计算机情报检索

（讲 义）

王能琴
景玉峰
编著

科 海 培 训 中 心
一九八七.三

前　言

本讲义主要结合一个实用的情报检索系统——中国计算机科学技术文献库系统”（以下简称 JKJ 文献库系统）介绍建立一个情报检索系统的全过程。该系统于 1984 年12月通过鉴定，先后荣获1986年度中科院科技成果进步奖，北京地区1986年优秀软件奖，并优选为1986年全国计算机应用成果展览会参展项目等。

鉴定以来，建库、服务等各项工作一直在积极、稳妥地进行。目前西文文献库已进库三万多篇文献题录，其中包括英、中、日、法、德、俄文种的四百多种计算机专业及其相关学科的期刊和会议。86年开始建中文计算机文献库。该系统为用户提供从分类号、文献题目、作者、出处、馆藏单位、年代、语种、文献类型、主题词、自由词等入口查找文献。

JKJ 文献库现在拥有一支稳定的标引队伍，保证每年万篇文献题录进库。文献库系统是我国自建的第一个投入实用的计算机专业文献库。它由《计算机科学技术英语主题词表》、《计算机科学技术汉语主题词表》、文献主题标引、文献库、检索软件、词库、词表自动管理系统、多语种词表管理系统等主要部分组成。

JKJ 文献库收录范围包括：

·数据通信有：通信系统、通信理论、远程通信、广播通信、通信控制系统、报文交换等；

·微程序设计有：动态微程序设计、模块微程序设计、微程序设计编译程序、微程序设计递归、微程序设计技术、微程序设计语言、微程序优化、微程序控制计算机等；

·逻辑设计有：计算机设计、控制逻辑设计、逻辑设计、逻辑设计语言、组合逻辑、顺序逻辑等；

·存储器有：存储工艺、存储控制器、存储器性能、设计、光存储器、随机存储器、只读存储器、相关存储器、虚拟存储器等；

·外围设备有：输出设备，输入设备伺服机构、磁带存储器、磁盘、磁头、光盘、软盘、打印机、计算机绘图系统等；

·计算机电路有：触发器、定时电路、逻辑电路、组合电路、寄存器、控制器、印制电路板等；

·中央处理部件有：运算器、运算控制器、监视器、控制台等；

·半导体有：半导体器件、半导体工艺、测试、封装等；

·组装技术有：电子封装、计算机组装、连接、冷却等；

·机房有：防火、空调等；

·电源系统有：电源保护系统、电源电路、电源散热、电源系统设计、稳压电源、稳流电源、非标准电源等；

·容错计算技术有：错误校验、容错计算系统、容错软件、冗余、自校验等；

·可靠性有：可靠性理论、可靠性分析、可靠性设计、可靠性模型、质量控制等；

·测试有：测试理论、测试模式、测试设备、测试语言、自动测试等；

·接口有：标准接口、程序接口、接口部件设计、输入输出接口等；

·网络系统有：报文分组交换、计算机网络、网络软件、协议等；

·实时系统有：实时系统应用、实时模拟、实时数据采集、实时材料系统等；

·计算机系统有：单机系统、多机系统、数据流计算机、阵列计算机、流水线计算机、向量计算机等；

·计算机体系结构有：固件、体系结构设计、系统结构、系统优化、系统分析、系统设计等；

·指令系统有：计算机字、指令格式、指令控制、指令表等；

·计算机安全有：保密控制、操作系统安全、数据库安全、通信管理、硬件安全等；

·兼容性有：软件兼容性、硬件兼容性、兼容性集、设备兼容性等；

·计算机性能有：性能分析、成本性能比、性能优化、系统性能等；

·计算机类型有：并行计算机、串行计算机、SIMD计算机、MIMD计算机、功能分布式计算机、流水线计算机、向量计算机、第四代计算机、第五代计算机、定点计算机、浮点计算机、并行操作计算机、同步计算机、异步计算机、巨型计算机、大型计算机、中型计算机、小型计算机、微型计算机、通用计算机、专用计算机等；

