

基于疾病症状权重信息的医疗知识图谱构建

王宁¹, 吴靖¹, 王远², 翟文彪¹, 傅新杰^{1*}

¹ 嘉兴职业技术学院互联网学院 浙江嘉兴

² 嘉兴大学生物与化学工程学院 浙江嘉兴

【摘要】针对互联网医疗信息冗余、真实性存疑及现有知识图谱疾病覆盖局限、症状权重缺失等问题,本研究提出一种融合多源异构数据的医疗知识图谱构建方法。通过整合权威医疗网站半结构化数据与开源医疗知识图谱,设计多阶段构建流程,并引入基于联合搜索频次的疾病-症状动态权重计算方法,最终构建了一个包含疾病症状权重信息的医疗知识图谱。该图谱不仅解决了疾病症状权重关系不明确的问题,还实现了对碎片化医疗信息的有效整合与深度挖掘,为医疗领域的相关研究提供有价值的数据资源。

【关键词】信息冗余; 疾病症状权重; 医疗知识图谱

【基金项目】本文受嘉兴职业技术学院校级科研年度项目资助,项目号:校 2024131、校 2024040、校 2024087; 嘉兴市青年科技人才专项,项目号:2024AY40025

【收稿日期】2025 年 4 月 18 日

【出刊日期】2025 年 5 月 28 日

【DOI】10.12208/j.ijcr.20250260

Construction of medical knowledge graph based on symptom weight information of diseases

Ning Wang¹, Jing Wu¹, Yuan Wang², Wenbiao Zhai¹, Xinjie Fu^{1*}

¹Internet College, Jiaxing Polytechnic, Jiaxing, Zhejiang

²College of Biological and Chemical Engineering, Jiaxing University, Jiaxing, Zhejiang

【Abstract】To address the challenges of redundant and unreliable internet-based medical information, as well as the limitations of existing knowledge graphs in disease coverage and symptom weight representation, this study proposes a medical knowledge graph construction method integrating multi-source heterogeneous data. By combining semi-structured data from authoritative medical websites with open-source medical knowledge graphs, a multi-stage construction framework is designed, and an innovative dynamic disease-symptom weight calculation method based on joint search frequency is introduced. The resulting knowledge graph incorporates disease-symptom weight information, resolving ambiguities in symptom-disease relationships while achieving effective integration and in-depth mining of fragmented medical data. This work provides valuable data resources for medical research and clinical decision support.

【Keywords】Information redundancy; Disease-symptom weights; Medical knowledge graph

1 引言

尽管科技的进步使人们可以方便地从互联网上获取医疗信息,但是数据量的爆发式增长使得用户很难从大量信息中找到自己所需要的信息,这导致了信息获取效率低下的问题。数据来源的多样性也增加了信息真实性的难度,因为不同来源的信息可能存在矛盾,而且部分来源的可信度也难以判断。此外,医疗信息的复杂性也给用户带来了困扰,对于普通人来说,许多医学术语和知识都是陌生的,很难理解医学领域的信息,

这进一步降低了用户获取准确医疗信息的能力。因此,需要构建一个医疗领域的知识图谱对互联网中的碎片化医疗信息进行整合,为智能问诊系统提供可靠的数据支撑。

然而,当前医疗知识图谱的研究主要集中在某个特定的疾病领域,开源数量有限且大多以 API 接口的形式向公众开放。此外,在医疗知识图谱的构建过程中,还存在着疾病症状权重关系不明确的问题。为了解决上述问题,本文以专业权威的医疗网站信息和开源新

第一作者简介:王宁(1995-)女,浙江嘉兴,助教,硕士,研究方向:知识图谱、自然语言处理;

*通讯作者:傅新杰(1990-)男,浙江嘉兴,讲师,硕士,研究方向:自然语言处理。

冠知识图谱中的数据作为数据来源, 通过数据采集和处理、知识抽取、知识融合、知识存储、疾病症状权重信息采集与存储的流程, 构建高质量的医疗知识图谱, 并将其作为智能问诊系统的知识库, 为用户提供可靠的医疗咨询服务, 为医疗领域的相关研究提供有价值的资源。

