第 11卷 第 1期　　辽宁省交通高等专科学校学报　　Vol. 11 No. 1
2009年 2月　　JOURNAL OF LIAONING PROVINCIAL COLLEGE OF COMMUNICATIONS　　Feb. 2009

文章编号: 1008 - 3812(2009)01 - 043 - 02

一种实现汉字转拼音编码的方法

孙文毅
(辽宁省交通高等专科学校,辽宁沈阳　110122)

摘　要　介绍了一种实现汉字转拼音编码的方法。给出了建立码表文件、通过查码表进行转换和维护码表文件的方法以及使用 Delphi 7编写的主要程序段。

关键词　汉字　同音字　拼音编码　Delphi

中图分类号: TP311.11　　**文献标识码**: B

1　引言

由于汉字中有很多同音不同形的同音字,给对用汉字表示的数据进行排序和查找等操作带来了一些困难。如生成按读音排序的学生名条、按姓名查找学生等操作。如果将汉字转换成拼音编码后按拼音编码进行排序和查找,这些困难就解决了。例如:在某个班中查找全部姓“YAN”的学生,我们知道“YAN”姓有“阎”、“闫”、“颜”、“言”、“延”和“严”六个字形,按汉字查找,需要查找六次才能完成。如果将输入的查找关键字和数据集中的学生姓名都转换成拼音编码,按拼音编码查找,就能一次将姓“YAN”的学生都查找出来。同样,将数据集中的学生姓名转换成拼音编码,再按拼音编码排序,这个班中姓“YAN”的学生,无论他的姓是六个字形中的哪一个,都会排在一起。

2　实现的方法

2.1　建立码表文件

笔者建立的是一种纯文本格式的码表文件,使用记事本就能建立。

码表文件的内容是将字典里的汉语拼音音节索引表(笔者使用的是新华字典,有的字典里这个表的名字是音节表)中的索引项(即拼音编码),按其在汉语拼音音节索引表中的顺序,每个索引项写在一行的行首上。然后,在字典里找出同音(不分声调)的汉字,将它们逐个地写在对应索引项的后面,索引项和汉字之间以及汉字和汉字之间用一个空格分隔。在最后那个汉字的后面加一个回车符。用于分隔的空格有两个作用:一是依据这个空格,可以很容易地将索引项从其所在的行中分离出来。二是这个空格和它后面的那个汉字组合在一起,构成了每一行中唯一的机内码组合,从而消除了比较时的误判断现象。

收稿日期: 2008 - 11 - 26

2.2　查表方法

笔者采用的查表方法是在查找关键字的前面加一个空格(与码表结构对应)后作为一个子串,使用查找子串的函数在码表的每一行上查找。在哪一行中查到了,就依据那一行的第一个空格,将行首的拼音编码分离出来,保存到存放查找结果的变量里。由于存在多音字,即使已查到了,也要继续在剩余的行中查找。

每一次查找的结果有三种可能:只查到一次、查到多次和没查到。对于这三种结果,笔者是这样处理的:只查到一次,是一个单音字,直接返回查找到的拼音编码。查到多次,是一个多音字,将查到的多个拼音编码显示出来,由操作者选择,然后返回操作者选择的拼音编码。没查到,是因为码表中没有这个字,将这个字显示出来,提示操作者需要维护码表文件。

2.3　维护码表文件

发现不能转换的汉字和转换的拼音编码不正确的汉字后,需要对码表进行维护。

在确定了汉字的正确读音后,使用记事本打开码表文件,如果是不能转换的汉字,将其添加到对应的行里。如果是转换的拼音编码不正确的汉字,将其移动到正确的行里。

3　使用 Delphi 7编写的主要程序段

下面的代码中,TForm2是一个在程序启动时建立的子窗体,mbStr是在这个窗体的 private段里声明的一个用于存放码表数据的 TStringList类型的变量。pymListBox是这个窗体上的一个 TListBox组件,用于存放查找的结果。button1是这个窗体上用于关闭模式窗体的按钮。

```
procedure TForm2.FormCreate(Sender: TObject);
begin
  mbStr:=TStringList.Create;
  mbStr.LoadFromFile('mb.txt');//读入码表
end;

procedure TForm2.pymListBox1Click(Sender: TObject);
begin
//如果是多音字,当操作者选择了一个读音后
button1.Enabled:=pymListBox1.ItemIndex>-1;
end;
```

下面的这段代码,是实现汉字转拼音编码的函数,待转换的汉字由形式参数 str传入函数中,函数的返回值是转换的结果。

```
function TForm2.hz2pym(const str:string):string;
var
  i:integer;
  Str1:string;
begin
  pymListBox1.Items.Clear;
  Str1:=''+str;
  for i:=0 to mbStr.Count-1 do
begin
  if pos(Str1,mbStr[i])>0 then
  pymListBox1.Items.Add(copy(mbStr[i],1,pos('',mbStr[i])-1));
  end;
//是多音字
  if pymListBox1.Items.Count>1 then
begin
  hzLabel1.Caption:=Str;
  button1.Enabled:=false;
  ShowModal;
  Result:=pymListBox1.items[pymListBox1.ItemIndex];
end else
begin
  if pymListBox1.Items.Count=1 then
  Result:=pymListBox1.Items[0];
end else
//没有查到对应的拼音编码
begin
  hzLabel1.Caption:=Str;
  button1.Enabled:=true;
  ShowModal;
Result:=str;
end;
end;
```

4 结束语

笔者在编写这个程序时,查阅了一些相关的资料,了解到两种实现的方法。一种是对汉字的内码进行运算,然后用运算的结果去查码表得到对应的拼音编码。测试后发现,使用这种方法,多个读音的汉字,只能转换其一个读音的拼音编码,而且还有一些汉字不能转换,如"谌"字。由于不清楚其码表是如何构建的,很难对码表进行维护。第二种方法是使用微软提供的 API函数,从提供了逆向搜索功能的拼音输入法的码表中查出拼音编码。这种方法测试没有成功。但是它使用了微软提供的 API函数,决定了它只能应用在微软的 Windows平台上。而且也同样不清楚拼音输入法的码表是如何构建的。

本文给出的这种方法,采用了纯文本格式的码表文件,所以建立和维护码表文件非常容易。多音字和不能转换的汉字的问题,通过维护码表文件很容易就解决了。没有使用操作系统提供的 API函数,向其它操作系统平台上移植也很容易。

参考文献

[1]新华字典(1990年重排本).商务印书馆.

A Method to Translate Chinese Cheracter into Pinyin Code

Sun Wenyi

〔**Abstract**〕this article introduces one kind of method that translate the chinese character to the Pinyin encoded. Some method are given, such as to create code table, to translate by code table look-up, to service code table and The main procedure segment by Delphi 7.

〔**Keywords**〕Chinese character; Homophone; Pinyin code; Delphi