基于 EMR 的乳腺肿瘤知识图谱构建研究

崔 洁[1,2] 陈德华[3*] 乐嘉锦[3]

[1](东华大学旭日工商管理学院 上海 200051)

[2](上海交通大学医学院附属瑞金医院计算机中心 上海 200025)

[3](东华大学计算机科学与技术学院 上海 200051)

摘 要 知识图谱作为一种描述实体及其联系的新方法,在医学领域也逐渐得到关注,出现了多种医学知识图谱。但是这些医学知识图谱的知识大多来源于公开的医学文献,较少涉及到 EMR 电子病历。EMR 电子病历涵盖了医院各科室各病种的患者诊疗全过程数据,具有丰富的医疗事实知识,是医学知识图谱的重要知识来源。为此,以乳腺肿瘤这一具体病种为应用实例,结合知识图谱技术的基本原理,给出了乳腺肿瘤知识图谱的定义;结合上海交通大学医学院附属瑞金医院的实际 EMR 电子病历数据集,通过知识抽取技术从 EMR 中提取乳腺肿瘤医疗事实知识。在此基础上提出乳腺肿瘤知识图谱的构建方法。

关键词 EMR 乳腺肿瘤 知识图谱 信息转化

中图分类号 TP3 **文献标识码** A DOI: 10.3969/j.issn.1000-386x.2017.12.023

STUDY ON THE CONSTRUCTION OF KNOWLEDGE GRAPH OF BREAST TUMOR BASED ON EMR

Cui Jie[1,2] Chen Dehua[3*] Le Jiajin[3]

[1](*Glorious Sun School of Business and Management, Donghua University, Shanghai 200051, China*)

[2](*Computer Centre, Ruijin Hospital Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China*)

[3](*Computer Science and Technology, Donghua University, Shanghai 200051, China*)

Abstract As a new method to describe entities and their relationships, knowledge graph has been paid more and more attention in the medical field. However, most of the knowledge of the medical knowledge graph is derived from the open medical literature, and less related to the EMR electronic medical records. EMR electronic medical records cover the whole process of patient diagnosis and treatment with a wealth of medical facts, which is an important source of knowledge of medical knowledge graph. Therefore, this paper takes the specific disease of breast tumor as an example. According to the basic principle of knowledge graph technology, we firstly gave the definition of knowledge of breast tumors. Combined with the actual EMR electronic medical records data set of Ruijin Hospital Affiliated to Shanghai Jiaotong University School of Medicine, the knowledge of breast cancer medical facts was extracted from EMR by means of knowledge extraction technology. On this basis, a method for constructing knowledge map of breast tumors is proposed.

Keywords EMR Breast tumor Knowledge graph Information transformation

0 引 言

EMR 电子病历记录了医院各科室患者在诊疗过程中所产生的各种数据,包括患者基本信息、诊断数据、检验数据、检查数据、用药数据、出院小结等。这些数据反映了医院内部真实发生的各种医疗事实,例如“某患者经超声检查被诊断为乳腺癌 IV 期”则包含了

收稿日期:2017-01-25。上海市科委科研计划项目(15511106902)。崔洁,高工,主研领域:医院信息化,信息管理。陈德华,副教授。乐嘉锦,教授。

两种类型的医疗事实,即该患者的检查事实和诊断事实。可见,EMR 经过记录数据间的关联,可形成各种医疗事实知识。这种医疗事实知识表现为各种医学实体如患者实体、基本信息实体、就诊实体、检查实体、诊断实体,以及各种实体之间的关系如患者实体与检查实体之间存在检查关系。EMR 电子病历有着丰富的医学事实知识,是医学知识的重要组成部分。

