基于爬虫技术和电子病历的糖尿病知识图谱的构建

杨美洁① 熊相超①

摘 要 目的：利用Python网络爬虫及电子病历数据构建糖尿病知识图谱。方法：利用Protege构建糖尿病本体库，Neo4j进行知识存储和知识推理，Django框架开发Web端和pyahocorasick模块实现自然语言的处理。结果：实现了糖尿病知识图谱问答系统的患者信息查询和知识查询功能。结论：辅助临床医生进行医疗决策，便于普通大众了解糖尿病基本知识。

关键词 糖尿病 知识图谱 知识存储 知识推理 问答系统

Doi:10.3969/j.issn.1673-7571.2020.02.002

[中图分类号] R319; TP301 **[文献标识码] A**

Construction of Diabetes Knowledge Map Based on Reptilian Technology and Electronic Medical Record / YANG Mei-jie, XIONG Xiang-chao//China Digital Medicine.-2020 15(02): 06 to 08

Abstract Objective: Use the data from Python web crawler and electronic medical record to build a diabetes knowledge map. Methods: It used protege to build a diabetes ontology library, Neo4j for knowledge storage and knowledge reasoning, Django framework was used for development web and pyahocorasick module for natural language processing. Results: The patient information query and knowledge query function of the diabetes knowledge map question answering system was realized. Conclusion: It assisted clinicians in making medical decisions and made it easier for the general public to understand the basics of diabetes.

Keywords diabetes, knowledge map, knowledge storage, knowledge reasoning, question answering system

Fund project Special Project of Science and Technology Innovation of Social Undertaking and People's Livelihood Guarantee of Chongqing (No. cstc2015shms-ztzx10003); University Student Innovation Training Project of Chongqing Medical University in 2018 (No. CXSY201825)

Corresponding author Medical Informatics College, Chongqing Medical University, Chongqing 400016, P.R.C.

1 引言

根据国际糖尿病联盟2017年的统计数据显示中国是糖尿病人数最多的国家[1]。如何高效地治疗和预防糖尿病已成为重要的问题[2]。

目前糖尿病的研究主要集中在并发症、治疗及护理等方面[3-4]，国外主要关注糖尿病的分型和护理[5]。知识图谱是用可视化技术描述知识资源及其载体，显示知识发展进程与结构关系，挖掘、分析和显示知识及其相互联系[6]。国内的研究主要集中于图书馆学、情报学等领域[7-8]。目前的研究均未实现糖尿病知识图谱可视化。

本文利用Python爬虫爬取寻医问药网、39健康网等网站以及从电子病历中获取糖尿病的症状体征、患者信息、并发症、食物、药物、检查检验等数据，利用本体建模工具Protege构建糖尿病知识模型，通过知识推理挖掘隐含的知识，推出实体间的语义关系。将处理过的数据和知识存储在Neo4j图数据库中，采用相关技术构建糖尿病知识图谱自动问答系统的Web端，后台通过对前台的信息进行自然语言处理和语义规则匹配，通过Neo4j图数据的Cyper语言获取结果数据，并通过前台将结果信息展示给用户。

2 相关技术介绍

2.1 Python爬虫 利用Python爬虫获取糖尿病简介、病因、预防、并发症、诊断、治疗等数据，并将数据存储在MongoDB数据库中。

2.2 糖尿病电子病历数据预处理 数据预处理主要包括数据清洗、集成和转

基金项目： 重庆市社会事业与民生保障科技创新专项（编号：cstc2015shms-ztzx10003）；2018年重庆医科大学大学生创新训练项目（编号：CXSY201825）

①重庆医科大学医学信息学院，400016，重庆市渝中区医学院路1号

换。本文利用SQL技术从重庆市某三甲医院电子病历中提取糖尿病数据1 209条，经过数据预处理后得到804条。

糖尿病知识图谱开发流程如图1所示。

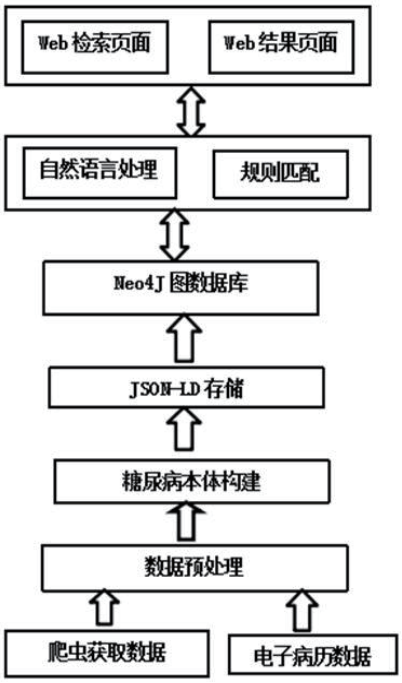

图1 糖尿病知识图谱开发流程

3 糖尿病知识图谱知识表示及知识建模

3.1 糖尿病知识图谱知识表示 知识表示是将知识数字化、系统化，便于计算机识别、存储和处理，知识表示是自然语言处理的基础[9]。本文采用了基于语义网络的表示方法和JSON-LD的知识表示方法。

