西文 DOS 环境下汉字交互式界面的设计

——一种汉字模糊输入方法

刘子斐 (陕西财经学院)

摘要:本文分析了西文环境中应用软件开发及运行的优势及汉字输入的需求,讨论了汉字输入及处理的原理和方案,设计并实现了一种实用的汉字索引模糊输入方法。

一、引 言

西文 DOS 图形方式下实现汉字的处理,以往在这方面做的较多的是汉字的显示输出,实现西文环境下汉字输入有重要的实际意义。

二、原理及方案设计

西文环境下汉字模糊输入基于图形方式,利用汉字内码与国标码、区位码之间的关系及拼音码与国标汉字符集排布规律(GB3212-80 方案),汉字库的存储规律,直接实现对汉字库的处理。

1.汉字内码、国标汉字编码、区位码之关系

各种汉字输入方法尽管其外部编码不同,而其内码的设置则是标准的,目前最常用的汉字编码(内码)标准遵循两字节内码方案,即一个汉字采用二字节表示,其基本特征是每字节的最高位(bit7)均为"1",以区别西文 ASCII 码,汉字信息的这种编码是以国标汉字字符集 GB12-80 为基础,由国标汉字演变而来,两者有着密切的联系:国标码+80H(十六进制)就形成了汉字内码,即高位变"1"。除国标码外最常用的是国标区位码,它是国标码的变形,对同一汉字其国标码比区位码大 20H。按国标 GB3212-80 的汉字编码方案,将常用汉字分为常用(一级)和次常用(二级)共约 6200 字,加上各种汉字符号及西文 0 字符(二字节),总数近 8000。这 8000 字符划分为 94 区,每区 94 个汉字(位)或其字字符,每一汉字及其它字符均有一固定区位码与其对应(区码、位码均用两位数描述)。汉字内码、国标码和国标区位码之间有下述关系:

汉字国标码=汉字内码-80H

汉字国标码=汉字区位码+20H

区位码=汉字内码-A0H

2.汉字拼音码、区位码与汉字库存储规律

一般汉字系统中汉字库都按区位码的大小顺序编排及存放,即遵守国标 GB3212-80 方案。因此知道了汉字

区位码就能找到字库中相应的汉字位置、从而读出汉字点阵数据。汉字区位码的分配还与拼音有密切关系,一级常用汉字约 3760 个按拼音顺序分配区位码,次常用汉字以偏旁部首为顺序分配。汉字从 16 区 01 位开始到 87 区止,16 区前各区存放各种字符数字及西文、日文、俄文字母等。按上述规则,对于一级汉字同一拼音音标的汉字排在一起,并且按拉丁文顺序 a、b、c…z 先后排列,如从 16 区开始的若干汉字及对应区位码和拼音码如下:

a		ai			an		ao	ba
啊	阿	埃	挨	艾…	氨	…	澳	芭
1601	1602	1603	1604	1605…1617…			1636	1637

利用区位码和拼音码这种关系及区位码同汉字库的对应关系,我们可以按区位码和拼音码直接读写汉字库,取得汉字的点阵信息并对其作相应处理(如放大),即可输出汉字。由于区位码与汉字是一一对应的,且汉字库中每一汉字也是按区位码顺序先后存放的,故直接输入区位码就可直接取得某汉字信息,对于区位码不清楚的汉字,可预先给定一区位码,当大致确定某一区位码后,程序按屏(每屏汉字数由程序设计时指定,一般为 10–15)输入汉字供挑选,并可翻屏。这对于不知确切区位码只知大致范围的输入很有用,同时还可用于对不常用汉字的输入和查询,这是一种直接的简单易行的方法,但缺点是必须熟记每一汉字的区位码,给使用者带来不便。

