中国图书馆学报(双月刊)1994年第6期
ZHONGGUO TUSHUGUAN XUEBAO

●陆长旭

后控词表的编制方法

ABSTRACT On the basis of differentiating various vocabulary systems the paper approaches the production and definition of the post-controlled vocabulary. The methods of compiling are methods of similarity matching, cluster anaiyses and artificial intelligence, etc. 14refs.

SUBJECT TERMS Subject thesauri--Compilation Post-controlled vocabularis--Patterns and types

CLASS NUMBER G254.2

自1975年起,我国图书情报界为实现文献处理现代化,开始编制《汉语主题词表》。它为以后的主题法研究与应用奠定了基础。到目前为止已编制出各种专业主题词表80多部。

作为检索语言之一的主题检索语言和由主题检索语言所形成的各种词表或词表系统的主要作用是为了更好、更科学地管理、控制与利用文献。随着计算机技术和检索方法的不断发展,对检索语言也不断提出新的要求,使主题法向更高、更准、更快的方向发展。"今后10年到15年,主题法可能发展成为叙词表和自然语言相结合的方式"[1]。最近有专家认为,叙词法可能向关键词法方向发展,80年代中期起,关键词法大有东山再起之势[2]。还有专家认为:"目前的主题词表(叙词表)应转向后控词表,使其由前台走向后台"(即由先控转为后控)[3]。后控方法研究和后控词表的出现正是符合了叙词表和自然语言相结合方式的发展趋势。

一、词表系统类型的区分

主题词表(叙词表)由手工编制走向计算机辅助编制,建立各种类型的词表管理系统(或称词表系统),为主题法的普及和应用铺平了道路。近年来建立的词表系统从方法上可以归纳为3种类型:

1. 传统叙词表的电子化

叙词表在实现电子化过程中,采用的方式和功能不同,又可分为3种。

(1)手工叙词表转换为机读形式

这是将手工编制的叙词表存放在计算机内或者出版叙词表的电子版。例如我国80年代出版的《汉语主题词表》(自然科学部分)经过90年代的修订后,出版了它的机读版,可用软件转入用户的系统中使用。《汉语主题词表》(自然科学部分)增订本在修改过程中增补了8221个新词,删除了5434个不适用词,纠正了27种4～5万条逻辑错误,并改正了

10多条词间关系。它成为中文自动标引系统后控词表的母表，可提供词源及词间语义关系，并对多个自动标引系统起协调作用[4]。

(2)利用计算机辅助编制叙词表

近年来，一些单位要建立自己的专业文献数据库，要求有本专业适用的叙词表，而人工编表费时费力，故采用自编软件或利用某些数据库软件的词典功能辅助编制专业叙词表。这种方式编制的词表往往是与他们建立的文献检索系统结合使用的。例如国家体委情报所利用CDS/ISIS的词典功能编制了《体育汉语主题词表》[5]。又如中国科学院计算机研究所建立的JKJ/TMS和ML-TMS系统，生成了《计算机科学技术汉语叙词表》，同时在建成的基于词表的计算机科学技术文献检索系统中，利用词表反映的词间关系和词义知识改进了检索系统的检索结果[6]。

(3)不断扩展功能的词表系统

在叙词表管理系统的基本框架下，充分发挥计算机的管理功能，建立起多文种多功能的词表管理系统。例如中国科学院电子学研究所为配合"中国无线电电子学数据库"建立了多文种多功能词表管理系统，并生成了专业叙词表。系统的特点是可处理多种文字，可管理多个词表，词条输入可有多种方式，不需要记忆各种功能命令，系统采用多功能模块的集成化结构等。在系统的6个模块中，可实现27种功能，对多个词表具有选择功能，并可不断扩充其功能[7]。

2. 赋词标引的词表系统

赋词标引是指在处理文献时用的词语(标引词)来自文献体外，使用一部有完整词形的参照词典。这种方法能够实现词转换，可正确区分并标引出相关的标引词(主题词或自由词)，并且标引词规范化的程度高。例如江西省国防科技情报所研制的《全国军转民数据库》自动标引系统中，分别采用词中单字"环"码形式对文献进行字处理取词，实现赋词标引；并用字处理控制切分抽词法，给出叙词表中没有的关键词，两者交替运算，实行互补，取词灵活，可提取新词，且切分词典容量小[8]。使用这种系统还可以不断自动更新其主题词表，并为补充与完善主题词间参照关系作好准备。从某种意义上说这种系统孕育着后控词表系统的思想与方法。

