语音模糊查询在信息管理系统中的实现

The Realization of Phonetic Fuzzy Query in Information Management System

阎红灿　刘保相　李丽红　（河北理工大学 理学院 河北唐山 063009）

摘要：本文详细讨论了信息管理系统中实现语音模糊查询的三个关键技术：获得汉字拼音数据字典，创建高效安全的汉字拼音数据库，设计和实现可靠的同音字检索算法。并给出了基于VB环境具体实现的核心代码。

关键词：语音模糊查询 汉字拼音数据库 同音字字符串

1 引言

在开发各类信息管理系统中，灵活快捷的信息查询无疑是开发人员和用户所追求的主要目标之一，在实际开发过程中，查询功能一般都是通过对字符进行比较、判断等方法来实现的，因此一般易于实现精确的汉字信息查询。如查找一个名叫“李明”的人，数据库中存在的叫“黎明”或叫“李铭”的数据就不能检索到，也就是说，采用常用的字符比较法，实现不了同音字的查询功能。然而在很多数据信息管理系统中，或者网络信息搜索中，需要具有新的查询方式，即只要知道某一信息的读音，并不知道每个字的具体写法，通过检索数据库，就能把所有符合这个读音的记录内容全部显示出来，即语音模糊查询技术。如输入“新郎”，应该检索出“新浪”等数据记录。

如果在数据信息查询时具有同音字查询功能，不仅方便用户使用，还会大大提高检索效率。据统计，汉语单字同音现象是非常严重的。以常用6763个汉字为例，没有同音字的汉字只有16个，其他汉字都有同音字，其中最多的有116个同音字。为了解决同音字的模糊查询问题，笔者借用WINDOWS系统下的输入法生成器，生成了一个拼音查询字典库，在VB环境下成功开发了《民用信息公用查询系统》，实现了按语音进行模糊查询的功能。

2 实现语音模糊查询的关键技术

实现关键字的语音模糊查询，首先就要找出每个汉字的同音字，其中拼音码最为关键。所以创建汉字拼音数据库成为关键技术的第一步；有了拼音字典库，如何在开发环境中实现可靠稳定的高效检索便是下一个解决问题，所以数据库的格式及安全性显得非常重要；通过汉字拼音数据库获得同音字字符串后，如何在大量数据中检索出与关键字读音相同的记录，即设计一个高效的查找算法成为关键技术之三。

2.1 汉字拼音数据库

WINDOWS系统提供了输入法生成器IMEGEN.EXE，直接运行便进入输入法生成器窗口，如图1所示。鼠标单击“逆转换”的页框，点击“打开文件”按钮，选中WINDOWS\SYSTEM文件夹下的WINPY.MB文件，（如果是WINGDOWS XP系统，WINPY.MB文件在SYSTEM32文件夹）在码表原文件中输入C:\WINPY.TXT，单击“逆转换”，此时系统对全拼字典库进行转换，最后将生成一个纯文本文件WINPY.TXT，利用这个纯文本文件即可生成一个拼音字典查询数据库。

由于VFP系统提供了将文本文件转换为数据库文件的功能，而且对数据记录的操作方便快捷，所以借助于VFP系统编制一段程序生成汉字拼音数据库。命令文件MyChange.PRG的代码如下：

```
CREA TABL B1 (NR C(60),HZ C(2),PY1 C(12),
PY2 C(12)) &&创建一个临时数据库
USE B1 &&打开生成的数据库
APPE FROM WINPY.TXT SDF &&将利用输入法生成器生成的字典码文件WINPY.TXT内容追加到数据库中
DELE FOR ASC(SUBS(NR,3,1)) > =128 &&在数据库中删除全部词组内容，只留下单字
```

DELE FOR RECN( ) < 13 &&在数据库删除编码库的头文件

PACK &&清除打了删除标记的记录。

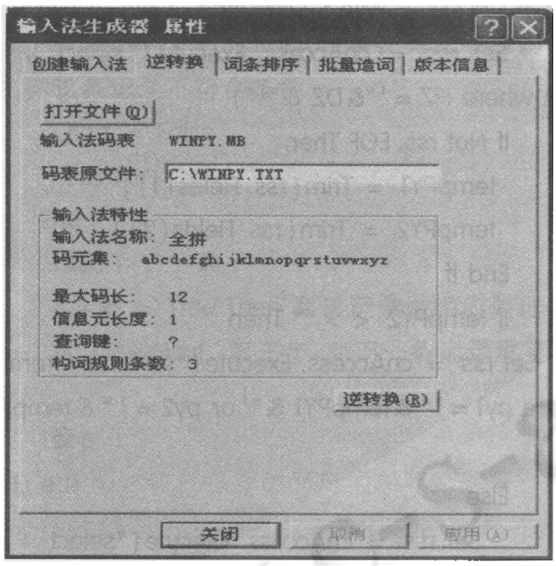

图 1　输入法生成器

REPL HZ WITH SUBS (NR,1,2),PY1 WITH SUBS (NR,3,AT('',NR) -2),PY2 WITH ;

