

多文种不同编辑方向的Windows'95处理平台的设计与实现

吾守尔·斯拉木 伊力哈木·牙生

(新疆大学计算机科学系, 乌鲁木齐市 830046)

[摘要] 文章叙述了具有不同编辑方向的维、哈、阿、汉和英多文种信息处理的Windows'95平台开发与应用, 不等宽民文信息处理在Windows环境中面临的问题及其解决方法, 以及民文显示技术、民文处理机制及其创新点。

[关键词] 多文种; Windows'95; 信息处理; 零宽度扩充编码; 民文处理系统

1 引言

Windows本质上是一个图形化的操作环境, 提供了多任务运行方式和多媒体支持, 得到广泛应用。它在内存管理、长文件名、虚拟机和多线程方面的特点, 让用户耳目一新。

计算机事业的发展给少数民族地区带来了学习和使用Windows的热潮。新疆是一个各民族聚居的地区, 使用维吾尔、哈萨克、柯尔克孜、蒙古等少数民族文字, 并且它们必须与汉字一起混合使用, 所以开发一套Windows'95民文处理系统有非常重要的价值, 也是时代的迫切需求。针对这一点, 1998年课题组开发了能在中西文Windows'95、Windows'97上处理维、哈、阿(阿拉伯)、汉、英等文字的民文处理系统——“民文视窗”。

2 设计原则和目标

为了在中西文Windows'95上开发民文处理平台, 总体设计原则和目标如下:

1) 通过外挂方式, 改变Windows内部与文本输出有关的GDI函数来实现汉、英、维、哈、阿

文正确混合显示、编排和打印;

2) 在输入、输出模块上解决民文的自动选型, 使得每个民文字符都能准确相连;

3) 民文字库应单独管理, 而不修改原有的中英文字库;

4) 设计并实现几种有效民文输入法, 以适合不同场合的需要;

5) 以动态过程库的形式提供应用程序接口, 使得维、哈、阿文应用程序的编制更为方便;

6) 为了该系统以后的扩充和应用, 应考虑兼容性和可靠性, 充分注意系统的模块化。

3 技术难点

由于民文自身的特点, 在多文种Windows平台开发过程中带来的技术难点表现在以下几个方面。

3.1 多文种编码上的矛盾

为了实现中英文的混合处理, 中文Windows 3.X在保留了前127个单字符特性的情况下, 对OXA1~OXFE的代码实施了双字节扩充, 在中文Windows'95中把这区域扩展到OX81~OXFE, 即首字节大于OX81都视为汉字, 并按等宽处理。

[收稿日期] 1999-08-02; **修回日期** 1999-11-20

[基金项目] 国家“八六三”计划资助项目(863-306-04-04)

[作者简介] 吾守尔·斯拉木(1942-), 男(维吾尔族), 新疆乌鲁木齐市人, 新疆大学教授, 博士生导师

众所周知,要实现多文种混合处理必须保证民文与英文没有代码冲突,否则会出现乱码。OX00~OX80是西文代码区域,不能被占用^[1],所以把民文放到OX81~OXFE的几个未用区域。

3.2 处理等宽汉字与不等宽民文的冲突

为解决这个问题提出了“零宽度扩充编码方案”,从而解决了等宽汉字与不等宽民文在混合处理中的难题。

3.3 书写方向相反所引起的矛盾

现代中文和英文一样,是从左到右输入(右向输入)^[2],但民文是个双向文本(BIDI文本),是从右到左输入(左向输入),两者之间书写方向上的这种差异是多文种混合处理遇到的又一个难点。

3.4 民文的自动选型

民文的每个字符必须根据两边的字符而选择其前连体、中连体、后连体和独立体四种形式之一。怎样实现计算机自动选型并前后准确连接,在输入时自动选型还是输出时自动选型是民文处理的又一个难点。

4 民文零宽度扩充编码技术

4.1 问题的提出

民文字库采用的是单字节索引,但民文编码采用的是双字节编码,即在内存中每个民文字符都是双字节,被显示之前,它们都被映射到单字节上,即双字节输入,单字节输出,这有助于系统的可靠性,但同时也带来了一些问题。Windows是根据内存中的字符来计算宽度,而不是根据被输出的字符串^[2]。在上述双字节输入单字节输出流程上,应用程序得到的字串宽度就不一定等于被输出的字串宽度。在这种情况下,会出现光标脱离和行未满之前换行等异常现象。

每当用户选择一种字体时,应用程序调用Windows函数Get Char Widths或Get Char ABC Width以得到当前设备描述表所用字体的宽度,并把此宽度保存到某缓冲区中。在下次计算某字符串的总宽度时,使用此缓冲区中的宽度值,而不再去调用Get Char Widths或Get Char ABC Width^[1,2]。即从第一次字体改变到第二次字体改变期间,字体宽度是不变的,不能动态地对它们进行修改,因此需要解决从用户程序得到的字符串宽度应等于要被显示的字符串宽度。