·处理机有：并行处理机、并行数据驱动处理机、单处理机、多处理机、处理机组织、分布式处理机、流水线处理机、微处理机、相联处理机、位片处理机、信号处理机等；

·操作系统有：操作系统理论，操作系统体系结构，磁盘操作系统、UNIX 操作系统、分时操作系统，实习操作系统等；

·设备管理有：设备分配、设备状态字、资源分配、资源管理程序、资源控制等；

·文件系统有：文件逻辑结构、文件物理结构、文件目录、文件维护、文件控制等；

·排队有：排队分析、排队规则、队列存取法、排队问题等；

·调度有：程序状态字、调度程序、调度算法、动态调度、进程、调度、资源调度、作业调度等；

·作业管理有：前台、后台、输出流、输入流、作业说明、作业装入等；

·存储管理有：存储分配、存储管理策略和算法、分段存储管理、覆盖、可重定位分区存储管理等；

·分时系统有：分时操作系统、分时软件、分时系统评价、分时语言等；

·多道程序设计有：并行控制系统、重入代码、优先权、装入控制程序、多维数组存取、分块、重入程序、再定位等；

·实用程序有：存储打印程序、存储转储程序、更新程序、求助程序、后处理程序、内务处理程序等；

·宏有：宏程序库、宏调用、宏汇编程序、宏系统、宏处理程序等；

·软件工程有：并发软件、软件复杂性、软件管理、软件开发、软件灵活性、软件模型、软件市场、软件维护、软件支援系统、自度量软件等；

·程序方法有：程序证明、检验、设计策略、设计技术等；

·软件包有：标准软件、程序包、例行程序、目标程序库、同余数生成程序、外存储程序等；

·程序设计有：程序设计技术、程序设计逻辑、分布式程序设计、函数程序设计、逻辑

程序设计、自动程序设计、智能程序设计环境等；

·程序设计语言有：ADA 语言、C 语言、COBOL 语言、PL/1语言、PROLOG 语言、BASIC语言、ALGOL 语言、FORTH 语言、FORTRAN 语言、RPG 语言、LISP 语言、SNOBOL 语言、数据流语言、判定表语言、可扩充语言、FORMULA 67语言等；

