

面向文献组织的重大公共卫生事件知识图谱构建研究^{*}

李晓瑛 蔡妙芝 李军莲 任慧玲 冀玉静 邓盼盼 夏光辉

(中国医学科学院/北京协和医学院医学信息研究所/图书馆 北京 100005)

〔摘要〕 提出一种融合中西医知识的新新型冠状病毒肺炎知识模型及基于自然语言处理的领域知识自动发现方法,从等级体系结构、主题分类、关系类型等多维度对语义知识进行组织与关联。收集新冠肺炎专题文献、诊疗方案及网络新媒体数据,在自动知识发现基础上进行人工审核和标注,构建新冠肺炎知识图谱并实现可视化呈现与在线检索,以期支撑生物医学科技文献资源深度组织和有效利用。

〔关键词〕 新型冠状病毒肺炎;知识图谱;知识发现;知识组织;自然语言处理

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.03.002

Study on the Building of Knowledge Graph of Major Public Health Events for Literature Organization LI Xiaoying, CAI Miaozhi, LI Junlian, REN Huiling, JI Yujing, DENG Panpan, XIA Guanghui, Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China

〔Abstract〕 A COVID-19 knowledge model integrating traditional Chinese and western medicine knowledge and an automatic domain knowledge discovery method based on Natural Language Processing (NLP) are proposed, and the semantic knowledge is organized and associated from the hierarchical structure, subject classification, relation types and other dimensions. Additionally, the COVID-19 literature, diagnosis and treatment regimens and the new media data are collected for knowledge discovery and manual annotation, and the COVID-19 knowledge graph is built and realized for visual presentation and online retrieval, so as to support the well organization and effective utilization of biomedical literature resources.

〔Keywords〕 COVID-19; knowledge graph; knowledge discovery; knowledge organization; Natural Language Processing (NLP)

〔修回日期〕 2020-12-07

〔作者简介〕 李晓瑛,博士,副研究员,发表论文 40 余篇;通讯作者:李军莲,研究馆员。

〔基金项目〕 国家社科基金项目“大数据时代的中西医知识组织与知识图谱构建研究”(项目编号:20BTQ062);中国医学科学院医学信息研究所新冠肺炎疫情专项“面向文献组织的新冠肺炎知识图谱构建研究”(项目编号:2020XG023);科技创新 2030-“新一代人工智能”重大项目课题“中文医学术语体系构建”(项目编号:2019AAA0104901);新闻出版业科技与标准重点实验室《医学融合出版知识技术实验室》。

1 引言

1.1 研究背景

随着新型冠状病毒肺炎（以下简称新冠肺炎）疫情全球化发展，新冠肺炎临床诊疗、发病机制、疫苗研发、疫情防控等相关科技文献迅速增长。如何对这些杂乱无序的文献大数据进行深度组织和语义知识关联，帮助科研人员精准地发现和获取所需资源和数据，推动我国新冠肺炎疫情尽早结束，为全球新冠肺炎疫情提供中国方案，具有重要意义。知识图谱是大数据时代下的一种新型知识组织系统，通常以语义三元组形式进行知识组织和表示，为人工智能等应用开发提供机器可理解的领域知识。在生物医学领域，北京大学计算语言学研究所开发的“中文医学知识图谱”^[1]，为我国智能咨询、智能诊断、智慧医疗等应用提供医学知识基础。

