

●严 峰

论现代信息技术条件下检索语言的应用及检索效率变化

摘 要 现代信息技术条件下,各类检索语言适应程度和用户的检索习惯等因素会对其检索效率产生影响。以自然语言为标识的检索语言最能与现代科学技术发展相适应。信息检索的现代化信息环境形成及信息载体的变革直接影响情报检索的方式和技术。人工创制的检索语言之间的融合,以及人工创制的检索语言与自然语言的和谐统一是一种检索语言变化趋势。

关键词 检索语言 检索效率 关键词语言 分类主题一体化语言 表达式检索语言
分类号 G354

ABSTRACT Under the conditions of modern information technology, the adaptability of various search languages and the usual practices of users will affect the efficiency of searches. The search languages marked by natural languages are best consistent with the development of modern science and technology. Modern information environment of information search and the revolution of information media have direct influence on the ways and techniques of information search. The author thinks that the integration of various artificial search languages and the unity of artificial languages and natural languages are the directions of the development of search languages.

KEY WORDS Indexing language. Search efficiency. Keyword language. Classification-subject integration language. Statement search language.

CLASS NUMBER G354

由于各类检索语言的功能是有区别的,特别是在现代信息技术条件下,各类检索语言适应程度是不同的,用户的检索习惯也会发生变化,这些因素会对检索语言的检索效率产生影响。本文试图结合网络上数字化文献信息资源的检索实践,来论述检索语言在现代信息技术条件下检索习惯和检索效率的变化。

1 现代信息技术条件下新型检索语言的应用

在现代信息技术条件下,人们对检索语言在传统手工条件下的一些要求,由于计算机功能的实现,已经变得不那么重要了。一般用户(读者)在检索上所具备的能力要求也有所降低,计算机检索界面已变得越来越“傻瓜”型了。传统检索语言的应用也发生着变化,分类检索语言和主题检索语言在一定领域里仍起着严肃的检索作用,而关键词语言和分类主题一体化语言在更加广泛的领域里的应用大有异军突起之势。

1.1 关键词语言将是越来越得到广泛运用的“大众化”检索语言

关键词语言是直接从文献的名称、正文或文摘

中抽出的具有检索意义的自然语词,它比较适用于计算机编制各种索引。关键词语言是一种基本上未经过规范化处理的自然语言,大量存在一词多义、多词一义或词义含糊的情况,检索用词无法一一对应,故会造成文献信息的漏检和误检。但是在计算机检索功能高效运行的条件下,人们对关键词语言的缺点有所“忽视”,反而充分发挥出了它的简便易用的优点,大量用于网络环境下的数字化信息的检索。

当前,关键词语言在 Internet 网站信息检索中得到广泛的应用。例如“搜狐”网站的中文检索、“Yahoo! 中国”网站的中文检索、“新浪”网站的中文检索、Excite Chinese 网站的中文检索等。

例如,“搜狐”网站()的中文检索说明就明确指出:“如果很清楚要找的网站(或新闻)主题,可以在检索框内键入关键字(Keyword),并单击旁边的搜索按钮,搜狐中文搜索引擎会返回检索结果。”“Yahoo! 中国”网站()的中文检索说明也明确指出:“可以在检索栏位内键入想要找的关键字串(Keyword),并按 Search 键,“Yahoo! 中国”搜索引擎会搜寻分类类目、资料库中的网站信息以及新闻资料库。”

在图书馆建立的网站上,包括国内一些著名的数

字图书馆网站上,关键词语言也在信息检索(其中包括题名、提要、文摘、全文等检索)中得到广泛应用。

例如,国家图书馆网站(<http://www.nlc.gov.cn>)的中西文书刊目录检索、中国数字图书馆网站(<http://www.d-library.com>)、超星数字图书馆(<http://www.ssreaders.com>)的书目检索和全文检索、华北制药数字图书馆(<http://lib.ncpcrd.com.cn>)的全文检索等。

