汉字编码技术及其在民航的应用

肖殷洪 (民航计算机信息管理中心)

1.汉字内码及汉字通讯

汉字内部码是汉字终端或汉字计算机系统在进行运算,存储等信息处理过程中应用的一种汉字代码.通常,在机器内部处理过程中需要对汉字信息进行各种频繁的操作,并且要求汉字代码与计算机系统原有的西文代码充分兼容.因此,必须正确设计和选用汉字内部码.

众所周知,人们主要通过字形来识别和区分汉字.汉字的字形代码是以数据字节串的形式存储在机内,这就是字形点阵数据码.用这种代码可以方便地控制显示器或打印机,以实现汉字的字形输出.但是,字形点阵数据码太长,会大量耗费机器资源(即机器的操作时间和内存空间),为了节省资源和提高转换效率,需用一种特殊的代码作为内部码来标识每一个汉字字符.

为了能够在两个或两个以上的汉字终端或汉字主机之间进行信息交换,需要同内部码既有联系又有区别的另一种汉字代码—交换码.内部码只是在本机内部使用,设计考虑的角度是如何使机内处理操作简便,并与系统原有的代码兼容;交换码涉及到不同的机器和系统,应有统一的标准.如果相距较远,还必须考虑到便于线路传输及通信的要求,这是内部码与交换码的不同之处.但是,由于这两种代码之间关系十分密切,相互转换非常频繁,因此,应尽可能使它们之间具有简单的一一对应关系,以提高机器操作效率.在不少西文设备和系统中,往往直接采用交换码作为内部码,就是这个缘故.但是,对于汉字系统来说,情况要复杂得多.因为现有的汉字系统是在西文系统的基础上进行二次开发而成的,而汉字交换码也是在西文交换码的基础上扩展编成的.如果直接用汉字交换码作为汉字内部码,在传输中中西代码易混淆,就不能做到汉字系统与西文系统的兼容,因此,必须进行汉字内部码的设计.

首先我们看看西文系统的内部码.它是表示西文字母,数字及常用符号(统称字符)以便于计算机处理的数据代码.人们把需用的字符排成一张表格,用一组有序的二进制数对应编号,成为一张代码表.由于常用的字符(加上基本控制功能字符)总数不超过 256 个,故用不超过 8 位二进制数(即 1 个字节)便可实现字符编码.

2.汉字交换码

汉字交换码是用于汉字信息系统或汉字终端之间交换信息的数据代码.众所周知,为了进行信息交换,交换码必须采用统一的格式.由于历史的原因,我国国内用于汉字通信的代码目前有两种格式.一种是电报码,是以<<标准电码本>>规定的汉字代码;另一种是国标码,是以国标 GB2312-80 规定的汉字代码.

汉字电报码用于汉字电报通信,它用 0-9 中的任意 4 个数字组成一个汉字代码,可以表示一万个汉字或非汉字图形字符.

汉字国标码又称为汉字国标交换码,是国家标准<<信息交换用汉字编码字符集—基本集>>GB2312-80 图形字符代码表的简称.这个代码表是在 7 位标准代码表的基础上扩展的双字节 7 位代码.它在 1981 年 5 月颁布实施,并已得到国际标准化组织(ISO)的认可,成为我国法定的信息交换用标准汉字代码.

汉字内部码的设计要求如前所述,汉字内部码是计算机进行信息加工处理时用的汉字代码.在一个汉字信息处理系统中,汉字内部码的设计和选择,对系统的性能和效率关系很大.目前,对汉字内部码还没有统一的规定,不同的汉字系统有各不相同的内部码设计方案.从现有的实践经验看,汉字内部码的设计和选择,应满足以下几点要求:应与国标码有简单明确的一一对应关系,以便于它们之间相互转换;应保证系统的中西文兼容;应有利于提高系统的效能.

国标码是我国法定的汉字信息交换用的标准代码,而内部码目前尚无统一规定.因此各个不同汉字终端要互相交换信息,就必须把内部码转换成国标码.目前大量生产的汉字字形发生器存储芯片,大多是以国标码作为读出地址的.总之,在汉字信息处理和传输交换过程中,汉字内部码与国标码之间的转换是十分频繁的.为了尽量提高系统

效率和减少转换中可能发生的差错,应使内部码与国标码之间的转换规则尽量简单;并且两种代码一一对应,没有二义性.因此,汉字终端的内部码设计都不是脱离国标码完全另稿一套,而是在国标码的基础上稍加改造,使两者保持简单明确的对应关系.但是,内部码与国标码是不能完全等同的,如果为了免去转换的麻烦而使两者相同,则会影响系统中西文兼容的要求.由于西文系统大多经过较长时间的开发和应用,具备丰富的软件资源,所以在西文系统基础上开发的汉字系统,必须中西文兼容.

