基于本体的医学影像信息整合①

张 娜, 王如龙, 王伟胜

(湖南大学 信息科学与工程学院, 长沙 410082)

摘 要: 计算机可理解的统一信息模型是基于语义的医学影像检索研究的数据基础. 讨论了医学影像及其相关信息使用中存在的数据异构、图像标注术语及语法不一致及数据格式不支持现有数据挖掘和图像语义检索的问题, 提出了一种基于本体的医学影像信息集成方案. 在分析医学影像信息来源及其关系基础上, 结合领域专家知识, 使用斯坦福大学提出的本体构建“七步法”设计了医学影像信息本体模型, 实现了本体模型的持久化、原始数据提取和数据整合, 解决了医学影像信息使用中存在的问题, 该信息模型已用于医学影像检索系统中.

关键词: 本体; 数据整合; 医学影像; 信息模型

Ontology-Based Information Model for Integration of Medical Imaging Data

ZHANG Na, WANG Ru-Long, WANG Wei-Sheng

(School of Information Science and Engineering, Hunan University, Changsha 410082, China)

Abstract: A computer readable unified information model is the data foundation in medical imaging semantic retrieve . In this paper, some challenges including lacking of unified information model for medical imaging information, the terminology and syntax for describing the semantic content in medical imaging varying were discussed, and an ontology-based information scheme for medical imaging information integrating was developed. Based on the analysis of medical imaging data source and the relationship of them, a medical imaging information ontology model was developed using "seven-step" method proposed by Stanford University, and the persistence of ontology model, extracting original data and data integration were realized. The information model was used in medical imaging semantic retrieve.

Key words: ontology; data integration; medical imaging; information model

1 引言

随着计算机断层扫描(CT)、核磁共振成像(MRI)、超声(US)等医学成像技术的发展, 医学影像在辅助诊断中得到大规模的应用. 如何充分利用和挖掘医学影像中重要的辅助诊断信息, 以支持深入的科学研究是亟待解决的关键问题, 也越来越受到医学影像研究领域的关注, 在基于医学影像的计算机辅助诊断、基于医学图像内容的检索以及医学图像标注领域开展了一些有意义的研究工作并取得了一定的研究成果.

但是目前还存在一些问题限制了医学影像的充分利用. 首先, 虽然医学影像提供了有关影像检测设备、病人、人体器官的生理结构及其病变的丰富信息, 但由于其采用医学影像特有的 DICOM[2]格式编码, 信息不能被计算机直接处理, 而且现有的有关数据检索、数据挖掘技术还不支持 DICOM 图像格式, 因此要利用医学影像提供的知识进行计辅助诊断并挖掘其包含的有用知识还存在一些障碍.

其次, 医学图像标注缺乏统一的语法和术语的支持. 医学图像、医学图像标注以及诊断结果组成了医学图像信息使用的整体. 其中标注是对医学图像中存在的解剖结构、病变以及诊断结果的语义描述, 是开展医学图像内容检索、数据挖掘的重要信息. 尽管在图像标注的其他领域已经制定了统一的标注术语及语法格式, 但在医学图像领域还没有一个被广泛采用的

① 基金项目:湖南省自然科学基金(12JJ6061);湖南省科技计划(2012SK3185);湖南大学青年教师成长计划
收稿时间:2012-08-18;收到修改稿时间:2012-10-20

统一的标注术语和语法格式, 医学图像标注普遍利用自然语言进行随意的书写, 限制了医学图像信息的共享和交换, 以及在语义 WEB 上的使用.

最后, 医学图像信息缺乏一个统一的信息模型. 病人及其检查信息被编码存储在 DICOM 图像文件的头部区域、标记信息以图像覆盖层(overlay)的形式存储为 DICOM 图像的显示对象、图像的语义信息和其他相关信息则多存储为非结构化的文本格式, 因此条块分割的信息组织方式给开展基于内容的图像检索研究带来了困扰, 进一步限制了医学图像信息的利用.

为此, 借鉴其他图像研究领域的研究成果和语义网技术对医学图像信息集成模型进行研究, 利用本体技术设计并实现了医学图像信息集成模型并用于医学图像内容检索研究中.

