·信息工作·

知识组织系统中关系模式的应用比较

牟冬梅 王丽伟 (吉林大学公共卫生学院 吉林长春 130022)

摘 要:文章以医学数字信息资源知识组织为案例,对领域本体、目前常用的医学数字信息资源知识组织的工具《医学主题词表》和统一医学语言系统的语义网络进行对比研究,深入分析了三者的联系和区别,在此基础上讨论了知识组织体系中的关系模式在知识组织中的应用特点。信息资源知识组织有赖于领域本体的完善,而领域本体高质高效的构建和应用需充分发扬和整合主题词表和语义网络的成果,即利用叙词表已经规范了的专业概念和语义网络定义的语义类型。

关键词:医学主题词表 语义网络 本体 知识组织

中图分类号:G250.76 文献标码:A 文章编号:1003-6938(2006)05-0058-04

Comparative Research on Practice of Relational Schema in Knowledge Organization

Mu Dongmei Wang Liwei (Public Health School of Jilin University, Changchun, Jilin, 130021)

Abstract: The paper takes adopted knowledge organization of medical digital information resource as a case, conducts comparative research and deeply analyzed the relations and differences among domain ontology, the current tools (MeSH- Medical Subject Headings and UMLS Semantic Network) in common use in medical digital information resource organization. On the basis of that, it discussed the characteristics of relational schema in knowledge organization system when applied in knowledge organization. Perfection of domain ontology is the premise of knowledge organization of information resources. Construction and application of high-quality medical ontology need to fully carry forward and integrate the advantages of first two tools, namely use the standard professional concepts in Subject Headings and semantic types defined in semantic network.

Key words: MeSH, semantic network, ontology, knowledge organization

CLC number: G250.76 Document code: A Article ID: 1003-6938(2006)05-0058-04

1 导论

知识组织是当前理论研究的热点话题,图书情报领域建立“数字图书馆”,人工智能领域建立知识库,信息科学领域建立语义互联网都离不开对知识组织的探讨。研究者发现这些领域的研究核心都是围绕着对数字信息资源再组织展开研究,试图将无序的、分散的特定信息资源,根据一定的原则和方法,整合成为有序、集中、便于检索的知识,以方便知识的提供、利用和传播,达到知识服务的层次。美国加州大学 Linda Hill 等人提出了数字图书馆的知识组织系统,主要由系统分类和大致分组模式(包括大致分组归类的类表、系统分类的分类表、标题表、知识分类表)、元数据的系统模式

基金项目:本文系吉林大学公共卫生学院攀登计划项目。

收稿日期:2006-03-05;责任编辑:王景发

（包括指南、地名辞典）、关系模式（包括本体、语义网络、叙词表）和词汇单（规范文档、字典、术语表）组成。[1] 从该知识组织系统可以看出，关系模式是数字图书馆的框架。另外，在 Tim Benners- Lee 提出的语义网七层体系结构中，本体（ontology）位于低层的 Unicode 字符集和 XML 语法结构之上，位于逻辑层和验证层之下，既是基于 XML，同时又为语义网的逻辑推理和验证功能提供基础，可以说本体是语义网结构中的关键部分。[2] 从这两个侧面可以看出关系模式对知识组织的重要作用，本文就关系模式进行讨论，对本体、语义网络、叙词表在医学数字信息资源知识组织中的应用进行比较，认为信息资源知识组织有赖于领域本体的完善，而领域本体高质高效的应用需充分发扬和整合主题词表和语义网络已经取得的成果。

2 三种关系模式在医学数字信息资源知识组织中的应用特点

2.1 叙词表

《医学主题词表》（Medical Subject Headings，简称 MeSH）是美国国立医学图书馆编纂的一部大型医学专业叙词表。[3] MeSH 的基本要素是叙词（亦称主题词），其理论依据是建立在叙词性质基础上的，在编制上吸取了多种情报检索语言的原理和方法，在医学数字信息资源的知识组织中具有明显特征。

