

●马张华

计算机分类检索

ABSTRACT The functions of computer-based classified retrieval are:
1) improving group retrieval performance of machine retrieval system; 2) constituting new retrieval strategy in a collocative way; 3) Using both the thesaurus and the key words of retrieval system to optimize the retrieval system; 4) increasing retrieval entrance; 5) promoting the compatibility and interchangeability between retrieval systems. China should combine classification retrievals as the key point of study.

SUBJECT TERMS Computer search; Classified search — Studies

CLASS NUMBER G354.4

图书馆自动化日益发展,分类法作为一种检索手段,是否能有效地应用于计算机检索系统,是世界图书馆界普遍关心的问题。各国对此进行了各种探索和努力。本文试图对计算机分类检索的进展、方法、意义等提出看法。

计算机分类检索的进展

计算机分类检索经历了一个探索、发展和深化的过程。

最先将分类法应用于机械系统的努力可以追溯到本世纪40年代。但是真正将分类法与计算机检索结合,则是从60年代开始的。其中影响最大的试验,是在美国物理学会支持下,由弗里曼和艾瑟顿进行的。1965年开始,历时数年,包括将《国际十进分类法》(UDC)类表转换成机读形式、利用UDC类号进行检索操作和对检索结果进行评价,1968年建成第一个联机检索系统。在试验的最后报告中,弗里曼对UDC作为一种机检语言加以肯定,认为尽管它不能象专为机检

设计的索引语言那样有效,但用于批处理和交互检索不存在问题。^[1]与此同时,美国 and 欧洲国家的一些学者分别进行了各种研究和探索,成功地将UDC应用于计算机索引编制和定题服务系统。印度文献研究和训练中心也对冒号分类法机检的可能性进行了实验。这一系列试验为分类法在机检系统中的应用提供了可能。

70年代是计算机分类检索的发展时期,分类检索的计算机应用成为文献领域自动化关注的焦点之一。由于美国长期形成的主题词检索的传统,弗里曼等人的试验未能在本国得到推广,但却在大洋彼岸的欧洲引起巨大的反响,英、德、瑞士等不少国家的图书馆纷纷利用UDC建立自己的检索系统。许多欧洲国家备有分类检索功能的机检系统、数据库等就是在这时逐步发展起来的。美国的一些文献单位,如美国的物理学会,则选择编制简单的分面类表作为机检工具。许多国家建立的机读目录一般都分别为《杜威十进分类法》(DDC)、《美国国会图书馆图书分类法》

(LCC)、UDC 等设立了可检字段,用户可以利用这些类号检索。到 70 年代末,单是利用 UDC 进行计算机试验或应用的国家就多达 27 个。

80 年代起,计算机分类检索的研究和应用表现出向纵深发展的趋势。随着计算机技术的发展,各国不再满足于对分类语言的简单应用,力图利用分类法的特点,通过与其它检索手段的结合,使机检系统高度完善。美欧各国的努力最为突出。美国 80 年代研究的重点之一是 DDC 和 LCC 的联机检索问题。经过多年准备,1984 年在 OCLC、美国图书资源委员会和森林出版社支持下,美国对 DDC 作为机检工具的有效性进行了一次联机试验^[2]。在这一试验的推动下,美国一些分类学者纷纷对 LCC 输入联机系统的可行性进行探讨^[3,4],欧洲国家则是在原有基础上对应用 UDC 机检加以完善。比较典型的做法是,在机检系统中建立分类号—主题词对应系统,促进两者的结合使用。如瑞士苏黎士技术学校、芬兰的一些大学图书馆大多采用此法作为改进检索系统的手段^[5]。

计算机分类检索的一个最新进展是美国国会图书馆 1987 年开始进行的 USMARC 分类格式的研制^[6]。这一格式是专为系统之间分类记录交流和作为存储分类数据的标准设计的,可用于联机系统中分类数据文档的控制、管理和应用。该格式于 1990 年被批准为暂用文本,目前正在国会馆进行试验。可以预言,它必将成为联机分类检索的有效工具。

