

基于临床指南的糖尿病本体构建及语义检索模型设计

于 凡 雷行云 高 星 万艳丽 吕欣航 胡红濮 王 青

(中国医学科学院/北京协和医学院医学信息研究所 北京 100020)

〔摘要〕 以糖尿病临床指南为基础,利用本体技术对糖尿病指南中的概念进行抽取并建立语义关联,通过 7 步法和骨架法构建糖尿病本体库,详细阐述构建步骤。在此基础上设计糖尿病本体语义检索系统模型,为糖尿病管理系统开发提供借鉴。

〔关键词〕 糖尿病;临床指南;本体;语义检索

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2018.05.011

Diabetes Ontology Building and Semantic Retrieval Model Design Based on Clinical Guideline YU Fan, LEI Xing-yun, GAO Xing, WAN Yan-li, LV Xin-hang, HU Hong-pu, WANG Qing, Institute of Medical Information, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing 100020, China

〔Abstract〕 Taking diabetes clinical guidelines as the foundation and making use of the ontology technology to extract concepts from diabetes guidelines and to establish semantic association. To build the diabetes ontology base with the seven-step method and skeleton method, the paper explicated the building steps. On this basis, it designs the diabetes ontology semantic retrieval system model to provide reference for the development of the diabetes management system.

〔Keywords〕 Diabetes; Clinical guideline; Ontology; Semantic retrieval

1 引言

据世界卫生组织 (World Health Organization, WHO) 数据显示,糖尿病已成为继肿瘤、心脑血管疾病之后的第 3 大威胁人类健康的慢性非传染性疾病^[1]。中国已成为全球糖尿病患者第 1 大国,约有 1.1 亿名糖尿病患者。据 WHO 预测,到 2025 年中国 2 型糖尿病患者将超过 1.3 亿^[2]。糖尿病的急、

慢性并发症尤其是慢性并发症可累及身体多个器官,致残致死率非常高,严重影响患者的生存质量和预期寿命。因此提高糖尿病的诊治水平,实现早预防、早治疗、早达标已成为当前专家学者和医生的研究重点和方向。

传统的医疗实践中医生多以教科书的理论知识和个人临床经验为主,一些主观判断错误会导致误诊或漏诊。为提高医疗过程的严谨性及诊疗效果,临床指南 (Clinical Guidelines, CG) 应运而生,它是在循证医学的概念基础上提出,按照循证医学的方法开发,旨在优化患者护理的一系列诊疗指导意见^[3],辅助医生做出符合患者病情的恰当处理和选择,从而提高医疗质量。但在实际应用中医生的临床

〔修回日期〕 2018-04-05

〔作者简介〕 于凡,硕士研究生;通讯作者:胡红濮,研究员,发表论文 60 余篇。

指南的利用率并不高,因为临床指南大多以文本形式进行保存,查阅起来比较耗时,且临床指南涉及众多医学专业知识,不便于记忆^[4]。因此将文本型的临床指南电子化,提高其利用率显得尤为重要。