4.2 零宽度扩充编码方案

中文Windows'95为了能让用户增加自己的一些符号,留下了两个未用区域OXAAA1~OXAFFE和OXF8A1~OXFEFE,并选用了此区域作为民文代码区。

经过研究和分析,对民文代码进行了如下创新。

1) 把维文、哈文、阿文分别安排了不同的区域同一个位码段。

2) OXFE为维文,OXFD为哈文,OXFC为阿文,这些码位刚好落在Windows'95的未用区域。

3) 原第二码为OXA1~OXFE,不能容纳维文的110多个字符,所以把尾码从OXA1~OXFE扩充到OX81~OXFE,这样可把代码码位扩展到128个,完全容纳所有维文字符并有余量。

4) 为了在民文与英文之间的自由切换,保留原OX00~OX80为西文字符区。

5) 让民文首码OXFE、OXFD、OXFC的宽度为零,并让尾码字符宽度任意的情况之下,增加宽度为零的首码,这就是“零宽度扩充编码”方案。

某个维文字符OXFEB0的宽度是两个字符OXFE和OXB0的宽度之和,其中OXFE的宽度为零,那么OXFEB0的宽度就等于OXB0的宽度,而OXB0就是要被输出的字符。这样可以解决民文与中英文没有代码上的冲突,又能解决双字节与单字节的宽度不等所带来的问题。

零宽度扩充编码方案简单地通过代码优化解决多文种代码冲突问题,解决了全字和半字处理(即双字节与单字节的宽度不等)所带来的一系列问题。

5 民文显示技术

民文是一种反向文本,从右往左写,并且每个字符都有前连体、中连体、后连体和独立体等不同形状,所以民文的显示相对来说比较麻烦。为了程序的简便性和系统的可靠性,在编程时,充分注意到了模块化。修改后的Ext Text Out和Text Out函数由下列模块组成:字符串识别(分隔)模块、映射模块、倒序模块、自动选形模块、字体选择模块和显示模块。

字符串分隔模块负责判断原字符串是否为民文:若不是民文,就直接把该字符串送显示模块显示;若是民文(即首字节为OXFC,OXFD,

OXFE), 就把该字符串送映射模块。

根据设计原则单独管理民文字符, 而不是去修改原有的中西文字符。该字符是标准单字节字符, 但给民文分配的是双字节代码。映射模块负责从原来的双字节与它相对应的单字节转换工作。这样保证了无代码冲突, 又保证了民文字库的独立性, 从而提高了系统的可靠性和兼容性。

在 Windows 中, 西文字符原有的输入方式是从左到右, 即第二次输入的字符应在第一次输入的字符右边^[2], 但与此相反, 民文是左向输入文本, 第二次输入的字符应在第一次输入的字符左边。因此, 内存中的民文字符串必须进行一次倒序。倒序模块负责把原民文字符串进行倒序, 然后再把该串字符送自动选型模块。

每个民文字符都有前、中、后、独立等 4 种不同形状。一个字符在字符串中的位置不同, 它所带来的表现形式也不同。自动选型模块接收倒序模块送来的民文字符串, 并对其中每个字母进行自动选型, 使它们能准确相连。

解决自动选型的方法有两种, 其一是输入时自动选型, 二是输出时自动选型。输入时自动选型是当用户击民文字符键时, 把每个字母以适当形式输入计算机, 应用程序得到的字符串是经过选型的字符串。这种方法的优点是可以得到民文字符串的精确宽度, 以便准确定位光标。但其缺陷是不能确定插入字符及识别两边的字符, 难以解决字符插入问题。输出时自动选型是当击民文字符键时, 只输入每种字母的一种形式, 即其独立形式。在字符串被显示之前对它进行一次调整, 选型后再输出。这种方法的优点是使自动选型容易实现, 并可以解决字符插入问题。其缺陷是一个输入字符不一定是输出字符, 并且民文的每一个字母 4 种不同形式的宽度不相等, 所以在内存中的字符串和显示字符串的宽度会出现差异, 而难以准确定位、插入字符。以上两种方法均有其优点和缺点。在制作本系统时充分考虑了这一点, 并结合了这两种方法进行优化。即输入民文时先对它们选型一次, 当这些民文字符串被显示之前, 再对它们进行一次选型。这样可以准确定位光标, 也解决了字符插入问题, 这是本系统的创新之处。

字符串经过自动选型模块之后, 全部发送到字符选择模块。该模块的主要任务是建立与当前字体大小相匹配的民文逻辑字体, 并把它设为当前字

体, 然后把字符串送显示模块中显示。本字符串显示后, 删除刚建立的民文逻辑字体, 以空出内存空间。

一个字符串的最后一个模块是显示模块, 该模块在指定位置上用当前字体显示指定的字符串。

6 True Type 字库与点阵字库的使用技术

True Type 字库和点阵字库是 Windows 中常见字库, 两者都有优点和缺点, 适合于不同的场合。如果字体选的太大, 点阵字库会带有锯齿形状, 显得难看, 但 True Type 字库则不会。如果字体选得太小, True Type 字库会显得非常模糊, 难看, 但点阵字库则不会。我们通过对前期开发的 Windows 3.1 的分析可知, 一个 True Type 字库, 如果字体小于 20 磅, 就自动映射到相应的点阵字库, 这样用户界面会更加美观和方便, 所以“民文视窗”采用了当字体小于 14 磅时, 映射到点阵字库“MS Sarts Serif”的实施方案。

