

对电子环境下分类检索应用的思考

马张华 陈文广 赵丹群

摘要 分类法作为一种系统检索方法,在知识组织系统中有其独特的价值;分类法在电子环境下的应用特点、形式,甚至管理方式等都发生了巨大变化,应在重视与检索实践结合的基础上汲取、探索;文献分类法在电子环境下的改造,包括改进专业领域的完备性、多维系统的构建和标引方式改进等多个方面,底层类表和上层类表结合应用是多维类表构建的基本方式。图1。参考文献11。

关键词 分类检索 文献分类法 多维类表构建

分类号 G254.335

ABSTRACT As a kind of systematic searching method, classification has its special value for knowledge organizing system. Substantial changes have occurred on the using characteristic, form and even the managing form of classification, which should be absorbed and explored with the involvement of retrieval practice. The change of documental classification under electronic environment includes the improvement of its completeness in special field, the construction of multi-dimensional system, the improvement on the way of marking, etc. The combined application of bottom-scheme and top-scheme is the basic way of constructing a multi-dimensional system. 1 fig. 11 refs.

KEY WORDS Searching by Classification. Document Classification. Construction of multi-dimensional scheme.

CLASS NUMBER G254.335

电子环境的出现对各种传统知识组织方法的应用和发展是一个巨大的促进和挑战。分类法作为知识组织系统的一种基本类型,在电子环境下的文献检索应用中是否仍然有其使用价值?其应用方式究竟有哪些发展和变化?为了适应电子环境下使用的需要,传统的文献分类法应该如何加以改进?本文试图对上述问题进行概要讨论。

1 分类法是不可缺少的检索方法

在当代学者中,美国学者 Gail Hodge 最早使用知识组织系统(KOS)这一术语概括各种信息组织的形式,并试图对其进行系统梳理和研究^[1]。但他在论及分类法时,只重点强调了它的排架功能,而未提及检索方面的应用。这种表述显然是与各国图书馆近二十年来分类法应用中重排架、轻检索的使用现状相联系的。需要指出的是,文献分类法的这种应用现状,并非完全是分类语言自然应用的结果,而是与个别国家分类法应用传统和使用定式强势扩展相联

系的。长期以来,不同国家的图书馆在分类法的应用方式上一直存在着差异。例如,美国图书馆习惯上只将分类法作为排架方法,同时建立字典式目录或分列式目录,提供主题及题名、著者等的检索系统,一般不再建立分类目录;而欧洲多数国家的图书馆通常在采用分类排架的同时编制手检分类目录,将它与其他目录并列使用,如著名的《国际十进分类法》(UDC)就是根据文献检索的需要编制的;中国的图书馆长期以来则在采用分类排架的同时,建立分类目录、著者目录、书名目录组成的目录体系,并将分类目录作为其中心目录。正是由于使用传统上的这种差异,美国图书馆在计算机检索系统发展过程中,一直未有效发展基于分类系统的检索途径。尽管在20世纪80年代进行的一系列分类检索试验中,美国学者发现分类法有助于改进检全率,认为应将分类法纳入电子检索系统^[2-3],但其主要的计算机检索系统中通常只提供分类号检索,一直未提供适合分类特点的浏览检索形式,极大影响了分类途径的使用

效果。美国在电子检索系统发展中的地位,深刻影响了各国图书馆文献检索系统的发展。到目前为止,我国从国外引进的一些大型图书馆检索系统,如国家图书馆、北京大学图书馆等的书目检索系统,其分类途径大体采用了相同的模式。使用这类系统进行分类检索,普通用户通常只能在以其他途径检出资源时,利用从资源数据中获得的分类号进行查找,且始终未显示类名,增加了用户使用的困难,既不能结合类目结构扩充,也没有提供主题索引角度查找使用的可能,使得分类途径形同虚设。大型图书馆的检索实践同时也深刻影响了国内图书馆检索软件的开发,尽管近年来一些书目检索系统开始在分类检索应用方面做出了一些努力,但仍然只处于起步阶段。图书馆作为分类传统最为悠久的文献机构,其对分类检索的这一使用状况对电子环境下分类法的使用和研究的负面影响是不言而喻的。据说有的网络公司为了判断网络分类目录的开发价值,曾到北京一些大学了解学生对分类目录的看法,发现多数学生不知道分类目录为何物,更不了解分类法在网上的使用价值。这对一个具有长期分类法应用传统的国家来说,显然是极不正常的。

