四角号码在中文题名检索中的应用

曹福元 沈 鸣 (南京大学)

摘要：本文介绍了对题名汉字四角号码进行 4-X-X-X 截取存储，实现四角号码中文题名检索的方法，并就如何截取给出了程序段，从而使读者很方便地将这一算法嵌入自己的应用系统中。

中文题名检索在中文编目系统中是一个十分重要的环节，选择汉字检索时，必须键入汉字检索项，这就涉及到汉字的输入问题。目前汉字的输入方法很多，如拼音方式、五笔字型输入方式等，尽管它们已被广泛使用，并且有这样和那样的优点，但由于诸多因素，有时也难以掌握，尤其是对多年从事图书馆编目工作的人员来说，渴望能通过输入四角号码来进行题名检索，以解决输入汉字所带来的不便。

实现四角号码检索，首先须在书目文件中增加一个相应字段，用以存放与题名相对应的四角号码值，每一个汉字对应四位十进制数的四角号码值。问题是新增的四角号码字段中，以一种什么样的方式来存放与题名相对应的四角号码值。如果采用与题名逐字对应的方式来存放四角号码的话，固然查准率较高，但在题名检索时存在几个方面的问题：

1.击键次数过多；

2.操作人员必须准确无误地键入四位一组的四角号码值，否则难以满足要求，这就势必给操作人员带来困难；

3.新增的四角号码字段长度是题名字段长度的两倍，空间开销过大。

出于上述几点考虑，可以采取 4-X-X-X 截取法来解决。即题名首汉字取其四角号码全码，其后每个汉字取其指定位数的四角号码，至多取若干个汉字，以 4-1-1-1 截取法为例，具体截取方法见表 1。

采用 4-X-X-X 截取法，在进行题名检索时，具有重码率低，易于掌握，便于记忆，击键次数少，操作简便等优点，但是在处理一些使用频率过高的词汇或词组时，譬如“中国共产党”、“中华人民共和国”、“大学英语”、“政治经济学”、“计算机”等等。对这些词汇采用 4-X-X-X 截取法，就无法达到降低重码率的目的。因此，有必要对这一类词汇进行压缩处理，以做为 4-X-X-X 截取法的补充。具体解决办法是建立一个适当规模的常用词组表，如在 FoxBASE 关系数据库环境下，可建立一个二维数组，若使用 C 语言编程可考虑建立一个结构数组（见程序实例），用以存放常用词表，并且给表中的每一个常用词组分配一个适当的四位十进制数的特定值。这个特定值的选取，以不与其它汉字的四角号码重复为原则，如“中华人民共和国，9991”、“政治经济学，9992”，这类词组的选择应根据系统的特点、具体专业范围等加以选择。

表 1 4-1-1-1 截取法

题名字数	截取方法
>=4	4-1-1-1
=4	4-1-2
=2	4-3
=1	4-口-口-口

检索时，如果欲检索以常用词表中的词为首的检索项时，则选用其特定值外加其后若干个汉字的指定位数的四角号码值进行检索，如检索“中华人民共和国大事记”一书，可采用 9991453 码进行检索，以 4-1-1-1 截取法为例，其中 4、5、3 分别是“大”、“事”、“记”三个汉字四角号码的首位数，这种方法不但降低了重码率，而且大大提高了工作效率，如果再辅之以当前记录的前、后记录查询，系统将更加灵活、实用和方便。

实现四角号码的截取，必须建立一个单个汉字及其对应的四角号码的数据文件，该文件结构见表 2。

表 2　汉字四角号码文件结构

字符类型	字符和汉字	四角号码

实现利用四角号码的检索，从编程角度出发，重点在汉字的四角号码截取。一旦截取完毕存放在相应的字段中，检索就显得十分简单。尤其是在关系数据库管理系统环境下编程，如 Fox BASE、informix 等，直接利用关系数据库管理系统所提供的查询语句，即可完成前方一致或完全一致检索。

四角号码的截取可通过以下程序段来实现，这个程序是采用 C 语言在 Informix 关系数据库管理系统环境下开发的，在 C 语言中，它实际上是一个函数，调用时可通过参数的选择，确定截取对象、截取的汉字数、以及除首汉字外的汉字四角号码的截取位数来完成截取。

整个截取过程，由函数 Cut__title()来完成，该函数分别调用 serch__sx()，get__serkey()，serch__phase()三个函数。cut__title()有三个参数，s 为要处理的题名，hz__numb 为要截取汉字个数，cut__numb 为除首汉字外、其余汉字截取的四角号码位数。

首先，调用 serch__phase()函数对题名字符串进行扫描，与常用词表比较，如题名首部与常用词表中的词组匹配，则取其特定值，否则取检索词首汉字的四角号码全码，然后调用函数 serch__sx()取其后汉字的四角号码，由函数 get__serkey()完成四角号码的拼接。

这种方法在实际使用中不但简化了操作，同时提高了查准率。本文所提供的源程序，只要稍加修改，将其嵌入应用系统中(如编目子系统、流通子系统、期刊子系统)的数据增加、更新模块中，便可完成检索对象的四角号码的截取，以实现四角号码的检索。

