基于医学社交媒体数据的多模态知识图谱构建*

王华琼 俞定国 钱归平

（浙江传媒学院媒体工程学院 杭州 310018） （1 之江实验室 杭州 311121
2 浙江传媒学院媒体工程学院 杭州 310018）

〔摘要〕 介绍医学领域传统知识图谱研究现状，分析医学社交媒体数据发展情况，详细阐述基于医学社交媒体数据的多模态知识图谱构建方法，提出多模态知识图谱构建成为必然趋势，多模态知识融合有助于进一步提高医学专家系统的智能性。

〔关键词〕 知识图谱；多模态；湿疹；本体构建；医学数据分析

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.04.006

Construction of Multi－modal Knowledge Graph Based on Medical Social Media Data *WANG Huaqiong, YU Dingguo, College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China; QIAN Guiping, 1Zhijiang Lab, Hangzhou 311121, 2College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China*

〔**Abstract**〕 The paper introduces the research status of traditional knowledge graph in the medical field, analyzes the development situation of medical social media data, elaborates the construction method of multi－modal knowledge graph based on medical social media data, and puts forward that the construction of multi－modal knowledge graph is an inevitable trend, and multi－modal knowledge fusion is helpful to further improve the intelligence of medical expert system.

〔**Keywords**〕 knowledge graph; multi－modal; eczema; ontology construction; medical data analysis

1 引言

随着医生和患者在社交媒体上活跃度的提升，网络上积累了大量医学社交数据，并以富文本形式产生、积累和显示，包含结构化数据、文本、表情、图片、音频、视频等多种不同模态信息[1]。类似于人类通过视觉、听觉、嗅觉、触觉等多种感官感知外界环境和实体，这些多模态信息通过不同的结构特征和表现形式对对象进行描述，进而形成更加完整而准确的表示和评价[2－3]。

在传统知识图谱构建过程中，往往根据特定研究任务选择一种占主导地位的模态数据进行分析[4]。Zhang Y 等[5]构建语义临床决策支持知识库，集成医疗保健本体知识和患者数据，但仍然以存储于医院信息系统中的结构化数据为主。面向社交媒

〔修回日期〕 2022－10－16

〔作者简介〕 王华琼，讲师，发表论文 17 篇；通信作者：钱归平，副教授，硕士生导师，发表论文 20 余篇。

〔基金项目〕 浙江省基础公益研究计划项目“个性化临床路径关键技术研究及应用”（项目编号：LGF20H180001）；浙江省教育厅科研项目“融合多模态知识图谱的新闻短视频自动生成技术研究”（项目编号：Y202146864）。

体中的多模态数据，如何解决模态异构性并建立跨模态关联是构建多模态知识图谱的核心问题[6]。单模态知识图谱中已经包含医学领域的诊断、药物、手术、临床路径等术语和规则，为多模态知识图谱建立奠定了医学本体基础。进一步将结构化数据、文本、表情、图片、音频、视频等模态信息进行归类，归纳为领域知识、文本知识和视觉知识 3 类。如何建立这 3 类模态知识与现有医学本体之间的映射关系是本文拟解决的核心问题。

根据医学社交媒体数据的多模态特征，本文基于现有医学本体进行补充和扩展，提出一种面向多模态信息的知识图谱构建方法。首先对医学社交媒体数据中的多模态知识进行分析和提取，归纳并提炼出 3 类模态信息；然后建立现有医学本体与这 3 类模态知识之间的语义关联，提高知识图谱的完整性和智能性。

2 医学社交媒体数据分析

医学社交媒体数据与医院信息系统中的结构化数据相比，内容更加丰富，形式更加多样，其中包含图片视觉知识、纯文本知识，以及医院、科室、医师级别等领域知识，见图 1。对普通用户而言，图文并茂能够帮助其更加直观、完整地理解文章内容；对知识图谱而言，视觉知识、文本知识的融入是对现有医学本体的重要补充，关联融合多个模态数据能够提升模型在分类或回归任务中的性能[7]。

