文章编号:1006-2475(2006)02-0088-03

汉语同音字和多音字处理方法研究

杨宪泽,谈文蓉,刘玉萍,张 楠,殷 锋

(西南民族大学计算机科学与技术学院,四川 成都 610041)

摘要:汉语同音字和多音字的存在给我国计算机应用增加了难度,本文分析了汉语关键词在计算机内存储、检索的过程,给出了同音检索算法。此外,还介绍了一种解决汉语系统中多音字判别和处理的方法。

关键词:同音字;多音字;检索算法;特征词典

中图分类号:TP391 文献标识码:A

Study on Processing Methods of Chinese Homonym and Polyphony Words

YANG Xian-ze, TAN Wen-rong, LIU Yu-ping, ZHANG Nan, YIN Feng

(College of Computer Science and Technology, Southwest University for Nationalities, Chengdu 610041, China)

Abstract: The existence of Chinese homonym and polyphony words gives our country the increased difficulty for applications of computer. This text analyzes the storing process and retrieval process of Chinese keywords in computers, giving homonym words retrieval algorithm. In addition, a kind of solveing polyphony words discretion and processed method in Chinese system, is also introduced.

Key words: homonym word; polyphony word; retrieval algorithm; feature lexicon

0 引 言

在我们研究的课题中,涉及到汉语同音字和多音字的处理。事实上,这两项处理的研究均具有重要的理论价值和应用价值[1]。例如,国内大量的管理信息系统,其检索方式通过汉语关键词来实现。即先按汉语关键词进行索引,然后通过定位函数确定信息记录地址。这一过程要求汉语关键词与索引内容一致。如果使用时在汉语关键词中误输同音字,就会导致检索失败。常见这样的例子,输入人名时,没有弄清每一个具体字,导致同音输入;有些字区别不大,把“检索”误输成“捡索”;由于定义不统一,把“电路节点”输成“电路结点”,把“存储器”输成“存贮器”等等;而多音字的处理在文字输入转换为语音输出的前沿研究中更显得重要。

统计资料表明,在 5.9 万汉语拼音词汇中,使用声调同音词占 9.6 %,不加声调时同音词达 27 %;而多音字出现的比例约占 24 %。这说明,同音问题和多音问题是干扰计算机有效应用的一大障碍[2~3],应该引起足够的重视。

本文设计了同音检索算法子模块,在检索方面力图解决同音问题。此外,还介绍了一种解决汉语系统中多音字判别和处理的方法。

1 同音字处理

1.1 计算机内中文存储方式及利用

在 MIS 中,计算机处理中文信息比处理西文信息难度大得多,其主要原因是:

(1) 中文是象形文字,字数多,字形复杂。西文是拼音文字,英文只有 26 个字母,加上大写小写及数字符号,总数不超过 128 个,用七位二进制码就可表达。而中文字成千上万,要用十几位二进制码才能把它们区别开来,这给存储乃至输入方式等都造成困难;

(2) 计算机内部只能处理二进制数。因此,汉语信息在计算机内部也要用二进制数表示.其字符集及其交换码标准根据 ASCII 码扩展而成,即把 94 个 ASCII 图形字符码中的任意两个加以组合代表一个汉字,总共可以表示 94 ×94 = 8836 个汉字。在同一系统中,ASCII 码和汉字代码之间的区别可以用特定的

收稿日期:2005-04-26

基金项目:教育部资助项目(0512226);西南民大重点项目(04NZ003)

作者简介:杨宪泽(1954-),男,四川成都人,西南民族大学计算机科学与技术学院教授,研究方向:自然语言处理和数据结构;谈文蓉(1968-),女,副教授,研究方向:系统结构和语音处理;刘玉萍(1956-),女,副教授,研究方向:自然语言处理;张楠(1972-),女,博士研究生;殷锋(1971-),男,博士研究生。

标识符,或用高位(第八位)是 0 或是 1 来区分。汉字系统用的控制功能码可以在国际标准字符集控制码基础上选用,若不够,可以用扩展符的方法加以扩充。即使如此,也可以看出汉语信息处理有不少西文信息处理所没有的额外课题。

汉字机内码是机器内部表示汉字的代码,是汉语系统体系结构设计的基础,也是同音检索算法实现的基础。汉字基本集标准 GB2312-80 包含一级汉字 3755 个,二级汉字 3008 个,各种图形符号 682 个。汉字按拼音字母顺序排列,同音字基本上在同一区中,少数跨越两区。每个汉字的机内码为两个字节,其高字节部分确定所在区号。我们构造了一个子模块,子模块中关键字内容以每一汉字所在区号进行索引,文献记录地址不变,见图 1。

