用汉语拼音检索数据库中的中文信息

王克宇,莫祥银,王 伟

(南京师范大学 分析测试中心,江苏 南京 210097)

[摘要] 介绍了利用汉语拼音实现数据库中文信息检索的技术,提供了该技术在Delphi语言下的应用实例.

[关键词] 拼音检索,辅助字段,拼音首字母,词条翻译,汉字库

[中图分类号]TP317.2, [文献标识码]B, [文章编号]1672-1292-(2004)03-0076-03

0 引言

在传统的数据库中文信息检索过程中,需要用户在检索框里输入被检索的中文词条,然后再点击相应的按钮启动检索.在检索框里输入中文词条时,其汉字输入速度因所选的汉字输入法而异.对于速度慢的微软全拼输入法来说,平均每输入一个汉字需击键5次;对于速度快的王码五笔字型输入法来说,平均每输入一个汉字仍需击键2次.

从本质上看,任何汉字输入法的输入过程本身就是一种检索过程,所以传统的中文信息检索方式是分两步完成的.首先,必须先通过某种汉字输入法从计算机操作系统的汉字库中检索出若干个汉字,在数据库的检索框中将其拼接成中文词条,然后再进行所需要的数据检索.

本文提出的数据库中文信息检索技术,可在数据库的检索框中直接输入汉字的拼音首字母,每个汉字仅需击键1次.由于采用的是英文输入法,所以在中英文混合信息检索中不必像传统的信息检索方式那样需频繁地切换输入法.本技术实现了直接检索,其检索速度是传统检索方式的2~5倍.

1 拼音检索的实现

本文以图书检索为例,为简单起见,仅对书名进行检索.图1显示了本例程序的运行情况.本例的3个可视控件中,用于数据表中字段book编辑的是控件Edit1,用于数据表中字段Sec_book的检索的是控件Edit2,用于显示数据表的内容以及检索结果的是控件Dbgrid.

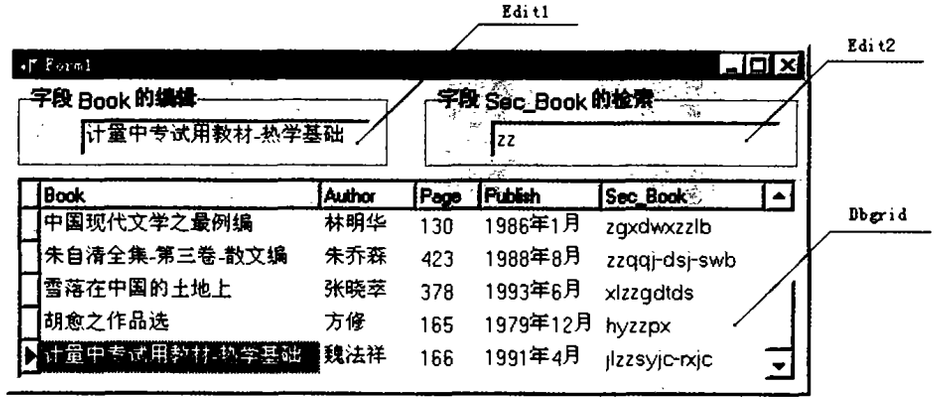

图1 图书检索的例子

图1中,检索框控件Edit2里只输入了"zz"两个字母,已检索出的图书名称中分别包含"之最"、"朱自"、"在中"、"之作"、"中专"等词条,它们的拼音首字母都满足检索条件"zz".若继续在检索框里添加字母,就可很快地定位到被检索的图书.

本例中还有3个非可视控件,分别是Query、Table和DataSource,它们用于连接、访问数据库.在检索过程中,Dbgrid经DataSource和Query联接数据表文件Book.db.由图1可见,Book.db中描述书名的字段是Book,而实际被检索的则是相应的辅助字段Sec_Book.辅助字段中存有的内容是由字段Book中的内容预先翻译而来的,其中的字符取

收稿日期:2003-08-20.

作者简介:王克宇(1954-),工程师,主要从事数据库的教学与研究.E-mail:wyky0619@sohu.com

自于字段 Book 中各个汉字的汉语拼音首字母,或者是其余非汉字字符.实际应用中,辅助字段 Sec_Book 是隐藏的,不必显示.

