

主题标引与索引技术

傅 兰 生 编

科学技术文献出版社

1 9 8 7

内 容 简 介

本书是指采用上题词作为检索标识的标引与索引技术。主题标引与索引技术广泛地应用于各种检索系统中,如各种文献检索系统、档案检索系统、管理数据检索系统等,尤其是期刊与电子计算机检索系统。

本书可作为大、专院校图书情报专业的师生、各级情报、文献、档案、图书馆等单位的工作人员的培训教材。

主题标引与索引技术

傅兰生 编

科学技术文献出版社出版

中国科学技术情报研究所印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

*

787×1092毫米 32开本 3.25印张 70千字

1987年3月北京第一版第一次印刷

印数: 1—10400册

科技新书目: 139—73

统一书号: 17176·515 定价: 0.70元

目 录

第一章 概论	(1)
1. 标引与索引的基本概念	(1)
2. 标引的对象和步骤与文献处理流程	(13)
3. 标引与索引工作的总体方案设计	(15)
第二章 主题分析	(17)
1. 主题的概念与形式	(17)
2. 主题分析的重要性	(18)
3. 主题概念的语法与逻辑关系	(19)
4. 主题分析的步骤	(20)
5. 主题类型的模式化	(23)
第三章 标主题词与标引规则	(27)
1. 标引方案	(28)
2. 直接标引与专指性标引规则	(30)
3. 组配标引与组配标引规则	(31)
4. 上位标引与上位标引规则	(44)
5. 近义标引与近义标引规则	(44)
6. 增词标引与增词标引规则	(45)
7. 标引方案的辨证选择	(45)
8. 标引记录	(47)
9. 分组	(51)
10. 选导词	(53)
11. 标分类号	(60)
12. 标引结果的著录	(63)
13. 标引规则总结	(66)

第四章 标引深度与标引误差	(69)
1. 不足标引与过度标引	(69)
2. 标引不一致性误差	(72)
3. 上位标引与组配标引中的标引误差	(73)
第五章 索引技术	(81)
1. 主题索引是检索系统的重要组成部分	(81)
2. 主题索引的典型款目形式与检索功能特点	(83)
3. 主题索引的一些表现手法	(93)

第一章 概 论

1. 标引与索引的基本概念

1.1 主题检索体系的检索标识

在检索系统中，赋予二次文献检索特征，并用来作为存贮和查询依据的标识，称为检索标识。如图1所示，检索标识可以分为下述各种：

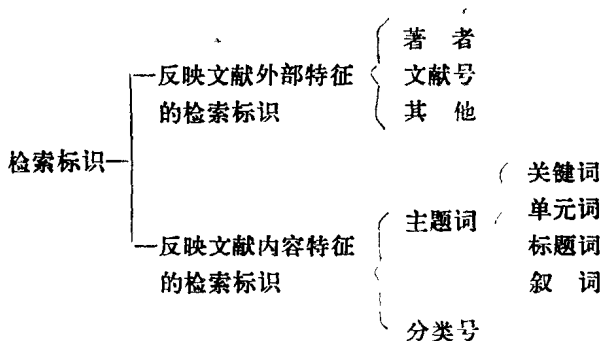


图1 检索标识的种类

根据图1所示，也可以给检索标识下这样的定义：描述文献外部特征和内容特征的专门用于情报检索的符号。其中，反映文献外部特征的检索标识比较简单，不难掌握；反映文献内容特征的检索标识则比较复杂，同时，大多数检索课题又都是从内容途径去进行检索的，因此，这种检索标识

成为人们研究的中心。习惯上，反映文献内容特征的检索标识称为检索语言，分类号称为分类法检索语言，主题词称为主题法检索语言，又称为主题检索体系的检索语言或检索标识。

分类号和主题词虽然都是反映文献内容特征的检索标识，但从反映文献内容的角度和表现方式来看却全然不同。主题词具有三个方面的基本特征：

- ①着眼于从事物的特性方面去揭示文献内容；
- ②采用文字符号；
- ③按字顺排列。

例如，“集装箱载重汽车”就是一个主题词。它从事物的特性方面去揭示文献内容，说明它不是别的，而仅仅只是集装箱载重汽车，采用文字符号“集装箱载重汽车”表示，如果是一个汉字检索系统，则按汉语拼音“Jizhuangxiang zaizhong qiche”的字顺位置排在检索体系中。