SUBS(NR,AT('',NR) +1) ALL &&将汉字与拼音存放在不同的字段里,拼音有两个字段,其中有一个为多音字。

&&为了照顾南方口音的人员使用,可将全部 zh,ch,sh 替换成 z,c,s。为此需增加两个字段 NR1 和 NR2

&&REPL NR1 WITH 's' + SUBS(PY1,3) FOR " sh" $ PY1

&&REPL NR1 WITH 'c' + SUBS(PY1,3) FOR " ch" $ PY1

&&REPL NR1 WITH 'z' + SUBS(PY1,3) FOR " zh" $ PY1

&&REPL NR2 WITH 's' + SUBS (PY2,3) FOR " sh" $ PY2

&&REPL NR2 WITH 'c' + SUBS (PY2,3) FOR " ch" $ PY2

&&REPL NR2 WITH 'z' + SUBS (PY2,3) FOR " zh" $ PY2

COPY TO PYZDK FIEL HZ,PY1,PY2 &&生成一个拼音查询数据库

USE &&关闭打开的数据库

ERASE B1. DBF &&删除生成的临过数据库

在 VFP 中执行上面这段程序,系统将自动生成一个拼音查询库,并将这个数据库表命名为 PYZDK. DBF,其中收录了 27954 个单字。

2.2　拼音数据库的高效性和安全性

VFP 数据库在 VFP 系统中使用方便,但是其他开发平台调用必须借助于 ODBC 数据源,这就为二次开发和用户的使用带来了不便,况且 VFP 数据库的安全性也没有保障。为了提高汉字拼音数据库的检索效率和安全性能,有必要将其转换成 Access 数据库,即便于二次开发调用,又可以实现加密保护。笔者在 VB 环境中实现了将 VFP 数据格式转换成 Access 加密数据库格式,部分核心代码如下:

Public Sub Vfp_DataLink( ) '建立 VFP 数据库连接的过程

Dim str As String ' 连接字符串,Provider 提供驱动,Data Source 指出 ODBC 数据源

str = " Provider = MSDASQL. 1; Persist Security Info = False; Data Source = PYVFP"

cnVFP. Open str 'cnVFP 为原始数据库的连接

End Sub

Public Sub Access_DataLink( ) '建立密码保护 Access 数据库连接的过程

str = App. Path

If Right(str, 1) < > " " Then str = str + " "

Rem 连接字符串中 OLE DB 提供者必须使用 Microsoft. Jet. OLEDB. 4. 0,密码通过 Jet OLEDB 驱动,然后通过 Database Password 项设置密码

str = " Provider = Microsoft. Jet. OLEDB. 4. 0; Persist Security Info = False; Data Source = " &str &

" PYDB. mdb; Jet OLEDB: Database Password = 2592457"

cnAccess. Open str '连接数据库 PYDB. mdb,密码为 2592457,cnAccess 为原始数据库的连接

End Sub

Private Sub Command1_Click( ) '数据库格式转换,DBF 格式转换为 MDB 格式

Set rsVFP = cnVFP. Execute (" select * from PYZDK") 'rsVFP 为原始数据库的记录集

rsVFP. MoveFirst

rsAccess. CursorLocation = adUseClient

Rem 打开 PYDB. mdb 库中的 PYM 数据表,其中

adOpenKeyset(键集游标)和 adLockPessimistic (悲观锁)为 CursorType 和 LockType 属性,以便更新数据。

```
rsAccess.Open "PYM", cnAccess, adOpenKeyset, adLockPessimistic 'PYM为拼音码数据表
Do While Not rsVFP.EOF
  rsAccess.AddNew 'rsAccess为原始数据库的记录集
  For i = 0 To 2
    rsAccess.Fields(i) = rsVFP.Fields(i)
  Next i
  rsVFP.MoveNext
Loop
rsAccess.Update '确认更新
End Sub
```

