第18卷 第1期 西昌农业高等专科学校学报 2004年3月
Vol.18 No.1 Journal of Xichang Agricultural College Mar. 2004

WINDOWS中DBCS的解决方案

柳 刚 吴德萍 单成海
(西昌学院 四川西昌 615013)

摘 要:本文通过对常见的三种字符集编码原理的介绍,然后专门针对在微软操作系统下,双字节字符集中怎样求一字符串的字符数和字节数的常用方法进行讨论。

关 键 词:字符集;SBCS;DBCS Unicode;字符数;字节数;API
中图分类号:TP312 文献标识码:C 文章编号:1008-4169(2004)01-0048-02

随着微软件公司操作系统不断更新,它所使用的字符集也在不断的变化,在DOS下使用单字符集(SBCS),而WINDOWS下却增加了双字节字符集(DBCS)和Unicode。由于这几种字符集的并存,程序员在编写程序时,在不同的字符集下求字符串长度和字符串字节数的方法根本不相同。特别是在DCBS下相对要困难一些。

一、问题的提出

在WINDOWS 98下主要是使用DBCS来处理字符,在该操作系统下处理字符串时会出现一些怪问题:1、在VF下用len函数测试字符串"abc12中国人民"的长度是13,用sbustr("abc12中国人民",7,1)得到是乱码。在C语言下使用标准的strlen("abc12中国人民")得到的长度也是13。2、在VB6下len("abc12中国人民")得到的长度却是9,用mid$("abc12中国人民",7,1)得到却是"国"。从上不难看出在第一种情况下得到的长度是字符串的字节数,而在第二种情况下长度是字符串的字符数。如果现在要在这几种编程环境中都需要得到字符串的字符数和字节数怎么办呢?见下面叙述。

二、字符集的编码工作原理

下面介绍常见的三种字符集编码原理。

(一)单字节字符集(SBCS)

它的所有字符都只有占一个字节的长度,也就是占8个位。因此它最多有256个字符。ASCII码就是SBCS,DOS下采用的就是SBCS。在SBCS中每个字符串的结束标志是值为0的字节。

(二)双字节字符集(DBCS)

它包含的每一个字符可能是一个字节或者是两个字节。Windows 中用得最多的字符是双字节字符集。在DBCS中是怎样来区分一个字符占一个字节或占两个字节的呢?在DBCS编码中,用一些保留值来指明该字符属于双字节字符从而达到上面的要求。例如,在Shift-JIS(通用日语)编码中,值0x81-0x9F 和 0xE0-0xFC 的意思是:"这是一个双字节字符,下一个字节是这个字符的一部分"。这样的值通常称为前导字节(lead byte),其值总是大于0x7F。前导字节后面是跟随字节(trail byte)。DBCS的跟随字节可以是任何非零值。与SBCS一样,DBCS字符串的结束标志是值为0的字节。

(三)Unicode编码

Unicode编码标准中的所有字符都是双字节长。即用一个16位的值来表示每个字符,因此总共可以得到65000个字符。Unicode字符串用二个零字节字符来表示字符串结束标志。

三、解决方案

在了解工作原理之后我们来看一下在DBCS下怎样来求一字符串的字符数和字节数。如果求字符串的字节数我们只要一个字节一个字节的遍历一次就可以。求字符串的字符数,只要分清哪些是前导字节并进行统计最后以字节数减去前导字节的个数即为该字符串的字符长度。很幸运的是在WINDOWS中为了能正确地对DBCS字符串进行操作,WINDOWS提供了一组函数(表1)。

CharNext和Charprev允许前向或逆向遍历DBCS字符串,IsDBCSLeadBye在字节返回到一个双字节符的第一个字节时将返回TRUE值。下面用VC++编写的一个函数来举例说明怎样返一个字符串的字节数和字符数。

收稿日期:2003-10-14

表1　对DBCS字符串进行操作的函数

函数原型	功能
PTSTR CharNext(PCTSTR pszCurrentChar);	返回字符串中的下一个字符的地址。
PTSTR CharPrev(PCTSTR pszStart,PCTSTR pszCurrentChar);	返回字符串中的上一个字符的地址。
BOOL IsDBCSLeadByteTRUE(BYTE bTestChar);	如果该字节是DBCS字符的第一个节字(即前导字节),则返TRUE。

```
//此处只列出类CTestDlg中的GetStrLen方法，该方法实现了上述的要求
int CTestDlg::GetStrLen(char * str, int flag)
//str要测试的字符串首地址,flag确定返回字节长度(flag=0)还是字符长度(flag=1)
{
    char *p=NULL, *temp=NULL;//定义两个临时指针
    int charlen=0,leadlen=0;//定义字符长度变量charlen和前导字节个数变量leadlen
    p=str;  //将字符串首地址赋给指针变量p
    while(1)
    {
      if(*str==NULL) return 0;  //如果为空串返回0
      charlen++;   //由于CharNext返回的是下一字符地址,所以变量charlen放在该函数的前面。
      if (IsDBCSLeadByte ((BYTE) *p)) leadlen++;  //如果该字符是双字节则leadlen加1
      temp=CharNext(p);  //返回下一字符地址
      p=temp; //将余下字符串首地址重新赋给指针变量p
      if(*temp==NULL) break; //如果到达字符串结束处则跳出循环
    }
    if(flag==1)
      return charlen;//返回字符长度
    else
      return charlen+leadlen;//返回字节长度
}
```

此函数使用了CharNext函数来遍历DBCS字符串,在遍历时,对每个字符又用IsDBCSLeadByte函数来测试该字符是否为双字节数，如果是则leadlen变量加1。当遍历完成后,循环的次数就是该字符串的字符长度。如果要得到字符串的字节长度，根据DBCS的编码规则可知,字符串的字节长度=字符串的字符长度+字符串中双字节字符个数。

四、小结

其实在VC中专门有针对DBCS操作的库函数。但是本文并没有对它们进行讨论，主要是因为这些库函数只能在C语言中应用，而上面讨论的三个函数是WINDOWS API,它们可以在其它的编程语言中使用,实现也比较方便。如在VB中可以通过VB自带的API文本浏览器将其它声明在VB程序中。本文只讨论了求字符串的字节数和字符数，其实还可以使用上面的三个函数进行准确地取子串，而不会出现乱码以及求某个子串在字符串的位置。

参考文献:

[1]王建华,张焕生,侯丽坤等.Windows 核心编程[M].11-12.

Resolution of DBCS in Windows

LIU Gang , WU De-pin and SHAN Cheng-hai

(Xichang Institute, Xichang Sichuan 615013)

Abstract:This paper gives an introduction to the three commonly used string coding principles and then a detailed discussion on popular methods of solving the number of characters and bytes under Microsoft operation system.

Key words:String;SBCS;DBCS;Unicode

(责任编辑:蔡光泽)