程序清单：

```
#define ischinese(c)      (c&0200)
#define NKEY (sizeof(phase) / sizeof(struct PHASE))
#include "stdio.h"
#include "dbio.h"
struct dbview hasx_db_1[]={
{"hzsx_type"},
{"hzsx_hz"},
{"hzxs_sjno"}
};
struct
{
char hzsx_type[1+1];
char hzsx_hz[2+1];
char hzsx_sjno[4+1];
}hzsxk;
struct PHASE{char * P_str;
                    char * p_value;
                    };
static struct PHASEphase[]={
{"中华人民共和国","9991"},
{"政治经济学","9992"},
{"马克思主义","9993"},
{"高等数学","9994"},
{"社会主义","9995"},
{"中华民国","9996"},
{"大学英语","9997"},
{"外国文学","9998"},
{"中国","9999"}
};
main()        /*调用函数 cut_title()程序示例(简化)*/
{
int cc;
char charptr[20], * cut_title();
opendb();      /*打开 hzsx_db 文件*/
strcpy(charptr,cut_title("中华人民共和国大事记",4,1));
aspace(charptr,7);   /*aspace()--字符串加尾空函数*/
printf("四角号码截取值:%s\n",charptr);
closedb();      /*关闭 hasx_db 文件*/
return(0);
}
/*********cut_title()**********
  函　数 1:cut_title()
  功　能 :取正题名前若干个汉字,按要求拼接四角号码.
  返回值:拼接的四角号码检索值,(4-.....)
  说　明:S--正题名 hz_numb--取汉字数 cut_numb--除第一
          个汉字外,每个汉字所取的四角号码位数.
*********cut_title()**********/

char * cut_title(s,hz_numb, cut_numb)
int hz_numb, cut_numb;
char * S;
{
struct{char hz[3];
     char hz_sx[5];
     }k[8];
int i=0,j,flag,number,len1;
static char serkey[30],psx[5];
char sp[5];
char * p, * pstr;
```

```
char * get serkey(), * serch sx(), * serch phase();
for (pstr = s;!ischinese ( * pstr)&& * pstr;pstr++);
if(! * pstr)return(NULL);
flag = 0;
for(j = 0;j < hz numb;j++){
    k[j].hz[0] = '¥0';
    k[j].hz sx[0] = '¥0';
  }
strcpy (psx, serch phase (pstr, &len1));
if(! isnull (psx)){   / * 函数: isnull()--判段字符串是否为空 * /
    flag = 1;
    number = hz numb-1;
    p = pstr+len1;
    }
else{p = pstr;
    number = hz numb;
}
for(; * p&&i < number;p++)
  {
   if(ischinese( * P)){
    k[i].hz[0] = * p++;
    k[i],hz[1] = * p;
    k[i],hz[2] = '\0';
    strcpy(sp,serch sx(k[i].hz)); / * 返回汉字四角号码值. * /
    if(!isnull(sp))
      strcpy(k[i++].hz sx,sp);
    }
  }
if(i = = 0)return(flag?psx:NULL);
strcpy(serkey, get serkey(k,i,hz numb,cut numb,flag,psx));
return(serkey);
}
/ * 函数 2:serch sx() * /
char * serch sx(s)
char * s;
{
int cc,len;
cc = dbfind)"hzsx db:,EQUAL,s,&len, &hzsxk);
return(!cc&&!isnull(hzsxk.hzsx sjno)?hzsxk.hzsx sjno:NULL);
}
/ * * * * * * * * * get serkey() * * * * * * * * *
  函 数 3:get serkey()
  功  能:拼接四角号码.(4-.....)
  返回值:接拼的四角号码检索值.
  说  明:cut numb1--除第一个汉字外,每个汉字所取的四角号
          码位数.
          numb--实际取到的汉字数
          type = 1 时,rstr 存放 phase 中列出的特定号码
 * * * * * * * * * * get serkey() * * * * * * * * * * /
char * get serkey(ptr, numb, hz tnumb, cut numb1,type, rstr)
int numb, type,cut numb1,hz tnumb;
char * rstr;
struct {char f1[3];
      char f2[5];
      } * ptr;
{
int i,n1 = 4;
static char str[30];
if(type)strcpy(str,rstr);
    else{strcpy(str,ptr++->f2);
        numb--;
        }
for(i = 0;i < numb;i++)
  {
  strcpy(&str[n1],ptr++->f2);
  n1+= cut numb1;
  }
str[hz tnumb * cut numb1+4] = '\0';
/ * aspace(str, HZ NUMB+3); * /
return(str);
}
/ * * * * * * * * * serch phase() * * * * * * * * * *
  函 数 4:serch phase()
  功  能:正题名与常用词组表比较.
  返回值:若存在,则返回其表中定义的四角特定值;否则返回
          NULL.
 * * * * * * * * * serch phase() * * * * * * * * * /
char * serch phase(s,v)
char * s;
int * v;
{
int i;
for(i = 0;i < NKE¥;i++)
    {
      if(!strncmp(phase[i]. p str, s, strlen(phase[i].p str)))
          {
          * v = strlen(phase[i].p str);
          return(phase[i].p value);
          }
    }
return(NULL);
}
/ * 因篇幅有限,函数 aspace(),isnull(),opendb(),closedb()程序未
列出. * /
```