基于上述特点，可以给主题词和主题检索体系这样定义：按主题概念的特性关系用文字符号表示的检索标识，称为主题词；以主题词作为检索标识的字顺检索体系，称为主题检索体系。主题检索体系与分类检索体系不相同之点主要在于：前者重于特性检索功能，后者重于族性检索功能；前者采用文字符号，后者采用数码符号；前者按字顺排列，后者按等级排列。

主题词可以分为四种，它们都具有主题词的共同特点，同时又有各自的特点。

关键词：表示文献关键内容但未经规范处理的主题词。又称键词。

例1: 原子能

说明: 这就是一个关键词。它具备主题词的三大特点, 用文字符号表示, 从特性的角度反映事物(即只反映原子能本身而不及其余, 分类号则是从族性的角度反映事物, 揭示此事物在一个族系中的逻辑次序), 按字顺排列在检索体系中。同时, 它又仅仅只有反映文献关键内容的功能, 未经规范处理, 规范处理后的主题词可能不是“原子能”而是“核能”。

例2: 坡莫合金(规范处理后的主题词可能是“铁镍系导磁材料”)

例3: 汽车(规范处理后的主题词可能也是“汽车”)

关键词是从文献中抽取出来的, 常因作者的用词习惯不同而使表示同一主题概念的关键词呈现多种字面形式, 而主题检索体系又是按主题词的字顺位置排列的, 关键词的字面形式不同, 将使同一主题的文献分散到不同的字顺位置, 从而使查全率受到影响。

在文献篇名中抽取的关键词称为题内关键词, 除题名以外, 还从文献正文中抽取的关键词, 称为题外关键词。

单元词: 在词义上已是单元概念不能再予拆分, 并已经经过规范处理的主题词。又称元词。

例1: 汽车

说明: 这是一个单元词, 它在词义上已经不能再予拆分, 如果拆分为“汽”与“车”, 就完全失去了原来的词义。“载重汽车”就不是单元词, 它可以再拆分为更单元的概念“载重工具”与“汽车”。

例2: 电动机

例3：温度表

标题词：用经过规范处理的一组标识，并予以固定组配而形成的主题词。

例1：计算机科学-发展趋势

说明：这是一对标题词。处在引导位置上的主题词称为主标题词，处在附属位置上的主题词称为副标题词。本例中，“计算机科学”是主标题词，“发展趋势”是副标题词。标题词与关键词和单元词都明显不同，不是用一个词而是用一对词来表现主题概念，而且，这一对词的组配关系是固定的。

例2：齿轮-强度

说明：例中，“齿轮”是主标题词，“强度”是副标题词。标题词对的表现方式有多种，最常见的有两种。一种就是例中的形式，主标题词在前，副标题词在后，中间用“—”连接。另一种是主标题词在上，副标题词在下，写成：

齿轮

强度

叙词：经过规范处理的关键词。

例1：载重汽车

说明：这是一个叙词。它与关键词不同，因为它已经进行过规范处理，避免了关键词那种可能把同一主题概念分散到不同字顺位置，从而降低检索精度的弊病；它不像单元词那样，要受单元概念的限制；它不像标题词那样，可以灵活组配。因此，叙词最能反映文献检索的实际情况。在四种主题词中，叙词的应用最为广泛。

例2：玻璃纤维增强塑料

说明：这~~也是一个叙词~~。可见，叙词的表现能力是比较强的，这是它为什么应用广泛的重要原因。在我国，泛指叙词法为主题法，泛指叙词为主题词，在非特定指明的情况下，主题词即指叙词。

主题词除了有特性检索功能强、用文字符号表示、按字顺排列的三大特点以外，还具有组配灵活、直观易读、增删修改方便的重要特点，因此，主题词最经常运用于：

- ①以论文等深主题文献作为存贮对象的情报检索系统；
- ②书本式检索工具；
- ③机读式检索工具。

在各种主题词中，以叙词的应用最为广泛，标题词的应用也较多。

1.2 主题检索系统

旨在存贮和查询情报信息的系统，叫做情报检索系统。以主题词作为检索标识的情报检索系统，称为主题检索系统。在各种情报检索系统中，最常见的是文献检索系统。图2为一般的主题型文献检索系统的示意图。