到目前,机检系统的主要检索途径仍然是主题法。但分类法的检索作用日益受到重视,逐步成为不可阻止的发展趋势。

计算机分类检索方法

计算机分类检索方法和分类数据的完备程度及采用的软件密切相关。从目前已建立的机检系统看,可按照分类数据的完备程度大致分为三种:只存入类号;除存入类号外,

备有类表、索引的机读文档;除存入类号外,同时建有分类号和主题词的对应系统。三种系统形成三种不同检索方法。

1. 类号检索。直接使用类号检索是分类机检的基本方法。机检系统通常根据输入的类号,首先显示该类的文献名称一览表。用户可有选择地查阅有关文献的详细著录,比手检目录更灵活、方便。一般须预先知道所需要的检索号码,使用时也不能通过机读类表对类目进行浏览和选择,但只要配有相应软件,也可以通过分类系统进行各种形式的检索。如弗里曼等人建立的物理学文献检索系统中,可通过 UDC 类号进行下述检索:

(1)扩充下位类号检索。如使用类似右截词的方法,输入“621. 039. 5?”,检索“621. 039. 5 核裂变反应器”及其下位类目的所有文献。

(2)扩充中间类号检索。如在主表号码和复分表号码组成的标记中,可使用类似中截词的方法,输入类号表达式“621. 035. 5—78”,检出有关各种核裂变反应器安全设备的文献。

(3)布尔逻辑检索。如输入“621. 039. 5? AND NOT 621. 039. 524”,可检出除热核反应器外其它各种核反应器有关的文献。

类号检索不须设立相应的类表机读文档,是目前各国使用最普遍的分类机检方法。

2. 结合类表文档检索。备有机读文档的机检系统通过文档中类名与号码的联系,可以从语词着手查找类号,对类表或索引进行显示浏览,具有较大的机动性、实用性。弗里曼等 1968 年建立第一个备有 UDC 文档的联机目录以来,随着机检技术的发展,这类系统得到了极大改进,其中最典型的是 1984 年美 OCLC 等建立的杜威联机目录。该目录通过将 DDC 输入机检系统,可以利用类表、注释、索引中的语词与分类号之间联系进行 4 种类型检索(见表 1)。

检索操作时,杜威联机目录一般先提供

表 1 杜威联机目录 4 种类型检索

检索类型及代码	含义	显示内容	下一步可能的选择
SA	输入语词浏览和检索相关索引中接近该语词的主题	显示一级包括该词的字母索引款目	1. 浏览索引。2. 选择索引款目直接检索文献。3. 从列出的索引款目查阅类表
SS	输入语词, 浏览和检索与该主题相关的类目	显示包括该主题的一组类目	1. 浏览类表 2. 选择类目检索文献
SD	输入语词, 直接查找该主题对应类目的文献	显示该主题的书目记录数量	检索该主题的文献
SC	输入类号, 浏览和检索与类号相关的类表	显示包括该类号的类列	1. 浏览类表 2. 选择类目检索文献

4 种主题检索选择的菜单, 用户可在输入检索代码及主题词或类号后加以使用。比如, 若用户选择“索引查找”, 在输入“SA 多相分析”后, 即可获得该主题所在的一组字母索引条目, 使用者可选择某一条目查阅文献或从索引条目转入类目表, 查核其类目结构或注释。

若用户选择“类表查找”, 在输入“SS 多相分析”后, 联机目录即显示该主题所在类目结构。用户可有选择地对上、下位类目进行浏览、选择, 查阅类目注释或按行号显示选定类目的文献。

结合类表文档可以使事先不了解类号的用户直接利用语词进行分类检索。类目索引和类表结构的显示, 则为用户利用其树型结构提供了可能, 便于用户通过浏览, 随时改变检索策略, 调整检索范围, 从而使分类法系统检索的优势充分发挥。

3. 结合分类号—主题词对应系统检索。这类系统的特点是以主题词为类目的替代物, 通过分类号和主题词的联结, 将分类检索

和主题检索结合成一体。以瑞士联合技术大学建立的 UDC—主题词对应系统为例, 该系统在 UDC 码后收入与该类号对应的英、德、法三种语言的主题词、同义词等, 以此为基础生成主题、分类、叙词环境三种文档。其中主题和分类文档为简要字顺叙词表和分类系统, 叙词环境档则用来揭示叙词与相应概念之间的全面联系。如“凝聚物质物理学”的概念环境为^[8]:

O	物理学	53	OQU
	凝聚物质物理学	538.9	OQU
Q	分子系统	539.2	OQU
Q	晶体学	548	OQU
U	结构/凝聚物质	538.91	OQU
U	迁移理论(物理学)	538.93	OU
U	量子态/凝聚物质	538.94	OU

