文章编号:1008-0570(2006)05-3-0205-02

笔顺输入法的汉字搜索算法研究

Search Algorithm of Chinese Character Based on Strokes Input Approach

(南京工业大学)常志玲 周庆敏 王 雷 肖 乐
Chang,Zhiling Zhou,Qingmin Wang,Lei Xiao,Le

摘要:结合开发实际,介绍了笔顺输入法中汉字搜索算法和字码表及词码表的生成过程。从排序和查找两方面考虑,首先将汉字字库生成汉字字码表,然后将字码表根据首笔进行分区,当用户输入首笔后由汉字字码表索引文件决定在哪个分区范围内进行查找。实例证明本方法满足查找速度要求。

关键词:搜索算法;字码表;词码表;汉字输入法;
中图分类号:TP311 **文献标识码**:A

Abstract: This paper introduces search algorithm of Chinese character and forming process of the character code table and the word code table. The algorithm considers both sort and search, firstly Chinese character library is produced to Chinese character code table, then the table is divided into five regions. The index file of Chinese character code table determines which goal region to be searched when user puts into the first stroke. The example proved that the algorithm satisfied the requiring speed.

Key words: search algorithm; Chinese character table; word table; input approach of strokes;

1 引言

现在微机已经普及,汉字输入是计算机应用的关键技术,也成为汉字信息处理的瓶颈之一。因此汉字输入法的研究吸引了很多人,新的输入法也不断提出。输入法中常用的汉字编码主要分四类:形码(笔顺码)、音码、形音码和音形码。

我们常用的输入法五彩缤纷。如五笔字型输入法是按照汉字的形体结构进行编码,紫光拼音输入法和微软拼音输入法等是按照汉语拼音进行编码。无论哪一种输入法的字库至少有几千字甚至上万字,因此都涉及到汉字的搜索算法效率问题。本文以笔顺输入法为例详细介绍了汉字的搜索算法,由于本输入法是按照国家语言文字工作委员会语言文字规范所规定的汉字笔画中五个基本笔画进行输入的,因此从排序和查找两方面考虑,采用一种结构相对稳定的索引算法。以 GB2312 中的 6763 个汉字实验,首先将汉字字库生成汉字字码表,然后将字码表根据首笔进行分区,当用户输入首笔后由汉字字码表索引文件决定在哪个分区范围内进行查找。实践证明本方法满足查找速度要求。

2 汉字搜索的前期处理

2.1 输入法码本的形成过程

首先考察目标环境的可用汉字字符集标准,选用符合自己要求的字符集,其次按照编码法的要求对选定字符集的所有汉字进行输入码的编码、编码信息标定、汉字属性信息如汉字常用度的标定等,获得字码本。最后收集词组,并按照编码法所确定的词组的编码原则,根据字码本确定词组编码,对词组确定其常用度等属性,得到词组码本。字码本和词码本为后来的字库和词库搜索做好准备。

