中成药数据图谱可视化与知识问答平台研究

周雪阳,廖诗雨,董泽华,程春雷,叶 青

(江西中医药大学 计算机学院,江西 南昌 330004)

摘 要: 中成药数据具有数量庞大、关系复杂等特点,如何对种类繁多的中成药临床、流通与标准规范数据进行有效存储、管理、跟踪与使用成为药品监管部门关注的重点。为实现中成药知识整合、提高数据关联性并挖掘数据的潜在价值,采用知识图谱存储结构结合可视化技术组织中成药临床技术、商业流通、标准规范等信息,构建中成药知识图谱数据库体系,搭建中成药知识图谱可视化平台。基于知识图谱技术的可视化平台能更好地挖掘中成药数据的潜在价值,而中成药知识图谱数据库可为智能问答研究提供数据基础,具有较好的知识服务前景。

关键词: 中成药;知识图谱;可视化;平台构建

DOI: 10. 11907/rjdk. 211690 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP391 **文献标识码:** A **文章编号:** 1672-7800(2022)005-0158-05

Research on Data Visualization and Knowledge Q&A Platform of Chinese Patent Medicine

ZHOU Xue-yang, LIAO Shi-yu, DONG Ze-hua, CHENG Chun-lei, YE Qing

(*College of Computer Science, Jiangxi University of Traditional Chinese Medicine, Nanchang* 330004, *China*)

Abstract: The data of Chinese patent medicine has the characteristics of large quantity and complex relationship. How to effectively store, manage, track and use the clinical, circulation and standard data of Chinese patent medicine has become the focus of the drug regulatory authorities. In order to help realize the knowledge integration of Chinese patent medicine, improve the data association, and mine the potential value of data. Methods the project used the knowledge map storage structure combined with visualization technology to organize the clinical technology, commercial circulation, standards and other information of Chinese patent medicine. Results the database system of Chinese patent medicine knowledge map was constructed, and the visualization platform of Chinese patent medicine knowledge map was built. The visualization platform based on knowledge mapping technology can better mine the potential value of the data. At the same time, the database of Chinese patent medicine knowledge mapping can provide data basis for intelligent question answering research, which has a great prospect of knowledge service.

Key Words: Chinese patent medicine; knowledge map; visualization; platform construction

0 引言

中医是中华民族优秀文化的重要组成部分,现代中医干预疾病的主要方式为中药[1]。中药在新冠肺炎的预防和治疗中发挥了重要作用,疫情爆发期间湖北省的中药救治参与率达91.05%,全国其他区域达96.37%[2]。

中成药是中药的重要流通形式。与中成药相关的数据多为半结构化或非结构化,具有4V特征[3]:数据容量大(Volume)、数据增速快(Velocity)、数据来源广(Variety)、真实性不高(Veracity)。传统的关系型数据库关联效率较低,且不易扩展,已无法适应关联性高的中成药数据。知识图谱是大数据时代用于大规模知识管理和智能服务的

收稿日期: 2021-05-11

基金项目: 国家级大学生创新创业训练计划项目(202010412022);江西省教育厅科学技术研究重点项目(GJJ201204);江西省教育厅科学技术研究项目(GJJ170727);江西中医药大学博士启动基金项目(2018WBZR021)

作者简介: 周雪阳(1998-),男,江西中医药大学计算机学院学生,研究方向为医学信息工程;廖诗雨(2000-),女,江西中医药大学计算机学院学生,研究方向为医学信息工程;董泽华(1998-),男,江西中医药大学计算机学院学生,研究方向为医学信息工程;程春雷(1976-),男,博士,江西中医药大学计算机学院副教授,研究方向为机器学习、知识表示与学习;叶青(1967-),女,江西中医药大学计算机学院教授、硕士生导师,研究方向为中医药信息学、计算机应用。本文通讯作者:叶青。

新兴技术，其可以捕捉和呈现领域概念之间错综复杂的关系，并将各种信息系统中分散的知识连接起来。知识图谱技术能有效解决中医药领域的知识岛问题，有助于整合知识资源，提高知识服务能力。基于此，本文构建基于知识图谱技术的中成药可视化与知识问答平台，以便更好地管理和存储关系复杂、种类繁多、结构多变的中成药数据。