本体作为一种知识表达模型的方法,可以做到知识表达与计算机的统一和互操作,还可为语义表达提供支持并实现知识共享,是对医学领域知识进行整合和组织的一种有效方法,众多专家学者都尝试将本体应用于医学领域的知识表示。如李晓瑛等人利用 PubMed 数据库和一体化医学语言系统(The Unified Medical System, UMLS),医学系统命名法-临床术语(SNOMED CT)搭建肿瘤本体,以扩充肿瘤本体知识库^[5];陈志云依托肝炎疾病知识库系统进行肝炎疾病本体构建并以此进行语义智能检索的探索^[6-7];叶青、李博等人依据高血压临床指南构建高血压临床指南本体,以期实现诊断推理^[8];刘何心利用医学主题词表尝试构建糖尿病领域本体模型^[9];方芳等人利用本体技术构建中医糖尿病医案本体^[10],更好地发挥中医医案的价值;S Elsappagh 等人参照开放生物医学本体(Open Biomedical Ontology, OBO)构建 2 型糖尿病治疗本体知识库,为患者提供最佳的药物和饮食等治疗方案^[11];也有通过利用标准医学本体,提出一种基于模糊本体的案例推理框架并将其应用于糖尿病的诊断^[12]。可以看出大多数研究主要集中在利用现有的一些词表知识进行相应本体的构建,再将其应用于疾病诊断和知识库建设方面,而基于权威临床指南构建领域本体并进行语义检索的探索研究较少。因此为充分发挥糖尿病指南的使用价值,使其更好地为临床医生提供决策支持,本研究以糖尿病临床指南为基础,利用本体技术将糖尿病临床指南知识进行数字化描述,构建基于糖尿病临床指南的本体库,同时在此基础上构建基于本体的语义检索模型,提出一种根据领域本体库和应用本体库将查询请求转换为查询本体以进行相似度匹配的方法,为糖尿病管理系统和检索系统的开发提供借鉴。

2 研究方法

本体是一种被定义为共享知识概念模型的明确

的规范说明^[13],能够对某一领域内的概念及其间的关系进行很好地描述。近年来本体技术被广泛应用于医疗领域^[14-18],目前本体构建方法主要有 7 步法、骨架法、METHONTOLOGY 法、TOVE、KACTUS 工程法、SENSUS 法等^[19]。本研究主要参考 7 步法和骨架法进行糖尿病本体的构建,同时根据研究领域的应用需求特点对方法步骤进行相应的调整和改进,最终确定糖尿病本体库的构建步骤主要包括:确定本体的专业领域和范畴、复用现有的本体、列出糖尿病领域中重要知识术语、定义糖尿病类及属性关系、创建实例、本体的检验与评价。由于针对中国人群,因此选择中华医学会糖尿病学分会主导编制的《中国 2 型糖尿病防治指南(2013 年版)》和《中国 1 型糖尿病诊治指南(2012 年版)》为主要依据,以其他相关指南为辅作为参考。

3 糖尿病领域本体构建

3.1 确定本体专业领域和范畴

是构建本体的第 1 步,即明确本体所覆盖的专业领域、本体使用人员和应用本体的目的。基于临床指南的糖尿病本体主要包括糖尿病的诊断、分类、检查、治疗、并发症以及特殊人群的糖尿病等内容。基于临床指南的糖尿病本体是对糖尿病领域知识进行组织和管理,形成统一的描述化语言,应用于实际系统中。基于临床指南的糖尿病本体是为临床医生和相关科研工作者提供学习的工具,为医生在临床实践中的辅助诊断提供便利。

3.2 复用现有本体

能够减少一些重复性的工作,加强词表的重用性,在这一步中需要搜集和查阅大量糖尿病相关的词表或文献。本研究主要参考的资料有一体化医学语言系统 UMLS、中英文医学主题词表 MeSH/CMesh、国际疾病编码系统 ICD-10。这些词表虽然都只提供相对分散的糖尿病知识框架,各概念术语间的关系也不是很清晰,并不能直接作为构建标准使用,但其具有很好的参考价值,对于后续工作的开展有所帮助。

3.3 列出重要知识术语

前期需要先构建出糖尿病的顶层本体知识框架。通过对《中国 2 型糖尿病防治指南（2013 年

版）》内容进行梳理，可将糖尿病知识本体分为诊断、分型、检查、治疗、并发症、特殊人群及患者等 7 大类内容，其涉及到的主要概念，见表 1，类的概念层次结构，见图 1。