3.2 糖尿病知识图谱知识建模 知识建模指从知识获取到知识完成的形式化表示的过程。主要包括知识获取、本体构建、基于本体的知识表示三部分[10]。本文使用Protege 5.0软件构建了糖尿病、并发症、药品、症状体征、患者、饮食和检查检验七大类，并通过定义实体对象属性和实体数据属性，实现糖尿病知识建模，如图2所示。

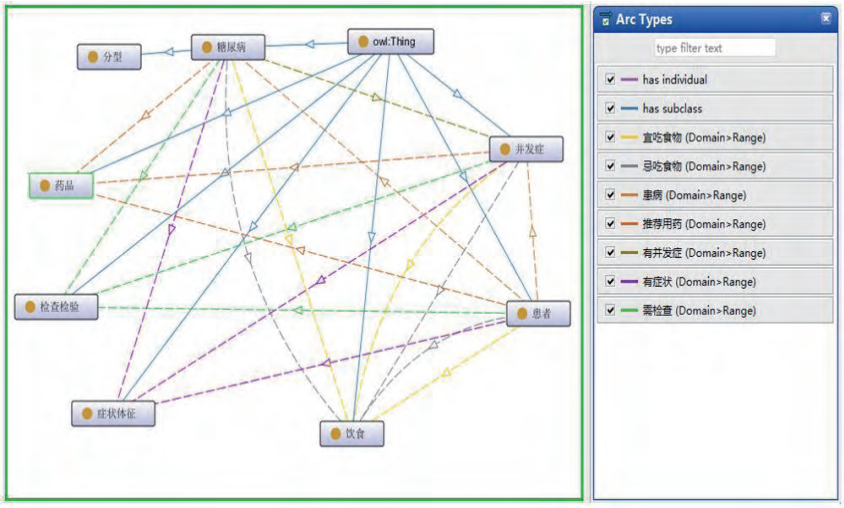

图2 糖尿病知识建模

对已构建的糖尿病实体类进行实体对象属性定义，包括：有并发症、有症状、患病、推荐用药、需检查、忌吃食物、宜吃食物和并发症；对已构建的糖尿病实体类进行数据属性定义，糖尿病和并发症属性：疾病名称、疾病描述、病因、预防、患病人群、治愈率、治愈周期、治疗方式；患者属性：年龄、性别、住院号、个人史、既往史、家族史、现病史、体格检查、科室；药品属性：药品名称、功能主治、用法用量、不良反应、禁忌；检查检验属性：检查检验名称、正常值、临床意义；症状属性：症状名称、症状描述；食物属性：食物名称、食物功效。

4 糖尿病知识存储及知识推理

知识图谱在逻辑上分为模式层和数据层，数据层由一系列的事实组成，知识图谱中的知识以事实为单位进行存储[11]。本文采用JSON-LD表达事实，知识图谱中的数据关系是图形

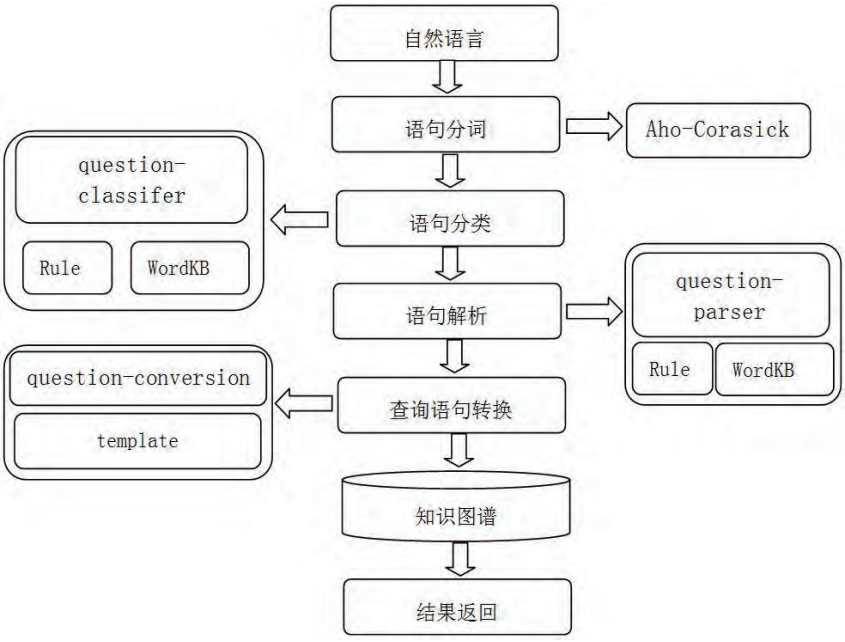

图3 基于自然语言处理的问答框架

的，利用图数据库Neo4j进行知识存储。知识推理注重知识与方法的选择和运用，通过知识推理可以对已有的知识进一步挖掘，推出缺失事实，完成问题求解[11]。图数据库Neo4j通过遍历各节点能够推理出糖尿病知识图谱中各节点间可能存在的关系。