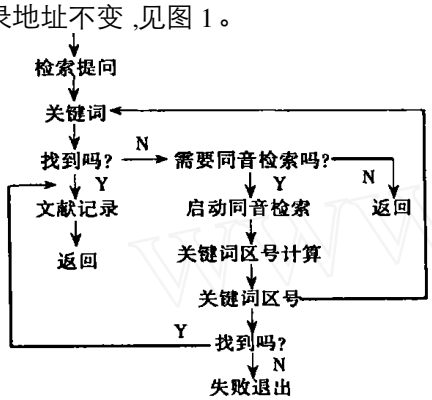

图 1　同音检索流程图

1.2　同音检索算法构造基点

同音检索的过程是将要检索的关键词每一字符的高字节 ASCII 码与存储区内已建立的关键词每一字符高字节一一比较,两者一致时存储区内关键词对应的文献记录为查询记录。同音字虽然机内码不相同,但高字节规定的区号是相同的。因此,子模块中首先建立关键词的每一汉字高字节区号构成的索引。例如,关键词“电路节点”的区号索引为:21-34-29-21。

关键词索引建立步骤:

(1) 初值 j = 1

(2) 求关键词(字符串)长度,M ←LEN(M $j),其中,M $j 为字符串

(3) 分划关键词为单一字符

do i from 1 to M

K $i ←MID $(M $j ,i ,1)

(4) 将每个字符的区号(高字节部分)连接起来

do i from 1 to M step 2

A $j ←A $j + K $i

(5) j ←j + 1,直至 j = N,实施(2)～(4),其中,M $1...M $N 为系统内已建立的 N 个关键词。

(6) 排序链接区号,并与原文献记录建立索引关系。

对于(6),按 GB2312-80 规定,一级汉字出现在 16～55 区。如果按关键字第一区号排序,就可采用分级技术。这里,关键词集所有第一区号作为一级索引,通过简单计算即进入口。以后,关键词比较采用效率较高的二分检索法。有少数汉字可能跨区,如“宋键义”和“宋健义”,模块允许两种定义:43-28-50;43-29-50,它们均与原文献记录索引。

此外,有可能出现区号完全相同的关键词,采用链接方式处理,检索结果将它们的文献记录全部输出,由用户判断需要哪一个。

同音检索算法的要点:

(1) 若常规检索失败,退出,以菜单方式询问用户是否要同音检索;

(2) 进入同音检索子模块,将检索的关键字求长度,分划,确定区号;

(3) 关键词第一字符区号简单计算,进入相应区域;

(4) 进行二分检索,第二字符区号以确定二分范围;

(5) 成功输出结果,失败退出。

1.3　算法实现描述

YJ1:输出待检索关键词 N $。

YJ2:进入常规检索。找不到,询问用户是否要同音检索?要,进入同音检索子模块(入口 YJ3);否,退出。

YJ3:求 N $长度,d ←LEN(N $)。

YJ4:分划 N $成单一字符 K $1,K $2,…,K $d。

do i from 1 to d

K $i ←MID $(N $,i ,1)

YJ5:区号连接 B $ = K $1 + K $3 + K $5 + …+ K $j (j ≤d)。

YJ6:分级入口,从 K ←ASC(K $1)转相应程序段。

YJ7:二分检索,以 K ←ASC(K $3)确定二分范围。

YJ8:二分检索子程序运行。

YJ9:若找到相应区号,其索引的文献记录输出,检索成功。

YJ10:若找不到相同区号,检索失败,退出。

2　多音字处理

2.1　方法构思

课题在研究语音信号识别时,提出了一种解决汉语系统中多音字判别和解决的方法。这种方法制成一软件模块,采用统计学习的思想,里面含一个基于特征的词典,该词典可以根据学习的语料动态更新。一字多音在汉语中是常见的问题,没有统一的规则可循。通常判别多音字的读音是根据经验,或者是约定成俗的读法。解决多音字判音问题应包括两个方面:词组形式出现的多音字和单字形式出现的多音字。

本文提出的方法是一个字音转换模块,它是文本串 $W = w_1 w_2 \ldots w_n$ 到音序列 $C = c_1 c_2 \ldots c_n$ 的映射。其中,w_i 代表一个字,c_i 代表该字对应的正确拼音。这一方法的实施过程是,有多音可读文本串进入字音转换模块,经分词处理后,程序自动进行判别,最后输出正确的读音。

2.2　多音字判别

多音字判别方法中技术的关键是基于统计特征,特征提取使多音字正确判音有效。特征包含在特征词典中,采用规则描述。共定义了以下特征:

(1) 词内左右邻接字,通式为:$x_{i-1}x_i$ 和 $x_i x_{i+1}$。x_i 是当前要判断读音的多音字,这是处理多音字在不同的词语中读不同的音的情况。例如“人参”与“参加”、

