·综 述·

医学领域本体研究现状

陈 焱[1]，姜慧敏[2]

（1. 吉林大学 公共卫生学院，吉林 长春 130021；2. 空军航空大学信息中心，吉林 长春 130022）

摘 要：领域本体描述了关于某个学科领域中的概念以及概念之间的关系，或者该学科领域的重要理论和基本原理。基于医疗领域的特殊需要，国内外均开展了医学领域本体的研究。本文介绍分析了国内外医学领域本体的研究现状，并提出未来需要解决的关键问题。

关键词：本体；领域本体；医学

中图分类号：GR－1　　文献标识码：A　　文章编号：1007-7634（2006）10-1587-04

Research on Medical Domain Ontology

CHEN Yan[1]，*JIANG Hui－min*[2]

（1. *School of Public Health JiLin University*，*Changchun* 130021，*China*；
2. *Information Centre*，*Airmen Aviation University*，*Changchun* 130022，*China*）

Abstract：Domain ontology describes relations between concepts of some field，or important theories and principles of this field. Medical domain ontology is researched both domestically and overseas based on specific needs of medical field. This paper introduces and analyzes the status quo of research on medical domain ontology internationally，and proposes the key issues which need to be resolved in the future.

Key words：ontology；domain ontology；medical

领域本体（Domain ontology）是专业性的本体，描述的词表关系到某一学科领域，描述该学科领域中的概念以及概念之间的关系，或者该学科领域的重要理论和基本原理[1]。领域本体可以有助于人与组织之间的信息交流[2]；分析专业领域知识[3]；了解该领域的演化发展。在医疗领域中，医院信息系统开发供应商希望有一种统一编码系统来满足临床电子病历发展的需要；行政管理部门及医疗质量研究人员缺乏一种可以理解和评价不同医院诊断/临床纪录的标准术语集；医疗保险部门也需要标准的术语编码实现对病人的临床医疗诊断和治疗信息的自动化处理。因此客观上需要一种新的术语集，它既可以满足用户结构化、智能化录入临床数据的需要，又能够用于优化自然语言处理，还能帮助医学信息的存储、提取与分析[4]。医疗行为随着生物－医学模式向社会－心理－生物医学模式的演变，生物医学再也不是“单纯”的自然科学了，而是渗入了人文科学的成份，因此该领域本体不仅包括临床医学术语，还应包括大量生命科学、人文科学领域的术语[5]。鉴于该领域的特殊性和复杂性，国内外图书馆情报学界、计算机科学界均展开了医学领域本体的研究。

1 国外医学领域本体主要研究项目

1.1 统一医学语言系统（UMLS）

1986 年，美国国立医学图书馆主持了一项长期研究和开发计划，即统一医学语言系统（Unified Medical Language System，简称 UMLS）[6]。该研究计划目的是要开发机读的“知识资源”，使许多不同的应用程序利用它来弥补不同机读资源的差异，以及用户在概念表达方式上的差异，从而识别与特定用户需求密切相关的信息资源；借助必要的电子通

收稿日期：2006－09－05

作者简介：陈 焱（1970－），女，吉林长春人，馆员，硕士研究生，从事医学信息管理研究；姜慧敏（1968－），女，吉林省四平市人，馆员，从事信息资源管理研究。

讯和查找方法，从这些信息资源中获取信息，使用户易于从电子病案系统、书目数据库、事实数据库、专家系统等信息资源中获取信息。

UMLS 目前收录概念超过 80 万个，覆盖了 60 多个词库。其中，NLM 独立开发维护的数据库近 50%。2006 年版的 UMLS 语义网络包括 135 个语义类型和 54 个语义关系。UMLS 由超级叙词表(UMLS Metathesaurus)、语义网络(UMLS Semantic Network)、信息源图谱(Information Source Map，简称 ISM)及专家词典(Specialist Lexicon)组成。超级叙词表是生物医学概念、术语、词汇及其涵义、等级范畴的广泛集成，是根据概念或涵义为中心组织起来的，是“一种将概念转换成词、将词扩展成概念的工具”，其根本目的是将相同概念的交替名称以及多种变形形式联系在一起，并识别不同概念之间的关系。

