传染病本体构建及其在知识服务平台中的应用*

方 安 洪 娜 高东平 李亚子 池 慧
(中国医学科学院医学信息研究所 北京 100020)

【摘要】针对国内传染病本体构建中存在一致性差和共享困难等问题,在参照UMLS、SNOMED-CT及MeSH等知识体系的基础上构建传染病本体,开发传染病知识服务平台,采集多种来源的信息资源,利用传染病本体对其进行结构化处理和标注,并以知识罗盘形式展示概念之间、概念和文献之间的关系,为科研人员和公众提供知识服务。

【关键词】传染病本体 知识服务平台 知识罗盘
【分类号】G350

Infection Disease Ontology Construction and Application in the Knowledge Service Platform

Fang An Hong Na Gao Dongping Li Yazi Chi Hui
(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

【Abstract】To solve problems of constructing and applying the infection Ontology, such as inconsistent and difficult reuse, this paper constructs infection disease Ontology based on legacy knowledge system, including UMLS, SNOMED-CT and MeSH, then develops infection knowledge service platform collecting various sources of information, and utilizes above Ontology to normalize and annotate the information, and shows the relationship between the concepts and information by knowledge compass, supplying knowledge service to researchers and consumers.

【Keywords】Infectious Disease Ontology Knowledge service platform Knowledge compass

1 引 言

目前,本体理论研究已逐步成熟,并在医学领域得到了较好的应用,如系统化临床医学术语集(SNOMED-CT)[1]、医学领域顶层本体一体化医学语言系统(Unified Medical Language System,UMLS)[2]、基因本体(Gene Ontology,GO)[3]等。上述本体对医学领域的术语概念进行明确定义和规范化,深入挖掘并构建概念间的语义关系,形成具有医学领域特点的知识体系,在医学信息的采集、存储、分析、标注、系统间的语义互操作等方面发挥了重要作用。国外较为著名的传染病本体有公共卫生领域疫情监测本体(BioCaster)[4,5]、开放生物医学本体(Open Biomedical Ontologies,OBO)中的传染病本体(Infectious Disease Ontology,IDO)[6,7]等,用于规范传染病概念和消除语义冲突等,促进医学领域知识整合,提高系统间的互操作性。国内对于传染病本体构建研究尚不成熟,所构建的本体规模较小或只针对某一特定应用[8,9],主要表现在构建疾病本体时一致性较差、共享困难、难以继承和重用、

收稿日期:2011-11-24
收修改稿日期:2011-12-18
* 本文系国家科技重大专项课题"艾滋病和病毒性肝炎等重大传染病研究信息化技术平台研究"(项目编号:2009ZX10004-215)的研究成果之一。

研究成果转化到实际应用较慢、缺乏大规模应用等。

本文借鉴 UMLS 的语义网络构建传染病知识体系和语义关联关系构建传染病本体，使其具有良好的可扩展性，利用传染病本体对相关信息资源进行有效的组织，搭建传染病知识服务平台，提供传染病相关的知识服务。通过抽取 OBO，尤其是 IDO 的概念、翻译及细化构建了传染病本体概念与对象，利用概念与 UMLS 语义网络中实体的映射实现概念之间关系的构建，通过继承和计算机辅助技术降低了构建本体的成本，并将语义关系用于标注中，丰富了不同文献之间的关联关系。

2 基于本体的传染病知识服务平台架构

通过整合网络、文献及数据库等知识载体，实现重大传染病科技信息集成管理与综合服务；平台采用统一标准，建立多种共享机制，实现了国内外重大传染病防治相关信息收集、分析、加工、整理、共享等功能。平台总体架构如图 1 所示：

