多维度临床知识组织方法及其知识库构建与平台开发

俞思伟 范 昊 王 菲

（1 武汉大学中南医院 （武汉大学信息管理学院 武汉 430072）
2 武汉大学信息资源研究中心 武汉 430072）

〔摘要〕 以语义技术中的本体分子理论为基础，研究如何利用本体分子对不同维度下语义内容产生变化的临床知识进行描述和组织，完成临床数据源的知识抽取、临床本体分子库构建及半自动建库工具开发，建立临床本体分子库综合应用平台，提出多维度临床知识组织与应用方法体系，设计并实现临床病程动态可视化系统、临床决策支持系统。

〔关键词〕 本体分子；多维度知识管理；临床知识组织与应用；平台开发
〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2016.04.001

Organization Method of Multidimensional Clinical Knowledge and Construction of Its Repository and Platform Development

YU Si-wei, 1Zhongnan Hospital of Wuhan University, 2The Center for the Studies of Information Resources of Wuhan University, Wuhan 430072, China; FAN Hao, WANG Fei, Information Management School of Wuhan University, Wuhan 430072, China

〔**Abstract**〕 Based on the ontology molecule theory in semantic technology, the paper studies how to utilize the ontology molecules to describe and organize clinical knowledge of which the semantic contents change in different dimensions. It extracts knowledge from clinical data sources, constructs clinical ontology molecules and develops semiautomatic tools for database building. Also, it establishes the comprehensive application platform of clinical ontology molecule base, puts forward the organization and application method system of multidimensional clinical knowledge, designs and implements the dynamic visual system of clinical course and Clinical Decision Supporting System (CDSS).

〔**Keywords**〕 Ontology molecule; Multidimensional knowledge management; Clinical knowledge organization and application; Platform development

1 引言

1.1 知识组织

知识组织是在传统文献信息环境下发展起来的信息组织和利用方法，1929 年由美国图书馆学家布利斯首次提出[1]，在百余年的理论与应用发展过程中，形成并完善以满足文献信息单元检索需求的知识组织工具，即分类法、主题词表等。目前的网络数字环境下，知识组织理论与系统需要进一步发展和创新，以满足海量异质异构且更新加快的数字化知识组织的要求。本体资源描述框架（Resource Description Framework，RDF）/网络本体语言（Ontology Web Language，OWL）结构作为万维网联盟（World Wide Web Consontium，W3C）公布的语义网内容组织的标准，以三元组主谓宾（<Object,

〔修回日期〕 2015-09-21
〔作者简介〕 俞思伟，博士，研究员，主编、参编教材和专著5 部，发表论文30 余篇。

Predicate, Object>）形式来表示更接近于自然语言的语义内容。由于本体知识模型良好的表达性、推理性和复用性，具有语义的形式化表示功能，还可作为系统互操作和分析的基础，在知识组织的理论和实践方面都具有十分重要的意义[2]。近几年，出现了很多关于将传统知识组织工具进行本体化改造的研究和成果，足见在知识组织和知识管理领域的学术意义和应用价值。

1.2 本体

本体在医学领域的应用已相当广泛[3]，相关的本体构建技术也十分成熟。国际著名的系统化临床医学术语集 SNOMED-CT 是医学领域成功利用本体的研究成果。SNOMED-CT 以现代医学理论为指导，将收录的与临床医疗相关的概念进行逻辑定义，保留了近 37 万条具有唯一定义的概念，将这些概念较为合理地分为 18 大类，其中以临床所见（Clinical-finding）和操作（Procedure/Intervention）为最重要的两类，而支持和解释临床所见与操作的概念，如发现与操作的工具、方法、部位、实体，则相应归入观察对象、身体结构、有机体等其他大类。这种概念分类方法，不仅符合临床医生的思考方式，也符合临床医疗的实际流程，更体现了现代医学的疾病认识观（即其本体论）。