实际上,各种知识组织系统各有其特点和适用性,是一种相互补充、结合应用的关系。Hodge在其论文中将知识组织系统的类型大致分为术语表、分类表和关系表三大类。其中,规范文档、词汇表和地名词典归入术语表;标题表与文献分类法、知识分类法(taxonomy)、范畴表属分类表;叙词表、语义网、本体(ontology)则归入关系表。就检索特点而言,规范文档、标题法、叙词法等直接以自然语言中的语词为标识,优点是更加灵活、通用、专指,并且标题法、叙词法有一定的相关检索能力;分类检索的特点则是它的系统性,是按照一定的关系完整揭示一个领域资源的能力,就系统检索功能而言,包括未建立与分类结合的传统叙词表在内,均无法与分类法相比。关系表中的语义网和ontology作为新出现的知识组织类型,更加重视知识关系的全方位揭示,比较符合当前使用的特点,尤其是ontology,是从推理应用的角度出发构建的,对关系的揭示更加充

分、完备,但目前只局限于某些领域,覆盖范围比较小,且仍处于发展和试验过程之中。分类法作为一种通用性的分类工具,则有更大的涵盖面。同时,尽管传统分类法只能从一定角度揭示知识之间的关系,但从后文的论述可以看到,在电子环境下,这一不足可以通过多维系统的构建予以解决。在图书馆和文献数据库中一直保有大量质量较高的分类数据的情况下,利用已有的数据资源开发适合分类特点的电子应用形式,无疑具有不可替代的作用。

显然,各种知识组织系统特点不同,各自承担不同的任务,从发展的角度看,应是一种并行和结合应用的态势。即使将来ontology等得到充分开发,分类法也仍将与其他组织形式并存,发挥适合其功能的作用。实际上,它们在电子环境下的分工以及结合使用,会随着应用不断融合、调整,切不要用传统的眼光看待,造成发展的不平衡。比较而言,目前图书馆对分类检索应用的研究,显然相对不足。

2 分类法在电子环境下的应用形式及其发展

在对分类法如何应用于文献检索的讨论中,一些专业人员往往习惯于按照传统分类语言的观念进行研讨,但实际上,检索语言的使用效果与应用环境联系密切,在电子环境下,尤其在网络检索中,分类法的应用特点、方式甚至管理形式等都已发生了巨大变化。

就应用特点而言,传统列举式类表作为一个等级系统,是否能够提供浏览以及从主题角度的查找可能,是有效使用分类结构的重要条件。与手检系统相比,计算机的处理能力以及屏幕显示界面为分类法的使用提供了巨大便利。就网络应用的情况看,其使用形式包括:①可以直接以类名标注类目,更加适合于浏览;②索引与类表联系密切,可以直接以语词检索类表,从而方便深度类表的使用;③可以同时采用多种浏览形式,包括在类目展开中结合采用从字顺角度浏览查找的形式,如Open Directory、维基百科的分类索引^[4]等;④可以采用超文本链

接的方式,通过重复反映、类目参照等形式,克服分类法按学科分散的不足,完整提供一个专业领域的资源;⑤可以使用多维揭示的方式,提供多种入口,满足用户从不同角度检索的需要。这一系列变化表明,分类法在电子环境下具有比传统手检目录更强的通用性和灵活性。

等级列举式类表的不足之一,是对于海量信息的适应能力问题。不少人因此对分类法在网络环境下的使用价值表示怀疑,认为分类法不适用于网上海量资源的处理。这种观点忽略了分类法应用的发展和多样性。实际上在网络环境下,分类法结合不同应用情况发展了多种形式,至少包括:

(1)将分类体系作为有层次的浏览检索工具,包括通用性检索工具与专用系统。前者如基于传统文献分类法的主题网关和主题指南,通常在精选资源的基础上编制,比较典型的有国外的 BUBLINK^[5]、Yahoo! Directory、Open Directory,以及国内的搜狐、新浪分类目录等。后者如维基百科的分类索引,该索引以维基百科的条目为对象,完全是根据百科条目的检索需要编制的。

(2)作为简约的分类工具用于通报性系统。如每日新闻资源的分类,由于分类对象数量有限,通常只需要区分一、二级,可在自动分类的情况下达到较高的准确度^[6];天网曾试图建立起一个3级层次的自动分类系统,但这类系统的标引准确度是随等级的增加递减的,效果不如前一种理想。

(3)作为范畴分类的工具,结合关键词检索应用。如 Scirus^[7]对科学资源进行自动范畴分类,将资源归入20个学科领域,作为检索匹配时提高检准率的手段;再如万方知识检索平台、清华同方的期刊论文数据库等也将概要分类作为关键词检索下提高检准率的基本手段。

(4)作为动态自动聚类的工具。这种方式可以通过对检索结果的聚类,提高检准率。包括用作检索优化和二次检索两种情况。前一种如 Ask^[8],后一种如 Vivisimo^[9]。后者相当于通过动态聚类提供一种潜在的深度分类体系,是海量环境中的一种新的分类形式。

(5)自由分类法(Folksonomy)。这实际上是一种在用户参与基础上建立的自由词标引。其优点是能反映用户的使用需要和检索特点,并可结合使用频率对语词标签进行选择,适用于特定社区、特别是多媒体资源的检索应用;不足是缺乏必要的关系揭示和同义控制。随着互联网的发展,如果将其作为网上词汇来源,用作关键词检索的排序因素,或作为构建分类系统的依据,将会更加有益。

上述各种形式中,按照层次方式进行深度分类的一个突出问题,是如何解决人工编制与互联网资源数量太大的矛盾。为了解决这一问题,ODP发展了在网络用户自愿参与的基础上形成的编制管理形式,成为最早出现的Web2.0形式之一,维基百科的分类索引也是采用相同的方式编制的。广大网络用户参与的结果,使得类表构建和发展具有开放性的特点,探索了众多新的设类和展示方法,为分类法在电子环境下的构建和应用编制方法改进等拓展了思路。目前,ODP已超过Yahoo!主题指南成为规模最大的网络分类系统;维基分类索引除具有良好的检索功能外,还具有优化条目设置、推进条目撰写的作用。图书馆领域则曾经探索了各种网络分类法编制的协作形式,比较典型的如荷兰图书馆曾采用的分工协作编制形式以及欧洲多个国家对于兼容系统的探讨^[10]。从发展的角度看,有针对性地建立专门(如多媒体资源)、专业目录是开发网络优质资源的一个有效手段。与此同时,自动分类则是分类体系应用的另一种方式,其中,动态自动聚类系统提供了与整体分类不同的应用形式,通过与关键词检索结合,使得分类能够以动态的方式灵活融入海量环境的应用,极大拓展了分类法应用的空间。可以看到,正是多种分类形式的结合,构成了分类法在网络环境下的整体应用。

在关键词检索系统的发展中,曾出现网络检索工具汲取文献数据库的技术,并结合网络特点探索、发展,其后文献数据库又反过来从网络的发展中受到启发,加以改进的过程,这可以从近年万方、同方检索系统的改进中明显感受到。目前图书馆分类检索系统的发展改进,可能也需要经历类似的过程。尽管在各种信息机构中,图书馆领域对分类语言的研究时间最长,