2 相关工作

作为知识图谱应用研究的热点之一, 医学领域涌现出了许多优秀的研究成果。国外将知识图谱的构建技术应用于生物医学领域, 实现对遗传信息的深入理解和利用, 从而提高个性化治疗效果和药物研发效率。Yu^[1]等通过从 miRNA 相关疾病数据库中提取三元组信息来构建知识图谱, 然后在此基础上训练深度图表示学习模型, 完成 miRNA 与疾病关系预测任务。Ye^[2]等提出一种融合知识图谱和推荐系统的 KGE_NFM 框架, 可以更加精准地预测药物靶点间的相互作用, 为药物研发和临床治疗提供帮助。国内注重于从具体的医疗应用场景出发, 构建以疾病为中心的知识图谱。Li^[3]等使用信息提取技术从生物医学文献、医学指南和临床试验等非结构化文本中提取实体信息, 经知识过滤与知识融合后, 最终构建出肝细胞癌知识图谱。Yin^[4]等通过对多个大型三甲医院的乙型肝炎电子病历中的数据

识图谱。刘勇^[5]等以糖尿病防治指南为核心, 采用多种自然语言处理技术, 构建了一套完整的糖尿病知识图谱。刘金奎^[6]等构建了冠心病知识图谱, 为中医的标准化、规范化诊疗提供了一种新的参考范式。

综上, 当前中文医疗知识图谱研究多聚焦单病种(如肝癌、冠心病), 与多病种综合应用场景匹配度低, 且开源资源匮乏(多为 API 形式开放)。此外, 现有医疗知识图谱普遍缺乏疾病-症状的权重量化, 导致问答系统无法精准排序关联疾病。

因此需基于半结构化医疗网站数据, 构建覆盖多病种且嵌入症状权重关系的知识图谱, 以弥补现有不足并满足临床辅助决策需求。

3 医疗知识图谱的构建

本文在构建垂直领域的医疗知识图谱时采用自底向上的构建方法, 构建流程如图 1 所示。首先通过分析页面结构来对爬虫程序进行设计与实现, 将采集到的数据进行过滤、补充、去重等数据清洗操作以及格式化操作得到格式化数据, 对处理后的数据进行知识抽取以得到实体、属性和关系, 通过知识合并与实体对齐技术将其与新冠图谱中的数据进行融合并将其保存到图数据库中, 最后, 采集计算疾病症状的权重信息并写入知识图谱的疾病症状关系边中, 为医疗领域提供必要的

