文章编号:1008-0570(2007)01-3-0255-03

中文数据排序与快速检索方法研究

Study on the Method of Alphabetization and Quick Search

(河北建筑工程学院)孙宏凯　王彦勋
SUN HONGKAI　WANG YANXUN

摘要:通过对 GBK 编码、全拼输入法、常用汉字拼音和 SQL Server 2000 排序音等的研究,制定出适用的选取汉字拼音的方案,进而编程获得汉字拼音,以实现中文信息的快速检索。

关键词:汉字排序;汉语拼音;输入法;快速检索
中图分类号:TP311　　**文献标识码:**A

Abstract: On code of GBK,full-spelling input method,spelling of Chinese characters in common use and spelling coming from the order by SQL Server 2000,formulate the applicable scheme to obtain the spelling of Chinese characters. In addition, program it to carry out the Quick Search on Chinese information.

Keyword:Alphabetization,Chinese Spelling,Input method,Quick Search

技术创新

1 引言

在很多应用程序中需要对汉字字串进行排序,如人事管理中的职工姓名、供应或销售管理中的供货单位与客户等。同时为了对这些信息进行快速检索,通常将汉字的拼音首字母生成检索码。如,"河北建筑工程学院"的检索码为"hbjy",即由前三个字与最后一个字的拼音码首字母构成。然而在程序开发语言中要很好地完成上述功能,并没有直接有效的办法。尤其的一些英文版的尚不支持 GBK 或 Unicode 码的编程语言,如 PowerBuilder9.0(以下简称 PB9)及以下版本。典型地,在 PB9 的数据窗口中,用 Sort 函数对中文字段"姓名"进行升序排序,结果如表 1 所示:

表 1 PB9 中用 Sort 函数依"姓名"升序排序结果

编号	姓名	首字拼音	首字区位码
0904	翟晓松	zhai	2152
0901	董宁	dong	2213
0905	郝东	hao	2634
0903	郝玢	hao	2634
0902	褚玉梅	chu	8150

事实上,这些编程语言中,汉字的排序是按区位码进行的,二级汉字总体排在一级汉字之后。特别地,"翟"按另一读音"di"存放在一级字库中。

表 2 支持 GBK 的软件依"姓名"升序排序结果

编号	姓名	拼音	首字区位码
0902	褚玉梅	chu	8150
0901	董宁	dong	2213
0905	郝东	hao dong	2634
0903	郝玢	hao bin	2634
0904	翟晓松	zhai	2152

如果使用 Excel、SQL Server2000、PowerBuilder10.0 等支持 GBK 或 Unicode 码的软件对上述信息进行排序 (拼音字母序),结果见表 2:

其中"玢"的常用音是"bin",且多用于人名,而排序时按发音"fen"排在"东"之后。

为了解决这种问题,同时方便快速检索,通常将汉字按拼音首字母进行编码(如前所述),自定义排序规则。问题是:①怎样得到汉字的拼音(或首字母)?②多音字问题如何解决?③排序音与常用字的常用音不同如何处理。

2 汉字编码与拼音方案的研究

GB2312-80《信息交换用汉字编码字符集》基本集,第 16-55 区:按照汉语拼音的顺序依次存放了 3755 个一级汉字(最常用的汉字);第 56-87 区:按照部首顺序依次存放了 3008 个二级汉字(次常用的汉字);一、二级汉字合计共 6763 个。

GBK 亦采用双字节表示,总体编码范围为 8140-FEFE,首字节在 81-FE 之间,尾字节在 40-FE 之间,剔除 xx7F 一条线。总计 23940 个码位,共收入 21886 个汉字和图形符号 (包括 GB 2312 汉字 6763 个,按原顺序排列。),其中汉字(包括部首和构件)21003 个,图形符号 883 个。

Windows 操作系统中的全拼输入法支持 GBK,共收录 20902 个不同汉字单字,并给出多音字的全部拼音码。其中一、二级字库中多音字 2144 个(一级字 1235 个,二级字 909 个);非一二级汉字共 14139 个,其中多音字 3392 个。

