●王 军

基于分类法和主题词表的数字图书馆知识组织*

摘 要 实现数字图书馆从信息管理向知识管理模式的转变的关键是知识组织,即将已积累的信息资源按照一定的知识体系组织起来。利用集成分类法、主题词表和语义元数据构造数字图书馆知识组织系统的方法,可以实现数字图书馆的知识组织。VISION 是基于《中国分类主题词表》的一个知识组织的原型系统。图 2。参考文献 4。

关键词 数字图书馆 知识组织 分类法 主题词表 语义元数据
分类号 G254

ABSTRACT The author thinks that the key to transform digital library from information management to knowledge management is knowledge organization, i.e. to organize acquired information resources according to certain knowledge systems. By using integration classification systems, subject thesauri and semantic metadata, we can realize the knowledge organization of digital library. The author also introduces VISION, which is a prototypical system based on *Classified Chinese Thesaurus*. 2 figs. 4 refs.

KEY WORDS Digital library. Knowledge management. Classification system. Theusaurus. Semantic metadata.
CLASS NUMBER G254

1 引言

新时期数字图书馆(DL)的历史使命是知识管理,国内 DL 的建设者和研究者都将知识服务作为新世纪图书情报的增长点。那么,如何在现有 DL 的架构上实现知识管理和知识服务,就成为亟待研究的一个课题。

国内 DL 的发展,目前取得成就主要是在以信息资源建设为主的基础设施建设方面。当前,国内的 DL 一方面继续加强资源的建设;另一方面,在已有相当资源积累的基础上,强调用户服务。特别是国家科学数字图书馆,提出了面向用户的信息服务体系,张扬知识服务的理念。但是,目前的 DL 体系,沿袭了传统图书馆以信息资源为中心的管理运行模式,所支持的仍然是以文献检索和传递为核心的信息服务,不能够有效支持面向用户的知识服务。正如张晓林博士指出的那样,要实现 DL 的知识管理,需要一个新型的技术机制,它"应该充分支持基于虚拟资源体系的服务集成,充分支持基于内容的数据检索、信息内容分析和动态集成,充分支持数据挖掘和知识发现,充分支持个性化、专题化和智能化服务,充分支持以用户为中心的信息交流、知识析取和知识应用,充分融合用户信息资源和信息系统"[1]。它不是以信息资源为中心,而是能将信息资源融入其中的一个知识管理系统。

那么,如何构建 DL 的体系结构,使得当前以信息资源为中心的 DL 能够平滑过渡到面向用户、提供知识服务的未来 DL,为以后的发展奠定基础并留下广阔的发展空间,就是我国 DL 当前必须关注的问题。实际上,从传统图书馆沿袭而来的信息管理模式已经成为当前 DL 利用和发展的障碍。主要体现在如下几个方面:

(1)缺乏有效的资源利用手段,特别是缺乏元数据资源。当前主要的检索手段——关键词检索和超文本浏览,远不足以满足用户的检索需求和充分挖掘信息资源的价值。这无疑是对业已累积的大量信息资源的浪费,特别是缺乏对于包含主题标引信息的元数据资源。关键词检索将主题标引信息(主题词、分类号)和普通字段同等对待,几乎完全浪费了标引员宝贵的智力劳动。

* 本文得到国家自然科学基金 70303002 和国家社科基金 03BTQ001 的支持。

(2)缺乏对DL内外各类信息资源统一组织和整合的能力。Web信息资源的丰富价值和急剧增长,使得人们期望DL在组织自身资源的同时,也能对Web上的网页、多媒体、地址等资源提供导航和管理。遍布DL之上的"信息门户"、"学科导航"便是对这一需求的响应。实现DL内外各类信息资源的整合和统一管理,需要一个一致的综合资源组织管理模型。

(3)不能充分发挥DL的教育功能。现代教育的高额费用和教育技术的复杂性,造成了教育机会的不平等和教育资源占有的两极分化。数字图书馆,全社会高价值网络信息资源的收集者、组织者和管理者,应当提供知识导航、虚拟参考等服务,以集成和发挥图书馆的教育功能,辅助用户自学习。但是当前DL的体系机构缺乏相应的支持机制。

造成这些问题的一个根本原因是DL中的信息资源缺乏组织[2]。就图书馆工作的三个基本环节(资源、组织和服务)比较DL和传统图书馆,DL中的资源是数字化的,服务是网络化的,唯有在资源的组织方面存在严重缺陷。免除了传统图书馆的分类排架和目录组织,也随之丧失了按学科门类和知识体系进行浏览和检索的手段;不能提供主题词表在检索阶段的辅助,关键词查找乃不得已而为。