其类目体系与文献标引实践的联系也最为紧密,几乎所有的网络分类工具,都十分重视汲取文献分类法的经验和方法,但在与检索实践的结合上,也许现在是图书馆向网络学习的时候了。与图书馆领域的研究相比,搜索引擎的研究者更重视分类法的检索应用,重视在检索语言与应用结合的情况下对分类语言加以改进,并采用开放的发展方式。显然前面提到的一系列实践发展都具有极大的参考、借鉴价值,但更为重要的是,应调整观念,重视分类语言与检索实践结合的探索,改变文献单位对检索语言应用研究的不足。应该看到,到目前为止,电子环境下检索界面的研制改进,尤其是控制语言的使用方式,仍存在较大的研究空间。传统文献检索系统的最大优势之一是高质量的元数据,事实上,许多网上的应用形式,依据标引数据同样可以实现,而且效果更好。以自动聚类为例,在分类检索的同时,于类下提供基于标题的聚类显示;在主题检索系统中同时采用分类优化、标题聚类浏览等形式,用以提高检准率,其效果将远优于基于文本数据的自动聚类,而难度远小于后者。在目前主题标引数据建立标题的情况下,上述应用通常只要对软件进行简单的改进就可以了。此外,文献单位资源类型、用户需求多样,存在着多样化发展需求,如果能采用图书馆2.0的方式,以开放式的态度加以探索,就一定能极大改进检索语言的编制和应用水准。当然,要做到这一点,应加强专业人员与软件人员的沟通,需要有相应的熟悉检索语言和具有中文开发能力的研发团队。由我国软件人员与检索语言专家结合来开发和优化我国大型文献单位的检索系统,有利于根据我国检索语言的特点充分开发检索功能,并必然具有更好的持续发展能力。

3 关于传统分类工具的改造与完善

电子环境下分类法应用中一个迫切需要探索和解决的问题,是如何将传统文献分类法基于排架和手工检索的单类表,改造为适合电子环境使用的多维类表,使其能更好地满足用户需要。

传统文献分类法是图书馆等文献单位组织

文献资源的基本形式,已经有数千年的历史,基本上是依据文献排架和手检工具的特点编制的一个从学科角度组织和揭示的线性系统。其优点是长期文献组织的实践基础,对知识关系揭示系统,严格遵循文献保证原则和用户使用需要,对手检工具有较好适用性。但直接将此法用于电子环境下检索的不足是:只能从学科角度检索,无法有效提供一个主题领域的完整资源;检索入口单一,不能满足不同角度检索的需要;原有的类目参照是根据手检工具的特点设置的,相关揭示不够充分;对复合主题资源的处理,如对应应用、比较关系主题等的分类,分别是按内容归入其中的一个方面,难以适应不同用户的需要,等等。因此有必要对这类处理进行相应的改造,发展出适合电子环境下使用的形式。显然在目前的使用背景下,应根据技术环境的变化和用户的使用需要,汲取网络分类目录类目构建的经验,在原有基础上对类目体系进行相应调整,改造的核心是通过超文本技术的使用,改进类表各门类的完备性和多维性,使其有更好的适用性。整体而言,这种改进至少应包括下述内容:

(1)通过重复反映,改进类表的专业领域的完备性,使用户可以同时从学科或主题的角度检索。以《中图法》为例,其类目体系的设置,是根据综合性类表的特点,按学科均衡展开类目的,与一个专业有关的内容,均按照所属的学科分散在相关门类之下。结合采用超文本链接,可以通过重复反映,将因按学科设类分散的某一专业领域的类目完备地加以揭示。这一改造涉及的对象包括:将原分散在相关学科门类的相应主题类目加以集中,使得专业领域相应完备;将原有交替类目加以增补和集中,使得多属性类目能得到充分反映;适当扩展类目参照,使相关主题类的揭示更加完整,从而可以完备检索一个领域的资源。

(2)通过多维系统的构建,改进类表的适用性和易用性。由于传统分类系统是按照单一的引用次序展开的,对复合主题的检索浏览只有一个途径,同时也只能从一个角度组织资源,这一状况必然会影响到系统的适用范围和检索效

