧义的问题，如同一名称对应不同实体或同一实体有不同名称。在中医领域，同一种疾病或中药材的名称会存在多种描述方式，需要在创建实体前消歧。鉴于药典中每种中药的名称、性味归经、基源以及功能主治等内容均采用结构化格式撰写，整体较为规范统一，本研究主要通过中医药学专业技术人员人工审核校正的方式对抽取的中药实体文本清洗、消歧和归一处理。

2.5 知识表示

通过 Protégé 软件，借助 Mapping Master DSL 基础语法批量创建实体并设置对象属性和数据属性与实体之间的关联关系，见图 3，可以清晰呈现中药饮片“海金沙”实体及其关联的属性信息。

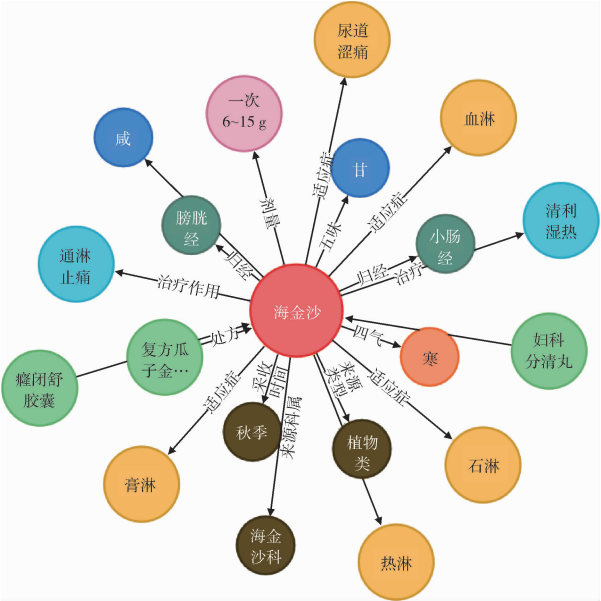


图 3 中药饮片“海金沙”属性关系

2.6 知识存储及可视化

知识存储主要探究将知识图谱以哪种方式存储。目前主流知识图谱大多以图数据库的结构存储，因此选取 Neo4j 图数据库作为中药知识的存储方式，即采用 RDF 数据格式表示中药的三元组关系数据，将 Protégé 中的属性关系和实体数据存储到 Neo4j，具体设置参数，见表 3。之后通过 Cypher 命令可以进行各种操作，输入 SQL 查询语句可以得到实体之间的关联关系，并以可视化的图谱形式展现查询结果。

表 3 Neo4j 图数据库参数设置

参数	值
Handle Vocab Uris	SHORTEN
Handle Multival	OVERWRITE
Handle RDF Types	LABELS
Keep Lang Tag	FALSE
Keep Custom Data Types	FALSE
Apply Neo4j Naming	FALSE
Base Schema Name space	neo4j: //graph. schema#
Base Schema Prefix	n4sch
Class Label	Class
subClass Of Rel	SCO
Data Type Property Label	Property
Object Property Label	Relationship
subProperty Of Rel	SPO
Domain Rel	DOMAIN
Range Rel	RANGE

3 结果及应用场景

本研究构建的中药知识图谱涵盖了中药饮片、中成药、中医治法、四气、五味、归经、剂型、禁忌、毒性、处方组成、中医证候、中医疾病、服用方式、用药剂量、活性成分 15 类实体。围绕中药饮片和中成药两大类本体层级体系，构建了 119 个类、9 136 个实体、25 种对象属性（包含 39 921 条对象属性关系）、5 种数据属性（包含 5 030 条数据属性关系），建立了中药特色的知识本体系统，全方位展示了中药

领域的药典关联知识，初步实现药典中药知识的重新组织和可视化表达。通过构建大型化、可扩展性强的中医药领域知识系统，实现知识的关联与融合，可以有效助力多种中医药场景的实际应用。

3.1 中药知识图谱应用场景一：为中药材种植产业提供指导

中医药知识图谱的构建可以实现知识汇聚，集成和关联各类碎片化的中医药知识。通过知识图谱系统分析《中国药典》中中成药的处方组成，得出甘草、当归、茯苓、黄芩、川芎、黄芪、丹参、地黄、白芍、陈皮、桔梗等几类中药材在中成药处方中使用频率较高，见图 4。相关数据信息可与市场中成药及中药材相关销量情况相结合，对于支持中药产业的药用植物选取及种植规模控制具有重要参考价值。



图 4 不同中药饮片在中成药处方中的使用频率

3.2 中药知识图谱应用场景二：阐明中药作用机制

利用中药知识图谱系统检索治疗风寒/风热感冒的中药性能关系图，通过桑基图（图中色块代表不同中药）分析发现同是治疗感冒，但使用的中药类别性质完全不同。治疗风寒感冒的中药性能主要是温、辛，归肺经、膀胱经，见图 5；而用于治疗风热感冒的中药性能主要是寒、苦，归肺经、肝经，见图 6。可见“感冒”病变的主要发生部位“肺”与治疗药物的属性定位“肺经”相统一，而治疗采用的药物四气属性与疾病的八纲征象是相反的。这是中医正治法“寒者热之，热者寒之”的治则体现，说明中药知识图谱的推理逻辑与中医辨证