表 1 糖尿病临床指南中的主要概念

1 级概念	2 级概念	3 级概念
糖尿病诊断	静脉血浆葡萄糖水平	(1) 典型糖尿病症状（多饮、多尿、多食、体重下降）加上随机血糖检测 $\geq 11.1\text{mmol/L}$ (2) 或加上 FPG 检测 $\geq 7.0\text{mmol/L}$ (3) 或加上葡萄糖负荷后 2 小时血糖检测 $\geq 11.1\text{mmol/L}$
糖尿病分型	按病因学证据	1 型糖尿病 2 型糖尿病 妊娠期糖尿病（GDM） 其他特殊类型糖尿病
检查	病史 体格检查 化验检查	家族史、病程、既往史、饮食、运动、营养状况等 身高、体重、BMI、腰围、血压、皮肤检查、足部检查等 空腹血糖、餐后血糖、HbA1c、总胆固醇、甘油三酯、LDL-C、HDL-C、尿常规、肝功能、肾功能
治疗	特殊检查 药物治疗 运动治疗 医学营养治疗 戒烟 减重手术治疗 （肥胖伴 2 型糖尿病）	眼底检查、心电图、神经病变相关检查、尿白蛋白和尿肌酐 口服降糖药：磺脲类、格列奈类、DPP-4 抑制剂、双胍类、TZDs、 α -糖苷酶抑制剂 皮下注射：GLP-1 受体激动剂（艾塞那肽、利拉鲁肽） 皮下注射：胰岛素 ≥ 150 分钟中等强度的有氧运动/周、2 次抗阻运动/周 均衡营养素结构：脂肪、碳水化合物、蛋白质等 - 腹腔镜袖状胃切除术（LSG） 胃旁路术（RYGB） 腹腔镜下可调节胃束带术（LAGB） 胆胰旁路术（BPD）
并发症	急性并发症 慢性并发症	糖尿病酮症酸中毒（DKA） 高血糖高渗综合征（HHS） 糖尿病乳酸性酸中毒 糖尿病慢性肾脏疾病 糖尿病视网膜病变 糖尿病神经病变 下肢血管病变 糖尿病足病
糖尿病的特殊人群	妊娠糖尿病 儿童和青少年糖尿病 老年糖尿病 糖尿病与阻塞性睡眠呼吸暂停低通气综合征（OSAHS） 围术期糖尿病管理 糖尿病与口腔疾病 糖皮质激素与糖尿病 抑郁症 精神疾病、人类免疫缺陷病毒/艾滋病（HIV/AIDS）和糖尿病	- - - 共患疾病 糖尿病大血管并发症、微血管并发症 口腔黏膜病变、龋齿、牙龈炎和牙周炎等 库欣综合征、外源性（激素治疗）糖皮质激素增多 相互影响 诱发或加重糖尿病
患者	基本信息	患者 ID、性别、年龄、症状等

