

基于中医药学语言系统的知识问答系统设计*

贾李蓉 高 博 刘 静 刘丽红

(中国中医科学院中医药信息研究所 北京 100700)

〔摘要〕 详细阐述基于中医药学语言系统的知识问答系统构建,包括系统设计及具体功能,指出该系统应用能够为临床诊断治疗提供辅助,可在一定程度上缓解医患信息不对等问题。

〔关键词〕 问答系统;中医药学语言系统;知识图谱

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.06.009

Design of the Knowledge Question Answering System Based on the Traditional Chinese Medicine Language System JIA Lirong, GAO Bo, LIU Jing, LIU Lihong, Institute of Information on Traditional Chinese Medicine, CACMS, Beijing 100700, China

〔Abstract〕 The paper elaborates the building of knowledge question answering system based on the Traditional Chinese Medicine Language System (TCMLS), including the system design and specific functions, points out that the application of this system can provide assistance for clinical diagnosis and treatment, and can alleviate the problem of information asymmetries between doctors and patients to a certain extent.

〔Keywords〕 question answering system; Traditional Chinese Medicine Language System (TCMLS); knowledge graph

1 引言

随着生活水平提高、人口老龄化逐年加快,慢

性病患者数量不断增加,医疗资源难以满足公众需求,患者就医率低等问题逐渐凸显。而互联网医疗产业的高速发展为由网络直接获取医疗知识提供更多便利^[1]。目前用户由网络获取医学知识信息的主要途径是百度等传统搜索引擎,只需输入疾病、症状等关键词便出现有关介绍和信息。但通过这种方式获得的信息通常数量巨大且缺乏针对性,其科学性也无法得到保障,智能问答系统应运而生。本文主要介绍中国中医科学院中医药信息研究所自主研发的中医药学语言系统(Traditional Chinese Medicine Language System, TCMLS)为知识基础,以“月经病”为范例构建的知识问答系统,该系统将中医药学语言系统知识与文献中的病案相结合,通过临床实际病历训练不断丰富后台知识库、修正问答算法,进一步提高回答准确度,帮助普通用户更

〔修回日期〕 2021-01-27

〔作者简介〕 贾李蓉,硕士,副研究员,发表论文 30 篇;通讯作者:刘丽红,副研究员。

〔基金项目〕 国家重点研发计划中医药现代化重点专项“多源异构中医科学数据汇交系统研发”(项目编号:2017YFC703505);国家重点研发计划“多维多层多态中医药知识图谱及时空演化模型研究”(项目编号:2017YFB1002302);中国中医科学院基本科研业务费自主选题项目“《中国中医药学主题词表》副主题词研究及平台升级”(项目编号:zz140301)。

为便捷地了解医疗知识、获取更有针对性的医疗信息。

2 系统设计

2.1 概述

中医药学语言系统中的中医知识问答系统是以中医常见病“月经病”为示范构建的智能知识问答系统。系统以自然语言对话形式与患者交流,患者以文本形式输入病情描述,系统据此给出可能患有疾病及其概率以及疾病相关信息,例如常见症状、证候、方药等^[2]。同时探索一种术语反馈模式,根据分词结果进行术语反馈,经专业人员审核后术语补充以不断丰富完善语言系统,使系统建设与应用实现良性循环。

2.1 文献采集与处理

2.1.1 文献类数据 文献采集范围主要包括期刊文献和医案类数据。文献类数据依据要求在中国中医药数据库、中国知网、万方医学数据库以检索式“主题检索=月经病”进行加权拓展检索,检索主题为“月经病”及其下位词“崩漏、闭经、功能性子宫出血、经前期综合征、逆经、痛经、月经过多、月经过少、月经失调”,得到文献约 15 000 篇。以“特征词=病例报告”进行二次检索、文献下载、人工筛选,最终选出符合要求的文献并以其中的病例记录作为本研究的数据基础之一。

2.1.2 其他数据 选取近现代 25 位妇科专家医案,共下载 57 部妇科医案专著。另外选取中医妇科学教材和临床指南以及治疗月经病的常见方剂内容。

2.1.3 数据处理 最终获得病案、指南、文献和电子病历 4 部分基础数据。将数据进行分类,利用文本转换技术对分类数据进行文本提取,部分模糊图片形式的数据采用人工方法进行清洗和整理。

2.2 知识图谱构建

2.2.1 构建原理及方法 基于中医药学语言系

统,利用知识抽取模块产生的知识实体信息形成知识图谱并进行有效存储和管理。随着垂直领域数据的不断积累和对智能化需求的不断提高,垂直领域知识库作用越来越重要。本研究中的数据大多为非结构化数据,针对此类中文数据的自动化知识提取方法较少,数据利用较困难,采用知识抽取、知识融合和图数据库管理等方法进行知识图谱构建^[3]。