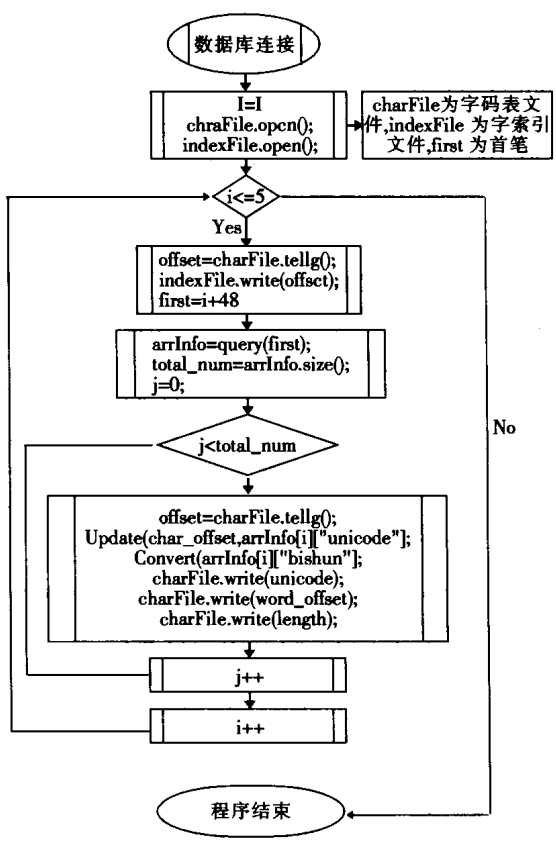

图 1 字码表创建流程图

常志玲:硕士研究生

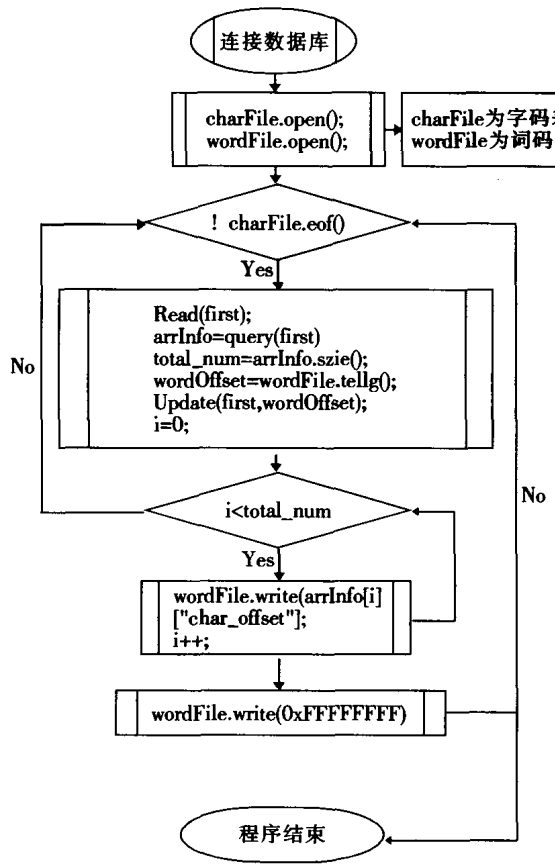

图 2　词码表创建流程图

2.2 汉字字符集的选择

从汉字编码规则知道,即使是汉字所在的字符集不同,对于某一固定的编码法相同形状的汉字应该具有相同的输入法编码。在 Windows98、NT 及以后的系统平台都完全支持 Unicode。Unicode 包含的汉字字符个数为 20902 个,远远多于其他汉字字符集。但 Unicode 在各个应用软件中并没有广泛的应用,因此在本输入法系统的底层,汉字采用 Unicode 内码,而在码本的前期处理时,我们选取 gb2312 字符集来进行编码。

2.3 字码表与词码表的格式和创建

2.3.1 字码表和词码表的格式

为了方便汉字的查找需要，我们创建汉字索引表,汉字按照笔顺的起笔将分为五个区,即“横(一)区、竖(丨)区、撇(丿)区、点(丶)区、折(乙)区”,字码表由两部分组成,字索引表和字码表。用 BNF 范式表示如下:

(1)<字索引表> ::= <分区字码表起始位置>[<分区字码表起始位置>]<字码表结束位置>

(2)<字码表> ::= <字码表项>[<字码表项>]

<字码表项> ::= <Unicode 值>(<词码表位置>|0xFFFFFFFF)

<笔顺长度><笔顺值>

(3)<词码表> ::= <词码表项>[<词码表项>]

<词码表项> ::= (<字码表位置>[<字码表位置>|0xFFFFFFFF])

2.3.2 字码表与词码表文件创建

在笔顺输入法中采用 GB13000.1 字符集汉字字序(笔画序)规范。《规范》中存在基础部件多,并且部件的文字定义也不够具体等难点,因此本汉字笔顺按照最基本的笔形共分五笔，用数字 1、2、3、4、5 进行标示,为了节省存储空间将五个笔画用三位二进制数进行表示,即 001、010、011、100、101。开始词码表没有创建,所以字码表项中词码表位置未知;同时词码表中字码表位置也是未知。所以先创建字码表,同时更新相应的字码表位置;再创建词码表,更新相应的词码表位置;最后重新创建字码表。创建字码表的流程图如图 1,创建词码表的流程图如图 2。