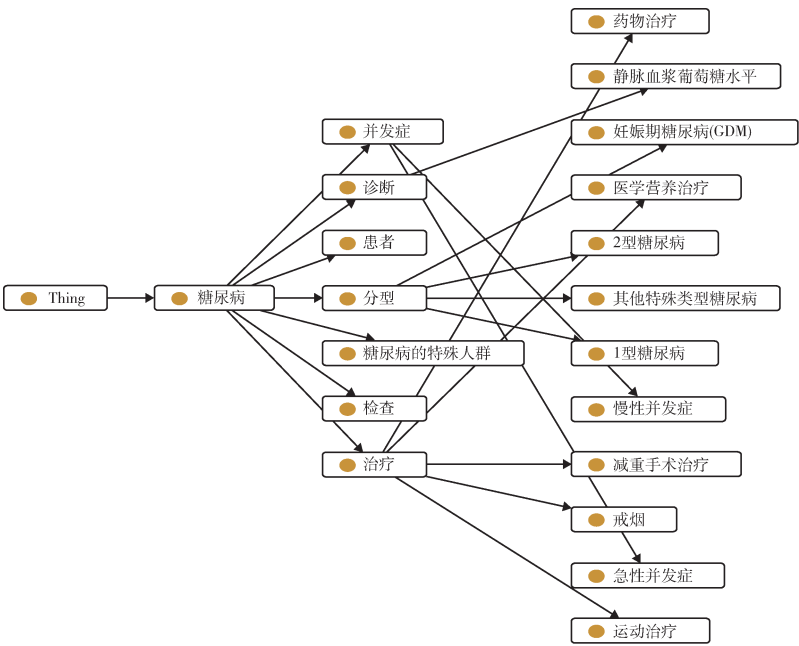


图1 糖尿病临床指南本体的类的概念层次结构

3.4 定义类及属性关系

在明确糖尿病本体的核心概念和基本层次结构后需定义糖尿病类的类及属性关系，包括对象属性和数据属性，对象属性描述的是两个不同概念间的关系，数据属性描述的是类与数据类型之间的关系。如糖尿病的分型（患者的诊断结果）和治疗（患者的治疗方法）是对象属性，患者的年龄和性别是数据属性。部分属性，见表2。

表2 糖尿病临床指南本体的部分属性

属性名称	属性类型	描述
糖尿病分类	对象	患者和糖尿病分型的关系
糖尿病治疗	对象	患者采用哪种治疗方法
化验检查	对象	HbA1c、总胆固醇、甘油三酯、LDL-C、尿常规、肝功能等指标
眼底检查	对象	患者的视力及眼底情况
心电图	对象	患者的心电图检查结果
糖尿病史	数据	患者是否有糖尿病患病史
身高	数据	患者的身高值
体重	数据	患者的体重值
年龄	数据	患者的年龄
性别	数据	患者的性别

3.5 创建实例

在确定糖尿病本体的类和属性关系后可以进行本体的实例化。实例是对类进行具体的赋值，使实例具有相应的属性值，从而展示出相应的语义关系。糖尿病临床指南中包含大量的实例，依据指南内容对本体进行实例化。实例化可以通过手工录入、机器导入和人工语义标注等方法实现，本研究通过 Protégé 的手工录入窗口在 Individuals 中录入相应的实例。将创建的实例进行保存，为后续检索使用。

3.6 本体检验与评价

领域本体知识内容既复杂又繁多，因此需要进行检验来确保所构建的本体具有合理性，Protégé 中自带推理机功能，点击“Reasoner”下的“Start reasoner”即可对构建的本体进行检测，经检测所构建的本体是合理的。对糖尿病本体进行检验后还需进行评价，通过评价对本体进行完善与修改，评价与本体构建的其他环节是一个多次迭代的过程。由于糖尿病临床指南的知识不断更新、丰富完善，对其本体的建设在稳定性和精确性的前提下要具有动态性。此外糖尿病本体库为医生和患者所使用，因此通过专家意见进行评价更为可行。糖尿病领域本

体进化及更新的内容包括以下方面：（1）概念进化。包括对本体概念的增加、删除、范畴调整。（2）概念间关系进化。包括在现有概念间建立新关系、现有关系的调整等。（3）概念实例的进化。增加或删除实例等。

4 语义检索模型设计

4.1 传统信息检索方法的不足

主要包括关键词检索和概念检索，关键词检索是将用户输入的关键词与信息库或文档库中的内容进行匹配对比，只有含关键词的内容能被检索出反馈给用户，而关键词所隐含的内容、关键词的同义近义内容无法被检索出，极大降低查全率和查准率。概念检索是通过概念相关关系拓展词汇概念，形成概念词典从而进一步扩展理解用户的检索意图，其主要是由专家制成的专业词表或是通过机器学习形成的关联概念网^[10]，可以看出概念检索一定程度上弥补了关键词检索的不足，但其表现出的依然是较初级的词汇概念间的相关关系，不能完全理解用户的检索意图。以上两种方法存在的共性问题是无法理解词汇的语义信息，不能判断出不同概念间的关系，无法挖掘出词汇所隐含的内容，如1型糖尿病患者的摄入食物“猕猴桃”，是富含维生素E的食物，属于营养物质中的维生素类，因此归属于糖尿病的营养治疗概念中。要解决此问题可在检索式中加入基于本体的语义推理，即对用户的检索条件加入语义并对检索内容进行推理。本体作为一种知识表示方法，能够将某一领域内的所有相关概念及其属性关系体现出来并进行语义推理，可以大大提高检索结果的检准率和检全率。