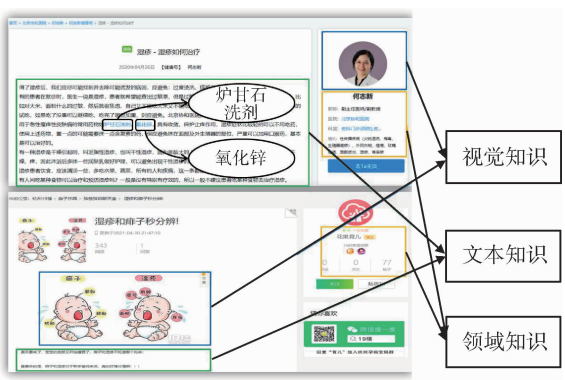

图 1 医学社交媒体数据中的多模态知识

3 医学领域知识图谱构建

3.1 总体思路

研究团队在构建传统知识图谱方面已有多年经验，在面向院内数据的单模态知识图谱构建方面提出“知识源确定 - 知识抽取 - 知识表达 - 模型评估”4 阶段法[4]。在此基础上，通过对医学社交媒体数据内容和特征的分析，归纳和总结多模态医学知识图谱构建方法[8]，提出基于医学社交媒体数据的多模态知识图谱构建 4 阶段法。多模态知识图谱的构建、评估，所采用的技术和方法与单模态知识图谱基本一致，最大区别在于多模态知识与现有本体知识之间的映射，即多模态知识融合。

3.2 构建方法

3.2.1 知识源确定 是指识别和确定医学社交媒体知识源。明确各类模态信息的数据来源对构建多模态图谱是非常重要的。以湿疹病种为例，从网络开放数据中获取“湿疹”相关的网页数据，将结构化数据存储到 MySQL 数据库中，并将图片等视觉知识载体以文件方式进行保存。

3.2.2 知识抽取 是在信息抽取的基础之上更加深入发现隐含知识的过程，为多模态医学知识图谱的构建提供原始素材和内容。传统知识图谱搭建过程中知识抽取更多针对领域专家的隐含知识，抽取过程主要依赖人工实现。基于社交媒体数据的结构和模态特征，多模态知识抽取强调对文本知识和视觉知识的抽取，需要将本体技术与深度学习技术、自然语言处理技术相结合，关注不同模态知识间关系的识别和抽取。

3.2.3 知识表达 是利用本体技术将多模态知识进行语义化表达。医学领域术语使用网络本体语言进行描述，明确定义湿疹病种相关知识的类、属性和实例。在 Protégé 中通过 OntoGraf 绘制类、实例关系。构建湿疹临床路径知识图谱，列出“临床路径信息库”“病种”“激素类药物”3 个类以及对应的 7 个实例，并通过属性建立实例“湿疹临床路径”和常见外用药物实例之间的关联，见图 2。知识图谱中的

规则使用语义网规则语言描述，支持智能推理。

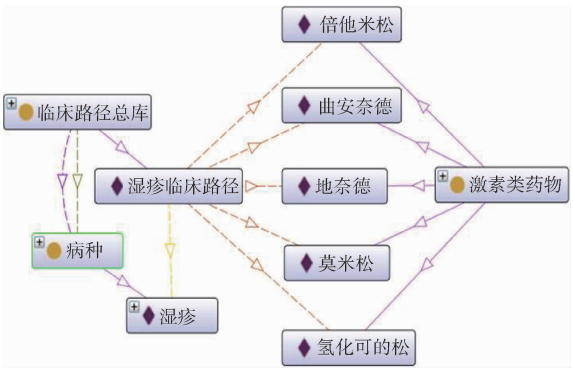

图 2 湿疹临床路径知识图谱

3.2.4 模型评估 该阶段采取面向任务的模型评估方法。研究结果中以单一病种“湿疹”为例，采用上述建模方法构建面向湿疹病种的多模态知识模型，并搭建多模态数据浏览展示界面，结合领域专家意见对知识模型进行评估与优化。

