

21世纪高等学校规划教材



XINXI JIANSUO ZONGHE JIAOCHENG

信息检索综合教程

费业昆 主 编
瞿其春 副主编



中国电力出版社
CHINA ELECTRIC POWER PRESS

21世纪高等学校规划教材



XINXI JIANSUO ZONGHE JIAOCHENG

信息检索综合教程

主 编 费业昆

副主编 瞿其春

编 写 张素芳 邵国川 江树青 张文忠



中国电力出版社
CHINA ELECTRIC POWER PRESS

内 容 提 要

本书为 21 世纪高等学校规划教材。全书主要内容包括信息检索基本理论, 信息来源与参考工具书, 搜索引擎、博客及 RSS, 中文电子文献信息的检索, 主要西文检索工具的使用, 全文型电子期刊数据库的检索与使用, 特种文献的检索与使用, 科技创新等。

本书内容全面、系统, 图文并茂, 尤其注重对实践知识的运用与指导, 为读者准确、快速、全面查询信息提供全方位的帮助, 使读者在了解并掌握信息检索基本理论和使用方法的同时, 能比较全面地了解在科研与学习过程中对所需学术信息的需求, 体会到在网络环境下新的学术信息交流环境和交流模式。

本书是一本既有理论又重实用性的教材, 可作为本科生、研究生学习和实训教材, 也可作为从事专业技术研究和应用的技术人员工作时参考。

图书在版编目 (CIP) 数据

信息检索综合教程 / 费业昆主编. —北京: 中国电力出版社, 2013.2

21 世纪高等学校规划教材

ISBN 978-7-5123-4049-7

I. ①信… II. ①费… III. ①情报检索—高等学校—教材 IV. ①G252.7

中国版本图书馆 CIP 数据核字 (2013) 第 027899 号

中国电力出版社出版、发行

(北京市东城区北京站西街 19 号 100005 <http://www.cepp.sgcc.com.cn>)

航远印刷有限公司印刷

各地新华书店经售

*

2013 年 4 月第一版 2013 年 4 月北京第一次印刷
787 毫米×1092 毫米 16 开本 19.25 印张 468 千字
定价 35.00 元

敬 告 读 者

本书封底贴有防伪标签, 刮开涂层可查询真伪
本书如有印装质量问题, 我社发行部负责退换

版 权 专 有 翻 印 必 究

序

1984年,国家教委印发了《关于在高等学校开设〈文献检索与利用〉》课的通知后,全国高等学校陆续开设《文献检索与利用》课程,近30年来全国图书馆界的学者编写出版了一系列教材,使得该课程得到了普及,教学水平逐步提高。但是随着信息技术和网络技术的迅猛发展以及存储技术、检索技术和数据库技术的发展,文献信息量呈爆炸式涌现,所有这些不仅丰富了信息检索的内容,也对本课程的教材建设提出了新的要求,也就是说,教材内容必须反映本领域的新进展和新创新思路。为此,由长期从事图书情报工作的部分学者合作编写了《信息检索综合教程》一书。

通览全书,这是一部既有理论又重实用的教材,读者主要面向本科生,也可以作为研究生和一般读者学习使用。该书以向读者指点迷津,提供文献资料线索为主导,内容新颖、实例丰富、层次分明,充分体现了当今科技创新的特色和当今数字化时代的全貌。该书有利于激发学生的信息意识和创新意识,培养他们在浩如烟海的知识海洋中快速准确地发现和利用有价值的知识信息能力,驾驭知识的创新能力,同时对广大研究生、教师、科研人员更好地把握本学科领域的现状和发展趋势,开辟广阔的思维空间,关注相关学科领域的发展,也有重要的参考价值。我认为该教材是一部具有较强可读性、可操作性和具有创新知识的信息检索读本,是一部大有裨益的好书。

我从事图书馆工作30余年,其中大部分时间是从事图书馆参考咨询和图书馆管理工作。目睹了中国图书馆从传统的手工管理到自动化网络化管理的演变过程。当我看到图书馆员和读者从繁琐的手工检索中解脱出来时,看到几代图书馆人的梦想成为现实时,感到由衷的欣慰。我真诚地希望这本书能使更多的人享受到现代文献信息检索带来的光明前景。

费业昆

2012年8月25日于安徽合肥

前言

21世纪是社会信息化的时代,信息化日益成为衡量一个国家、民族、地区竞争力的主要标志。以计算机技术和网络技术为核心的现代信息技术的进步,不仅加快了信息交流与传递的速度,而且也增强了人们获取、存储、分析、处理以及利用信息的能力。随着社会大众信息素养的逐步提高,人们越来越意识到信息资源是人们进行学习、研究、知识创新、科学发现、技术发明和事业发展的基础和重要支撑,而信息的有效存储和检索是科学管理和有效获得信息资源的技术手段。对于在校的大学本科生的教学科研人员来说,掌握现代信息社会的资源存储、检索原理与基本技能尤为重要。

