杨振山等
编著

DANGANGUANLI
QINGBAOJIANSUO
YU
JISUANJI

档案管理情报检索与计算机

同济大学出版社

档案管理、情报检索与计算机

杨振山　等编著

同济大学出版社

内 容 提 要

本书是一本介绍如何使用电子计算机进行档案管理、情报检索的书籍。本书由原理和应用两部分内容所组成。原理部分主要介绍汉语主题词表的管理方法、汉语主题词的计算机标引和检索方法等。应用部分结合在微型计算机上已实现的《主题词智能标引和检索集成化系统》(DINIRIS)以及它在档案管理方面的应用，介绍了档案的前处理工作以及 DINIRIS 的使用方法。

本书不但可以作为 DINIRIS 的使用手册，而且可以作为从事档案管理、图书情报检索和办公自动化的科技工作者以及大学有关专业的参考书和实验指导书。

本书备有软件盘片，同济大学出版社可向音像部购买。

责任编辑　王建中
封面设计　陈益平

档案管理、情报检索与计算机
杨振山　等编著
同济大学出版社出版
(上海四平路 1239 号)
新华书店上海发行所发行
上海市印刷四厂印刷
开本：787×1092　1/32　印张：3.875　字数：99 千字
1989 年 1 月第 1 版　1989 年 1 月第 1 次印刷
印数：1—4500　定价：2.10 元
ISBN 7-5608-0152-8/TP·8

前　言

现代科学技术的发展，对档案管理和情报检索工作提出了更高的要求，而档案管理和情报检索工作的现代化又促进了科学技术的进一步发展。使用电子计算机是实现档案管理和情报检索的主要手段。

在我国，用电子计算机进行档案管理，目前尚处于初级阶段。用计算机进行情报检索虽然起步较早，主要还是用于检索。至于标引主题词这项极其重要的前处理工作，还是靠人工来完成。用计算机标引主题词目前还处于研究阶段。同济大学与上海市档案馆共同研制的主题词智能标引和检索集成化系统——DINIRIS 这个软件于 1987 年 10 月通过鉴定，可以用来标引主题词。这是我国第一个公开发表的标引主题词的软件。可以预见，不久的将来会有更多更好的用计算机进行档案管理和情报检索（包括机标主题词）的软件出现。本书是结合 DINIRIS 的设计和实现，介绍如何用计算机进行档案管理和情报检索，以飨读者。

全书共分五章。第一章介绍主题词表的作用以及管理方法。第二、三章介绍主题词的标引和检索方法。第四章介绍档案文献的前处理工作。第五章介绍了 DINIRIS 的使用说明。

本书可以作为 DINIRIS 的使用手册。对从事档案管理和情报检索的工作人员是一本比较好的参考书。对于从事办公自动化以及其它计算机应用工作者来说也是很有参考价值的。

参加本书编写工作的人员有杨振山，路贵增和蒋宁。其中路贵增编写了第二章前半部分，蒋宁编写了第四章，杨振山编写了其它部分。在编写过程中还得到了同济大学计算中心 DINIRIS 课题组其它一些成员的大力帮助，他们是：孙山东、周广敏、谢燕菊、诸其良、朱建国、徐效敏、陈晓萍、郑红、陶碧云、王灵。在此谨表深切谢意。

作者　　一九八八年一月

目　录

第一章　主题词表的管理

§ 1　主题词表的作用和发展概况

主题目录是一种主要的读者目录，它是以揭示档案和图书资料的主题为目的，选用一些规范化了的词、词组或短语来表达主题。而按此顺序组织起来的目录就是主题目录，在档案管理、图书情报检索中有着广泛的应用，在按内容主题进行检索时也离不开主题目录。

主题目录主要涉及到两个问题，即主题目录的建立和主题目录的使用。前者主要是指主题词的标引问题，这将在本书第二章中介绍。后者主要是指主题词的检索问题，这将在本书第三章中介绍。这里必须指出，两者之间必须要密切联系，相互匹配。建立主题目录时要照顾到检索的需要，检索时则要考虑建立主题目录的规定，否则将会产生前后脱节的现象，严重地影响检索效果。用主题词表可以沟通主题目录的建立和使用之间的联系，即标引和检索时均使用同一主题词表，使得主题词表成为著录、标引和检索之间的一条纽带。

主题词表有各种叫法，比如“叙词表”、“检索词库”、“检索词典”、“关键词表”、“类属词表”、“关联词表”、“主题词典”、“主题表”等等。就主题词表的定义来说也是众说纷云。本文不准备详细介绍这些定义，也不准备为其重新加以定义，现仅引用这样一个定义为例来说明主题词表的作用：

“主题词表，从结构上看，就是按照情报检索查全和查准的要求，通过明确的概念（即主题）之间的相互关系的方法，组织和展示主题词，作为标引，存贮和检索文献的共同依据；从功能