知识图谱(Knowledge Graph)作为一种新的知识表示方法,属于语义网范畴[1],其基本原理是借助图模型来刻画和描述现实世界中存在的各种实体或概念,建立这些实体或概念之间的关联关系,表达相关领域实体或概念之间的语义关系。目前,业界已提出许多通用的知识图谱,以谷歌公司的搜索知识图谱最为典型 Google Knowledge Graph[2]。与此同时,由于知识图谱具有知识语义化、数据易关联、易扩充等特性[3],国内医疗信息学领域也开始逐渐开展医学知识图谱的构建工作,也提出了多种医学知识图谱,包括中国中医科学院中医药信息研究所基于已有的中医药学语言系统构建的中医药知识图谱[4],基于知识图谱的基因组流行病学可视化分析[5]和生物医学信息可视化分析[6]。纵观这些医学知识图谱,其知识来源主要是公开的医学文献,但是较少涉及到 EMR 电子病历的医疗事实知识。利用知识图谱来描述 EMR 中的医疗事实知识,可以更好地刻画 EMR 电子病历数据中存在的实体和属性分类,并通过实体间的关系揭示临床数据间的内在联系,从不同层次的形式化模式上给出这些实体和实体间相互医疗事实关系的明确定义,从而避免来自不同数据源的信息的语义异构。

乳腺肿瘤是女性主要恶性肿瘤之一,其发病率和死亡率不断上升,对女性的健康造成严重危害[7],乳腺肿瘤患者基数较大,国内大中型医院均已积累了大量的乳腺肿瘤 EMR 记录,其中包含着大量关于乳腺肿瘤诊治的医疗事实知识。因此,本文以乳腺肿瘤为具体病种应用实例,借鉴知识抽取的技术思想[8],提出一种基于 EMR 的乳腺肿瘤知识图谱构建方法。该方法分别从概念层和实例层两个层次对乳腺肿瘤知识图谱进行设计,支持乳腺肿瘤医疗实体及关系的抽取,实现从乳腺肿瘤 EMR 数据向医疗事实知识的转化。具体而言,本文的乳腺肿瘤知识图谱构建方法由两个阶段组成: 第一阶段即乳腺肿瘤知识图谱概念层设计阶段,主要实现乳腺肿瘤 EMR 中各种医学实体的抽取,并提取出各种实体之间的关系。第二阶段即乳腺肿瘤知识图谱实例层设计阶段,主要实现由乳腺肿瘤 EMR 记录向知识图谱的转化,完成乳腺肿瘤知识图谱的自动构建。

1　相关工作

1.1　通用知识图谱

由于中文知识图谱的构建对中文信息处理和检索具有重要的研究和应用价值,近年来吸引了大量的研究[8]。例如在业界出现了百度知心、搜狗知立方等商业应用。在学术界,清华大学建成了第一个大规模中英文跨语言知识图谱 KLore、中国科学院计算技术研究所基于开放知识网络(OpenKN)建立了“人立方、事立方、知立方”原型系统、中国科学院数学与系统科学研究院陆汝钤院士提出知件(Knowware)的概念、上海交通大学构建并发布了中文知识图谱研究平台 zhishi.me、复旦大学 GDM 实验室推出的中文知识图谱项目,等等[9]。这些项目具有较大规模的知识库,覆盖广泛的知识领域,能够为用户提供一定的智能搜索及问答服务。

1.2　医学知识图谱

近些年来,国内对医学信息学领域知识库的研究也逐渐活跃。医学知识库(NKIMed)[10]是中科院计算机研究院 1995 年所研发的用于检索和挖掘医学信息的本体知识库,包括了多达 52 个医学概念分类,1 691 种医学属性,19 595 个知识概念,共计录入 78 013 条知识。医学知识库是国家基础知识设施(National Knowledge Infrastructure)的一个分集合,对医学知识的分析和推理具有重要作用。

