MIS的同音检索问题研究

杨宪泽　（西南民族学院）

摘要：本文以科研成果的*MIS*为基础，分析了中文关键词在计算机内存储，检索的过程，给出了同音检索算法。为配合这一算法的广泛应用，还介绍了词义辅助分析法。该算法和方法的引入，在用户输入关键词误输同音字后，*MIS*仍可完成检索工作。

一、引言

国内大量的管理信息系统，其查询方式通过中文关键词来实现。即先按中文关键词内容进行索引，然后通过定位函数确定信息记录地址。这一过程要求中文关键词与索引内容一致。如果使用时在中文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“检索”；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。

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%[1]，不加声调时同音词达27%[2]。这说明，同音问题是干扰MIS有效应用的一大障碍，应该引起足够的重视。

本文的工作以科研成果的MIS为实验框架，设计同音检索算法子模块，在MIS的检索方面力图解决同音问题。此外，对于输错字，或由于地方口音差异而得到的所谓同音字，我们设计、实验了词义辅助分析方法。

二、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[3,4]，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，中文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$94\times94=8836$个汉字。在同一系统中，ASCII码和汉字代码之间的区分可以用特定的标识符，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，也可看出中文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，是中文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，各种符号图形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，子模块中关键词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。

使用时，按常规查询误输同音字，导致检索失败，可进入这个子模块。子模块中进行关键词每一汉字区号比较，两个关键词若每一汉字区号完全相同，则存储器中关键词对应的文献记录就是要检索的内容。这样，误输同音字导致检索失败的问题就得以解决，或在很大程序上缓解。

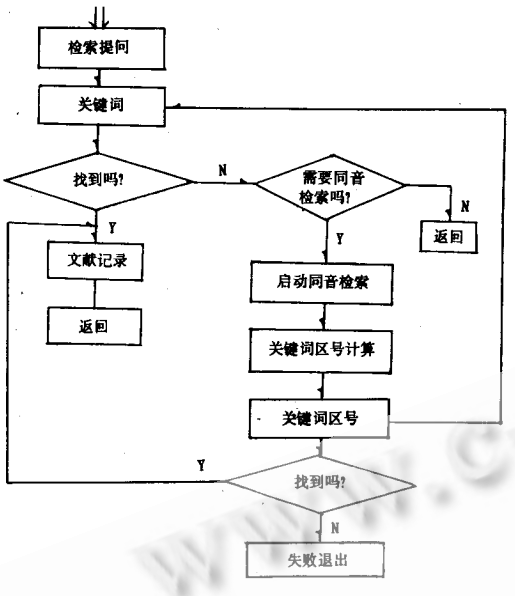


图 1

三、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 ASCII 码一一比较，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，但高字节部分规定的区号是相同的。因此，子模块中首先建立关键词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)切分关键词成单一字符：

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。如果按关键词第一区号排序，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，见图 2。

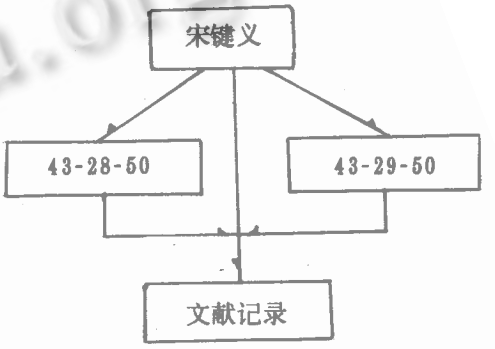


图 2

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均输出，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，将检索的关键词求长度，切分，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。

四、算法实现描述

A1:输入待检索关键词 N $。

A2: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 A3)；否，退出。

A3:求 N $ 长度，d←LEN(N $)。

A4:切分 N $ 为单一字符 K $ 1，K $ 2，… K $ d。

do i from 1 to d

K$i←MID$(N$,i,1)