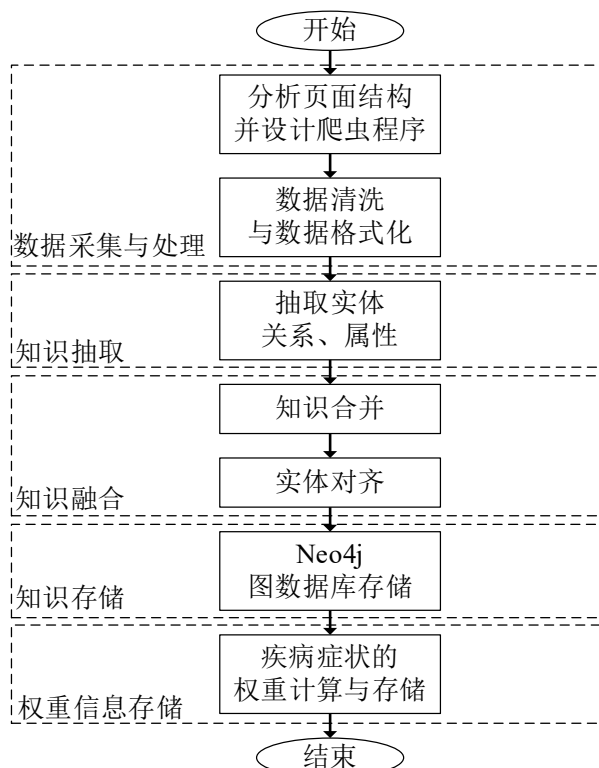


图 1 医疗知识图谱构建流程图

3.1 数据采集与处理

3.1.1 数据来源

本文使用爬虫方式获取医疗健康网站中的半结构化数据,并将其作为主要数据来源,此外,为了提高医疗知识图谱的实用性,以开源新冠知识图谱中的结构化的数据作为补充。在众多医疗健康网站中,经过仔细调研,最终选择了专业权威的医疗健康网站为爬取对象,该网站中的疾病条目由专业医生编写审核,因此其可信度高,专业性强。

3.1.2 数据爬取

为了实现信息的自动化爬取,需要遵循一定的流程对网页信息进行获取和解析。数据爬取的基本流程如图2所示,包括三个步骤:获取疾病URL队列、确定页面爬取元素、编写爬虫脚本。

首先,通过分析疾病百科页的URL规律,利用

Python的requests库循环遍历页码参数,结合parsel库的XPath解析技术提取每个百科页内嵌的疾病详情页链接,构建完整的疾病URL队列。其次,针对疾病详情页,通过开发者工具分析页面结构,确定需要抓取的13个核心字段(如疾病名称、挂号科室等),并提取对应HTML元素的XPath路径。最后,编写自动化爬虫程序,采用多级抓取策略:先批量获取所有疾病入口链接,再并发请求详情页并执行结构化解析,最后将提取的标准化疾病数据存储。

3.1.3 数据处理

经过数据爬取以后,从医疗网站上获得了大量疾病相关的数据,然而,这些数据中存在着一些混乱、无效的脏数据,会对医疗知识图谱的构建产生影响。因此,需要对采集的数据进行清洗和格式化,以得到格式化数据,具体流程如图3所示。

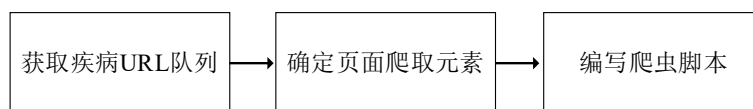


图2 数据爬取流程

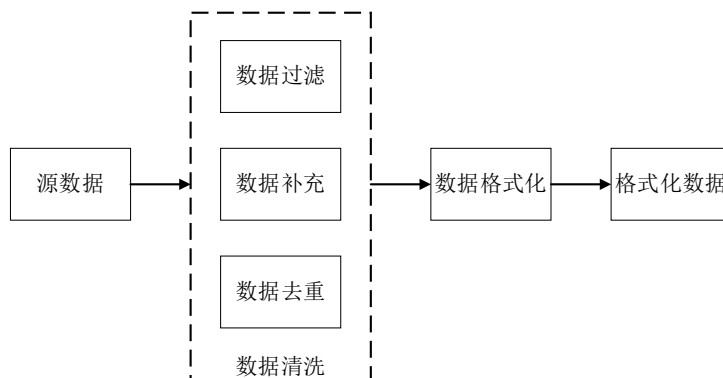


图3 数据处理流程图

经过上述步骤后最终确定了13种疾病标签,分别为:疾病名称、疾病简介、挂号科室、治疗方法、治疗周期、治愈率、多发人群、典型症状、临床检查、并发症、常用药品、病因、预防。

3.2 知识抽取

在完成数据采集与清洗后,将半结构化的疾病、药物数据转化为知识图谱三元组。依据JSON标签特征划分两类数据:8种疾病属性(如病因、预防等非结构化文本)解析为<实体,属性,值>形式,5种实体关系(如症状、药品等列表数据)转化为<实体1,关系,实体2>结构。通过遍历数据条目,将属性存储为字典、关系存储为列表,统一转换为CSV格式的三元组,并