上看，它不仅是文献处理人员和不同读者之间思维的桥梁，是自然语言（文献使用语言）和系统语言（检索系统规范化语言）之间的媒介，同时也是人和系统（即人机）之间联系的工具。”

不难看出，主题词表的质量好坏将直接影响到标引和检索的效率。衡量主题词表的质量通常是与其规范化程度和完备性相关联的。所谓规范化，是指主题词表中所选择的词是否符合作为一个主题词所应具备的条件，比如概念明确，词语精炼，具有单义性，以及研究和检索价值等等。所谓完备性是指使用的主题词是否满足专业范围内的标引需要，即主题词表的专业复盖面。因此，国内外的情报检索界、档案界都非常重视主题词表的建立工作，不惜花费大量的人力、物力、财力以建立一些高质量的，适合于各种应用的主题词表。

现在，已出版的主题词表种类很多。据不完全统计，仅在美国就有 300 多种。在我国也出版了不少主题词表，如《汉语主题词表》、《国防科技主题词典》、《电子技术汉语主题表》、《航空科技资料主题表》、《原子能科技资料主题词典》、《机械工程主题词表》、《农业主题词表》、《铁路汉语主题词表》、《常规武器专业主题词表》、《水利水电科学技术主题词典》等。这些主题词表，其设计目的、词表性能以及使用范围不尽相同，词表的结构也有差异。其中《汉语主题词表》最有代表性，它是我国第一部综合性的大型主题词表，集 191,158 条正式主题词和 1,740 条非正式主题词于一书，其规模之大，条数之多居世界主题词表之冠。它是由全国 505 个单位的 1,378 位专业工作者花了整整四年时间，花费了大量的艰巨劳动才完成的。国家档案局即将公布的《中国档案主题词表》，是我国第一个用于档案管理的主题词表，它集中了将近 24,000 条主题词，为我国档案界建立统一的档案资料检索数据库，实现联机检索网络和档案资料自动检

索以及档案信息进一步开发、利用和交流创造了必要的条件。

§ 2 主题词表的结构

一个主题词表，就其结构来说，它应是以规范化的、受控的和动态的主题词作为基本成份，以参照系统来反映词间的语义关系，通过字顺表及其它一些辅助索引工具的使用，用于标引、存贮和检索档案以及图书资料的词表。从主题词表的管理角度来看，上述内容有三点特别值得注意，即动态词表、参照系统和辅助索引。

2.1 主表和辅表

考虑到词表的完备性，那么词表的词汇量应是越多越好，但是词表过分大，就又显得臃肿庞大，不利于使用，因此有的主题词表分成主表和辅表。主表是主题词表的全体部分，它具有一套比较理想的结构形式和一批规范化的常用主题词。辅表是主题词表的附属和补充部分，以弥补主题词汇量的不足。主表和辅表的词汇量构成了整个主题词的总量。划分主表和辅表的原则不尽相同，有的是根据专业领域，使辅表中的主题词是一些个别学科的专业名词；有的是根据不同的应用，辅表中的主题词是一些备用待选的词，如此等等，不一而足。当然，可以将主表和辅表中的词在两表之间相互流动。

在 DINIRIS 系统中，主题词表也分为主表和辅表，主表中存放着一些经常使用的主题词（包括正式主题词和自由词），而将一些暂不使用或增加一些可能使用的主题词放在辅表中。这样做不但有利于词表管理，而且还可以节省宝贵的计算机存贮空间。

一个具体的应用检索系统，无论是档案管理或情报检索，其专业面都是有其局限性的。以文书档案管理为例，使用国家档案

局即将出版的《中国档案主题词表》当然是非常理想的，但是24,000条主题词，再加上一些辅助成分，对于微型计算机来说，其所占有的存贮空间是相当可观的。对于一个具体的档案管理部门来说，这24,000条主题词并非都能用到，有相当多的主题词可能根本不用，相反却还要增补一些能反映地方特色和个别应用领域的主题词或自由词。这就要求主题词表的主表和辅表中的词经常流动。辅表中的备用词一旦选用，则可上升到主表，而主表中的一些不甚理想的词亦可下放到辅表中去。这种流动情况并不是对等的，随着系统逐步推广使用，主表中的词汇量呈逐渐增加的趋势，辅表向主表流动的词汇量将日趋减少，即对于这种具体应用而言，主表的完备性程度越来越高，到一定程度，可以认为主表基本趋于饱和，可以停止辅表向主表的词汇流动，这时也就没有必要再存贮辅表。这时的主表应该认为是一个适合于这个具体应用的比较理想的主题词表了。

2.2 参照系统

主题词的含义是通过主题词的词间关系加以体现的。词间关系通常分为三种：同义关系，属分关系和相关关系。主题词表中的主题词，通过这些词间关系，形成一个复杂的有助于扩大检索途径的网络结构。