2.1.1 规范的科学用语，灵活的使用方法

MeSH 中的术语是经过严格规范的科学语言，即进行了严格的同义规范、词义规范、词类规范、词型规范，明确词的含义及所涉及的范围，使得每个叙词在词语的形式和语义上只能有一个概念，不允许一词多义和一义多词，规范了标引人员和编目人员的使用，减少人为造成的误差。最为标引人员熟悉的是Neoplasm这个科学用词，它包含了 Tumors、Tumor、Benign Neoplasms、Neoplasms、Benign、Benign Neoplasm、Neoplasm、Benign、Cancer、Cancers 这些同义词，这些同义词以入口词的形式出现在 MeSH 表中，用 MeSH 对医学知识进行标引、组织，表达肿瘤概念的只能是科学语言 neoplasm，不是平时常用的 tumor、cancer 等自然语言。但当进行计算机检索时，入口词自动转换为优先选用的叙词，从中可以看出，MeSH 对于数字信息资源组织的独具匠心。

2.1.2 完美的聚类方式

MeSH 的主题字顺表和树状结构表即词汇的字顺表示和词汇的系统显示是两个互相补充的部分，是以事物聚类和以学科聚类的完美结合。[4] 主题字顺表是将所有的主题词、副主题词、非主题词全部按字顺排列，每个主题词下设该主题词建立的年代、树状结构编码、历史注释及各种参照系统。MeSH 的树状结构表，把所有的主题词按词的范畴和学科属性分为十五大类，有的大类按需要再依次划分为一级类、二级类，最多分至九级。树状结构的每个类目中，主题词按等级从上位词到下位词逐级编排，表达主题词之间逻辑上的隶属关系。

2.1.3 特殊的词组性主题词设置

MeSH 的词组性主题词以两种形式出现，一是按正常的自然语序排列的顺装形式，如：Hypothalamic Disease；另一种是将名词型中心词提前，形容词类的修饰语置后，中间用逗号相隔的倒装形式。[5] 它相对集中地排列一组中心概念相同的主题词，在一定程度上弥补了美国医学索引没有分类检索途径的不足。

2.1.4 简单的语义关系

MeSH 表的参照系统包括用代参照（“See”和“X”参照），即非主题词见主题词，非主题词以入口词的形式出现，入口词包括同义词、近义词、缩写、不同的拼写形式及其他用代形式。如 Outbreak See Disease Outbreaks。以此来处理主题词与非主题词之间的相互关系；此外还有一种 consider also 参照系统用来引见意义上相近、但词根不同的词，如 BRAIN consider also terms at CEREBR- and ENCEPHAL- .；MeSH 在主题参照系统中删除了 XU（属分）参照；添加了主题词/副主题词组配参照，该参照系统严格规定用户不得进行无效的主题词/副主题词组配，而且当有先组的主题词时，用先组主题词而不用主题词/副主题词组配来表达同一概念。如 Abdomen/Injuries，See ABDOMINAL INJURIES。MeSH 从几个简单的方面揭示叙词间的相互关系，提示知识的内涵，对知识进行组织管理。

2.2 医学语义网络

1970 年西蒙正式提出了语义网络（Semantic Network）的概念，按照数学的观点，语义网络是一种带有标记的有向图。节点表示物理实体、概念或状态，连接节点的“边”用于表示实体间的关系。[6] 由于语义网络表示知识简洁、直观，因此在专家系统、自然语言理解等领域获得了广泛的应用。

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）。[7] 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。UMLS 中的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。2005 年版的语义网络包括 134 种语义类型和 54 个语义关系，语义网络给出了这些语义类型和语义关系的定义。

2.2.1　概念高度结构化

UMLS 用语义网络作为表示医学知识的手段，其最大特点是将医学知识结构化，并发展出语义的描述机制，以及着重突出知识间的关联性，让知识如滚雪球般不断自动地累积。[8] 语义网络给出了这些语义类型和语义关系的定义，包括文本描述、层次关系和应用规则。Alexa T. McCray 等[9] 将 134 个语义类型聚为 15 个组（groups），从而使概念高度结构化。在全部概念中，仅有不足 5%（4913）的概念同时属于 2 个组，只有 6 个同时属于 3 个组。这 15 个组分别是：活动和行为、解剖、化学物质和药品、概念和思想、设备、失调、基因和分子顺序、地理区域、生命体、物体、职业、组织、现象、生理、过程。