果。例如,按照《中图法》经济类的类目设置方法,“工业经济”是按部门为主线展开的,有关各个部门的工业计划与管理的资料,必须在相关部门下查找,而不能在“工业计划与管理”类下找到,这一设类方式既不能集中显示工业领域各种“计划与管理”的资源,也不利于普通用户检索查找。通过重复反映,可以使其同时在相关工业部门与“工业计划与管理”下加以提供,使分类系统具有同时从不同角度集中资源的能力,例如,既可以从各个工业部门的角度,又可以从“计划与管理”的角度同时组织和揭示资源,从而可以满足不同的使用需要。此外,检索入口的增加也极大改进了检索系统的易用性。

(3)结合文献内容进行组配标引,改进类表的使用形式。一些复合主题文献,例如应用关系、影响关系、比较关系等主题的文献,目前在分类处理时基本上是依据排架特点,按其重点分类的。对这类情况,应结合检索需要加以调整,使其得以同时在相关主题下反映。实际上,过去文献分类法中发展的多种组配符号,如UDC设置的“:”、“/”等众多辅助符号,正是供这类情况使用的。例如,主题“计算机在图书馆的应用”,可以标引为G25:TP399(图书馆:计算机应用),检索使用时,以轮排的形式同时在相关类下加以显示。类似的辅助符号,国内《中图法》等大多也已经引入,目前需要解决的只是结合电子环境的使用特点制订规则、以适合的方式加以贯彻的问题。

上述三个方面中,第一方面改造相当于将各个主题领域的类目改造成一个专业类表;第二方面改造着重解决多角度集中、多维入口的问题;第三方面则主要是通过规则的调整,改进复合主题的多元揭示问题。三者结合,有助于将系统改造成适合电子环境使用的、具有全方位揭示主题关系能力的分类结构。

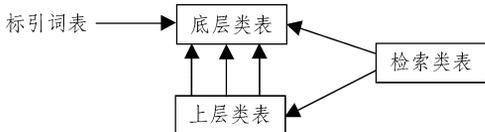


图1 电子类表结构及功能示意图

电子环境下多维类表的结构,本质上是底层标引类表与多层次的上层检索类表的结合。电子类表的结构及其功能可以简单图示(如图1)。其中,底层类表仍然是一个线性的单维系统,它是资源标引的依据,同时也可直接用以检索,通常可以直接使用原有类表;上层类表则是根据检索需要增添的一系列与底层结构联结的专用检索类表,这一类表不用来标引,只提供新的检索途径,其类目通过与底层类表对应类目的链接向用户提供资源,使得系统灵活、多元的揭示成为可能。电子类表的这一结构方式的典型例子是目前网上的主题指南或网络分类目录构建的多维系统^[1]。使用这一方式,可以在不对类表作过多变动的基础上实现。试以《中图法》“F40 工业经济理论”下“工业计划与管理体制”的类目为例:原表中各个工业部门经济的“工业计划与管理体制”的资源归入部门经济,无法于“F402 工业计划与管理体制”下进行集中。在进行多维改造时,只要在F402下加入部门经济下的类目,并将其指向工业部门经济下的对应类目,就可以在原有类表基础上实现多维检索。

- F40 工业经济理论**
- F401 工业结构与体制**
- F402 工业计划与管理体制**
 - F402.1 地质、矿业@ (F407.12)
 - F402.2 能源工业、动力工业@ (F407.22)
 -
- F403 工业建设和发展**
- F404 工业劳动与工资,劳动生产率**
- F405 工业品供销与果例**
- F406 工业如业组织和经按管理**
- F407 工业部门经济**
 - F407.1 地质、矿业**
 - F407.11 工业结构与体制
 - F407.12 工业计划与管理体制 ←
 -
 - F407.2 能源工业、动力工业**
 - F407.21 工业结构与体制
 - F407.22 工业计划与管理体制 ←
 -