3 搜索算法

3.1 字的搜索算法

通过创建字码表，得到字码表中的汉字六千多个,因此选择一个高效的查找算法很重要。开始使用 STL 中的 MultiMap 结构,当字码表中有二十几个汉字时输入相应的笔顺要查找的汉字可以立即显示出来,但是当字码表中的汉字增加到上百个时就显示不出来了。经过仔细考虑,从排序和查找两方面考虑决定采用一种结构相对稳定的索引算法。用户输入笔顺的首笔后有可能通过翻页一直查找下去,用户不输入笔顺自然也就无需查找。考虑到首笔就要确定汉字,决定将汉字字码表根据首笔进行分区,即按“横、竖、撇、点、折”分成 5 个区,当用户输入首笔后由字码表索引文件决定在哪个分区范围内进行查找。其索引结构图如图 3。实践证明这样的搜索算法能够满足输入法下对查找速度的要求的。

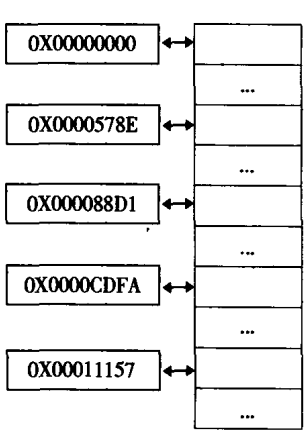

图 3　索引结构图

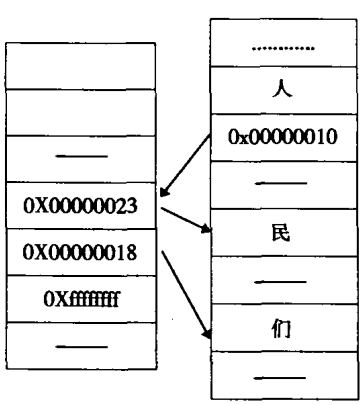

图 4　词搜索示意图

3.2 词的搜索算法

在汉字的字码表项中包含一个词码表位置项,如果在词码表中存在可与该字组成词的字,则该项的值为词码表相对位置的地址,否则为 0xFFFFFFFF。在词码表中存储的是字码表中的位置,返回字码表获得 Unicode 码和笔顺值进行比较。其搜索示意图如 4,可以组成词“人们”和“人民”。(转 193 页)

个码字。另一组文件出现的错误符号大于16个码字的情况更多,比例约占1/3。

实际上通过分析多条接收的卫星轨道数据,我们可以得出以下结论:

(1) 卫星接收仰角在5度以下时,干扰严重,存在大量的无法纠正CADU帧数据;

(2) 卫星数据在仰角5°到7°存在较大干扰,7°度以上则接收的数据质量变好。

(3) 在可统计的CADU中,每223+32 RS字节中错误符号在16个以下的帧与错误符号在16个以上帧的比例在4:1和8:1之间,由于EOS采用了RS编码并且交织深度I=4,增强了纠错能力,获得很好的数据接收质量。

所以当接收仰角在7°以下时,确实存在着严重的突发性信道衰落,导致大片码块出现误码。根据北京地面站接收EOS数据统计,错误中大约有1/8的错误帧存在错误RS符号大于16,如果采用交织深度I=1,就会有1/8的错误帧无法纠正(这里未统计多个出错的RS码字是否在同一帧CADU内的情况),将直接影响接收质量。

最后需要说明的是:对地面设备而言,取I=1和I=4,均可以考虑采用软件纠错或硬件纠错两种方式,通过更改设置是可以适应的,但必须相应地增加软、硬件设备。如果采用软件纠错,和硬件纠错相比,需要增加存储设备,并且需要多达10倍或更长的处理时间,这对气象卫星数据传输这种时效性要求较高的系统来说可能是难以忍受的。如果卫星采用编码采用I=1,则纠错能力比较差,在同样情况下达不到I=4的纠错效果。因此在气象卫星数据传输的RS编码方案中应尽量采用I取值比较高的编码方案,同时考虑和国际同类卫星的相互兼容,在当前技术条件下,选取交织深度I=4是比较好的选择。