对点阵字库的映射, 在美化用户界面方面起了积极作用, 但也带来一些问题。因为在 True Type 和点阵字库的相应字符间存在着宽度差, 所以映射到点阵字库时会出现光标脱离等异常现象^[3]。这个问题的解决途径是: 当字体大小或字体名改变时, 应用程序都调用 Get Char Width 或 Get Char ABC Widths 等函数, 得到一个字库中该字符的宽度。所以修改这两个函数来确保在民文字体小于 14 磅时, 上述函数得到的字符宽度是点阵字库的宽度, 而不是 True Type 字库的宽度。

7 系统实验与应用

于 1998 年 8 月国家“八六三”计划智能机主体专家组对系统组织技术鉴定, 通过了测试, 并在该系统平台上研制成功的大型集成化维文多媒体软件及其推广应用表明, 该系统的性能指标达到了原中西文 Windows '95 系统的性能指标。不仅能混合处理不同编辑方向的多文种信息, 而且对各种高层应用软件如 VB5.0、VisualC++ 等各类高级语言、Access 95 等数据库类、Notepad 等字表处理类、photo shop 等图形处理类、Atho Ware 等多媒体工具类、3DSMAX 等动画制作类等都能提供多文种支持。对多文种的输入、输出提供了直接输入法、提示行输入法、智能输入法、鼠标输入法等。目

前，这种多文种信息处理 Windows 平台及其应用软件正在广泛地推广应用，并对民族地区多文种信息处理事业的发展有着重大意义。

8 结论

在“民文视窗”中首次提出了“零宽度扩充编码方案”，以编码结构优化来解决民文与中西文的编码冲突问题，并解决了等宽中文和不等宽民文混合处理问题。从系统最底层核心级扩展系统功能，建立了民文处理机制；支持一切高层应用的民文处理，提高了兼容性；提供了 4 种输入法，以适合不同的环境；并提供了 7 种民文字库以满足不同的用户需求。在解决自动选型问题上，提出了输入时自动选型和输出时自动选型两者结合的方法，解决了

字符插入，删除等编辑问题。为了使系统界面更加美观，提供了 True Type 和点阵字库的自动切换技术。为了多文种用户使用方便，屏幕上的一些常见提示信息可动态翻译成母语。系统一开始就考虑了模块化工作，有着较高的可靠性和兼容性。

参考文献

- [1] Microsoft 公司. Microsoft Windows 3.1 程序员参考大全 (二)、(三)、(五) [M]. 吴之美, 余立新译. 北京: 清华大学出版社, 1993
- [2] Maruzzi S. Microsoft Windows '95 开发者必读 [M]. 翟炯, 石秋云译. 北京: 电子工业出版社, 1997
- [3] Richter J. Ms Windows NT Server 3. 51 程序员参考书 [M]. 郑全战, 王毅译. 北京: 清华大学出版社, 1997

Designing and Implement of A Multi language Windows '95 Platform

Wushour·Slamu, Iliham·Yasin

(Dept. of Computer Science, Xinjiang University, Urumqi 830046, China)

[Abstract] The development of a multilanguage, multiedit direction Windows '95 Platform is introduced, which can process Uigur, Kazak, Arabic, Chinese, as well as English. The technique difficulties and the solution in processing different width letter of multilanguage information under Windows system are described in detail. Some renovation methods of the multilanguage processing and display technique are proposed in this paper.

[Key words] multilanguage; windows '95; zero width extension code; multilanguage processing system

《中国工程科学》2000 年第 3 期要目预告

齿轮技术的创新和发展趋势 梁桂明

21 世纪中国水资源持续开发利用问题 陈志恺

加速胚胎移植技术产业化促进我国养牛业发展 郭志勤

视频分层组织方案和技术 章毓晋等

论航空发动机发展的“渐改法” 温俊峰

开发浅海滩涂资源 拓展我国的生存空间 陈吉余

成像光学工程面临的光学问题 潘君骅

影响微电子机械系统成品率和可靠性的粘合力 and 摩擦力 王渭源

航海避碰智能系统 郝燕玲等

长江水质酸化与黄河水质浓化趋势及成因探讨 陈静生等

钨矿物原料 NaOH 分解过程中抑制杂质的研究 李洪桂等

测量与信息技术 张国雄

动力气垫地效翼船的设计特点及其发展前景 恽良等

测量用互感器及其测试仪器的发展和创新 赵修民等

发展我国新型轻钢结构建筑体系 陶忠等

油料中功能性成分的研究开发与油脂工程技术的发展 张根旺

农业信息工程的理论、方法和应用 孙九林