通过输入法提供的系统函数(API)可以编写代码反查汉字的"全拼输入法"拼音码(实现方案见后),但对多音字只能得到一个拼音码 (多个发音按字母序排列后的第一个,下称"得到音")。将全拼输入法提供的 20902 个不同汉字单字用支持 GBK 的软件排序,可得到其"排序音"。此外,《字符集和信息编码国家标准汇编》中给出了部分常用字的"常用音"。需要指出的是,这三种读音之间存在一定的差异。

(1)对于 GB2312-803 中的 6763 个汉字 (即一二级字库,包括部首或构件),《字符集和信息编码国家标准汇编》对其中的 3 个单音字"盱、虔、筠"和 37 个部首或构件没有注音;补充上这

孙宏凯:副教授　硕士

3个字的读音,并称为“常用音”。此外,“斜”的常用音为“tou”,根据《新华字典》的注音及该字的解释,实用软件编写中修改为“dou”,并达到三种拼音一致。

(2)按“全拼输入法”,一、二级字库中多音字 2114 个(不包括 37 个部首或构件,其中一级字 1238 个,二级字 876 个)。

(3)一、二级汉字(去掉 37 个部首或构件)中:

①得到音与排序音不同汉字个数:完整拼音 1117 个(不包括 37 个部首或构件),其中一级字 674 个,二级字 443 个;拼音首字母不同的有 843 个,其中一级字 512 个,二级字 331 个。

②常用音与排序音不同汉字个数:完整拼音 76 个(不包括 37 个部首或构件),其中一级字 22 个(见表 3),二级字 54 个;拼音首字母不同的有 46 个,其中一级字 6 个,二级字 40 个。

表 3　一级字中常用音与排序音不同汉字

汉字	常用音	排序音	汉字	常用音	排序音	汉字	常用音	排序音
耙	ba	pa	嚼	jiao	jue	能	neng	nen
剥	bao	bo	桔	jie	ju	蔫	nian	yan
薄	bao	bo	咯	ka	lo	属	shu	zhu
称	cheng	chen	潦	liao	lao	嗽	sou	shuo
幢	chuang	zhuang	滤	lv	lu	她	ta	jie
闯	chuang	chen	氓	mang	meng	嗅	xiu	xu
囱	cong	chuang	呐	na	ne			
都	du	dou	内	nei	na			

其中首字母不同的有 6 个字:“耙幢咯蔫属她”。

③得到音与常用音不同的 1079 (一二级分别为 671 和 408)个,而首字母不同的 808(一二级分别为 508 和 300)个;部首或构件得到音不是“pianpang”的共 19 个,分别为:“丨匚刂冂勹氵宀阝阝凵艹亠辶彐屮灬尹钅扩”。

需要说明的是,得到音与常用音虽相同,但仍可能与排序音不同。共有 37 个,如“薄”字,得到音、常用音、排序音依次为:bao、bao、bo;当然,也有三种音均不相同的汉字,如“她”字,三种音依次为:chi、ta、jie。

(4)全拼输入法提供的非一二级汉字共 14139 个,其中多音字 3392 个,单音字 10747 个;多音字中有 1822 个得到音与排序音不同,而其中 1503 首字母不同。

3　获得汉字拼音的方案

将全拼输入法提供的全部汉字及拼音存储于数据库,并添加常用音和排序音,使用时逐个检索便可解决问题。但这种方法效率较低,尤其是在后台数据库管理的情况。理想的方法是,通过输入法先获得拼音,再按要求进行个别校正。

(1)用于生成全拼音:

先用程序获取“得到音”,对一二级汉字,并校正不同于“常用音”的部分字;部首或构件中“得到音”不是“pianpang”的,用“pianpang”替换;非一二级汉字不作修正而直接使用“得到音”。

(2)用于生成检索码或排序,仅考虑拼音的首字母:

先用程序获取“得到音”的首字母,对一二级汉字(不包括 37 个部首或构件),并校正不同于“常用音”首字母的部分字;部首或构件中“得到音”的首字母与“排序音”的首字母不同的,共 17 个,分别是“丨亻艹讠廴凵扌彡犭夂饣忄氵纟礻衤”,均用“排序音”的首字母替换;非一二级汉字,全用“排序音”首字母,同样地,替换“得到音”与“排序音”不同的部分。