论治的思维特点是一致的，可以辅助解释中药的作用机制及作用特点。

“玉泉胶囊”“糖脉康颗粒”等，见图 7。

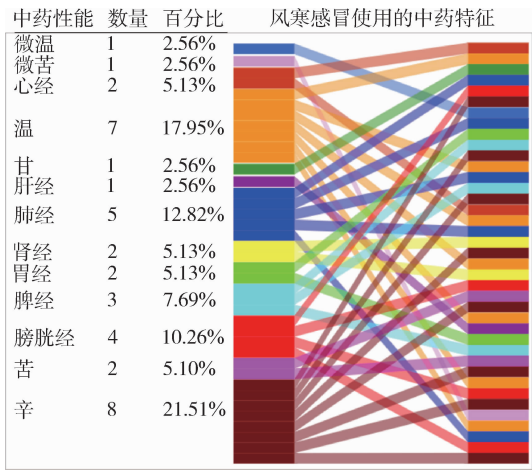


图 5 治疗风寒感冒的中药特征

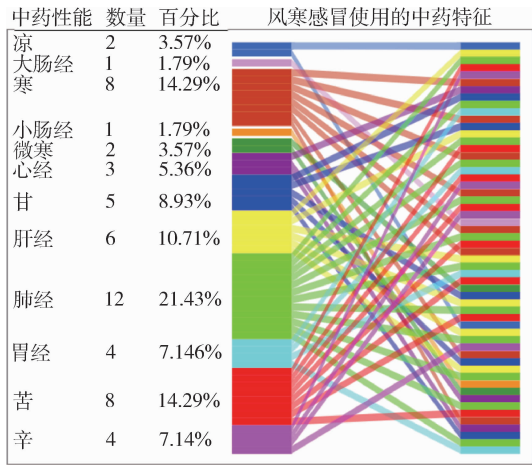


图 6 治疗风热感冒的中药特征

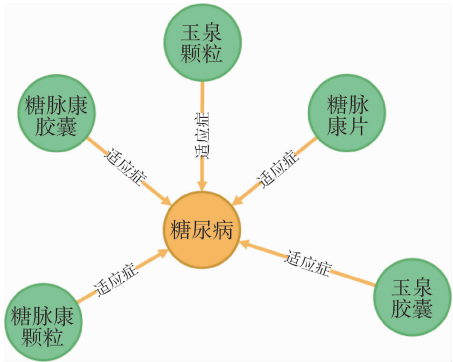


图 7 治疗“糖尿病”的智能问答推荐结果

知识图谱系统根据医生输入的临床症状，输出该病症适合使用的中成药推荐选项，每个中成药同时与其对应的剂型、用法用量、禁忌证、处方等通过属性关系直接关联，根据每个中成药的属性分布特点，结合患者的实际体征，可综合评估从而确定最适合的药品。知识图谱系统实现中药推荐结果的基本逻辑是根据疾病、证候、治疗作用与中药之间的关系，搜寻临床表现与中药之间的关联，这些关联代表临床表现与中药之间的潜在路径关系。因此可以通过这些路径找到在某个场景下可以适用或不适用的中药使用结果。中药知识智能问答功能可以助力患者了解中医药治疗思路及用药指南，进而提升就医效率并增强了患者的自我管理能力和同时，其还能模拟临床场景，为医学生提供诊疗案例，促进学生的临床思维培养。

3.3 中药知识图谱应用场景三：实现智能问答

中药知识问答系统的处理流程为，针对用户提出的中药知识相关问题，采用 Neo4j 图数据库的查询语言 Cypher 构造查询条件，得到符合条件的答案，最后以可视化图文形式输出^[18]。基于中药知识图谱可以实现基本的知识问答，能回答用户用自然语言提出的中医药临床问题，直接给出答案或推荐结果。例如：想了解治疗糖尿病的中成药有哪些。输入匹配查询命令：“MATCH (n: 中成药) - [r] -> (m: 中医疾病 {uri: ‘糖尿病’}) RETURN n, m, r;”，查询结果为适应证对应糖尿病的中成药结果，如

4 讨论

本研究将药典中药知识结合 SNOMED CT 顶层设计，建立中医药本体层级结构。将中成药和中药饮片单独分类构建，使中药概念更为具体详尽，也更有层次性；构建了中药与中医疾病、中医证候之间的属性关系，可以在一定程度上解释中药对中医疾病的作用机制，使对中药知识的理解和临床应用更加深入。利用知识图谱系统进行进一步知识推理，设计相应的中药使用场景，可以解决中药对于不同病因病机的用药方案选择问题^[19]。