如中国中医科学院中医药信息研究所基于已有的中医药学语言系统构建的中医药知识图谱,哈工大信息检索研究中心(HIT CIR)在文本智能化检索领域进行了深入研究,主要包括文本过滤、篇章理解和知识分析等,其研究成果已应用于文本智能化检索、机器翻译、自动分类、自动文摘等系统。除了这些综合类的比较全面的医学领域知识系统外,国内的研究还有些专门针对具体某种疾病或者某一具体领域的知识体系。比如专门用于诊断肾脏疾病的 PIP(Present Illness Program),PIP 采用框架语义网结构,框架涵盖生理状态、临床表现、典型的病症等,它主要使用匹配技术来进行诊断并给出相应的治疗方案[11]。

但是现有的各种对医学信息学领域知识库的研究大多是基于互联网上公开的医学文献,以及各种开放数据库和电子资源,这类知识虽然获取比较方便。由于知识来源比较局限,如何利用真实医学数据来构建知识图谱,获取更准确、更全面、更权威的知识成为医学知识图谱领域的研究需求。

2 乳腺肿瘤知识图谱相关概念

本文专注于乳腺肿瘤这一特定病种的知识图谱构建研究。下面给出乳腺肿瘤知识图谱相关概念的形式化定义。

定义 1 (乳腺肿瘤医学实体 E):乳腺肿瘤医学实体 E 指的是乳腺肿瘤 EMR 记录中各种可唯一标识的医学实体。

一般在医院 EMR 中,乳腺肿瘤医学实体包括了乳腺肿瘤患者实体、基本信息实体、乳腺肿瘤诊断实体、乳腺肿瘤检查实体、乳腺肿瘤检验实体等。

定义 2 (乳腺肿瘤医学事实关系 R):乳腺肿瘤医学事实关系表示不同乳腺肿瘤医学实体之间所发生的医疗事实联系即 $R\{E_i, E_j\}$,其中 E_i、E_j 为乳腺肿瘤医学实体。

结合乳腺肿瘤 EMR 记录,在医学领域专家的帮助下,一共整理出以下几种乳腺肿瘤医学事实关系类型,具体包括了:

(1) has_a 关系:表示实体 A 和实体 B 之间的隶属关系。

(2) instance_of 关系:表示实体 A 与实体 B 间的实例关系。换言之,实体 B 是实体 A 的一个实例。

(3) attribute_of 关系:表示实体 A 是实体 B 的属性值。

(4) part_of 关系:表示整体与部分的关系,例如,检查报告中的特征描述实体 A 是检查报告实体 B 的一部分。

(5) owns 关系:表示病人实体 A 拥有检查报告实体 B 或者病理报告实体 C。

(6) diagnosis 关系:表示诊断结论实体 A 与患者实体 B 之间是诊断关系。

(7) detect 关系:表示仪器实体 A 与患者实体 B 是检测关系。

在定义了上述乳腺肿瘤医学实体和医学事实关系的基础上,乳腺肿瘤知识图谱的形式化定义如下。

定义 3 (乳腺肿瘤知识图谱 G):乳腺肿瘤知识图谱为一张有向标签图 $G = (E, R, T)$,其中 E 为知识图谱的顶点集,用于表示乳腺肿瘤医学实体集合;R 为知识图谱的边集,用于表示乳腺肿瘤医学事实关系;T 为 $EXE \rightarrow R$ 的函数,表示了知识图谱中的所有元组。

表 1 为一位乳腺肿瘤患者的具体 EMR 记录。从中可见,该 EMR 记录中蕴含着患者实体,基本信息实体(其下包含了性别实体、年龄实体和地区实体),检查实体(其下包含了超声检查实体、CT 检查实体、MRI 检查实体和病理检查实体)和诊断实体(包含了超声诊断实体、CT 诊断实体、MRI 诊断实体和病理诊断实体)。这些实体间具有不同的关系,例如患者实体与基本信息实体之间存在 has_a 关系,患者实体与检查实体之间存在 Detect 关系,检查实体与诊断实体之间存在 Diagnosis 关系。乳腺肿瘤医学实体及其关系表示了乳腺肿瘤知识图谱的模式结构,类似于关系数据库的概念模式。图 1 所示为乳腺肿瘤知识图谱模式结构图即概念层结构。