·处理程序有：编辑程序、翻译程序、汇编程序、解释程序、生成程序等；

·程序执行有：程序运行、程序步、程序间通信、调用、迭代、动态流程图、活动程序、装入程序、异常结束等；

·元语言有：埃威逊记号、巴科斯范式、元—元语言、元语言变量等；

·交互系统有：交互程序、人—机系统、交互终端等；

·中断有：中断处理、中断系统、中断源等；

·结构程序设计有：结构程序设计环境、结构语句、由底向上程序设计、逐步求精法、自顶向下设计等；

·仿真有：仿真程序，仿真技术，仿真计算机等；

·联机系统有：联机信息系统、联机设备等；

·数据结构有：数据抽象、数据集、线性数据结构、分类法等；

·数据转换有：模数转换、数模转换等；

·记录有：记录格式、记录管理、记录结构、逻辑记录、物理记录等；

·自动机有：多自动机、概率自动机、模糊自动机、无限自动机、有限自动机、自动机理论等；

·形式语言有：变换文法、迭代定理、LR(K)文法和语言、上下文无关语言、上下文有关语言、形式语义学、形式语言理论、线性方法等；

·开关理论有：开关函数、模糊逻辑、组合分解理论、组合逻辑等；

·算法有：半数值算法、并行算法、近似算法、马尔可夫算法、模糊算法、判定算法、算法复杂性、组合算法等；

·计算数学有：代数、概率、函数、数理逻辑、数理统计、数论、数学分析、数学模型、拓扑学等；

·数值分析有：常微分方程、插值、迭代法、非线性方程解、函数逼近、函数求法、函数分析、有限元分析、最小二乘方逼近等；

·图论有：割集、矩阵表示、路径、平面图、图枚举、图论应用、图论算法、向量空间、有向图等；

·判定有：决策系统、判定函数、判定方法、判定分析、判定规则、判定理论等；

·布尔代数有：布尔表达式、布尔差分、布尔方程、布尔函数、布尔运算、真值表、双稳态变量等；

·动筹学有：博奕论、排队论、数学规划、优化、预报理论等；

·情报检索系统有：情报检索方法、信息科学服务等；

·数据库有：分布式数据库、集中式数据库、关系数据库、数值数据库、事实数据库、数据库逻辑组织、数据库物理组织等；

·信息处理有：数据简化、表处理、查表、符号处理、数据处理等；

光数据处理有：光处理机、光数据处理理论、光纤维、光信号处理、光信息存储器、

滤光技术等；

·汉字信息系统有：汉字处理软件系统、汉字输入系统、汉字输出系统、汉字信息处理等；

·机器翻译有：翻译机、机器翻译算法、机器翻译系统、语义学等；

·评价有：评价标准、数据评价、数据库评价、性能评价、情报检索系统评价等；

·人工智能有：定理证明、机器人、计算机视觉、人工智能应用、问题求解、学习、知识工程、自然语言处理、自组织系统等；

·计算机图形学有：剪裁、彩色显示、光栅图形、计算几何学、三维计算机图形学、事件处理、图形系统、显示装置等；

·图象处理有：图象编码、图象恢复、图象增加、连续图象表征、离散式二维线性处理等；

·模式识别有：边缘识别、笔划、分段、群聚、形状识别、语音分析、语音识别、字符识别、字识别、特征选择等；

·模拟有：混合模拟、数字模拟、控制模拟、模拟应用、模拟语言等；

·CAD 有：计算机辅助工程、辅助制造、辅助概念、辅助开发、辅助设计服务、辅助设计语言、辅助解题系统等。

·自动控制有：闭环控制、城市交通控制、程序控制、生产控制、实时控制等；

·计算机应用在工程上有：材料工程、电机工程、电子工程、工程设计、航天工程、核工程、环境模拟工程、化学工程、交通运输工程、军事工程、矿产工程、声学工程、摄影技术、通信工程、土木工程、冶金工程、制药工程等；

·计算机在工业方面、自然科学方面、社会科学方面、行为科学方面的应用。

该库收集内容广泛，比较适用于从事计算机及其相关学科的科研、生产、教育、应用、管理等方面各层次工作人员要求。

该系统的主要特点有：

1. 总体设计合理，功能完善，系统配套，实用性强；

2. 本系统数据源均立足于本库协作单位的馆藏，因此凡从本库检索出来的文献题录均能容易地找到原始文献，方便用户复制；

3. 本系统检索功能较强，易学易用；

4. 本系统服务途径多，多用户联机和检索、定题服务、自动编排出版检索刊物和专题文献题录等；

5. 本系统易于推广和移植。它已在 PDP—11系列的34机、23机、24机和73机上运行，部分软件和数据已移植到 VAX 系列机的750机和 PC 机上使用，效果良好。

随着计算机科学技术的迅速发展和应用范围越来越广泛、图书、情报等部门业务管理自动化工作随之发展，从事这个方面的工作人员也日益增多。他们是科技战线、生产战线、经营管理的前头兵。他们迫切需要进行知识更新，以适应四化建设的需要。我们编写这本讲义，奉献给我们的同行，且愿它能为你们参考使用。全书内容按建立文献库的工作顺序，从编主题词表等基础工作开始，一直讲到文献库提供利用，进行服务等方面内容。目的想使读者通过学习，对建立一个实用的、计算机化的文献库系统有一个完整的概念，并对图书、情报、资料、档案等自动化工作有启发引导作用。

由于我们水平有限，编写时间仓促，难免有考虑不周的地方，欢迎大家批评指定。

编者 87.3.1

目　录

第三篇　文献主题标引

第四篇　介绍PDP－11小型机的一般使用

第五篇　文献检索系统

第六篇 论文选登

第一篇　编制主题词表

王能琴

第一章　概　述

主题词表是根据情报检索需要而创造的一种人工语言，又称情报语言、检索语言、标引语言或索引语言等。它用于表达文献（或提问）主题，贯穿于情报的存储和检索的全过程，是沟通标引（或分类）人员和检索人员双方思路的桥梁。