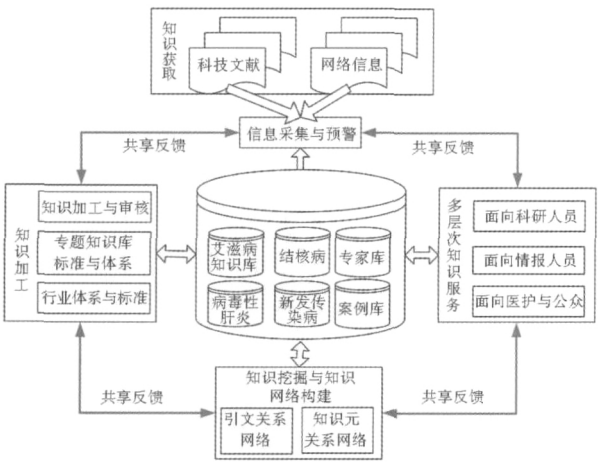

图 1 基于本体的传染病知识服务平台架构

基于本体的传染病知识服务平台主要由知识获取、知识加工、知识挖掘以及对外服务等模块构成，这些模块之间具有适当的耦合性，作为平台的重要组件协同作用，同时提供开放接口，供第三方调用。

(1) 知识获取模块

知识获取模块采集传染病领域的信息资源，加工资源库。传染病信息资源库内容包括中英文传染病相关新闻、政策法规、文献资源等相关信息。信息资源的来源包括传染病相关权威机构网站、网络数据库等。可针对特定网站进行定向数据采集，确保信息资源库质量。

知识获取模块按照资源库定义的分类和每一类数据库的结构，将信息资源存储到对应的资源库中。例如，将针对艾滋病的信息按照资源库结构分解为作者、主题、摘要、全文、发表时间、数据来源等信息。其中 CBM 是重要的知识来源之一，通过调用 CBM 的接口实现作者、主题、标题及关键词等过滤条件，返回相关记录，丰富知识库内容。

(2) 基于本体的知识加工模块

基于本体的知识加工模块主要实现对各种知识体系的管理、知识的加工、审核与标注等功能，知识体系管理实现各种知识体系、规范、标准的定义、获取和构建。知识体系的元数据具有较好的可移植性，支持导入和导出，并能建立和其他知识体系之间的对应关系。基于本体的知识加工模块可实现对特定的重大传染病书籍、文献、资料、相关病例、网页内容等知识进行抽取和标注，并规范知识加工流程，建立不同类型的知识库。本文将重点介绍基于知识体系构建传染病本体的过程。

知识加工与审核模块提供知识编辑和维护功能，支持知识节点的移动和可视化。支持多种类型知识库的创建以及多维、多级的结构维护；支持完整的知识生命周期管理过程。

(3) 知识挖掘模块

知识挖掘模块梳理分析知识加工模块产生的文献、知识单元(节点)之间的关系，构成系统的知识网络，从而有效地发掘和利用资源，描绘不同层面的知识网络结构。

(4) 对外服务模块

基于本体的传染病知识服务平台通过门户提供国内外传染病信息的动态浏览，用户可检索在线数据库，主要包括中国医学科学院医学信息研究所购买的电子期刊、文摘和开放获取类型数据库，根据定制的用户 IP 地址范围获取部分全文；对于没有全文的题录，可链接到全文传递系统 SinoMed 中，通过文献传递获取全文。用户可获得重大传染病相关的政策法规，也可以对艾滋病知识库、病毒性肝炎知识库、结核病知识库、新发传染病知识库中的信息进行浏览及下载。门户最大的特点是以知识罗盘形式展示知识库内容，提供重大传染病知识网络导航图，所有资源在加工时使用知识体系进行了标注，将信息资源内容和相应的知识点进行

了映射，从而可由一个关键词拓展到更丰富的信息，并以轮辐式的信息分支来展示。

知识获取模块从多种信息来源中摄取信息，是整个系统处理的对象，是对外提供服务的主要内容；知识加工模块是基础和支撑，首先通过构建标准的知识组织体系为系统内术语等知识表达的规范化提供重要依据，是计算机可理解各种信息和词汇的保障，加工和审核过程通过计算机实现知识获取模块中获取信息的预处理，辅助进行标注，最后通过人工干预对其进行调整，保证标注的正确性与权威性；知识挖掘和知识网络构建模块对标注内容的深入分析和利用，通过不同知识体系之间的映射与关联实现信息之间的映射和关联，从而形成知识网络；对外服务模块面向不同类型用户提供服务，将知识库中丰富的内容以不同的视图展现给用户，提高知识库内容的可视性，并通过知识罗盘和知识漫游尽可能全面地获取知识点的相关内容。