为了适应建设科技创新型国家和高等学校文献检索课教学的需要,满足在校大学生和教学科研人员的学习需求,特组织编写本书。本书从系统的角度出发,对于理论性检索原理做了较大篇幅的增加,重点介绍了国内外重点常用数据库特点、检索方法及应用,并增添了最新的数据库及其原理的扩展介绍和若干比较新的信息检索方法,对图书馆界最新的理论和实践的创新观点论述也都做了较为详细的描述。本文内容丰富、图文并茂,尤其注重对实践知识的运用与指导,为读者准确、快速、全面查询信息提供全方位的帮助,使读者在了解并掌握信息检索的基本理论和使用方法的同时,能比较全面地了解科研与学习过程中对所需学术信息的需求,体会到网络环境下新的学术信息交流环境和交流模式。

本书由费业昆主编、瞿其春副主编,全书共八章,其中第一章、第二章由江树青编写,第三章、第七章由邵国川编写,第四章、第八章由张素芳编写,第五章、第六章由瞿其春编写,张文忠为本书编写提供了资料,费业昆拟定提纲,并负责统稿审查全书。

本书在编写过程中,借鉴了同行专家的研究成果和部分内容,在此一并向他们表示由衷的感谢!由于本书编者的学术水平有限,书中难免有疏漏之处,恳请广大读者多提宝贵意见或建议,以便于今后的修订和完善。

编者

2012年8月

目 录

序

前言

第一章 信息检索基本理论	1
第一节 检索语言	1
第二节 检索理论及计算机检索系统	10
本章小结	25
本章思考题	25
第二章 信息源与参考工具书	27
第一节 信息源	27
第二节 参考工具书	32
本章小结	38
本章思考题	38
第三章 搜索引擎、博客及 RSS	39
第一节 搜索引擎	39
第二节 博客	57
第三节 利用 RSS 获取信息	62
本章小结	72
本章思考题	73
第四章 中文电子文献信息的检索	74
第一节 中文检索工具概述	74
第二节 中文图书的查检	82
第三节 中文期刊的检索与使用	101
本章小结	119
本章思考题	119
第五章 主要西文检索工具的使用	121
第一节 《SCI》、《Ei》、《CPCI》(原 ISTP) 三大检索系统	121
第二节 《Inspec》、《CA》、《CSA》、《PQDT》检索	145
本章小结	167
本章思考题	167
第六章 全文型电子期刊数据库的检索与使用	168
本章小结	207
本章思考题	207
第七章 特种文献的检索与使用	208
第一节 专利文献检索	208

第二节 会议文献检索	223
第三节 学位论文检索	232
第四节 标准文献概述	237
第五节 科技报告	247
本章小结	255
本章思考题	255
第八章 科技创新	256
第一节 文献信息的收集、整理、分析与利用	256
第二节 科技查新	278
第三节 学术规范	285
第四节 学术不端检测	293
本章小结	295
本章思考题	295
参考文献	296

第一章 信息检索基本理论

第一节 检索语言

检索语言在信息检索过程中起着十分重要的作用,是沟通信息存储和信息检索两个过程的桥梁,也是沟通标引者和检索者的桥梁。

一、检索语言概述

1. 检索语言的概念

检索语言是指信息标引和检索所使用的一种专门语言,是用以表达概括文献信息特征和检索课题内容特征的概念及其相互关系的一种概念标识系统。所谓“信息标引”是指获取信息内容检索标识的基本方式,所以,检索语言在存储时用以描述文献信息的内容和外部特征,而在检索时则用以表达提问的语言。

任何一种检索语言,无论其表达形式是词、词组或符号,都是一种用概念及其相互关系来表达文献内容特征和信息需求提问特征的标识系统,因此概念及其相关逻辑是检索语言的基础。在信息检索过程中,为了揭示一种事物与其他各种事物之间的区别与联系,达到检索的目的,检索语言不仅要表达事物的不同概念,还要揭示概念之间的逻辑关系。

概念是事物本质属性的概括,是人们对事物的认识上升到理性认识阶段的产物。概念间的逻辑关系有等同关系、属种关系、相关关系三种。其中,相关关系包含交叉关系、矛盾关系和对立关系。检索语言有效地利用了概念的划分与概括、分析与综合这两种逻辑方法来建立自身的结构体系。

综上所述,信息检索语言包含以下三个基本要素:①有一套用于构词的专用字符;②有一定数量的基本词汇用来表述各种基本概念;③有一套专用语法规则来表达由各种复杂概念所构成的概念标识系统。

2. 检索语言的类型

随着信息检索的发展,检索语言也发展出不同的种类,不同的检索语言构成不同的信息标识和索引系统。为用户提供不同的检索途径。

检索语言的类型有多种,不同的划分标准有不同的检索语言,如图 1-1 所示是用来描述检索语言类型的结构图。

(1) 按检索语言的规范化(或受控)程度来划分,可分为规范化的人工语言和非规范化的自然语言。

1) 人工语言。人工语言(Artificial Language)又称规范化语言(Controlled Language)。它是根据文献信息检索的需要,由人工采用规范词而创建的一种受信息检索控制的语言。人工语言使用的控制、规范词是用来专指某个指示宽度适当的概念,供检索时选择。它可以将同义词、近义词、多义词和缩略词等规范在一起,由人工控制。

