

哈萨克文手机输入法的研究

杨志敏,袁保社

(新疆大学 信息科学与工程学院,新疆 乌鲁木齐 830046)

摘 要:新疆是个多民族聚居的地区,但是支持哈萨克文信息处理的手机却一直都是市场的空白。通过研究哈萨克文手机输入法,哈萨克族用户可以很方便地操作手机的方寸键盘,快速、高效地输入文本信息,实现和家人、朋友的交流与沟通;同时这对发展少数民族地区通讯和经济也有着非常重要的意义。文中结合手机中多种文字输入的基本技术和方法,对哈萨克文手机输入法进行了研究。文章首先介绍了哈萨克语言的特点、手机输入法设计中的关键技术和根据哈萨克文的特征设计的哈萨克文手机键盘,接下来研究了词频的动态调整和词库的动态更新,并给出了实现其关键模块功能的程序流程图,最终实现了支持哈文和数字混合显示的智能手机输入法。

关键词:哈萨克文;智能输入法;手机键盘;Unicode 编码;词频

中图分类号:TP393

文献标识码:A

文章编号:1673-629X(2013)06-0151-04

doi:10.3969/j.issn.1673-629X.2013.06.039

Research on Input Method of Kazakh Language for Mobile Phone

YANG Zhi-min, YUAN Bao-she

(College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

Abstract: Xinjiang is a multi-ethnic region, however, its cellphone market has a vacancy in Kazakh information processing. Through developing Kazakh input method, Kazakh users could easily operate inch square key-boards of cellphones to input text information fast and efficiently, which also realizes communications with family members and friends; meanwhile, it's of significance to improve the communication and economy in ethnic regions. It includes research of input methods of the Kazakh mobile phone, combined with the basic techniques and methods of input of mobile phones in a variety of languages. First introduce features of written Kazakh, the key technology in the design of mobile phone input method, and the keyboard design for mobile phone suited to the characteristics of Kazakh language, then study the dynamic adjustment of word frequency and vocabulary updating, and give the flow chart of the program that realizes the key module function, finally achieve the intelligent input method for mobile phone which supports Kazakh language and digital display.

Key words: Kazakh language; intelligent input method; keyboard; Unicode encoding; word frequency

0 引言

随着中国通信业的发展,移动通信逐渐成为最为重要的通信业务,人们也越来越多地将目光从个人电脑转到更小、更快并极具个性化的移动智能设备上,智能手机的发展无疑代表了未来手机的发展趋势。而文字输入技术又是手机系统的重要组成部分,一直是信息处理的热点问题之一。特别是随着短信息应用的日益火爆,用户对快速文字输入技术的需求日趋强烈。我国是一个多民族的国家,随着社会的不断发展,用不同语言交流的人们对掌握各种语言的需求不断增长,尤其是在新疆地区和哈萨克斯坦之间。哈萨克文手机

输入法的研究,对发展我国手机产业,对促使少数民族与其国家的交流,加速少数民族地区科技教育事业的蓬勃发展,以及加快语言文字的信息化建设步伐是非常有必要的。

现有的哈萨克文字输入方法都是通过逐个字母来敲入单词的。由于哈萨克文字母多,在手机数字键盘上布局时出现一键多字母的情况,给用户操作造成非常的不方便。文中借鉴了手机中汉字、英文单词快速输入的一些原理,提出了一种手机中哈萨克文单词快速输入的方法,有效地提高了手机中哈萨克文的输入效率。

收稿日期:2012-09-09

修回日期:2012-12-10

网络出版时间:2013-03-05

基金项目:工信部2009年度电子信息产业发展基金项目(工信部财[2009]453)

作者简介:杨志敏(1984-),男,河北石家庄人,硕士研究生,研究方向为嵌入式系统;袁保社,教授,硕士生导师,研究方向为嵌入式系统、图像处理、电路理论、信号处理。