3 传染病本体构建

3.1 传染病本体概念及语义关系构建

传染病本体构建模块提供构建和利用传染病本体的工具，能够实现传染病本体中的类、对象属性、数据属性、实例的构建，并通过服务接口进行本体查询。

传染病本体中的医学概念类以及实例组成医学概念空间。虽然传染病相对于整个医学领域而言范围较小，但其涉及的医学基本概念的范围较广，所以在概念与语义关系的设计中，需在医学基本概念范畴内细化传染病相关概念和语义关系。在构建传染病本体概念时所有概念均来自于医学本体或词表，并结合现有资源诠释概念，充分借鉴原有知识体系中概念的定义及继承关系，并对实例进行细化。在细化的基础上建立实例与实例之间的关系，对于在参照的本体、词表或其他知识体系中未收录的概念，则遵循书籍中通用的名称，再由专家甄别，在传染病本体概念构建过程中参照的主要知识体系有：SNOMED－CT、MeSH、UMLS 以及 OBO 的 Infection Disease Ontology 等。

本文构建的传染病本体概念之间的语义关系继承了 UMLS 语义网络中的语义关系，以概念为核心，利用概念的语义网络构建相关概念的关联关系。为了描述基本概念类之间的语义关系，在 UMLS 的语义关系的基础上，添加了新的关系，如表 1 所示。

表 1 传染病本体概念语义关系及注释

关系				注释
继承（isa）*	是实例			是一种
	是一种			类与实例关系
				类间继承关系
相关关系（associated with）	物理相关于			有联系与物理相关
		是部分*		
		是组成*		
		包含*		
		连接于*		
		内部连接于*		
		是分支*		
		是属性*		
		是成员*		
	空间相关于			空间相关
		位于*		
		邻近*		
		包围*		
		横切*		
	功能相关于			功能相关
		影响		
			管理*	
			治疗*	
			干扰*	
			并发（并发于）*	表示一种疾病引起另一种疾病（表示一种疾病由另一种疾病引起）
			继发（继发于）*	表示一种疾病由另一种疾病引起（表示一种疾病由另一种疾病引起）
			交互*	
			预防*	
			易发人群	疾病易发于人群
		产生		
			导致*	表示一种疾病引起另一种疾病
			致病（由…致病）*	由病毒所致疾病
		执行		
			实践*	
		发生于		
			是过程*	
		使用*		
		是表现*		表示体征和疾病关系
		指出*		
		是结果*		
		鉴别		表示疾病的鉴别关系
		侵及		表示疾病发生后影响到另外部位
		分布于		表示疾病发生在身体的哪些部位
		原发于		表示疾病原发部位
	时间相关于			时间相关
		与并发*		
		先发于*		
	概念相关			概念相关
		是评价*		
		是程度*		

（注：表 1 中* 项目表示继承自 UMLS 语义关系，非* 表示在 UMLS 基础上扩展的语义关系。）

基本概念类间共定义了 50 种关系，其中保留了 UMLS 原有的 32 种关系。在这 32 种关系中，最基本的关系是“是一种”关系，如艾滋病是一种传染病，“是一种”关系描述了基本概念类之间隶属关系[17]。

本文构建的本体主要是中文本体，一方面是引入英文本体，将其翻译为中文，另一方面自建中文本体，自建本体主要为了细化英文本体以及补充英文中缺少的概念和对象，同时中文本体名都具有对应的英文名称。在借用 UMLS 语义关系时主要通过英文名称实现概念之间的关联，首先中文本体概念根据英文名称映射到 UMLS 的实体中，UMLS 实体之间已由语义关系建立了关联关系，通过中文概念→英文概念→UMLS 实体→UMLS 语义关系一系列的传递关系构建中文本体概念之间的语义关系网。