2.2.2 信息抽取 知识抽取主要是实现知识概念、属性、关系以及知识分类等实体信息的抽取和融合。知识抽取技术是信息抽取、信息检索、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分。基于中医药学语言系统,结合中文分词、命名实体抽取、关系抽取等自然语言处理关键技术对中医文献进行知识实体信息抽取,为中医领域知识图谱构建提供数据支持。

2.2.3 知识融合 知识抽取实现从非结构化和半结构化数据中获取实体、属性、关系等信息,然而这些信息中可能包含大量冗余和错误信息,数据之间关系扁平化,缺乏层次性和逻辑性,需要对其进行清理和整合。知识融合包括实体链接和知识合并两部分内容,通过知识融合可以消除概念歧义、剔除冗余和错误概念,从而确保知识质量。

2.2.4 知识库管理 基于中医药学语言系统,利用知识抽取模块产生的知识实体信息形成中医药知识图谱并进行有效存储和管理。图数据库是基于数学中图论算法实现的可高效处理复杂关系网络的新型数据库系统。其善于处理大量、复杂、互联、多变的网状数据,效率大幅超越传统关系型数据库。

2.2.5 知识存储、查询和展示 采用 Neo4j 图数据库。知识图谱通常是由异质节点和边组成的具有语义关系的网络,其中节点表示各种知识实体,包括领域基本概念、专业术语、专有名词等,知识实体可以具有多重属性信息,边则表示知识实体间的多种关系,由节点和连接节点的边共同组成网状知识图谱。

2.3 自动问答

2.3.1 概述 自动问答技术通过中医药知识图谱

中的知识实体和关系发现用户提问中的知识关联,通过一定算法进行推理进而获得用户问题答案。自动问答系统主要由问题理解、信息检索和给出答案3部分组成^[4]。用户输入病情描述后,系统首先进行问题理解,基于中医药学语言系统概念表示,利用中文分词、实体命名等技术将用户提问的自然语

言转化为规范知识表示出来;然后基于中医药知识图谱查找相关知识体系,判断用户可能患有的病症及其概率,给出问题相关信息并通过与用户交互反馈形成较为可信的结论,同时利用信息搜索功能搜索具备一定相似度的病案,输出处方建议并为用户提供辅助诊疗建议,见图1。

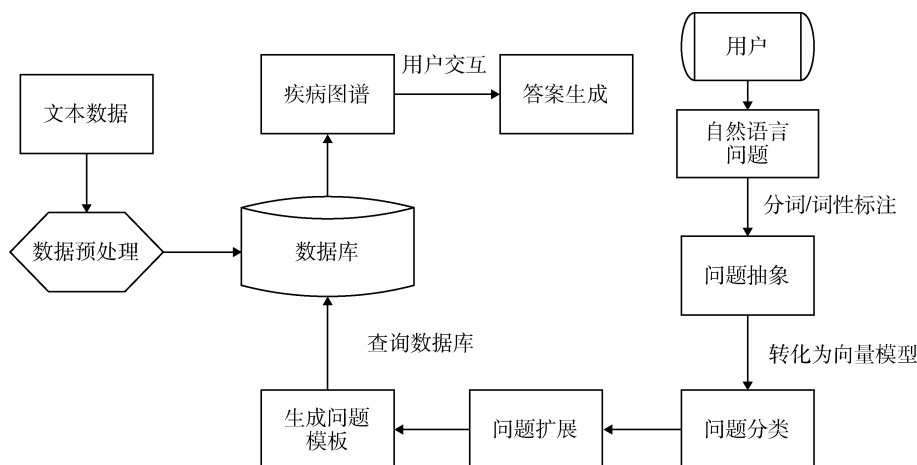


图1 自动问答流程

2.3.2 问题理解 自动问答的关键在于问题理解,其利用分词等技术将用户输入的自然语言转化为计算机可以识别、理解的结构化语言。过程包括智能分词、实体识别、问题类型确定、问题关键词提取等,对后续信息检索和返回答案具有重要影响^[5]。

2.3.3 答案检索 计算机将理解后的语序转化为 Neo4j 的查询语句,在知识库中进行查询。由于查询语言中键值查询是准确匹配的,而用户输入往往是模糊查询,主要利用计算查询词与词之间的相似度完成答案检索工作。检索得到相关实体和属性,通过对相关实体或属性的选择给出更多输入,进一步增强给出答案的可靠性。

2.3.4 提问-回答 为知识问答系统的主要部分。与用户进一步交互,通过交互得到更加精确的答案,最后给出最佳答案。一般用户可在系统给出初步答案后对问题进行修订和补充,经过反复交互和递进式的提问最终获得最佳答案。