语义网络为超级叙词表所有概念提供了语义类型、语义关系和语义结构。语义类型是通过计算机程序指定的或在人工复审过程中增加的，它既是超级叙词表与语义网络之间的连接，也是超级叙词表与情报源图谱的连接之一。语义类型是语义网络的节点，节点与节点之间的关系即为语义关系。由语义类型和语义关系构成了网状的语义结构。情报源图谱是一个关于生物医学机读信息资源的数据库，其目的是利用超级叙词表和语义网络实现以下功能：测试情报源与特定提问的相关性，以便选取最合适的信息源；为用户提供特定信息源的范围、功能和检索条件等人工可读的信息；自动连接相关信息源；在一个或多个情报源中自动检索并自动组织检索的结果。专家词典是一个包含众多生物医学词汇的词典，它是在 NLM 自然语言处理专家系统项目上开发出来的[7]。

1.2 标准医学参考术语(SNOMED RT)

标准医学参考术语(Systematized Nomenclature of Human and Veterinary Medicine Reference Terminology[8]，简称 SNOMED RT)由美国病理学家学会所属的 SNOMED 编委会开发和推广，SNOMED RT 的测试版包括了拥有各自编码的 180,000 个词条，每个词条都与一个概念相连接，概念的总数是 110,000 条，包括了 260,000 显式的关系，分别组织在 12 个不同的轴和章节中，它包括解剖学、形态学、正常与非正常的功能、症状及疾病体征、化学制品、药品、酶及其它体蛋白、活有机体、物理因素、空间关系、职业、社会环境、疾病/诊断和操作。SNOMED RT 的每一个术语(词条)均有一个编码与之对应，在疾病/诊断轴内，很多疾病概念同时提供与其它术语的交叉参照关系[9]。SNOMED RT 包括概念表(Concepts Table)、术语表(Terms Table)、层次和关系表(Hierarchies & Relationship Table)、描述逻辑(Description Logic)。①概念表：使拥有同一个实质性内涵(概念)的不同术语拥有相同的概念码和不同的术语码。概念码与 SNOMED RT 中的 TermCode 是相同的。每个概念都有一个唯一的概念码，但每一个概念码可能与多个术语相对应，概念表由三部分组成的：概念码、全名和状态(状态“OK”代表在用，“RET”代表已退役)。②术语表由四部分组成：术语码(ID)、术语的同义词关系(Eclass)、术语的字符串(Term)、状态(Status)。③层次与关系表：为了提供一种连贯、稳定、多层次的关系，SNOMED RT 提供一张关系表来显式地表达医学概念之间复杂的、多层次的关系。它由三部分组成的：概念码、关系及该概念的上层概念码。④描述逻辑：SNOMED 编辑委员会选择了类似知识表达系统规则(KRSS，Knowledge Representation System Specification)的方法来表述 SNOMED RT 的概念。

1.3 人类发育解剖学本体(HUMAT)

人类发育解剖学本体(Ontology of Human Developmental Anatomy，HUMAT)是 J. Bard 建立的有关人类解剖学数据库，分为标准解剖学和详细解剖学两个目录及小鼠和人类发育阶段的比较图谱。此外，还提供大量的有关人类胚胎发育和相关信息的网页。内容包括：Standard Anatomy，Detailed Anatomy，Comparison Chart，Note on ontology construction，People，Links[10]。人类发育解剖学本体按胚胎发育每一个卡内基分期(Carnegie Stages，下称 CS)形成结构化树状表，在每一个 CS，胚胎组织按层次并归类到上级组织中。在人类发育解剖学本体中将人类胚胎的发育分为两个阶段，CS9 以前为一个阶段，CS10 及以后为一个阶段，在 CS9 以前在主要描述是胚胎外成分即组织的三个机体层和它们早期衍生物，以及各种机体腔。在 CS10 以后为胚胎成熟期，围绕主要器官系统进行组织和描述。

1.4 基因本体(GO)[11]

2000 年，基因本体联盟(The Gene Ontology Consortium，GOC)开始进行基因本体(Gene ontology，简称 GO)项目的研究。GOC 的目的是要创建

一套动态的受控词表。GO 旨在建立一个适用于各种物种的，对基因和蛋白功能进行限定和描述的，并能随着研究的不断深入而更新的语言词汇标准。GO 是多种生物学本体语言中的一种，提供了三层结构的系统定义方式，用于描述基因产物的功能。随着分子生物学技术手段和显微设备的发展，细胞中基因和蛋白质作用方面的知识处于不断积累和变化中，因此研究真核生物的基因序列需要动态性工具。为了达到这一目的，GO 构建了三部相互独立的本体，它们是生物学过程本体（Biological Process）、分子功能本体（Molecular Function）和细胞成分本体（Cellular Component）。一个基因产物有一个或多个分子功能，可能被用于一个或多个生物过程当中；基因产物可能辅助一个或多个细胞成分，这就是基因本体之所以构建三个本体的原因[12]。