3. 后控词表系统

20多年前美国国家医学图书馆的MEDLARS用户就采用了后控词表的基本成份，并将检索策略中难构造部分存贮起来备用。它是一个具有"逻辑或"关系的词语一览表，覆盖了《MeSH标题表》中词语的全部口语形式，也就是同义词表，并与《MeSH标题表》的树型结构联系起来，对MEDLARS系统检索实行部分后控。而BIOSIS的关键词表则是一个较完整的后控词表，它是由标引者增加一些相容性词语编制而成，为检索者提供有关概念，用这种词表系统检索比单纯自然语言检索更有效。我国在80年代后期建立的中国化学文献检索系统，使用《汉语主题词表》进行标引与检索，同时积累与建立了"汉英/英汉化学文献标引检索用语词库"，通过词间关系文件反映后控规范的主题词表语义关系，利用主题词的单汉字索引文件及地址文件和词序关系文件作为匹配处理工具，对主题词进行控制，并不断完善主题词和自由词之间的语义网。这样的后控表系统不仅可以提高标引与检索的效率，也为以后进行自动标引奠定了很好的基础。

二、后控词表的产生和定义

由于计算机技术的逐步引进，对主题词表的编制与管理技术和方法也不断提出新的要求。60年代中期国外利用计算机辅助编制词表并进入实用阶段。近10年来推出的词表管理软件不断进入市场，它们有：INDEX、BASIS、MeSH、TMS、DOMESTIC、PROTERM-T等。据报道，国内一些单位也先后开发了各种类型的词表管理系统，如北京大学、中国农业科学院文献情报中心、中国

科学院文献情报中心等，但多限于自己单位使用，尚未进入市场，也有些单位利用国外的软件进行汉语主题词表编制、更新工作，如中国科技情报所、国家体委情报所等。

到目前，人们对主题检索语言中词语的控制方法，无论是手工编制的主题词表（叙词表），还是进入计算机的词表系统，它们都是"先控"（或"前控"）的。先控的方法是在标引（输入）阶段进行的，因而在检索（输出）系统中有新概念（词语）更新慢、控制的滞后性、标引失控和误差的产生以及用户使用困难等不足之处。国外从20多年前就开始采用同义词典法、存贮构造难度大的检索式法、存贮完整检索式法和自动构词法等方法来弥补先控之不足，奠定了后控词表的雏型。人们研究与借鉴了自然语言的特点，采用规范化语言与自然语言相结合的后控方法，产生了赋词标引的词表系统和后控词表系统。下面探讨一下后控词表的定义。

美国情报专家兰开斯特，F. W. 早已指出，"尽管自然语言检索在联机检索中变得越来越通用……但现在对后控词表的概念却研究得很少"。他认为："后控词表只是把相同含义的词编辑在一起，但不只是有常用的叙词表结构。即使它没有什么有效的结构，它也能在检索系统中起到非常大的辅助作用，并把含义相同的词替代了。其控制施加在输出阶段而不是输入阶段。"所以又称为它是一种简易型叙词表。"一部真正的后控词表是由众多的可供联机网络内自然语言数据库用户查找的名称及识别号码所组成"[9]。

后控词表是编制很严谨的主题词表，包括词的各种关系，甚至词的语法属性、关系之间能相互参照，对用户提问词进行各种控制，包括同义词扩充、相关词扩充，上下位按等级扩充，还可带有智能型（即联想功能、自学习功能、自我完善功能）的一种词表系统[10]。

后控词表只用于检索不用于标引，也称为只供检索的词表，其编制方法是利用检索表达式中用词由计算机自动积累而成的，不断增长（词语）是所有后控词表的特点[11]。

采用自然语言标引的后控规范（词表），是在计算机中保持叙词法的骨架体系，通过后控规范来不断扩充这一词表体系的语义网，使词表体系能充分反映科学技术发展过程中不断出现的新术语[12]。这种词表体系形成的语义网词库就是后控词库。

通过专家们对后控词表定义的阐述、研究与实践，人们对后控词表的概念正逐步明确，现可以归纳出后控词表应具有的特点：

1. 必须有以存贮于计算机内的词表体系为基础；

2. 词表系统在标引（输入）时可使用自然语言；在检索（输出）时使用机内的只供检索的词表；

3. 有通过提问词不断自动积累新词语的词表体系，并可及时更新；

4. 词表体系中的词间关系组成的语义网也是可以不断更新的。

三、后控词表的实现方法

由于后控词表具有上述特点，它的建立是有一定条件的。张琪玉先生提出了后控词表的5种编制方式，笔者认为这些编制方式均是可行的，但如何选择则要结合各自系统的特点来决定。下面探讨几种具体的实现方法。