利用控件 Edit2 的 OnChange 事件句柄进行查询的代码如下:

```
procedure TForm1.Edit2Change(Sender: TObject);
begin
  query.close;
  query.sql.clear;
  query.sql.add('select * from book where sec_book like ("%' + edit2.text + '%")');
  query.open;
  datasource.dataset := query;
end;
```

2 词条翻译

由上述可见,拼音检索是基于辅助字段中预先翻译好的内容实现的.为了充分提高检索速度,我们把词条翻译安排在慢速的字段 book 编辑过程中.编辑框控件 Edit1 用于字段 book 编辑,每当用户通过 Edit1 输入及编辑数据表的字段时,启动词条翻译,及时将用户编辑的字段中的词条逐字翻译成拼音首字母,且存入对应辅助字段中以备检索.

在编辑字段时,Dbgrid 经 DataSource 和 Table 联接被访问的表 Book.db.Query 则联接自建的汉字库 hzk.db(见图 2).Hzk.db 中的字段 Word 存储单汉字,字段 Letter 存储其拼音首字母.词条翻译的过程就是逐个挑出字段 Book 中的汉字,通过 hzk.db查找对应的拼音首字母,再顺序填写到字段 Sec_Book 中去.

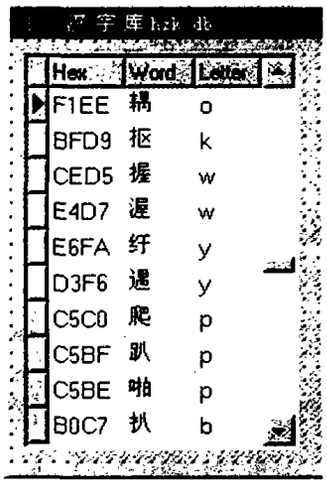

图 2 自建的汉字库

利用 Table 的 AfterScroll 事件句柄使 Edit1 跟踪显示 Dbgrid 的当前记录的 Book 字段:

```
procedure Tform1.TableAfterScroll(DataSet: TdataSet);
begin
  edit1.Text := table.fieldbyname('book').asstring;
end;
```

利用控件 Edit1 的 OnChange 事件句柄编辑字段 Book,且同时启动了词条翻译:

```
procedure Tform1.Edit1Change(Sender: Tobject);
var
  n:integer;
  s1,s2,s3:string;
begin
  table.Edit;      //table 的当前记录进入编辑状态
  table.fieldbyname('book').asstring := edit1.Text;
  table.post;      //Edit1 的内容写入 table 当前记录的 Book 字段
  s1 := edit1.Text;      //把当前记录的 Book 字段中的词条读入 s1
  n := 1;      //指向第一个字符
  s3 := '';      //清空准备写入 Sec_Book 的内容
  while n <= length(s1) do      //循环直到每个字符处理完毕
  begin
    if byte(s1[n]) > 127 then
      begin      //汉字处理
        s2 := s1[n] + s1[n+1];      //读入汉字 s2
        n := n + 2;      //跳过一个字符
        query.Close;
        query.sql.Clear;
        query.sql.Add('select * from hzk where Word like "' + s2 + '"');//通过 hzk.db 查询对应的拼音首字母
        query.Open;
        s3 := s3 + query.fieldbyname('Letter').asstring;//添加到 s3 中
      end
    else
      begin      //非汉字字符处理
        s3 := s3 + s1[n];      //添加到 s3 中
        n := n + 1;      //指向下一个字符
      end;
  end;
  table.Edit;      //s3 写入辅助字段
  table.fieldbyname('sec_book').asstring := s3;
  table.Post;
  datasource.dataset := table;
end;
```

3 建立汉字库

为了实现词条翻译,必须拥有图 2 所示的汉字库,手工建立该汉字库的工作量很大,且完备性及正确性都难以保证.我们通过编程从微软全拼输入

法字库中提取出6763个国标一、二级简化字,方法如下:

(1) 首先利用操作系统提供的输入法生成器将微软全拼的码表文件 winpy.mb 逆转换成码表原文件 winpy.txt(其转换方法参见输入法生成器的帮助说明).用微软的写字板观看该文件的情况见图3所示.码表原文件的头部是说明部分,正表部分将近56000行的内容中,每行由汉字及对应全拼音组成.码表原文件有如下特点:有的行描述单个汉字,有的行描述词组;有简体字,也有繁体字;描述单个汉字的行中,凡是多音字,都由以空格隔开的两个拼音描述,且常用读音置于后面.