1.2 相关研究

疫情发生以来，国内外卫生保健机构积极开展新冠肺炎相关术语命名规范与发布。国家卫健委与世界卫生组织同步发布新病毒及其所引发疾病的中英文名称，即 2019 新型冠状病毒（2019 - nCoV）、新型冠状病毒肺炎（COVID - 19）。同期各类信息媒介中随处可见“飞沫传播”、“气溶胶”、“康复者血浆治疗”、“清肺排毒汤”等高频词。中国翻译研究院按疫病名称、传染防控等类别搜集梳理，翻译审定新冠肺炎疫情相关词汇中英文表达；随后中国医学科学院医学信息研究所在国家人口与健康科学数据共享服务平台发布“新型冠状病毒肺炎术语集”^[2]，数据包括疾病、病毒、症状体征等相关术语中英文对照、同义关系、释义和来源；美国国立医学图书馆将 2019 - nCoV、COVID - 19 作为独立主题词纳入 2021 版《医学主题词表》（Medical Subject Headings, MeSH）并建立等级体系及与其他主题词、副主题词的相互关联；武汉科技大学计算机学院、浙江大学计算机学院、哈尔滨工业大学社会计算与信息检索研究中心、清华大学、南京大学、IBM 中国研究院等机构及团队，在中文开放知识图谱 OpenKG 提交 13 个与新冠肺炎相关知识图

谱，内容涉及论文集、专利、百科、科研、事件、健康、概念、流行病、临床、英雄、防控、物资以及图谱之间的映射等多个方面，从多维度支撑新冠肺炎相关智能系统应用开发。

1.3 研究目的

本研究立足于长期的中文医学知识组织系统和文献服务系统建设实践，在《中文医学主题词》（Chinese Medical Subject Headings, CMeSH）等医学术语资源基础上，从中西文科技文献、诊疗方案等多渠道收集、发现新冠肺炎相关主题及其语义关系，从等级体系结构、主题分类、语义知识类型等多维度对相关主题与知识进行深度组织与语义关联，在此基础上完成知识图谱构建及其可视化展示，支撑生物医学科技文献主题标引、语义标引及主题分类，以及生物医学文献信息服务系统的主题、语义、分类检索。

2 研究方法

2.1 新冠肺炎知识模型构建

在充分借鉴疾病本体（Disease Ontology, DO）结构基础上^[3]，从现代医学、传统医学两种学科体系对新冠肺炎相关知识进行揭示。具体而言，前者涉及新冠肺炎病原学特点（病因、疾病属种）、流行病学特点（传染源、传播途径）、诊断（病理改变、临床表现、实验室检查、胸部影像学）、临床分型、并发症、鉴别诊断、治疗及预防相关知识，而后者主要涵盖中医病因、范畴、病症、治法治则等信息；两者互为补充，从两种不同哲学思维方式呈现当前对新冠肺炎的认识、临床诊断及防治。新冠肺炎知识模型，见图 1。

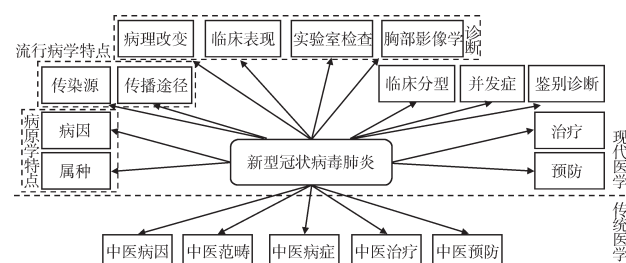


图 1 新冠肺炎知识模型

2.2 新冠肺炎知识发现

2.2.1 概述 人工智能时代，人与机器均可理解的领域知识通常以主谓宾语三元组（Subject - Predication - Object, SPO）形式表示，其中主语及宾语为领域概念（术语），谓语为语义关系类型。新冠肺炎知识发现首先从公开发表的科技文献、诊疗方案等多渠道收集相关概念及术语（即主语或宾语），再依据概念所在上下文的动、名词谓语，发现病因、传播、临床表现、诊断、治疗等语义关系。此外为提高知识发现可靠性及全面性，邀请专业人员对计算机自动发现结果进行审核，采用人工标注方式及时补充重要语义知识。

