●袁国发 杨莉英 杨 健 袁立军

中文期刊刊名的计算机快速检索 *

摘 要 利用中文期刊刊名每个字汉语拼音的首字母,为中文期刊刊名排序和同类号期刊排序,并在计算机上进行,排序和检索速度快。尤其是利用刊名的前 4 个拼音首字母,计算机检索极为便捷。

关键词 中文期刊 刊名检索 计算机检索
分类号 G252.7

ABSTRACT In this paper, the author thinks that the computer sorting by initials of Pinyin Romanization of Chinese periodicals titles will be more efficient in search, and can be applied in library services.

KEY WORDS Chinese periodicals. Title search. Computer search.
CLASS NUMBER G252.7

"中文期刊刊名汉语拼音首字母 - 全拼排序法",是利用中文期刊刊名的每一个汉字汉语拼音的第一个字母和中文期刊刊名汉字的汉语拼音及其声调等作为排序条件而实现中文期刊的刊名排序和同类号期刊排序的。该排序法简单、方便、快捷,好掌握,易操作,极易推广普及。用计算机排检,不需要编制任何辅助程序,在任何装有汉字系统的计算机上都可实现,只需为每种期刊配一对应的刊名汉语拼音首字母即可。这些刊名首字母在计算机数据库中作为一个与刊名匹配的独立字段存在,属于字符型字段,专为期刊的排序、检索使用。它和中文期刊数据库并存,其输入一劳永逸。

实践中,我们在用该排序法对中文期刊进行刊名计算机检索时,为了更加快捷,遇到较长的刊名,常常有意识地少打后边的几个首字母,结果,同样可以实现检索的目的,使工作变得更加轻松自如。如《生物化学与生物物理学报》,只要打出前 4 个首字母 SWHX 就行了,甚至再少打一个字母也可达到同样效果。很明显,减少后边刊名首字母,是提高中文期刊刊名检索速度的有效途径。但使用几个首字母比较理想?为此,我们又对占我国 1998 年中文期刊总数的 98 %的 7840 种期刊进行了一次考察。

我们首先将中文期刊数据库中与刊名匹配的首字母由前往后分别按取前 2 个、前 3 个、前 4 个、前 5 个、前 6 个 5 种情况进行统计。以所取首字母相同的所有期刊为一个"品种",称为重首字母期刊。同一个品种的重首字母期刊,有几种就被视为"重几种"。以重的种数数量相等的所有品种的重首字母期刊划为一"类"。以取前 4 个刊名首字母的期刊为例,前 4 个刊名首字母为 GYWS 的共 3 种,它们是《贵阳文史》、《工业微生物》和《工业卫生与职业病》。这 3 种期刊则被视为"重 3 种"的同一个"品种"的重首字母期刊。而《职业医学》、《中医药学报》、《中医药信息》和《遵义医学院学报》这 4 种期刊,则是前 4 个刊名首字母为 ZYYX 的"重 4 种"的另一个"品种"的重首字母期刊。在整个数据库取前 4 个刊名首字母的期刊中,"重 3 种"的期刊共 226 个品种 678 种,它们被视为"重 3 种"的一"类"。"重 4 种"的期刊共 104 个品种 416 种,被视为"重 4 种"的一"类"。无相同首字母的期刊则被称为无重首字母期刊,为另一"类"。所有的"类"再按所重首字母期刊种数的多少分为无重、重 2 种、重 3 种、重 4 种……,出现以下结果:

取前 2 个首字母时,所考察的期刊分为 72 类。其中,无重首字母期刊为第 1 类共 54 种,占总测刊数的 0.69 %;重首字母期刊共 71 类 410 个品种 7786 种为第 2 —72 类,占总测刊数的 99.31 %。重 2～20 种的第 2 —20 类期刊共 19 类 252 个品种 1893 种,占总测刊数的 24.15 %。前 20 类期刊共计 1947 种,它们在 DOS 环境下和在 Windows 环境下检索都不用翻屏(在 DOS 环境下,一屏幕可显示 20 种;在 Windows环境下,一屏幕可显示 29 种),共占总测刊数的 24.83 %。重 20 种以上的期刊共 52 类 104 个品种 5893 种,占总测刊数的 75.17 %,这些期刊在 DOS 环境下检索均需要翻屏。重 29 种以上的期刊共 44 类 73 个品种 5101 种,占总测刊数的 65.06 %,这些期刊在 Windows 环境下检索均需要翻屏。在取前 2 个首字母的期刊中,一类重首字母期刊涵盖 1～40 个品种的重首字母期刊,同一个品种的重首字母期刊所重种数最多的达 839 种。

当取前 3 个首字母时,所考察的期刊共分 39 类。其中,

* 本文为河北省教育厅指令性社科科研项目的研究成果。

无重首字母期刊共 1380 种,占总测刊数的 17. 60 %;重首字母期刊共 38 类 1403 个品种 6460 种,占总测刊数的 82. 40 %。重 2～20 种的共 18 类 1379 个品种 5395 种,占总测刊数的 68. 81 %。前 20 类的 6775 种期刊检索时都不用翻屏,它们共占总测刊数的 86. 41 %。重 20 种以上的共 20 类 23 个品种 1065 种,占总测刊数的 13. 58 %,它们在 DOS 环境下检索均需要翻屏;重 29 种以上的共 17 类 19 个品种 970 种,占总测刊数的 12. 37 %,这些刊在 Windows 环境下检索均需要翻屏。在取前 3 个首字母的期刊中,同一品种的重首字母期刊所重种数最多的达 98 种。取前 3 个首字母的期刊虽比取前 2 个首字母的期刊需要翻屏检索的品种都大幅度减少,翻屏次数较多的品种也大为减少,但需要翻屏检索的期刊仍大量存在,不需要翻屏检索的期刊中,在同一屏幕上显示期刊种数较多的品种也大量存在,查找不便,不够理想。

当取前 4 个首字母时,所考察的期刊分为 19 类。其中,无重首字母期刊 4156 种,占总测刊数的 53. 01 %;重首字母期刊共 18 类 1274 个品种 3684 种,占总测刊数的 46. 99 %。这些重首字母期刊多集中在 2 —16 类,共 3583 种,占总测刊数的 45. 70 %,且绝大部分在 2 —8 类中,是重 2～8 种的,共 3345 种,占总测刊数的 42. 67 %。占总测刊数 95. 68 %的期刊一屏幕仅显示 1～8 种。重 9～16 种的 9 —16 类中都分布很少,8 个类一共才 20 个品种 238 种,仅占总测刊数的 3. 04 %。1 —16 类不用翻屏检索的期刊占总测刊数的 98. 71 %。重 16 种以上的只有 3 类 3 个品种,即 ZGSY、ZGYX 和 GWYX,分别为 24 种、30 种和 47 种,共计 101 种,仅占总测刊数的 1. 29 %。

当取前 5 个首字母时,所考察的期刊共分 8 类。其中,无重首字母期刊 6056 种,占总测刊数的 77. 24 %;重首字母期刊共 7 类,多分布在 2 —7 类中,为重 2～7 种的,共 1764 种,占总测刊数的 22. 5 %。1 —7 类期刊共 7820 种,占总测刊数的 99. 74 %。仅 ZGYXW 这一个品种的期刊种数最多,为 20 种,占总测刊数的 0. 26 %。取前 5 个首字母时,所有期刊在 DOS 环境下和在 Windows 环境下检索都不用翻屏。

