# 基于中医医案的知识图谱构建*

羊艳玲 李 燕 帅亚琦 陈月月

（甘肃中医药大学信息工程学院 兰州 730000）

〔摘要〕 介绍知识图谱构建过程、数据来源、数据标准化方法、人工序列标注方法及关系定义，详细阐述中医医案知识图谱构建方法，包括知识表示、知识存储、知识可视化分析，为相关研究提供参考。

〔关键词〕 知识图谱；中医医案；可视化展示

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.10.009

**Construction of Knowledge Graph Based on Medical Records of Traditional Chinese Medicine** *YANG Yanling, LI Yan, SHUAI Yaqi, CHEN Yueyue, School of Information Engineering, Gansu University of Chinese Medicine, Lanzhou 730000, China*

〔**Abstract**〕 The paper introduces the construction process of knowledge graph, data sources, data standardization methods, manual sequence annotation methods and relationship definition, and elaborates the construction methods of medical record knowledge graph of Traditional Chinese Medicine (TCM), including knowledge representation, knowledge storage and knowledge visualization analysis, so as to provide references for related research.

〔**Keywords**〕 knowledge graph; medical records of Traditional Chinese Medicine (TCM); visual display

## 1 引言

经过几千年的发展中医学积累了丰富的临床经验，形成众多经典理论。如何运用信息科学技术挖掘、整理与分析中医学知识体系以及隐含在医案文献中的学术思想、临床经验和辨证方法是值得探讨的重要课题。随着信息技术不断发展，可视化技术越来越成熟，在知识工程领域引进知识图谱概念，使用知识图谱的主要目的是描述现实世界的概念、实体及其之间的相互关系，从而实现对知识的共建、共享以及重用[1]。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。于彤、刘静和贾李蓉等[2]以中医药学语言为骨架构建大型中医药知识图谱；张德政、谢永红和李曼等[1]提出基于本体的中医核心知识图谱及其构建方法；聂莉莉、李传富和许晓倩等[3]基于自然语言处理方法自动构建基于“疾病－症候－特征”3层结构模型的医学诊断知识图谱。本文拟在已有研究基础上进一步利用知识语义化、数据易关联的特性将中医医案中蕴藏的知识结构或相互关系予以可视化展示，主要围绕中医诊疗路径展开，完整的诊疗路径以症状为出发点，依次为证候、治法、处方、药物，具有逻辑鲜明的层次关系特征，以期为名老中医传承经验提供参考。

〔修回日期〕 2021－10－14

〔作者简介〕 羊艳玲，硕士研究生，发表论文4 篇；通信作者：李燕，副教授。

〔基金项目〕 国家中医药管理局项目“甘肃省基层医疗卫生机构中医诊疗区健康信息平台”（项目编号：2305181101）；甘肃中医药大学研究生创新基金“基于深度学习的高血压中医医案知识图谱的构建”。

## 2 相关知识及研究基础

### 2.1 知识图谱概述

知识图谱是大数据时代背景下针对海量知识的一种新型管理与服务模式，被视为一张巨大的图，其中节点表示实体，边代表实体间的语义关系。知识图谱通过对结构分散的知识进行重新组织、汇聚整理，提高知识资源关联与整合程度，为解决“知识孤岛”问题提供理想的技术手段[4]。目前知识图谱构建过程主要包括数据获取、知识抽取、知识融合和知识加工 4 个步骤[5]。其中数据获取是基础，数据源包括结构化、半结构化及非结构化数据，知识图谱应用于医疗领域时，主要的数据来源为医学专业论文、书籍文献、医案和电子病历等。知识抽取的基本原理是将已有非结构化和半结构化数据中的知识用不同种格式或表示方法提炼出来，清晰展示数据中包括的主要内容，再将其处理为相同形式数据的过程，主要包括实体抽取、关系抽取和属性抽取 3 个部分。在获取实体、关系及属性信息后，要对其进行清理和整合，即知识融合，包括共指解析和实体消歧，保证知识的正确性和逻辑性。最后通过知识加工，包括本体抽取、知识推理、知识发现和质量评估，最终得到结构化、网络化的知识体系形成的知识图谱，见图 1。