怎样编制和选用合适的主题词表是保证提高情报检索效率的重要条件之一。一本理想的主题词表必须满足如下要求：

① 应具有必要的词汇—语法手段，能准确地表达某一科学技术领域方向的任何文献（或提问）主题。

② 每个词（或类名）必须概念明确，具有单义性（唯一性）。

③ 主题词表应便于将文献检索标识(主题词或分类号)同提问检索式中的项目相比较。

④ 主题词表体系应是科学、合理的，便于使用。

任何书面语言，包括主题词表，一般由符号体系、词汇和语法组成。符号体系是指表示某种语言词汇所使用的代码系统。这种代码是由字母、阿拉伯数字或字母数字混合组成的一串符号（如主题词表中的词号、范畴号、分类表中的分类号等）。词汇是文献检索中使用的主题标识（如主题词、标题词或类名等）。这种词汇是规范化的。以表的形式按规定的体系或顺序汇集在一起（如分类表、主题词表、标题表等）。语法是指如何创造和运用上述标识来正确表达文献（或提问）主题，以实现有效检索的一套方法和规则(如组配标引规则等)。

目前国内外最常用的情报检索语言分：分类语言和主题语言。分类语言是用分类号来表达文献主题概念的语言，以分类表作为文献分类的工具，主题语言是用主题词或标引词来表达文献主题概念的语言。本篇下面主要讲主题词表的基本概念；主题词为什么要规范化；规范化主题的选择；主题词表的组织和展示；主题词表的若干规则；主题词表的参照关系；主题词表的增删改；主题词表数据的计算机处理等；国内外几种主要词表；另附《ISO2788—1974文献工作——编制单语种叙词表规则》。本讲义特点主要结合编制《计算机科学技术英语主题词表》和《计算机科学技术汉语主题词表》加以论述。通过这部分学习，使大家有一个基本概念且能动手编制专业词表。较深一些的内容可以在干中学习，理解。

第二章　主题词表基本概念

主题词表，目前在国内外有好几种叫法，如叙词表、主题词典、索引词典、关键词词典、检索词典等。我国情报人员习惯用主题词表。有关术语定义：

1. 款目词：所有特指词都是款目词，所有主题词都是款目词（是指其自身）。但不是所有的款目词都是特指词或主题词。

2. 主题词：系指以概念为基础的规范化的词语或词组，专门用于标引和检索文献资料的情报检索语言词汇。

3. 主题词表是由能够全面反映某学科领域并由语义相关，概念等级相关的词汇所组成的一种规范化动态性词表。词表按照等级体系和字顺来收录所有主题词及其间的关系符号。词表规范化就是规定一个词只能代表一个概念，也就是说，要求在同义词中选出一个主题词来唯一的代表某一概念；同时，对于多义词必须明确其使用意义。词表的动态性是指它经常不断地更新，或者是增加新的主题词以表示某学科的新生概念；或者根据标引或检索中积累的资料改造成现有的词义及其结构。

主题词表就其性质分有专业性主题词表和综合性主题词表。我们这次主要偏重于专业性主题词表。

4. 字顺表（即主表）

字顺表是主题词表的主体，所以又称主表。主表中的全部主题词款目是按款目主题词的字顺排列的，主题词款目是字顺表的基本单位，分为简单款目和复杂款目两种，一个款目凡是由一个主题词组成的，叫简单款目；凡是由具体参照关系的几个主题词组成的，叫复杂款目。

主题词分为正式主题词和非正式主题词两类。正式主题词用于标引和检索文献；非正式主题词作为正式主题词的同义词，起参考和指引作用。主题词款目包括款目主题词及其有关资料。款目主题词的有关资料是指限义词、主题词范畴、注释和各种关系词等。

5. 限义词即说明语句。它是正式主题词的一部分，借以识别多义词，如：上层建筑（船舶）、上层建筑（政治经济学）。

6. 主题词范畴是指款目主题词的概念分类。

7. 注释是为指明用法而给款目词加以注释，写在款目主题词的下面。

8. 关系词是指在语义上与款目词有密切关系的其它主题词。

9. 参照关系指明与款目主题词构成的各种关系。大多数词表采用三种关系：同义关系、等级关系和相关关系；五种符号：USE（称为“用”）、UF（即Used For的缩写，称为“代”）、BT（即Broader term的缩写，称为“广义词”，简称“属”）、NT（即Narrower term的缩写，称为“狭义词”，简称“分”）和RT（即Related term的缩写，称为“相关词”，简称“参”）。

10. 同义关系这是把从情报检索的角度称为同义的非正式主题词指向正式主题词，保证等同概念所用的主题词唯一性。非正式主题词指向正式主题词，采用“用”（USE）符号。正式主题词指明它所代替的非正式主题词，采用“代”（UF）符号。“用”和“代”互为反参照。

