

文献信息知识组织与内容揭示方法探究

潘 薇¹, 喻 浩²

(1. 中国标准化研究院国家标准馆, 北京 100088; 2. 中国农业大学图书馆, 北京 100094)

[摘 要] 简要论述了国内外文献信息知识组织的现状与发展情况, 介绍了文本与网络信息知识内容揭示的主要方法, 包括基于主题词和知识分类的信息组织与揭示、基于元数据及信息资源内容描述框架的信息组织与揭示以及基于知识组织体系的内容揭示, 为网络环境下信息资源的知识与内容揭示提供了有益的参考。

[关键词] 知识组织; 内容揭示; 元数据

[中图分类号] G254 **[文献标识码]** A **[文章编号]** 1003-725X(2009)03-0046-03

Research on Methods of Knowledge Organization and Content Revelation of Document Information

PAN Wei¹, YU Hao²

(1. Library of China National Institute of Standardization, Beijing 100088, China;

2. Library of China Agricultural University, Beijing 100094, China)

Abstract: This paper briefly expounds the current status and the development of knowledge organization in foreign and domestic information and document management. It introduces the main methods of knowledge and content revelation of document and network information resources including information organization and revelation based on subjects and knowledge classification, based on information content description framework and metadata, as well as based on knowledge organization system in order to provide valuable reference for knowledge and content revelation of network resources.

Keywords: knowledge organization; content revelation; metadata

CLC number: G254

文献内容的揭示工作是图书馆的基础工作之一。随着信息量的急剧增加和图书馆数字化、网络化的全面推进, 用户希望能查阅到具有详尽内容的书目数据。网络化、数字化时代的文献信息组织与揭示, 可以从以下三个层次来进行: 一是基于主题词的信息组织与揭示; 二是基于元数据的信息组织与揭示; 三是基于知识组织体系的内容揭示。这些内容揭示方法为多元化的用户提供了直接和便捷的文献信息查找途径。

1 基于主题词的文献信息内容揭示

1.1 主题描述

主题(词)法是标题(词)法、单元词法、关键词法和叙词法的统称。国家标准 GB13190—91《汉语叙词表编制规则》将叙词表定义为: “将文献、标引人员或用户的自然语言转换成规范化语言的一种术语控制工具, 它是概括各门或某一学科领域并由语义相关、族性相关的术语组成的可以不断补充的规范化的词表。”主题法可追溯到明代永乐年间编制的《永乐大典》, 它比英国克里斯塔多罗提出的字顺标引系统早 448 年^[1]。有研究指出, 主题法在 20 世纪 50 年代随着计算机在图书情报领域的应用而逐步产生的, 它吸收了元词法、标题法、分面组配分类法的优点。将受控的表达文献主题的语词编成专业或综合性的词表, 是叙词法的基础和前提条件。1959 年, 美国杜邦公司编制了第一部叙词表, 其后, 随着计算机的应用, 叙词表得到迅速发展, 叙词语言成为受控信息组织和检索的主要语言。到目前为止, 国内外叙词表的数量不下千种,

我国叙词表也已超过 130 种。20 世纪 60 年代初, 我国开始应用主题法。1978 年 1 月国防科技信息中心编辑出版的《国防科学技术主题词典》, 是我国第一部英文叙词表。1980 年 3 月, 在中国科技情报所和国家图书馆的组织下, 《汉语主题词表》出版。该表按自然科学和社会科学两个子系统分别编制, 全表共 3 卷 10 分册, 收录主题词十万余条, 其中, 正式主题词九万余条, 非正式主题词一万余条。该表的问世标志着我国主题词表的编制与主题语言的研究达到了新的高度。1982—1985 年我国相继出版了有关农业、计算机科学、国防、交通、化工等汉语主题词表。

叙词表的术语由叙词和非叙词组成, 叙词是在文献标引与检索中用以表达文献的主题而规范化的词, 叙词表中的词间关系由等同关系、属分关系、相关关系组成^[2]。叙词法与自然语言相结合, 将叙词语言的应用范围扩展到网络化信息的组织和检索, 促进了语义网的智能知识检索等。

1.2 主题标引

主题标引是揭示文献内容最直接、最有效的手段, 是利用主题词表, 对文献内容进行主题分析, 赋予主题词标识的过程。1984 年我国实施的国家标准 GB3860—83《文献主题标引规则》对文献主题分析、主题词的选定、主题款目标目的拟定、审校工作和质量管理等均作了规定。标引规则和组配规则规定得比较细。1985 年《汉语主题词表标引手册》详细地介绍了《汉语主题词表》的结构和使用方法, 重点阐述了文献主题标引的步骤和方法, 并以实例

