

马张华

论主题检索系统中先组词的选择和使用

ABSTRACT The advantages and the problems of the use of pre-coordinate vocabularies are briefly recorded, the bases and the methods of choosing pre-coordinate vocabularies are discussed and the setting up of the build-in conversion mechanism among the decomposition forms of different vocabularies in the retrieval systems is put forward. 1 tab. 5 refs.

KEY WORDS Subject retrieval systems pre-coordinate vocabularies Choices Uses

CLASS NUMBER G254.2

先组词亦称词组,通常指主题检索系统中直接以词组形式标引和检索文献的词。与先组词对应的是后组词即单元词,是主题检索系统中以单词形式标引和检索文献的词。

先组词问题实质上是语词组配检索系统中词汇组配单元的选择问题和句法问题,它不仅直接影响关系受控标引系统中词表编制、标引词转换及检索等方面,也牵涉到自然语言标引和检索系统中词汇的选择和使用。先组词的使用增加了词汇分割的变化和检索用词的多样性,对标引和检索中词汇的选择以及检索系统的功能、使用效果等具有深刻影响。但目前我国图书情报界对先组词的选择和使用还缺乏足够的重视。本文试图对先组词使用的意义、先组词选择规则及其在检索中的使用等进行研究和探讨,以求教于图书情报界同行。

1 使用先组词的必要性和问题

在采用后组词的同时结合使用一定数量的先组词,是现代主题检索系统的一个基本特点,是由主题检索系统的需要决定的。它既

有一定优点,也带来相应的问题。其优点是:

(1) 有助于增加标引和检索的直接性。例如在手检系统中,“图书馆业务辅导”远比“图书馆—业务—辅导”更为直接;“大学图书馆—图书流通率”远比“大学—图书馆—图书—流通率”直接。

(2) 有助于增加标引和检索的准确性。例如“橡胶工业”比“橡胶—工业”准确,后者除可以表示橡胶工业外,也是工业用橡胶的多面组配形式。又如“学校图书馆”比“学校—图书馆”准确,后者除可用于表示学校图书馆外也有图书馆学校的意思。

(3) 有助于揭示专业术语的相关性,建立相应的语义网络。如在不使用先组词的情况下,下述词间关系不可能建立:

化学工业

S 重工业

F 发酵工业

合成工业

合成洗涤工业

石油化学工业

橡胶工业

制药工业

(4) 有助于改进总论性资料的查找。例如在元词系统中,关于各类图书馆的文献均集中于“图书馆”一词下,要检出总论图书馆的文献十分困难,而在采用先组词的情况下,各类图书馆如公共图书馆、国家图书馆、科学图书馆、大学图书馆、儿童图书馆等均分别用专词标引,从而改善了总论图书馆文献的查准率和可检性。

(5) 有助于概念组配的进行。概念组配不可避免要使用一定数量的先组词。例如:小儿肺炎的概念组配形式为“儿童疾病——肺炎”;建筑用水泥的概念组配形式为“建筑材料——水泥”;均使用了先组词。

采用先组词的问题主要是:

(1) 多数先组词表达的概念可以通过系统中已收入的后组词组配表达。以先组词形式代替元词组配,必然会增加检索系统的词汇总量。而且先组词比例越高,先组词越专指,检索系统的词汇总量就越大,从而增加系统的管理费用。

(2) 增大标引和检索时词汇选择的难度。按照叙词组配规则,在词表收入多个相关词的情况下,叙词的组配标引必须使用词表中两个或两个以上与被标引概念关系最密切的词。如在词表已收入“眼科学”、“历史”、“自然科学史”、“医学史”的情况下,有关眼科学史的文献必须标引为“眼科学—科学史”。又如在词表收入“制糖”、“工厂”、“食品厂”等词

的情况下,有关制糖厂的文献必须标引为“制糖——食品厂”。显然,后组词、先组词的并存造成了标引中选词的不确定性,增大了标引工作的难度。对检索用户来说,困难更大。很难想象,一般检索用户会使用“制糖——食品厂”、“眼科学——医学史”去检索有关制糖厂和眼科学史的文献。

以上分析说明,先组词的使用是不可避免的。在使用先组词的过程中,应充分发挥其长处,限制和避免其可能带来的问题。尤其是在词表编制或标引阶段,应根据检索系统的特点和要求做好先组词的选择;在检索使用时,针对先组词可能带来的问题,改进使用方式,增加系统的易用性。

2 先组词的选择

先组词的选择与语词特点和检索需要密切相关,具有较大灵活性。要做好先组词的选择,就要制订适用的先组词控制规范。

对于主题检索系统中先组词的选择,我国图书情报界已进行过不少探索,取得了一定的进展,但仍存在不少问题。从有关研究和词表编制实践看,目前我国主题词表收入的先组词可分为两类:一是从主题揭示和词汇组成特点的角度,直接采用先组形式的复合词(见表 1)^[1,2];二是可以通过限定组配或交叉组配表示,但从检索系统的实际需要出发,