4　程序实现与应用

GBK 编码中汉字及图形符号分类:

(1)汉字区。包括:① GB 2312 汉字区,即 GBK/2: B0A1-F7FE;②GB 13000.1 扩充汉字区,包括:GBK/3: 8140-A0FE 和 GBK/4: AA40-FEA0。

(2)图形符号区。包括:①GB 2312 非汉字符号区,即 GBK/1: A1A1-A9FE;②GB 13000.1 扩充非汉字区,即 GBK/5: A840-A9A0。

4.1 利用全拼输入法获得“拼音”

由于参数数值类型的匹配性、PB 的不同版本获取汉字编码的方式不同,直接用 PB 调用 API 函数来实现相对较困难,因而不妨用 C 语言自己编写动态库供 PB 调用。关键性代码如下:

```
#include "imm.h"
#pragma comment(lib,"imm32.lib")
```

①得到所采用的输入法的句柄,并激活该输入法,同时返回其句柄

函数:HKL chime(const char *imename)

主要利用 GetKeyboardLayoutList、ImmEscape、ActivateKeyboardLayout 三个函数

②获得拼音码,并返回其指针

函数:char *getcode(HKL pt ,const char *hz)

```
{ int   dwg;
tatic   char   buff[256];
char   *p;
dwg = ImmGetConversionList (pt,NULL,hz,NULL,0,GCL_REVERSECONVERSION);
if  (dwg<=0)      //是否可反查编码
return  NULL;
ImmGetConversionList(pt,NULL,hz,(PCANDIDATELIST) buff, dwg, \
GCL_REVERSECONVERSION);
p = (char *)buff;
p += buff[24];
return p;  }
```

③供外部调用的函数:返回拼音码的指针

函数:int WINAPI GetQpy(unsigned char *pch)

```
{    HKLpt;
HKLpa;
Char*p;
pa = GetKeyboardLayout(0);
pt = chime("全拼");
……
//标点、制表符等图形符号,返回“?”
//如果是汉字(按 GBK 编码中汉字及图形符号分类范围确定),返回获得的“全拼音”
p = getcode(pt,(char *)pch);
strcpy((char *)pch,p);
//英文、数字等键盘符号,照原字返回
……}
```

为方便调用程序使用及灵活性处理,函数仅处理单个字符或汉字。

4.2 PB 中自定义函数:获得字符串的拼音或首字母

外部函数声明:function int GetQpy(ref string ls_p) LIBRARY "Getpy.dll"

①获得字符串的拼音

ls_chars = "捭鞴臂……最蕞柞唑怍"　//得到音与常用音不同的一二级汉字(共 1079 个)

ls_chars_p = " 丨匚刂冂勹讠宀阝阝山艹宀辶彐屮灬聿乍扩"　//得到音不是"pianpang"(共 19 个)

ls_chars = ls_chars+ls_chars_p

ls_pym[] = {' bai' ,' bei' ,' bi' ,……,' pianpang' ,' pianpang' ,' pianpang' }　//用于校正的常用音

如果给定汉字在 ls_chars 中，在数组 ls_pym 中查找其拼音,否则用函数 GetQpy 获得。

②获得字符串的拼音首字母

ls_chars = "厂呆单……躜最蕞唑"//得到音与常用音首字母不同的一二级汉字(共 808 个)

ls_chars_p = "丶亻亠讠辶口扌彡犭夂饣忄刂氵纟礻礻"//得到音与排序音首字母不同(共 17 个)

ls_chars_3 = "稠鲋单……僔蒮捽椊"　//非一二级汉字得到音与排序音首字母不同(共 1503 个)

ls_chars = ls_chars+ls_chars_p +ls_chars_3

ls_pym[] = {' c' ,' d' ,' d' ,……,' z' ,' z' ,' z' ,' z' ,' z' }//用于校正的拼音首字母

同样,如果给定汉字在 ls_chars 中,在数组 ls_pym 中查找其拼音首字母,否则用函数 GetQpy 获得拼音,并取首字母。

4.3 应用

在 PB 中对汉字信息排序，可以依据检索码进行。上述方法在"中小型民营制造业企业管理与成本控制系统"软件中得到成功运用,该项目通过河北省科技厅组织的专家鉴定。

5 结束语

本文作者创新点:针对汉字多音及编程语言拼音排序中的弊端,结合汉字常用音,给出汉字拼音自动生成及排序的方法。软件设计上没有采用数据库搜索,而是利用输入法本身并加以修正,很大程度地提高了排序效率。

参考文献:

[1]常志玲,周庆敏等.笔顺输入法的汉字搜索算法研究[J].微计算机信息,2006,5-3:205-206.