2 基于本体的医学图像信息集成模型

2.1 医学图像信息

目前, 医学图像信息分别存储在以 DICOM 为标准的医学图像和以格式化或非格式化存储的文本中. 这些信息包含了患者、影像设备、影像模态、成像参数、影像、影像标注、影像语义以及影像病理诊断信息. 其中患者信息包含了患者的标识信息、其人口统计学信息及医疗信息, 这些信息按照 DICOM 图像的编码格式存储在 DICOM 图像文件的头部; 影像信息按照影像检查、序列、图像文件的层次组织为一系列的 DICOM 图像, 包含了影像检查的设备信息、成像参数等数据, 也保存在 DICOM 图像文件的头部; 影像标注以直线、矩形、圆形、箭头、曲线、多边形等图形方式标记了图像中的感兴趣去(ROI), 以文本形式给出了这些 ROI 的注解或其他相关信息, 这部分信息以覆盖图层的形式存储在 DICOM 图像文件中; 影像语义给出了放射学专家对图像的理解或诊断, 属于图像的语义信息, 这部分内容多存储为文本形式, 以影像诊断报告形式存在; 同时以文本形式存在的还有和医学影像相关的其他信息, 例如患者的病理学诊断结果等.

由于文本信息和影像信息之间在检索和数据存储间存在语义鸿沟, 因此需要将医学影像领域的相关知识和概念整合在一起, 形成一个完整的知识体系.

2.2 本体

本体是对共享的概念进行形式化的显示规范说明[1]. 本体包括概念模型、显式、形式化和共享四个基本的层次含义, 首先本体由概念组成, 而概念是对客观世界共性存在的抽象描述; 显式表示本体的概念具有明确的语言表述, 对计算机可读; 共享说明本体表达的是被广泛公认和接受的知识. 在本体的组成中包括概念、关系、规则、公理和实例五个元语. 通过使用这五个元语对概念进行描述, 定义概念的属性, 通过关系、规则和公理来表达概念之间的内在联系, 通过实例来表示概念在客观世界中实际存在的个体, 从而实现知识的表示.

目前, 不少医学研究组织都致力于医学领域本体的研究, 通过构建医学领域本体将隐性的医学知识显性化, 并通过形式化的本体描述语言描述、发布和共享这些医学知识.

因此在本研究中通过建立基于医学影像领域知识的领域本体, 收集信息源中的数据, 并把数据按规定的格式和统一的术语整合为统一的信息模型, 形成完整的知识体系.

2.3 本体构建及知识表示

本体的构建一般遵循: 1)明确性, 本体中的概念应当具有明确的含义, 且是客观的; 2)一致性, 本体中的概念应当一致, 当由公理、规则进行推导时, 不应出现矛盾; 3)可扩展性, 在本体中添加新的概念时, 不应修改原有的概念; 4)编码倾向程度最小, 本体概念的定义应当在知识层面, 而不应依赖于符号层面的某一编码体系; 5)本体约束最小, 本体中对于客观存在进行建模的约束应尽可能地少, 以便于不同的成员能够自由地细化或实例化本体中的概念.

斯坦福大学医学信息学中心提出的领域本体构建七步法被广泛采用, 该方法包括确定本体的领域和范围, 列举该领域中的重要术语, 定义本体中的概念、概念层次以及概念之间的关系, 创建本体中的实例等七个步骤. 同时该中心还基于七步法开发了本体构建工具 Protégé[3]可用于构建本体的概念、属性和实例, 并可以将本体构建结果转换为以 OWL、RDF 本体描述语言描述的文本格式, 或持久化到数据库中.

通过与领域专家的合作, 采用斯坦福大学的本体构建七步法, 对需要整合的医学图像信息进行分析, 确定了本体构建的目标和语境, 收集和分析了医学影像相关的术语、相关知识和概念.

首先确定使用的专业术语: RadLex(Radiology Lexicon)[5]作为一个由北美放射学会为实现放射学信息, 包括放射图像的统一索引和检索而开发和维护的

受控词汇表, 其包含了近 12000 个解剖学概念和病理学概念, 以及成像技术和诊断图像质量等信息, 其不仅包含领域知识还包含了词法信息. 因此在医学影像信息的本体模型中引入 RadLex 本体, 使用其提供的术语和概念来构建本研究的本体模型.

其次确定相关概念及概念间的关系: 模型定义的顶级概念包括疾病分类、解剖结构、影像模态等一般性的医学专业知识概念和影像、影像标注、影像元数据等医学影像相关的概念. 其中疾病表示临床诊断中的疾病分类, 其能使用的术语限制为 RadLex 提供的疾病分类; 解剖结构描述了人体生理结构的相关概念及其之间的关系; 影像模态则描述了 CT、MRI 等影像模态的相关知识;

影像概念的定义参考了 DICOM 图像信息模型用来表示由不同成像设备产生的一副或多幅(序列)图像, 一个影像由一个检查组成, 一个检查可以有一个或多个序列, 每个序列有一副或多幅图像组成. 其中检查包含检查标识号及检查日期; 序列概念中包含序列标识号、协议名称以及模态信息; 图像包含了图像的 SOPID、病人访问、像素空间及图像生成时间等成像信息.