3.拼音模糊输入方案设计

拼音模糊输入是直接输入区位码的一种改进,这种方法利用拼音和区位码之间的关系,只需输入标准拼音码的前一个或若干个码即可。拼音码模糊输入是一种间接的方法,这是一种准拼音码,即不完全的拼音码。综上所述,拼音码与常用汉字的区位码有一种范围对应关系,只要预先详细描述出这种关系,将其形成一个或若干个分级的索引文件,即拼音码与区位码(汉字位置)的对照表,则可模糊输入拼音码由系统转换为区位码。这两种方案的实质都是取得某一个或某一屏汉字的点阵数据,并对其每一位做"描点"输出(位值为"1"时在对应象素点输出点或彩点)。处理数据"描点"时,需注意点阵的各位与屏幕象素的对应关系。一般显示用 16×16 点阵汉字库是按行存放的,汉字节按顺序为每二字节为一行象素位,而某些高密度字库如 24X24 点阵由于作打印字库,故按列顺序存放,即象素列阵,72 字节按顺序为,每三个字节对应于某一列象素点,0～2 为第一列,3～5 为第二列……69～71 为第 24 列。点阵数据与屏幕象素点对应关系如图 1 所示。因此在处理点阵数据做描点输出时对于 16×16 点阵,顺序处理每一字节(0,1,2,…30,31),每两字节下移一个 Y 坐标(一个象素)而对 24×24 点阵,根据其存储模式,每一汉字的一横笔(一行)象素点对应的是每三个字节为一个单位的第一字节的对应位,如第 0 字节,第 3 字节,第 6 字节…的相同位值,因此顺序处理输出其 72 字节数据时,每三个字节为一个单位,每输出一位都下移一行,共 24 行,之后再回到起始行,同时 X 坐标右移一个象素位。

需要注意的是读取汉字点阵时,汉字库中的汉字地址并不完全对应于区位码,这是因为有些汉字库为了节省存储空间,省去了 10–15 区的扩展部分,即汉字存储由 16 区开始提前至第 10 区,如 CCDOS 的 CCLIB,而 UCDOS 的 CCLIB.DAT 则保留了全部 10–15 区。因此,对于某一具体字库计算时应有一个调整量。

4.模糊拼音码输入处理过程

(1)建立索引文件。根据 GB3212–80 汉字集拼音与区位码的关系,构造一个拼音码的对照索引表并以文件形式存入 PY.IND,该索引可采取多级索引方式以适应不同的输入要求,第一级索引采用 1 位拼音码,共 23 个(i,u,v 无拼音声母),第二级索引为 2 位拼音码(取拼音头两个字母),以此类推,其索引文件结构见表 1。使用者可根据需要采用一级、二级…等进行汉字选择。索引表的设置应考虑以下几个要点:

①索引表中拼音索引键应包括所有可能的音节、约 392 个;

②下一级索引包含上一级的索引值、即索引键值有第 N+1 级必须有与其对应的 N 级索引,如 bang 为四级索引,其上应有对应的三级索引 ban、二级索引键 ba 和一级索引 b,这样根据查询对象的模糊度,选择上一级或下一级均可。

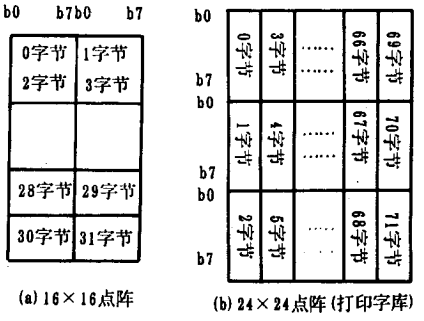

图 1　屏幕汉字点阵数据字模存储格式

③索引值的设置应根据级别所含汉字数多少确定是否设置下级索引(指非标准汉字音标)以减少击键次数加快输入速度。如标准汉字音标中无 bia 但有 bian 根据要点②设置 bia,即输入"b, bi,bia 或 bian"均能找到,鞭、边、编…等。