表 1 从主题揭示和词汇组成角度应采用先组方式的复合词类型

应采用先组方式的复合词类型	举 例
1 分解后,一分解概念失去检索意义的复合词	剩余价值、激光照排
2 分解后,一分解概念改变其含义的复合词	雪崩、二极管、猎户星座
3 经组配,可能出现二义的复合词	橡胶工业、分析化学
4 经组配,可能会转变其含意的复合词	北京大学、北京图书馆

习惯上直接选用的常用复合词。比较而言,前一种情况规则明确,容易操作;后一种情况则

不够具体,灵活性较大,是造成主题词表中先组词选择过多的重要原因。据蔡润的调查,我

国 62 部专业叙词表的先组度平均值高达 64.8%, 部分词表的先组度超过 70%^[3], 其原因主要是对常用复合词的选入缺乏必要的控制。因此, 要做好我国主题词表中先组词的选择, 目前的重点是制订常用复合词的选择规范。必须明确其选择依据和范围, 不断改进选择方法, 使选择具有可操作性。笔者以为先组词的选择依据应包括下述几个类型或来源:

其一, 按照用户保证原则, 为方便用户使用, 将用户检索中经常使用的、检索频率高的常用概念, 以复合词的形式选入系统。

其二, 按照文献保证原则, 对已有较大文献量的特定主题对象, 增设先组形式的特称叙词, 改进系统对总论性资料的检索能力。如在“图书馆”一词下视需要增设“大学图书馆”、“科学图书馆”、“儿童图书馆”等, 在“历史”一词下视需要增设“哲学史”、“经济史”、“民族史”、“文学史”等。

其三, 根据参照系统或词汇分类系统建立语义网络的需要, 适当增设必要的、对语义网络结构具有关键联结作用的复合词。

对常用先组词选择的一个突出问题是选择缺乏客观依据, 带有较大的主观任意性。这是造成先组词数量多、质量失控的主要原因。要保证常用先组词的质量, 必须建立先组词选择的控制规范, 至少应包括以下内容:

(1) 逐步建立入选的量化指标。即对复合词选取中的文献保证或用户提问保证, 规定一定的数量标准作为系统入选的依据。这有利于避免先组词选取的任意性, 也便于在机检系统中实施。

(2) 规定词长限制。目前有的词表中先组词最长的超过 15 个字, 不利于充分发挥机检系统的组配检索能力。除少数专有名词, 多数先组词应保持在 6 个汉字以内, 一般不超过 10 个汉字。

(3) 严格控制因建立语义网络需要增设的先组词。一般限于少数在语义网络中起关

键联结作用的复合词, 不应为参照系统或词汇分类系统末端的完整而增设复合词。在词汇分类系统展开的中间环节也可适当采用组配形式代替复合词。

(4) 应明确不宜采用先组词的组配类型。如根据主题因素之间关系, 下述类型词汇一般不应选作先组词^[4]: 事物及部分组成的复合词形式; 事物与方面包括材料、性质、过程、操作、工艺等组成的复合词形式; 专有名词与通用概念组成的复合词形式; 出版物类型与出版物内容构成的复合词形式; 联结关系组成的复合词形式。

3 检索系统中先组词的易用性处理

一个理想的网络检索系统应能使用各种词汇和不同语词切分形式, 以组配的方式自由检索相应的主题内容。要达到这一目标, 必须对不同的词汇分解形式进行必要的处理。先组词使用中的最突出问题是容易在标引和检索中造成词汇选择的多样性和不一致性。

在机检系统中改进先组词的易用性, 必须从方便用户的角度出发提供各种可能的方法和机制, 主要包括以下几个方面:

(1) 提供有关先组词使用的联机帮助信息, 向用户介绍组配检索的基本知识, 包括介绍词汇分解、先组词选择的基本规则和方法, 供用户使用时参考。在可能的情况下规范用户的检索用词。

(2) 提供丰富的组代词机内转换机制, 使用户可以通过各种不同的词汇分解形式, 直接进行查找。

(3) 提供与先组词对应的常用组配词汇的屏幕显示。它在形式上类似《中国分类主题词表》中的词串, 但表达的对象则是系统中收入的先组词或部分常用的组配形式。表中所有词汇组配形式按字顺方式显示, 以便用户检索或系统人员管理时浏览参考。

(4) 建立检索用组代词或先组词的增补

机制。即依据用户保证原则,将用户检索提问中出现频率较高的词汇组配形式或先组词通过人工帮助纳入检索系统,使之与相应标引词建立联系。

在上述几个方面中,建立丰富的组配词转换机制是改进先组词易用性的关键。在元词系统中,词汇组配单元为元词,组配检索有明确的规律可循。先组词的引入使词汇分解增加了某种不确定性。要求终端用户能按标引规则的要求,以标准的组配形式检索显然是不合理的。因此,在使用先组词的情况下,要使终端用户能方便检索,就必须将词汇的各种不同分解形式像处理同义词一样收入机内转换系统,使得用户可以以各种分解形式直接查找。这是一个成熟的用户友好的文献检索系统必须具备的条件。国外词表对此较为重视,如《美国国会标题表》对“馆际互借”这一词条,就使用了多种用代关系:

Inter-Library Loans	馆际互借
U F Book Lending	代图书借阅
Book, lending of	图书, 借阅的
Loans, Inter-Library	借阅, 馆际

这就使得用户得以从多种词汇形式入手,对该条目进行查找。目前我国已出版的专业词表中的等同率平均值仅 15%,远低于国外词表的 62%^[5],已建立组代关系的词表更是寥寥无几。这是与检索系统将用户方便放在首位的宗旨格格不入的。在我国主题检索系统的建设中,首先应该转变观念,从方便用户出发,建立丰富的入口词,并为先组词使用可能带来的多种形式的组配用词提供检索入口。

要解决先组词使用带来的词汇组配形式的不一致性,检索系统应提供的机内转换机制至少应包括下列形式:

提供已入选先组词与相应检索用的后组词组配形式之间的转换。如:

企业+ 改革	(检索词)
见 企业改革	(入选词)

提供已入选组配形式与检索用先组词之间的转换。如:

女法官	(检索词)
见 女性+ 法官	(入选词)

提供已入选先组词与多种可能出现的组配形式的检索用词之间的转换及多个入选词组配形式与检索用先组词之间的联系。如:

基本建设+ 会计	(检索词)
基本+ 建设+ 会计	(检索词)
见 基本建设会计	(入选词)

文艺下乡	(检索词)
见 文艺演出+ 农村	(入选词)
文艺演出+ 牧区	(入选词)

提供一个已经入选的组配形式与检索用的各种可能出现的组配形式之间的对应转换。如:

眼科学史	(检索词)
眼科学+ 自然科学史	(检索词)
眼科学+ 历史	(检索词)
见 眼科学+ 医学史	(入选词)

淡水养殖+ 鱼类	(检索词)
淡水鱼类+ 养殖	(检索词)
淡水+ 养殖鱼类	(检索词)
淡水+ 养殖+ 鱼类	(检索词)
见 淡水鱼类+ 养殖鱼类	(入选词)

上述转换形式的建立,应以用户检索提问为基础,依据文献保证原则。这在计算机检索系统中不难做到。上述转换形式既可用于受控检索系统,亦适用于在自由标引的基础上建立的主题检索系统。在后一种情况下,表现为不同标引词、检索词之间的连结,标引词与入口词之间无明显界限。上述转换形式采用了概念组配和字面组配并存的方式。以第四种转换形式为例,按照叙词标引规则,淡水养殖鱼类这一主题只能使用入选词“淡水鱼

邱均平

信息资源对社会发展的影响和作用

ABSTRACT The influence of information resource on social development is shown in 8 aspects: change of human society, social and economic development, prospering the nation by science and education, progress of science and technology of society, construction of spiritual civilization, cultural development of society, citizen's quality and the comprehensive influence of information technology on social development. 3 refs

KEY WORDS Information resources Social development Actions

CLASS NUMBER G352.1

信息资源之所以在当今社会受到人们的青睐,得到普遍的重视和广泛利用,其根本原因在于它对人类社会的生存和发展具有十分重要的作用。信息,作为构成客观世界的三大要素之一,其基本作用就是消除人的认识的不确定性,增强世界的有序性。如果没有物质,那么就没有人类生活的世界;没有能量,世界就会消亡;而没有信息,则物质和能量也只能形成一个混沌、杂乱的空间。信息资源对社会发展的影响就是这种基本作用在特定人类社会形态中的具体表现。以此为出发点,下面考察

和探讨信息资源在与社会发展关系十分密切的几个主要方面的重要影响和巨大作用。

1 信息资源与人类社会变革

在人类进化和社会发展过程中,信息资源是一种不可缺少的前提条件和推动社会进步的重要因素,这主要表现在以下几个方面:第一,信息资源是人类生存的必备条件。人类在生存、延续和繁衍的进化过程中离不开信息。在信息环境之中,人的活动实际上是一个

类”、“养殖鱼类”标引,其他组配形式均不要求。但如规定用户必须按这一要求才能检出,必然影响检索系统的实际使用效果。终端用户采用的各种组配检索用词包括字面组配形式,只要其含义是明确的,均应当允许作为合理的检索形式收入系统,这样才能彻底解决使用多元词带来的检索困难。至于在自然语言标引系统中,则应建立更加丰富的不同词汇切割形式之间的联系,使用户可以用各种想到的词汇和组配形式方便地检索。

参考文献

- 1.4 侯汉清,马张华 主题法导论 北京:北京大学出版社,1991
- 2 国际标准化组织,文献与情报工作国际标准汇编 北京:中国标准出版社,1980
- 3.5 蔡润 我国汉语专业叙词表的分析评价. 北京大学硕士学位论文,1996

马张华 北京大学信息管理系教授. 通讯地址:北京市中关村,邮编 100871

(来稿日期:1996 10 15. 编发者:徐苇.)