1. 相似性匹配法

它是利用计算机半自动处理方法，找出概念上有某种联系的词语。在大多数情况下，词形相似的主题词间存在着概念上的某种联系。例如："空气污染"与"土壤污染"两词语在字面上不一致，但有部分重合，在一般倒排档检索中两词语是不匹配的。此法根据相似性匹配算法设计了一种不完全一致匹配方法，找出词形上相似的词语部分，并按相似程序排序，简化了查找词间关系的过程。这种方法在自建的中国化学文献库中进行了试验和使用[13]。通过后控规范处理不断扩大、补充词

表系统的语义网，对提高检索效率大为有益，同时还可以改善标引工作效率，提高文献数据库的质量。

2. 聚类控制法

聚类分析是将事物按其某些属性的相近程度归至各个群体。因为同一类群的事物一般都具有相同的特性，而不同类群之间都有着不同的差别[14]。采用聚类分析方法对用关键词或自由词标引的检索系统中的词表系统的词建立词间关系，形成语义网。有了这样的语义网可以提高系统的检索效率，达到语义控制的目的。“一般来说，系统聚类法比动态聚类法更易反映数据间的关联程度，且不受其他因素的影响，客观性较好”[15]。绝大多数的聚类方法首先计算类之间的相似性或相异性或距离的测度的矩阵。然后进行并类成为一个新类，再计算新类与当前各类间的近似度，反复循环直至类的个数等于 1 为止[16]。可以预见，系统聚类法能更好地实现词间关系的聚类控制。目前这种方法正在试验中。

3. 人工智能法

在传统的信息检索系统中，对用户的提问系统只向用户提供一串抽象概念的词语。这些离开具体语言环境的词语将给用户不同理解留下充分的余地，使对事物概念的准确控制造成困难。人工智能法是利用语义网理论和约束性传递激活法，使用户与系统对话中让激活的一组词显示到满意为止，同时用户也可要求显示词间关系等，供加深理解用。这样的系统用显式的非线性超文本组织，将事物（对象）及其关系显式表达，为用户提供了与标引人员大致相同的文本环境，无疑奠定了标引与检索一致性的基础，减少了检索的不确定性，提高了系统的效率。例如在国外近年来建立的 GRANT 系统中采用距离约束、扇出约束和推理模式实现系统的传递激活。而 METACAT 系统则是建立在用联机词表构成的语义网知识表达基础上，系统使用 LCSH 词表，利用启发式传递激活达到精炼需求的目的。在另一个 IR 智能检索专家系统中则利用文献间的引证和被引证关系，建立了文献的语义网络结构，通过超文本技术，用户可以从文献作者、索引词和相跟随的链，实现对文献数据库的检索与扫描[17]。这些系统都可达到对信息检索系统实现智能型后控的目的。

除以上 3 种方法外，还有词频统计法和神经元网络法等。

参考文献

1 陆长旭. 主题法的发展. 计算机与图书馆，1980，(1)：38～43

2 顾耀芳. 主题语言的应用、分析及预测. 现代图书情报技术，1993，(3)：2～6

3,10 陈道豪. 检索系统与后控制词表. 国防科工委情报所第四届学术年会论文. 1992

4 庄玉君.《汉语主题词表》的新进展及其对中文自动标引的影响. 现代图书情报技术，1994，(3)：21～24

5 张忠友等. 利用 CDS/ISIS 词典功能编制更新汉语主题词表. 现代图书情报技术，1992，(3)：11～14

6 景玉峰等. 基于词表的情报检索系统 JKJ/TMS. 现代图书情报技术，1991，(1)：19～24

7 周文表. 多文种多功能词表管理系统. 现代图书情报技术，1992，(2)：12～13

8 欧金森等. 汉语主题赋词标引. 现代图书情报技术，1993，(2)：8～11

9 Lancaster F. W. Vocabulary Control for Information Retrieval, 2nd ed. Virgiria, Information Resources Pr, 1986

11 张琪玉. 论后控制词表. 图书情报工作，1994，(1)：1～4

12,13 王源等. 后控规范的计算处理. 现代图书情报技术，1993，(2)：4～7

14,15 董建成. 聚类分析在情报研究中的应用. 情报理论与实践，1992，(3)：29～31

16 张永奎. 聚类分析在自然语言处理中的应用. 情报学报，1993，(5)：352～358

17 贾同兴. 智能情报检索中的超文本技术. 现代图书情报技术，1994，(1)：42～46

陆长旭 路学。现为中科院文献情报中心副研究馆员。已发文 10 余篇。通讯地址：北京中关村科学院南路 8 号，邮码 100080。

（来稿时间：1994-04-19。编发者：刘喜申。）