综上所述,无论是基于肩负的历史使命还是解决目前DL发展和建设中的问题,都需要改造DL从传统图书馆中继承来的以信息资源为中心的信息管理模式,进化到知识管理。实现这一改造的关键是将DL的信息资源按照知识体系进行重组——知识组织。那么,能否将传统图书馆中以分类法和主题词表为主的知识组织工具应用于数字化、网络化的信息资源的知识组织呢?分类法和主题词表是传统图书馆中最重要的知识组织工具,是数代图书馆员智慧和经验的积累,它们的知识组织能力在两百多年的发展和应用过程中得到了充分证明和不断的丰富。当前Web社区对词表、知识本体、网络知识组织工具(NKOS)的热烈讨论说明了它们在新的信息环境下的生存能力和应用前景。但是,起源并应用于传统图书馆的知识组织工具——分类法与主题词表,直接应用于DL中,尚有许多缺陷。第一,DL所管理的对象是动态的、海量的、分布的数字化信息资源。规范化的、严格受控的分类法和主题词表的编制和修订都要依赖专家。其结构和内容相对于网络信息资源的迅速更新和变化,结构和内容滞后,难于自动更新。第二,分类法和主题词表是面向图书馆员的,体系和规则都较为复杂。DL直接面向分布在整个因特网上的最终用户,他们的职业不同、年龄不同、教育背景不同,体系庞大、结构复杂的分类法和主题词表难于被普通用户所掌握。第三,二者的功能都侧重于对文献的标引和组织,对检索服务的应用要求考虑得较少。而这正是DL所需要的。第四,大多数DL中的信息资源,都有自己的收藏特色,面向特定的领域,服务于特定的用户,而通用的分类法和词表缺乏对所应用资源的针对性。第五,分类法的一大功能是藏书排架和目录组织。这一功能能否继续用于DL的资源组织和服务呢?

因此,将分类法和主题词表应用于DL资源的知识组织,必须对它们进行改造。本文提出了一个将分类法、主题词表与语义元数据集成起来,构造DL的知识组织系统(DLKOS)的方法,以解决上述的所有问题。国内DL已经取得的建设成就和国际上网络知识组织系统的研究成果为DLKOS的实现提供了必要的物质和技术保障。下面,首先简介国内外知识组织相关领域的研究现状,然后讨论对改造分类法和主题词表,并集成元数据以构造DLKOS的全过程。其次是我们据此开发的原型系统VISION的介绍,它以《中国分类主题词表》为基础,集成北京大学图书馆提供的5千多条书目数据。最后是全文的总结。

2 集成分类法、主题词表和语义元数据构造DLKOS

上文已经提到,国内DL的建设已经积累了大量的元数据资源,而且目前DL资源利用不足的问题主要体现在元数据资源上。元数据是DL中最重要信息资源,是DL建设的重中之重"元数据是关于数据的数据",其中所包含的原始文献的内容标引信息(如分类号、主题词等)是标引员在理解文献内容的基础上,根据分类法、主题法的知识体系和标识系统来表示的。它凝聚着标引员宝贵的智力劳动。为了强调这一点,我们称其为语义元数据。由于元数据资源没有像传统图书馆中的馆藏那样,进行分类排架和目录组织,从而肢解了隐藏其中知识系统,因此,构建DL知识组织系统的关键是使元数据资源中被掩盖的知识体系显现出来,发挥它资源组织和检索服务的功能。这就是集成分类法、主题词表和语义元数据,构造数字图书馆的知识组织系统

(DLKOS)的基本方法。这种方法的要点是:首先改造分类法和主题词表,形成一个由类目或同义词集合作为概念节点、以学科等级关系或概念语义关系作为边的概念网络;然后将各元数据记录按照它们的主题标引信息分配到对应的概念节点下,作为对应概念节点的文献实例,相当于元数据的"上架"。这样,结合了具体元数据记录的概念节点不仅包含抽象的概念,而且包含具体的文献实例,成为一个知识节点。上述概念网络就成为一个知识的网络——DLKOS。

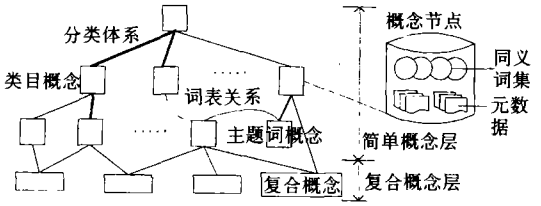

图 1　DLKOS 的结构

DLKOS 的具体构造方法如下(如图 1 所示):