1.2 在计算机支持下的检索公式已成为一种新型的表达式检索语言

人们在传统手工方式条件下创制的标识单元检索方式设想,在计算机网络检索技术的支持下得以实现。在采用关键词语言进行文献信息检索的同时,Internet 信息网站在搜索引擎中充分使用代码和布尔逻辑符号来表达比较复杂的检索公式,从所起作用和功能上看,这种检索公式实际上已成为一种新型的检索语言,虽然它的标识是以公式化的形式出现的,但是符合检索语言“能表达一系列文献信息主题概念及其相互关系的人工创制的概念标识系统”的定义,是非常值得研究的。表达式检索语言主要有以下种方式:

(1) 同时使用几个关键字,通过选择适当的逻辑操作符“与(and、AND、&、+、同时)”、“或(or、OR、|、或者)”和“非(-)”等得到相应的结果。例如:

搜狐中文检索系统搜索引擎的语法和逻辑操作符:

AND:表示前后两个词是“与”的逻辑关系。如关键字:“中国 AND 北京”、“中国 北京”以及“中国 and 北京”都会将所有包含“中国”并且包含“北京”的页面搜索出来。

OR:表示前后两个词是“或”的逻辑关系。如关键字:“中国 OR 北京”会将所有包含“中国”或者包含“北京”的页面搜索出来。

“Yahoo! 中国”网站的中文检索,查询时逻辑操作符的使用为:

使用空格、逗号(,)、加号(+)和 & 表示“且”的关系。例如想查询关于女歌手的网页或新闻,则输入关键词“女性 歌手”或“女性+歌手”或“女性 & 歌手”。

使用减号(-)搜索表示“非”的关系。例如想查找服装方面网页或新闻,但不包含内衣,输入关键词“服装-内衣”。

使用字符“|”表示“或”的关系。例如想查询关于乒乓球或网球方面的网页、新闻,则输入关键词“乒乓球|网球”。

表示表达式是一个整体单元,使用字符“()”。例如想查找计算机方面的网页或新闻,但不包含“软件”与“硬件”,输入关键词“计算机-(软件 硬件)”。

Excite 中文搜索引擎使用布尔检索符号“AND”、“OR”、“NOT”和“()”进行检索。

检索结果必须含有所有用“AND”连接起来的提问词。例如:如果你想要检索所有含有“经济”和“贸易”这两个词的文献;

检索结果必须至少含有一个用“OR”连接起来的检索词。例如:如果你想要检索有关体操或游泳的文献;

检索结果不能含有紧接在“NOT”后面的提问词。例如:如果你想要检索有关汽车但又不包括本田在内的网页;

如果要求检索结果含有“水果”一词以及“香蕉”与“苹果”任一词,可以使用检索表达式:“水果(香蕉 OR 苹果)”。如果要求检索结果要含有某个关键词或关键字,可在这个关键词前面加上“+”;相反,如果希望这个关键词不要出现在检索结果中,那么在这个关键词前面加上“-”。例如:“+储蓄+定期-活期”。

国家图书馆网站的中西文书刊目录检索使用“同时”、“或者”来表达关键词之间的逻辑关系:

例如想查询关于女歌手的文献信息,则分别输入关键词“女性”、“歌手”,并确定为“同时”。

例如想查询关于乒乓球或网球方面的文献信息,则分别输入关键词“乒乓球”、“网球”,并确定为“或者”。

(2) 在用某一关键词进行检索时,计算机系统会通过内置的相关词表提示出相关的检索词,以扩大检索路径。

这是 Excite 中文检索的特点之一,它可以根据所输入的提问式,动态地推荐一些相关的提问词(字),目的是帮助用户更正确地表达检索需求,更容易地添加检索词到检索提问式中。例如:如果你使用“宾馆”一词进行检索,检索智囊会推荐给你“hotels”、“饭店”、“酒店”、“三星”、“大厦”等词。

在“新浪”搜索一个关键词时,在结果页面上方会出现与被搜索词相关的一些关键词的链接,

点击一个链接,就可以得到相应关键词的搜索结果。譬如,当用户搜索“手机”时,结果页面上方会出现:手机价格、手机炒股、手机音乐、手机报价、手机铃声等和手机相关的关键词链接,点击某个链接,譬如“手机音乐”,接下来就会出现搜索“手机音乐”这个词的结果页面。