如：自行车
　　代　脚踏车
脚踏车
　　用　自行车

11. 等级关系是指同一族内各主题词之间的不同概念等级的专指性。下位主题词指向上

位主题词，采用“属”（BT）符号。上位主题词指明下位主题词，采用“分”（NT)符号。“属”（BT）和“分”（NT）互为反参照。

如：磁盘

分（NT）固定盘

固定盘

属（BT）磁盘

12．相关关系是指等级关系以外的密切相关的关系。具有这种关系的主题词之间可以互相参照，采用“参”（RT）符号。

如：电子计算机

参（RT）自动化

自动化

参（RT）电子计算机

上述“用、代、属、分、参”符号在汉语主题词表中可以分别用汉语拼音的第一个字母，即Y、D、S、F、C表示。

13．范畴表（或叫范畴索引，或叫分类索引）

范畴表是对全部主题词概念范畴的划分，主要作用是：①满足从学科、专业的分类角度组织和查找主题词，作为字顺表（主表）的辅助工具；②作为资料分发用的类目。为此主题词概念的划分应考虑各学科主题词分类的数量的分布和馆藏文献数量的分布。

范畴表多数采用二级分类型，类号采用四位数字、字母—数字或字母三种编号。类目分为大类和小类。在小类之下，主题词按字顺排列。大部分主题词是一词分入一类，少数主题词是一词分入两类以上。

1200　计算机和数学

1201　计算机

存取时间

计算机辅助设计

⋮

中央处理机

数字字母混合码形式

1202　计算机程序

汇编程序

程序设计语言

⋮

1203　计算机应用

34　化学

34A　基础化学

34AA　一般概念

⋮

34B　化学分析

⋮

14．词族表（也有叫词族索引、等级表等）

词族表是词间等级关系的族系展示，按词族首词字顺排列。在同族内主题词按其等级关系排列；在同级内，主题词按字顺排列。词族表的作用：①检索时供选词参考；②编词表时借以校对主题词间的等级关系。词族表是把主表中具有属分关系、整部关系和包含关系的正式主题词，按其本质属性展开全显示和一种词族系统。族首词是一族之首（最上位概念）

如：

automata	自动机
.automata theory	.自动机理论
.finite automata	.有限自动机
..asynchronous finite automata	..异步有限自动机
..generalized finite automata	..广义有限自动机
.	.
.	.
.	.
.fuzzy automata	·模糊自动机
.infinite automata	·无限自动机
.	
.	
.	

15.语言对照表（即语言对照索引）

语言对照表是指一种语言的主题词与另一种语言的主题词之间建立对照关系，以便按需要将一种语言的主题词转换为另一种语言的主题词。其主题词是按第一种语言词的音顺或字顺排列。

如：白光	white	light pen	光笔
光笔	light pen	white	白光
.	.	.	.
.	.	.	.
.	.	.	.

第三章　主题词为什么要规范化

情报检索的一切复杂处理过程总要牵涉到类目—文献的这种或那种分类处理。我们按主题内容标引文献时就是把它归入一个或几个类里。为了便于文献分类工作并能熟练处理这些类目，每一类必须有一个类名或类标。给这些类目起的类名（我们也可以把它们叫作“类标”）通常叫做主题词，而这些主题词的全体则叫做主题词表。为了满足一个特定课题所需的某种情报而在一种情报检索体系中进行查找时，我们所要做的是：①判定哪些类目里最有可能有与所需情报有关的文献；②查找这些类目；③检出特定文献或所有文献。

一个情报检索系统的效率主要取决于该系统的类目规模和类目结构，同时也取决于准备查找的类目（即准备采用的查找方案）。一般来说，如果我们准备查找的都是大类，就能查全某一个特定课题的全部文献（即做到查全率高），但都难于检出特定的文献（即查准率低）。反之，如果我们准备查找的是许多小类，就可能做到较高的查准率，但我们却会发现难于进行广泛查找而做到较高的查全率。