5 结束语

本文作者的创新点是提供了一种方法,用真实的数据对卫星信道拟采用的编码方式在实际电磁环境下进行直观评估,进一步可直观进行气象卫星图像质量评估,具有较强的实际工程参考价值。

参考文献:

[1]Timothy Pratt, Charles Bostian, Jeremy Allnutt著,甘良才译,卫星通讯(第二版),电子工业出版社,2005.7

[2]Dennis Roddy著, 张更新等译,卫星通讯,人民邮电出版社,2002.5

[3]Consultative Committee for Space Data Systems, Recommendation for Space Data System Standards, Telemetry Channel Coding, CCSDS 101.0-B-6, BLUE BOOK, Oct.2002

[4]EOS维护文档资料,Seaspace Corporation,2003

[5]杨忠立; 王纪栋; 刘玉君著,基DSP的任意码长RS编码及算法优化[J]微计算机信息, 2005.16:104-105

作者介绍:娄志平,女,1970年生,北京大学物理学院大气科学系在读研究生 主要研究方向为卫星遥感与卫星通讯系统;李万彪 男,北京大学物理学院大气科学系,博士,研究方向为大气辐射、遥感、气象卫星;郎宏山 男,国家卫星气象中心 高级工程师 主要研究方向为极轨气象卫星系统设计;魏彩英 女,国家卫星气象中心 研究员,主要研究方向为气象卫星地面系统总体规划

Author brief introduction: Lou Zhiping, a graduate student of School of Physics, Peking University. Major research field: satellite remote sensing and communications systems

通讯地址:(100094 北京5113信箱)娄志平

(投稿日期:2005.9.13) (修稿日期:2005.10.15)

(接206页)

4 结束语

通过将字码表按照汉字首笔进行分区,把六千多汉字分成5个区,达到了化整为零的目的。然后在每个分区建立搜索索引文件,当用户输入首笔后由字码表索引文件决定在哪个分区内进行查找,大大提高了查找的效率。然后使用IME API(应用程序编程接口)实现输入法编程。实践证明,对于Unicode码中20902个汉字按照本算法搜索,也能够满足查找速度。因此本算法是可行的。

参考文献:

[1] 刘维光,陈立伟. 一种基于DHT的P2P搜索方法[J]微计算机信息,2005,3:131-133

[2]张小衡.《信息处理用GB13000.1字符集汉字部件规范》在输入法应用中的难点讨论[J].中文信息学报,2004(4):60-65.

[3] 侯捷. STL源码剖析[M]. 西安:华中科技大学出版社, 2003

[4]红梅. 基于Windows 2000/XP平台蒙古文输入法的设计技术[J]. 内蒙古师范大学学报,2005(1):40-43.

作者简介:常志玲(1976-),女,河南濮阳人,汉族,硕士研究生,主要研究方向为数据挖掘、粗糙集理论等。E-mail:czlhnpy@163.com;周庆敏(1963-),女,河北唐山人,汉族,教授,硕士生导师。主要研究方向为数据挖掘、粗糙集理论等。

Author brief introduction: CHANG Zhi-ling (1976-); Gender: female; Master; Major in Data minning and Rough Sets etc. ZHOU Qing-min (1963-) ; Gender: female; Professer at College of Information Science and Engineering, Nanjing University of Technology; Major in Data Minning and Rough Sets Rough.

(210009 江苏南京工业大学信息科学与工程学院)常志玲 周庆敏 王 雷 肖 乐

(College of Information Science and Engineering, Nanjing University of Technology, Nanjing 210009, China) Chang,Zhiling Zhou,Qingmin Wang,Lei Xiao,Le

通讯地址:(210009 南京工业大学丁家桥校区213信箱) 常志玲

(投稿日期:2005.9.6) (修稿日期:2005.10.16)