通过本体构建模块构建的本体包括艾滋病本体、结核病本体、肝炎病本体、新发传染病本体、机构本体、专家本体等。

3.2 艾滋病本体构建

根据艾滋病领域信息资源的特点以及对使用该系统的用户检索行为的分析，构建了艾滋病本体。该艾滋病本体用树状结构表示，其中每个节点表示定义的一个概念，每条边表示概念之间的关系，如图 2 所示：

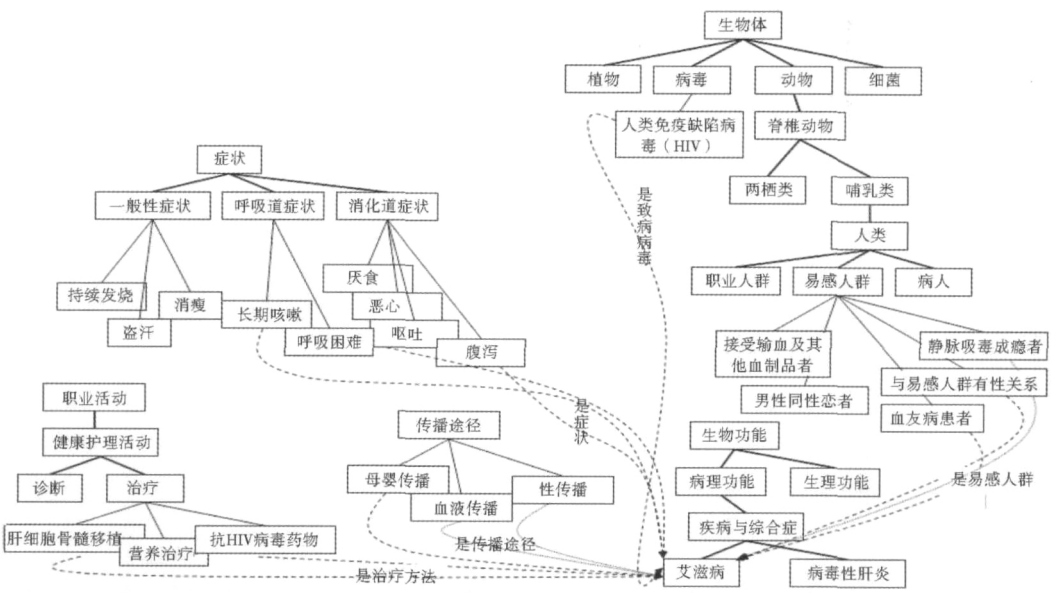

图 2　艾滋病本体的树状结构示意图(片段)

图 2 中树的根节点即第一层节点为“艾滋病”，与“艾滋病”关联的第二层节点包括“症状”、“治疗”、“传播途径”、“易感人群”、“世界艾滋病日”、“全球首个艾滋病病例”、“专家”；与“症状”关联的第三层节点包括“肿瘤”、“一般性症状”、“呼吸道症状”、“消化道症状”、“神经系统症状”、“皮肤和粘膜损害”；与“治疗”关联的第三层节点包括“营养治疗”、“肝细胞骨髓移植”、“水果治疗”、“抗 HIV 病毒药物”；与“传播途径”关联的第三层节点包括“母婴传播”、“血液传播”、“性传播”、“公用针具传播”；在“易感人群”下与其关联的第三层节点包括“血友病患者”、“静脉吸毒成瘾者”、“男性同性恋者”、“接受输血及其他血制品者”、“与易感人群有性关系”；在“世界艾滋病日”下与其关联的第三层节点包括“12 月 1 日”；在“全球首个艾滋病病例”下的第三层节点包括“1981、6 美国”；在“专家”下与其关联的第三层节点包括专家的姓名，每个专家的姓名对应该层的一个节点；在“肿瘤”下与其关联的第四层节点包括“卡波希式肿瘤”；在“一般性症状”下与其关联的第四层节点包括“持续发烧”、“盗汗”、“淋巴肿大”、“消瘦”；在“呼吸道症状”下与其关联的第四层节点包括“长期咳嗽”、“胸痛”、“呼吸困难”；在“消化道症状”下与其关联的第四层节点包括“厌食”、“恶心”、“呕吐”、“腹泻”；在“神经系统症状”下与其关联的第四层节点包括“头晕”、“头痛”、“反应迟钝”；在“皮肤和粘膜损害”下与其关联的第四层节点包括“弥漫性丘疹”、“带状疱疹”、“口腔和咽部粘膜炎症”；在