表 1 乳腺肿瘤患者 A 的 EMR 记录

实体		实体	实体属性值
基本信息		性别	女
		年龄	45
		地区	上海
检查	超声检查	位置	右
		方位	11 – 12 点钟
		大小	10 mm × 7 mm
		形态	类圆形
		表面	欠光整
		境界	较清
		内部回声	低回声
		分布	不均
		回声	散在粗大斑状强回声
		团块后回声	明显声衰减
		血流信号	较丰富
		血管	较粗大
		走形	扭曲
	病理检查	大小	3 cm × 2 cm × 0.8 cm
		形状	不规则型
		颜色	乳白色
		质地	偏硬
诊断	超声诊断	诊断结论	BI-RADS V
	病理诊断	诊断结论	恶性

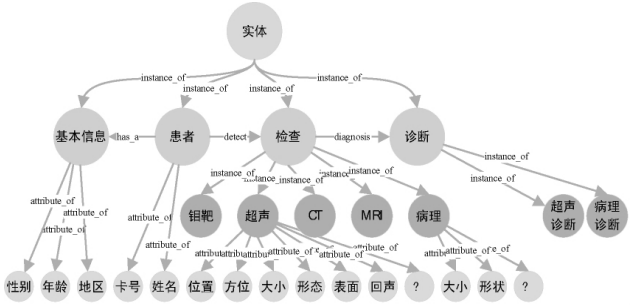

图 1 乳腺肿瘤知识图谱概念层结构

基于上述的乳腺肿瘤知识图谱概念层结构,建立 EMR 记录各项值与概念层实体及关系之间的对应关系,构建 < 主语,谓语,宾语 > 三元组,完成乳腺肿瘤知识图谱实例层。以 EMR 的乳腺肿瘤患者基本信息为

例,患者基本信息表中的列名“姓名”可以转化成 RDF 数据中的谓词,表中对应的取值为 RDF 宾语,如 ID 为“102413148”的患者姓名可以用三元组 < 102413148,姓名,张三 > 表示。图 2 所示为乳腺肿瘤知识图谱实例层结构。

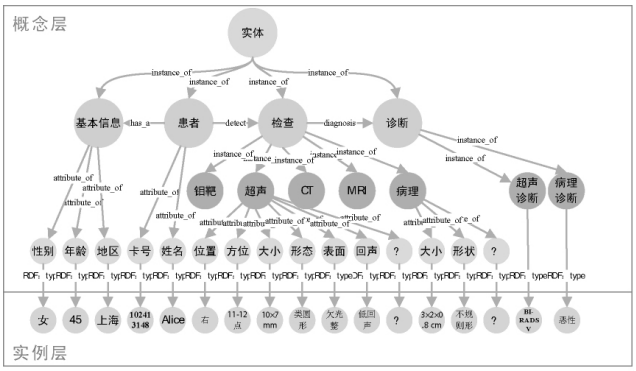

图 2　乳腺肿瘤知识图谱实例层

3　乳腺肿瘤知识图谱构建方法

如 2 节所述,EMR 记录蕴含着丰富的医疗事实知识,是医学知识图谱的重要数据来源。本节提出一种基于 EMR 的乳腺肿瘤知识图谱构建方法,该方法由概念层设计和实例层设计两个阶段组成。下面分别给出两个阶段的具体流程。