上表中 F402 下的子类为重复反映类目,通过超文本链接将其与对应类目联结,即可在原表的基础上对“工业计划与管理体制”的资源加以集中。使用这样的方式,就可以将文献分类法从原来的单维表,改造成多维表,改进其揭示功能。

在文献分类法的多维化改造中应予关注的问题包括:①主题领域相关类目的增补,需要与用户的使用需要相结合,通常应在对知识关系充分了解和对类表编制技术有较好把握的基础上进行,其处理方法类似专业表。②多维系统的充分揭示,在一些情况下需要有相应类目的细化相配合。如上例中,假若“F407.1 地质、矿业”下原表未设置下位类,就无法建立起对应的多维揭示机制。因此在进行改造时,必要时还应根据需要对原表进行细化,使得这类多维揭示可以在适用的层次上得以实现。③必要的度的问题,即对于多维揭示的“维度”加以把握,使多维系统的构建适合使用需要,保持较高的有效性。④在这一应用中,分面理论和分面标记技术的结合应用显然是十分有价值的,应结合电子环境下的使用特点和需要进一步研讨。为保持下层类表的稳定,分面技术通常应更多用于优化上层类表的构建。

此外,复杂主题内容的组配标引,在目前的系统中是依据相应的标引规则确定的。例如应用关系的资源,通常只将其归入应用到的类目;比较关系的资源,通常只按重点归入其中的一个方面,无法从理论、方法的角度或被忽略的比较方面加以揭示,对这类情况,应设法进行补充标引,可探索以人机结合的方式予以解决。例如,利用相应关键词的匹配结合人工判断,发现这类主题的文献。对有关方法及其可能性,我们拟在其他论文中讨论。

我国是一个具有文献分类传统的国家,两千多年来,分类目录一直是我国主要检索工具。近年来分类法在电子环境下的应用现状,尤其是文献分类法在图书馆和文献数据库中未能有效开发的状况,极大制约了分类法检索功能的发挥,使得图书馆与文献数据库拥有的大批高质量的分元数据无用武之地,十分可惜。显然,对传

统分类法改造的研究,不仅有利于推进分类法在电子环境下使用规律的研究和探索,而且具有极高的实践价值。相信随着我国分类学界的努力和民族软件的发展,分类系统会逐步成为信息检索的基本途径之一,而适合电子环境特点的分类检索系统也会在这一过程中得到完善。

参考文献:

- [1] Hodge, G. Systems of Knowledge Organization for Digital libraries. Beyond traditional authority files [C/OL]. Washington, DC: the Council on Library and Information Resources. [2008-08-10]. <http://www.clir.org/pubs/reports/pub91/contents.html>.
- [2] Markey, K. Findings of the Dewey Decimal Classification On-line Project [J]. International Cataloguing, April/June 1986: 15 - 19.
- [3] 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 [J]. International Cataloguing, July/September 1986: 31 - 34.
- [4] [2008-07-10]. <http://zh.wikipedia.org/wiki/Wikipedia:分类索引>.
- [5] BUBLINK [OL]. [2008-07-10]. <http://bubl.ac.uk/>.
- [6] 马张华. 论自动标引的实际应用 [J]. 图书情报工作, 2003 (2): 48 - 51.
- [7] [2008-07-10]. http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf.
- [8] Ask [OL]. [2008-07-10]. <http://www.ask.com/?o=10182>.
- [9] Vivisimo [OL]. [2008-07-10]. <http://vivisimo.com/>.
- [10] 马张华. 文献分类法在网络资源组织中的应用 [J]. 图书情报工作, 1999 (12): 24 - 29.
- [11] 马张华. 网络分类法类目体系研究 [J]. 图书情报工作, 2001 (2): 36 - 40.

马张华 北京大学信息管理系教授。通讯地址: 北京。邮编 100871。

陈文广 赵丹群 北京大学信息管理系副教授。通讯地址同上。

(收稿日期: 2008-07-20)