1.3 本体的三元组结构

它是对客观世界的全部或某一部分的概念化和结构化的抽象，主要是从知识客体或对象出发，通过建立知识客体之间的概念联系和等级关系，将对知识客体的揭示深入到知识内涵的层次并实现对其内在联系的推理。相对于传统的知识组织理论和系统，本体具有表达性和推理的优势。但是由于简单的 RDF 三元组结构，本体仅能解决静态知识问题，即不随任何情景维度的变化而变化的知识内容的描述和组织。对于不同角度且语义内涵不同的知识。对于不同时间、不同领域和不同粒度等多维度临床知识，本体技术无法通过描述逻辑进行表达。这对本体在临床知识领域的应用造成了障碍。OWL 是在逻辑推理的需求中，使用一系列的描述逻辑词表，建立分类、约束等推理机制，在知识的描述能力上仍等同于基于 RDF 的三元组描述，也不能解决知识语义内涵变化的描述问题[4]。因此，多维度临床知识组织问题，也是 RDF/OWL 三元组结构的本体技术无法解决的问题。

1.4 本体分子理论

即针对本体技术遇到的知识演化问题而提出的解决方案，是在本体理论的基础之上，结合描述逻辑、图论等相关理论，用于解决动态知识组织管理和控制的问题。本体分子模型可以为知识管理中的本体提供动态响应变化的能力，即演化能力。本体分子是实现本体演化的途径，本体演化则是本体分子动态变化的结果。本体分子不仅探索知识演化理论问题的解决方法，还可在多维度知识组织与管理系统中得到应用。

2 临床本体分子建模方法

2.1 临床本体建模方法

2.1.1 国际系统化临床医学术语集 SNOMED-CT

将术语分为概念完全指定的名称（Fully Specified Name）、概念的首选术语（Preferred Term）、同义术语（Synonym）3 类，理清了其收录的近 100 万条术语与 37 万条概念间的对应关系，较为理想地解决了由于一个概念对应多条术语或一条术语对应多种概念而可能引起的歧义。SNOMED-CT 还利用了 40 余个连接词，将具有内在关系的概念两两联结起来，组成能够明确描述一个临床事件，形如概念+连接词+概念的三元组结构化短句。用这种方法建立起数量高达 146 万组的语义关连，基本上涵盖了现代临床医学事件描述的需要，提高了计算机识别处理能力。SNOMED-CT 发布多年来，通过世界上 30 余个国家和地区引进使用的反馈信息看，SNOMED-CT 已在电子病历书写、电子处方及医嘱录入、疾病样本、化验检验报告、文献编码、临床研究等诸多方面得到广泛应用，而且未发现其结构和内容具有影响使用的重大缺陷。

2.1.2 《医学主题词表》（Medical Subject

Headings, MeSH) 美国国立医学图书馆编制的权威性主题词表。它是一部规范化的可扩充的动态性叙词表。美国国立医学图书馆以它作为生物医学标引的依据，编制《医学索引》（Index Medicus）及建立计算机文献联机检索系统 MEDLINE 数据库。MeSH 汇集约 18 000 多个医学主题词。

2.1.3 一体化医学语言系统（Unified Medical Language System, UMLS） 又称为统一医学语言系统，是对生物医学科学领域内许多受控词表的一部纲目式汇编。UMLS 提供位于这些词表之间的映射结构，使这些不同的术语系统之间能够彼此转换；同时，UMLS 也被看作是生物医学概念所构成的一部广泛全面的叙词表和本体。UMLS 还进一步提供若干适用于自然语言处理的工具，旨在供医学信息学领域的系统开发人员使用。当然还有很多其他临床医学的本体模型，如国际疾病分类 ICD、RxNorm 临床药物标准命名系统等。