从图2可以看出，文献检索系统应具备三要素：二次文献集合、检索词典及标引规则、检索手段。

所谓二次文献，是指以文献处理为目的，对一次文献（即原始文献如各种论文、科技报告等）进行著录、文摘编写等加工后形成的文献。二次文献是文献检索系统中的基本信息单元，一个检索系统当然应当拥有大量的二次文献。有序排列的二次文献集合及其检索标识，称为检索工具。记录在卡片上的检索工具，叫做卡片式检索工具，记录检索标识

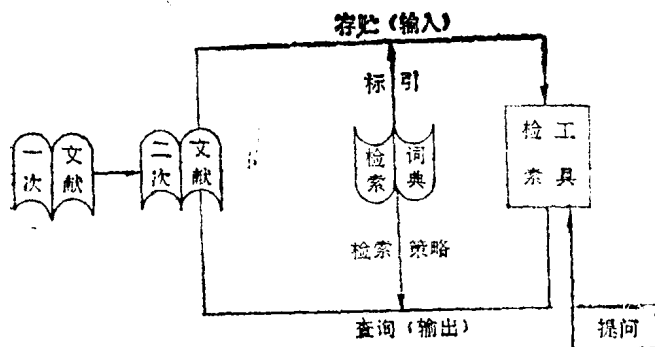


图2 文献检索系统示意图

的卡片称为导卡，记录二次文献的卡片称为文献卡。记录在书本上的检索工具称为书本式检索工具，在书本式检索工具中，二次文献与检索标识一般分开排列，二次文献部分按分类编排，检索标识部分按主题词、著者或其他要求编排。定期的书本式检索工具即检索期刊，不定期的以某专题为内容的书本式检索工具叫做专题文献索引。记录在计算机可读介质上的检索工具叫做机读式检索工具，机读式检索工具一般按记录顺序排列为顺排文档或按主题词等检索标识排列为倒排文档，也可按其他要求排列。

记录检索标识的词典称为检索词典，记录主题词的检索词典叫做主题词典或主题词表。检索词典或本身附有标引规则，或与标引规则配套使用，使检索系统中的二次文献有序化。

检索手段是指检索用的方式方法。手工检索中主要是指

各种导卡、导词的设置方法及检索款目的组成形式，机器检索中主要是指检索程序的功能和检索策略的安排等。

1.3 主题标引

把文献的主题内容和某些具有检索意义的特征、用检索标识记录下来的文献处理过程，称为标引。用主题词作为检索标识的标引，称为主题标引。所谓主题标引技术就是指研究主题标引的原则、特点、方法和技巧。

一个检索系统，不管是采用机器检索方式，还是采用手工检索方式，系统中存贮的每条信息或文献，都必须具备检索条件，即能给检索者提供一定的检索途径。主题标引就是赋予二次文献主题词的过程，因此，可以认为，构成一个检索系统有诸多环节，标引是其中最重要的环节。

标引有各种方式，采用何种标引方式应由检索系统的设计者根据系统的功能要求决定。

① 深标引与浅标引

详细描述主题的标引方式称为深标引方式，一般描述主题的标引方式称为浅标引方式。

例

标引概念：球墨铸铁曲轴的疲劳强度计算

深标引：曲轴，球墨铸铁，疲劳强度，计算

浅标引：曲轴，疲劳强度

说明：深标引与浅标引是相对而言的，例中的浅标引方式较之深标引方式在表现主题的深度上更浅，故称之为浅标引。

② 全面标引与针对标引

全面描述主题的标引方式称为全面标引方式，针对某种特定需要重点描述部分主题的标引方式称为针对标引方式。

例

标引概念：汽车与拖拉机的液压传动系统

全面标引：汽车，拖拉机，液压传动系统

针对标引：拖拉机，液压传动系统

说明：针对标引方式一般都是考虑到检索工具的专业性而选择的，不管一篇文献的主题有多宽，对于某种专业检索工具而言，都只有与它相关的主题才具有标引意义，其余部分则可以舍弃。例如，一种拖拉机专业的检索工具，尽管在它标引的文献中有汽车的主题，也可以不予标引，而采用针对标引方式。对于综合性的检索工具，则应采用全面标引方式。