当前知识图谱相关智能问答应用尽管较为广

泛,但也存在一些问题,实际运用相对比较初级,大规模中医药多模态知识图谱的构建和应用仍具有较大挑战^[20],例如中医药领域,目前始终缺乏高质量、高精度的知识图谱系统。后续随着 AI 技术、智能问答以及计算机技术的发展,将拓展中医药标准的参考来源范围,深度挖掘和完善已建立的中医药知识库,持续优化和修正中药知识推荐结果。

5 结语

本研究初步实现了药典中药知识的重新组织和可视化表达,为进一步结合大规模中药数据集,建立数据更全面、内容更丰富的中药知识图谱奠定了基础。中药知识图谱在实际构建过程中也存在一些问题和不足之处,一方面对于中医疾病的属性关系未做进一步的定义和细化,另一方面是中药本身的性味归经与疾病的发生机制、性质未详尽分析。未来将进一步细化中药性能属性,关联补充中医药的其他相关概念和关系,并对中医证候、中医疾病和中医治则治法进行归纳和分析,实现中医药本体系统整合联动,扩大知识图谱的关系维度和知识规模,使中医药知识图谱的应用更加全面具体。

作者贡献: 付涛涛负责论文撰写;陈艳梅负责文献调研;李庆娜负责数据分析;邵义明负责研究设计;苏国彬负责项目协调、论文修订;弓孟春负责论文指导与审核。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

1 朱彦,徐静雯. 中医药本体研究思考与展望——从术语集到本体集 [J]. 医学信息学杂志, 2022, 43 (8): 8-14.

2 HAO X, JI Z, LI X, et al. Construction and application of a knowledge graph [J]. Remote sensing, 2021, 13 (13): 2511.

3 ATAEVA O M, SEREBRYAKOV V A, TUCHKOVA N P. Ontological approach to a knowledge graph construction in a semantic library [J]. Lobachevskii journal of mathematics, 2023, 44 (6): 2229-2239.

4 赵静,肖勇. 基于知识图谱的中医药信息化研究可视化分析 [J]. 中国卫生信息管理杂志, 2023, 20 (5): 838.

5 Gene Ontology Consortium. Gene ontology consortium: going

forward [J]. Nucleic acids research, 2015, 43 (D1): D1049-D1056.

6 LIU Y, ZHANG Y Z, IMOTO S. Microbial gene ontology informed deep neural network for microbe functionality discovery in human diseases [J]. Plos one, 2023, 18 (8): e0290307.

7 EL-SAPPAGH S, FRANDA F, ALI F, et al. SNOMED CT standard ontology based on the ontology for general medical science [J]. BMC medical informatics and decision making, 2018, 18 (1): 76.

8 LU S, WANG H, ZHANG J. Identification of uveitis-associated functions based on the feature selection analysis of gene ontology and Kyoto encyclopedia of genes and genomes pathway enrichment scores [J]. Frontiers in molecular neuroscience, 2022, 15 (9): 1007352.

9 朱木易洁,鲍秉坤,徐常胜. 知识图谱发展与构建的研究进展 [J]. 南京信息工程大学学报(自然科学版), 2017, 9 (6): 8.

10 ZHAO Z, HAN S K, SO I M. Architecture of knowledge graph construction techniques [J]. International journal of pure and applied mathematics, 2018, 118 (19): 1869-1883.

11 龙海,朱彦. 论 GFO 的基本框架及顶层本体比较研究 [J]. 中国中医药图书情报杂志, 2015, 39 (5): 18.

12 徐静雯,朱彦. 中医药领域本体研究现状 [J]. 医学信息学杂志, 2023, 44 (1): 36.

13 杨美洁,熊相超. 基于爬虫技术和电子病历的糖尿病知识图谱的构建 [J]. 中国数字医学, 2020, 15 (2): 6-8.

14 孙华君,李海燕,贾李蓉,等. 基于国际标准 ISO/TS 16843—6: 2022 的针刺效应本体构建研究 [J]. 中国针灸, 2023, 43 (1): 73.

15 国家药典委员会. 中华人民共和国药典(一部) [M]. 北京: 中国医药科技出版社, 2020.

16 贾李蓉,于彤,崔蒙,等. 中医药学语言系统研究进展 [J]. 中国数字医学, 2014, 9 (10): 57.

17 LIU Q, CHENG G, GUNARATNA K, et al. Entity summarization: state of the art and future challenges [J]. Journal of web semantics, 2021, 69 (1): 100647.

18 李伟,王竣生,秦鹏. 基于知识图谱的医疗问答系统研究 [J]. 长江信息通信, 2023, 36 (6): 107-109

19 杨郑子衿,徐倩,王安莉,等. Protégé 在构建中医药本体中的应用 [J]. 医学信息学杂志, 2021, 42 (6): 37.

20 李荣耀,徐倩,吴雨璐,等. 基于《本草纲目》的多模态知识图谱的构建研究 [J]. 现代计算机, 2022, 28 (13): 10.