(1) 同义关系

同义关系又称等同关系，它是反映同义词或准同义词间的语义关系。由于地域差异，人们认识事物的角度和习惯不同，在表现同一事物时采用多种名称。比如称为自行车的，又可称脚踏车，单车等。这就为标引和检索工作带来很大的麻烦，处理不妥就会影响检索效率。解决的办法是在这些词义相同或近义的词中选择一个比较规范的作为正式主题词，而其它皆为这个正式主题词的

同义词。比如，选定自行车为正式主题词时，脚踏车、单车则为同义词。在词表中设置相应的两个项目："Y"和"D"。"Y"表示"用"关系，在本条目中列出的词为相应的正式主题词。"D"表示"代"关系，在本条目中引出的词为相应的自由词。相应的自由词（同义词）可能不止一个。不难看出，对于主题词表中的任何一个主题词，它的相应的"Y"和"D"不可能同时非空。一般地，一个主题词如果它的"Y"条目中已有词出现，那么这个主题词一定是自由词，所以它的"D"条目中自然不会有其它别的什么词出现。比如，自行车、脚踏车、单车这些词在主题词表中为下列形式：

自行车

Y

D 脚踏车、单车

脚踏车

Y 自行车

D

单车

Y 自行车

D

(2) 属分关系

属分关系是反映主题词之间的上下位语义关系，即主题词间的语义关系的等级性。在主题词表中设置两个条目来反映这种关系。反映上位概念的条目表示为"S"（即"属"）下位概念的条目表示为"F"（即"分"）。例如：

自行车

S 交通工具

F 载重自行车，轻便自行车，花式自行车

载重自行车

S 自行车

轻便自行车

S 自行车

花式自行车

S 自行车

和同义关系不同，对一个主题词，其“S”和“F”可能两个条目均不空，但“S”条目中如果不空时，只能有一个主题词，因为一个上位概念的主题词只能有一个。但“F”条目下如果不空时，就有可能出现多个主题词，这是因为一个主题词其下位概念往往不止一个。

一个主题词的上位概念仍可能有它的上位概念，一直追溯上去，直到再也找不到有上位概念的这个主题词称为族首词。以这个族首词为首，找出它的所有下位词，再进一步找出其下位词的所有下位词，一直到找不出下位词为止，所有找出的这些主题词组成一个主题词族。族首词在词表中通常是在其后面加上一个“*”标识的。以族首词开始，将该族中所有主题词按其上下位关系，依层次列出，可以形成一个树结构。

(3) 相关关系

相关关系是反映除同义关系、属分关系以外的主题词间的其它关系，如交叉关系、矛盾关系、对立统一关系、形式与内容关系、本质与现象关系，原因与结果关系等。在主题词表中设置一个“C”(即“参”)条目来反映这些相关关系。例如：

想象

C 表系　幻想　理想　虚构

在“C”条目中，如果非空时，通常也不只一个主题词，因为具有相关关系的词往往不只一个而是一组。

2.3 索引结构

主题词表的索引结构是主题词表的组成部分，它是标引和检索的辅助工具。一般主题词表中设置字顺表、范畴索引、轮排索引和多语种索引等。

§ 3 主题词表的机内表示

上节中介绍的词表结构适用于一般情况下使用主题词表的需要。在用计算机存贮主题词表时，由于计算机具有运行速度快的特点，其词表结构有所改变。现以DINIRIS中所使用的主题词表为例，介绍其结构情况。该词表共设有以下几个条目：

主题词号：这是用以唯一标识主题词的编号，一旦确定后就不再改变，直到该主题词被删除为止。

分类号：根据内容范畴的分类号。

主题词：汉语主题词。

汉语拼音：主题词的汉语拼音。

用：主题词的“用”关系，存放相应主题词的词号。

属：主题词的上位概念，存放相应主题词的上位主题词。

参：主题词的相关关系，存放该组相关主题词的组号，组号是任取该组主题词中的任一个词号。

频率：该主题词的使用频率。

使用这种结构的主题词表，可以很方便地找到同义的正式主题词、上位词以及相关关系，查找“代”关系时（即查找与正式主题词同义的自由词），可以通过查找“用”关系而得到，也就是说，所有具有相同“用”主题词的那些词为该主题词的同义自由词。查找下位词也是通过类似办法来实现的，即具有同一上位词的那些词就是该上位词的下位词。查找“参”关系主题词可以