[2]中国标准出版社.字符集和信息编码国家标准汇编[M].中国标准出版社,2004.

[3]本书编写组.Windows API 函数参考手册[M].人民邮电出版社,2002.

[4]郑阿奇.PowerBuilder 实用教程(第 2 版)[M].电子工业出版社,2004.

作者简介:孙宏凯(1964-),男(汉族),河北省张家口人,河北建筑工程学院数理系副教授，硕士，主要从事信息系统开发;E-mail:hongkaisun@163.com。

Biography: Sun Hongkai (1964-),male (the Han nationality), Zhangjiakou Hebei,associate professor at Department of Mathematics and Physics,Hebei Institute of Architectural Engineering, Master, Research area: information systems.

(075024 张家口市河北建筑工程学院数理系)孙宏凯　王彦勋

通讯地址:(075024 张家口市河北建筑工程学院数理系)孙宏凯

(收稿日期:2006.10.25)(修稿日期:2006.11.24)

(上接第 200 页)

图 4 GPS 信息显示

5 结束语

LBS 是地理信息系统、通讯技术和计算机技术的有机结合。它充分利用了无线移动的方便性、灵活性,也体现了大部分信息与位置有关的客观事实。基于位置的服务将成为人们日常生活中一种重要的信息服务，并成为未来信息服务业的重要组成部分。本论文提出的基于 Java 的 LBS 服务架构支持互操作、分布式运算,能满足移动用户任何时间、任何地点获取感兴趣的空间信息。其核心的服务具有平台独立性,可跨操作系统部署。

创新点:

1. 本文提出基于 GML 的服务器端多源数据融合技术和客户端地理信息服务的实现技术;

2.基于 Java 的 LBS 框架具有开放性,可扩展性,互操作性等特点;

3.本文提出的框架最终在智能手机上成功实现,实现了当前位置查询,地图漫游,位置查找等功能。

参考文献:

[1]Cheng Chengqi. Developing location-based services architecture integrated with GIS [J].International Conference on Computer Graphics and Spatial Information System, 2002.

[2]陈能成.2003. 基于 J2EE 的分布式地理信息服务研究.武汉大学博士论文.

[3]陈镧敏.1997.基于 GML 的空间信息共享及其在 WebGIS 服务器端的实现.北京大学硕士学位论文.

[4]卢军.J2ME 应用程序开发-手机、PDA 程序开发捷径.北京:中国铁道出版社,2002.

[5]张瑞江,齐华,韩卫杰等.基于 J2ME/Mobile SVG 移动 GIS 设计与实现[J].微计算机信息,2006,3:164-166.

作者简介:王盛校(1980-),男,汉族,中国测绘科学研究院地图学与地理信息工程专业 2003 级硕士研究生,主要研究方向为动态数据库和地理信息系统应用.E-mail: wsxiao@163.com;林宗坚(1942-),男,教授,博士,博士生导师,原中国测绘科学院院长,主要从事摄影测量与遥感图形图像学和地理信息系统等方面的研究工作。

Biography: Wang Shengxiao (1980-),Male,Han nationality,Postgraduate of Chinese Academy of Surveying and Mapping,Speciality:Cartography and GIS;Major in GIS Application study.E-mail: wsxiao@163.com.

(100039 北京 中国测绘科学研究院重点实验室)王盛校　唐新明

(430079 武汉 武汉大学测绘学院)范钊

通讯地址:(100039 北京海淀区北太平路 16 号中国测绘科学研究院重点实验室)王盛校

(收稿日期:2006.10.25)(修稿日期:2006.11.24)

技术创新