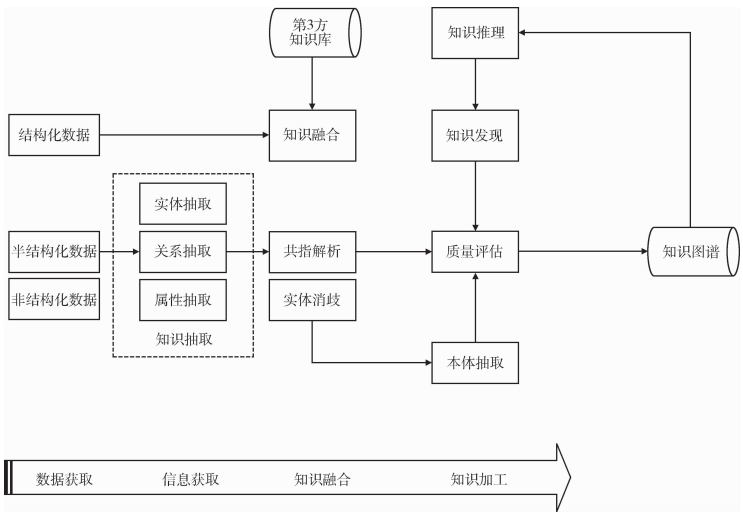


**图 1 知识图谱构建过程**

### 2.2 数据来源

本文研究数据主要来源于中国中医科学院中医药信息研究所研制的古今医案云平台软件[6]，整理平台上所有与高血压疾病相关的中医医案。纳入标准如下：医案中明确记载诊断为高血压或眩晕的患者；就诊时的主诉辨治以高血压为主；数据完整，包含临床表现、病机分析、治法和用药等内容。依照权威诊断标准和名师指导意见对平台中高血压疾病相关医案进行手动检索与筛选，并对其内容进行规范，按照序号、ID、患者姓名、性别、年龄、医案内容、中医疾病、证候和医案来源等类别录入到 Excel 中。研究过程中对中医医案中用作训练的数据集进行整合，用单字切分原始文本，对训练集中的所有语句按照疾病、症状、证候、治法、处方进行分类，最后共录入 435 条医案数据。

### 2.3 数据标准化

中医医案是医者在诊疗过程中自然语言的描述，其表述缺乏规范性和标准性。目前中医医学词典和知识库较少，增加了学者研究医学知识图谱的成本和难度。此外由于中医医案尚未统一，具有多样化特点，对于医案术语、计量单位等未做明确要求，同一个实体有多种表达形式，难以适应信息时代要求，也为医学实体消歧带来困难。针对上述问题进行以下处理：首先将已整理的医案导入古今医

案云平台进行标准化，黑色字体代表与标准表完全匹配，已被标准化；原始值红色，标准值黑色代表模糊匹配标准值，提示可查看是否匹配正确；皆为红色表示匹配不到标准值，可进行选择操作。如“心虚肝郁、痰火扰心 = 肝郁证，痰火证”“化瘀涤痰 = 化瘀”等，依据标准替换不规范的术语，把握图谱节点内容的一致性。

### 2.4 人工序列标注

序列标注即对给定序列中的元素进行标注，赋予对应标签，并在这些标签基础上对序列做进一步深度分析，是自然语言处理过程中常需解决的问题。对于实体识别的等量标注任务，标签由两部分组成：实体类别和实体中的位置。采用 BIO 表示实体类别和位置，将每个元素标注为“B - X”“I - X”或者“O”，再以字符作为最小标注单元。在 BIO 表示中，B 代表实体头部，I 表示中间实体，O 代表实体尾部，X 表示实体类型。在标注过程中，对中医实体以“标签，实体”形式将其归属到对应的中医类别，见表 1。

表 1 BIO 标签集

| 实体标记 | 头部标记 B | 中间标记 I | 尾部标记 O |
|---|---|---|---|
| 疾病 | B - disease | I - disease | O - disease |
| 症状 | B - symptom | I - symptom | O - symptom |
| 证候 | B - syndrome | I - syndrome | O - syndrome |
| 治法 | B - treatment | I - treatment | O - treatment |
| 处方 | B - prescription | I - prescription | O - prescription |

### 2.5 关系定义

知识图谱本质是定义实体和实体之间联系知识的关系。实体作为图谱知识节点的一种表现方法，主要目的是用来表达知识结构与概念之间的关系。知识图谱集中每个实体都包含其名称、定义和注释。通常将实体关系定义为 < 实体、关系、实体 >，其中实体是疾病、症状、证候、治法、处方和药物，并且关系可用于连接两个实体[7]。最终共确定 632 个实体、495 种关系，其之间的关联，见图 2。