③ 概括标引与分析标引

只根据题名的主题概念标引的方式称为概括标引方式，在分析全文内容基础上标引的方式称为分析标引方式。

例

题名：计算机在机械工程中的应用

概括标引：计算机应用，机械工程

分析标引：计算机辅助设计，计算机辅助制造，机械工程

说明：概括标引用得不多，但概括标引方式与分析标引方式联合使用却是经常的。

④ 受控标引与自由标引

采用规范主题词标引的方式称为受控标引方式，采用自由词标引的方式称为自由标引方式。前者又称为赋词标引，

后者又称为抽词标引。

例

标引概念：重型卡车的转向性能

受控标引：重型载重汽车，转向性

自由标引：重型卡车，转向性能

说明：自由标引就是采用关键词的标引。

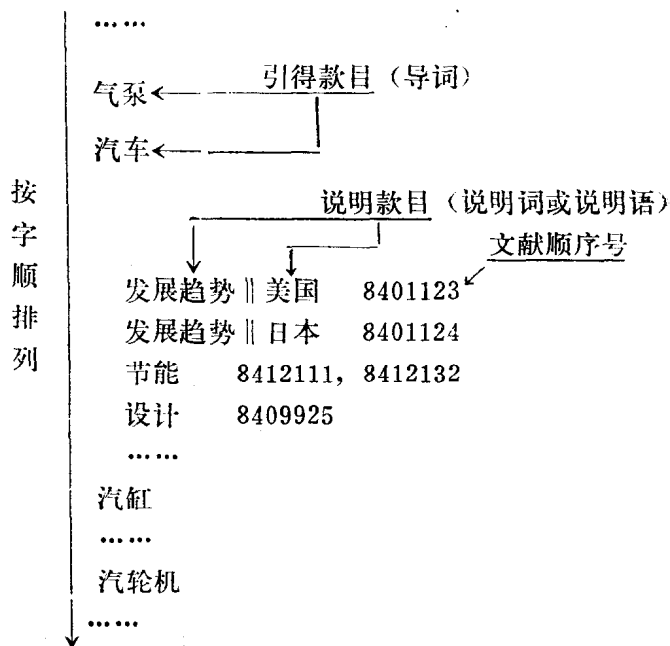
1.4 主题索引

检索工具中对照或引导二次文献的检索标识排列表，称为索引。用主题词作为检索标识并按字顺排列形成的索引，称为主题索引。在专著、论文集等图书或手册、企业名鉴等文献中用主题词作为检索标识，对照或引导文献中主要内容的索引，也是主题索引。

在主题索引中，又可视所用的标识分为关键词索引、单元词索引、标题词索引及叙词索引。

主题索引：一般由引得款目、说明款目和二次文献或其代号——文献顺序号组成。在索引中作为检索入口的款目，称为引得款目，组成引得款目的主题词称为导词。在索引中作为对引得款目起说明、修饰、限定等作用的款目，称为说明款目，组成说明款目的主题词称为说明词，如果说明款目不是由主题词组成的，则称为说明语。二次文献在检索工具中的排列顺序号，称为文献顺序号，在特定的情况下简称为文献号。

主题索引的一般型式如下：



主题索引可以起到下述几方面的作用。

①为从主题途径进行检索提供便利。检索工具的二次文献部分一般都是按分类编排的，编制主题索引即提供了从主题进行检索的途径。按分类编排、按主题编制索引是目前检索期刊最常见的形式。