2.4 信息检索

采用基于 Elasticsearch 的信息检索,通过搜索

引擎全文索引功能可以快速定位用户所需文章信息,从而提高知识利用效率^[6]。自动问答模块在获取用户数据后提炼出规范知识表示并将其作为输入调用信息检索模块,信息检索模块根据索引信息获得相关病案和知识实体信息并根据相似度进行排序输出。

2.5 术语反馈

2.5.1 新词发现 在信息检索领域,为节省存储空间和提高效率会在处理自然语言数据时过滤掉部分字词,即停用词。传统停用词定义仅考虑一个词在语料中出现的频率,但如果两个词出现及共同出现的频率较高则有可能构成新词。术语反馈是通过知识抽取结果进行分析识别新术语信息,经过专家审核确定新词在术语库中的添加。本研究中术语识别通过关键词抽取、停用词和规则过滤、词频统计完成术语中的新词发现。

2.5.2 短语提取 中医术语大多由两个或两个以上的词组成,例如“小便短黄”“小腹胀痛”等。短语在语法和语义结构上较单词更完整,提取中文短语即固定多字词表达串的识别,从陌生文本中自

动发现固定短语。短语提取经常用于自动推荐、新词识别等领域。关键短语识别常采用抽取式和生成式方法。

3 功能介绍

3.1 知识图谱

3.1.1 人工修正 包括两部分：一是标识未识别的实体，二是修正已识别的实体和关系。如文档中的“苔薄白”“经行错后”“舌边尖紫暗”等实体未被识别，可人工将其标出并归入“症状”类；可人工修改已标注错误的实体类型，如“王不留行”被识别为“症状”类，点击修改为正确类别。

3.1.2 知识图谱生成 通过对文献或医案内容进行分析及人工修正后得到单篇文件的知识图谱。将每篇文献产生的知识图谱进行人工标注修正后上传至系统，作为自动问答的基础资源。系统提供每篇文献知识图谱，标识该篇文献知识内容与结构，也可展示整个系统所有文献的图谱，可点击每个点查看具体信息。

3.2 自动问答

自动问答系统尚未实现用户完全自由提问，而是为用户提供可提问类型选择，如疾病、治疗方法、方药、病案等。例如用户可输入问题“什么疾病有小便短黄、月经量等症状”，点击“提交”按钮后系统首先返回知识图谱，其中显示与“小便短黄、月经量少”相关的疾病和症状。同时用户根据给出的图谱提示补充病症、病情描述以获得治疗建议。治疗建议按可能性大小排序给出疾病名称和治疗方案。

3.3 信息检索

中医药知识问答系统后台由庞大知识库支撑，其中蕴含大量中医药知识。系统为用户提供全文检

索功能。在关键字检索框内输入自由词即可检索包含该自由词的所有文本，点击文本标题即可查看详细文本信息。

4 结语

本文对基于文献采集、知识图谱构建、信息检索和术语反馈等关键技术构建的自动问答系统介绍。该系统以中医常见病“月经病”为示范，从电子病历、医案、医学指南、医学文献中抽取症状、用药等知识实体构建知识图谱。系统根据用户所提问题和辅助信息，通过知识图谱查找和智能问答技术给出答复，包括患者可能患有的疾病、参考病案及其概率等，为临床诊断治疗提供辅助。中医药知识问答系统可进行智能搜索，通过检索为用户提供有价值的信息，用户可高效、准确、快速地查询相关信息，帮助用户自我诊断，一定程度上缓解医患信息不对等问题；该系统还可用于中医临床教学，为年轻医生提供一定学习参考。

参考文献

- 1 陆森, 徐霄玲, 陈园园, 等. 基于人工智能技术的医学知识问答系统研究 [J]. 软件导刊, 2019, 18 (12): 65-68.
- 2 梁耀波. 智能医疗诊断系统的研究与实现 [D]. 北京: 北京理工大学, 2016.
- 3 贾李蓉, 刘丽红, 刘静, 等. 基于中医药学语言系统的知识问答系统的设计与构建 [J]. 中华医学图书情报杂志, 2019, 28 (5): 11-14.
- 4 罗鹏, 张宾, 高亚斌, 等. 中文知识图谱在商品咨询自动问答系统中的应用 [J]. 自动化技术与应用, 2020, 39 (6): 173-176.
- 5 朱淑媛, 罗军. 基于本体的领域自动问答系统研究 [J]. 计算机应用与软件, 2019, 36 (8): 98-105, 154.
- 6 郭鹏. 基于 LUCENE 技术的知识库搜索引擎设计与实现 [D]. 北京: 北京邮电大学, 2011.