对于以后检索具有重要意义的是，在标引时，文献归类要遵循前后一致的原则。分类表

要能起作用，它就必须把有关的文献归到一起。这就是说，我们必须使标引工作规范化。大多数检索系统都要求标引人员按事先的规定对文献进行归类，而不是给每个标引人员以全权而随意给文献设立新的类目。因为那样做会产生许多交叉类目，而使有关文件分散。这种事先的规定就是给标引人员编出一个标引时必须使用的主题词表。标引人员判定手中文献与什么问题有关，大体上对于哪些检索提问能提供有用的情报，就从正式词表里选出主题词给文献标上。

规范化的主题词通常叫做规范化词汇或权威词汇。词表中的主题词是规范化的，具有权威性。

词汇规范化的趋向是改善标引的前后一致性。两个标引人员（或同一标引人员在不同的时候）在表述一个特定课题的时候，比较可能选用较为互相一致的一个词或几个词，如果这些词是从事先编妥的词表中选出来的，而不是在标引时独自编的话。另外，当我们用一个检索系统进行查找的时候，如果我们能从一个确定了的类目里查找，就较为可能查到正确的类目（也就是包含了我们关心的类目）。

规范化词汇在一个情报系统里的重要性，在标引过程和查找过程中都很突出。但规范化词汇不能，至少也不应该影响对文献进行的概念分析和检索提问进行的概念分析。概念分析阶段同语言转换阶段是两个不同的阶段。我们首先要制定一篇文献或一个检索提问是关于干什么的，然后把我们的概念分析转换成主题词表中的正式主题词。概念分析和语言转换两个阶段对检索系统起着不同的作用。例如，我们能确切地判定一篇特定文献是涉及“氩弧焊”的。这就是我们的概念分析。当我们把它转换成词表中的主题词的时候，我们可能要用含义更广泛的词(较不精确的词)来表达。检索系统的词汇不足以使我们精确定出涉及“氩弧焊”文献的类目。于是我们必须求助于含义较为广泛的词，例如“保护弧焊”，或“弧焊”，或“焊接”那样的词。

主题词表在以下两点上影响着检索效能；由于它确定了查找者能以多大的精确程度表述检索提问者感兴趣的事，它影响着查找策略，由于它确定标引者能以多大的精确程度表述文献内容，它影响着标引作业。主题词表在检索工作中起着很重要的作用。对检索系统全部任务的完成有着重要影响。

主题词表的存在主要是为了使标引者的语言与查找者的语言趋向一致。正常情况下，它对同义词和近同义词给予控制，以防止不同标引者对同一内容用不同的词来表述。词汇里指出优选的同义词，这就防止了同类文献的分散，并告诉查找者要查什么词、不要查什么词。正常情况下，它也把同形异义词区分开，如“Plant”这个词就有“植物”的和“工厂”的以及其他各种可能的含义。

主题词表也有使我们易于进行族性检索的功用。它需以某种形式把相关的归到一起，以便对较广泛的主题进行查找。假定我们需要查找“甾族化合物”（Steroids）这个主题，词表就应该能够把以某种方法联结到一起的全部有关词展示出来。这就节约了查找者的精力。不然的话，他就是想出词汇里可能有的所有“甾族化合物”的词。这样，减少了他对相关词漏查的可能性，保证他实际上的查寻是全面的。为了有助于标引者和查找者按上述那样把词的关系展示出来，它会超出通常的属一种关系而扩大到其他类型的关系，包括部分一整体关系以及物质或工具同它的可能应用的关系。

总之，主题词规范化确保文献标引的一致性，从而保证查准、查全、以提高文献检索的效率和性能。

第四章　规范化主题词的选择

在编制一部规范化词表时，我们必须：（1）确定课题所包括的确切范围；（2）选择描述这些范围的适当的词；（3）当这些词选入词表时要确定其准确的形式；（4）采用某种有效方法组织这些词；（5）采用某种有效方法展示这些词。首先，考虑一下待选词的汇总方法。

以往所用的至少有四种选词方法：（1）标引一批有代表性的文献，以此为依据，凭经验选择词汇。（2）转换现有的词汇，如将一部标题表转换成主题词表。（3）从现有的较综合的主题词表中抽词，或者在较综合的主题词表的范围内产生出专业主题词表，即产生一部微型词表。（4）从各种来源，包括术语汇编和其他出版物，以及学科专家那里，把词收集在一起。