当取前 6 个字母时,所考察的期刊也分为 8 类。其中,无相同首字母期刊 6393 种,占总测刊数的 81. 54 %。重首字母期刊也共 7 类,多分布在 2 —7 类之中,为重 2～7 种的,共 1428 种,占总测刊数的 18. 21 %。1 —7 类期刊共 7821 种,占总测刊数的 99. 76 %。仅 ZGYXWZ 一个品种的期刊种数最多,为 19 种,占总测刊数的 0. 24 %。取前 6 个首字母时,所有期刊在 DOS 环境下和在 Windows 环境下检索也都不用翻屏。

从以上情况看,取前 2 个首字母检索无采用价值,取前 3 个首字母检索不够理想,取前 5 个首字母和取前 6 个首字母相比,品种分布差别不大,而取前 6 个首字母要比取前 5 个首字母多敲 1 键。取前 5 个和取前 6 个首字母检索都不需要翻屏,而且一屏幕显示的期刊种数也都不多,取 6 个以上首字母检索就无必要了。取前 5 个首字母检索和取前 4 个首字母检索相比,后者虽存在翻屏检索的问题,但所占比例不高,需翻屏检索的 3 个品种的重首字母期刊才占总测刊数的 1. 29 %;不需翻屏检索的重首字母期刊一屏幕所显示的期刊虽比取前 5 个首字母检索时要多一些,但一屏幕最多才显示 16 种,且比取前 5 个首字母多出现的重 8～16 种的数量很小,仅占总测刊数的 4. 06 %,而占总测刊数 98. 71 %的期刊检索时却都少敲了 1 键,相比之下取前 4 个首字母检索要优于取前 5 个首字母检索。由此可见,取前 4 个首字母检索刊名是最佳的选择。

在取前 4 个首字母时,占总测刊数 53. 01 %的期刊检索时一屏幕只显示 1 种,占总测刊数 95. 58 %的期刊检索时一屏幕只显示 1～8 种,占总测刊数 98. 71 %的期刊在 DOS 环境下和在 Windows 环境下检索都不需要翻屏,只有两类 54 种刊在 DOS 环境下检索需各翻一次屏,1 类 47 种刊在 DOS 环境下检索需翻两次屏。而在 Windows 环境下检索,只有两类 77 种刊需各翻 1 次屏,且很好记,1 个是代表刊为《中国医学》的 ZGYX 类,1 个是代表刊为《国外医学》的 GWYX 类,绝大部分为医学专业类系列刊。在 DOS 环境下检索需翻 1 次屏的期刊为 ZGSY 和 ZGYX 类,代表刊是《中国摄影》和《中国医学》;需翻两次屏的期刊是 GWYX 类,代表刊是《国外医学》。若不愿对这些期刊翻屏检查,只要记住这 3 类刊,随后再增加 1 个首字母,用 5 个首字母检索,就可大幅度减少重首字母期刊的显示,它们在 DOS 环境下和在 Windows 环境下检索就都不再需要翻屏,这些期刊的总量才占被测刊数的 1. 29 %。

被测期刊中 1～4 字刊分别为 3 种、364 种、236 种和 2718 种,首字母数分别为 3 个、728 个、708 个和 10872 个,共计 3321 种 12311 个首字母。其中,1～3 字刊共计 603 种,为了在检索 1～3 字刊时避免不必要的 3 字以上刊名的出现,每种期刊检索时其刊名首字母后需再补敲一空格键,603 种 1～3 字刊共需多敲 603 键,相当于增加了 603 个首字母,1～4 字刊的首字母计算数就成了 12914 个。5 字以上刊(含 5 字刊)为 4519 种,其前 4 个刊名首字母共计 18076 个。需翻屏检索的 101 种重首字母期刊,按最多的击键次数计算再补加 148 键,相当于再补加 148 个首字母。全部期刊的前 4 个首字母计算数为 31138 个,平均每种刊 3. 97 个首字母,平均不到每刊敲 4 键。而 7840 种期刊的全部首字母数是 45408 个,平均每种刊 5. 79 个首字母。这样,使用该法检索一种期刊要比全首字母检索平均少击 1. 82 键,相当于在原来通用的汉字全拼的基础上已经提高了近 3 倍的情况下又提高了 31. 43 %。如果 1～3 字刊检索时除输入刊名的全首字母外不愿再补敲那一多加的空格键也可,其速度会相对更快些,一屏幕显示完的还可少敲 1 键。一屏幕显示不完的,头屏最前边已显示出了要检索的期刊,第二屏不需再