可以看出 GO 项目旨在定义出一套结构化的、定义精确的通用受控词表，可用来描述任何有机生物体中基因和基因产物的作用。生物学过程本体的任务是描述有序的生物化学反应的全过程，如有丝分裂、嘌呤代谢等。分子功能本体的任务是描述每个基因产物发挥作用的全过程。细胞成分本体的任务则是描述亚细胞结构、细胞器定位、大分子复合物的结构等。

GO 中的术语都是经过结构化组织的，是定向非环化曲线结构。这种结构与通常的等级体系不同，在这种体系关系的结构中，一个子节点可能会有多个父节点。GO 的定义法则已经在多个合作的数据库中使用，这使在这些数据库中的查询具有极高的一致性。这种定义语言具有多重结构，因此在各种程度上都能进行查询[13]。

2　国内医学领域本体主要研究项目

2.1　医学知识库（NKIMed）

医学知识库（NKIMed）是国家基础知识设施（National Knowledge Infrastructure，简称 NKI）的子集，NKI 是中国科学院计算技术研究所博士生导师曹存根于 1995 年提出的一个具有基础性和前瞻性的课题。NKI 是一个庞大、可共享、可操作的知识群体，它不仅集成了各个学科的公共知识，而且还融入了各学科专家的个人知识[14]。课题组在 NIKmed 的构建过程中研究了医学领域本体，研究了医学本体中的类结构设计、从文本知识中获取医学知识的方法以及从半结构化文本中自动获取知识的方法；建立了 52 个医学概念类，整理出了 1691 种医学属性和 107 条医学关系，并利用这些概念类获取到 554 个临床检验指标的知识，以及 19595 个医学概念的知识，共计 78012 条医学知识。同时，获取了相应的医学公理共 812 条，并用这些公理对医学知识进行一致性分析和知识推理。

2.2　中国科学技术信息研究所的研究[15]

中国科学技术信息研究所的庞景安和他的硕士李毅进行了基于多层次概念语义网络结构的中文医学信息语义标引体系和语义检索模型研究。这项研究主要进行多层次概念语义网络结构的理论性探讨；中文医学信息三层概念语义网络结构的设计和各网络层次语义类型和语义关系的完善；语义标引体系的建立和语义标引方法的确定；语义检索模型的建立和不同检索方法检索效率的比较。

该研究通过对中文医学信息的语义研究，在目前语义网络的基础上，设计了专业概念语义网络层、普通概念语义网络层和基本概念语义网络层。其中专业概念语义网络层满足了医学专业领域的需求；普通概念语义网络层增强了医学信息语义表达的广度，以此可对评价、态度、思维、认识、修辞等语义信息进行标引并实现检索；基本概念语义网络层加大了医学信息语义表达的深度，使得关于秩、序、时、空、数、量、质、类、度以及趋势性、对比性、可能性、存在性、真伪性、判断性等属性的范畴信息标引和检索成为可能。三层概念语义网络结构的检索模式完善并丰富了现有语义网络及其语义类型和语义关系，构建了 142 种语义类型和 81 个语义关系。

2.3　中医药一体化语言系统

国家科技部基础性工作——《中医药一体化语言系统》项目以中医药学科体系为主导，遵循中医药学科理论体系，旨在建立中国第一个计算机化的、可持续发展的、包含中医药学及其相关学科语言的中国医药学检索语言集成系统，形成中医药学的语言系统平台[16]。现已初步：（1）建立了整体结构与原则、规范了术语加工标准与方法，收词量达 60 余万条。完成了 11 万中医药术语概念词条的加工与关联，其中：中医基础理论类 3700 条；中药学类 3777 条；疾病－症状类 6543 条；中医各家学说类 3200 条；医学人物类 10434 条；诊断学类 5008 条；中成药

及方剂类 44271 条; 中医文献类 11980 条; 药用动植物类 16356 条; 中医治法类 2562 条等。(2)建立了中医药学科结构分类体系、中医药语义类型与中医药语义关联系统, 展示了中医药语言的特性与内在的关联关系。建立了本系统一级类目, 及中医基础理论、中药学、方剂学、药用动植物学、中医各家学说、中医诊断、疾病证候、治则治法、医学人物、中医文献等专题类目的体系分类。(3)实现了各类目间的语义关联及语义描述浏览功能, 实现了网上同义词查询检索功能。在设计上参照美国 UMLS, 将中医药学语言系统中隐含的各种语义关系全部提取出来, 形成语义关联, 并以此为中心, 建立学科术语概念与概念、概念与名词、名词与名词之间的内在联系, 形成一个网状的信息表示结构。