2.2 本体分子定义与特征

在解决可变知识管理和控制问题时，本体属性的变化类似于物理分子的特性，即属性存在不变和可变的特征，不变的部分正如分子的原子核，核变化其物质特性也变化，同样地，根本属性改变，原来的本体就不存在了。本体分子[5]的概念由此而产生。本体分子具体是指在本体基本元素（本体实例、三元组）基础之上，用唯一标识符标注，根据语义或者语用划分、无缺失、最小冗余的本体知识单元。本体分子是在本体基本元素和本体库之间的一个平衡点，它使得相对不同粒度知识管理成为可能。在形式化描述中表示不变知识的是本体分子核子 c（core），表述可变知识的本体分子离子 o（outer）。本体分子核子 c 为：c = func_ c（id_ c, g_ c），id_ c ∈ U，g_ c ∈ G。其中 id_ c 为本体分子核子 c 的唯一标识符，g_ c 表示本体分子核子的范围，func_ c 是 id_ c 和 g_ c 的映射函数。同理：本体分子中一个离子 o 为：o = func_ o（id_ o, g_ o），id_ o ∈ U，g_ o ∈ G。其中 id_ o 为本体分子一个离子 o 的唯一标识符，g_ o 表示本体分子一个离子的范围，func_ o 是 id_ o 和 g_ o 的映射函数。

2.3 本体分子临床领域应用

本体分子是解决传统本体中知识动态性表达而产生的。传统本体的建模技术表达领域静态知识具有优势，而在表达动态知识的过程中存在困难。例如在不同的时间空间等维度下，知识体系会不断发生演变，环境变化导致本体知识库也需要做出相应变化，由此提出本体分子理论。因此，对于动态知识的建模相关领域，本体分子建模技术是一种不错的选择方案。在临床医学领域，存在大量的本体知识库，在具体的临床实践中，患者的病情是动态变化的，可能好转也可能恶化。为了捕获患者的病情发展状况，需要对患者临床医疗指标进行跟踪记录，基于这些动态的临床诊疗数据，使用本体分子技术进行语义表达与建模，就可以构建出患者整个患病过程的周期模型，使用语义分析与挖掘技术，以及查询与推理技术就可以对患者的病情发展进行可视化、预测，找出影响病情发展的关键因素，为下一步的治疗提供可信的决策支持等，进而促进患者早日康复和医疗技术的快速发展[6]。

2.4 临床本体分子建模

实现基于临床本体分子模型的医疗辅助与决策的核心问题是临床本体分子知识库的建立，该过程分为 3 个阶段。临床诊断信息捕获与收集阶段，对各种医疗设备的实时监控记录信息、病人的病历数据、医院诊疗人员的病情记录数据等相关的医疗数据进行收集、电子化处理；临床诊疗信息的标准化与形式化，原始的信息比较粗糙繁杂，需要进行结构化和标准化处理，借助文本处理以及本体建模技术，剔除数据中的错误与冗余，引入规范的医学数据，便于临床信息的处理与互操作；使用本体分子建模技术，将前一个阶段处理得到结构化语义数据中的动态变化部分使用本体分子建模词汇进行动态语义表达，完成本体分子临床知识库的构建。这些本体分子知识库构建成一个庞大的语义网络，基于该语义网络能够开展临床疾病的趋势预测、用药指导等临床辅助决策。

3 多维度临床知识库构建方法

3.1 构建过程及其 3 个层次

依据临床本体分子模型，可按照语义或语用的标准设计多维度临床知识库结构及其应用平台，其构建过程，见图 1，多维度临床知识库的构建方法是自底向上的，共分为临床数据源的知识抽取、领域临床本体分子库构建和临床本体分子库应用平台构建 3 个层次。