通过查找相同组号（参）的主题词而得到。

为了方便地标引和检索主题词，词表设置了下列索引结构：

主题词字顺表：根据汉语主题词建立索引结构，用于检索主题词。

按分类号的范畴索引：根据分类号建立索引结构，便于范畴检索。

“用”关系索引：按“用”条目建立索引，便于检索“代”关系。

“属”关系索引：按“属”条目建立索引，便于检索“分”关系。

“参”关系索引：按“参”条目建立索引，便于检索相关主题词。

汉语拼音索引：按主题词的汉语拼音建立索引，便于按汉语拼音检索主题词。

§ 4 主题词表的管理系统

为了使主题词表更好地发挥作用，适应各种情况的变化需要，一个主题词表通常至少具有以下的管理维护能力：

4.1 检索功能

检索主题词表是为了了解主题词表组成情况的一个重要手段。可以从不同的角度进行检索，比如：

按汉语主题词检索；

按汉语拼音检索；

按组成部件检索；

按同义词检索；

按上下位词检索；

按词频检索；

按族首词检索；

按分类号检索，等等。

所谓组成部件是指组成主题词的一些成分，比如按组成部件“经济”检索，是指检索出所有包含“经济”这个成分的主题词。

4.2 编辑功能

这是维护词表的一项重要功能，它可以对词表中各个项目进行添加、插入、修改、删除等操作。比如：

添加主题词（包括相应的词号，汉语拼音以及参照关系）

插入主题词（包括相应的词号，汉语拼音以及参照关系）

修改主题词（包括相应的词号，汉语拼音以及参照关系）

删除主题词（包括相应的词号，汉语拼音以及参照关系）

主表和辅表间主题词的流动、转换。

4.3 统计主题词表

对主题词表进行统计是从不同角度进行数量分析的重要手段。执行统计功能可以完成下列各种统计任务：

按类统计；

按使用频率统计；

按族首词统计；

按同义词统计，等。

4.4 词表输出

可以将主题词表依不同的需要输出不同的形式，包括，

整表输出；

按同义词系列输出；

按族首词输出；

按组成部件输出，等。

§ 5 主题词表管理的一个例子

这里选用部分主题词表形成组织管理的一个例子，供参考。

词号	主 题 词	拼 音	Y(用)	S(族)	C(参)	AjGS	wjGS
110010	师范教育	shi fang jiao yu		110040	110020	88	4
110020	幼儿师范教育	you er shi fang jiao yu		110100	110010	12	1
110030	高等师范教育	gao deng shi fang jiao yu		110110	110170	17	9
110040	专业教育	zhuan ye jiao yu			110140	4	13
110050	中等专业教育	zhong deng zhuan ye jiao yu		110040	110090	6	17
110060	中等教育	zhong deng jiao yu			110070	7	19
110070	中学	zhong xue		110080		9	23
110080	学校	xue xiao			110110	11	24
110090	专业学校	zhuan ye xue xiao		110080	110130	14	26
110100	师范学校	shi fang xue xiao		110080	110010	18	29
110110	高等教育	gao deng jiao yu		110080		22	34
110120	技术教育	ji shu jiao yu		110040	110140	34	37
110140	进修学校	jing xiu xue xiao			110010	61	82
110160	民办中学	ming bang zhong xue		110080	110060	43	46
110170	干部学校	gang bu xue xiao		110080	110040	12	72
110180	高等学校	gao deug xue xiao			110010	37	18
110190	实验学校	shi yan xue xiao		110080	110040	88	2
110200	职业学校	zhe ye xue xiao	110090		110050	51	1
110210	技术学校	ji shu xue xiao		110090	110130	21	3

第二章　标引主题词法

§ 1　概述

标引主题词法是19世纪50年代随着计算机在图书资料、文献档案检索中的应用而发展起来的一种新的标引方法。其目的是向用户提供一种按主题词检索的工具。

情报检索一般包括信息存贮和信息查找两大部分。在信息存贮部分，所谓标引主题词，就是通过对图书资料、文献、档案等的标题、摘要、乃至全文主题内容的分析，参照主题词表，确定出若干个能表达主题内容的主题词或其组配，然后按照著录格式要求，把它们填写到相应的存贮条目中，并用索引的方法编辑成主题词目录，以供用户按主题词检索资料使用。在信息查找阶段，也需要对主题词进行标引，这是因为一般用户并不具有主题词检索的很多知识，不可能要求他们用主题词检索语言向检索系统提问。换句话说，应允许他们用自然语言向检索系统提问。因此，对他们的提问必须进行分析，选择出能表达用户提问要求的主题词，并进行组配，然后通过主题目录，查找到用户所需要的资料。

这种方法比分类法、标题法、单元法和关键词法有很多的优越性。主要表现在：

1.1 直观性好。由于主题词法以规范化的自然语言——主题词作为主题内容的标识符，使人们从字面上就能知道其表达的主题内容。不象分类法那样，以字母、数字或字母数字作为表达主题内容的标识符，使不了解分类法的人，只看到一些A，B，C或1，2，3，不知它们表达了些什么内容。