在某些条件下，可以使用第二和第三种方法，但其他两种方法一般更为适用，是一般情况下所遵循的方法。这两种方法基本上可称为经验法和汇编法。对于前者，最初的词汇是通过文献的自由标引得到的。由此得到待选词，然后加以评定、归并、编制成有用的体系。这是编制杜邦公司叙词表使用的方法，也是穆尔斯建立他的Zator系统所使用的方法。伍斯特，H.称这种方法为“石笋式”方法。实际上我们是从原始待选词着手，并把我们组织起来。

对于汇编法，一般从术语汇编和其它出版物来源抽待选词，尤其是从学科专家委员会那里经讨论来抽取待选词。由此产生的词汇，再由编辑人员精选。只有当词汇处在完成或半完成形式时，才可用来标引文献。这基本上是工程师联合会编制的《工程术语叙词表》所使用的方法。伍斯特以下述的话来描述这种方法：

编制词表有两种方法：钟乳石式和石笋式。石笋式是从文献的山洞底慢慢向顶部发展。钟乳石式是召集一批专家，凑在一起选词。《计算机科学技术英语主题词表》是采用树形结构方法，这实际上就是石笋式方法。

石笋式词表可以通过人工或计算机处理文章中的实际词汇进行编制。而钟乳石式词表只由专家委员会编制。如果一部词表有着一刀切那样的规则性，所有词都等同地扩充，这种词表是由注重本行业价值的学科专家编制出来的；而如果一部词表充满了迷乱的不规则性，如有些词几乎忽略，其它词则扩展得很专指，则词汇可能是反映了作者的迷乱的不规则性的机器所编出的。

虽然学科专家们可以作为顾问起有效的作用，特别是在词汇的结构方面。但是，只依靠汇编法，而不通过有代表性的文献的标引进行经验上的验证，那是不适当的。汇编法的缺点是由于：

1．学科专家可能不完全熟悉文献，而更重要是不熟悉用户的要求和系统可能的使用者；

2．学科专家可能提出对检索目的并无用处的过细的繁琐的区分，这将使标引和查找任务复杂化；

3．学科专家可能为抬高自己专业的重要性，而使词表作为一个整体失去平衡。

规范词汇应是一种实用的工具，对它的基本要求是能够描述文献中和向系统提问中出现

的概念。显然，这就需要以文献和提问的特征为依据，并需要在同样的基础上不断地改进。往往被忽视的一个重要原则就是文献依据（也称为目录依据），这个原则由休姆，在1911年就提出了的。按照休尔米（Hulme）的说法，由文献组成类目不应当以任何理论上的“知识分类”为基础，而取决于文献本身在逻辑上可能形成的组。这就是说，类目取决于已有的文献。换句话说，文献本身的特征，将决定在系统中所确定的类目。那么，系统的词汇应以本系统所涉及的各个主题中所实有的文献的类目的准确调查及其范围为基础。用休尔米的话来说，这就是“预先划定现有文献的范围”。

显然，根据经验由文献标引产生的词汇有充分的文献依据。反之，用汇编法编制的词汇文献依据很少。美国国会图书馆的分类表就是一个有充分文献依据的规范化词汇的例子，因为它是以图书馆书架上出现的图书进行有效地分类编出的。

在讨论《教育资源情报中心叙词表》时，古德曼，F.就是一个文献依据论的得力鼓吹者，他说：

涉及到使用词表中尚未出现的任何词的考虑，必须是从标引具体文献的人员开始。不能仅仅因为一个词汇编辑人员认为这个词有用，就加一个；不能仅仅因为教育术语小组的成员认为这个词有用，就加上一个；除非有一篇文献属于词表的范畴，否则谁也不可以给词表任意加入那怕是一个词。

他还说：值得强调的是应让标引人员来补充他认为可能重要的词，而不是把他限制在那些与实际文献标引中关系不大的想象的词汇范围内。

美国全国标准化学会（ANST）在其最近的一份关于词表结构标准方案中，把经验法称之为分析法，而汇编法描述为格式塔法。一般来说，应优先选用前者，特别是对于专业知识领域更是如此。但是，该标准的编制者指出，格式塔法一般更适用于包括好几个学科的泛指的主题范畴。并且在某些应用中希望结合采用这两种方法。

《计算机科学技术英语主题词表》和《计算机科学技术汉语主题词表》的词汇来源主要采用经验法（分析法）。选词的最主要依据是我们几年建库标引中遇到的词汇，然后请专家帮忙补充修改，也参考了有关词表和辞典等。