3.1　概念层设计

乳腺肿瘤知识图谱概念层设计的主要任务是在领域专家的帮助下,根据领域知识创建乳腺肿瘤知识图谱的概念模式结构。

目前,乳腺肿瘤知识图谱概念模式结构是以上海交通大学医学院附属瑞金医院的实际 EMR 记录结构为基础,结合美国国家综合癌症网络(NCCN)的乳腺癌临床指南[12],构建了乳腺肿瘤知识图谱的概念层。其中,以患者实体为中心,在同一层次与基本信息实体、检查实体和诊断实体之间存在不同的联系。具体而言,患者实体与基本信息实体之间存在 Has_a 关系,患者实体与检查实体之间存在 Detect 关系,患者实体经检查实体与诊断实体之间存在 Diagnosis 关系。患者实体有医疗卡号和姓名两个属性,而基本信息实体的属性则有性别、年龄和地区等。由于乳腺肿瘤检查有不同检查手段,所以检查实体包含了超声检查实体、CT 检查实体、钼靶检查实体和病理检查实体等子层次实体。检查实体与这些子层次实体之间存在 Instance_of 关系。不同子层次检查实体还具有不同的属性,例如超声检查实体具有位置、方位、大小、形态、表面、回声分布、血流信号等属性。对应于不同的检查手段,诊断实体也包含了超声诊断实体、CT 诊断实体、钼靶诊断实体和病理诊断实体等子层次实体。

3.2　实例层设计

实例层设计属于知识抽取范畴,其主要任务是从无语义信息的 EMR 记录中抽取与概念层相匹配的医疗事实知识。实际的 EMR 记录既有结构化数据如患者基本信息、就诊信息、处方信息、检验信息等,也有半结构化数据如出院小结,还有非结构化内容如超声文本报告等。实例层设计的目标就是从不同格式的 EMR 记录内容中提取乳腺肿瘤医学实体及关系,并表示为主谓宾三元组形式。

1) 乳腺肿瘤医学实体提取

乳腺肿瘤医学实体提取是构建乳腺肿瘤知识图谱的首要步骤,目的在于从 EMR 记录中找到用于表示乳腺肿瘤医学实体或属性的相关术语或标记集合。其中,EMR 记录中的结构化和半结构化数据由于具有较好的模式结构,实体提取的规则相对容易制定; 而对于非结构化文本数据由于格式较为自由,在实体提取规则上需要借助自然语言处理技术对 EMR 文本内容进行结构化处理。

下面结合瑞金医院实际的 EMR 记录,阐述如何实现上述概念层各种实体的具体操作步骤。

(1) 患者实体提取: 从 EMR 记录的患者 ID 和姓名两个字段,提取每位乳腺肿瘤患者的 ID 和姓名字段值作为患者实体的属性值。

患者基本信息实体提取: 从 EMR 记录的患者性别、年龄和地区三个字段,提取每位乳腺肿瘤患者的性别、年龄和地区字段值作为患者基本信息实体的属性值。

(2) 检查实体提取: 每个患者根据不同的病情需要,进行不同类型的检查,从检查实体中,提取出钼靶、超声、CT、MRI、病理等不同检查类型,作为检查实体的子类实体。

(3) 检查实体属性值提取: 由于检查报告为文本格式,本文首先采用作者提出的临床文档结构化处理方法[13]对各种检查文本报告进行结构化处理,提取文本报告中的指标和指标值,以提取出来的指标和指标值作为检查实体属性值。

(4) 诊断实体提取: 患者所做的每项检查均有对应的诊断结论,从诊断实体中,提取出钼靶诊断、超声诊断、CT 诊断、MRI 诊断、病理诊断等不同检查的诊断结论,作为诊断实体的子类实体。

2) 实体间的关系类型

如前所述,乳腺肿瘤患者的 EMR 记录经抽取后形成五类医学实体。这些实体可与概念层的概念属性相关联,作为这些属性的属性值。

结合概念层的概念间关系,可知患者实体与患者基本信息实体之间的关系为 has_a 关系; 患者实体与检查实体之间的关系为 detect 关系; 检查实体与诊断实体之间的关系为 diagnosis 关系。患者的姓名实体与患者之间的关系为 instance_of 关系; 患者的性别、年龄、地区等实体与基本信息实体之间的关系为 instance_of 关系; 超声检查、钼靶检查、CT 检查、MRI 检查、病理检查实体与检查实体之间的关系为 instance_of 关系; 超声诊断、钼靶诊断、CT 诊断、MRI 诊断、病理诊断结果实体与诊断实体之间的关系为 instance_of 关系。将该患者的患者实体和基本信息实体之间为 has-a 关系。患者的姓名实体与患者实体之间的关系为 instance_of 关系; 基本信息实体与性别、年龄和地区实体之间为 instance_of 关系。患者实体和检查实体之间为 detect 关系。检查实体与超声检查和病理检查实体之间为 instance_of 关系。患者实体和诊断实体之间为 diagnosis 关系。诊断实体与超声诊断和病理诊断实体之间为 instance_of 关系。

