中国图书馆学报(双月刊)1996 年第 1 期
ZHONGGUO TUSHUGUANXUEBAO

●张琪玉

自然语言与人工语言对应转换

——情报检索语言走向自动化之路

ABSTRACT After analyzing 19 document indexing models of computer retrieval system, the author points out that in the information retrieval language, the key link for moving toward automation is to add the corresponding transformation of natural language and artificial language. Classification schemes and thesauri should be developed toward this direction.

SUBJECT TERMS Information retrieval language - Trends Artificial language - Transformation - Natural language

CLASS NUMBER G254.0

1 文献检索当前使用的两类语言工具

人类的巨大知识积累需要控制，否则不能有效利用。控制的主要方法是建立检索系统。检索系统的基本原理，一是对文献进行登录，二是对文献进行标引(揭示其主题内容)，三是对文献中的知识进行组织。

目前对文献内容进行揭示和组织的语言工具分为人工语言和自然语言。自然语言是在计算机检索系统出现后才流行起来的。

人工语言即情报检索语言，是根据情报检索的需要创制的，包括分类检索语言(分类号)、主题检索语言(检索词，在本文中称控制词)、代码检索语言(代码)。代码检索语言适用范围窄，故使用不多。

自然语言检索用词一般取自文献本身(题名、摘要、各级小标题、全文，但对于文献标引而言，以题名最为可取)，个别由标引人员自主赋予(即自由标引，一种非依据词表的主题标引方法)。

在表达文献主题概念方面，自然语言词专指性最好，控制词次之，分类号更次之。

自然语言词、控制词、分类号之间可以互相转换。这种转换，可以是等义转换，也可以是广义与狭义的转换，或近义和意义密切相关的转换。

一般而言，分类号可控制控制词(一对一或一对多)，控制词可控制自然语言词(一对一或一对多)。

2 计算机检索系统的多种文献标引模式

在计算机检索系统条件下，对文献的标引模式可能有 19 种(见表)。

现对其做简略说明。

模式 1 是对文献用人工进行主题分析，将析出的主题依据某种分类号—控制词对应表(如《中国分类主题词表》)用人工转换成控制词和分类号。

表 计算机检索系统中的文献标引模式

抽取主题概念	所用语言工具	转换方式	检索时可使用的检索标识
1 人工主题分析	分类号—控制词对应表	人工转换	控制词,分类号
2 自动抽词	有自动赋词功能的分类号—控制词对应表	自动赋词	关键词,控制词,分类号
3 人工主题分析	词表	人工赋词	控制词
4 人机结合抽词	有自动赋词功能的词表	自动赋词	关键词,控制词
5 自动抽词	有自动赋词功能的词表	自动赋词	关键词,控制词
6 自动抽词	有自动赋词功能的词表	自动赋词	控制词
7 人工主题分析	分类表	人工赋号	分类号
8 人机结合抽词	有自动赋号功能的分类表	自动赋号	关键词, 分类号
9 自动抽词	有自动赋号功能的分类表	自动赋号	关键词, 分类号
10 自动抽词	有自动赋号功能的分类表	自动赋号	分类号
11 人工主题分析并自由标引	编成后控制词表		自由标引词,控制词,分类号
12 人工抽词	编成后控制词表		关键词,控制词,分类号
13 人机结合抽词	编成后控制词表		关键词,控制词,分类号
14 自动抽词	编成后控制词表		关键词,控制词,分类号
15 人工主题分析并自由标引			自由标引词
16 人工抽词			关键词
17 人机结合抽词			关键词
18 自动抽词			关键词
19 单汉字系统			无确定检索标识

模式2是从文献中自动抽出关键词,保留关键词作检索标识,并依据某种具有自动赋词功能的分类号—控制词对应表给出相应控制词。因为控制词与分类号是对应的,所以也可用分类号进行检索。

模式3是对文献用人工进行主题分析,将析出的主题依据某种词表用人工转换成相应控制词作检索标识。

模式4是从文献中用人机结合方式抽出关键词,保留关键词作检索标识,并依据某种具有自动赋词功能的词表给出相应控制词。

模式5是从文献中自动抽出关键词,保留关键词作检索标识,并依据某种具有自动赋词功能的词表给出相应控制词。

模式6是从文献中自动抽出关键词,将关键词依据某种具有自动赋词功能的词表转换成控制词作检索标识。

模式7是对文献用人工进行主题分析,将析出的主题依据某种分类表用人工转换成相应分类号作检索标识。

模式8是从文献中用人机结合方式抽出关键词,保留关键词作检索标识,并依据某种具有自动赋号功能的分类表给出相应分类号。

模式9是从文献中自动抽出关键词,保留关键词作检索标识,并依据某种具有自动赋号功能的分类表给出相应分类号。

模式10是从文献中自动抽出关键词,将关键词依据某种具有自动赋号功能的分类表转换成分类号作检索标识。

模式 11 是对文献用人工进行主题分析，将析出的主题由标引人员赋予自由标引词（自由标引词属自然语言），并将其编成后控制词表（该表为自由标引词→控制词→分类号三级，起初用人工编制，以后可用机助增补）。