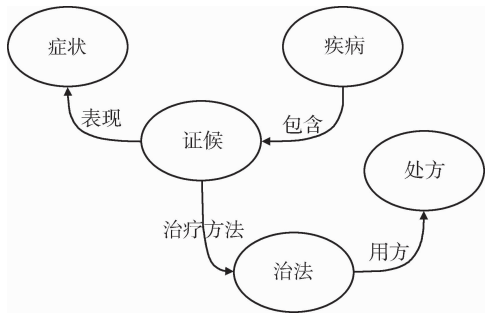


图 2 中医实体关系层

## 3 中医医案知识图谱构建

### 3.1 知识表示

知识图谱是一种可以使用属性图模型来表示的图数据结构，属性图模型主要是由节点和连边组成，节点在知识图谱概念中表示现实世界中的实体，连边用来表示实体与实体之间的关系，而且节点和连边可以包含多个属性，即通过节点集合和边集合构造关系图。其中节点表示数据集中识别出的命名实体，其具有唯一的标识符和若干条属性值；边表示数据集中抽取的命名实体之间的关系，其具有唯一标识符和若干条属性值[8]。在简单的属性图模型中，“眩晕”包含“肝肾亏虚，血络瘀阻”，继而表现“头晕目眩”。节点表示数据集中识别的实体，“眩晕”为疾病实体，具有别名、并发症、证候等属性值；“肝肾亏虚，血络瘀阻”为证候实体，具有症状、类型等属性值；“头晕目眩”为症状实体，具有类型、表现部位等属性值。边集合中关系表示为 D Include S Represent S′，其中 D 表示疾病（Disease），S 表示证候（Syndrome），S′表示症状（Symptom），见图 3。


图 3 疾病、症状、证候属性图模型

### 3.2 知识存储

知识图谱的最大优点是可以利用空间形象的表现来展示知识点间的联系。在可视化展示方面，以图结构存储知识并通过 Neo4j 实现可视化阶段，在众多数据库系统中 Neo4j 具有高性能、设计灵活、开发便捷等优势，用户可以使用 Cypher 语言操作数据[9]。Neo4j 最重要的两个元素是实体和实体之间的关系，分别为节点和连边。

### 3.3 知识可视化分析

可视化是指将知识单元之间的关系转化为能够更好理解的图形形式，用以表现抽象的事物。Neo4j 控制台的图形界面具有将存储的知识单元和知识单元之间的关系转换为知识图的功能，可以方便地查看知识图中的关系信息[10]。Neo4j 批量导入前文提取的实体和关系后，采用 Cypher 查询语言获取满足条件的数据，以可视化图形展示出来。数据可视化用于基于知识图的查询结果可视化，包括中医知识查询和中医诊疗路径。图 4、图 5 分别展示本文提取的实体及关系的部分可视化图，在图谱中可以自定义图谱内容以显示更为清晰的内容，在图谱中关系图中连边表示不同类别实体间的语义关系，图 4 为疾病 - 症状 - 证候（Disease Include Syndrome，Syndrome Represent Symptom）可视化图，西医高血压在中医中主要以“眩晕”和“头痛”进行表述，证候主要是“肝肾阴虚”“肝火上炎”“脉络瘀阻”“气机不畅”等。图 5 展示治法 - 处方，每个治法对应相应的处方，也可以看到不同处方之间药物也有所关联。从图中看到高血压的中医名称不具有唯一性，一个具体的疾病实体关联着多个不同的症状实体，且一个具体症状实体关联着不同疾病。因此将疾病实体与症状实体对应后，可以根据患者表现出的症状推断患者可能患有的疾病，根据症状信息，基于知识图谱，结合多种中医方法进行辨证论治策略的推荐。知识图谱能够帮助用户快速发现所关注的知识扩展及衍生，更好地掌握中医药知识体系，并在浏览中发现具有潜在关联的“知识孤岛”。