网络出版地址:<http://www.cnki.net/kcms/detail/61.1450.TP.20130305.0816.026.html>

1 哈萨克语的介绍

1.1 哈萨克字母的书写形式和书写规则

哈萨克民族语言属阿尔泰语系突厥语族,其语言按语言系属分类,属阿尔泰语系突厥语族克普恰克语支;按形态结构分类,属粘着语类型。哈萨克文共有 33 个字母,哈萨克语有 9 个元音音位,其中前元音 5 个,后元音 4 个;辅音音位有 24 个,其中清辅音有 10 个,浊辅音有 14 个。哈萨克字母是从右向左书写,词与词之间必须保留有一定的空隙^[1,2]。另外每一个哈萨克字母按照在单词中所处位置的不同,写法略有变化,有的字母有两种书写形式,有的字母有四种书写形式。字母的每种形式应该出现在词的哪个位置上,都有一定的规则。每个字母还有一定的规格,字母的大小有一定的比例,每个字母的基部都必须落在基点上。书写时先写字母的基部,然后点符号。

1.2 哈萨克语言的语音

哈萨克的音可以分为元音和辅音,哈萨克单词可由一个或几个音节构成,而一个音节则由一个或者几个语音组成。元音可以单独地构成音节,而辅音不能单独构成音节,原音哈萨克语言除了 33 个字母,还有一个特殊的软音符号“.”,在哈萨克单词中如出现 к,к',к' ,三个字母中的任何一个,整个单词的前面就不用标写软音符号,这样的词也按前元音读,如单词“ ومورگ ,”(在人生中)。在多音节结构中,只有几音节要读前元音时,软音符就标写在词首,其他地方不标写。

一般地,哈萨克单词中不会连续出现两个或两个以上相邻的元音字母。根据音节中元音和辅音结合的关系,哈萨克语的音节类型可以分为以下三种:

(1) 开音节:由一个元音构成或者辅音起首元音结尾的音节叫开音节。

(2) 促音节:由元音起首、辅音结尾的音节叫促音节。

(3) 闭音节:由辅音起首并由辅音结尾的音节叫闭音节。

1.3 哈萨克语言的语法

在哈萨克语中,一个词从结构上可分为词根、词干和附加成分(词缀和词尾)等部分。

词根是一个词中体现主要词汇意义的核心部分。词根具有明确的、独立的意义,能单独构成词。由词根单独构成的词叫根词。

词干是一个词中体现词汇意义的部分,一个词除了词尾所剩部分就是词干。词干=词根+词缀。当词缀为零(即没有词缀)时,词干就等于词根,这时的词干称之为零缀词干。

一个词中不体现具体的基本词汇意义,不能独立

成词的部分就是附加成分,附加成分包括词缀和词尾。附加成分与词根相对立,一个词中,词根以外的部分都是附加成分。词缀是缀接在词根或者词干上增添抽象的词汇意义从而构成新词的附加成分,它也叫构成附加成分。词缀按其在词中所处位置,通常分为前缀、中缀和后缀三种。词尾是缀接词干后表达种种语法意义的附加成分^[3-5]。

2 移动平台哈萨克文输入技术分析

哈萨克文输入跟英文输入有些类似,早期在移动平台上的文本输入法主要还是通过以下两种方法来实现:一是运用费兹定律(Fitts' Law)减少文本输入过程中按键移动的时间和距离;二是减少按键的次数,另外还有其它的一些输入法综合了上述这两种方法,下边这几个数字键盘文本输入法是当前性能比较好的。

Multitap:目前手机和 PDA 移动设备的文本输入普遍采用的都是 Multitap 输入方式。用户为了输入一个字母需要重复按下相应的数字按键,直至显示出欲输入的字母(如,按数字键‘5’4 次输入‘ н ’,按数字键‘8’3 次才能输入‘ ж ’,)。Multitap 输入方式最大的困难就是,当用户连续输入同一个数字按键上的字母就会存在歧义问题(亦即分割问题)。解决分割问题的方法有两个,使用计时器或者通过增加额外的功能键来移动光标。对于专业的人员来讲,虽然后者方法的输入速度比前者要快^[6],但目前大部分移动设备还都是采用的前一种方法。