②反映相关检索款目之间的关系。相关索引款目之间时常采用“见”、“参见”注释等方式表示其联系，可以起到扩大检索的作用。

③提供某一主题领域的发展趋势。某一科学技术领域的发展一般都与该领域所发表的文献数量成比例，对比历年索

引中某一主题词下记录的文献号，就可以看出该主题领域所发表文献的增减情况，从而判断其发展趋势。

④为某一领域的名词术语提供指南。

1.5 标引深度、引得深度与检索深度

用主题词反映文献主题的深度，称为标引深度。标引深度可以表示为：

$$I_{\text{标}} = S + W$$

式中： $I_{\text{标}}$ 标引深度

S 反映文献主题的专指度

W 反映文献主题的网罗度

专指度和网罗度都是用主题词的数量来表示的，因此，一篇文献的标引深度实际上就是指其标引的主题词的个数。

一篇文献在检索工具中的检索入口数，称为其引得深度，一个检索工具的引得深度就是该检索工具中全部文献引得深度的平均值。由于检索入口的多少就是导词的多少，因此，一个主题型检索工具的引得深度可以表示为：

$$I_{\text{引}} = \frac{L}{M}$$

式中： $I_{\text{引}}$ 引得深度

L 检索工具主题索引中导词之和

M 检索工具中存贮的二次文献的总篇数

检索中应当采用的导词数与实际采用的导词数之差，称为检索深度。检索深度可以表示为：

$$I_{\text{检}} = L_1 - L_2$$

式中: $I_{\text{检}}$ 检索深度

L_1 检索中应当采用的导词数

L_2 检索中实际采用的导词数

在实际中, 标引深度、引得深度、检索深度之间总是存在着下述关系:

$$I_{\text{标}} \geq I_{\text{引}} \geq I_{\text{检}}$$

例

标引概念: 在高速重载螺旋锥齿轮的设计和制造中采用电子计算机技术。

标引词: 高速齿轮, 重载齿轮, 螺旋锥齿轮, 计算机辅助设计, 计算机辅助制造

说明: 对这篇文献标引的五个主题词中, 前三个主题词反映了文献主题的专指度, 后两个主题词反映了文献主题的网罗度。假定选出的导词是“高速齿轮”、“重载齿轮”和“螺旋锥齿轮”, 而在实际检索中只采用了“螺旋锥齿轮”这一个导词。那么, 此例的 $I_{\text{标}} = 3 + 2 = 5$, $I_{\text{引}} = 3 \div 1 = 3$ (对于一篇文献而言, 其引得深度就是其导词数), $I_{\text{检}} = 1$ 。在用顺排档进行机检时, 是不考虑导词的, 即标引的全部主题词都可以作为导词, 此时 $I_{\text{标}} = I_{\text{引}}$ 。

1.6 标引频率与检索频率

某主题词在标引中使用的次数, 称为标引频率, 在检索中使用的次数, 称为检索频率。

标引频率是衡量主题词检索价值和提供标引参考的重要依据, 检索频率也可以起到同样的作用, 然而两者统计的出发点不同, 前者是通过标引考察, 后者是通过检索考察, 显

然，后者还具有考察标引正确性的意义。

2. 标引的对象和步骤与文献处理流程

一般而言，可以直接依据二次文献进行标引，因为，多数文献的题名及其文摘已经能够充分地反映它的主题。对于那些不能确切反映文献主题的条目，当然应当进一步参阅原始文献再进行标引。然而，不必要求所有的标引都根据原始文献，那样会使标引工作过于繁琐。

一般的文献处理工作流程见图3。涉及到标引的几个步骤是流程中最重要的部分。

从图3可以看出，标引与索引工作有以下步骤：

①主题分析，确定标引概念。每篇文献都有个主题，所谓主题分析就是从专业的角度分析出应当传达给用户与读者的主要情报信息，并从检索的角度分析出应当标引的概念。

②标主题词。分析出来的标引概念需要转换为主题词，在转换主题词时，一方面要查主题词表选词，同时又要遵守标引规则的规定。

③标引记录。标引中应当记录标引频率和主题词的增删修订情况。标引频率是指标引中某一主题词的次数，使用一次就是频率为1。主题词的增删修订内容广泛，无论是主题词的字面形式还是其参照项目，凡有增删修订都应记录。标引记录的目的是对主题词进行审核，累积定量性意见。

④分组。为了检索的需要，有时应将同一篇文献的若干主题词分为两组或若干组，分组的目的是为了消除主题词与主题词之间可能形成的错误逻辑关系。