4 多模态知识融合

4.1 领域知识提取

医学社交媒体数据中包含部分领域知识，例如医生姓名、所属科室、所属医院等，还包括行业认证、用户评价等信息，在一定程度上反映其发布内容的权威性和认可度。领域知识在医学社交媒体数据中通常以类似于结构化数据的形式表达，通过对爬取网页的正则表达式解析，快速、便捷地提炼出该领域知识，利用 Jena 语义框架写入到现有知识图谱语义数据模型中。网页中的医生“何 * 青”经过语义转换变成语义模型中“门诊医师”类的一个实例，其所在医院、科室等信息对应到语义数据模型中的属性值。

4.2 文本知识分析

医学社交媒体数据中包含大量文本信息，除了医生等医务人员发布的文章，还包含大量评论和回复信息。以 19 楼网站为例，通过主题网站爬虫获取湿疹相关开放数据，合计包含文章 1 130 篇，评论和回复 7 990 条。

文本分析主要包含词频统计和情感分析两个方面。词频统计主要统计文章和评论文本中关键词出现频次[9]。将湿疹常见外用药物作为关键词进行词频分析，反映出该药物在湿疹病种治疗方案中的受关注程度。通常，人们在社交媒体上描述药物时会有不同的表达方式，例如“糠酸莫米松”“艾洛松”“莫米松”等。药物实例通过定义其 rdfs：label 属性和 rdfs：comment 属性，设置药物的中文名称、主要成分名称和别名。药物不同表达方式通过 rdfs：label 属性关联到同一药物实例。假定，对病种 D 存在 n 种外用激素类药物可供选择，每种激素类药物在数据中可能存在 m 种不同表达方式，每种表达方式在社交媒体数据中出现的频次为：$X_{i,j}(i = 1,2,\ldots,m;j = 1,2,\ldots,n)$，那么，某种激素类药物在文章中的词频（word frequency，WF）计算式表示为：

$$WF_j = \sum_{i=1}^{m} x_{i,j} \tag{1}$$

词频统计仅能反映药物在社交媒体数据中的受关注程度，却无法体现发布者对该药物的态度。因此，进一步使用情感极性分析方法挖掘文本知识，提炼各种激素类药物的受推荐程度。情感极性分析主要有两种方式：基于情感字典的分析[10]和基于机器学习的分析[11]。由于语料数据量限制，采用情感字典分析方法[12]。

将每种激素类药物的每种表达方式作为一个关键词，提取关键词所在上下文进行情感极性分析。通过正负情感字典分别统计该关键词的正向情感得分 $Np_j(j = 1,2,\ldots,n)$ 和负向情感得分 $Nn_j(j = 1,2,\ldots,n)$。通过程度级别字典统计程度等级，假定上下文中共有 k 个程度级别词，每个词的程度级别为 $D_{l,j}(l = 1,2,\ldots,k;j = 1,2,\ldots,n)$。综合正负向情感得分和程度级别，关键词情感得分的计算式表示为：

$$Y_{i,j} = 1 + (Np_j - Nn_j) \times (1 + \sum_{k=1}^{l} D_{k,j}) \tag{2}$$

为了避免出现关键词所在上下文的程度等级为零的情况，在上下文的程度等级上加 1。同样地，关键词的情感得分也加 1。对具有 m 种表达方式的某种激素类药物，其情感得分计算式为：