“银行”与“行程”、“重量”与“重复”等等。

(2) 左右邻接词,通式为: W_{i-1} x_i 和 x_i W_{i+1}。x_i 是当前要判断读音的多音字,W_{i-1}和 W_{i+1}是多音字的左右邻接词,这是处理多音字与不同的邻接词读不同的音的情况。例如“相当长”、“大队长”、“长方形”等等。

(3) 当前词的词性,例如“数”作名词的读法和作动词的读法,“更”作名词的读法和作副词的读法等等。

(4) 边界条件,该特征是有的字在句首、句末或不同位置读音不同,更多地体现在一些语气助词上面。例如“了”在句中和句末时读音往往不会相同。

2.3 方法实施简介

方法的实现首要条件是大规模的语料,即好的特征词典,并具有学习功能;其次是规则描述,建立一个可以不断扩充的规则集。一般情况下,多音字正确读音根据相应规则去匹配特征词典就能够确定,这方面处理技术已经成熟[4~5],不再赘述;但也有特殊的情况,这主要表现在变调和部分难判断的字,例如“为”、“曾”、“解”等等。因为“为”的高低频度的读音相差不是很大,语言环境都类似,造成无法明显的区分,这时规则反复推理就显得尤为重要,当然这也需要进一步研究。

3 结束语

对于同音检索算法的实验发现,对于输错字或由于地方口音差异而得到的所谓同音字会跨越多个区,算法无能为力。但这时,可以采用词义辅助分析的方法予以解决[6]。

在多音字处理中,也可以研究有权值的特征词典。即特征集记为 F,当前多音字为 x,它的一个拼音得分可以表示成$\Theta_x^i(F)$。上标 i 代表多音字 x 的第 i 个读音,匹配特征时,有一个匹配分,记为 $V_i^k(F)$,它的取值为 0 或 1。0 表示 F 匹配不到该拼音的特征词典,1 表示匹配到,k 表示第 k 个特征。最后计算$\Theta_x^i(F)$的公式为:

$$\Theta_x^i(F) = \sum_{k=1}^{8} W_{i,x}^k V_i^k(F)$$

其中,$W_{i,x}^k$是多音字 x 的第 i 个读音的第 k 个特征的权值,最后找出最大的$\Theta_x^i(F)$。如果权值定义得好,得到多音字拼音正确率就会高得多。

参考文献:

[1] Yael Karov, et al. Similarity-based word sense disambiguation [J]. Computational Linguistics, 1998, 24(1): 41-59.

[2] 符淮清. 现代汉语词汇[M]. 北京: 北京大学出版社, 1985.

[3] 万建成. FPY 中的同音词智能识别方法[J]. 中文信息学报, 1993, 7(2): 23~26.

[4] 杨宪泽. 基于规则的索引算法和排序算法[J]. 中文信息学报, 1993, 7(2): 67~72.

[5] 杨宪泽. 自然语言处理的句法分析和规则索引算法[J]. 科技通报, 2002, 18(6): 470~473.

[6] 杨宪泽. 近似检索的一些处理方法[J]. 计算机工程, 1998, 24 (1): 34~36.

(上接第 87 页)

```
    mov mul_l ,a
    mov a , # 0
    addc a ,mul_h
    mov mul_h ,a
    lcall ram_add
    mov a ,r2
    rr a
    mov r2 ,a
    jnb acc.7 ,LB
    inc r3
    cjne r3 , # 24 ,LB
    ret
    ram_add:
    push dpl
    push dph
    mov a ,mul_l
    mov r0 ,a
    lcall lcd_data
    mov a ,mul_h
    mov r0 ,a
    lcall lcd_data
    mov r0 , # 24h
    lcall lcd_cmd
    pop dph
    pop dpl
    ret
```

3 结束语

本系统已通过了硬件调试,达到了快速旋转显示汉字的目的。与 16 ×16 点阵汉字显示相比,16 ×16 点阵汉字一屏汉字的显示的信息量为 120 个汉字,而采用 12 ×12 点阵汉字显示一屏的汉字显示的信息量为 200 个汉字。在汉字未旋转以前,显示一屏的汉字为 10 行,而将汉字旋转后,显示一屏汉字为 20 行,这样大大地提高了汉字显示的层次感。与一般的汉字显示技术相比,一般的显示技术显示一个汉字要 30ms,显示一屏要 6000ms。而本系统中采用的快速显示技术显示一个汉字的时间为 5ms ,显示一屏汉字所花时间为 1000ms,大大提高了显示速度。

参考文献:

[1] 杨立辉,等. 点阵式液晶显示模块与单片机的接口技术[J]. 微计算机信息, 1997(7).

[2] 李朝青. 单片机原理及接口技术[M]. 北京 :北京航空航天大学出版社, 1998.