说明如何进行主题分析,如何选用标引词以及标引规则。

各科技领域主题标引规则在推动文献内容揭示方法应用、加强内容标引工作规范性方面,均起到了重要作用。1995年出版的《中国分类主题词表》是一部大型综合性的、分类法与主题法语言兼容的文献标引工具。其主题词对应表包括22个大类,该表把分类法与主题法相结合的研究推向新的水平。

1.3 网络信息的概念与主题揭示

网络信息组织是通过网络对数字资源的信息组织,它比传统信息组织更复杂。网络的交互性、文件格式的多样性、信息载体的多样性、储存地点的多样化等,形成了数字资源信息组织新的特点。目前,网络信息组织方式主要是按照实际信息内容,设置适宜的主题栏目,形成一定知识体系的信息组织方法。

自20世纪90年代起,因特网得到世界范围内的普及和使用。网络信息的指数增长特点,使网络信息的信息组织直接采用了关键词的全文检索方式,最典型的就是网络搜索引擎的研制和使用。目前,世界上使用最为广泛的是Google。它是由美国斯坦福大学两名博士研究生Larry Page和Sergey Brin于1996年开始的研究项目。大约搜索引擎利用市场的60%由Google控制,这一数字仍在增加^[2]。Google为使用者提供了117种语言界面,成为世界上最优秀的网上内容揭示搜索引擎之一。利用Google可查找到不同格式的网络文件,包括PDF文件、Flash文件、Microsoft Office(doc,ppt,xls,rtf)和其他类型文档,以及不同类型文件的“HTML版”,使不同用户即使在没有安装相应应用程序的情况下,也能方便检索和阅读各种类型的文件。

基于主题词的文献信息内容揭示方法利于用户从主题词的角度检索和利用文献。

2 基于知识分类的文献信息内容揭示

2.1 文献内容的知识分类

分类法在图书情报领域的应用,是在图书馆的藏书、排架、文献检索等具体需求的基础上发展起来的。它作为一种人工编制的信息语言,按照一定的科学知识体系,把相关的规则编制成可供利用的分类表,并将其作为一种标准执行,形成文献的知识分类系统。

文献分类法从编制形式上大体分为:(1)列举式。即对某一概念进行分类时,采用把主要因素列举出来,将下位概念按顺序排列起来的方法;(2)组配式。即把复合主题用较为简单的概念组合起来。用这种组配的方式来表达具体类目。也称为分析合成型分类;(3)混合式。在按等级列举类目的基础上,运用组配式的原理,将列举式和组配式加以混合使用。

文献分类法在号码编制上有十进制与非十进制之分。《杜威十进分类法》和《国际十进分类法》采用十进制编码方法,《美国国会图书馆分类法》采用非十进制编码方法。《杜威十进分类法》将知识体系分为10个主要的学科(main classes),每个大类下细分10类(divisions),每个类又分成10小类(sections)。类是以三个数字中的第二个数字表示,例如:500表示科学,510则表示数学,520是天文学,530则是物理学。它被世界一百三十多个国家的二十多万图书馆使用^[3]。《国际十进分类法》因其分类详细、灵活性强、通用性好等特点,被广泛应用于科学论文、标准文献的分类。《国际十进分类法》同样将知识体系分为10个大类(Main numbers),并且使用了大量的符号表达概念间的组配关系,如用“+”表达两个不连贯的类目标配成一个综合类号,用“:”表达两个类目的交叉重叠部分等^[4]。

《美国国会图书馆分类法》是一部综合性等级列举式分类法。

该分类法目前广泛应用于北美大、中型图书馆。它的类号是采取字母和数字混合制配置的。如J代表Political Science、K代表Law、L代表Education、M代表Music and Books on Music、Q代表Science等^[5]。

《中国图书馆分类法》是新中国成立后编制出版的一部具有代表性的大型综合性分类法。该法吸取了国外分类法的编制理论和技术,普遍应用于全国各类型的图书馆,采用拉丁字母与阿拉伯数字相结合的混合标记符号。

分类法一般包括编制使用说明、分类表、类目索引及手册四个部分。分类表的基本单元是类目,包括类名、类号、类级(隐含)及注释(含类目参照)四部分。分类表往往分为主表和副表(即辅助表),主表由类目大纲、简表、详表构成,是对知识结构的具体描述,分类号是由纯数字或字母与数字混合组成,将分类主题进行排序,通过简明的符号,容易体现出相应的主题知识。随着世界信息资源数字化、网络化的飞速发展,在传统图书资料基础上发展起来的分类法,在对网络信息资源的组织中,将会展现出新的优势和特色。