Two-Key Input:顾名思义,此文本输入方式要求用户在输入任何一个字母时只需按键两次,首次按键是为了选择欲输入字母所在的数字键;第二次是根据该字母在数字键的位置^[5]选择相应的数字完成输入,例如要输入字母‘ б ’,则需要按键‘2’和‘3’完成输入。对比前一种方法,该方法就不存在分割的问题,但是该方法在输入速度方面要比 Multitap 慢^[7]。

上述这些在移动设备的按键上进行文本输入的方法和思想,对于设计哈萨克文智能输入法有很大的帮助,文中的研究和设计正是基于第二种方法。

3 哈萨克输入法的设计

3.1 哈萨克输入法的核心思想

当用户输入时,根据组成单词的首个字母对应按键及其在此按键上的顺序,需要按键两次完成该字母的输入,如要输入字母‘ и ’,则需要按键“4”和“3”,输入框里就会显示出该字母,候选框中便会显示词库中以此字母开头的所有候选字词集,并且这些候选词都是按词频从高到低依次显示的,输入单词的其他字母,

过程以此类推。随着用户键入字母个数的增多,其候选列表框所列举的单词就会越少,用户在输入的过程中可以使用上下选择键和左右翻页键进行选择。用户在使用 的过程中,每当输入某个单词一次,其词频就会自动加 1,若以后再次输入该词时,由于其词频较大,用户就能在候选列表中很快找到该词,从而减少了翻页次数,提高了输入效率^[8]。

为哈萨克词库生成了索引表文件,用户输入的手机按键编码,需要与索引区的关键字进行匹配,匹配成功后返回对应 在词库中的起始位置和哈萨克词语块的长度。索引表的结构如表 1 所示,Index 中存放的是手机按键编码的关键字,Start 表示哈萨克词库中数据块的起始位置,Length 对应数据块的长度^[9]。

表 1 索引表的结果

Index	Start	Length
-------	-------	--------

哈萨克文输入法的处理流程图如图 1 所示,该流程图描述了用户从开始击键到哈萨克单词上屏的系统设计与处理过程。

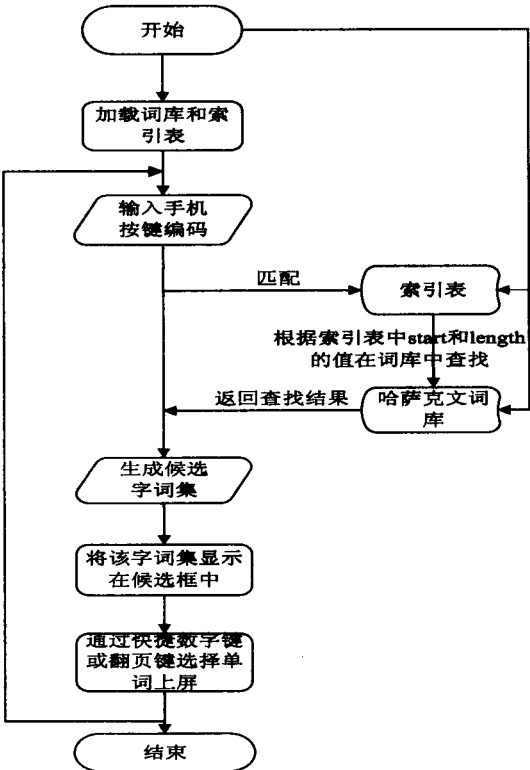


图 1 哈萨克文手机输入法处理流程图

3.2 界面设计

虽然说输入法的功能是程序设计中最核心的地方,但是友好的手机界面也是必不可少的。输入法系统的界面设计主要包括:输入法状态指示图标、输出框、输入框(输出框和输入框合称为显示窗口的编辑部分)和候选框(即显示窗口的候选部分)。

3.3 键盘布局

到目前为止,哈萨克语手机数字键盘布局还没有

一个统一的国家标准和行业标准,哈萨克语数字键盘布局如表 2 所示。该布局充分考虑到用户对市场现有 哈萨克手机键盘的使用习惯和哈萨克字母发音等因素,按哈萨克字母的自然排序,最终将 33 个哈萨克语字母以每个数字按键放置 4 个字母(其中 7、9 按键放置了 5 个字母)的方式进行布局^[10]。