2.2.2 新概念及新术语自动发现与补充 CMeSH 是目前生物医学领域规模最大的中文主题词表，广泛用于中文医学文献、书目等资源的主题标引与检索、主题分类与导航等应用中。本研究中 CMeSH 作为基础词表，与科技文献中的作者关键词进行词形比较，词形不同的关键词称为候选词，进一步通过语义相似度比较，判断其为新概念（新主题）或新术语，见图 2，相应补充到 CMeSH 中。就语义相似度算法而言，本研究前期已开展相关探索，提出一种针对 CMeSH 词表的生物医学领域新术语语义映射方法^[4]，公式为：

$$SemSim(D, K) = Max(Sim(T_i, k)) \quad (1)$$

其中： D 表示 CMeSH 中某个已有主题词， T_i 代表该主题词的第 i 条术语（ $i = 1, 2, \dots, N$ ， N 为包括主题词及入口词在内的术语总数）， k 为来自文献的作者关键词， $Sim(T_i, k)$ 表示术语 T_i 与 k 的字面相似度； $SemSim(D, k)$ 表示主题词 D 与关键词 k 的语义相似度，取自 D 的所有术语中最大值。关于字面相似度，经对比经典的 Dice^[5] 与 Jaccard^[5] 系数法后选择前者，公式为：

$$Sim(T_i, k) = 1 \times \frac{len(T_i \cap k)}{len(T_i) + len(k)} \quad (2)$$

其中：公式右端的分子为术语 T_i 与关键词 k 的共同

子串长度，分母为 T_i 与 k 的词长之和。Dice 字面相似度的取值在 0 ~ 1 之间，待比较的两个短语相同部分越多，则取值越高。这种方法能够忽略字顺差异带来的影响（如“先天性遗传性疾病”与“遗传性先天性疾病”），在实际应用中效果良好。

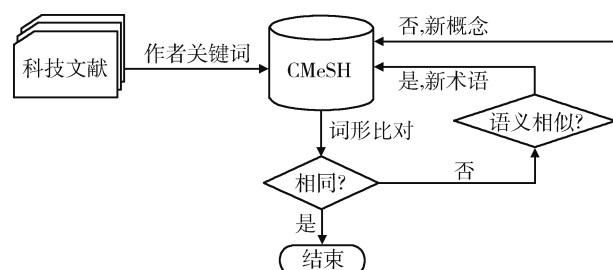


图2 新术语自动发现流程

2.2.3 语义关系自动发现 本研究所采用的语义关系自动发现方法综合应用分句、词典分词、特征（模式）匹配等自然语言处理技术，以及语义类型约束等优化方法。具体而言，首先对于科技文献、诊疗方案等资源的题目、摘要等自然语言文本（段），以句号、问号、感叹号、省略号、换行符等作为分隔符进行分句预处理；其次，对切分后的每条文本句，将 CMeSH 词表（概念及入口词）作为领域专用的切词词典，采用双向最大匹配法进行词典分词；再次，对于同时出现至少两个 CMeSH 概念的文本句进一步应用预先抽取的特征进行模式匹配（以“药物治疗关系”为例，特征词包括“治疗”、“药物疗法”、“阻止”、“干预”等）；一旦匹配成功，则利用 CMeSH 词表中概念的等级结构（树号）作为语义类型，对候选的语义关系进行约束和优化，例如“药物治疗关系”存在于“化学物质和药物”（树号以“D”开头）与“疾病”（树号以“C”开头）之间，主语及宾语为其他语义类型的关系三元组则需要滤除。最后，输出高可靠性的语义关系三元组及其所在的文本句供人工审核。语义关系自动发现流程，见图 3。