但是,现用的汉字国标码(即 GB2312 码)是在西文字符代码(即 ASCII)的基础上制订的,它们之间相互联系,我们看到汉字国标码是由两个 ASCII 码组成的,在机器处理中并不知哪两个 ASCII 字符的组合是汉字,所以必须设计一种可区别于西文 ASCII 码的汉字内部码.

汉字内部码的设计没有统一的标准,视具体使用的机器而定,但一般的设计原则是与国标汉字码保持较简单的对应关系,这样既可保留国标码的特性,又可提高两种码的转换效率.这样派生出多种内码方案,主要有八位码和七位码两种.

八位码是将 7 位 ASCII 码中的高位置位,以视同 ASCII 字符内码的区别.当操作系统读到两个连续的高位置 1 的 ASCII 字符后,将这两个字符转换成汉字.这种方案的优点是占用字符少,内码与国标码转换简单,效率高,但需占用八位,与机器硬件的扩充能力有关.

七位码是在表示汉字的两个 ASCII 字符的前后加上引导码,标识这两个字符是汉字码.这种方案需增加引导码来标识,并且引导码必需是特殊字符,但它可用 7 位 ASCII 码来标识.我们汉字终端就是采用的这种方案.主要是在通讯转输时,一个 BYTE 中,用 7 位表示 ASCII 字符,1 位做奇偶校验的奇校验且不可改变.采用这种方案当终端读到引导字符时,自动将后面两个 ASCII 字符转换成汉字,直到读到标识汉字结束的引导码;在主机中只是将其当做一个字符串来处理.由此而实现了在西文主机上的汉字处理.

兼容性问题是汉化的一个非常重要的问题,汉字终端的使用及汉字在主机内的相应处理,应尽可能地减少对主机原系统的影响,最理想的情况是对主机无影响,保留系统中原有的所有功能.七位码方案较易达到这种要求,系统中的西文不变,对原有西文处理无影响,汉字字符只当一个特定的西文字符串来处理,这样的处理方式对操作系统来说是透明的,只是对应用系统中的相应程序进行修改.但需要修改应用系统程序的基本输入/输出部分,使其能够进行汉字的输入和输出,这部分的修改对应用系统影响较大,因涉及到基本输入/输出所以有可能对西文的输入/输出也产生不良影响.但操作系统升版本时,对应用程序的影响较小.

3.汉字终端在民航服务系统的应用

现有的民航旅客订座服务系统中,有近 2000 台西文设备,包括:西文终端,西文打印机以及电传机,并且主机还有特殊的接口进行国际间电传电报的信息交换.主机处理的汉字如何在这些西文设备上传输,显示是一个很大的问题,如果将这些西文设备全部丢掉未免代价太高,并且将汉字信息转到国外航空公司的主机上,外国人是不认可的.所以在设计代码时必须考虑这部分设备的使用问题.既保证西文时的正常使用,又得在汉字显示时能够正确地表达汉字信息.所以解决冲突是我们重点要考虑的问题.

根据我们系统主机的硬件结构和通讯转输的要求,我们采用了七位码方案.将汉字加上一个特殊的双引导字符.这一引导字符序列不可能由键盘输入,但通讯系统又可将其正确地转入主机应用程序之中.它保证了系统中汉字识别的唯一性.之后,我们对应用系统的基本输入/输出,应用系统程序进行了大幅度的修改,保证了汉字处理的正确性和完整性.对于汉字信息的西文显示问题,根据系统对汉字处理的要求,在汉字由终端转向主机时,加入其拼音码.在主机中对每个终端设备的定义中加入终端类型的标识,在显示一个旅客信息时,如果定义此终端设备为西文设备,只将其拼音码送入终端显示;如定义为汉字终端设备,则将汉字码送给终端显示;如果是电传设备,就只送拼音码.这样,在西文设备上我们只能看到旅客姓名的拼音码,一般来说对国际航班的控制就足够了;在汉字终端上,我们就可读到旅客的汉字姓名,这对国内旅客的控制和管理提供了必要的信息;当国内旅客转到国际航班上或由国际航班转到国内航班,需要电报查询时,我们也能检索到该名旅客,提供正确的查询信息.

几年来的经验证明,我们较好地选用了汉字终端技术,开发出了较完善的终端产品.这一产品受到航空公司用户的级大欢迎,如今已有近 4 千台汉字终端在民航的各个计算机系统中使用,为国家节省了大量的外汇.这一成果证明我们的汉字终端方案上是可行的,技术是成熟的,应用效果是很好的.