●吕 娟 袁相琴

论第四种情报检索语言系统

摘 要 第四种情报检索语言是自然语言与人工语言结合的一体化语言。第四种情报检索语言系统是一种基于网络的信息检索系统,比分类主题一体化情报检索语言系统更高级更新颖,是我国 21 世纪情报检索语言系统研究的方向。加快我国第四种情报检索语言系统研究的关键,是解决汉语分词技术问题。参考文献 14。

关键词 情报检索语言 第四种情报检索语言系统 中国分类主题词表 一体化语言系统 词素

分类号 G354.2

ABSTRACT The fourth information retrieval language is a language integrating natural and artificial languages. The fourth information retrieval language system is a network-based information retrieval system, more advanced than the classification-subject integration information retrieval language system, and is the direction of the research of information research language system in China in the twenty-first century. In this paper, the authors think that the key to accelerate the research in China is to solve the problem of Chinese word segmentation. 14 refs.

KEY WORDS Information retrieval language. Fourth information retrieval language system. *Classified Chinese Thesaurus*. Integrated language system. Morpheme.

CLASS NUMBER G354.2

20 世纪,分类语言、主题语言和分类主题一体化语言被应用于传统情报检索系统(包括联机检索系统和光盘检索系统);21 世纪,自然语言与人工语言相结合并应用于网络环境。第三种情报检索语言即分类主题一体化语言,实质是人工语言内部的一体化[1]。笔者称自然语言与人工语言结合的一体化语言为第四种情报检索语言。

1 第四种情报检索语言产生原因

迅猛发展的全球计算机网络系统应用环境和多数据库资源共享体系的实现是第四种情报检索语言产生的前提,情报用户社会化使其蓬勃发展。

1.1 人工语言向易用方向发展的必然性

人工语言词汇量少,标引难度大,更新速度慢,词汇发展滞后,直接影响网络资源的快速高效流动。同时,在一个大的数据库中,每个主题词下都集中了数百数千条记录,要获得所需数据还得逐条搜索,费时费力。自然语言系统则不然,它专指性高,词汇量

看,可敲 Esc 键终止显示,击键次数相当于补敲 1 空格键,只是屏幕上显示出的不相关的期刊太多,显得杂乱。为了检索时逐屏显示,显示命令应使用 DISP 而不使用 LIST。

每个汉字的汉语拼音字母平均 3.2 个,最容易检的汉字,除打完拼音字母外,在 DOS 系统下的智能全拼至少要加击 1 选字键,而在 Windows 系统下的智能 ABC 输入法最少也要加击 2 键才能把字打到光标处。就是说,要把 1 个汉字打到光标处,不连换行选字的击键次数,最少也要平均击 4.2 或 5.2 键。而使用我们设计的这种检索方法检索一种刊名,把所需输入的首字母打到光标处还平均不到 4 键。可见,我们设计的这种检索刊名的方法,其速度比检索一个普通汉字的速度还要快。

结论:利用中文期刊前 4 个刊名汉语拼音首字母检索刊名,是计算机快速检索中文期刊刊名的一条捷径。

袁国发 副研究馆员。通讯地址:石家庄市裕华东路 186 号,河北科技大学图书馆。邮编 050018。

杨莉英 女,河北科技大学图书馆馆员。通讯地址同上。

杨健 河北科技大学图书馆助理馆员。通讯地址同上。

袁立军 石家庄乐仁堂医药经营公司工作。通讯地址:石家庄市中山东路 161 号。邮编 050011。

(来稿时间:2001-07-13)