上述款目显示了“凝聚物质物理学”的上位类(O)、下位类(U)和相关类(Q)。类号后的标记揭示各类的概念环境。有的有上位类、下位类, 但无相关类。通过语词与类号的联系为用户提供了该类及相关类的完备信息。

利用该系统可进行下述类型检索: (1) 输入主题词, 利用主题文档检索。(2) 输入分类号, 利用分类文档检索。(3) 通过不同文档之间的转换, 利用叙词环境档或分类档进行扩检或缩检。还可以通过语词表达式, 直接利用类目系统进行下述类型检索^[9]: 检索一语词对应的类目及其下位类; 检索一语词对应的类目及其下位类与另一语词的组配; 检索两语词对应类目及其下位类的组配; 检索一语词除某类外的其它下位类。

结合分类号—主题词对应系统检索不仅提高了从语词角度应用分类体系的便利程度, 而且扩大了分类语言在机检系统中的使用领域。

显然, 上述三种检索类型中, 后两种是对第一种优化和改进。虽然这些检索方式在实际使用中尚属少数, 但它们体现了分类机检的发展方向。

计算机分类检索的意义

在机检系统中,分类检索有下述作用:

一、有利于增加机检系统的族性检索能力。可以利用输入的类目文档进行系统浏览、显示和检索,通过类目的等级系统进行扩检或缩检。

二、有利于通过逻辑组配方式,构成新的检索策略,满足各种形式的检索要求。这类检索无论是使用类号进行,还是在分类主题对应系统中通过语词的组配直接进行,其实质都是逻辑组配和分类结构的结合,可以提供主题法不能提供的检索便利。

三、有利于通过叙词、关键词的结合使用,优化检索系统。其典型方式是,可以利用不同检索形式的互补性,建立起一种可以融合分类主题之长的结构,利用计算机的数据处理能力,提供揭示主题概念之间联系和多途径检索的可能,从整体上提高检索系统的功能。

四、可以增加检索入口。国外对联机目录的研究表明,主题检索迫切需要增加更多的检索点。机读类表本身就是提供检索用主题概念的丰富来源。现代机检系统中,类表、注释及索引中收入的所有语词,都可以通过计算机的处理成为检索入口。利用这些语词与类号及通过类号与文献记录的联系,可以进行多种形式的检索操作。

五、可作为媒介语言促进检索系统之间的兼容和互换。现代通讯技术的进步和联机检索系统的发展,使检索语言的兼容问题日益突出。同一种语文的数据库之间也存在着衔接问题。使用权威性分类法,尤其是那些被翻译成多种语言的分类法,有利于建立起一种媒介机制,通过它避免检索语言不同造成的障碍,促进资源共享。

对于分类语言在机检系统中的作用,各国分类界曾进行多次试验,得出大致相同的结论。以80年代OCLC等进行的DDC联机

试验为例,该项目备有两个目录,其一为传统主题目录,可利用标题词、题名、丛书名、注释中的关键词及类号检索;其二为杜威联机目录,采用除主题法外前文所示的各种检索功能。同时组织用户和馆员为对象进行了两组对照试验:用户按自己选择的主题连续在两个不同目录上检索;馆员则按预先设定对应主题分别在两个目录上进行受控试验。共160个用户在主题目录上检索162次,在杜威联机目录上检索158次。共60个馆员分别在两个目录上各进行检索180次。从检索结果看,用户组分类检索与主题检索相比差距稍大,馆员组则相当接近。两组平均,分类检索的成功率(86%)略低于主题检索(90%),

表2 主题联机目录和杜威联机目录试验结果比较

单位:次, %

检索工具	用户检索试验				馆员检索试验				平均检索效果			
	检索数量	成功率	失败率	共同失败率	检索数量	成功率	失败率	共同失败率	检索数量	成功率	失败率	共同失败率
主题联机目录	162	89	4	7	180	91	6	3	342	90	5	5
杜威联机目录	158	81	12	7	180	90	7	3	338	86	9	5

但仍有5%在主题目录中检索失败的主题在分类检索中获得成功。两者结合,可使检索的成功率达到95%^[10]。这证明了分类法对提高机检效率的作用及将分类语言结合进联机系统的价值。