A5:区号连接 B$ ＝K$1+K$3+K$5+…+K$j(j＜d)。

A6:分级入口,从 K←ASC(K$1)转相应程序段。

A7:二分索索,以 K←ASC(K$3)确定二分范围。

A8:二分检索子程序运行(鉴于有关书刊[5]中这种算法均有介绍,本文略)。

A9:若找到相同区号,其索引的文献记录输出,检索成功。

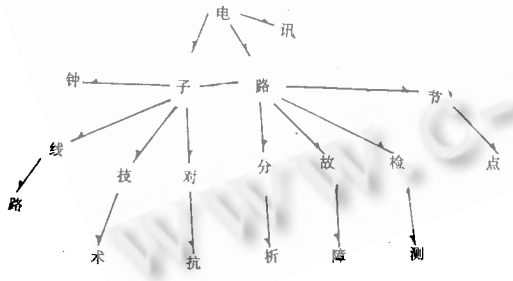
A10:若找不到相同区号,检索失败,退出。

五、词义辅助分析

上述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。为此,我们设计、实验了词义辅助分析模块。

该模块以科研成果 MIS 搜集的关键词的基础,定义了每一关键词中的字与词的相关关系,即字与字之间的搭配关系。对于每一关键词来说,总是从第一个字入口。依次检查,只要词之间任何一个字出现同音字,模块将不予认可从这里中断,显示出已通过的部分和造成同音中断的这个字,由用户进行修改,否则视为检索失败。

模块中,相关的关键词构成了词义辅助分析树,关键词集由若干多叉树组成。如,我们的实验模块搜集的以"电"为入口的关键词辅助分析树为:



若把"电子对抗"输成了"电子对杭""抗"在 31 区,"杭"在 26 区,这属于输错字,靠同音检索算法无能为力,但通过词义辅助分析树,将发现"杭"字有问题,屏幕显示出

电子对[杭]

这说明,程序模块对"杭"字质疑,请求用户修改。

如果跨区同音字或错字发生在关键词中间部分,那么屏幕上也会出现上相似内容,但关键词后部分不出现,只有修改了同音字或错字后,辅助分析方能继续进行。

如果跨区同音字或错字发生在关键词第一个字,屏幕会显示第一个字或第一、二个字。因此,屏幕要求修改第二个字时并不排除第一个字是错的。

辅助分析程序模块连接在图 1"失败退出"框之前,见图 3。

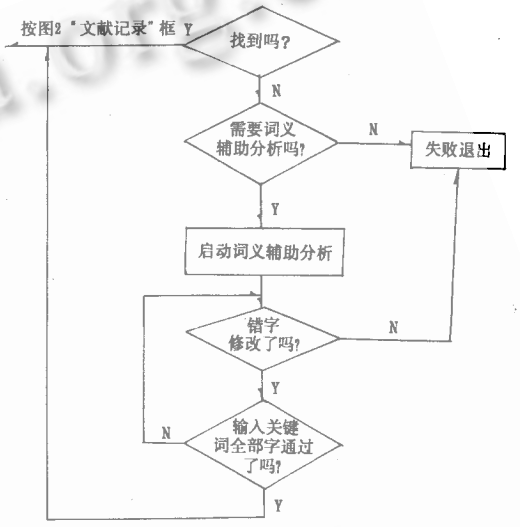


图 3

希望本文的算法和方法能视为 MIS 在常规检索方面发展的一个环节。或从某种意义上讲,本文的算法和方法可能为 MIS 在常规检索方面的发展提供启示。

感谢我院秦文海同志为本文同音检索算法所做的大量实验工作。

参考文献:

[1] 符淮青,现代汉语词汇,北京大学出版社,1985。

[2]万建成,FPY 中的同音词智能识别方法,中文信息学报,第七卷,1993 年第 2 期。

[3]赵珀璋等,计算机中文信息处理,宇航出版社,1989。

[4]郭平欣等,汉字信息处理技术,国防工业出版社,1985。

[5]严蔚敏等,数据结构,清华大学出版社,1988。