4.2 基于本体的语义检索模型设计

语义检索是一种基于知识的分析检索，通常在对自然语言理解的基础上借助计算机与统计学相关知识在知识关联模型下完成检索^[20]。而基于糖尿病本体的语义检索是通过对糖尿病检查、治疗相关的语义属性进行语义关联，然后可对糖尿病症状的搜索结果进行语义属性和语义关联的自动分类和匹

配，而不再是字面上的匹配，使检索结果更为全面和准确。语义检索的核心在于对提问词和检索资源的语义处理，首先利用糖尿病本体中的类（概念）的信息对检索词进行规范化处理，然后查询糖尿病指南本体库，借助语义推理引擎分析检索词之间的语义关系，使计算机真正理解用户的检索用途，本体库的查询结果将作为新的检索词对资源库进行检索。本研究提出的基于糖尿病本体语义检索系统模型主要由本体管理模块、提问词处理模块以及查询处理模块组成，见图2。本体管理模块主要是对糖尿病指南本体库的构建、更新与维护，本体库是整个检索模型的核心，因此需要准确描述糖尿病相关的每个概念及其属性关系，同时可以利用所构建的本体库建立新的信息资源（新的指南路径、医学术语等）的应用本体库（新的医学症状或指南中的实例数据），对本体库进行补充与完善。提问词处理模块是提取医生和用户所输入的关键检索词，利用糖尿病本体的类的概念对关键检索词进行标注，将用户的检索请求转换为本体的查询请求。查询处理模块是利用推理引擎或推理规则，将第1步所得到的查询本体和糖尿病指南本体、指南应用本体进行语义关系判断，揭示检索词所隐含的语义关系，实现查询本体与领域本体的相似度匹配，然后将匹配结果按相关度排序返回给用户，同时用户也可以通过多次检索进行信息反馈，对本体库进行更好地完善。

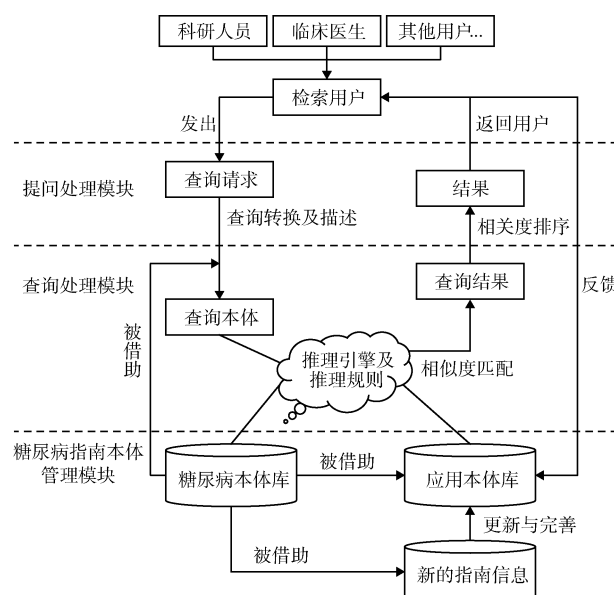


图2 基于糖尿病指南本体的语义检索框架模型

5 结语

临床指南是临床医生在诊疗过程中所遵循的规范化文件,糖尿病防治指南较为完整地总结我国糖尿病防治和诊疗的实践经验,为糖尿病及其并发症的防治起到关键作用。为进一步提高临床指南的作用和价值,本研究利用本体技术构建糖尿病指南的本体库,将其中的知识概念进行语义关联,在糖尿病本体的基础上设计基于语义的检索模型,实现语义推理与语义检索,为糖尿病管理系统的开发提供借鉴。