在提取出实例层的实体及关系之后,即可将 EMR 的乳腺肿瘤数据转换成 RDF 形式的链接数据 D2R (Relational Database to RDF)[14]。乳腺肿瘤知识图谱中主谓宾三要素关系如表 2 所示。

表 2 乳腺肿瘤患者 A 构建知识图谱的主谓宾三要素

主	谓	宾
患者	has_a	患者基本信息
患者	detect	超声、病理检查
超声检查	diagnosis	超声诊断
病理检查	diagnosis	病理诊断
超声检查报告的指标描述	part_of	超声检查报告
病理检查报告的指标描述	part_of	病理检查报告
检查报告的描述的每个指标	attribute_of	每个指标的值
超声诊断	attribute_of	超声诊断结论
病理诊断	attribute_of	病理诊断结论

4 结 语

本文以医院内部实际的 EMR 记录为基础,选择乳腺肿瘤为具体病种,提出基于 EMR 的乳腺肿瘤知识图谱的构建方法,特别对其中的概念层设计和实例层设计进行了详细阐述。乳腺肿瘤知识图谱的构建为后续疾病知识学习和推理奠定了数据基础,因此下一步工作将是基于乳腺肿瘤知识图谱的辅助诊断、智能问答。

参 考 文 献

[1] Zhang L. Knowledge graph theory and structural parsing [D]. Enschede: Twente University,2002.

[2] Singhal Amit. Introducing the Knowledge Graph: things, not strings [EB/OL]. Official Google Blog. [2012-5-16]. http://googleblog.blogspot.co.uk/2012/05/intro_ducing-knowledge-graph-things-not.html.

[3] 阮彤,孙程琳,王昊奋,等. 中医院知识图谱构建与应用[J]. 医学信息学杂志,2016,37(4):8-13.

[4] 贾李蓉,刘静,于彤,等. 中医药知识图谱构建[J]. 医学信息学杂志,2015,36(8):51-53,59.

[5] 王俏,王伟. 基于知识图谱的国际基因组流行病学可视化分析[J]. 中华医学图书情报杂志,2013,22(4):2-9.

[6] 黄鑫,胡榜利,邓莉,等. 基于知识图谱的生物医学信息可视化研究进展[J]. 中国临床新医学,2012,5(11):1090-1093.

[7] 叶华蓉,杨怡,林萱,等. BP 神经网络在高频彩超特征诊断乳腺癌中的应用[J]. 中国卫生统计,2016,33(1):71-72.

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

[9] 程学旗,靳小龙,王元卓,等. 大数据系统和分析技术综述[J]. 软件学报,2014,25(9):1889-1908.

[10] 周肖彬. 医学本体和医学知识获取的研究[D]. 中国科学院研究生院(计算技术研究所),2003.

[11] 黄小燕. 基于潜在语义关系的更年期综合症知识图库的构建及其应用研究[D]. 四川: 电子科技大学,2015.

[12] 周斌,刘世伟,高国璇,等. 2016 年 NCCN 乳腺癌临床实践指南(第 1 版)更新与解读[J]. 中国实用外科杂志,2016,36(10):1066-1027.

[13] 田驰远,陈德华,王梅,等. 基于依存句法分析的病理报告结构化处理方法[J]. 计算机研究与发展,2016,52(12):2669-2680.

[14] Bizer C, Seaborne A. D2RQ-Treating Non-RDF Databases as Virtual RDF Graphs [C]//International Semantic Web Conference,2005.