$$S_j = \sum_{i=1}^{m} Y_{i,j} \quad (3)$$

综上，使用词频统计公式（1）计算每种激素类药物在社交数据中的出现频次，使用情感评分公式（3）计算各药物的情感评分，得到词频统计和情感评分结果，见图 3。

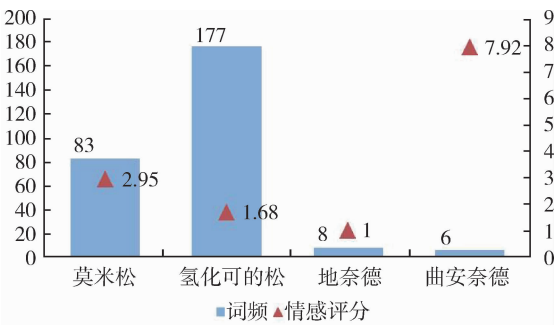

图 3 基于文本的关键词词频和情感分析

从图 3 可以看出，情感分析结果和词频统计结果并没有呈现正相关，“氢化可的松”词频远高于“莫米松”，情感评分却低于“莫米松”。在社交媒体数据中，受关注程度高表明存在较高的讨论度或较大争议，也可能出现负面评价，并不能完全代表该药的使用率，与受推荐程度无直接关系。因此，两个指标不存在正比关系。在药物评价时，体现药物推荐程度的情感评分指标具有更为重要的参考价值；但单一的情感评分指标又是不充分的，需要词频指标进行样本量支持。综合考虑两个指标，有助于为家庭护理场景中的药物选择提供辅助决策支持。将图 3 所得结果写入知识图谱是对现有医学本体的重要补充。文本知识极大地丰富和完善了知识图谱，为后续药物推荐系统或者辅助诊疗专家系统的研发提供智能性基础。

4.3 视觉知识融合

社交媒体数据中的图片通常包含网址、标题以及简单的文本描述信息。首先将该部分信息作为知识图谱中的视觉模态信息保存，建立图片与已有实体之间的关联。例如，湿疹通过属性 hasImage 建立与图片之间的关联。属性关联只是一种初步关联，不同模态信息交互性较低。为了提高视觉信息对知识图谱智能性的促进作用，本文通过细粒度视觉实体抽取，形成面向实体的视觉模型，辅助分类诊断。通过对选取的图像实体进行不同的视觉细节特征计算实现图像实体的多样化选择。图片特征提取过程采用多种现有深度学习方法来进行特征提取融合和领域泛化，例如 Koohbanani N A 等[13]提出的自监督多任务学习网格框架 Self－Path、Meng Q 等[14]提出的基于互信息的解纠缠神经网络 MIDNet 等，通过在现有模型基础上调整参数实现。然后通过语义技术将提取的特性信息关联到知识图谱的实体之中。通过特征提取明确图片中红疹的部位，并建立与已有实体之间的细粒度关联。

5 讨论

5.1 多模态知识图谱构建方法

以湿疹知识图谱构建过程为例，描述从医疗社交媒体数据中提取的领域知识、文本知识和视觉知识对现有医学本体的补充过程，通过建立现有医学本体与这 3 类模态知识之间的语义关联，提高知识图谱的完整性和智能性。尽管本研究以湿疹单一病种为例，但是所提出的多模态知识图谱构建方法对其他疾病同样适用，3 类模态信息的提取和融合过程在技术方法上是一致的。

5.2 多模态知识图谱构建方法适用范围

本文提出的知识图谱构建方法，与已有知识图谱构建方法相比，在数据来源和技术方法两方面存在明显差异。本文提出的知识图谱构建方法以医学社交媒体数据作为来源，从数据来源上摆脱了对医疗机构的依赖性，具备开放性特征，并提取领域知识、文本知识和视觉知识 3 种不同模态信息，综合运用文本分析、视觉实体识别等技术对现有医学本体进行补充。从医学社交媒体数据的开放性特征出发，基于本知识图谱的应用，不受限于医疗机构，有利于开放给公众、服务于家庭医疗，因此所提出的方法对依赖家庭护理的慢性疾病管理具有更重要的推广价值。社交媒体数据在具备开放性优势的同时存在表达不准确、知识矛盾冲突等局限性，从网络数据中挖掘知识，

最终将知识保存到图谱中，而网络数据本身并不保存到知识图谱，降低错误样本对整体疾病知识库的影响。由于医学社交媒体数据依赖网络表达，患者对专业治疗手段并不能清晰、准确地进行描述，因此本方法对依赖医疗机构处理的急性或者复杂病种并不适合。