表 1 模糊输入索引文件

索引文件(表) PY.IND

索引键	起始区位码	结束区位码	索引级
a	1601	1638	1
b	1637	1832	1
c	1833	2077	1
d	2078	2273	1
.	.	.	.
.	.	.	.
z	5249	5589	1
ai	1603	1615	2
an	1616	1627	2
ao	1628	1636	2
.	.	.	.
.	.	.	.
ch	1869	2034	2
.	.	.	.
.	.	.	.
zi	5540	5554	2
zo	5555	5565	2
zn	5566	5589	2
ban	1663	1659	3
bao	1690	1712	3
.	.	.	.
.	.	.	.
cha	1869	1200	3
.	.	.	.
.	.	.	.
zho	5448	5472	3
zhu	5473	5539	3
bang	1678	1689	4
beng	1732	1737	4
.	.	.	.
.	.	.	.
zang	5263	5265	4
.	.	.	.
.	.	.	.
zhou	5459	5472	4
zhua	5505	5521	4
chang	1893	1911	5
cheng	1937	1951	5
.	.	.	.
.	.	.	.
zhong	5448	5458	5
zhuan	5508	5521	5

上述设计的特点是,只要输入拼音的前任何位数都能找到需要的汉字,最大模糊度的输入是一级索引即输入:a、b、c…,按上述方法设计五级索引可包括全部一级 3755 个常用汉字,索引值约 400 个,具体实现时我们可以按索引级别、分别建立一级、二级、三级索引文件(表),根据输入的键值并选择一种进行查询,则可加快索引表的搜索速度。可按数组成树结构把索引文件组织到内部内存中形成索引表。

①先将某一级索引文件或全部索引文件读入内存形成几个索引表(数组、键表或树结构.)

②输入一个索引键(模糊拼音码)(一级、二级…),程序进行合法、合理性检查(除去非字母键)程序按索引级别查找相应的索引表,找到与其相符的索引表项,得到起始和结束区位码并计算出下列几个基本数据:(以 16×16 点阵汉字为例)。

• 索引表项的起始区、位码和结束区、位码:

区码 1=索引值 1、100;

位码 1=索引值 1(模除)1、100;

区码 2=索引值 2/100;

位码 2=索引值 2(模除)100;

• 欲检索的汉字段在字库中的起始位置:(其中调整量根据不同字库取值,即字库中省略的区数如 CCDOS 中的 CCLIB 取值为 6)

((区码 1−1−调整量)×94 位码−1)×32

• 汉字段的汉字总数,其算法如下:(汉字段为该索引所表达的汉字总数)

当区码、等于区码 2 时,表示全部汉字在同一区:

则汉字总数=位码 2−位码 1+1

当区码 1 不等于区码 2 时,表示汉字段跨区:

则汉字总数=95+位码 1+(区码 2−区码 1)×94

• 可输出页数

页数=汉字总数/页汉字数

其中页汉字数可由用户定义,一段为 10…15。

• 剩余汉字数=汉字总数模除页汉字数。即取余数

③打开汉字库,按被检索汉字段的起始地址,读入一屏汉字点阵数据到内存缓冲区,其缓冲大小为:32×页汉字数(字节)

④设置屏幕为图形方式,输出点阵数据到显示行并在每汉字前设置代码及后读字总数区汉字选择。

⑤选择输入显示行汉字或翻屏(前或后),此时程序将动态显示后续汉字的数目。

输入的索引键级越高(模糊拼音码越多),越接近真实码,其搜索的范围就越小,汉字的确定性越高,输入速度越快;反之级别低搜索范围就越大,速度就越低,但输入拼音外码过程简单,如欲输入汉字"毕",我们可只输"b"即可,则有 184 个汉字供选择,按 10 个一屏则有 18 屏零 4 个汉字,需要再后翻 10 屏才能找到,而输入"bi"则一屏就可找到。

三、设计实例

针对上述几种输入方式,笔者编写了一个综合示例程序。该程序有三个功能:

1.单字区位码汉字输入

2.多屏区位码汉字模糊输入

3.拼音码多屏模糊输入

程序主屏见图 2:

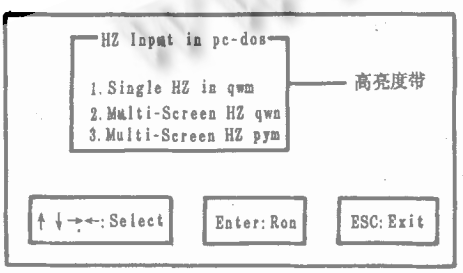

图 2　主功能屏幕

进入系统后,缺省状态为 1 即单字区位码输入,击↑,移动光带至 3,Multi-screen HZpym 回车选择 3;程序进入主屏字库选择菜单,用户通过↑ ↓ →← 键选择汉字库,用 Enter 确定选择,缺省状态为第一项即 UCDOS16×16 点阵字库 cclib.dat 如图 3,汉字库选择,选 UCDOS 后回车。

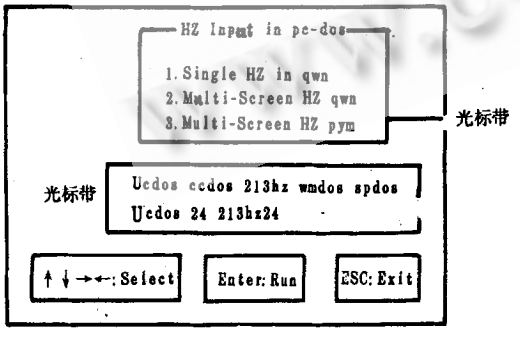

图 3　选择汉字库

系统进入图形方式,在屏幕底部出现拼音索引输入屏,提示用户输入拼音索引,如果希望输入"王"字,最大模糊度时可以只输入索引键"W"并回车后,拼音输入屏上方出现一屏 10 个汉字:[挖,哇,洼,娃,瓦,袜,歪,外,碗]及后续字数[110]当用户键击[+] 键两次后,便出现图示的一屏汉字含有"王"字,[090]表示其后还有 90 个汉字。选定后(其代码为 8),"王"字出现在屏幕中央的编辑屏,见图 4。汉字可重复选择,按 ESC 退出选择,按[+]或[−]键可向前后翻屏,当输出最后一屏时,向前翻屏操作会引起响铃报警,同样回翻至第一屏时,键[−]也将报警。

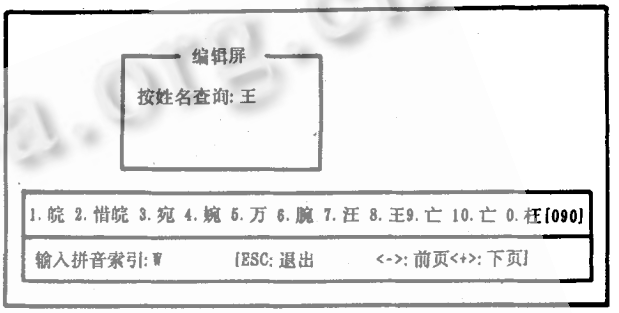

图 4　汉字输入屏

汉字输入结束后,返回所选择输入的汉字串供应用系统作进一步处理,例如作关键字查询汉字数据库文件。下面给出程序的处理流程图(图 5)。该程序用 C 语言写成,

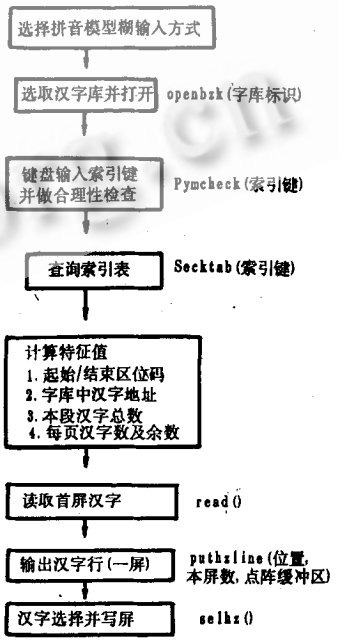

图 5　拼音模糊输入处理流程

在 Turbo C2.0 环境下编译。可在 PC / XT,286、386、486 系列微机上 MS-DOS(PC-DOS)下运行。(由于程序较长,未予列出,需要源程序的同志来函联系。)