(3)采用“前方一致”、“任意一致”、“等于”、“不等于”等限定选择功能,提高检索效率。

在国家图书馆网站的中西文书刊目录检索和中国数字图书馆网站的书目检索页面就有此功能,使检索者能选择确定所用检索词在检索式中的位置。

(4)采用限定时间范围的方法,以缩小检索范围,提高检索效率。

如在“搜狐”网站中,可以指定查询结果中网页的生成时间,目前有四种选择:“任何时间的网页”(默认选择)、“三个月内的网页”、“六个月内的网页”、“一年内的网页”。生成时间是一个下拉菜单,可以从中选择。

(5)采用“进阶查询”和“相关检索”的方法,逐步提高检索专指度。

如在“新浪”网站中,要求检索者开始时不要把条件限制的过于严格,最好是检索出一些结果后,再使用“进阶查询”和“相关检索”的方法来排除检索结果中不需要的内容。当使用关键词查询的网页检索时,除得到相应的检索结果之外,还提供“进阶查询”和“相关检索”功能,它能大大提高搜索效率,减少找到有效信息所花费的时间。“进阶查询”和“相关检索”的具体含义如下:

在网站查询结果页面中继续查找信息,有三种选择:重新查询、在结果中再查、在结果中去除。

重新查询:用本次键入的关键词重新进行查询(是默认的选项);

在结果中再查:在上一次的关键词查询结果中再用本次键入的关键词进行查询,这两次的查询结果是同时包括这两次的关键词的;

在结果中去除:在上一次的关键词查询结果中去除本次键入的关键词再进行查询,这样的查询结果是包括上一次的关键词但不包括本次的关键词的。

(6)利用双引号“”进行词组检索。

“词组检索”也称“完全符合检索”,检索结果必须含有与提问式完全一样(包括次序)的字串。当检索者要搜索比较专指的文献,可使用双引号“”进行

词组检索。

例如:“数据库管理”、“中国人民银行”等。

需要说明的是,用这种方法在搜索中有的网站用关键词反而查找不出来,这种情况通常是因为使用了不恰当的限定条件,导致有用的信息被滤掉。例如输入的关键词是:“金庸武侠小说天龙八部”,新浪搜索中是有金庸的武侠小说“天龙八部”的,但用了这么长的关键词,就会导致搜索时找不到完全匹配的网站。如果直接输入关键字“天龙八部”就能得到想要的结果了。

(7)网页检索结果中重复内容的合并处理。

为提供给用户最有价值的信息,避免数量过多且重复的检索结果影响用户的使用,新浪搜索对网页检索的结果进行了技术上的处理,包括两种情况:一是站点类聚,它是指在检索结果中,如果来自同一站点的网页多于一篇,则除了最相关的一篇外,其余均被隐藏起来;同时会为这个站点提供一个相关链接,以检索在这个站点上的其它相关网页,如果用户需要此站点上更多的信息,则可点击“此站点上的更多结果”链接来得到所需信息;二是内容类聚,指在检索结果中,如果某几个结果的网页内容相同,则只保留一篇,其余被隐藏起来。新浪搜索同时采用了这两种类聚技术方案,这样处理得到的检索结果更方便用户选择自己所需要的信息。

1.3 分类主题一体化语言

我国第一部综合性分类主题一体化检索语言工具是《中国分类主题词表》。这是一种人工创制的、将分类检索语言和主题检索语言的功能加以融合的新型检索语言。分类主题一体化语言由两大部分组成,即“分类号—主题词对应表”和“主题词—分类号对应表”。“分类号—主题词对应表”是从分类到主题、从分类号到主题词的对照索引体系;“主题词—分类号对应表”是从主题词到分类号、从标题到分类号的对照索引体系。

分类主题一体化语言的功能主要体现在:分类和主题可以互为索引,降低标引难度,提高标引质量,节省人力和物力,适合于计算机和网络上的应用,检索的检准率和检全率指标都较高,实现分类号和主题词的混合检索。《中国分类主题词表》的整体功能远远超过了《中图法》和《汉语主题词表》,实现了先组式检索语言和后组式检索语言结构语义等方面的兼容。例如,当检索者搜索一个关键词时,计算机能自动通过内置的分类主题词表,与分类号标