(1)改造主题词表,并和分类法结合起来,构造一个嵌入了分类体系的抽象的概念网络。词表中一个主题词和所有与它具有等同关系的关键词形成一个同义词集合,作为 DLKOS 中的一个抽象概念。概念沿袭原词表中的等级关系,具有等级关系的若干概念形成一个概念族,对应于原词表中的词族。若两个或多个概念共同参与了一篇文献的标引,则它们相互之间具有组配关系。这样,词表被改造成一个抽象的概念网络。随后,再将分类法的学科体系结构嵌入到这个概念网络中,充当概念网络的主干结构。若是分面主题词表,类表和词表已经完全结合在一起,每一个类目都对应着一个概念,类目间的学科等级就是概念间的等级关系;若是分类主题对照表,二者没有完全的等值对应关系,需要另外创建类目概念节点,并建立类目概念和主题词概念间的双向索引。

(2)根据元数据的语义标引信息,将它们分配到上述的抽象概念网络中去,形成一个知识的网络。这是构造 DLKOS 的关键。假设 m 是一条元数据记录,若 m 是由一个概念 A 标引的,则将 m 作为 A 的文献实例,置入 A 的概念节点之下;如果 m 是由若干个概念标引的,例如 A 和 B,则建立 A 与 B 之间的边"A—B"以反映 A、B 间的组配关系,并将 m 作为边 A—B 的实例。在 DLKOS 中,组配关系边实现为一个同时以 A 和 B 为上位的复合概念(如图 1 所示)。组配关系可以认为是对词表中相关关系的替代。词表中的相关关系是非常松散的和不确定的:在 DLKOS 中,当且仅当存在相应的文献实例时,即同时用 A、B 标引的元数据记录,才确认 A、B 间的相关关系。元数据充当了概念间相关关系的验证。进一步,当出现了新的专业术语来表示 A—B 间的组配关系时,边 A—B 可以进而演化为新的概念节点,从而提供了 DLKOS 的自丰富机制。结合了元数据记录的概念节点不再是一个抽象的概念,即有表达概念内涵的同义词集合,又有表达概念外延的元数据集合,成为一个知识的节点,抽象的概念网络成为附有文献实例的知识网络,完成了 DLKOS 的构造。DLKOS 中的概念间有三类联系:来自于词表的概念间的等级关系;来自于分类法的概念间的学科相属关系;来自于元数据的概念间的组配关系,即相关关系。这三类关系用统一的上下位关系表示。

(3)从元数据中自动提取专业术语,并添加到 DLKOS 中,实现 DLKOS 的自丰富与增强。要保持 DLKOS 的可应用性与生命力,必须及时地对 DLKOS 中的词汇、概念以及概念间的关系进行更新和补充。元数据是原始文献的替代物,能够及时地反映学科的最新进展。特别是科技文献的元数据资源,在标题字段中包含有大量的专业术语。而且科技文献的标题具有鲜明的语法特征,语义上能忠实地反映文献的主题内容,从而和主题标引信息有着较直接的对应关系。基于此,采用统计语言学的方法从元数据的标题字段提取专业术语,并且计算它们在 DLKOS 中的语义位置,确定和已有概念间的关系。需要解决的关键技术问题有三个:标题的切分及关键词的提取;提取的关键词的筛选;筛选出的关键词在 DLKOS 中的定位。这部分是 DLKOS 构造的关键技术和突破所在。

如此构建的 KOS 具备五大功能:

(1)分类法和主题词表为 DL 资源的组织和管理提供了一个知识框架:下层的元数据作为概念网络的数据实例继承了概念间的所有关系,原本离散、孤立的元数据单元相互间拥有了丰富的语义联系,加入到一个统一的知识体系中。以此为组织框架,还可以吸纳馆外的信息资源,形成以 DL 为中心的大一统 KOS,实现 Web 信息资源的泛 DL 化管理。

(2)为 DL 业已累积的元数据资源提供有效的利用手段:由于所有的元数据资源都组织到 DLKOS,一方面,用户可以循着学科等级和概念间的语义关系进

行浏览,实现了知识导航;另一方面,定位了相应的概念就检出了所需的文献,无须对数据资源进行遍历式的搜索,检索过程的关键不再是检索词的匹配,而是用户检索需求的概念化表达,实现了概念检索。