人工语言的规范处理重在两个方面:一是使一个概念只用一个词汇来表达,这样就避免了多词一义的混淆情况出现;二是使一个标引或检索词只能表达一个概念,这样就排除了一

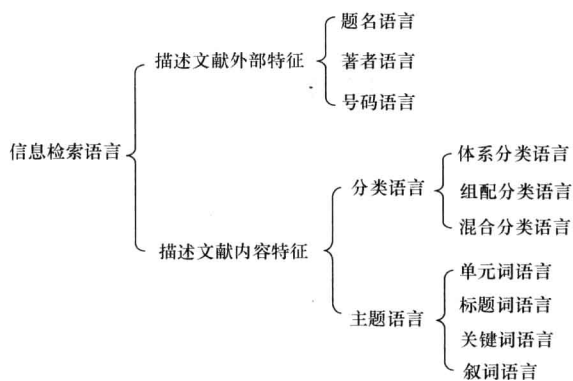


图 1-1 检索语言类型的结构图

词多义的现象产生。用户在检索时可省略对其概念的全部同义词或近义词的考虑，因为在标引时规范词将表示这一概念的其他的非规范词全部替代了，它提供了一种能够有效避免漏检、误检的高效查找检索功能。

人工语言的具体表现形式往往是某种规范语言的词表。凡是有规范词表的检索工具，在主题检索时应先到词表中对检索词进行规范后再检索。人工语言的查全率和查准率均比较高，它的检索效率要高于自然语言。但是，人工语言对于标引来说，工作量大，成本高；对于信息检索来说，用户不容易掌握。

2) 自然语言。自然语言指检索用词是从信息内容本身抽取的，不受人工控制的自然形态词语，即使用非规范词或称自由词。自然语言丰富而复杂，它存在一词多义、多词一义及词义交叉的现象。自然语言有较大的灵活性，使用随意，专指性强，查准率高。它能及时反映最新出现的词汇，反映规范词难于表达的特定概念或新概念，且易被用户掌握。但是，自然语言缺乏对词的控制能力，因此检索效率比较低。

(2) 按检索语言的词汇组配方式来划分，可分为先组式语言和后组式语言。

1) 先组式语言。先组式语言是指在检索实施前已事先组配好的一种检索语言。用户只能使用这种已固定好的检索词组形式去完成检索，它有较好的直接性和专指性，但灵活性差。例如，标题词语言就属于此类。

2) 后组式语言。后组式语言是指在检索实施前未事先组配好的、以单元词等形式出现的一种检索语言。用户在检索时将它们临时组配起来，表达一定的概念，来完成信息检索。后组式语言能提供灵活的组配方式，在计算机检索中得到了广泛的应用。

(3) 按描述文献信息内容特征来划分，可分为分类语言和主题语言。

1) 分类语言。分类语言是用分类号来表达文献主题概念，并将文献按学科性质分门别类地系统组织起来的一种检索语言。它又分为体系分类语言、组配分类语言和混合分类语言三类。利用分类语言编制的分类表主要有《中国图书馆分类法》、《中国图书资料分类法》、《中国科学院图书馆图书分类法》，美国的《十进制图书分类法》等。

2) 主题语言。主题是指文献具体论述的对象和研究的问题。主题语言是直接以代表文献内容特征和科学概念的词语作为检索标识，并按其字顺组织起来的一种检索语言。按选词原则，主题语言又分为单元词语言、标题词语言、关键词语言和叙词语言四种。著名的检索工具《工程索引》(Ei) 的“工程信息叙词表”、《科学文摘》(SA) 的“Inspec 叙词表”及《汉语主题词表》的“轮排索引”，就是根据主题语言编制而成的。

本节接下来重点介绍表述文献内容特征的分类语言与主题语言。

二、分类语言

所谓“类”是指具有许多共同属性的集合，每一个事物都有许多属性，用其某一种属性作为划分依据来对一类事物进行划分就称为“分类”。分类是人类逻辑思维的一种最基本方式。信息检索中的分类检索语言是按学科范畴和知识之间的相互关系列出类目，并用数字、英文

字母对类目进行标识的一种检索语言。在这种检索语言中,各级类目都一一列举,事先予以固定组配,每一级类目都有一定的标记表示,这个标记就是分类号。分类号连同它的类目一起构成分类表。这一套号码便称为“分类表”或“分类法”。它集中反映文献内容的学科系统性及学科与学科之间的相关、从属、派生等关系。

我国图书馆情报界目前广泛使用的是《中国图书馆分类法》。

《中国图书馆分类法》简称《中图法》,是目前我国各类文献信息机构使用最广、影响最大的一部分分类法。1975年正式出版了《中图法》第1版,《中图法》问世后迅速在全国推广应用,不仅图书情报部门类分文献使用,而且在图书发行、各类数据库乃至互联网中也得到应用。分类法的应用是推动其不断发展的重要因素,根据用户的意见和科学发展,《中图法》始终处于不断修订和完善之中。1980年出版第2版,1990年出版第3版,1999年出版第4版。《中图法》,2011年出版了第5版,每一版修订都在不断提高对文献信息整序和检索的适应能力。