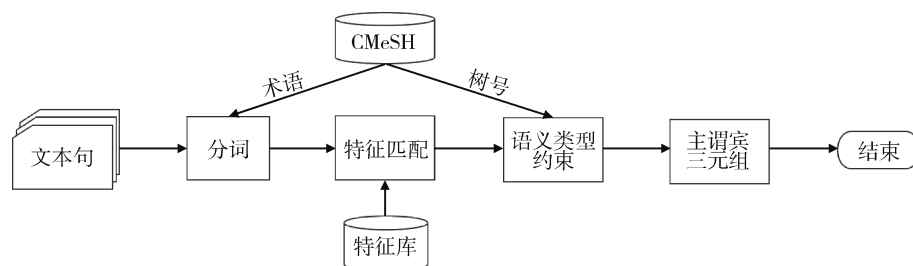


图 3 语义关系自动发现流程

2.2.4 人工审核与标注 为提高新冠肺炎知识发现的准确性，由专业人员对计算机自动发现结果进行审核确认。审核原则主要归纳为两点，为最大程度方便审核及时调整算法。一是作者关键词应根据其含义映射到最确切的主题词。如“一次性口罩”，按公式（1）、（2）计算出语义相似度最大的主题词为“一次性设备”（相似度为 0.6），但依据词义内涵人工将其调整到主题词“面罩”（与入口词“口罩”的相似度为 0.57）。基于此调整新术语自动发现算法，输出计算机找出的语义相似度最大的前 3 个主题词，供人工审核时对比参考。鉴于新冠肺炎相关文献数量与日俱增，在一段时间内审核完所有已公开发表文献中的作者关键词极具挑战，本研究根据这些关键词在文献中出现频次（以文献篇数为单位）从高到低进行排序筛选，优先审核高频词，及时将审核结果应用到语义关系发现中。二是语义关系需结合其所在的文本句判断主语、谓语及宾语是否准确有效。例如从文献摘要“…古代医家对类似新冠肺炎的时令发热性疾病积累了丰富的临床经验，故按照方证相应的原则提出本次疫情的经方对策，以小柴胡汤为基本方变化运用…”，可自动发现语义关系三元组“经方 - 治疗 - 新冠肺炎”及“小柴胡汤 - 治疗 - 新冠肺炎”，而“经方”指以《伤寒杂病论》中的方剂为主的古方，仅用于总论，具体方药需用专指词，因此审核时认为“经方 - 治疗 - 新冠肺炎”的语义关系过于笼统，无实用价值，予以删除。此外本研究对知识发现算法未揭示的诊疗方案中的语义关系进行人工标注，将语义关

系三元组中主语或宾语涉及的新概念或新术语依据其含义补充到 CMeSH 词表中。例如对国家卫健委发布的《新型冠状病毒肺炎诊疗方案（试行第八版）》标注后发现“新冠肺炎 - 并发 - 儿童多系统炎症综合征”的语义知识，将“儿童多系统炎症综合征”识别为新概念。

2.3 新冠肺炎知识深度组织

2.3.1 概述 对于计算机自动发现、人工审核标注的新冠肺炎知识进行有序组织，有助于挖掘知识之间的关联，为用户提供精准有效的知识服务。本研究对于源自科技文献、诊疗方案等多来源的语义知识重点从两个维度进行深度组织。

2.3.2 依据 CMeSH 结构对新概念进行深度组织 CMeSH 词表收录主题词中英文名称及入口词、树状结构号、标引注释、范畴注释、可组副主题词、唯一编号（MeSH ID）、《医学专业分类表》中的分类名与分类号、树状等级结构等信息。以“新型冠状病毒”为例，对发现的新主题概念按照 CMeSH 词表的内容结构进行组织，见表 1。不仅有助于开展医学文献等信息资源主题标引与主题分类，而且支持按等级结构从广度上扩展知识主题（如同位主题“SARS 病毒”、“中东呼吸综合征冠状病毒”，上位主题“乙型冠状病毒属”等）。另外鉴于 CMeSH 主题词的颗粒度较粗，少数具体细分概念则参照《系统化医学术语表 - 临床术语》（Systematized Nomenclature of Medicine - Clinical Terms, SNOMED CT）内容结构进行组织。