3 结 语

综观国内外的研究现状, 可以发现目前关于医学领域本体理论的研究日趋成熟, 理论体系正在逐步完善; 关于医学领域本体应用的研究国内外还处于不断发展阶段, 国外研究重点在于构建医学领域规范化工具, 对医学术语语义概念和语义类型进行表述, 这些描述都只是涵盖的大量概念的具有简单推理功能扩展化词表, 而缺少智能知识库所需要的推理功能和智能检索功能, 国内医学领域本体的研究与知识库和医学信息系统结合比较紧密, 试图构建基于本体的具有智能推理功能的医学知识库, 但目前尚未现出可以应用的成熟产品。

医学生物信息学家常常掌握着一些特殊的数据, 他们需要找到服务来操作这些数据以便产生期望的结果, 或者希望在这些数据之上应用一些任务。另一方面, 他们必须准确无歧义地表达他们的需求, 以便同可用的服务匹配, 并考虑服务的功能、接收和产生的数据以及用于完成其目标的资源, 从而匹配用户的需求[17]。

目前, 这些功能的实现还有如下需要迫切解决的关键问题[18]: (1) 应用有效的知识标识语言, 对医学资源进行语义描述, 通过单一语义映像和多层语义互联, 将医学知识库群从多个不同类型的语义空间变换并整合到一个统一的资源空间。(2) 针对各种医学信息用户需求表达具有语种的多样性和语义的复杂性, 利用以 ontology 为核心的语义网技术将用户需求有效地整合到一个统一的用户需求空间, 使得需求具有语义相关性的用户能够实现在共享资源上的协同工作。(3)应用虚拟组织机制, 将已经整合到统一资源空间知识库和整合在统一用户需求空间的用户需求整合在一个虚拟语义空间中, 实现知识库与用户需求的语义匹配, 达到用户需求与知识库的互理解。

参考文献

1 牟冬梅, 崔艳玲. MeSH、本体论在医学知识组织中的作用[J]. 情报杂志, 2005, (7): 120—122.

2 李善平, 尹奇韡, 胡玉杰, 等. 本体论研究综述[J]. 计算机研究与发展, 2004, (7): 1041—1052.

3 McGuinness, D. L., Fikes, R., Rice, J. and Wilder, S.. An Environment for Merging and Testing Large Ontologies[EB/OL]. http://www.ksl.stanford.edu/people/dlm/papers/kr2000—camera—ready—copy.doc, 2005—01—01.

4 朱礼军. 万维网环境下基于领域知识的信息资源管理模式研究[D]. 北京: 中国农业大学博士论文, 2004.6.

5 Alxea T. McCray. An upper-level ontology for th biomedical domain[J]. Comparative and Functional Genomics, 2003, (4): 80—84.

6 http://www.nlm.nih.gov/pubs/factsheets/umls.html, 2005—02—15.

7 李 毅. 基于多层次概念语义网络结构的中文医学信息语义标引体系和语义检索模型研究[J]. 情报学报, 2003, (4): 403—411.

8 http://www.snomed.org/, 2005—10—16.

9 SNOMED International, Introduction, College of American Pathologists, April 1993, page 6.

10 Edinburgh Human Developmental Anatomy[EB/DL]. http://www.ana.ed.ac.uk/anatomy/database/humat/, 2005—02—16.

11 GO 简介[EB/DL]. http://db.sgst.cn/index2_20_b.jsp, 2005—09—16.

12 An Introduction to the Gene Ontology[EB/DL]. http://www.geneontology.org/GO.doc.shtml, 2005—02—15.

13 DNA replication Grap hical view[EB/DL]. http://www.godatabase.org/cgi-bin/amigo/go.cgi?action=dotty&view=details&search_constraint=terms&query=GO:0006260&session_id=1435b1108519687, 2005—03—01.

14 http://www.ict.ac.cn/3-2-23.htm, 2005—02—18.

15 李 毅. 基于多层次概念语义网络结构的中文医学信息语义标引体系和语义检索模型研究[J]. 情报学报, 2003, (4): 403—411.

16 张汝恩. 建立《基于本体论体系的中西医药一体化语言系统》[DB/DL]. http://test.cintcm.com/sjgx/database/content/yyxt/article/yitihuayuyexitong.doc, 2005—09—26.

17 刘 炜. 学习"语义网格"[EB/OL]. http://meta.blogchina.com/2447032.html, 2005—10—26.

18 毕 强, 牟冬梅, 刘 昆. 语义网格环境下数字图书馆知识组织研究论纲[J]. 图书情报工作, 2006, (6): 28—33.

(责任编辑: 滕代娣)