1　研究现状

目前已上市的中成药有接近 1 万种[4]，相关大数据亟待开发与利用，但存在以下困难：①中成药数据来源广泛，但却没有统一标准，数据质量不能保证；②中成药数据共享不足，不能充分实现数据价值；③中成药数据没有统一管理规范，存在数据滥用现象[5]。

数据可视化是指将海量数据以图像的形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程[6]。目前，数据可视化不再是简单地利用各种图表对实体及其之间的关系进行展示，通常需要从多维数据、层次关系、文本数据 3 个方面进行可视化研究[7]：①多维数据：使用不同形式对数据进行多维关系展示，使用户能够通过简单操作实现数据的观察与分析，从而获得所需信息；②层次关系：大数据关注的重点往往是不同实体之间的联系，这就要求采用不同图形，尽可能丰富地呈现数据的层次关系；③文本数据：必须结合文本数据帮助用户理解信息，因此应注意文本信息的视觉效果，以便发挥知识问答的作用。

目前，各大医学数据可视化平台多以传统的表格、折线图、直方图等形式展示数量庞大的中成药数据[8]，具有简单直观的特点，但也存在许多问题[9]：①展示视角不够全面，无法综合文献、临床、商业等多背景数据；②图形比例设置不当，导致用户产生视觉误差；③平台过分追求界面简洁，文字说明少，很多数据只通过简单的图形进行展示，导致用户难以解读内容；③没有中医专业人士的参与，平台只能展示提前设定的数据，用户难以找到所需信息。为此，许多学者尝试应用知识图谱技术对中医辨证、案例分析等进行可视化展示，取得了一些研究成果。例如，王菁薇等[10]利用中医经典古籍《伤寒论》中的数据文本进行中医药知识图谱构建；贾李蓉等[11]结合可视化技术研发出中医药知识图谱可视化平台；秦锦玉等[12]基于可视化技术对中医药知识进行整合与可视化展示，开发出一个中医药知识图谱可视化交互平台；陈姗姗等[13]利用知识可视化软件梳理国内有关中医药国际化发展的文献，呈现出该领域的知识图谱；郑懿鸣等[14]将知识图谱与知识卡片相结合，开发出中医药知识图谱用药推荐系统；孙华君等[15]详细分析了知识图谱在中医基础、中医临床、中医养生保健等领域中的应用。在此基础上，本文采用知识图谱技术结合可视化技术对中成药大数据进行研究分析，以期解决中医药数据利用不足、管理不规范等问题。

2　系统设计

首先从中成药数据的实际需求关系出发，选取必要的实体、属性等；然后使用爬虫技术从互联网上获取相关数据，存入对象关系型数据库 Postgresql，通过对中成药数据进行实体识别、关系抽取、整理分析，构建中成药知识图谱并存入图形数据库 Neo4j 中；最后利用 Python 中的 Flask 框架搭建可视化平台，实现数据的多角度展示以及智能问答。具体实现流程如图 1 所示。