(1)《中图法》的基本部类及其序列。

基本部类和基本序列是我国分类法编制中特有的概念。基本部类是文献分类表中为便于各种类目的展开,对知识门类所进行的最能概括最本质的划分与排列,是确立全部类目关系的基础。基本部类的排列次序成为基本序列。

《中图法》也是属于等级列举式分类法,采用五分法设立了5大部类。它以自然科学、社会科学、哲学作为人类知识领域的基本构成建立基本序列,将作为社会主义分类法编制指导思想的“马列主义”和结合文献特殊性设置的“综合性图书”作为部类,形成5个部类的基本序列。

基本大类,又称基本大纲。《中图法》是在基本部类的基础上,将社会科学展开为9大类,将自然科学展开为10大类,连同其他三个部类各自为相对应的大类,共有22个大类,是分类法中划分的第一类目。

(2)《中图法》的结构。

分类法是一个具有内在联系的逻辑系统,其系统功能是通过其结构实现的。分类法的结构分为宏观结构和微观结构。

1)《中图法》的宏观结构。

《中图法》的宏观结构是指它的各个组成部分及其之间的组织方法、相互联系和作用,配合而构成的一个有机整体。《中图法》的宏观结构包括:编制说明、基本大类表、基本类目表、主表、附表、索引、使用手册等。这些部分共同构成《中图法》的整体。

编制说明是对《中图法》的编制目的和适用范围、编制原则、体系结构、辅助表、标记符号、注释等的介绍。

基本大类表是指分类表中列出的第一级类目,它代表这较大学科和领域。

基本类目表,亦称简表,是由基本大类及进一步区分出来的二、三级类目所组成的,是分类法类目体系的框架。

主表,亦称详表,是在简表或大纲的基础上进一步展开的,是各级类目组成的一览表。它是文献分类标引的依据,是分类法的核心。主表按功能分为术语(类名)系统、标记系统、注释系统。

附表,亦称辅助表。实际上是复分表,由分类法的8个通用复分表组成,是主表类目进

行总论复分、地区复分、时代复分、民族和种族复分、通用时间地点复分的依据。

索引,亦称类目索引。它是按类名及类目相关概念名称字顺查找相应类号的工具。

使用手册是详细阐述分类法的编制理论与技术、各类文献规则与方法,是分类法的使用指南。

2)《中图法》的微观结构。

《中图法》的微观结构是指类目的构成要素及其组织。类目是构成分类法的最基本要素,每个类目通常代表某种共同属性的文献集合。《中图法》的类目结构由类号、类名、类级、注释和参照组成,其中类号、类名、类级是必需的。

标记符号,又称分类标记或分类号,是类目的标记符号或代号,是决定类目在分类体系中的位置。

类名是类目的名称,是文献分类法的基本单元之一。类名的作用是不仅规定类目的含义,使其区别于其他类目,而且还规定类目的内容范围。

类级是指类目的级别。代表该类目在分类体系中的等级(划分的层次),显示类目间的等级关系。

类目注释和参照是对类目的含义及内容范围、分类方法、复分方法、特殊书次号编制方法、该类与其他类目关系等进行说明。

(3)《中图法》的类目关系。

《中图法》的类目关系主要有从属关系、并列关系、交替关系、相关关系等。

1) 从属关系。

类目的从属关系是指类目体系中一个类与其直接区分出来的子类之间的关系,即上位类和下位类的关系,又称隶属关系或等级关系。在具有从属关系的类目中,被区分的类,称为上位类,亦称母类;区分出来的类,称为下位类,亦称子类。凡不具有直接从属关系的各级上位类和下位类则通称为上级类或下级类。根据类目划分的等级,人们习惯上将具有从属关系的各级类目分别称为一级类目、二级类目、三级类目等。

从属关系又分为属种关系、整部关系、方面关系。

属种关系是包含和被包含的关系,如化学与无机化学、种子植物与被子植物等。它的特点是:上位类必然能包括下位类的外延,下位类必然具有上位类的属性。

整部关系是整体与部分的关系,是指表示事物整体的上位类和表示事物某部分的下位类之间的关系,如拖拉机与底盘构造、华东地区与上海市等。

方面关系是全面和某一方面的关系,是指表示事物总体的上位类与表示事物某一方面的下位类之间的关系,如中国经济与中国经济政策、物理学与物理学史等。

2) 并列关系。

类目的并列关系指由一个上位类依据某一分类标准划分出来的若干个平级的下位类之间的关系。并列关系是通过同位类来体现的。由一个上位类区分出来的一组下位类互称同位类,如“B95 道教”类下的“B951 道藏”、“B952 经文”、“B953 戒律”、“H954 布教、仪注”、“B956 宗派”等五个类就是同位类,一组同位类总称为一个类别。