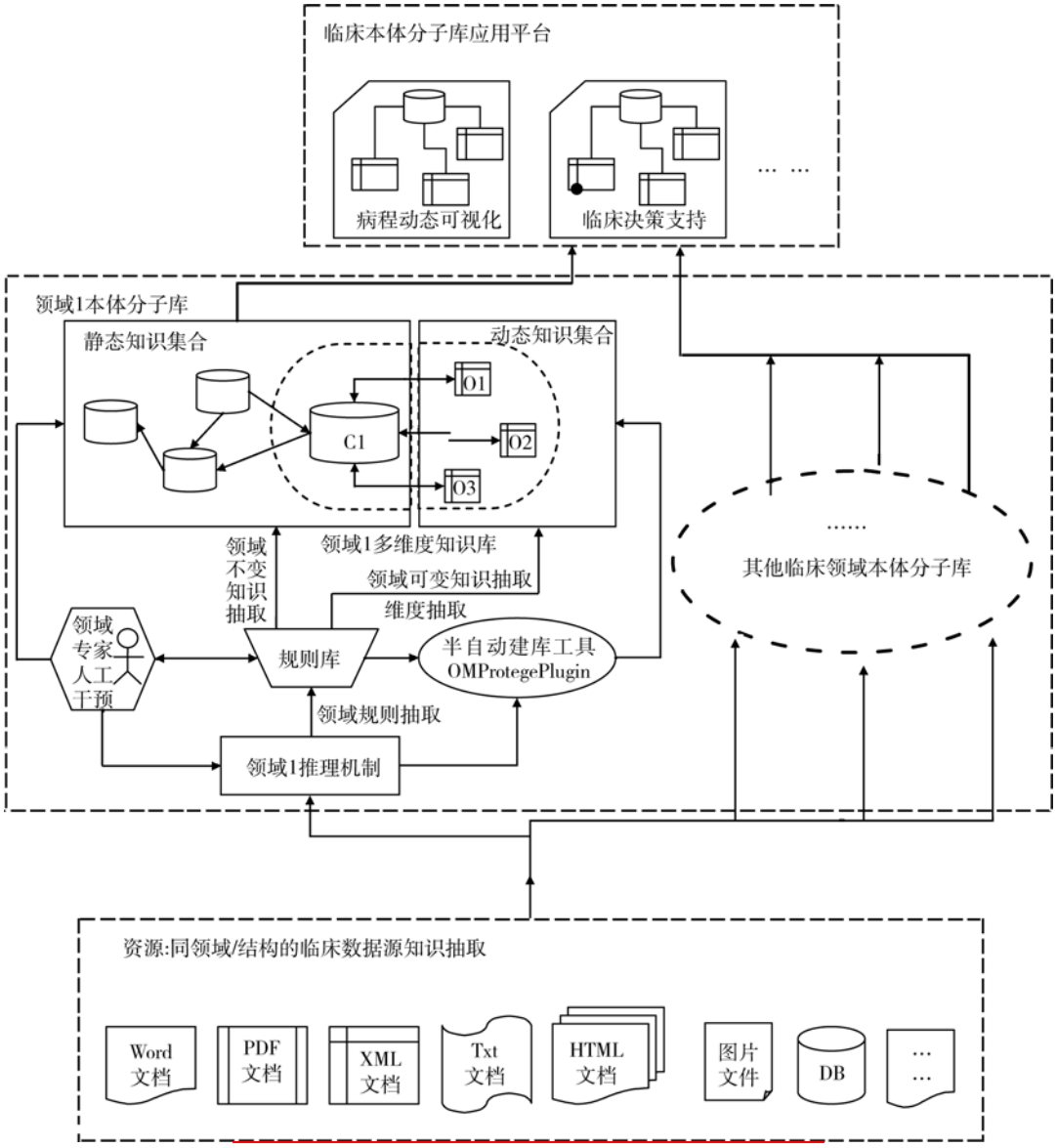

图 1 多维度临床知识库及其应用平台

首先是知识抽取层，对不同领域，或不同结构的临床数据来源进行收集，为进一步的领域分析和复杂语义关系的挖掘和抽象奠定数据处理的基础。几乎所有的结构化、半结构化、非结构化的数据都在此范围之内。其次是针对具体领域的临床本体分子库的构建，主要包括领域中的数据对象的规则抽取和知识抽取，从而形成不同领域的规则库和不同维度的知识库，其实现手段是半自动建库工具和领域专家人工干预相结合。该步骤中最重要的结果是领域规则库和多维度知识库，而半自动建库工具的设计与应用也会直接影响多维知识库建设的效率和准确性。第 3 层基于临床本体分子库的应用平台构建，这是临床知识组织应用的实现，也是最终目标和结果，是由不同领域临床本体分子库组成的松散

耦合的知识资源体系，是针对临床不同领域的异质异构知识资源的大规模集成和应用。临床本体分子库应用平台可以利用“import”机制，开发满足不同知识需求的应用程序接口，即病程动态可视化系统、临床决策支持系统等，形成既可整体使用又可以根据特定需求调用部分知识库的临床本体分子应用体系。

