文章编号:1001-5132(2006)04-0430-05

3G终端拼音输入法字库搜索算法的设计与实现

屠秋萍[1], 曾兴斌[1,2], 何加铭[1,2]

(1.宁波大学 通信技术研究所, 浙江 宁波 315211; 2.宁波市通信芯片与射频技术重点实验室, 浙江 宁波 315040)

摘要: 手机现有拼音输入法在字库容量较大时, 输入效率比较低, 搜索算法和结构还不是很合理. 本文设计的编码采用了树型数据结构, 并根据此结构设计了1种优秀的搜索算法, 添加的辅助信息, 可提高用户的文本输入效率, 且所占空间少. 该新的拼音输入法已成功移植到3G终端开发板中.

关键词: 3G; 拼音输入法; 搜索算法; 树型结构; 嵌入式系统

中图分类号: TN916 **文献标识码**: A

随着3G和下一代网络NGN的应用正在喧嚣尘上, 短信市场又将面临崭新业务带来的良好发展机遇. 要想在手机上编辑文本, 输入方式至关重要. 一种合理、方便的输入方式将让用户轻松便捷地编辑文本, 进行交流. 这是扩展短信业务的一个十分重要的方面. 然而, 目前手机输入法存在严重的重码率和搜索算法效率不高、字库占用空间较大等问题, 使手机键盘的基本文本输入变成一个缓慢的过程. 为此, 本文提出了一种高搜索效率、低空间占用率、低复杂度的算法. 并采用MFC编写了可视化界面, 最终移植到3G终端开发板中, 完成对硬件嵌入式系统及其存储器件的研究.

1 设计方案确立[1,2]

在拼音输入法的设计中, 汉字编码系统的设计要做到编码简单、搜索快速. 下面对从搜索效果、字库的压缩以及设计的复杂度3个方面, 对3种常规方案进行比较.

1.1 方案1

方案1很直观, 直接对每一个汉字进行编码, 编码结构分为编码、地址2个部分, 如图1所示. 其编码形式是按每一汉字的按键顺序进行编码. 为了便于搜索, 此方案的编码采用定码长, 编码中的每一个数字必须至少4个bit存放, 所以整个编码大小为3个byte. 地址段用于存放汉字在字库中的地址, 如果字库为5 000个汉字, 则地址编码需要13 bit, 实际占用2 byte, 此时编码的字库大小为:

$$(3+2)\times 5\,000 = 25\ \text{KB}.$$

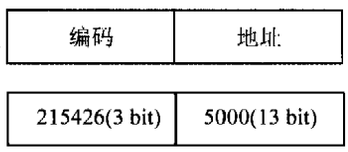

图1 方案1的编码结构

可见, 方案1编码方便简单, 但是占用空间比

收稿日期: 2006-09-13.
基金项目: 国家自然科学基金(60372026); 浙江省高校青年教师资助项目(2003643); 宁波市工业攻关项目(2003B10012); 宁波大学科研基金(Z0110014).
作者简介: 屠秋萍(1981-), 女, 浙江上虞人, 在读硕士研究生, 主要研究方向: 无线移动通信. E-mail:tuqp_0814@163.com

较大，效率不高.

1.2 方案 2

方案 2 在方案 1 的基础上作了一些改进. 由于有些汉字的拼音相同，即汉字的拼音编码有重码，因此不需要对每一个汉字进行编码，从而可节省存储空间. 改进后的编码结构如图 2 所示.

编码	地址	长度
215426(3 bit)	5000(13 bit)	125(7 bit)

图 2　方案 2 的编码结构

方案 2 没有对每一个汉字进行编码，而是将所有发音相同即编码相同的汉字编为 1 组. 在编码结构上相对于方案 1 增加了 1 个长度字段，用于表示同 1 组汉字的多少，其长度为 7 bit(这已经足够大了). 但是其编码字段的编码方式仍然和方案 1 相同. 仍假设对 5 000 汉字进行编码(仍然为 3 byte)，地址字段也一样(13 bit). 并假设所有汉字的编码有 500 个，那么该编码方案所占空间大小为：

$$(3+2+1)\times 500 = 3\ \text{KB}.$$

由上述分析可知，该方案所占空间大为减少. 但是此方案缺少搜索信息，搜索效率低. 同时，采用定码长的方法，很多汉字的编码很短就可以. 此方案的编码结构不利于提高搜索效率.

1.3 方案 3

针对以上 2 种方案在存储空间和搜索效率方面的缺陷，提出了第三种方案. 其编码结构如图 3 所示.