类目的从属关系和并列关系在等级结构中使用不同的字体和排列的缩格来显示,也可以根据分类标记的数位来显示。

3) 交替关系。

交替关系是指表达相同主题概念的正式使用类目与非正式使用类目之间的关系。交替关系是用交替类目来显示的,交替类目是相对于正式类目而言的,它是指在类目体系中具有多种从属关系的非正式使用的类目。交替类目一般不直接用来类分文献,只起查找使用类目的作用。手工检索工具中,交替类目通常以参照的方式指向使用类目,满足用户从不同的角度检索的需要。在机检系统中,一般可以通过链接的方式直接将交替类目作为检索入口,查找对应的信息资源。

4) 相关关系。

相关关系是指类目之间除从属、并列、交替等方式以外的其他联系,是类目间一种横向联系的关系,指类目间虽不存在等级关系,但内容上相互关联。相关关系主要是通过采用类目参照的方式来表示。

上述四种类目关系类型中,从属关系、并列关系是揭示类目体系展开的主要线索,反映了类目的纵向联系;交替类目、类目参照则是对类目体系显示的主要关系的补充,揭示被类目体系所分散了的横向联系,通过两者的结合,文献分类法按照类目之间关系建立起了纵横交叉的、严密又实用的分类体系。

三、主题语言

20 世纪后半期,主题法得到迅猛发展,成为分类法以外的另一大主流方法,与分类法成为相对应的、互为补充的信息标引与检索语言。主题语言含有以下四个方面共同特征:①直接以语词作为检索标识;②以字顺作为主要检索途径;③以特定的事物、问题、现象,即主题为中心来检索信息资源;④往往是通过参照系统等方式揭示主题词之间的关系。

1. 标题词语言

所谓标题词,是指从自然语言中选取并经过词汇控制,用来表示事物概念的词、词组或短语。标题词语言就是经过规范化处理的名词术语来直接表达文献所论及的事物或主题,并将全部标题词按字顺排列起来而形成的一种检索语言。它是一种以标题词作为主题标识,以词表预先确定的组配方式标引和检索的主题语言。

标题词语言一般是将标题列举出来,供标引和检索选用。标题除了采用单词和词组形式的单词标题、词组标题外,还有倒置标题、带限定词的标题以及大量在标题词表中预先组配好的多级标题形式。

标题词语言具有主题语言的优点:以事物为中心集中信息,便于从事物出发的检索,特性检索功能较强,标识直观。此外,它还有标识含义明确,易于使用等优点。它的缺点是由于采用标题的列举方式和先组方式,概念表达能力受限,概念难以多向成族,检索途径少;由于采用自然语词,标识的通用性较差。

作为一种传统的主题语言类型,标题词语言开创了主题语言的最初形式,探索了标题法词汇控制的一系列方法,包括:制定了标题的选择和确定的准则,规定了标题的形式,标题之间关系的揭示,标题标引和检索过程中组配方法的使用等。

2. 单元词语言

单元词语言又称单元词法,是为克服标题法的不足而发展起来的一种主题法类型,它是单元词作为主题标识,通过字面组配的方式表达文献主题的主题法。所谓单元词,是指用来标引文献主题的、最基本的、字面上不能再分的语词。如“建筑”和“设计”就属于单元词,因为它们不能再进行分解;而“建筑设计”与“信息检索”可进一步分解成“建筑”“设

计”、“信息”“检索”，就不属于单元词。在使用单元词的情况下，对复合主题信息的标引和检索是通过单元词的组配进行的，如“儿童图书馆读者服务”这一主题，就必须通过“儿童”、“图书馆”、“读者”、“服务”四个单元词来进行组配表达。

与标题词语言相比，单元词语言具有词表体积小，标引专指度高，便于从不同主题词角度检索等优点。单元词法的不足是：直接性差；不适宜于查找论述基本单元主题的文献，例如，不适宜于对论述“经济”“图书馆”等基本主题的文献进行查找；采用字面组配方法，在字面分解与语义分解不一致时容易造成误差，例如，按字面组配，“雪崩二极管”应分解成“雪崩”和“二极管”，而“雪崩”一词则与“雪崩二极管”主题毫不相关；单元词之间较少存在词间关系，因此，单元词语言一般缺乏完善的参考系统，难以满足族性检索的需要。

在主题语言发展过程中，单元词语言率先探索了后组式检索方法，是后组式检索语言的前驱。单元词法使用的反记法，开目前机械检索系统中倒排档之先河，后来为叙词语言等主题语言类型所采用。此外，单元词法还广泛探索了后组式检索中的规律和问题，探索了联号和职号等辅助符号的使用方法及其对各种检索系统的适应性，为叙词语言的发展和使用开辟了道路。单元词法目前已被叙词法所取代。

3. 关键词语言

所谓关键词，就是从文献的题名、摘要和内容中抽取出来的，能够表达文献主题内容的具有检索意义的关键性词汇。关键词是一种未经过规范化的自然语言词汇，所以用关键词表达事物概念，具有直接、准确等特点。凡在概念上有意义并且能够代表文献信息的主要内容的词都可用作关键词。但是，为了确保检索用词与文献记录中的词汇完全一致，和控制计算机自动分词和抽词，有一些词被列为禁用词（stop-term），如一些冠词、介词、副词或连词等。