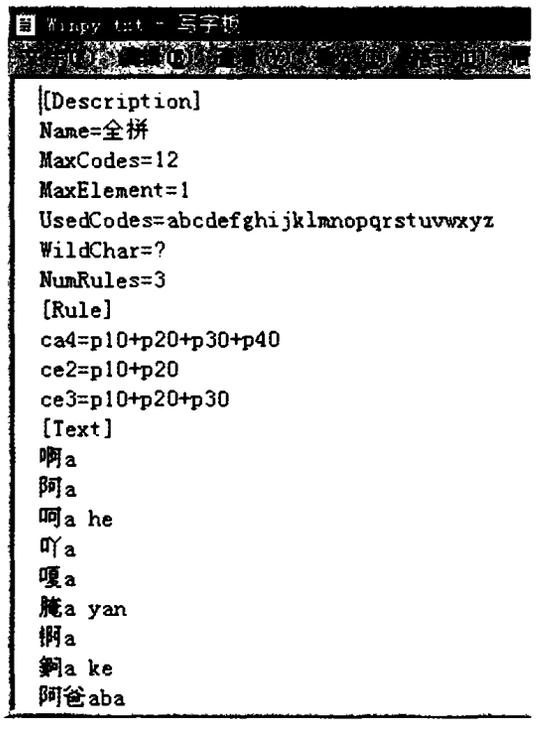

图3 微软全拼音输入法的码表原文件

(2) 编制程序,从 winpy.txt 中提取国标一、二级简化字,写入标准数据库的表文件 hzk.db 中.程序流程见图4.

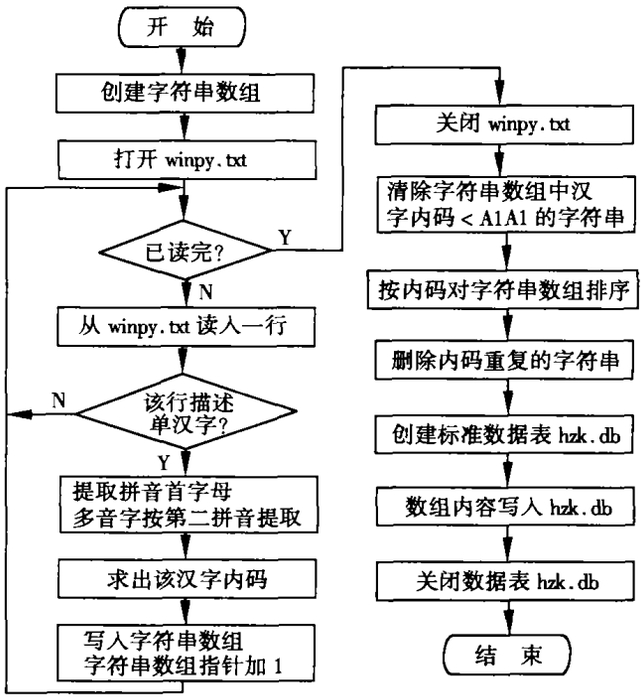

图4 提取国标一、二级简化字的程序流程

4 结束语

本技术已成功地应用于化工原材料信息检索以及医院挂号、站台售票等需要快速检索的数据库系统中.本文的叙述中,采用是 Delphi 语言,其原理也可在其它语言平台下实现.

[参考文献]

[1] 张大年,廖智勇,刘剑锋. Delphi 数据库应用开发技术与实例[M]. 北京:清华大学出版社,2002.

[2] 蒋方帅. Delphi 程序员指南[M]. 北京:人民邮电出版社,2000.

Phonetic Retrieval for Chinese Information in Database

WANG Keyu, MO Xiangyin, WANG Wei

(Analysis and Testing Center and Materials Science Laboratory, Nanjing Normal University, Nanjing 210097, China)

Abstract: This paper introduces the technology of phonetic retrieval for Chinese information in Database. An actual example in Delphi for applying this technology is given.

Key words: phonetic retrieval, subsidiary field, first letter of phonation, word translation, Chinese character library

[责任编辑:刘健]