2.2.2　两大类型语义关系揭示医学领域的内容关系

UMLS 中的语义网络定义两大类型的语义关系，即：是（isa）、相关（associated with）。在相关关系中又细分为物理相关、空间相关、功能相关、时间相关和概念相关，为了精确地反映医学各概念、术语、事件、行为的关系，对各种相关关系进一步细化，从而形成 53 种相关语义关系，这样将疾病的分类、流行、预防、诊断、治疗、并发症等内容加以揭示。

2.3　本体

本体（ontology）是一个关于一些主题的清晰规范的说明。它是一个规范的、已经得到公认的描述，它包含术语表和关系集，术语表中的术语全是与某一学科领域相关的，术语表中的逻辑声明全部是用来描述那些术语的含义和术语间关系的；关系集把握着词表中这些术语间的联系。[10]

2.3.1　术语的自然语言化

在本体（ontology）中的术语、主题、概念可以是自然语言和半自然语言，而不是受控的科学语言，使得本体（ontology）更新更容易，并为人和机器互相理解提供基础。

2.3.2　概念结构的网状化

本体（ontology）用概念（类）、子类、实例表示概念的上下位关系，用属性描述类的性质，用函数表示概念间的联系。这样对概念的描述形成三维结构。

2.3.3　关系描述系统化

ontology 对概念、术语间的关系描述得更为广泛、细致和全面，这也是其作为知识组织立足点最重要的特质。在本体（ontology）中可以描述的概念间关系如下：①反义关系（antonym）②上位关系（hypernymy）③下位关系（hyponymy）④整体-部分关系（holonymy）⑤部分-整体关系（meronymy）⑥转指关系（metonymy）⑦近义关系（near-synonymy）⑧同义关系（synonymy）⑨动作关系（troponymy）。

2.3.4　知识组织体系完备化

本体（ontology）可以将医学领域的信息资源进行组织，组织成为知识组织体系完备的知识库，该知识库具有智能查询、回答用户问题等功能，具有进一步知识挖掘的能力。

3　ontology 与 MeSH、Semantic Network 的比较

3.1　ontology 与 MeSH、Semantic Network 的联系

（1）本体（ontology）、MeSH、Semantic Network 都是知识组织工具。

（2）MeSH 对概念进行规范性说明、具有层次结构和分类等级，Semantic Network 具有网状结构和简单的推理功能，因此 MeSH 和 Semantic Network 是广义意义上的“本体（ontology）”。

（3）MeSH、Semantic Network 是医学专家和图书情报专家的智慧结果，其组织的科学性、合理性为构建完善的医学领域本体提供了基础。

3.2　ontology 与 MeSH、Semantic Network 的区别

ontology 与 MeSH、Semantic Network 在描述对象和范围、知识表示深度等都存在着差异，具体见表 1。

4　结束语

本体的核心概念是知识共享，领域本体可以减少概念和术语上的歧义，本体描述为某一组织或某工作小组提供了一个统一框架或规范模型，使来自不同背景、不同观点、不同目的人员之间的理解和交流成为可能，并保持语义上的一致性。[11] 在进行复用现有领域本体和扩展这些本体的工作中，对术语进行规范地分析是极有价值的，并且避免了领域知识的重复分析。[12] 从这三种关系模式在知识组织中的应用特点可以看出，领域本体无论是在描述对象和范围上、概念描述语言使用上、结构安排上，还是在描述的关系、知识表示的深度、对断言的表达深度上都有巨大的优势，并且有自备知识库，能实现智能查询功能。因此知识组织体系的关系模式应完善领域本体的构建。

领域本体的构建应结合叙词表和语义网络的优势，而不能完全摒弃这两种知识组织体系。叙词表是几代图书馆人和医学专家智慧和经验的结晶，并且随着领域知识本身的变化，叙词表还在不断增删主题词完善语义关系，从叙词表收词的变化上，可以发现该领域的发展方向和领域热点。