5.3 研究创新性

在医学领域“多模态”并不属于新名词，对于多模态医学影像研究非常广泛，侧重图像识别与融合。但是多模态医学数据融合属新兴研究领域，近两年出现结合医学信息系统和医学影像的模态融合研究，多采用深度学习算法，应用于实体识别、图文转换等。本文应用现有多模态融合深度学习算法，侧重多模态知识图谱构建，目标是提供一个多模态医学知识图谱构建的技术框架，多模态融合深度学习算法这一技术点上直接采用现有前沿算法，因此并没有展开篇幅详细介绍。

6 结语

基于医学社交媒体数据的特征和内容分析，提出医学多模态知识图谱的 4 阶段构建方法。研究结果以“湿疹”病种为例展开介绍，综合运用本体构建、词频统计、情感极性分析、视觉实体识别等技术，描述现有本体构建过程，重点阐述文本知识和视觉知识的抽取方法以及其对传统知识图谱的扩展和补充作用。伴随社交媒体数据的日益积累和标准化，多模态知识图谱构建成为必然趋势，多种模态知识融合有助于进一步提高医学专家系统的智能性。

参考文献

1 ALPER B S. Usefulness of online medical information [J]. American family physician, 2006, 74 (3): 482 – 488.

2 WANG M, WANG H, QI G, et al. Richpedia: a large – scale, comprehensive multi – modal knowledge graph [J]. Big data research, 2020, 22 (10): 100159.

3 ATREY P K, HOSSAIN M A, SADDIK A E, et al. Multimodal fusion for multimedia analysis: a survey [J]. Multimedia systems, 2010, 16 (6): 345 – 379.

4 WANG H Q, LI J S, ZHANG Y F, et al. Creating personalised clinical pathways by semantic interoperability with electronic health records [J]. Artificial intelligence in medicine, 2013, 58 (2): 81 – 89.

5 ZHANG Y, TIAN Y, ZHOU T, et al. Integrating HL7 RIM and ontology for unified knowledge and data representation in clinical decision support systems [J]. Computer methods and programs in biomedicine, 2016, 123 (1): 94 – 108.

6 王树徽, 闫旭, 黄庆明. 跨媒体分析与推理技术研究综述 [J]. 计算机科学, 2021, 48 (3): 79 – 86.

7 GAO J, LI P, CHEN Z, et al. A survey on deep learning for multimodal data fusion [J]. Neural computation, 2020, 32 (5): 829 – 864.

8 BERNASCONI A, MASSEROLI M. Biological and medical ontologies: systems biology ontology (SBO) [J]. Encyclopedia of bioinformatics and computational biology, 2019, 1 (1): 858 – 866.

9 AIZAWA A. An information – theoretic perspective of tf – idf measures [J]. Information processing & management, 2003, 39 (1): 45 – 65.

10 PEI J. A dictionary – based maximum match algorithm via statistical information for Chinese word segmentation [J]. International journal of electronics and information engineering, 2020, 12 (1): 24 – 33.

11 THIRUTHUVANATHAN M M, KRISHNAN B. Multimodal emotional analysis through hierarchical video summarization and face tracking [J]. Multimedia tools and applications, 2021, 4 (3): 1 – 20.

12 WANG H, QIAN G. Guideline – driven medical decision support methods for family healthcare [J]. IEEE access, 2021, 9 (7): 116612 – 116621.

13 KOOHBANANI N A, UNNIKRISHNAN B, KHURRAM S A, et al. Self – path: self – supervision for classification of pathology images with limited annotations [J]. IEEE transactions on medical imaging, 2021, 40 (10): 2845 – 2856.

14 MENG Q, MATTHEW J, ZIMMER V A, et al. Mutual information – based disentangled neural networks for classifying unseen categories in different domains: application to fetal ultrasound imaging [J]. IEEE transactions on medical imaging, 2020, 40 (2): 722 – 734.