字母	地址	同级	有无下级	长度
21(6 bit)	5 000(13 bit)	125(7 bit)	0/1(1 bit)	125(7 bit)

图 3　方案 3 的编码结构

方案 3 最大的特点是对编码方式做了极大的改善. 此方案是对每一个汉字的汉语拼音的最后 1 个字母进行编码. 对 26 个字母进行编码，只需要 6 bit 就足够了，这是由于数据结构采用了树型结构的缘故. 有些汉字拼音不同，但其输入时按键顺序是相同的，在此称之为同级. 同时有些汉字拼音后再加 1 个字母即可变为另 1 个拼音，比如：can 和 cang. 因此在编码结构中还添加了同级字段和有无下级字段，其中在“有无下级”字段中 0 表示“有”，255 表示“无”. 这些辅助信息的存在，将大大提高搜索的效率. 其他字段与前 2 种方案相同，所有编码结构的集合称之为索引表. 可以算出其所占空间大小：

$$(1+2+2)\times 500 = 2.5\ \text{KB}.$$

显而易见，其所占空间比第二种方案还小.

从搜索时间和所占空间上可以看出，第三种方案的编码结构所占空间最小，而且其搜索信息全面，因此，本设计采用了第三种方案的编码结构.

2　系统总体框架描述

功能关系[3]如图 4 所示，输入模块可以说是 1 个接口，它一方面要与用户相联系，时时接收用户的输入. 另一方面，输入模块必须为后面的搜索做准备，起到 1 个预处理的作用. 比如当输入“2”键时，输入模块将其转化为“abc”3 个字母，并存储于数组中，以供搜索时使用. 由于存在同一编码不同发音的汉字的现象，这就需要一对上下按钮来让用户选择何种发音. 同时由于同一拼音的汉字可能很多，在一页中无法显示，因此需要一对左右按键，以供用户翻页查找. 为了让后面模块完成其功能，输入模块必须记录用户按下这 2 对按键的次数.

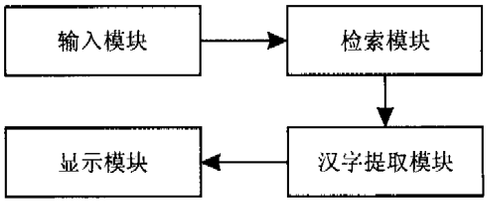

图 4　系统框架图

检索模块是整个系统的核心模块. 当用户输入时，检索模块根据输入模块提供的字母进行查找. 整个编码体系称之为索引，而检索模块的搜索也在索引中进行. 为了缩小搜索范围，字库内的汉字根

据汉字拼音第一个字母被归为 8 组. 如“abc”，“def”，“ghi”等，如图 5 所示.

字母	地址	同级	有无下级	长度
a	0	5	0	7
……				
i	1	2	255	13
……				
b	5	24	0	17
……				
c	21	255	0	3
……				
……				
d	……			
……				

放在同一个数组中

图 5 编码结构

汉字提取模块的功能[4]主要是根据检索模块提供的结果，在字库中提取汉字以供显示. 检索模块检索所得到的结果是当前输入字母在索引表中的地址. 需要根据编码结构中的地址字段来找到对应的地址. 显示模块也就是输出模块，本文重点在于搜索算法的研究，因此使用MFC编辑的可视化界面进行显示，需要显示的内容有 3 项：拼音、同一拼音的汉字列表以及正在编辑的文本. 在实际的手机系统中，汉字是经字模转换程序转换后，以 16×16 点阵形式存放的，所以本文最终在3G终端实现时，是液晶片通过点阵形式显示的.

3 系统软件设计[5]

一个合适的数据结构配合优秀的搜索算法，将极大提高搜索的效率.

3.1 数据结构

本设计中，采用了树形结构作为数据结构. 树形结构是一类重要的非线性数据结构. 其中以树和二叉树最为常用，直观看来，树是以分支关系定义的层次结构，其整体结构如图 6 所示.

图 6 是一个树形的层次结构，这里假设第一次按下的是“2”键，因而此结构的根尾“2”. 由于

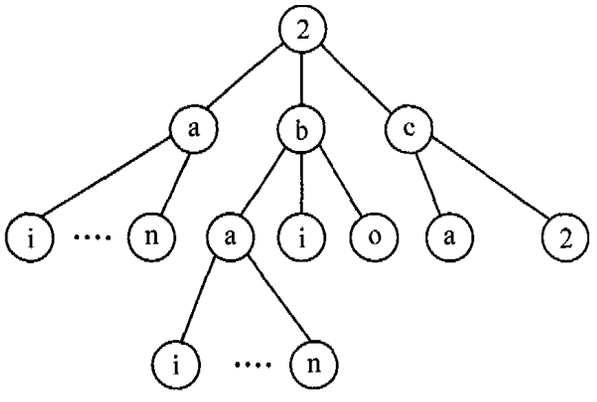

图 6 树形结构

“2”键对应的字母“abc”在汉语拼音中均可作第一个字母，因此其子树有 3 个. 同样下面的结构也是如此，都是本结点以及前面的父结点组合后可能的拼音，比如“ai”、“an”都是有可能的拼音，因此“i”，“n”作为“a”的子结点；而“bai”、“ban”也是成立的，所以在“b”后放一“a”结点，在“a”结点后放“i”和“n”作为“a”的子结点. 但是这样的关系结构在搜索时是无法实现的，因此实际各结点的关系如图 7 所示.