关键词语言是将文献中原有的、能描述其主题概念的具有检索意义的词汇抽出，并按字顺排列起来的检索语言。采用关键词语言编制的检索工具很多，如美国《化学题录》（CT）中的“题内关键词索引”、《化学文摘》（CA）中的“关键词索引”、《国际学位论文文摘》（Dissertation Abstracts International）及《应用力学评论》（Applied Mechanics Reviews）等。

利用关键词语言编制的关键词索引主要有三种类型：单纯关键词索引、题内关键词索引和题外关键词索引。

（1）单纯关键词索引（Keyword Index, KWI）。

单纯关键词索引是指把从文献的正文、摘要和题名中分析出的一组关键词，依据其字顺轮排，且在索引条目中不包含非关键词。

例如，有一篇题为《计算机在神经生物与行为学中的应用》的文献，分析题名可归纳出三个关键词：计算机、神经生物、行为学。在单纯关键词索引中，这三个关键词领头轮排，构成单纯关键词索引的三条款目（假定其文献代号为 03851）：

计算机	神经生物	行为学	03851
神经生物	行为学	计算机	03851
行为学	计算机	神经生物	03851

这三条款目依其字顺被排列在纯关键词索引中，检索者无论从哪一个关键词入手都会检索到文献代号为“03851”的文献。

单纯关键词索引的优点是索引的编制无需用词表,可广泛地用于计算机检索;出版速度快,满足快速传递文献信息的需要;关键词是未经规范化的自然语言词汇,表达概念直接,符合人们的习惯。其缺点是关键词没有经过语义规范,各种同义词、近义词及多义词都被作为关键词编排在索引中,造成索引的混乱,检索时也必须从各种可能的角度进行全面检索,否则会造成漏检;单纯关键词索引中各个关键词之间没有语义关系的显示,不易判断文献的主题内容,容易导致概念混淆,造成误检。所以说,单纯关键词索引的质量比较粗糙,一般只用于检索工具的期索引,没有累积索引。

(2) 题内关键词索引(Keyword in Context Index, KWIC)。

它又称上下文关键词索引。编制题内关键词索引的目的是:克服单纯关键词索引只有并列的关键词,而无词间语法关系,易造成误检的弊端。所以它和单纯关键词索引的本质区别是在每条款目中保留了非关键词(如冠词、介词、连词等),用以反映关键词之间的语义关系,使之成为一个具有完整意义的短语,从而准确地表达了文献的主题内容。

还以上例为例,常见的 KWIC 如下:

左栏	中栏	右栏
与行为学中的应用	/ 计算机在神经生物	03851
中的应用 / 计算机在	神经生物与行为学	03851
/ 计算机在神经生物与	行为学中的应用	03851

题内关键词索引就是把关键词包含在文献的题名之内,关键词的上下文和词序都不变动。在编制索引款目时,将题名全部排印出,题名中的每一个关键词依字顺轮流排在中栏首位,作为标引和检索的标识。索引中的“/”为起读符号,阅读一条索引时,先从“/”符号处往右读,然后再读“/”左侧部分,这样便可完整地读出该条索引。

题内关键词索引除具有单纯关键词的优点之外,还有一个显著的特点即“含义明确”。因为题内关键词索引除关键词之外,还保留了非关键词(如例中的“与”、“的”、“在”),用以规定关键词间的语法关系,使得每条索引款目的含义十分明确。这是它和纯关键词索引的本质区别。但是作为索引编排和检索主要依据的关键词,置于索引的中栏,排在上下文之间,直观性和易读性都较差。

(3) 题外关键词索引(Keyword out Context Index, KWOC)。

题外关键词索引是为了克服题内关键词索引的易读性差,但纯关键词索引没有关键词语法的显示,易造成误检而产生的一种关键词索引。它实际上是上述两种索引编制方法和原理的结合。

常见的 KWOC 著录格式如下:

计算机	《计算机在神经生物与行为学中的应用》	03851
神经生物	《计算机在神经生物与行为学中的应用》	03851
行为学	《计算机在神经生物与行为学中的应用》	03851

在这三个索引款目中,不仅著录了关键词,还著录了文献的题名,使用起来一清二楚,让检索者能更好地理解文献的内涵,但是 KWOC 篇幅庞大,编制起来费时费力。

关键词语言的优点是:标引时不需要主题分析和查看词表,简便易行,因而降低了对标引人员的要求,节省了大量人力;标引和索引编制易于实现自动化,从而大大缩短了检索系统信息组织和报道时差,保证了信息报道和传递的及时性;关键词是文献中使用的自然语词,