提取5类实体(科室、症状等)进行去重存储为txt文件。

3.3 知识融合

为了提高医疗知识图谱的完整性,有必要将与新冠肺炎有关的知识融合到知识图谱中。知识融合分为实体合并和实体对齐两个过程。

3.3.1 实体合并

基于清华大学新冠知识图谱(含疾病、症状、药物等结构化JSON数据),通过解析其Schema定义与实体信息生成CSV文件,利用Python的csv库将其与原医疗网站数据合并,形成统一知识库。

3.3.2 实体对齐

实体对齐的目的是实现对不同来源知识的整合, 构建出更为全面的医疗知识图谱。在实体相似度计算中, 通常采用字符串相似度和语义相似度这两种不同的方法并通过加权平均的方式综合考虑两者的影响, 以获得更为准确的计算结果, 相似度计算公式如式(1)所示:

$$Sim(e_1, e_2) = \beta Sim_{cos}(e_1, e_2) + (1 - \beta) Sim_{dice}(e_1, e_2), \quad (1)$$

其中, $Sim_{cos}(e_1, e_2)$ 代表实体 e_1 和 e_2 之间的语义相似度, 这里采用的余弦相似度来度量。 $Sim_{dice}(e_1, e_2)$ 代表的是实体 e_1 和 e_2 之间的字符串相似度, 这里采用的是系数来度量。 $Dice$ 代表加权权重, 通过少量别名数据集验证可得其值约为 0.7。

在实体相似度计算后, 设定阈值 0.9 作为对齐判断标准: 若候选实体集中所有实体对的相似度均低于阈值, 则保留原实体; 若存在高于阈值的实体对, 则选取相似度最高的候选实体进行对齐操作, 将新冠知识图谱中的实体名称统一替换为医疗网站中的标准化命名。

3.4 知识存储

经过前面小节的处理, 已经得到了结构化的疾病知识, 接下来, 需要将数据存储到数据库中, 构建较为完整的医疗知识图谱。由于本文构建的知识图谱包含

的数据量比较大, 故将 CSV 格式数据文件通过 load csv 命令进行一键导入将数据快速导入到 Neo4j 数据库中。

3.5 疾病症状权重的计算与存储

在医疗知识图谱构建中, 需量化症状与疾病关联权重以提升查询准确性。例如“流鼻涕”可能关联鼻炎、荨麻疹等疾病, 但鼻炎关联性更高。通过计算症状-疾病权重 (如基于临床频率或医生经验), 将权重值嵌入知识图谱关系边。本文通过必应搜索引擎中<症状, 疾病>对出现的频率来计算两者之间的权重, 计算公式如式(2)所示:

$$Weight < s, d > = \frac{count(< s, d >)}{\sum_{d_i \in D_s} count(< s, d_i >)} \quad (2)$$

其中, $count(< s, d >)$ 代表症状 s 和疾病 d 在必应搜索引擎中进行联合搜索所匹配到的网页结果数量, 同理, $count(< s, d_i >)$ 为症状 s 和疾病 d_i 在匹配到的网页结果数量, D_s 为与症状 s 相关的所有疾病组成的集合。

4 结果分析

本文使用 Neo4j 存储的医疗知识图谱共包含了 7439 个实体, 39622 对实体关系。已经构建好的医疗知识图谱的部分可视化效果如图 4 中所示, 图中, 不同颜色表示不同类型实体的节点, 比如紫色表示疾病实体节点, 棕色表示症状实体节点。