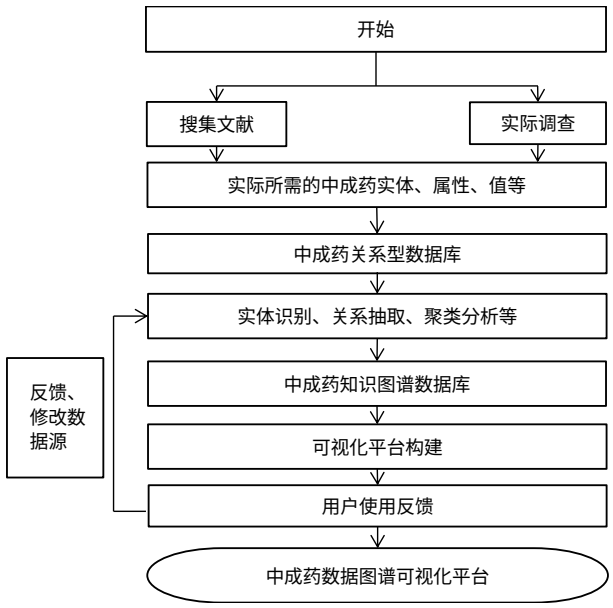

Fig. 1　System implementation flow

图 1　系统实现流程

2.1　数据选取与获取

根据实际调查，将中成药的实体属性分为基本属性、安全性、经济性 3 大类，细分属性如表 1 所示。

Table 1　Detailed properties of Chinese patent medicine

表 1　中成药细分属性

实体属性	属性内容
基本属性	商名、主要规格、用途（主治）、用法用量、组成、产品分类、生产厂家（数量）、性状
安全性	文献研究（数量）、药品禁忌、不良反应
经济性	基要目录、医保目录、标准来源

以中成药名称为检索关键词，使用 Python 爬虫技术从站内众多开放网站中爬取相关网页源码[16]，利用正则表达式技术提取所需数据并存入关系型数据库 Postgresql 中，以实现中成药数据的动态更新。以中成药莱阳梨止咳糖浆为例，其属性如表 2 所示。

2.2　中成药可视化与知识问答平台设计

使用 Python 中 Web 开发模块的 Flask 框架进行中成药知识图谱可视化平台开发。Flask 框架是一个轻量级 Web 开发框架，较其他同类型框架更灵活、安全且容易上手，可开发出功能强大的网站。同时，使用 Ajax 技术进行网站交

Table 2 Property examples of Laiyang pear cough syrup

表 2 莱阳梨止咳糖浆属性示例

属性	内容
商名	莱阳梨止咳糖浆
主要规格	100mL
用 途(主治)	镇咳祛痰,用于伤风感冒引起的咳嗽多痰
用法用量	口服,一次 10mL,一日 4 次
组成	莱阳梨、麻黄、杏仁水、桔梗、远志、北沙参、百合、薄荷脑,辅料为蔗糖、苯甲酸钠
产品分类	药品/中成药/肺系病症
生产厂家(数量)	10
性 状	棕黄色至棕褐色的浓稠液体;气香,味甜、微酸、凉
文献研究(数量)	0
药品禁忌	忌食辛辣、油腻食物
不良反应	卫生部药品标准中药成方制剂第十四册
基要目录	否
医保目录	非医保
标准来源	卫生部药品标准中药成方制剂第十四册

互设计,Ajax 即 Asynchronous JavaScript and XML(异步 JavaScript 和 XML),是一种创建交互式、快速动态应用的网页开发技术,无需重新加载整个网页便能实现部分网页的更新。基于 Ajax 技术的异步交互方法[17]可以实现前端页面的无等待实时刷新,提高人机交互水平,提升用户体验。