表达主题比较直观、专指度高,可以保证较高的查准率。

但关键词语言也存在着以下几个方面的不足:对自然语言中大量存在的等同关系词不加规范统一,也不显示等同关系,使相同主题文献常因作者用词不同而被分散,导致漏检;不显示关键词之间的等级关系和相关关系,难以进行族性检索,特性检索的查全率也不高;由于关键词多限于从文献中抽取,当这些关键词对文献内容的表达不充分或不准确时,会使关键词检索有一定的漏检和误检;在机编索引情况下,由于机械地抽词和轮排,其中有不少关键词款目不能反映文献主题而被抽取,不起检索作用却徒增篇幅。

4. 叙词语言

叙词语言是20世纪50年代随计算机在文献信息工作中的应用发展而逐渐产生的,也是在分类语言、标题词语言、关键词语言和单元词语言(已逐渐被淘汰)的基础上产生出来的一种新型的信息检索语言。

它吸收了以上各种检索语言的优点,克服了其不足之处:吸收了单元词法单词组配的基本原理;采用了标题词语言对名词术语进行严格规范的方法,以达到词与概念的一一对应;吸收了标题词语言的语义参照,充分揭示主题词之间的语义关系;采用了关键词索引的索引编排方法,用以编制各种索引;采用分类语言的基本原理编制叙词分类索引。

由此可见,叙词语言是多种信息检索语言的综合。取长补短,体现了检索语言的发展趋势。它既适合于手工检索,又适合于计算机检索。目前,世界上利用叙词语言编制的“叙词表”很多,如英国《科学文摘》的“INSPEC叙词表”和我国的《汉语主题词表》等。

叙词语言就是以叙词作为文献内容标识和检索依据的一种主题词语言。叙词(Descriptor)是从自然语言中选取的从概念上不可再分的,并经过规范化处理,使其变成词和概念一一对应、具有单义性特点的词汇。叙词的有序排列构成叙词表(Thesaurus),它是一种规划化的、动态性的词表。

叙词语言与单元词语言的不同主要在于采用了概念组配代替字面组配。即检索时,检索者需对一些词表中没有的复合词或短语,用词表中已有的叙词进行概念组配。所以说,概念组配是叙词组配的基本原理,是叙词语言的灵魂。

叙词语言集多种检索语言的功能于一身,从而成为一种具备优越性能的现代检索语言。其特点如下:

- (1) 结构完备,词汇控制严格,可以根据检索系统的需要对词汇进行有效控制。
- (2) 组配准确,标引能力强,能够准确、专指地标引和揭示各种主题内容。
- (3) 检索效率高,可以通过灵活组配方式进行多途径检索,达到更好的检索效果。
- (4) 检索系统适应能力强,可以同时适用于标识单元和文献单元检索方式,既能适应计算机检索系统的要求,又能适应手工检索的要求。

叙词语言的不足主要是:由于词汇控制严格,词表编制和管理的要求高,需要花费较多人力、物力;文献标引必须在概念分析的基础上进行,标引难度大,要求高。

四、检索语言的几点补充

1. 分类主题一体化语言

分类主题一体化语言是在一个检索系统中,对它的分类表部分和叙词表部分,就术语、参照、标识、索引等实行统一的控制,即一个分类系统与一个主题系统互相兼容,既能充分发挥各自独特的功能,满足分类或主题标引和检索的需要,又能融为一体,发挥最佳的整体

效应。简而言之,分类主题一体化语言是分类语言与主题语言的有机结合。《中国分类主题词表》即是在《中图法》类目与《汉语主题词表》主题词对应的基础上,将分类法与主题法、先组与后组融为一体的文献标引和检索工具。其中的《分类号—主题词对应表》和《主题词—分类号对应表》是分别从分类角度和主题角度进行分类主题一体化标引的工具。

2. 指南型分类体系

为满足检索飞速发展的网络信息资源的需要,雅虎(Yahoo)首创了指南型分类检索体系。它是一个分类字顺体系,属于等级体系型分类语言。这种分类形式已成为网络检索工具普遍使用的一种模式。以“雅虎”为代表的分类体系的使用目的是类分和检索网络信息资源,以适应网络资源的特点及其用户需求。与传统的文献分类体系有所不同:①不用分类号作为类目的排列标记而直接使用类名,按类名的字顺排列类目;②不采用以学科为中心的传统文献分类法的展开模式,而是建立以事物对象为中心的更直接、更符合网络用户查找习惯的分类体系;③打破传统分类法单线结构的展开方式,采用了多重划分、多元展开的类目结构,便于用户从不同角度出发都可以进行查找;④使用网络特有的信息组织技术——链接方式,对多重从属关系重复揭示,有效反映知识之间的联系,使一类下的类目的范围更加完整,相关类的揭示也更为充分,有利于网络用户对资源的选择查找。

3. 引文语言

引文语言是利用文献之间引用与被引用的关系作为文献内容主题标识,并以此标引和检索文献的语言。科学研究是在借鉴前人研究成果的基础上的一种创新活动,人们在创作科学文献的过程中往往会参考前人的成果信息来提供某种佐证、前例或背景材料。作者会在文献最后列出参考文献,这就是引文。该作者所创作的文献完成后也将汇入知识集合中,再后来的人又可将其作为引文加以利用和记录。文献间的这种引证关系,不仅反映了科学技术的交流和反馈,也启发人们通过追溯这种引证关系,找到一系列内容相关的文献。