过去有种看法,认为随着机检的广泛使用,分类法作为检索手段将逐渐被叙词法、关键词法代替。这种观点其实是和迄今为止的检索发展实际背道而驰的。就美国而言,19世纪以来,分类目录在图书馆逐渐消失,分类法实际上已沦为单纯的排架手段。但本世纪60年代后,随着机检系统的发展,分类法的族性检索能力及它和主题法互补的特点使人们日益认识到分类语言在计算机时代的价值。DDC、LCC等的机检试验正是在这一情况下进行的。其实,正是计算机系统的巨大数

据处理能力为分类法发挥检索语言的作用提供了新的可能。机检系统的使用不仅没有使分类检索过时,而且大有促使其“回归”的趋势。可以预言,随着 MARC 分类格式的逐步完善,分类法在机检系统中的使用将不断推向深入。

关于我国的计算机分类检索

我国是个有较长分类传统的国家。新中国成立后,分类语言建设取得了一定成果。有全国统一使用的、基础较好的分类法《中图法》,有在《汉语主题词表》与《中图法》对应基础上编制的《中图分类主题词表》,为进行分类检索打下了良好基础。但对于如何有效地进行计算机分类检索,迄今为止的探索仍是远远不够的。70年代以来逐步建立起来的一些专业数据库,一般均以主题检索为主,不少系统虽然也输入了分类号,但多数只限于按类号进行简单检索。就总体而言,对分类机检的研究和发达国家比较还存在着较大差距。应当在汲取国外经验的同时结合我国的特点加以改进。

根据目前情况,我国的分类机检应当逐步把结合类表检索作为研究重点,集中力量对在机检系统中使用《中图法》检索进行探索。采用结合类表检索是改进分类检索效能、优化检索系统的有效手段。我国绝大多数检索系统都已输入了《中图法》号码,采用这一方式必将为用户利用《中图法》检索带来极大便利,改善机检系统的实际效果。

同时,应当尽快建立《中图法》的机读文档。类表机读文档不仅有助于类表的修订和改进,而且是和机检密切联系的。目前国外的一些大型类表,如 DDC、UDC 等都已建立起机读文档,甚至连 LCC 这样体积庞大的列举式类表,也正在为建立机读文档努力。《中图法》作为全国性权威分类法,理应组织力量,尽快加以解决。在这一过程中,应注意对 USMARC 分类格式进行研究。该格式是专

为分类法在联机环境中的应用设计的,曾在编制过程中广泛参考了 LCC、DDC、UDC 等类表,可根据建立《中图法》文档的需要加以汲取。

还应有计划地进行分类机检目录的研制和试验。应特别重视分类机检目录的设计。可选择目前机检系统中采用《中图法》类号的文献记录和机读类表结合,在使用其它项目检索的同时,探索利用语词和类号对类目结构进行浏览和检索的规律性,逐步获得经验,使研究不断深入。

参考文献

- 1 Foskett, A. C., The Universal Decimal Classification, Clive Bingley, London, 1973
- 2, 7 Markey, k. and Demeyer, A. Dewey Decimal Classification On-line Project; Integration of Library Schedule and Index into the Subject Search capabilities of an On-line catalogue. International Cataloguing July/September. 1986. pp31-34
- 3 Lois Mai chan, Library of congress Classification as an On-line Retrieval Tool: Potentials and Limitations. Information Technology and Libraries 5(3): 181-192
- 4 Williamson, N. J. The Library Congress Classification: Problems and prospects in On-line Retrieval. International Cataloguing October/December 1986 45-46
- 5 Gilchrist, A. and Strachan, D., eds. The UDC essays for a new decade London: Aslib, 1990, 35-46
- 6 Guenther, R. S. The Development and Implementation of the USMRC Format for Classification Data. Information and Technology and Libraries. June 1992 pl20-131
- 8, 9 Loth, K. and Funk, H. Subject Search in ETHICS on the Basis of the UDC. In Gilchrist, A. and Strachan, D., eds. The UDC: essays for a new decade. London: Aslib, 1990, 35-46
- 10 Markey, K. Findings of the Dewey Decimal classification On-line Project. International Cataloguing April/June 1986. PP15-19

马张华: 1976年北京大学图书馆专业毕业,现为该校信息管理系讲师。已在多种刊物上发文8篇。参编专著2种。通讯地址: 北京大学, 邮编: 100871。(来稿时间: 1993-06-05。编发者: 黄文田)