表 1 基于 CMeSH 词表内容结构的新冠肺炎知识深度组织（以“新型冠状病毒”为例）

组织维度	具体信息
中文主题词	新型冠状病毒
英文主题词	SARS – CoV – 2
入口词	新型冠状病毒（2019 – nCov）；2019 新型冠状病毒（2019 novel coronavirus）；COVID – 19 病毒（COVID – 19 virus）；严重急性呼吸综合征冠状病毒 2 型（severe acute respiratory syndrome coronavirus 2）；SARS 病毒 2 型（SARS2）；新冠病毒（COVID19 virus）
树状结构号	B04. 820. 578. 500. 540. 150. 113. 968
标引注释	infection = COVID – 19
范畴注释	新型冠状病毒属于 β 属的冠状病毒，有包膜，颗粒呈圆形或椭圆形，常为多形性，直径 60 – 140nm。——《新型冠状病毒肺炎诊疗方案（试行第八版）》
副主题词	化学（CH）； 分类（CL）； 药物作用（DE）； 酶学（EN）； 生长和发育（GD）； 遗传学（GE）； 免疫学（IM）； 分离和提纯（IP）； 代谢（ME）； 生理学（PH）； 致病力（PY）； 辐射效应（RE）； 超微结构（UL）
MeSH ID	D000086402
分类名	R373. 1：呼吸道传染性病毒
树状结构	病毒 RNA 病毒 巢状病毒属 冠状病毒科 冠状病毒属 乙型冠状病毒属 乙型冠状病毒 1 型 冠状病毒，大鼠 中东呼吸综合征冠状病毒 鼠肝炎病毒 SARS 病毒 新型冠状病毒

2.3.3 从现代医学、传统医学分别对语义关系进行组织 新冠肺炎知识语义关联，见图 4。其中语义关系类型与图 1 中的知识模型基本保持一致，此处不再赘述。

2.4 新冠肺炎知识图谱可视化系统设计

语义关系三元组是知识图谱中最基本的知识表示与知识存储单元，为知识图谱可视化提供便利。一般而言，数据可视化技术往往将三元组中的主语与宾语直观地表示为计算机图形中不同节点，而谓语则形象化为节点之间的连边。然而这种方式仅适

用于数据量较少的知识图谱可视化；当同一概念的不同关系类型语义三元组较多时，则由于连边过多导致图形发散、可读性极大降低。因此设计新冠肺炎知识图谱可视化系统时，同一概念的不同关系类型语义知识将进行捆绑展示，从而减少图形发散性，提高知识图谱易读性。此外为帮助用户深入理解概念内涵与外延，在可视化系统中显示概念在 CMeSH 词表中的详细信息，包括中英文主题词名称及入口词（用于支持语义知识查询检索）、范畴注释、分类名、语义类型（等级结构族首词）。