2.2 分类标引

分类标引是根据事先规定使用的能够体现知识逻辑系统的分类法,对各种文献信息的内容、体裁、写作目的及其表现形式等各种特征进行分析,并将分析结果转换为相应的分类语言标识符号,汇集、组成科学的知识系统。对信息资源进行分类,可促进信息资源的有效组织和高效利用。通过将信息资源涵盖的知识内容的析出,用简单通用的符号进行排序,并体现其一定的相互关系,来表达不同层次结构的信息主题,便于用户准确查询检索文献信息,是知识分类体系进行信息组织和揭示的主要目的。

3 基于元数据与信息资源内容描述框架的信息组织与揭示

3.1 MARC元数据

关于元数据的定义有很多种,ISO15489中对元数据的定义是:元数据是描述文件的背景、内容、结构及其整个管理过程的数据。美国图书馆协会认为:“元数据是资源内容和格式信息的摘要,它可以描述著作权人、出版日期或更详细的情况,以利于信息的提供与服务。图书馆卡片、图书目录、索引、图书的版权说明、磁盘的标签、MARC(机读目录格式)和AACR(英美编目条例)等,都是元数据的一种形式。”元数据是以数据去描述数字资源的结构模式。

元数据有多种分类方法。从功能角度出发,元数据分为描述性、结构性和管理性三种元数据。描述性元数据用于揭示和描述一个事物,如MARC和都柏林核心数据集都是这一类型的元数据。结构性元数据是指关于数据之间的关系的数据,通过结构性元数据,可以将各类数字资源按一定顺序连结起来,以一种合适的结构将数字资源展现给用户,如在不同页面间翻转和切换的电子图书,在图像和文本说明间的相互切换等数据信息。管理性元数据用于揭示数字资源的管理特征,如版权信息、作者信息等。

MARC是Machine Readable Catalogue的缩写,意即“机器可读目录”,即以代码形式和特定结构记录在计算机存储载体上的、用计算机识别与阅读的目录。MARC可一次输入,多次使用,是信息技术发展和资源共享要求的产物。

MARC数据最早产生于美国。1961年,美国国会图书馆开始图书馆自动化的设想,随着计算机技术的进步,1963年,美国国会图书馆组织了在内部工作中采用电子计算机技术的可行性调查,1966年1月,产生了《标准机器能读目录款式的建议》,即MARC-1格式,1967年提出MARC-2,它是目前使用的各种机读目录格

式的母本。1969年开始向全国发行 MARCII 格式书目磁带,并将 MARCII 格式称为 US-MARC,即美国机器可读目录。作为一种计算机技术发展早期形成的数据格式,这一格式在定义时比较充分地照顾到图书馆书目数据在文献形式描述、内容描述、检索等方面的需要,表现为:字段数量多;著录详尽;可检索字段多;定长与不定长字段结合,灵活实用;保留主要款目及传统编目的特点;扩充修改功能强;并能实践中不断发展完善。

CNMARC 是中国机读目录 (China Machine-Readable Catalogue) 的缩写,是用于中国国家书目机构同其它国家书目机构以及中国国内图书馆与情报部门之间,以标准的计算机可读形式交换书目信息。CNMARC 格式为我国机读目录实现标准化、与国际接轨,从数据结构方面提供了保障。

鉴于 MARC 格式使用方法相对复杂,不适于对网络信息资源进行著录,所以近年来图书馆界广泛应用都柏林核心元数据集 (Dublin Core, 简称 DC) 来对网络信息资源进行著录。它已成为与 MARC 一样重要的一种元数据标准。

3.2 DC 元数据集和可扩展置标语言

都柏林核心元数据集 (Dublin Core Metadata Element Set, 简称 DCMES) 的目的是试图使用简单的元数据,建立一套描述网络文献资源的方法,以便实现对网页信息的准确检索。目前,DC 已被翻译成三十多种语言。1989年9月,因特网工程任务组 (Internet Engineering Task Force, 简称 IETF) 正式接受 DC 这一信息描述方式为其正式标准,2003年4月8日,DC 已被批准为国际标准 ISO15836-2003。随着网络通信技术的发展,面对数字图书馆海量信息资源及其丰富的内容特征和属性,需要更加丰富的语言来描述各种不同的属性和特点,SGML、HTML 和 eXML 置标语言随之出现,成为信息时代网络和数字化资源内容揭示的新手段。HTML 可用标签有限,数据格式无法表达其内在含义,数据信息不适合再利用,信息检索不精确,浏览器间数据互换有可能失去某些机能,所以人们又开发出了可扩展标记语言 XML (EXtensible Markup Language: 可扩展的标记语言),XML 可以看作 SGML 的简化版本,比 HTML 功能强大,比 SGML 简单易用。多数人认为,XML 将会成为下一代互联网的主要描述语言。