参考文献

- 1 闫冠韞,陈洪恩,李舜,等. 大数据视阈下糖尿病患者管理模式探析 [J]. 中国全科医学, 2018, 21 (9): 1066 - 1069, 1084.
- 2 廖涌. 中国糖尿病的流行病学现状及展望 [J]. 重庆医科大学学报, 2015, 40 (7): 1042 - 1045.
- 3 Peleg M. Computer - interpretable Clinical Guidelines: a methodological [J]. J Biomed Inform, 2013, 46 (4): 744 - 763.
- 4 赵志娟,赵玉虹. 文本临床指南转变为数字化临床指南发展现状及趋势 [J]. 中国数字医学, 2017, 12 (1): 32 - 34.
- 5 李晓瑛,李丹亚,夏光辉,等. 肿瘤本体构建研究 [J]. 数字图书馆论坛, 2015, (8): 37 - 42.
- 6 陈云志. 肝炎本体构建及语义相似度研究 [D]. 杭州: 浙江大学, 2017.
- 7 李博,李科,曾东,等. 基于语义关系的高血压临床指

- 南知识库构建 [J]. 西安: 中国数字医学, 2013, 8 (9): 64 - 67.
- 8 叶青. 高血压防治指南本体开发与电子文档构建 [D]. 西安: 第四军医大学, 2012.
- 9 刘何心. 糖尿病本体的构建与检索研究 [D]. 长春: 吉林大学, 2014.
- 10 方芳,徐天馥,沈同平. 糖尿病医案本体库的构建及应用研究 [J]. 中医学报, 2016, 31 (11): 1680 - 1683.
- 11 S Elsappagh, D Kwak , F Ali , KS Kwak. DMT0: a realistic ontology for standard diabetes mellitus treatment [J]. J Biomed Semantics, 2018, 9 (1): 8.
- 12 ES Shaker, E Mohammed, AM Riad. A Fuzzy - ontology - oriented Case - based Reasoning Framework for Semantic Diabetes Diagnosis [J]. Artif Intell Med, 2015, 65 (3): 179 - 208.
- 13 胡兆芹. 医学本体的构建及检索研究 [D]. 武汉: 华中科技大学, 2006.
- 14 Unified Medical Language System (UMLS) [EB/OL]. [2016 - 04 - 20]. <https://www.nlm.nih.gov/research/umls/>.
- 15 GO Database [EB/OL]. [2017 - 09 - 14]. <http://geneontology.org/page/go-database>.
- 16 Open GALEN [EB/OL]. [2017 - 09 - 14]. <http://www.opengalen.org>.
- 17 叶青. 高血压防治指南本体开发与电子文档构建 [D]. 西安: 第四军医大学, 2012.
- 18 杨春媛,李满生,朱云平. HuLDO: 人类肝脏疾病本体的构建及应用 [J]. 军事医学科学院院刊, 2015, 39 (2): 111 - 116.
- 19 王美琴,吴庆斌. 基于本体的医学知识库构建方法综述 [J]. 医学信息学杂志, 2017, 38 (3): 73 - 76.

(上接第 41 页)

- 7 Ratnasingham S, Hebert P D N. BOLD: The Barcode of Life Data System ([www. barcodinglife. org](http://www.barcodinglife.org)) [J]. Molecular Ecology Notes, 2007, 7 (3): 355 - 364.
- 8 陈士林,宋经. 中草药 DNA 条形码物种鉴定体系 [J]. 药学进展, 2017, 41 (2): 87 - 88.
- 9 Hui Y, Song J, Chang L, et al. Use of ITS2 Region as the Universal DNA Barcode for Plants and Animals [J]. Plos

- One, 2010, 5 (10): e13102.
- 10 薛瑞超. 中药综合数据库 [D]. 上海: 华东师范大学, 2013.
- 11 Perl. Perl 语言介绍 [EB/OL]. [2017 - 10 - 25]. [http://www. perl. org/about. html](http://www.perl.org/about.html).
- 12 Catalyst. Perl MVC framework [EB/OL]. [2017 - 11 - 06]. [http://www. catalystframework. org/](http://www.catalystframework.org/).