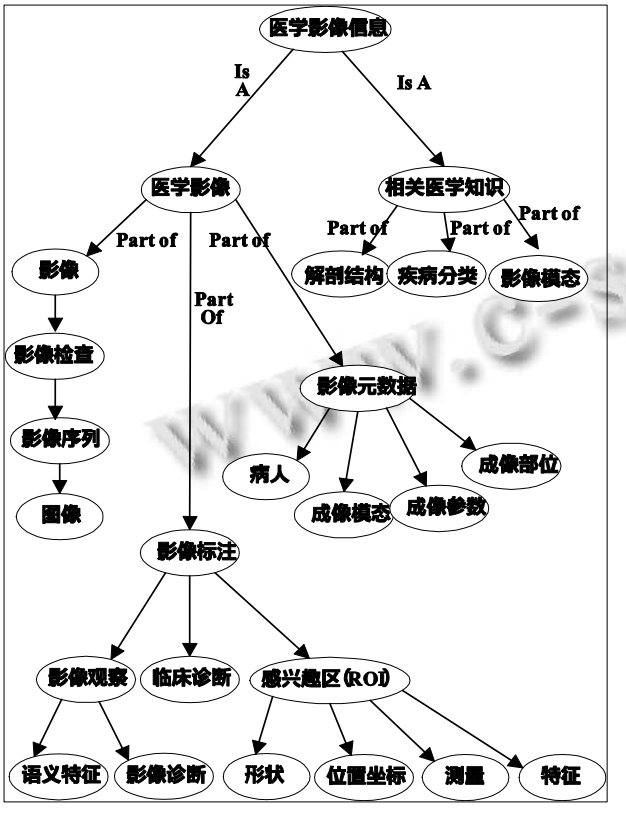

图 1 医学信息本体

影像标记概念描述了以覆盖图形式存储在 DICOM 文件中的放射学专家在 DICOM 图像上的标记, 一副图像上可以有 0 个、1 个或多个标记, 每个标记用直线、箭头、矩形、椭圆等形状直观地表示图像中的 ROI, 同时还可以包含对 ROI 的文字注解, 标记图形以二维或三维坐标的形式存储图形在图像中的位置和大小; 同时还需要提取 ROI 的大小、直径、面积、体积等作为 ROI 的测量结果, 同时还包括专家对 ROI 其他相关特征的描述来作为数据挖掘的基础数据.

影像观察描述了放射学专家或计算机程序对影像包含的语义信息, 是对影像临床诊断结果的描述. 其中语义特征概念描述了放射学专家对影像观察结果的语义描述, 可包含 0 个或多个影像诊断相关特征的描述, 例如从密度、透明度、边界等方面描述影像中观测到的解剖结构或病变组织特征.

系统提取的概念本体如图 1 所示.

3 实现

系统的主要处理流程包括本体库构建、本体库持久化和数据整合.

3.1 本体库构建及持久化

本体库构建中我们使用斯坦福大学医学信息中心开发的 Protégé工具按照图 1 的设计进行了本体库的构建, 描述为 OWL 文件形式.

对由 OWL 描述的本体文件进行解析, 将概念、关系持久化到数据库表中. 本体库需要存储包括本体模型中的概念、概念之间的关系以及概念的属性, 因此分别设计了概念表、属性和概念关系表来存储这些信息, 概念表中只存储概念的名称、定义以及同义词信息, 概念的属性则存储在属性表中, 通过外键关系来描述概念和属性的关系; 概念间存在的关系信息则存储在概念关系表中, 通过外键、主键关系和概念表联系起来. 设计的本体库中主要包括 Concept 表、Attribute 表、ConceptLinks 表. Concept 表用来存储概念, 表中的 Concept_ID 用来表示每个概念的唯一标示号, name 表示概念的名称, definition 字段描述了概念的定义, source 则描述了概念的来源, 考虑到当前的放射学领域使用的专业术语不统一, 使用 acr_id、acr_term、snomed_id、snomed_term、umls_id、umls_term 分别用来描述和概念相对应的 ACR(America College of Radiation Index) 、 SNOMED CT(Systematized

Nomenclature of Medicine--Clinical Terms) 和 UMLS(Unified Medical Language System)中的术语,作为概念的同义词或近义词. Attribute 表存储了概念所具有的属性和取值范围. ConceptLinks 存储概念间关系,CID、PID 描述两个相关的概念的标识, relationtype 字段则描述概念间的关系. 表之间的关系如图 2 所示.

