零死角玩转STM32

LCD—液晶显示中英文

淘宝：firestm32.taobao.com

论坛： www.firebbs.cn

扫描进入淘宝店铺

主讲内容

01 字符编码

02 什么是字模？

03 各种模式的液晶显示字符实验

参考资料:《零死角玩转STM32》

“LCD—液晶显示中英文”章节

LCD—液晶显示中英文

Unicode字符集和编码

由于各个国家或地区都根据使用自己的文字系统制定标准，同一个编码在不同的标准里表示不一样的字符，各个标准互不兼容，而又没有一个标准能够囊括所有的字符，即无法用一个标准表达所有字符。国际标准化组织(ISO)为解决这一问题，它舍弃了地区性的方案，重新给全球上所有文化使用的字母和符号进行编号，对每个字符指定一个唯一的编号(ASCII中原有的字符编号不变)，这些字符的号码从0x000000到0x10FFFF，该编号集被称为Universal Multiple-Octet Coded Character Set，简称UCS，也被称为Unicode。最新版的Unicode标准还包含了表情符号(聊天软件中的部分emoji表情)，可访问Unicode官网了解：http://www.unicode.org。

Unicode字符集只是对字符进行编号，但具体怎么对每个字符进行编码，Unicode并没指定，因此也衍生出了如下几种unicode编码方案(Unicode Transformation Format)。

LCD—液晶显示中英文

UTF-32

对Unicode字符集编码，最自然的就是UTF-32方式了。编码时，它直接对Unicode字符集里的每个字符都用4字节来表示，转换方式很简单，直接将字符对应的编号数字转换为4字节的二进制数。

由于UTF-32把每个字符都用要4字节来存储，因此UTF-32不兼容ASCII编码，也就是说ASCII编码的文件用UTF-32标准来打开会成为乱码。

字符	GBK编码	Unicode编号	UTF-32编码
A	0x41	0x0000 0041	大端格式0x0000 0041
啊	0xB0A1	0x0000 554A	大端格式0x0000 554A

LCD—液晶显示中英文

UTF-32

对UTF-32数据进行解码的时候，以4个字节为单位进行解析即可，根据编码可直接找到Unicode字符集中对应编号的字符。

UTF-32的优点是编码简单，解码也很方便，读取编码的时候每次都直接读4个字节，不需要加其它的判断。它的缺点是浪费存储空间，大量常用字符的编号只需要2个字节就能表示。其次，在存储的时候需要指定字节顺序，是高位字节存储在前(大端格式)，还是低位字节存储在前(小端格式)。

LCD—液晶显示中英文

UTF-16

针对UTF-32的缺点，人们改进出了UTF-16的编码方式，它采用2字节或4字节的变长编码方式(UTF-32定长为4字节)。对Unicode字符编号在0到65535的统一用2个字节来表示，将每个字符的编号转换为2字节的二进制数，即从0x0000到0xFFFF。而由于Unicode字符集在0xD800-0xDBFF这个区间是没有表示任何字符的，所以UTF-16就利用这段空间，对Unicode中编号超出0xFFFF的字符，利用它们的编号做某种运算与该空间建立映射关系，从而利用该空间表示4字节扩展，感兴趣的读者可查阅相关资料了解具体的映射过程。

字符	GB18030编码	Unicode编号	UTF-16编码
A	0x41	0x0000 0041	大端格式0x0041
啊	0xB0A1	0x0000 554A	大端格式0x554A
崾	0x9735 F832	0x0002 75CC	大端格式0xD85D DDCC

注：崾 五笔：TLHH(不支持GB18030码的输入法无法找到该字，感兴趣可搜索它的Unicode编号找到)

LCD—液晶显示中英文

UTF-16

UTF-16解码时，按两个字节去读取，如果这两个字节不在0xD800到0xDFFF范围内，那就是双字节编码的字符，以双字节进行解析，找到对应编号的字符。如果这两个字节在0xD800到 0xDFFF之间，那它就是四字节编码的字符，以四字节进行解析，找到对应编号的字符。

UTF-16编码的优点是相对UTF-32节约了存储空间，缺点是仍不兼容ASCII码，仍有大小端格式问题。

LCD—液晶显示中英文

UTF-8

UTF-8是目前Unicode字符集中使用得最广的编码方式，目前大部分网页文件已使用UTF-8编码，如使用浏览器查看百度首页源文件，可以在前几行HTML代码中找到如下代码：

```
<meta http-equiv=Content-Type content="text/html;charset=utf-8">
```

其中“charset”等号后面的“utf-8”即表示该网页字符的编码方式UTF-8。

LCD—液晶显示中英文

UTF-8

UTF-8也是一种变长的编码方式，它的编码有1、2、3、4字节长度的方式，每个Unicode字符根据自己的编号范围去进行对应的编码。它的编码符合以下规律：

- 对于UTF-8单字节的编码，该字节的第1位设为0(从左边数起第1位，即最高位)，剩余的位用来写入字符的Unicode编号。即对于Unicode编号从0x0000 0000-0x0000 007F的字符，UTF-8编码只需要1个字节，因为这个范围Unicode编号的字符与ASCII码完全相同，所以UTF-8兼容了ASCII码表。

- 对于UTF-8使用N个字节的编码(N>1)，第一个字节的前N位设为1，第N+1位设为0，后面字节的前两位都设为10，这N个字节的其余空位填充该字符的Unicode编号，高位用0补足。

LCD—液晶显示中英文

UTF-8

Unicode(16进制)	UTF-8（2进制）				
编号范围	第一字节	第二字节	第三字节	第四字节	第五字节
00000000-0000007F	0xxxxxxx				
00000080-000007FF	110xxxxx	10xxxxxx			
00000800-0000FFFF	1110xxxx	10xxxxxx	10xxxxxx		
00010000-0010FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx	
…	111110xx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx

UTF-8解码的时候以字节为单位去看，如果第一个字节的bit位以0开头，那就是ASCII字符，以单字节进行解析。如果第一个字节的数据位以“110”开头，就按双字节进行解析，3、4字节的解析方法类似。

UTF-8的优点是兼容了ASCII码，节约空间，且没有字节顺序的问题，它直接根据第1个字节前面数据位中连续的1个数决定后面有多少个字节。不过使用UTF-8编码汉字平均需要3个字节，比GBK编码要多一个字节。

LCD—液晶显示中英文

BOM

由于UTF系列有多种编码方式，而且对于UTF-16和UTF-32还有大小端的区分，那么计算机软件在打开文档的时候到底应该用什么编码方式去解码呢？有的人就想到在文档最前面加标记，一种标记对应一种编码方式，这些标记就叫做BOM(Byte Order Mark)，它们位于文本文件的开头，见下表。注意BOM是对Unicode的几种编码而言的，ANSI编码没有BOM。

BOM标记	表示的编码
0xEF 0xBB 0xBF	UTF-8
0xFF 0xFE	UTF-16 小端格式
0xFE 0xFF	UTF-16 大端格式
0xFF 0xFE 0x00 0x00	UTF-32 小端格式
0x00 0x00 0xFE 0xFF	UTF-32 大端格式

由于带BOM的设计很多规范不兼容，不能跨平台，所以这种带BOM的设计没有流行起来。Linux系统下默认不带BOM。

THANKS

论坛： www.firebbs.cn

淘宝：firestm32.taobao.com

扫描进入淘宝店铺