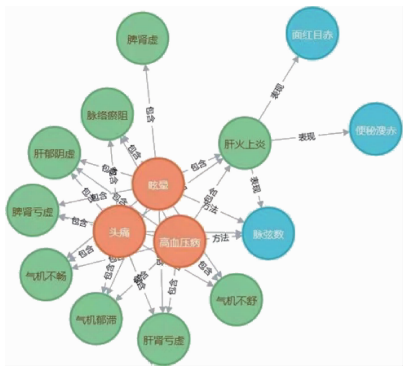


图 4 疾病 - 症状 - 证候可视化

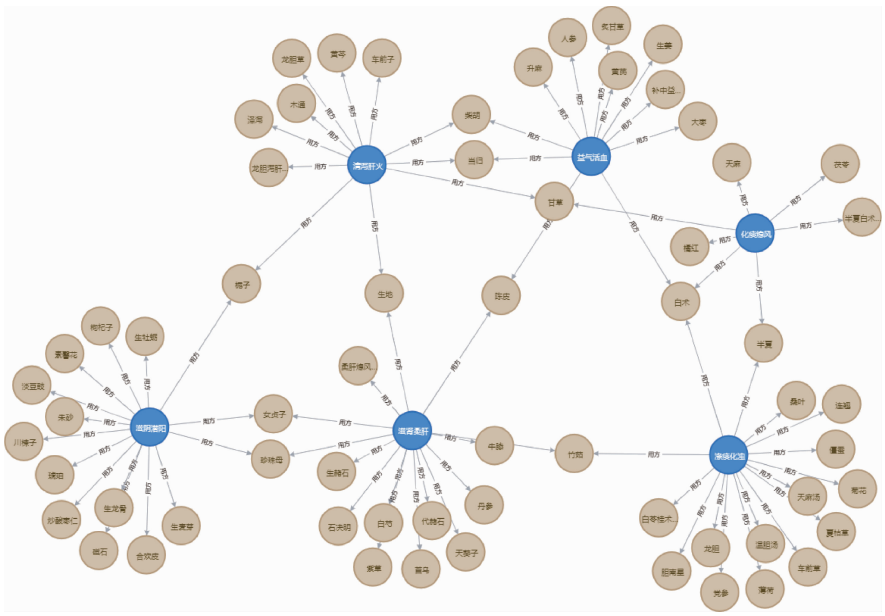

图 5 治法 - 处方可视化

## 4 结语

中医学的诊疗主旨为“辨证论治”，在中医医案中有充分体现。中医医案的记录以辨证思路为核心，强调名老中医之间的差异性[11]。中医药知识图谱的构建实质是中医医案到知识图谱的知识转换，是一个知识抽象和归纳的过程。在这个过程中，一方面基于中医医案等临床知识源，通过疾病、证候、症状、治法、处方等核心概念对医案文本进行分析和标注，完成知识抽取；另一方面，构建中医医案知识图谱实现医药知识的结构化表示。将知识图谱应用于中医临床，可实现智能化、个性化的中医药服务，促进与中医临床互融互通，揭示中医实体间的相关关系，辅助医生临床研究与决策。但是当前不同医案相对零散且大多基于非结构化数据，较难对医案中的知识进行高效管理。针对上述问题，本文将医案中疾病、症状、证候、治法、处方、药物实体进行命名实体识别和抽取，在此基础上以知识图谱的形式将其关联起来探索其中关系，以“病证症”结合的方式探讨高血压相关的中医疾病名称对应的症状所关联的治法；以“方药”结合探索该治法所涉及处方以及对应的药物组成。本文采用的中医领域知识有限，构建的中医知识图谱只是一个实验性知识库，相较于大型知识图谱，本文所构建的知识图谱中的实体以及实体间的关系较简单，需要更多中医专家参与进一步完善；且因不同医家对疾病的具体证型和划分标准存在差异，在对不同名老中医医案进行收集和整理时较难实现标准化和规范化。随着医案数量增加和中医药临床知识划分标准的形成，知识图谱与中医药文献、医案、电子病历等的知识联系，在中医药事业发展、全方面医学领域知识图谱构建方面将发挥更加重要的作用。

## 参考文献


1 张德政，谢永红，李曼，等. 基于本体的中医知识图谱构建［J］. 情报工程，2017，3（1）：35－42.

2 于彤，刘静，贾李蓉，等. 大型中医药知识图谱构建研究［J］. 中国数字医学，2015，10（3）：80－82.

3 聂莉莉，李传富，许晓倩，等. 人工智能在医学诊断知识图谱构建中的应用研究［J］. 医学信息学杂志，2018，39（6）：7－12.

4 于彤，李敬华，朱玲，等. 中医临床知识图谱的构建与应用［J］. 科技新时代，2017（4）：51－54.

5 Suchanek F M，Kasneci G，Weikum G，et al. Yago：a Large Ontology from Wikipedia and WordNet［J］. Journal of Web Semantics，2008，6（3）：203－217.

6 李永苗. 基于 BiLSTM 的中文电子病历知识图谱构建及实现［D］. 杭州：浙江工业大学，2020.

7 邓宇，周卫强，张振铭，等. 基于名老中医医案的知识图谱构建［J］. 湖南中医杂志，2019，35（7）：186－187.

8 Wu Q，Kuang Y，Hong Q，et al. Frontier Knowledge Discovery and Visualization in Cancer Field Based on KOS and LDA［J］. Scientometrics，2019，118（3）：979－1010.

9 黄梦醒，李梦龙，韩惠蕊. 基于电子病历的实体识别和知识图谱构建的研究［J］. 计算机应用研究，2019，36（12）：3735－3739.

10 于琦，李敬华，李宗友，等. 基于本体的中医医案知识服务与共享系统构建研究［J］. 中国数字医学，2017，12（5）：103－105.

11 李园白，崔蒙. 关于中医医案的综合性分析研究近况［J］. 中国中医药信息杂志，2006（2）：91－93.