中成药可视化与知识问答平台主要分为 8 个模块,详见图 2。用户在搜索框中输入中成药名称,点击搜索后,平台将自动加载出相应信息。

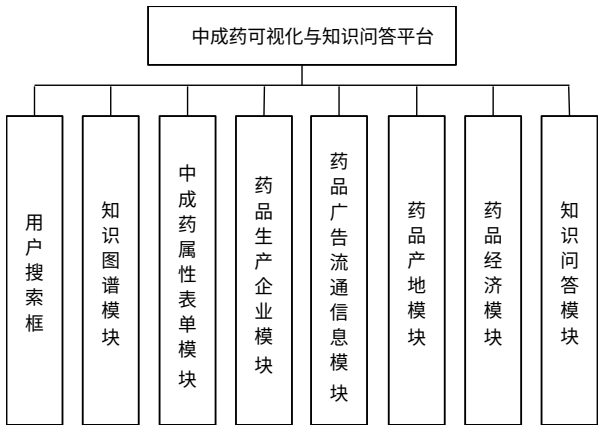

Fig. 2 Design of visualization and knowledge Q&A platform of patent medicine

图 2 中成药可视化与知识问答平台设计

(1)用户搜索框。用户可在搜索框中输入中成药名称,点击搜索按钮即可进行该中成药相关数据的查询。

(2)知识图谱模块。网站以知识图谱的形式将被搜索中成药的相关数据展示出来,用户根据实际需求进行选择性节点展示、刷新知识图谱以及保存下载相关图片等操作。

(3)中成药属性表单模块:网站以表格的形式呈现被搜索中成药相关属性数据,方便用户查看。

(4)药品生产企业模块:网站以词云的形式呈现被搜索中成药的生产厂家信息,用户通过点击某个生产厂家可以在药品广告流通信息模块单独查看该企业生产的药品流通信息。

(5)药品广告流通信息模块:该模块以折线图或柱状图的形式呈现被搜索中成药所有广告在不同年份的流通数量,用户亦可以查看不同生产企业的广告流通信息。

(6)药品产地模块:该模块以饼状图的形式呈现被搜索中成药的产地,用户可查看不同省份生产该中成药的比例。

(7)药品经济模块:该模块以折线图的形式展示被搜索中成药的市场售价信息。

(8)知识问答模块:用户可通过输入中成药名称、药材、症状等关键词检索出相关信息。

3 数据可视化与知识问答应用

3.1 中成药知识图谱可视化应用

基于知识图谱对中成药数据进行存储能有效避免传统关系型数据库的弊端,使非结构化的中成药数据具有更好的关联性,为文本处理提供更为全面的语义特征[18]。

选取中成药的商品名作为实体结点,以基本属性、经济性、安全性作为一级属性结点,其中主要规格、用途、用法用量、组成、产品分类、厂家数量、性状为从属基本属性的二级属性结点;文献研究、药品禁忌、不良反应为从属安全性的二级属性结点;基药目录、医保目录、标准来源为从属经济性的二级属性结点;从数据库中提取的中成药相关数据作为对应属性下的三级实例结点。基于以上信息建立中成药知识图谱数据库,具体示例见图 3、图 4。

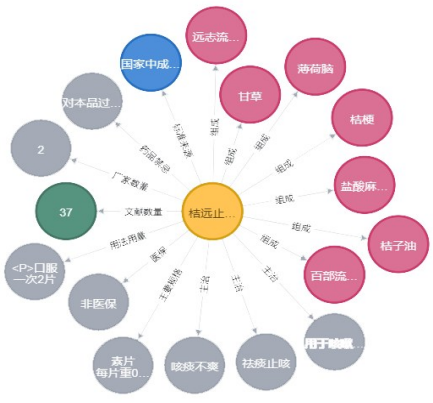

Fig. 3 Example of knowledge map of "Juyuan Zhike Tablet"

图 3 桔远止咳片知识图谱示例

运用 Neo4j 数据库存储中成药相关数据,将单个中成药的相关属性分为基本属性、安全性、经济性 3 大类,分别以不同颜色展示,不同中成药实体根据其相同属性连接起来,形成中成药知识图谱体系,运用图匹配技术实现基于

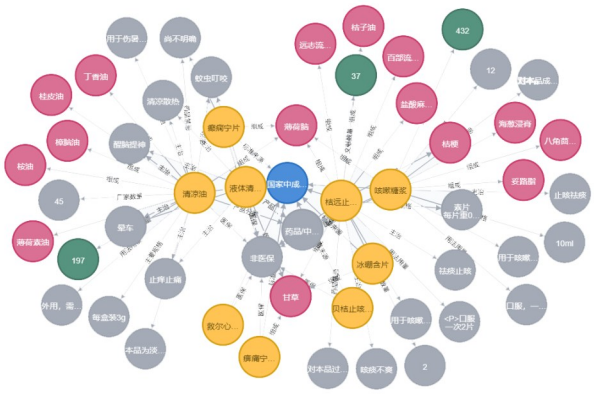

Fig. 4　Knowledge map example of relationship between different Chinese patent medicines