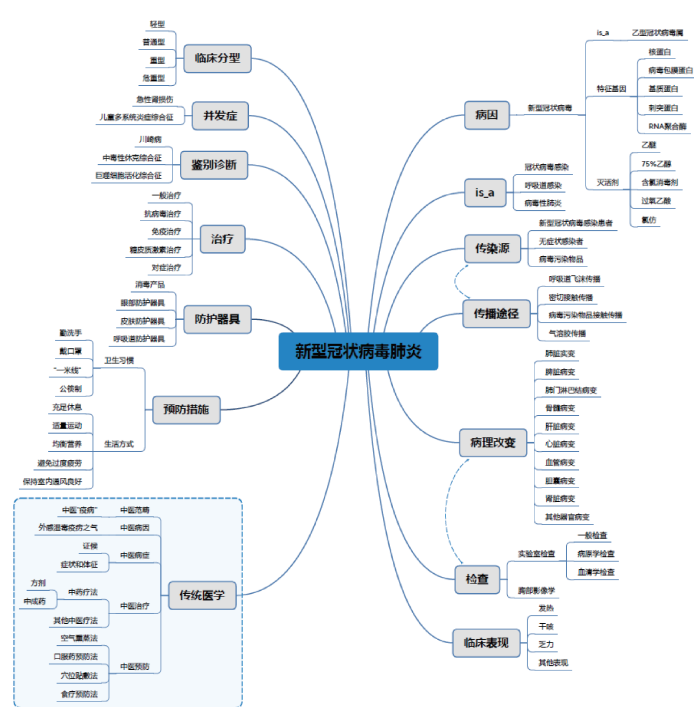


图 4 新冠肺炎知识语义关联

应用。

3 研究成果

3.1 知识自动发现与审核系统

SpringBoot 是 Java 平台上一种开源的轻量级框架，具有简单易用、快速便捷、可移植性高、良好支持数据库应用服务等优势；MyBatis 为一款优秀的持久层框架，支持多种数据库简便配置、存储和映射。本研究搭建基于 SpringBoot 和 MyBatis 的知识发现与审核系统，通过连接 MySQL 数据库，支持从科技文献、诊疗方案等多种信息资源发现医学知识。系统分为两个核心功能模块，可单独使用或联合应用。以基于科技文献数据的知识发现为例，用户首先上传文献题目数据（含文献 ID、题名、摘要和作者关键词等），在“主题发现”模块，系统将按照新词发现算法从作者关键词发现新概念或新术语，自动推荐补充到 CMeSH 词表，专业人员审核确认后，新词发现结果将增加到 CMeSH 中，然后通过“语义关系发现”模块，发现疾病病因、治疗药物等语义关系，最后系统将输出人工审核及标注后的语义关系三元组，供知识图谱构建等进一步处理与

3.2 新冠肺炎知识图谱

本研究对 SinoMed 中国生物医学文献服务系统收录的 7 000 余篇新冠肺炎相关核心期刊文献以及《新型冠状病毒肺炎诊疗方案（试行第 8 版）》《医疗机构内新型冠状病毒感染预防与控制技术指南（第 1 版）》《新型冠状病毒感染的肺炎防控中常见医用防护用品使用范围指引（试行）》《北京市新型冠状病毒肺炎中医药防治方案（试行第 5 版）》《海南省新型冠状病毒感染的肺炎中医药防治方案（公众版试行第 2 版）》、《陕西省新型冠状病毒感染的肺炎中医药预防方案》等国家和地方诊疗方案，采用计算机自动推荐与人工审核标注相结合的方式，发现新冠肺炎相关语义知识。此外使用主流的动态图形可视化技术 D3.js，开发新冠肺炎知识图谱可视化系统，支持对图谱的浏览、检索、详细信息展示。具体而言，图谱以新型冠状病毒肺炎为中心，依据新冠肺炎知识组织维度，逐级展开与关联。目前初步构建的知识图谱内置语义关系三元组 2 000 余条（包括中英文同义关系），后续将进一步补充完善。

（下转第 19 页）

警、防控等提供数据全景支撑。该系统为新冠肺炎及类似传染病疫情防控提供了新思路。

参考文献

- 1 中国新闻网. 习近平主持中央政治局常委会会议 研究疫情防控工作 党中央成立应对疫情工作领导小组 [EB/OL]. [2020 - 01 - 25]. <https://www.chinanews.com/gn/2020/01-25/9069630.shtml>.
- 2 焦玲艳, 王颖, 韩啸宇. 2013 - 2018 年天津市滨海新区流感监测分析 [J]. 现代预防医学, 2019, 46 (19): 3612 - 3615.
- 3 李多, 徐闻, 伏晓庆, 等. 2017 年云南省乙型流感病毒监测结果及其血凝素基因特征分析 [J]. 病毒学报, 2019, 35 (4): 606 - 611.
- 4 范家诚, 姚永明, 郑敏, 等. 2011 - 2016 年贵州省少数民族艾滋病疫情流行特征分析 [J]. 中国艾滋病性病, 2020, 26 (1): 40 - 42.
- 5 张桂欣, 许玉成, 周志峰, 等. 2008 - 2018 年深圳市福田区流行性感冒聚集性疫情流行特征分析 [J]. 现代预防医学, 2019, 46 (19): 3471 - 3474, 3486.
- 6 祝姗姗, 高洁. 基于动态网络生物标志物的甲流疫情早期预警算法研究 [J]. 病毒学报, 2018, 34 (1): 9 - 15.