5 基于糖尿病知识图谱的问答系统关键技术

本文采用Python的pyahocorasick模块实现自然语言的处理，该模块基于AC自动机（Aho-Corasick Automation）的分词模块，使用Trie（前缀树）数据结构的多模式匹配算法，通过AC分词系统，对用户输入的语句进行自然语言处理[12]。自然语言处理的问答框架如图3所示。

6 基于糖尿病知识图谱的问答系统实现

利用Python Web框架（Django）、Neo4j数据库、腾讯云服务器等相关工具，开发糖尿病知识图谱的自动问答系统。并用图数据库Neo4j存储糖尿病数据。Neo4j糖尿病数据库如图4所示。

糖尿病知识图谱自动问答系统的首页主要包括系统导航、关于我们、支持问答形式三个模块，供用户查看、了解及使用。

在该页面中用户输入问题，系统通过自然语言处理解析问题，然后从Neo4j数据库中匹配与问题相关的数据，匹配成功的结果返回到该页面。

7 结语

通过Python爬虫及电子病历获取数据，利用Protege构建糖尿病本体库，利用相关技术开发糖尿病知识图谱平台。糖尿病知识图谱的构建，能够辅助临床医生进行医疗决策，便于普通大众了解糖尿病相关知识。糖尿病知识图谱的构建也为其他慢病（高血压）的知识图谱构建提供借鉴。

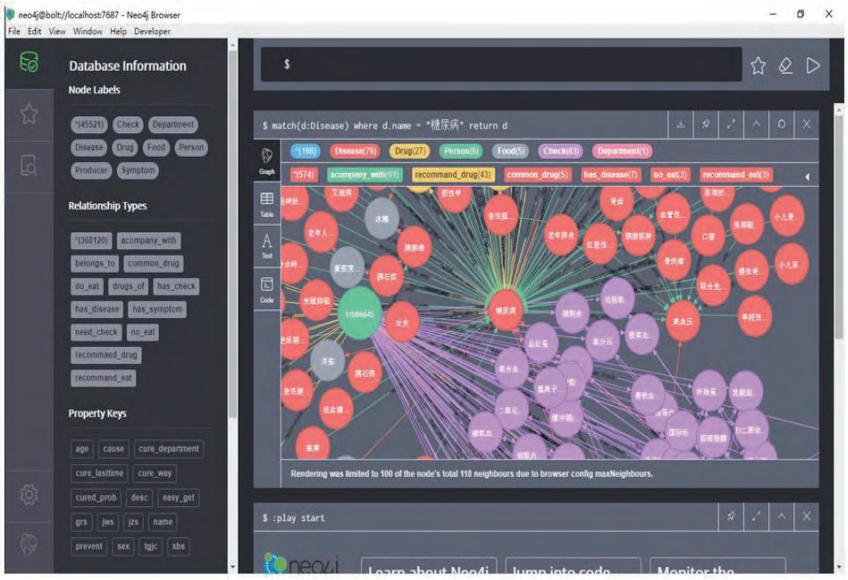

图4 糖尿病数据的Neo4j图

参考文献

[1] International Diabetes Federation.IDF Diabetes Atlas[S].8th ed.Brussels,2017.

[2] 侯清涛,李芸,李舍予,等.全球糖尿病疾病负担现状[J].中国糖尿病杂志,2016,24(1):92-96.

[3] 周海龙,杨晓妍,潘晓平,等.中国人群糖尿病疾病负担的系统评价[J].中国循证医学杂志,2014,14(12):1442-1449.

[4] 彭笑笑,马秀华.妊娠合并糖尿病诊断及早期筛查研究进展[J].中国医刊,2018,53(10):1096-1098.

[5] Hempler NF,Joensen LE,Willaing I.Relationship between social network, social support and health behaviour in people with type 1 and type 2 diabetes: cross-sectional studies[J].BMC Public Health, 2016(16):198.

[6] 刘峤,李杨,段宏,刘瑶,等.知识图谱构建技术综述[J].计算机研究与发展,2016,53(3):582-600.

[7] 曹倩,赵一鸣.知识图谱的技术实现流程及相关应用[J].情报理论与实践,2015,38(12):127-132.

[8] 张思琦,张文兰,李宝.国外近十年深度学习的研究现状与发展趋势——基于引文分析及共词矩阵的知识图谱分析[J].远程教育杂志,2016,35(2):64-72.

[9] 文亮,李娟,刘智颖,等.基于概念层次网络的知识表示与本体建模[J].中文信息学报,2018,32(4):66-73.

[10] 徐增林,盛泳潘,贺丽荣,等.知识图谱技术综述[J].电子科技大学学报,2016,45(4):589-606.

[11] 袁凯琦,邓扬,陈道源,等.医学知识图谱构建技术与研究进展[J].计算机应用研究,2018,35(7):1929-1936.

[12] 徐懿彬.基于Aho-Corasick自动机算法的概率模型中文分词CPACA算法[J].电子科技大学学报,2017,46(2):426-433.

【收稿日期：2019-04-23】
【修回日期：2019-08-04】
（责任编辑：肖婧婧）