识和主题词标识进行比较。检索词如果是正式词时,计算机系统会自动将已标引的有关文献信息显示出来,同时还会将与这个检索词相对应的分类标识下的文献信息也显示出来;检索词如果不是正式词,计算机系统会通过词表中的入口词(或分类表中的交替类目)的指引,显示正式词条目下的文献信息。用户要检索“红楼梦”,分类主题词检索语言系统不但显示出与“红楼梦”有关的文献信息,同时还会将与“红楼梦”这个检索词相对应的分类标识下的“红楼梦研究”、“中国古代小说”、“中国古代文学研究”等文献信息也显示出来,对研究型读者来说是非常方便的。

2 检索语言的发展趋向简便实用

在计算机与通讯技术日异发达的今天,人们更加深刻地认识并体会到检索语言无论是依赖结构复杂的叙词,还是依赖标识直观性差的分类表,都越来越难以适应数量骤增、层次不同、需求各异的数字化信息资源用户检索要求,检索语言的发展必然要向简便实用的方向发展。

2.1 以自然语言为标识的检索语言最能与现代科学技术发展相适应

当前返朴归真已成为检索语言研究中的一条重要原则,分类语言和叙词语言存在的标引难度大、速度慢、词汇更新滞后,对标引和检索人员要求过高弊端,在联机环境中明显地暴露出来。而某些“规范词”虽然很完美,但繁琐、生僻,用户很不适应,采用贴近自然语言的检索语言,如关键词等形式更容易被用户接受。联网环境用户的变化呼唤自然语言登台亮相。

2.2 信息检索的现代化信息环境形成及信息载体的变革直接影响情报检索的方式和技术

情报检索正在由以往的对文献线索的检索,发展到全文检索、数值检索、事实检索、超文本检索、多媒体检索;由文献检索发展到信息检索;由相关性检索发展到直接性检索,使检索结果更贴近用户

的需求。超文本、多媒体、全文检索等技术及计算机网络技术的发展使检索语言的联机显示、采用各种表达式或表达符号、多数据库查询中的词汇兼容与互换、自然语言检索接口、自动标引以及语义网络推理、识别和转换等问题的研究和应用成为现实。

2.3 人工创制的检索语言之间的融合,以及人工创制的检索语言与自然语言和谐统一是一种检索语言变化趋势中的“回归”,是情报检索语言更高层次的和谐与统一

源于计算机技术和网络技术的飞速发展以及人类对情报检索语言实际应用要求的变化,对情报检索语言的应用市场产生了极其深刻的影响,促进了情报检索语言的改造和创新。这种改造和创新的实质是对概念词汇控制的科学性、规范化程度以及方法等规范标准的创新。分类主题一体化检索语言和“表达式检索语言”就是人工创制的检索语言之间的融合和人工创制的检索语言与自然语言相结合的成功典范之一,这些都有助于情报检索语言的自然语言理解和自然语言处理,使情报检索语言趋于自然语言化、智能化和“大众化”。

综上所述,新的信息环境引起了情报检索系统的大变革,同时也使作为情报检索系统重要组成部分的情报检索语言得到发展和创新。这种发展和创新是顺应应用市场的需要的。

参考文献

- 1 张琪玉. 情报语言学基础. 增订2版. 武汉:武汉大学出版社,1997
- 2 陈树年.《中国分类主题词表》标引手册. 北京:北京图书馆出版社,1998
- 3 “搜狐”网站中文检索说明(<http://search.sohu.com>)
- 4 “yahoo! 中国”网站检索说明(<http://cn.yahoo.com>)
- 5 “新浪”网站搜索帮助信息(<http://search.sina.com.cn>)

严峰 金陵图书馆副研究馆员,副馆长。通讯地址:南京市长江路262号。邮编210018。

(来稿时间:2001-06-04)

程焕文教授当选 OCLC 成员委员会代表

2001年9月10日,从美国俄亥俄州都柏林 OCLC 总部传来消息:中山大学图书馆馆长、网络教育学院院长、信息管理系教授程焕文先生以最高的选票当选为 OCLC 成员委员会代表。这是程焕文教授继1997年当选国际图书馆协会联合会(IFLA)国际书目控制专业组通讯委员和图书馆史圆桌会议执委会委员之后,又一次出任国际图书馆界的重要职位,也是我国大陆图书馆界人士第一次当选 OCLC 成员委员会代表。(日月谭)