图 4　不同中成药之间关系的知识图谱示例

知识图谱技术的中成药智能问答研究。

采用 Echarts 可视化技术融合大屏可视化背景元素完成中成药知识图谱可视化平台的构建,平台以知识图谱、表格、柱状图、折线图、散点图、词云、饼状图等形式将中成药相关数据呈现给用户,界面展示如图 5 所示。

Fig. 5　Visualization platform of Chinese patent medicine knowledge map

图 5　中成药知识图谱可视化平台界面展示

由图 5 可知,平台以知识图谱和表单的形式呈现被搜索中成药的基本属性、安全性、经济性 3 大类信息,以折线图、柱状图的形式呈现流通信息(以广告招标数量表示),以散点图的形式呈现其市场流通价格信息,以饼状图的形式呈现生产地区信息,以词云的形式呈现生产企业信息。

同时,平台的交互功能可以帮助用户更详细地查看相关数据,具体表现为:①用户通过点击知识图谱模块中的对象、对象属性、关联对象、所属类别图标,可以选择性隐藏部分结点信息;②用户通过点击某个生产企业,可以在药品广告流通信息模块中单独查看该企业生产的药品流通信息。

3.2　知识问答应用

基于中成药知识图谱数据库和中成药知识图谱可视化平台,可进行中成药知识问答研究。通过融入自然语言处理技术,抽取与表征中医药领域文本特征,采用基础的模式匹配与 Transformer 深度学习模型相结合的方式,实现平台的中医药智能问答功能。具体实施流程如图 6 所示。

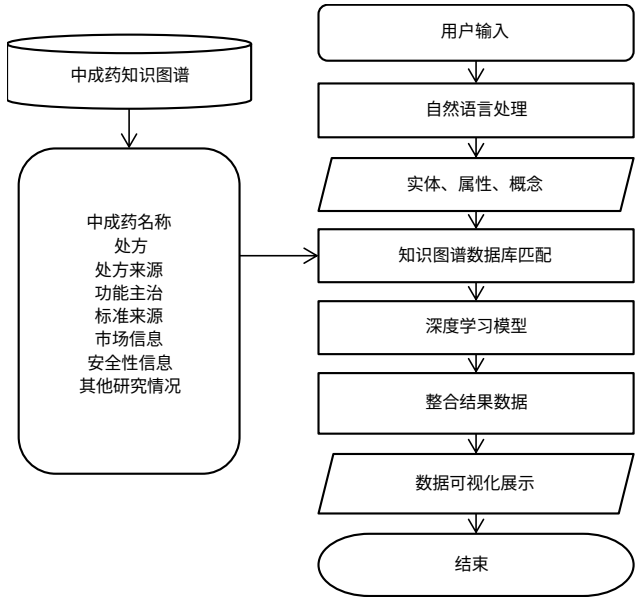

Fig. 6　Implementation process of knowledge Q & A platform

图 6　知识问答平台实施流程

采用 Python 进行自然语言处理,构建中医药问答词典,增加分词精确度,构建相关问题模板,提高问答效率。通过分词、实体识别、关键词提取将用户输入的自然语言文本形成问题列表,通过与已有问答模板进行关键词匹配分析问答类型,再跳转至相应的数据查询模块进行查询,并将所得数据整合成答案传到前端。对于未匹配到的问题文本,采用 Transformer 深度学习模型对问题进行解析,通过自定义中医药知识问答素材抽取问题答案呈现给用户。具体处理流程如图 7 所示。

由图 7 可知,系统通过对输入文本进行自然语言分词处理、文本特征分析、模式匹配等选择合适的模板回答搜索问题,针对未能匹配到的文本,使用 Transformer 深度学