知识组织体系应采用叙词表来表达术语，采用语义网络的语义类型表达术语间的关系，用本体（ontology）的六要素

表 1 ontology、MeSH、Semantic Network 区别表

	ontology	MeSH	Semantic Network
描述的对象和范围	本体描述的医学领域公认的概念，既可以是简单的事实，又可以是信念、假设、预测等抽象的概念；既可以是静态的实体，又可以是与时间推移相关的概念，如治疗、病理功能等。	医学检索工具中专用于主题词检索的所有概念，均为静态的实体。	UMLS 超级叙词表定义的所有概念，及概念间的关系。
概念描述语言	自然语言、半自然语言	科学用语	科学用语
结构	网状结构	树状结构	网状结构
描述的关系	所有的 MeSH、Semantic Network 中能表示的关系，以及事件发展过程的时间关系	用、代、属、分、参	两大类 54 种语义关系
知识表示的深度	有 6 个要素“概念、类、关系、函数、公理和实例”，它通过这 6 个要素来严格、正确地刻画所描述的对象	只能用概念、概念间的组配来刻画所描述的对象	深度上不如本体，对建模没有特殊要求。
对断言的表达深度	最深。能表示出‘乙型肝炎实验室诊断方法的组成和数值意义’这样的整体断言。	不表示断言。用‘肝炎，乙型/诊断’这样的主题词/副主题词组配来表示文献描述了这部分的内容。	深。只表示‘实验室检查对乙型肝炎的诊断有作用’这样的断言。
是否有自备知识库	有	没有	没有
是否有智能查询功能	有	没有	与专家系统配合，有智能查询功能
使用难度	简单	需要经过专业培训	简单
备注	① 这里的本体（ontology）是理想化全部功能都实现的本体。 ② 这里的 Semantic Network 为 UMLS Semantic Network		

来组织关系模式。既发挥叙词表在术语、主题和概念描述的精准特点，同时扩展叙词表单一的线性树状结构为网状构架，增加对术语属性和术语、概念之间关系的描述，将科学的概念表达融入到本体（ontology）模式的知识组织体系中。

参考文献：

[1] Linda Hill,Ph.D., Olha Buchel, MLS, Greg Janee, MS, 曾雷. 在数字图书馆结构中融入知识组织系统[J].现代图书情报技术，2004，(1)：4-7.

[2] Tim Benners-Lee. Semantic Web by XML 2000 [N]. 2000-12-06.

[3] 中国医学科学院医学信息研究所. 医学主题词字顺表：2002 年版[M].北京：中国计量出版社，2002.

[4] 严青利，张勇.医学主题词表（MeSH）评述[J].情报杂志，2001，(8)：64，66.

[5] Welch J.Searching the medical literature [M].London：Chapman and Hall，1985：17.

[6] 赵瑞清，王晖，邱涤虹.知识表示与推理[M].北京：气象出版社，1991：19-21.

[7] http://www.nlm.nih.gov/pubs/factsheets/umls.html [2005-02-15]

[8] 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.

[9] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity [EB/OL].http://www.uclac.uk/kmc/news/stories/medinfo2001_report.pdf/2005-03-01.

[10] 李景，钱平.叙词表与本体的区别与联系[J].中国图书馆学报，2004，(1)：36-39.

[11] 李善平，尹奇韦华，胡玉杰，郭鸣，付相君.本体论研究综述[J].计算机研究与发展，2004，(7)：1041-1052.

[12] McGuinness, D. L., Fikes, R., Rice, J.and Wilder,S..An Environment for Merging and Testing Large Ontologies [EB/OL].(http://www.ksl.stanfordedu/people/dlm/papers/kr2000-camera-ready-copy.doc/2005-02-01)

作者简介：牟冬梅，吉林大学公共卫生学院信息管理与信息系统专业副教授，图书馆学专业硕士生导师；王丽伟，吉林大学公共卫生学院信息管理与信息系统专业讲师，图书馆学专业硕士生。