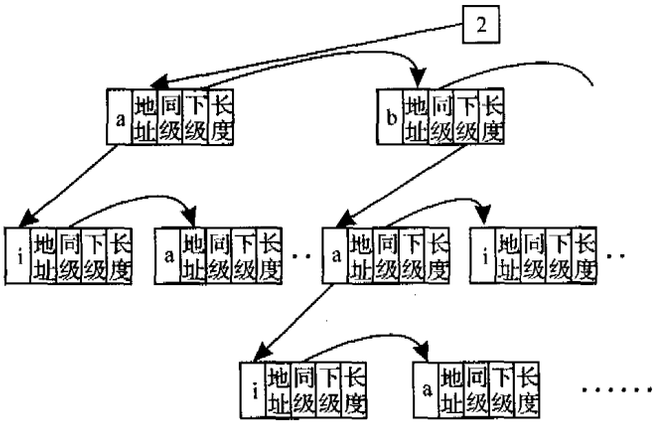

图 7 各结点的实际关系

根结点“2”只和其子结点“a”发生直接联系，而其子结点“b”只与结点“a”直接发生联系，其子结点“c”只与结点“b”直接发生联系. 结点“a”、“b”、“c”与其各子结点之间的关系亦是如此. 在这些联系中子结点与父结点的联系是通过各节点的“下级”字段来完成的. 如果本节点没有子结点了，就以 1 个特殊符号“255”表示. 而兄弟结点之间的联系，是通过各结点的“同级”字段来完成的，若本节点后再无兄弟结点，就以特殊符号“255”

表示.

3.2 搜索算法[6-8]

搜索过程可参照图 7. 其流程图如图 8 所示.

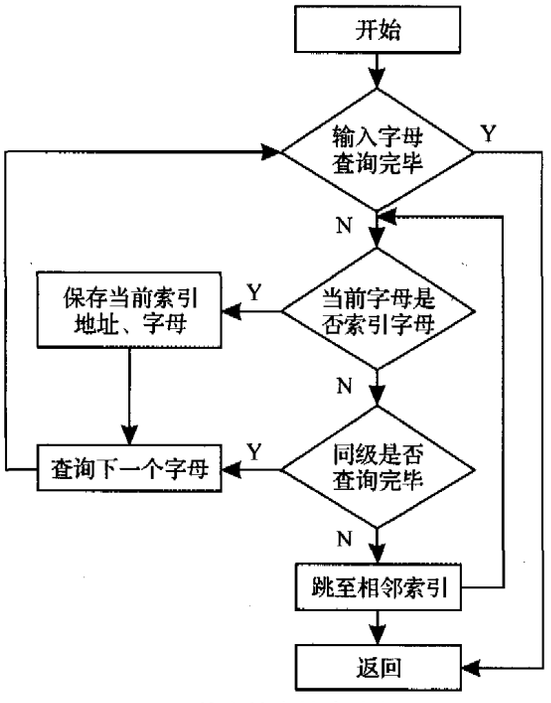

图 8　单个结点搜索流程图

以上讨论是对单个结点的子结点进行搜索的过程. 由于手机数字键盘输入时，一个按键对应多个字母，所以在输入时就有可能记录多个结点. 如果用户第一次按下“2”键，经按键与字符转换程序后，就必须对“a”、“b”、“c”查找. 显而易见，其结果必定是根结点的子结点“a”、“b”、“c”均符合要求. 故搜索程序记录了这 3 个结点. 若现在用户需要查找字母“a”，则必须分别从这 3 个结点出发来查询其各子结点.

3.3 系统可视化效果[9]

如图 9 所示，在本设计中用 MFC 中的按钮控件作了 18 个按钮.

图 9 中上下键：选择列表中拼音；左右键：对同一拼音的汉字序列翻页；“1”键：输入符号或用于确认时的选择；“2”到“9”键：用于输入拼音或确认时的选择；“确认”键：找到目标汉字后的确认输入；“取消”按钮：取消输入的拼音或删除输入文本.

下面就左右键作简单介绍，流程如图 10 所示.