(3)为用户提供了一个检索、服务、教育一体化的知识空间:DLKOS除了提供知识浏览和概念检索的功能,它还提供了一个共同的知识空间,用户在浏览和检索的过程中,可以定位感兴趣的知识点及其在知识体系中的位置,并了解相关的知识点,帮助、指导用户自学习。

(4)一个自丰富、自增强、自适应的知识系统:元数据是原始文献的替代物,包含丰富的专业词汇。可以从中自动提取新的术语丰富到DLKOS中,解决分类法和主题词表不易更新、只能依赖专家修订的问题,同时也增强了DLKOS的检索能力。结合了元数据的DLKOS,凡是有元数据实例存在的概念节点就显现出来;没有的则隐藏起来,通用的概念网络根据资源的学科领域和规模自动地适应和调整,减轻了用户的负担。

(5)实现DL知识管理的技术基础:DLKOS为DL资源的知识组织提供了现实可行的解决方案,为DL开展知识服务,实现以资源为中心的信息管理模式向以知识体系为中心的知识管理模式的演进提供了技术基础。

3 原型系统VISION

在《中国分类主题词表》的基础上,我们实现了这样的一个知识组织系统—VISION[3]。VISION集成了北京大学图书馆提供的5000余条计算机领域的书目数据,DLKOS在服务器端,用Oracle9i实现;前端是一个概念检索系统,采用Java来实现。Oracle9i中丰富的面向对象技术,如嵌套表和可变长数组,为VISION中复杂对象的实现提供了支持。

建立DLKOS的数据流程如图2所示。首先,将分类主题词表中所有的款目导入到数据库中,具有等同关系的所有主题词形成一个概念,概念之间通过等级关系构成概念树;然后建立分类法的类目和概念间的对应关系,并利用类目节点间的学科等级关系将所有的概念树连缀起来,形成一个由概念节点和类目节点组成的概念网络。最后将书目记录按照它们的主题组织到相应的概念节点中去,完成DLKOS的构造。在此之前需要对原始数据作些规范化的处理,如全角半角的转换,英文大小写的转换,标点符号的处理等。

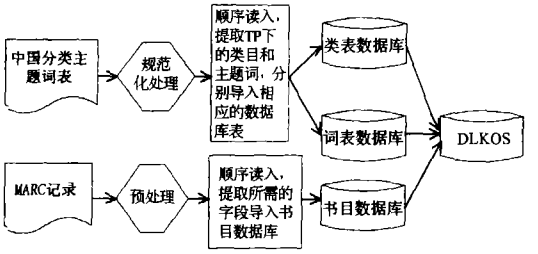

图2 DLKOS的建立流程

在DLKOS的基础上,我们开发了一个概念检索系统。目前提供的主要功能有:词汇辅助、知识导航和单概念检索,其它功能尚在完善之中。

左侧窗口显示了概念网络,可以从学科分类、字顺、概念族三种方式查看概念网络;右侧中部窗口显示用户选取或查询的概念的信息,包括属于这一概念的词汇,上下位概念,所属类目等;右侧左上窗口用图形化的方式显示了给定概念和其它概念间的关系;右侧下部窗口是属于这个概念的所有书目数据。用户在VISION以知识点为基本单元(包括属于一个概念的所有词汇以及以此概念为主题的所有元数据)进行知识导航,定位了一个概念,也就定位了该概念在知识体系中的位置;用户可以从任何一个同义词出发,检索对应的概念和以之为主题的文献。查询不再需要在数据集合中进行关键词匹配,而是在概念网络进行概念检索。

VISION的另一个技术突破是DLKOS的自丰富机制。科技文献的标题通常能忠实地反映文献的内容,包含了丰富的反映学科最新进展的专业技术术语。标题通常都是名词性短语,具有鲜明的语法特征。根据这些特点我们应用基于bigram[4]的统计方法提取候选词,并利用元数据记录中主题标引信息和标题间的语义对应关系,从候选词中筛选出有价值的、专指度高的词,并计算新词在DLKOS中的合适位置。从5000余条书目记录中,在阈值为5的情况下,我们提取出554个词,抽词准确度达到89.5%;从这554个词中,正确选出356个专业词汇,准确度达到95%,其中341个被正确定位在DLKOS中,准确度达95.8%。实验结果是相当令人满意的。DLKOS的自丰富机制充分利用了图书馆员在标引过程中的智力投入。可以说,VISION所做的是将编目员的手工劳动挖掘和显现出来。限于篇幅所限,DLKOS的自丰富机制和抽词算法待另外撰文介绍。