3.3 其他文献信息内容揭示方法

(1) 知识库 (Knowledge Base)。它是系统的有组织的知识集群,是针对某一(或某些)领域问题求解的需要,采用若干知识表示方式或框架,通过先进的计算机系统,进行组织、加工、存储和使用的知识数据系统。这些知识可包括理论知识、事实数据、专家经验性知识以及常识性知识等。(2) 提要。它是揭示文献内容最常用也是最基本的一种方法。提要须揭示文献的内容、著者的情况、版本的沿革等情况,有些还有历代人士的评价、编著者的评价等内容。(3) 文摘。根据《文摘编写规则》国家标准,文摘编写应遵循以下几条原则:忠实原文原则;逻辑性原则;新颖性原则;规范性原则;简明性原则。(4) 综述。它是对某一时期内的某一学科或专题的研究成果和技术成果进行系统的较全面的分析研究,进而归纳整理进行综合叙述。其包括综述、述评以及专题研究报告、动态趋势研究等,为指导科学研究提供综合信息,反映现状趋势和提供决策依据。

4 基于文献信息知识体系的内容揭示

关于本体论在信息组织中的研究和应用,目前在国内外均处于研究和实验阶段。本体论作为网络信息时代需求的产物,主要特点是包含一个领域的术语、对术语的定义以及术语间的关系。目前,虽然已经开发了一些本体标准语言和工具,如 RDF、OWL

等知识描述语言,但目前还没有开发出使用这些语言的应用软件,没有开发出基于这些语言数据的搜索引擎等。但本体论在图书情报领域的研究和应用,逐渐成为国内外的一个热点领域。Ontology 可以是在叙词表基础之上,借助语义相关和扩展标记语言 (XML) 等信息技术,在增加术语相关性的基础上形成一个知识系统。如德国卡尔斯鲁厄大学开发的 KAON 本体编辑工具。使用 KAON 可以获得扩展的 RDF 数据^[6]。在本体构建中,选词需领域专家参与。通过一定的概念、属性和实例的建立以及其间关系的描述,来表示专业领域的知识体系。本体论是一个含有语义的知识组织系统,可以基于本体论进行网络信息的组织和检索,甚至进行自动标引和主动进行专题信息推送等工作。现代信息系统正在从“数据加工”朝“概念加工”转变,这意味着加工的基本单元的数据越来越少,而带有内容概念的文中信息和语义概念的解析变得越来越多。由此,人们认识到,本体研究和相关技术发展将为“语义网”(SW)的实现提供基础。在知识体系概念下,SW 可以浓缩信息核心内容,将知识集成、处理,在网络系统中发挥更具“语义”导向的作用^[7]。

随着信息技术的飞速发展及其在学术领域的广泛应用,网络化、数字化信息资源的数量迅速增长,表现形式多样化,发布渠道多元化,因此,对资源的描述正变得越来越困难,越来越复杂。同时,信息技术的发展使用户可以广泛参与到学术资源的组织和描述活动中,用户不再仅仅是资源的创建者和使用者,更成为新一代的资源描述者和组织者。他们的参与将对信息资源的深度内容揭示起到不可低估的作用。而基于本体论的文献信息组织方法为用户参与信息组织提供了可靠的技术支持。

[参考文献]

- [1] 常春.数字图书馆信息组织[M]//潘淑春,常春.数字图书馆研究与发展.香港:中国中外新闻出版社,2006:187-228.
- [2] Dewey Decimal Classification [EB/OL].[2008-05-04].www.oclc.org/dewey.
- [3] UDC Consortium [EB/OL].[2008-05-07].http://www.udcc.org/outline/outline.htm.
- [4] LIBRARY OF CONGRESS CLASSIFICATION OUTLINE[EB/OL].[2008-05-08]http://www.loc.gov/catdir/cpsol/lcco/lcco.html.
- [5] Google Introduction-What was Google doing before it became this big?[EB/OL].[2008-05-08].http://www.astahost.com/info.php/google-introduction_t12942.html.
- [6] 常春.大型 ontology 构建工具 KAON 的使用和评价[J].现代图书情报技术,2004(8):14-17.
- [7] GROBELNIK M. Automated knowledge discovery in advanced knowledge management [J], Journal of Knowledge Management, 2005(5):132-149.

[作者简介] 潘薇(1979-),女,北京人,馆员,硕士,研究方向:信息管理;喻浩(1975-),男,北京人,馆员,硕士,研究方向:网络技术应用与系统开发。