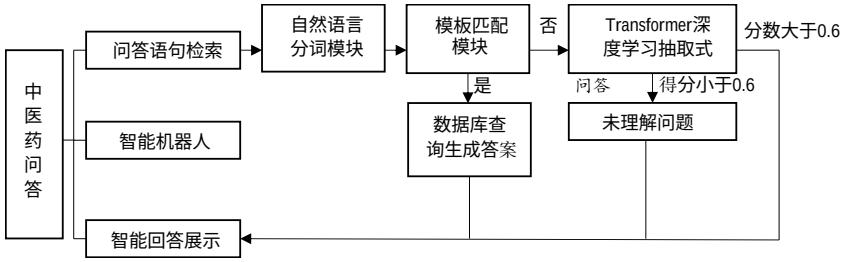

Fig. 7　Process of knowledge Q & A

图 7　知识问答处理流程

习抽取式问答生成答案及得分，分数大于0.6的回答视为正确答案输出，反之则表示未能理解[19]。

4 结语

本文采用Python爬虫技术采集中成药开源数据，通过实体抽取、属性抽取、知识融合等技术成功构建了基于Neo4j图形数据库的中医药知识图谱体系，同时结合Flask框架与Echarts可视化技术搭建了中成药数据可视化平台。目前，本文设计的智能问答功能模块尚且只注重模板匹配以及简单的素材文本抽取式问答，难以达到医药类问答对于精确度的要求。在后续研究中将深入挖掘中医药古籍文献数据，解析中医药多维度数据，并采用深度学习技术构建可泛化计算的智能问答模型[20]，以更好地挖掘中成药数据的应用价值，为大众提供中成药大数据服务。

参考文献：

[1] CHEN K J. Innovative development of traditional Chinese medicine modernization and industrialization [J]. Science&Technology Review, 2020, 38(6):1.
陈可冀. 创新性发展中医药现代化、产业化事业[J]. 科技导报，2020，38(6):1.

[2] ZENG Y, ZHAO M. Deep practice and system construction of traditional Chinese medicine in combating COVID-19 epidemic [J]. Lishizhen Medicine and Materia Medica Research, 2020, 31(4):951-954.
曾予，赵敏. 中医药抗击新冠肺炎疫情的纵深实践及制度构建[J]. 时珍国医国药，2020，31(4):951-954.

[3] CHENG X E, WEN C B, XU Q, et al. Discussion on typical characteristics of TCM big data based on TCM artificial intelligence technology [J]. Modemization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2020, 22(4):1243-1248.
程小恩，温川飙，许强，等. 基于中医药人工智能技术探讨中医药大数据的典型特征[J]. 世界科学技术-中医药现代化，2020，22(4):1243-1248.

[4] LU J W, WANG F, YAN D M, et al. Review and prospect of technology development of proprietary Chinese medicine industry [J]. China Journal of Chinese Materia Medica, 2012, 37(1):5-8.
陆建伟，王芳，颜冬梅，等. 中成药工业科技发展回顾与展望[J]. 中国中药杂志，2012，37(1):5-8.

[5] SHI K L, XIE Q Y, MENG Q G. Research on small data of traditional Chinese medicine in the era of big data [J]. Chinese Archives of Traditional Chinese Medicine, 2019, 37(2):372-377.
石康乐，谢晴宇，孟庆刚. 大数据时代的中医药小数据研究探讨[J]. 中华中医药学刊，2019，37(2):372-377.

[6] QIAN Q. The topic preface of "health care big data management and application" [J]. Data Analysis and knowledge Discovery, 2020, 41(12):1.
钱庆. “健康医疗大数据管理与应用”专题序[J]. 数据分析与知识发现，2020，41(12):1.

[7] JIANG Y, MENG X, WANG Y J. Based on high-quality clinical research, establishing multi-dimensional big data of medical and health [J]. Science Foundation in China, 2021, 35(1): 81-84.
姜勇，孟霞，王拥军. 基于高质量临床研究，建立医疗健康多维度大数据[J]. 中国科学基金，2021，35(1):81-84.

[8] XU X, HUANG Z J, CAI J, et al. Visualization of medical data based on big data research [J]. Chinese Journal of Health Statistics, 2017, 34(2): 347-349.
许茜，黄子杰，蔡晶，等. 基于大数据研究的医学数据可视化[J]. 中国卫生统计，2017，34(2):347-349.