- 7 王思嘉, 聂绍发. 湖北省潜江市手足口病疫情不同预警模型效果比较 [J]. 中国公共卫生, 2017, 33 (12): 1760 - 1763.
- 8 魏明珠, 张海涛, 刘雅姝, 等. 多维属性融合的社交媒体高影响力人物画像研究 [J]. 图书情报知识, 2019 (5): 73 - 79, 100.
- 9 张杨斌, 谢辉, 毛进, 等. 面向城市数据画像构建的多源数据需求与融合方法研究 [J]. 情报理论与实践, 2020, 43 (6): 88 - 96.
- 10 韩梅花, 赵景秀. 基于“用户画像”的阅读疗法模式研究——以抑郁症为例 [J]. 大学图书馆学报, 2017, 35 (6): 105 - 110.
- 11 杨晓宇, 王秀民, 李黎. 以患者画像识别技术为核心的 CRM 系统实践精准医疗 [C]. 珠海: 2016 中华医院信息网络大会, 2016: 1 - 4.
- 12 陈旭, 郑浩毅. 基于手机定位数据的个体出行行为特征分析综述 [J]. 综合运输, 2019, 41 (9): 38 - 44.
- 13 程文炜, 闫晓芳, 史静琚, 等. 基于 GIS 的中国中老年人糖尿病空间分布及其影响因素分析 [J]. 中华疾病控制杂志, 2017, 21 (11): 1082 - 1087.
- 14 查文婷, 王玲, 郑剑, 等. 基于 GIS 技术分析长沙市 2006 - 2013 年手足口病流行病学特征分析 [J]. 中华疾病控制杂志, 2017, 21 (4): 345 - 348.

(上接第 13 页)

4 结语

领域知识图谱是支撑智能应用的重要基础。本研究利用自然语言处理技术及计算机辅助人工构建的方式, 从新冠肺炎相关科技文献、诊疗方案等多来源发现语义知识, 从等级体系、主题分类、关系类型等多维度进行深度组织, 探索出一种专题知识图谱构建方法, 初步建成的新冠肺炎知识图谱提供新冠肺炎相关概念等级分类、主题分类及病因、传播、诊疗、防护等多种语义关联, 可用于支撑生物医学文献主题标引、语义标引及主题分类, 以及文献信息服务系统的主题、语义、分类检索等应用。今后的工作重点包括新知识发现及补充完善、与 OpenKG 中新冠肺炎开放图谱的知识融合、基于所建图谱的知识推理等方面。

参考文献

- 1 奥德玛, 杨云飞, 穗志方, 等. 中文医学知识图谱 CMeKG 构建初探 [J]. 中文信息学报, 2019, 33 (10): 1 - 9.
- 2 中国医学科学院医学信息研究所. 新型冠状病毒肺炎术语集 [EB/OL]. [2020 - 11 - 20]. <https://www.ncmi.cn/phda/dataDetails.do?id=CSTR:A0006.11.A0003.202002.000590-V1.0>.
- 3 Schriml L M, Arze C, Nadendla S, et al. Disease Ontology: a backbone for disease semantic integration [J]. Nucleic Acids Res, 2012 (40): D940 - D946.
- 4 孙海霞, 李军莲, 李丹亚, 等. 基于 CMeSH 语义系统的领域自由词 - 主题词语义映射研究 [J]. 现代图书情报技术, 2013 (11): 46 - 51.
- 5 Pavel Shvaiko, Jerome Euzenat. A Survey of Schema - based Matching Approaches [J]. Journal on Data Semantics IV, 2005 (4): 146 - 171.