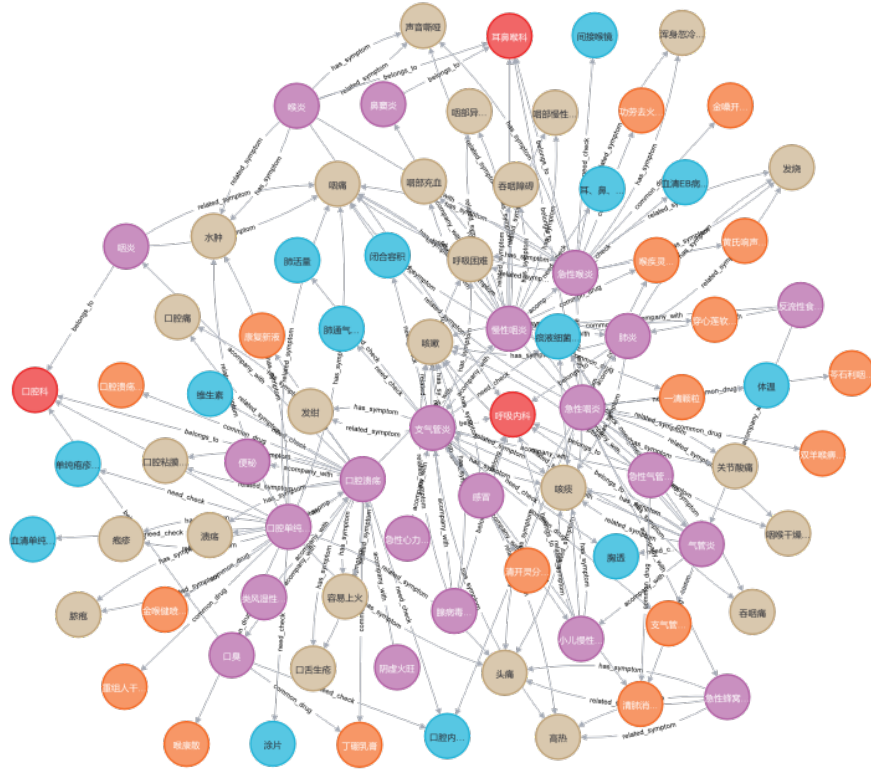


图 4 医疗知识图谱的可视化图

5 结论

本研究针对医疗知识图谱开源性不足及疾病-症状关联权重缺失问题, 构建了一套完整的医疗知识图谱框架。流程分为五阶段: 首先通过爬虫采集权威医疗网站数据并清洗结构化; 其次抽取实体、属性与关系形成三元组; 随后融合新冠知识图谱数据, 通过实体合并与对齐整合异构信息; 接着将结构化知识存储至 Neo4j 图数据库; 最终基于必应搜索的疾病-症状共现网页数量计算关联权重并嵌入图谱边属性。该方法通过自底向上的构建策略与动态权重机制, 既提升了图谱的覆盖度与标准化程度, 又强化了症状-疾病关联的临床指导价值, 为智能诊疗提供精准可靠的知识支撑。

参考文献

- [1] YU S, WANG H, LIU T, et al. A knowledge-driven network for fine-grained relationship detection between miRNA and disease [J]. 2022, 23(3): bbac058.
- [2] YE Q, HSIEH C-Y, YANG Z, et al. A unified drug-target interaction prediction framework based on knowledge graph and recommendation system [J]. 2021, 12(1): 6775.
- [3] LI N, YANG Z, LUO L, et al. KGHC: a knowledge graph for hepatocellular carcinoma [J]. 2020, 20: 1-11.
- [4] YIN Y, ZHANG L, WANG Y, et al. Question answering system based on knowledge graph in traditional Chinese medicine diagnosis and treatment of viral hepatitis B [J]. 2022, 2022(1): 7139904.
- [5] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: Representation, acquisition, and applications [J]. 2021, 33(2): 494-514.
- [6] HUANG L, GUAN H, LIANG Y, et al. Covid-19 knowledge graph for drug and vaccine development; proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), F, 2021 [C]. IEEE.
- [7] 刘勇, 齐梦霁. 基于糖尿病防治的医学知识图谱构建的研究 [J]. 医学信息, 2020, 33(18): 11-4.
- [8] 刘金垒, 惠小珊, 张振鹏, et al. 基于中医诊疗指南的冠心病知识图谱构建 [J]. 中国实验方剂学杂志, 2023, 29(07): 208-15.

版权声明: ©2025 作者与开放获取期刊研究中心 (OAJRC) 所有。本文章按照知识共享署名许可条款发表。

<https://creativecommons.org/licenses/by/4.0/>



OPEN ACCESS