表 2 哈萨克文手机数字键盘布局

数字键位	键盘哈萨克文字母				
2	А	Б	В	Г	
3	Д	Е	Ж	З	
4	И	Й	К	Л	
5	М	Н	О	П	
6	Қ	Ғ	Ҝ	Ҡ	
7	Җ	Ҙ	Һ	Ү	Ұ
8	Ҝ	Ҡ	Ң	Ҥ	
9	Җ	Ҙ	Һ	Ү	Ұ

现行的哈萨克文是以阿拉伯文字母为基础的拼音文字,哈萨克字母共有 33 个,其中有 9 个元音字母,24 个辅音。在 Unicode 代码标准中,将哈文字符的范围分为基本代码区和扩展区。基本代码区的范围是 0600 ~ 06FF,扩展区的范围是 FB50 ~ FDFF 和 FE70 ~ FEFF^[11]。

3.4 哈萨克文词库的数据结构

算法是由数据结构来实现,所以设计一个程序之前首先要设计程序实现中所使用的数据结构。数据结构是算法的基础,数据结构支持算法。良好的数据结构有利于算法的编写和执行。

智能输入法所使用的词库是将每一条词语按照手机数字按键编码和词频高低等规则组合而成,放置于一个文件中供输入法系统使用。表 3 所示为一个词条在计算机中的数据结构示意图。

表 3 哈萨克文词语的数据结构示意图

Unicode 十六进制编码	分隔符
----------------	-----

Unicode 十六进制编码:在计算机中,哈萨克文字符是由 Unicode 数字编码来表示的,诸如А, (0x0646), Б, (0x0649)等形式的编码,所以该字段的编码长度是不固定的,其长度取决于词语的长度。

分隔符:占用 1 个字母元素的空间,也称之为结束位,表示一个哈萨克词语的结束。

由以上的分析可得到存放哈萨克文词语的词库文

件的数据结构(见表 4):

表 4 哈萨克文词文件的数据结果

手机按键编码	哈萨克文单词	词频
--------	--------	----

3.5 词频的动态调整

为了使哈萨克文输入法更加智能、更加人性化,输入法系统需要动态地调整哈萨克词语的词频信息,用户在输入哈萨克文单词的过程中,如果欲输入的词语在候选框首页,则直接通过手机的上下方向键进行选择,否则的话,用户需要通过翻页键去查找某个哈萨克词语。用户输入后,将该哈萨克词语的词频增加 1,(然后将哈萨克语词库按词频从高到低重新排序)这个动态调整哈萨克词语词频的过程具有实时性,如图 2 所示。

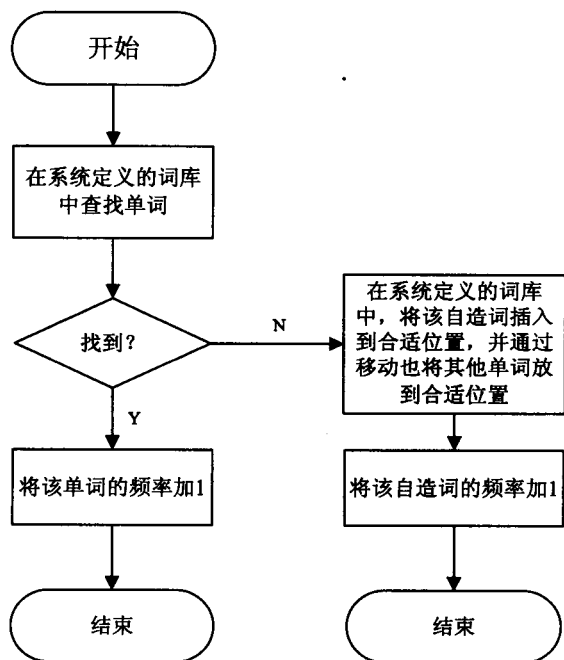


图 2 哈萨克文单词频率调整流程图

通过以上分析和研究,用户在使用本哈萨克文智能输入法时,只需按下哈萨克文字单词中每个字母所对的数字键,并根据字母在候选列表中的序号选上该哈萨克字母。在输入的过程中,系统会依据所输入的字母组合,动态地匹配出用户可能输入的一序列哈萨克文单词,并会显示在手机屏幕的候选框部分。候选框部分显示的完整哈萨克单词,都是按照单词的由短到长、词频的从高到低依次顺序显示的。