3.2 临床领域知识抽取的内容

针对相对复杂的临床领域知识组织问题，在知识抽取过程中，除了提取与建立临床本体库相同的不变静态知识（即不随任何条件的变化而变化的知识）外，还需要对相对的、可变的多维度知识内容进行抽取。确定不变知识和可变知识的范围与抽取方法，是正确进行基于本体分子的多维度知识组织的关键。当然这种范围与方法的确定也取决于临床领域知识管理的具体语用需求[7]。

3.3 确定临床本体分子维度

3.3.1 维度的概念 在临床领域可变知识的抽取过程中，确定临床知识组织的维度是非常关键的步骤。维度是基于本体分子的多维度知识库中衡量知识是否为真的基本手段，是表达知识成立条件的基本工具。若知识所对应的维度与用户查询中的维度限定相匹配，则知识为真。否则，知识的正确性不受保障。通俗的理解，维度就是判断知识是否为真的另一个变量，只有当这个变量为真时，才能保证知识的正确性。例如“糖皮质激素为克罗恩病治疗药物”，这条语句并不是永远成立，只有在克罗恩病急性炎症期诱导缓解才有效。影响到这条语句成立与否的关键变量是时间，因此时间就是这条语句的维度。维度的确定有利于将可变知识、相对知识清晰化，对语句添加维度的限定使知识表达更准确。

3.3.2 维度的确定过程 维度的确定过程也是基于本体分子的多维度知识组织过程中不变知识（核子）和可变知识（离子）的区分过程。若知识库中定义的所有维度都不影响到三元组表达知识的正确性，则该三元组表达的为绝对的不变知识。基于本体分子的多维度临床知识组织的维度提取应符合两个原则：冲突消解和标准化。冲突消解是维度抽取应与本体构建过程中的描述冲突记录相结合。另一方面，维度的提取应符合标准化的原则，体现在两个方面：细分和重用。细分是指应该将知识库的维度进行尽可能的拆分，如条件“早期大肠癌”，一种解决方式是将时空限定融合在一个维度里。这样做的缺点是不利于维度的复用，而且会导致维度实例数量的急剧增长。更合理的做法是设立两个维度，一个用来进行时间限定，另一个用来进行疾病限定。其中时间值为早期，地点值为大肠癌，两个维度都被添加进维度容器来对知识的陈述进行限定。这样“早期”和“大肠癌”两个维度被重用的概率就大大增加了。维度确定的方法是结合通用维度（如时间、地点、学科等）的基础上，针对可能产生相对复杂的可变知识的陈述，结合领域专家的干预确定维度类型和数量，不同的领域可能会提出不同的知识维度。实际上基于本体分子的知识组织维度主要是根据知识组织的需求和特定的领域问题来设计的，无论是单一维度，还是复合的多个维度都可以较好地利用维度类、维度和维度容器的设计和配置表现出来。

3.4 临床本体分子半自动建库

本体分子半自动建库工具要比本体相关工具的开发困难，但可以借鉴本体半自动构建的思路。全手工建设本体库的工作量太大，受到诸多方面的限制。另一方面本体自身的复杂性和严格的建模要求，使得其创建离不开领域专家的参与，目前无法实现完全自动化的本体知识库构建已经成为共识[8]。在借鉴某些本体知识库构建工具的基础上进行二次开发，生成基于本体分子的多维度知识库的半自动建库工具 OMProtégéPlugin，本质上是一个 Protégé 插件，通过扩展 ProtégéOWLPlugin 的方式来提供本体分子建库的功能，OMProtegePlugin 开发出实验性的版本，可以实现部分功能，如多维度知识库中不变知识三元组和可变知识三元组的实例添加，根据特定领域规则自动生成 ID 将多维度知识库中的核（不变知识）与离子（可变知识）相关联等。另外半自动构建本体分子库的平台 OMConPlatform 也初步开发完成并得到了一定的应用，减少了大数据环境下人工判断的工作量[9]。

