

转 单字节字符集，多字节字符集，Unicode

2013年10月15日 18:51:04 阅读数：5292

我们在这里介绍一下字符类型。这里有3种编码模式对应3种字符类型。

第一种编码类型是单字节字符集 (single-byte character set or SBCS)。在这种编码模式下,所有的字符都只用一个字节表示。ASCII是SBCS。一个字节表示的0用来标志SBCS字符串的结束。

第二种编码模式是多字节字符集 (multi-byte character set or MBCS)。一个MBCS编码包含一些一个字节长的字符,而另一些字符大于一个字节的长度。用在Windows里的MBCS包含两种字符类型,单字节字符 (single-byte characters)和双字节字符 (double-byte characters)。由于Windows里使用的多字节字符绝大部分是两个字节长,所以MBCS常被用DBCS (double-byte character set or DBCS) 代替。

在DBCS编码模式中,一些特定的值被保留用来表明他们是双字节字符的一部分。例如,在Shift-JIS编码中(一个常用的日文编码模式), 0x81-0x9f之间和 0xe0-0xfc之间的值表示"这是一个双字节字符,下一个字节是这个字符的一部分"。这样的值被称作"leadingbytes",他们都大于0x7f。跟随着在一个leading byte子节后面的字节被称作"trail byte"。在DBCS中, trail byte可以是任意非0值。像SBCS一样, DBCS字符串的结束标志也是一个单字节表示的0。

第三种编码模式是Unicode。Unicode是一种所有的字符都使用两个字节编码的编码模式。Unicode字符有时也被称作 *宽字符* (Wide Character), 因为它比单字节字符宽(使用了更多的存储空间)。注意, Unicode不能被看作MBCS。MBCS的独特之处在于它的字符使用不同长度的字节编码。Unicode字符串使用 两个字节表示的 0 作为它的结束标志。

单字节字符包含拉丁字母表, accented characters及ASCII标准和DOS操作系统定义的图形字符。双字节字符被用来表示东亚及中东的语言。Unicode被用在COM及Windows NT操作系统内部。

你一定已经很熟悉单字节字符。当你使用char时, 你处理的是单字节字符。双字节字符也用char类型来进行操作(这是我们将会看到的关于双字节字符的很多奇怪的地方之一)。Unicode字符用wchar_t来表示。Unicode字符和字符串常量用前缀L来表示。

文章标签： [单字节字符集](#) [多字节字符集](#) [Unicode](#)

个人分类：[纯编程](#)

此PDF由spygg生成,请尊重原作者版权!!!

我的邮箱:liushidc@163.com