[9] LIU L, ZHU X M, QIU K, et al. Visual comparative analysis of research direction and hot spot of medical big data at home and abroad [J]. China Digital Medicine, 2020, 15(12): 98-101.
刘莉，朱勋梅，邱珂，等. 国内外医疗大数据研究方向及热点可视化对比分析[J]. 中国数字医学，2020，15(12):98-101.

[10] WANG J W, XIAO L, YAN J F. Study on the construction of knowledge map of Treatise on Febrile Diseases based on Neo4j [J]. Computer and Digital Engineering, 2021, 49(2): 264-267, 396.
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.

[11] JIA L R, LIU J, YU T, et al. Construction of knowledge map of traditional Chinese medicine [J]. Journal of Medical Informatics, 2015, 36(8): 51-53, 59.
贾李蓉，刘静，于彤，等. 中医药知识图谱构建[J]. 医学信息学杂志，2015，36(8):51-53，59.

[12] QIN J Y, ZHAI J, CHEN C, et al. Research on visualization technology based on knowledge map [J]. Electronic Design Engineering, 2018, 26 (14): 1-5.
秦锦玉，翟洁，陈程，等. 基于知识图谱的可视化技术研究[J]. 电子设计工程，2018，26(14):1-5.

[13] CHEN S S, SHAO Y J. Visual metrological analysis of the trend and hot spots of internationalization of Chinese medicine in China [J]. Journal of Traditional Chinese Medicine Management, 2017, 25(12): 1-4.
陈姗姗，邵英俊. 国内中医药国际化研究趋势和热点的可视化计量分析[J]. 中医药管理杂志，2017，25(12):1-4.

[14] ZHENG Y M, ZHAI J, HU X L, et al. Intelligent question answering and drug recommendation system based on knowledge map of traditional Chinese medicine [J]. Electronic Technology and Software Engineering, 2019, 8(20): 134-135.
郑懿鸣，翟洁，胡晓龙，等. 基于中医药知识图谱的智能问答与用药推荐系统[J]. 电子技术与软件工程，2019，8(20):134-135.

[15] SUN H J, LI H Y, NIE Y. Knowledge mapping and its application in the field of traditional Chinese medicine [J]. Modemization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2020, 22(6): 1969-1974.
孙华君，李海燕，聂莹. 知识图谱及其在中医药领域应用研究进展[J]. 世界科学技术-中医药现代化，2020，22(6):1969-1974.

[16] PAN X Y, CHEN L, YU H M, et al. Summary of research on theme crawler technology [J]. Application Research of Computers, 2020, 37 (4): 961-965, 972.
潘晓英，陈柳，余慧敏，等. 主题爬虫技术研究综述[J]. 计算机应用研究，2020，37(4):961-965，972.

[17] CHEN L L, ZHANG L, LIU Z L. State based Ajax dynamic Web page extraction in search engine [J]. Computer Applications and Software, 2013, 30(7): 217-220.
陈莉莉，张丽，刘正龙. 搜索引擎中基于状态的Ajax动态网页提取研究[J]. 计算机应用与软件，2013，30(7):217-220.

[18] MA Z G, NI R Y, YU K H. The latest progress, key technologies and challenges of knowledge mapping [J]. Chinese Journal of Engineering, 2020, 42(10): 1254-1266.
马忠贵，倪润宇，余开航. 知识图谱的最新进展、关键技术和挑战[J]. 工程科学学报，2020，42(10):1254-1266.

[19] WANG Z Y, YU Q, WANG N, et al. Review of intelligent question answering based on knowledge mapping [J]. Computer Engineering and Applications, 2020, 56(23): 1-11.
王智悦，于清，王楠，等. 基于知识图谱的智能问答研究综述[J]. 计算机工程与应用，2020，56(23):1-11.

[20] YANG X H, WAN R, ZHANG H B, et al. Knowledge map representation learning algorithm based on symbolic semantic mapping [J]. Journal of Computer Research and Development, 2018, 55(8): 1773-1784.
杨晓慧，万睿，张海滨，等. 基于符号语义映射的知识图谱表示学习算法[J]. 计算机研究与发展，2018，55(8):1773-1784.

（责任编辑：尹晨茹）