4 临床本体分子库综合应用平台开发

4.1 概述

由于临床本体分子库不仅兼容了不变的临床知识组织，对更为复杂的动态知识、相对知识和多粒度知识进行了抽象和表达，因此基于本体分子理论的多维度临床知识组织体系是对临床知识语义内涵深层次的挖掘和更精确的表达。其中，多维度的临床本体分子库的集合并不是各临床领域知识的简单叠加，而是针对不同领域的知识管理需求，形成多维度的知识体系。基于临床本体分子库的多维度知识库体系既可以整体综合运用，也可以根据具体需求，通过 API 调用并生成某一个或多个领域的多维度知识体系加以应用。

4.2 病程动态可视化系统

由于临床知识的复杂性、相对性、多维度和动态化，目前在临床可视化方面，医学影像检查的资料可以展示病程的形态变化，但还无法与治疗过程的文字描述和数据关联起来完整地展示病程的动态发展。基于临床本体分子的病程动态可视化系统可以完整的、全方位多媒体的展示病程的动态发展。

4.3 临床决策支持系统

临床决策支持系统是指将临床数据作为输入信息，将推论结果作为输出，有助于临床医生决策并被用户认为具有一定“智能”的软件。能够有效提高医疗质量和效率、减少医疗差错、降低医疗费用[10]。尽管临床决策支持系统有很多优点，真正能为医生所接受并投入实际临床使用的为数不多，其主要原因是：不确定知识的表示与推理的困难、知识更新的困难、知识库的透明性问题；其次，与其他医学信息系统以及医生的工作模式难以融合，无法从大量临床数据的处理中获得决策的参考依据。本体分子理论的应用正是针对这些问题的有效解决方法，基于本体分子理论和大数据处理的临床决策支持系统将更好、更精准地满足临床医生诊疗的知识服务需求。

5 结语

本文以本体分子理论为基础，对临床知识采用多维度的组织与描述方法，探讨如何利用本体分子理论对不同维度下语义内容产生变化的临床知识进行描述和组织，为多维度临床知识管理与利用提供理论依据和应用基础。在此基础上构建基于本体分子理论的多维度临床知识组织模型，完成临床数据源的知识抽取、临床本体分子库构建及半自动建库工具开发，建立临床本体分子库综合应用平台，提出多维度临床知识组织与应用方法体系，设计并实现临床病程动态可视化系统、临床决策支持系统。其意义在于扩展和完善本体分子的理论体系和应用领域，丰富知识组织的理论方法，对如何在语义网环境下进行多维度临床知识组织与应用提供新的理论依据与适用方法。

参考文献

1 Gail Hodge, Linda Hill, et al. Next Generation Knowledge Organization Systems: integration challenges and strategies [C]. ACM New York, NY, USA: Proceedings of the 5th ACM/IEEE - CS Joint Conference on Digital Libraries, 2005.

2 董慧. 本体与数字图书馆 [M]. 武汉：武汉大学出版社，2008.

3 李劲松，黄智生. 生物医学语义技术 [M]. 杭州：浙江大学出版社，2012.

4 吴刚，唐杰，李涓子，等. 细粒度语义网检索的研究 [J]. 清华大学学报（自然科学版），2005，(9)：1865-1872.

5 董慧，姜赢，高巾，等. 基于数字图书馆的本体演化和知识管理研究Ⅰ-本体分子理论 [J]. 情报学报，2009，(3)：323-330.

6 俞思伟，姜赢，董慧. 基于本体的面向多用户知识管理模型研究 [J]. 医学信息学杂志，2009，(12)：46-50.

7 董慧，徐雷，王菲，等. 语义分析系统研究（I）——史籍语义分析流程 [J]. 情报学报，2014，(2)：183-194.

8 苗壮，张亚非，陆建江. 本体的半自动构建技术 [J]. 解放军理工大学学报(自然科学版)，2006，(10)：426-431.

9 董慧，聂曼曼. 中文本体的半自动构建研究 [J]. 情报杂志，2009，(11)：145-149.

10 俞思伟. 临床决策知识管理中语义技术的应用研究 [J]. 中国数字医学，2013，(4)：25-28.