第五章　主题词表的组织和展示

族性分析

对通过各种途径收集的主题词进行族性分析，首先要定出几个大的基本类别，进行粗分组，定好族首词，再找出主题词间的等同关系，确定好正式和非正式主题词，然后根据正式主题词的概念大小，列出树形结构的等级关系。将族首词下的主题词等级关系按族首词字顺排列成词族表，采用阶梯点数表示法对同一族内的检索词按字母顺序排。至此一个词族表初稿编成了。

主题词表的字顺表组织和展示

1．一般主题词表的展示，以NABA词表为例（图1），从中可以指示出典型词表的主要特点：

①在按字顺展开的每条主题下，列出属（BT）项、分（NT）项和参（BT）项；

②每条主题词款目下的三组词的每一组一般都按字顺排列；

③非正式主题词也排在字顺表中，同时标上“用”（USE）的那个相应的正式主题词。

④“用”项的反参照项“代”（UF）项要标在相应的主题词下面；

以《计算机科学技术英语主题词表》为例（图2-1和图2-2）。

```
electron energy
  0604 2304 2311 2402 2403
  UF  electron temperature
      electronic levels
  BT  particle energy
  NT  eletron states
  RT  activation energy
      electron density (concentration)
      electron mass
      electron precipitation
      electron pressure
      energy
        .
        .
        .
electron flux
  USE  electrons
       flux (rate)
```

图 1

```
8896  paged memory management
6461   USE  paged storage management
6461  paged storage management
8896   UF  paged memory management
6435   TT  storage management
6462   NT  demand paging
6463       demand paging system
4839  pager loop
4687   TT  printout
4818   BT  photo detector
           .
           .
           .
```

图2-1

```
2253  emulation
2254  .emulation computer
2255  .emulation technique
2256  ..emulation system orga-
        nization
2258  .emulator system
```

图2-2

除上述举的两个词表例子是常用的方法。此外还有图解展示，如果关系不复杂，当然展开以后，一目了然。这一点字顺展示法是做不到的，但是当等级层次关系复杂时，图解法就不容易表达清楚。

第六章　主题词表的若干规则

主题词表的编者必须对主题词的形式和标点符号作出明确规定。1962年的沃尔，E.主题词表、1966年的 LEX 编辑计划、1967年的 COSATI 主题词表、1966 年教育局的教育资源情报中心主题词表（ERIO）及其1969年的再版，已积累和记载了主题词表结构方面的主要经验。对这些和其它主题词表所介绍的规则和惯例加以研究将是有益的。

主 题 词 词 形

主题词无论在何处出现，都尽可能采用名词形式；比如，采用“粗糙度(roughncss)”，而不采用“粗糙的（rough）”。形容词或类形容词用于概念要求比较精确而又没有恰当名词的场合；比如，“空中的（airborne）”，“机动的（mobile）”，“轻便的(Portable)”。从来不用动词做主题词；“分析”宁可用其名词“analysis”，而不用其动词“analyse”。“倒、灌、注”宁可用动名词“Pouring”，而不用动词“Pour”。ERIC 主题词表用一部分形容词代替名词；比如，用“社会地位低下的(disadvantaged)”代替“社会地位低下的人(disadvantaged people)”，用“盲人的(blind)”代替“盲人(blinb people)”。《美国全国标准学会主题词结构标准草案》建议，名词短语中尽可能不用介词；比如，“糖类代谢作用”用“(carbohydrate metabolism)”，而不用“melabolism of carbohydrates”。

主 题 词 的 数

LEX 编辑计划作了以下规定：

1.“可数(count)名词”用复数形式（就是说，其对应的提问是“多少个(how many?”），比如，“电子管（tubes）”，“滤光器(filters)”和“城市(cities)”。

2.“集合(mass)名词”用单数形式(就是说，其对应的提问是多少(how much)?”)，比如，“辐射(radiation)”和木材(wood)”。

3. 具体过程、性质或状态，采用单数形式。

表1列出了 LEX 关于主题词的数的整套规则。

我们在两部词表的编制过程中，凡是一般字典里取单数的也取单数，究竟取单数还是复数，可以根据编制者所用的检查系统定。