链滴

字符集与编码

作者：helly
原文链接：https://ld246.com/article/1501379287049
来源网站：链滴

一、基本概念

1. charset（character set）：字符集

2. encoding（character encoding）：字符编码，简称**编码**

3. MBCS。。。

双字节字符集（DBCS，Double Byte Charecter Set），指一系列汉字字符集，包括GB 2312，GB，GB 18030，BIG5 等；

4. ANSI（American National Standard Institite），美国国家标准协会。各个国家（非拉丁语系家）自己制定自己的字符集，符合ANSI的标准（即兼容 ASCII 字符集，在此基础上扩展），得到ANS的认可，全世界在表示对应国家文字的时候都通用这种编码就叫ANSI编码。

ANSI问题：编码重叠问题--要同时显示多种语言，可能会出现一个编码两个字符集里面都有对，计算机无法判断该显示哪一个。

代码页

5. 简单编码模型

按照惯例，人们认为字符集和字符编码是同义词，因为使用同样的标准来定义提供什么字符并且些字符如何编码到一系列的代码单元（通常一个字符一个单元）。由于历史的原因，MIME和使用这编码的系统使用术语**字符集**来表示用于将一组字符编码成一系列八位字节数据的整个系统。

在简单编码模型里，一个字符集定义了这个字符集里包含什么字符，同时把每个字符在计算机中比特表示也进行了定义。例如 ASCII，在 ASCII 里直接定义了 A -> 0100 0001。

6. 现代编码模型

现代编码模型由统一码（Unicode）和通用字符集（UCS，Universal Character Set）构成。

现代编码模型自底向上分为五个层次：

（1）抽象字符表 ACR (Abstract Character Repertoire)

抽象字符表是一个系统支持的所有抽象字符的集合。

（2）编码字符集 CCS (Coded Character Set)

将抽象字符表中的每一个字符用一个非负整数表示。抽象字符表及映射的码位值称为编码字符集。

编码空间（encoding space）：包含所有字符的表的维度，例如 ASCII 的编码空间为 128。

码位（code point）：编码空间的一个位置。

码位值（code point value）：一个字符映射到编码空间的码位。

Unicode属于这一层。

（3）字符编码表 CEF (Character Encoding Form)

将编码字符集的非负整数值（即码位）转换成有限比特长度的整型值（称为**码元**code units）的列。

UTF-8、UTF-16、UTF-32等属于这一层。

（4）字符编码方案 CES (Character Encoding Schema)

（5）传输编码语法 TES (Transfer Encoding Syntax)

二、 各种字符集及其编码

1. ASCII 字符集

美国信息交换标准代码（ASCII，American Standard Code for Information Interchange），18 个字符（26 个拉丁字符，10 个阿拉伯数字，59 个英式标点符号，33 个无法显示的控制字符）用显示现代美国英语。

0 - 127 表示 128 个 ASCII 字符，每个编码占 7 位。

2. EASCII 字符集

延伸美国标准信息交换码（EASCII，Extended ASCII），256 个字符支持现代美国英语和部分欧语言。

0 - 255 表示 256 个 EASCII 字符，每个编码占 8 位。

3. ISO 8859 系列字符集

ISO 8859，全称 ISO/IEC 8859。

ISO 8859 字符集是一组字符集的总称，其下共包含了 15 个字符集，即 **ISO 8859-n**，其中 n=1,..,11,13,...,16。

（1）ISO 8859-1 字符集(Latin-1) - 西欧语言

（2）ISO 8859-2 字符集(Latin-2) - 中欧语言

.....

（11）ISO 8859-11 字符集(Thai) - 泰语，从泰国的 TIS620 标准字集演化而来

（13）ISO 8859-13 字符集（Latin-7 或 Baltic Rim） - 波罗的语族

.....

（16）ISO 8859-16 字符集(Latin-10) - 东南欧语言

4. GB 2312 字符集

GB 2312，中华人民共和国国家标准简体中文字符集，又称为 GB 2312-80 或 GB0。

GB2312 是对 ASCII 的中文扩展。

GB 2312 标准共收录 6763 个汉字，其中一级汉字 3755 个，二级汉字 3008 个；同时收录了包拉丁字母、希腊字母、日文平假名及片假名字母、俄语西里尔字母在内的 682 个字符。

GB 2312 的出现，基本满足了汉字的计算机处理需要，它所收录的汉字已经覆盖中国大陆 99.75 的使用频率。但对于人名、古汉语等方面出现的罕用字和繁体字，GB 2312 不能处理。

小于等于 127 的字符的意义与原来 ASCII 编码相同；

两个大于127的字符连在一起时，就表示一个汉字，前面的一个字节（称之为高字节）从0xA1用0xF7，后面一个字节（低字节）从0xA1到0xFE；

在 ASCII 里本来就有的数字、标点、字母都重新编了两个字节长的编码，这就是常说的”全角”符，而原来在127号以下的那些就叫”半角”字符了。

5. GBK 字符集

GBK，汉字内码扩展规范

GBK 向下完全兼容 GB 2312 编码。支持 GB 2312 编码不支持的部分中文姓，中文繁体，日文名，还包括希腊字母以及俄语字母等字母。不过这种编码不支持韩国字，也是其在实际使用中与 unicode 编码相比欠缺的部分。

向下兼容 GB 2312；

6. GB 18030 字符集

GB 18030，国家标准 GB 18030-2005《信息技术　中文编码字符集》。

增加了几千个新的少数民族的字。

向下兼容 GBK；

7. BIG5 字符集

Big5，又称为大五码或五大码，是使用繁体中文（正体中文）社区中最常用的电脑汉字字符集标，共收录13,060个汉字。

Big5虽普及于台湾、香港与澳门等繁体中文通行区。

8. Unicode 字符集

Unicode（中文：万国码、国际码、统一码、单一码）是计算机科学领域里的一项业界标准。它世界上大部分的文字系统进行了整理、编码，使得电脑可以用更为简单的方式来呈现和处理文字。

Unicode伴随着通用字符集的标准而发展，同时也以书本的形式对外发表。Unicode至今仍在不增修，每个新版本都加入更多新的字符。目前最新的版本为2017年6月20日公布的10.0.0。

（1）Unicode 编码

用两个字节，也就是 16 位来统一表示所有的字符，对于 ASCII 里的那些“半角”字符，Unicode 保持其原编码不变，只是将其长度由原来的 8 位扩展为 16 位，而其他文化和语言的字符则全部新统一编码。由于"半角"英文符号只需要用到低 8 位，所以其高 8 位永远是 0 。

问题：

保存英文文本时会多浪费一倍的空间；

16 位远远不够；

（2）UTF-8（8-bit Unicode Transformation Format）编码

编码规则：

0xxxxxxx

110xxxxx 10xxxxxx

1110xxxx 10xxxxxx 10xxxxxx

11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

(3) UTF-16 编码

编码规则：

(4) UTF-32 编码

CodePage

base64

参考：

维基百科

https://stackoverflow.com/questions/2281646/whats-the-difference-between-encoding-and-harset

https://my.oschina.net/goldenshaw/blog/304493

https://www.zhihu.com/question/27562173

https://yq.aliyun.com/articles/63036

https://zhuanlan.zhihu.com/p/19857727

http://blog.csdn.net/softman11/article/details/6124345