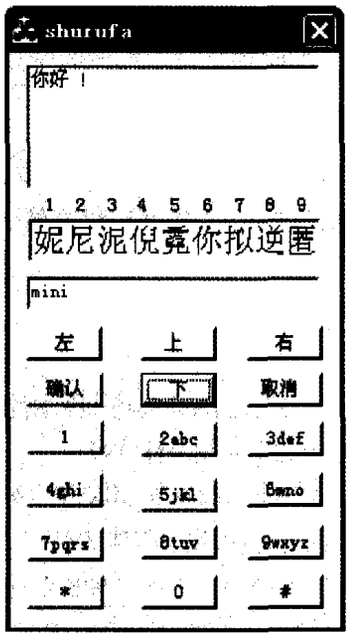

图 9　模拟界面图名

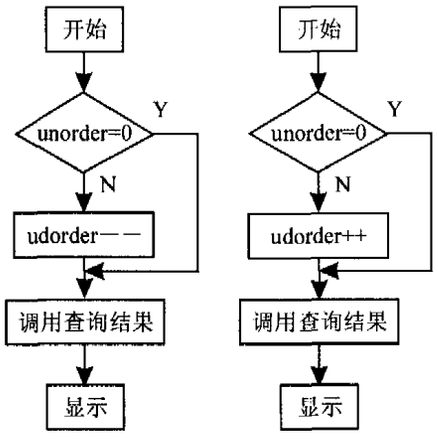

图 10　上下按键流程图

由于数字键盘中 1 个数字键对应多个字母，因此在用户输入的过程中必然会生成许多键盘按键方式相同，但拼音不同的汉字. 而用户需要的拼音又往往不是第一个，因此需要用 1 对左右键进行选择. 主要是使用了 1 个标志 UDorder，通过它对保存在数组中的查询结果进行选择.

4　硬件实现[10]

在本设计中，使用了 1 套 3G 终端开发系统. 整个系统以 ARM 作为中央控制器，以系统自带的键盘作为输入设备，以液晶屏(167 × 220)作为显示设备. 系统具有运算速度快、存储容量大、宽带传输

的特点，能够处理图像，文字等多种数据，完全符合3G终端的要求.

将编写的程序经过编译后，使用超级终端Hyper Terminal传输、加载到开发板，ARM系统接收到用户指令后可正确进行拼音输入，实现了本文的要求.

参考文献：

[1] 刘路放. C 语言的汉字处理与图文数据库技术[M]. 西安: 西安交通大学出版社, 1995.

[2] 严洪华, 强寿松, 倪旭东. 用C语言编制文字处理软件[M]. 北京: 人民邮电出版社, 1994.

[3] Min Lin, Andrew Sears. Chinese character entry for mobile phones: a longitudinal investigation[J]. Interacting with Computers, 2005, 17(2):121-146.

[4] Kevin Curran, Derek Woods, Barry O R. Investigating text input methods for mobile phones[J]. Telematics and Informatics, 2006, 23(1):1-21.

[5] 严蔚敏. 数据结构(C语言版)[M]. 北京: 清华大学出版社, 1996.

[6] Jiao C Z, Dai W H. Research on the procedure design of input method [J]. Transaction of Xianning Teacher-training Acad, 2001, 21(3):73-77.

[7] 焦翠珍, 戴文华. 输入法程序设计技术初探[J]. 咸宁师专学报, 2001, 21(3):73-77.

[8] 谭浩强. C 程序设计[M]. 2 版. 北京: 清华大学出版社, 1999.

[9] 李凤霞. Visual C++ 6.0 实用教程[M]. 北京: 电子工业出版社, 2001.

[10] 李驹光. ARM应用系统开发详解: 基于S3C4510B的系统设计[M]. 北京: 清华大学出版社, 2004.

Design and Realization of a New Words Database Searching Algorithm for Pinyin Input Method in 3G Terminal

TU Qiu-ping[1], ZENG Xing-bin[1,2], HE Jia-ming[1,2]

(1.Communication Technology Institute, Ningbo University, Ningbo 315211, China; 2.Ningbo Communication Chip and RF Technology Key Laboratory, Ningbo 315040, China)

Abstract: The current pinyin input method for mobile phone is not sound as expected. As a result, when database is large, the input efficiency would seem comparatively low. This paper presents a tree-like data structure and a corresponding searching algorithm. With some additional information for each word, the new search algorithm achieves a higher efficiency with less memory space. The new IME has been validated on PC and successfully transplanted to 3G terminal developing board.

Key words: 3G; pinyin input method; search algorithm; tree structure; embedded system

CLC number: TN916 **Document code:** A

（责任编辑 史小丽）