“抗HIV病毒药物”下与其关联的第四层节点包括“叠氮胸苷、双脱氧胞苷、双脱氧肌苷”；在专家姓名下与其关联的第四层节点包括该医生擅长的项目和所在医院名称。

传染病本体的概念之间关系带有权值，该权值可以根据概念之间的关联程度定义，关联越紧密权值越小。例如，通过对文献的分析统计，包含“艾滋病高危人群”的文献60%也会包含“同性恋者”，则将1与该比例之间的差值0.4定义为这两个概念之间的权值。如果将本体定义的概念作为图中的节点，概念之间的关系作为图2中的边，那么任意两个节点之间可以算出最短路径。用户进行检索的概念与检索出的文献所包含的概念之间可以求得最短路径值，该最短路径值可以作为计算检出文献相关度的依据。通过相关度对文献进行排序，可以使用户获得比较精准的信息。

3.3　基于传染病本体的资源标注

基于传染病本体对资源库内的信息进行机器辅助标注，传染病领域的特定概念可以与传统的分类方法结合进行文献信息的标注。例如，一篇艾滋病相关的文献按照自定义的概念，包含“治疗”、“易感人群”等入口，同时也可以按照传统的文献分类方法包含“期刊文献”、“网络”等入口，按照传统的对医学文献的分类包含“病理学”、“病原学”入口。

对传染病本体的定义是进行标注的依据，同一个文献可以包括多个标注入口，例如同时涉及“易感人群”和“治疗方法”的文献，可包含多个对应内容的标引。系统还可以为每个概念设置优选词，定义同义词表，包括每个概念的同义词及英文译名，以使文献的标注更加准确。本文资源标注对象主要针对中文和英文信息，系统中内置了多种知识体系，其中包括中英文MeSH词表以及UMLS等，通过知识组织体系相互之间的映射，例如同一个概念的不同表达方式、不同语种表现等方式进行关联和映射，从而实现中文和英文信息的标注，以及标注入口词的映射等功能。

4　基于本体的传染病知识服务平台应用

基于本体的传染病知识服务平台以门户的形式发布资源，为研究人员提供传染病领域最新动态、文献等资源，为公众提供传染病传播与预防的科普知识等，本系统已经在线提供服务，其运行的硬件环境为IBM X3650，操作系统为RedHat Linux AP5，数据库管理系统为Oracle 10g，网络服务器为Tomcat 6.0。在门户中知识罗盘提供了基于本体概念及关系的可视化内容展示。关于艾滋病的内容展示如图3所示：

图3　知识罗盘

图3中知识罗盘展示出了与艾滋病相关的致病病毒、高危人群、传播途径、疗法等属性与关系。点击其中的一个知识点，如“传播途径”，知识罗盘会继续显示以“传播途径”为关注点的新的关联节点，包括母婴传播、血液传播等，用户可逐级进行点击。随着用户点击知识点的变化，知识罗盘及知识罗盘下面的文献列表均会产生变化。同时，页面左侧“病原学”、“病理学”等也会根据用户点击的知识点，重新计算文献书目。用户可点击右上角“知识罗盘使用教程”，观看知识罗盘使用说明视频，以更好地了解、应用知识罗盘。罗盘左边展示出了艾滋病分类信息，用户可根据分类点击查看，例如用户点击罗盘中艾滋病节点后再点击左侧“病原学”，则在页面中会显示出艾滋病知识库中病原学相关的信息列表。

5　结　语

本文针对传染病本体构建中存在的不足和用户对传染病知识共建共享的迫切需求，构建了传染病知识本体，对传染病概念的语义一致性等方面进行规范化表达。利用传染病本体搭建了知识服务平台，采集多种途径的资源，并提供基于本体的信息资源标注功能，向不同用户提供知识浏览和知识检索等相关的知识服务，实现了对传染病信息资源的知识组织和集成共享服务。