(下转第64页)

存原始数字资源的比特流。

3.2 用于界定数字图书馆的功能

对于数字图书馆功能,国内外研究机构给出了不同的界定。国内比较典型的一种说法是数字图书馆的功能包括5项:(1)各种载体数字化;(2)数据的存储和管理;(3)组织对数据的有效访问和查询;(4)数字化资料在网上发布和传送;(5)系统管理和版权保护。

OAIS的功能参照模型被用于指导确定建立数字图书馆的功能模块。比较典型的项目是欧洲国家版本图书馆(NEDLIB, Networked European Deposit Library),由欧洲7个国家图书馆(荷兰、法国、挪威、德国、葡萄牙、瑞士、意大利)和3家主要出版社(Kluwer, Elsevier和Springer Verlag)参加,其内容包括建立欧洲版本图书馆网络的基础结构,保证电子出版物的长期保存和利用。他们研制的DLS(Digital Library System)系统借鉴了OAIS的参考模型,包括编目、信息采集、DSEP、读者检索权限控制、信息服务等11个模块。其中,DSEP(Deposit System Electronic Publication)负责存储处理和保存功能。DSEP以OAIS作为框架,但它的获取、数据管理和检索功能比OAIS的相应功能范围小,其中有些功能被分到了其他模块里。DSEP通过输入和输出接口来和外界保持联系。输入和输出接口负责将外界接收来的数据格式转换成DSEP规定的格式,并根据用户需求将DSEP的内部格式转换为读者需要的格式。

4 结束语

综上所述,OAIS参考模型对于制定数字图书馆的元数据标准和功能标准都具有十分重要的意义,一旦元数据标准和功能标准得以确定和执行,数字图书馆将不仅在横向上而且在纵向上都具有开放性,数字信息的长期保存和利用问题也得以解决。

值得一提的是,国内对于OAIS的研究目前仍然很少。在具体的数字图书馆实践中,对于OAIS的借鉴将会带来什么样的实际问题,仍然有待于进一步研究。

参考文献

1 张晓林.数字信息的长期保护问题.图书馆,2001(5)
2 杨宗英,郑巧英.数字图书馆研究.大学图书馆学报,2000(1)
3 刘嘉.元数据:理念与应用.中国图书馆学报,2001(5)

余传明 武汉大学信息管理学院情报学2002级博士生。通讯地址:武汉。邮编430072。

董 慧 武汉大学信息管理学院教授,博士生导师。通讯地址同上。

(来稿时间:2003-05-12)

(上接第44页)VISION下一阶段的开发工作将扩展实验数据集。目前VISION中仅仅集成了计算机领域的5000余条书目数据,新词的提取和定位也是在这些书目数据上完成的。我们希望能在更多的领域、规模更大的其它元数据类型上进行实验,最终将VISION推向实用。

4 结语

集成分类法、主题词表和语义元数据构造DL的知识组织系统,为DL提供一个现实可行的知识组织模型,为DL从信息管理向知识管理的过渡提供技术基础。它为当前我国DL业已累积的信息资源提供了基于内容、面向知识的利用和服务手段。VISION原型系统的成果充分说明分类法、主题词表等传统知识组织工具在网络信息环境下仍然有着重要的价值,为了适应数字化、网络化的信息环境,传统图书馆的理论和方法需要不断进行变革与发展。正如国际著名信息学家奈斯比特(J. Naisbit)所指出的:"我们正受信息淹没,但却渴求知识。"DLKOS将为人们提供一个驾驭海量的、日益增长的网络信息资源的知识框架,为解决信息爆炸和信息污染的问题作出图书馆学领域的贡献。

参考文献

1 张晓林.走向知识服务.中国图书馆学报,2000(5)
2 Wang Jun. A Knowledge Network Constructed by Classification, Thesaurus and Semantic Metadata in digital libraries. ASIST Bulletin, 29(2), 2003
3 王军.VISION:集成分类法、主题词表和语义元数据的概念网络.情报学报,2003(4)
4 Sproat, R., and Shih, C. L. A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese & Oriental Languages, 4, 4(1990)

王 军 北京大学信息管理系副教授,计算机科学博士。研究方向为数字图书馆、知识组织、信息检索。通讯地址:北京大学信息管理系。邮编100871。

(来稿时间:2003-11-19)