3.6 词库的动态更新

输入法系统设置了自造词库和普通词库两个字库,当用户在输入字母组合时,系统会优先去自造词库中匹配单词,若匹配成功,系统会将匹配成功的单词显示在候选框中。若在自造词库中匹配不成功的话,那就到系统定义的普通字库中去匹配。如果输入了全部的字母,但在候选部分没有相应的哈萨克词语,当用户把刚刚在编辑框自造的哈萨克词语输入到输出框时,

系统会自动将该哈萨克词语添加到自造词库中。另外,系统也会自动调整自造词的频率^[12]。


4 结束语

文中研究了哈萨克文手机输入法的关键技术,考虑了输入法中的联想性,即当用户输入单词时,不用全部输入单词的每一个字母,只需按下单词的前几个字母所对应的按键,该系统候选框中便列举出相关单词供用户选择,方便了用户输入,提高了输入效率。文中实现了哈萨克文和数字混合显示的功能,此方法可以直接用到柯尔克孜少数民族文字手机输入法上。另外,文中没有考虑智能语句级输入技术,这样可以进一步提高输入法的智能程度,当用户在输入一个单词以后,候选框除了提示一些以此单词开头的词语以外,还会提示某些以其开头的句子。

参考文献:

- [1] 王 花. 基于语料库的哈萨克文统计研究[D]. 乌鲁木齐: 新疆大学, 2010.
- [2] 伊力亚尔·加尔木哈买提. 哈萨克文语料库词汇校对研究[D]. 乌鲁木齐: 新疆大学, 2008.
- [3] 达吾勒·阿布都哈依尔, 古丽拉·阿东别克. 哈萨克语词法分析器的研究与实现[J]. 计算机工程与应用, 2008, 44(19): 146-149.
- [4] 杨振明. 哈萨克语[M]. 乌鲁木齐: 新疆人民出版社, 1983.
- [5] 耿世民. 现代哈萨克语语法[M]. 北京: 中央民族学院出版社, 1989.
- [6] MacKenzie I S, Kober H, Smith D, et al. LetterWise: Prefix-based Disambiguation for Mobile Text Input[C]//Proc. of ACM Symp. on User Interface Software and Technology. New York: ACM, 2001: 111-120.
- [7] Butts L, Cockburn A. An evaluation of mobile phone text input methods[C]//Proceedings of the Third Australasian Conference on User Interfaces. Melbourne, Victoria, Australia: [s. n.], 2002: 55-59.
- [8] Silverberg M, Mackenzie L S, Korhonen P. Predicting Text Entry Speed on Mobile Phones[C]//Proceedings of the ACM Conference on Human Factors in Computing Systems. New York: ACM, 2000: 9-16.
- [9] 王云琴, 袁保社. 基于嵌入式 Linux 和 Qtopia 平台维文输入法的实现[J]. 计算机应用与软件, 2011(9): 151-153.
- [10] 程新方, 吾守尔·斯拉木. 维吾尔语手机智能输入法的研究与实现[J]. 新疆大学学报(自然科学版), 2011, 27(1): 99-100.
- [11] 热依曼·吐尔逊, 吾守尔·斯拉木. 维吾尔文手机输入关键技术研究[J]. 中文信息学报, 2006, 20(2): 72-74.
- [12] 刘必强. 基于 Smartphone 的智能手机输入法的研究和实现[D]. 哈尔滨: 哈尔滨工业大学, 2006.

哈萨克文手机输入法的研究

作者： 杨志敏， 袁保社， [YANG Zhi-min](#), [YUAN Bao-she](#)
作者单位： [新疆大学信息科学与工程学院, 新疆乌鲁木齐, 830046](#)
刊名： [计算机技术与发展](#) 
英文刊名： [Computer Technology and Development](#)
年， 卷(期)： 2013, 23(6)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjfz201306039.aspx