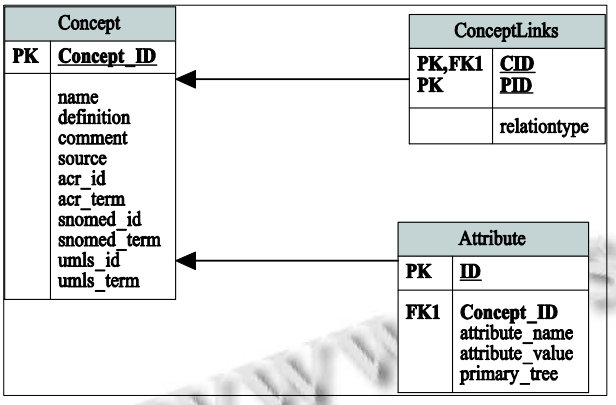

图 2 本体持久化化后的数据库表结构

3.2 数据整合

为了解决医学影像数据组织的异构性需要进行数据整合, 在医学影像信息的整合中, 数据组织的异构性主要体现在两个方面: 1)数据本身的异构性. 需要整合的数据包括 DICOM 图像、放射学报告、临床诊断报告等多种类型的数据和信息. 2)数据表示的异构性. 医学影像信息涉及到的数据表示形式有 DICOM 图像格式、放射学结构化报告及非结构化的文本信息. 针对数据的异构性, 使用 vc++和 DCMTK 工具包[4]开发了数据整合程序来实现从原始数据提取所需信息并存储到数据库的功能. 首先通过调用 DCMTK 开发包中提供的DICOM文件解析功能按照DICOM标准提取了有关病人、设备、影像信息按照相对应的本体和与其他概念之间的关系实例化.

在 DICOM 图像中的 overlay 用来表示在图像中人工添加或机器自动产生的标记图形, 用于指定 ROI、参考符号和注解, 其形式可以是位图、图形或者文本. 根据 overlay 的存储编码规则, 用 c++和 DCMTK 编程实现了嵌入图像像素和独立存储两种方式的 overlay 数据的提取. 提取的 overlay 数据组织成相应的影像标记的本体实例, 其中提取的用于注解的文本信息映射为标记本体的 text 属性, ROI 的形状、颜色和位置则分别映射为对应的 ROI 形状本体及其下级本体, ROI 的位置则映射为 ROI 本体的坐标本体实例. 同时提取了 ROI 的图形特征, 本系统中提取了 ROI 图形特征主要包括直径、面积、体积、圆形度等图像的低层特征. 提取的图形低层特征则映射为本体模型中的标记计算等相关本体实例.

医学图像的语义信息主要来源于影像的诊断报告, 其描述形式一般为自然语言. 在语义信息的提取与整合过程中首先进行概念识别: 对于自然语言文本按照语法规则进行分词, 将RadLex本体中的概念及其同义词作为词库对分词结果进行匹配. 匹配采用最长字符串原则, 即对于文本中某个词语, 若存在本体概念或概念的同义词与该词语字符串的部分相匹配, 则取匹配最长的本体概念识别该词语. 提取的概念作为实例存储, 同时记录词汇间的关系.

4 结论

通过对医学影像及其相关知识进行本体建模, 可以将医学影像中包含的隐性知识线性化, 使得这些知识能够被方便地共享和重用, 并无缝地应用于语义网络和医学知识数据挖掘领域, 提高了信息的使用效率; 本体模型可以对异构数据进行整合并实现基于语义的查询, 有效提高了查全率和查准率. 在下一步的工作中, 对于整合后的数据开展医学数据挖掘相关研究, 通过数据发现本体概念之间的潜在关系.

参考文献

1 Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition, 1993,5(2):199–220.

2 Digital imaging and communications in medicine(DICOM) 2003. American College radiology &National Electrical Manufactures Association, 2000,3: 32.

3 Rubin DL, Noy NF, Musen MA. Protégé: A tool for Managing and Using Terminology in Radiology Applications. Journal of Digital Imaging, 2007, 20(1): 34–36.

4 Trigo JD, Alesanco A, Martínez I, García J. A Review on Digital ECG Formats and the Relationships Between Them. Information Technology in Biomedicine, IEEE Trans. on, May 2012,16(3):432–444.

5 Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics, 2006 Novdec; 26(6): 1595–1597.