参考文献：

[1] SNOMED CT [EB/OL]. [2011-07-21]. http://www.nlm.

nih. gov/research/umls/Snomed/snomed_main. html.
[2] UMLS [EB/OL]. [2011 - 07 - 21]. http://www. nlm. nih. gov/research/umls/quickstart. html.
[3] Gene Ontology [EB/OL]. [2011 - 07 - 21]. http://www. gene-ontology. org/.
[4] BioCaster [EB/OL]. [2011 - 07 - 21]. http://biocaster. nii. ac. jp/.
[5] Collier N, Kawazoe A, Jin L, et al. A Multilingual Ontology for Infectious Disease Surveillance: Rationale, Design and Challenges [J]. *Language Resources and Evaluation*, 2007, 40(3 - 4): 405 - 413.
[6] Open Biological and Biomedical Ontologies [EB/OL]. [2011 - 07 - 21]. http://obofoundry. org.
[7] Sintchenko V. Infectious Disease Informatics [M]. New York: Springer, 2010: 389.
[8] 方安,王惠临,王军辉,等. 临床疾病领域本体构建方法研究——以手足口病本体为例[J]. *情报杂志*,2009,28(11):180 - 184.
[9] 高珊,王文俊,杜磊,等. 传染病应急案例共享本体模型研究[J]. *计算机应用*,2010,30(11):2924 - 2927.

(作者 E - mail: chi. hui@ imicams. ac. cn)

Elsevier 和 Ex Libris 合作在 Scopus 和 ScienceDirect 中推出 bX 推荐服务

Ex Libris 集团宣布与 Elsevier 达成合作协议,为双方客户提供集成到 SciVerse 界面的 bX 文章推荐应用。该应用将使得 Scopus 和 ScienceDirect 的用户能够在阅读从 Scopus 和 ScienceDirect 数据库中找到文章时查看 bX 推荐,即根据研究人员使用数据做出进一步阅读建议。

Elsevier SciVerse Scopus 是世界上最大的同行评审文献和高质量网络资源的摘要和引文数据库。Scopus 包含4 550万条记录,来自世界各地5 000个出版商,其中 70% 的记录都有摘要信息。Elsevier SciVerse ScienceDirect 是一家领先的全文科学数据库,提供超过 950 万期刊文章和图书章节,来源于2 500多家同行评审期刊和11 000多本图书。

bX 是第一个提供根据研究人员的使用数据生成推荐的服务,只关注学术领域,已应用于 1 000 多家学术机构。文章推荐基于对数亿份使用日志的分析,这些使用日志是由世界各地 Ex Libris SFX OpenURL 链接解析器用户中的 bX 应用订阅者贡献。bX 是基于云的,能够集成到 SFX 和 Primo 发现和交付解决方案之中,还能够使用应用程序接口集成到其他系统之中。

Elsevier 应用市场和开发网络副总裁 Rafael Side 指出 "我们很高兴在 SciVerse 应用中推出 bX 推荐应用,也很高兴看到这一举动将会为 SciVerse ScienceDirect 和 Scopus 的用户带来附加价值。我们期待为大家带来基于融合了全球研究人员集体智慧的实际使用数据的学术文章推荐应用。"

"我们很高兴能够与 Elsevier 合作推出这项突破性的成果。"Ex Libris 发现和交付解决方案副总裁 David Beychok 指出,"bX 服务在发现服务的舞台上是独一无二的,因为它能够在用户通过自己搜索找到的材料之外,还为用户提供高度相关的材料。这一重要的合作关系将使得 bX 推荐能够应用于更多平台上的更多的研究人员。"

Ex Libris 致力于与客户保持密切合作关系并为他们提供创造性的解决方案。它使得学术图书馆、国家图书馆和研究型图书馆最大限度地提高了生产力和工作效率,同时,大大提高了用户体验。通过使用户发现并获取他们所需的信息,图书馆确保了其作为通往知识的桥梁的地位。

(编译自: http://www. exlibrisgroup. com/de/files/Germany/PressRelease/2011/October/Elsevier_ExLibris. pdf)

(本刊讯)