模式 12 是从文献中用人工抽出关键词，并将其编成后控制词表（该表为关键词→控制词→分类号三级，起初用人工编制，以后可用机助增补）。

模式 13 是从文献中用人机结合方式抽出关键词，并将其编成后控制词表（该表为关键词→控制词→分类号三级，起初用人工编制，以后可用机助增补）。

模式 14 是从文献中自动抽出关键词，并将其编成后控制词表（该表为关键词→控制词→分类号三级，起初用人工编制，以后可用机助增补）。

模式 15 是对文献用人工进行主题分析，将析出的主题由标引人员赋予自由标引词作检索标识。

模式 16 是从文献中用人工抽出关键词直接作检索标识。

模式 17 是从文献中用人机结合方式抽出关键词直接作检索标识。

模式 18 是从文献中自动抽出关键词直接作检索标识。

模式 19 是对文献中的（主要是题名或文摘中的）自然语言词逐字自动作单字索引，事实上等于不标引，因而无确定检索标识。

至于机辅标引，可归为人工主题分析—人工赋词、人工赋号一类。

3 增加自然语言与人工语言对应转换功能是情报检索语言走向自动化之路

人工主题分析的质量高于从文献本身抽词。人工抽词和人机结合抽词的质量高于自动抽词。

分类号—控制词对应表比单独的词表和分类表功能多。后控制词表严密性稍差。

人工赋词、人工赋号的准确性决定于标引人员的水平，而自动赋词、自动赋号的准确性则决定于自然语言与人工语言的对应质量。

自动赋词和自动赋号基于自动抽词，因此标引质量有些欠缺，在处理速度和成本上则有明显优势，但有小部分文献在处理时需人工辅助。

自由赋词标引与后控制词表结合，标引较易，标识专指性较好，检索效率有保证，处理速度比使用分类表和词表快，但比自动抽词、自动赋词慢。

编制后控制词表对标引人员的要求较高。

自动抽词＋自动转换（自动赋词、自动赋号）可以有两种结果：一种是在自动转换后不再保留抽出的关键词；另一种是在自动转换后仍保留抽出的关键词。显然是保留关键词好，可多一种文献原用专指词的直接检索途径。既有关键词，也有控制词和分类号的系统，在检索中可随需要灵活选用检索标识。关键词有助于提高检准率和方便新学科、新概念的检索，而控制词和分类号都有助于提高检全率。

有不同程度的自动化。上述 19 种模式都属自动化范围。可归纳为：(1)人工标引，自动检索；(2)人机结合标引，自动检索；(3)自动标引，略加人工辅助，自动检索；(4)自动标引，自动检索。总之，在计算机检索系统中，检索（查找）过程都是在程序控制下自动进行的，但标引过程却可以是多种多样的，有许多种标引模式。

自动或半自动标引（无论是自动抽词或自动赋词还是自动分类）的处理对象都是文献本身的用词（即自然语言）。所以，自动抽词和自动转换是自动标引的主要内容。

对汉语来说，自动抽词即汉语自动分词，

此项技术目前在许多学科、专业领域已达到或接近实用水平，关键在于抽词词典的编制。抽词词典看来只能分别编成专业性的，最好既有“用词词典”(抽取检索有用词)，也有“非用词词典”(排除检索无用词)。抽词词典越丰富和完善，抽词的完全率和正确率越高。

自动转换必须以自然语言与人工语言的对应为前提。通过对应表将自然语言转换成人工语言。所以，把分类表和词表改造成自然语言与人工语言的对应表，是情报检索语言走向自动化的必由之路。

把词表改造成自动赋词适用的词表，实质就是增加大量自然语言词作为入口词。最好把出现在文献(特别是文献题名)中正规的和非正规的词和词组(非正规的词和词组在文献题名等特定语言环境中是可以理解的，但在普通词典和专业词典中却查不到，特别是那些作者自造的词和词组)都收录进词表，对应于控制词之下。对应可以采取等同关系和等级关系的形式。近义关系一部分可视为等同关系，一部分可视为相关关系。采取相关关系的对应，则应慎重。若一个自然语言词或词组可对应多个控制词，应特别注出。在这种词表中，自然语言词可能数倍于控制词，而且要不断补充，故最好做成数据库形式，可不断升级出新版。

体系分类表改造成自动赋号适用的分类词表——词—分类号双向对应表，较为复杂，而且还要制定自动分类规则，但实践证明是可行的。关于自动赋号分类的原理和将体系分类表改造成自动赋号适用的分类词表的方法，详见我的《分类法主题法一体化自动标引系统的基本原理和方法》一文(《图书馆论坛》，此处从略。

后控制词表实质上也是自然语言与人工语言的转换工具，与具有赋词功能的分类号—控制词对应表相似，但编制方法有所差异，其结构可多一种词的轮排表(请参看我的《论后控制词表》，《图书情报工作》1994年第1期)。

《中国分类主题词表》的机读版，无论从增加自动标引功能考虑，还是从增强其机辅标引功能和在计算机系统中更充分地发挥检索功能考虑，都有必要作这种改造。若要使其具有自动赋词功能，只要将第二表改造成为自然语言词—主题词—分类号对应表即可。若要使其具有选定主要分类号的功能，则还需对第一表进行改造，但不必改动原有分类号。

人工语言在检索中的控制作用(产生于其规范化和显示概念关系)是自然语言所无法替代的。所以，自然语言检索系统并不排斥人工语言，高级的自然语言系统必然是与人工语言(或其原理)结合的。在信息高速公路上，自然语言检索将广泛应用，人工语言将成为对自然语言的强有力后控制手段，依然有它的发展前途。人工语言必须与自然语言结合起来。其结合的基本方式就是两者的对应转换。

所以，我以为，情报检索语言中增加与自然语言的对应转换功能，是它走向自动化的关键性环节，今后分类表和词表应向这个方向发展。

张琪玉 1954年北大图书馆学系业业。现为上海空军政治学院信息管理系教授。发文150余篇，出版专著20多种、译著18种。其中《情报检索语言》获1985年国家科学技术进步一等奖。通讯地址：上海五角场，邮码：200433。

(来稿日期：1995—08—21。编发者：徐苇。)