假设有文献A和B,B文献在完成过程中引用了A文献,并把它列在参考文献中,此时就称A是B的“引文”(Citation),而B因为提供了包括文献A在内的若干引文,故被称为“来源文献”(Source Item 或 Source Document)。无论是哪种检索语言都要完成集中内容相关文献的功能,引文是这样完成集中文献的功能的:如文献A被文献B引用,又被文献C引用,也就是说文献A同时是文献B和文献C的引文,那么这个作为共用引文的文献A就通过文献之间的引用关系,集中了来源文献B和C。如果把文献A作为检索入口,就可以检索到在它之后发表的引用了它的所有来源文献。就好像是用某个主题词,就可以检索到含有这个主题词的所有文献一样。由于来源文献和引文之间在内容上是相关的,所以引文就完成了文献集中的功能。所谓引文检索语言就是根据上述原理,以引文标识标引或检索文献,而引文标识的全体就构成了引文检索语言。

使用引文作为检索语言完成标引和检索工作的技术就称为引文索引法。引文索引法的使用最早开始于美国的法律界。1873年,美国出版了一种供律师和法学家查阅法律案例的被称为“*Shepard's Citation*”的检索工具。该检索工具采用从一个已知案例查询出引用过该案例的所有案例的检索方法。20世纪50年代美国的尤金·加菲尔德(Eugene Garfield)从中受到启发,研制出用计算机辅助编制的引文索引。他主办的“科学情报所”(Institute for Scientific Information, ISI)先后出版了《科学引文索引》、《社会科学引文索引》和《艺术与人文科学引文索引》3种著名的引文索引刊物。

引文检索语言的独特性形成了引文索引的许多优势。从标引角度看,引文索引的标引工作很大一部分是由作者本人完成的,而且引文中所包含的内容均为自然语言写成,易于实现自动化标引和检索工具的自动化编制,大大节省了标引成本,且使检索系统更新速度快。因为作为标引员的作者是该学科的专业人员,这就克服了外行标引员在标引过程中容易产生主观片面性和不一致性。引文唯一代表一篇文献,不存在词义上的干扰,也不必因为学科的发展更新标引词,因此标引是稳定和专指的。相比于传统检索语言,描述文献更深入,涉及的范围也更广。从检索角度看,引文检索语言的检全率高,每一次的检索结果均可作为下一次检索的起点,如此循环往复,仅以一篇文献线索,就可以找到大量相关文献。检索结果的灵活性好,由于引文语言的伸缩性强,可跨学科检索,因此有利于用户根据检索需求随时调整检索范围,特别有利于边缘学科、交叉学科文献的检索。

引文检索语言也有其局限性。首先,标引者即作者选择引文的动机是多种多样的,造成引文与来源文献的关系有的高度相关,有的则相差甚远,容易导致误检。其次,引文索引文献收录的不完全性使相关文献难以查全,漏检率较大。

第二节 检索理论及计算机检索系统

一、信息检索方法

所谓检索方法,简言之就是利用检索工具依照一定的方法从不同的途径查找课题所需的文献信息。信息检索方法很多。目前,常用的检索方法大致可归纳为如下几种。

1. 常用法

常用法是利用检索工具查找文献信息的方法,是在文献检索中经常使用的方法,故称为常用法。在具体操作中,由分为顺查法、倒查法和抽查法。

(1) 顺查法。

顺查法是指利用选定的检索工具,以检索课题的起始年代为起点,按时间顺序由远及近(从过去到现在)地查找文献信息的方法。当需要了解某一课题的发展全过程时,就可用顺查法从最初的年代往近期查找。这种方法能系统、全面、可靠地检索到某一课题的文献信息,可获得较高的查全率,但工作量大,比较费时费力,适用于较大课题的信息检索。

(2) 倒查法。

此法与顺查法相反,指利用选定的检索工具,按照时间的逆序,由近及远(从现在到过去)的查找文献的方法。一方面,由于近期的文献信息新颖性和及时性好,可能已经包含和吸收了过去的文献信息;另一方面,由于一旦基本掌握了所需要的信息,或检索出的文献信息已经符合用户的要求,可随时终止检索。所以,倒查法检索效率高,获得的信息新颖及时,工作量小,省时省力,一般用于新开课题。新开课题比较注意近期的文献,以便掌握最近一段时间该课题所达到的水平及研究动态。只需查找若干年的文献,即可解决问题。

(3) 抽查法。

抽查法是根据研究课题的实际需要,利用检索工具检索某个特定时期文献信息的方法。使用该方法的前提是用户必须清楚有关该课题的文献信息最可能出现或最多出现的时间段。它适合于检索某一学科研究高潮很明显的、某一专业的发展阶段很清晰的、某一事物出现频率在某一阶段很突出的课题。此法能快速、有效地检索到某一课题的文献信息,检索效率高,