过程 Vfp_DataLink( ) 和 Access_DataLink( ) 调用后,单击 Command1 按钮即可完成加密保护的拼音数据库 PYDB.mdb 的创建。这里 Access 数据库 PYDB.mdb 的空表 PYM 已经提前创建,并添加了密码保护。

2.3 设计可靠的查询算法

在系统数据库中对关键字进行同音字模糊查询的算法分三步:

第一步:根据要查询的关键字(如姓名 strName),首先在生成的汉字拼音数据库中找出第一个汉字(单字 DZ)在字典库中的位置,利用这个汉字的读音(tempPY1 和 tempPY2),对字典库记录内容进行过滤,通过循环,将全部与这个汉字同音的字相加生成一个字符串送到一个变量中(同音字数组 TYZ(i))。继续对另外输入的汉字进行同样的处理,最后根据输入汉字的多少,生成 n 个同音字字符串变量。

第二步:开始检索待查询的系统数据库,将数据库中查询字段拆开分别与生成的同音字字符串进行判断,只有当数据库中该字段全部字符都能在相应的字符串中查找到,那么这条记录就符合查找的条件,打上一个标识,或者根据需要作其他处理。

第三步:比较下一条记录,循环到数据库的结尾,这时就可以将所有做了查询标记的数据库内容显示出来,即实现了按语音模糊查询的方法。

笔者在 VB 环境下成功实现了以上算法,现将部分代码提供给读者。第一步的核心代码如下:

```
ReDim TYZ(n) '存放每一个待检索汉字的所有同音字,初值为空
For i = 1 To n 'n为关键字汉字个数
  DZ = Mid$(strName, i, 1) '如果一个汉字2个字节,则 DZ = Mid$(strName, i*2-1, 2)
  Set rss = cnAccess.Execute("select * from PYM where HZ = '"&DZ &"'")
  If Not rss.EOF Then
    tempPY1 = Trim(rss.Fields(1))
    tempPY2 = Trim(rss.Fields(2))
  End If
  If tempPY2 <> "" Then
    Set rss = cnAccess.Execute("select * from pym where py1 = '" &tempPY1 &"' or py2 = '" &tempPY2 &"'")
  Else
    Set rss = cnAccess.Execute("select * from pym where py1 = '" &tempPY1 &"'")
  End If
  If Not rss.EOF Then
    rss.MoveFirst
    While Not rss.EOF
      TYZ(i) = TYZ(i) &Trim(rss.Fields(0)) '构成同音字字符串
      rss.MoveNext
    Wend
  End If
Next I
```

第二步和第三步的核心代码如下:

```
Set rs = cn.Execute("select * from worker") 'worker为待查询的数据表
If Not rs.EOF Then
  rs.MoveFirst
  Do While Not rs.EOF
  flag = True '标志是否匹配同音字
  For i = 1 To n
DZ = Mid$(rs.Fields("name"), i, 1) '如果一个汉字2个字节,则 DZ = Mid$(rs.Fields("name"), i*2-1, 2)
    For j = 1 To Len(TYZ(i))
      If DZ = Mid$(TYZ(i), j, 1) Then '该字在同
```

```
音字字符串中找到则退出本循环
            Exit For
          End If
      Next j
      If j > Len(TYZ(i)) Then 该字在同音字符串中不存在,则该条记录不符合条件
          flag = False
          Exit For
        End If
       Next i
       If flag = True Then 对该记录做相应处理
       rs.MoveNext '第三步,对下一条记录重复第二步
     Loop
   End If
```

3 结束语

本文系统介绍了在 VB 环境中,实现按语音模糊查询的三个关键技术,读者根据以上实现语音查询的原理,在实际工作中可以针对具体需要实现不同形式的语音模糊查询方式,比如关键字直接输入汉语全拼,或者继续变更拼音数据库,构造同音字数据库等等,都可以在此基础上稍加修改便可实现,并希望通过提供的代码实例,为广大计算机同行起一个抛砖引玉的作用,让我们开发的程序更加便捷、好用。

参考文献

1 李东、张湘辉,汉语分词在中文软件中的广泛应用[J/OL],微软中国研究开发中心,2004。

2 夏昆,在 Delphi 中用拼音首字符序列来实现检索功能[R],乌鲁木齐:中国人民银行乌鲁木齐中心支行。

3 阎红灿,VB6.0 调用密码保护的 Access 2000 数据库[J],计算机应用与软件。

4 宋伟、吴建国, Visual Basic 6.0 高级编程[M], 北京 清华大学出版社,1999.4。

5 成功、杨佃福, VC 中几种数据库访问技术的比较与选择[J], 计算机应用研究,2000.2。