常见字符编码

字符集

■ 常用字符集分类：

➢ ASCII 及其扩展字符集

作用：表语英语及西欧语言。

位数：ASCII 使用 7 位表示，能表示 128 个字符；其扩展使用 8 位表示，表示 256 个字符。

范围：ASCII 从 00 到 7F，ASCI 扩展从 00 到 FF。

➢ ISO-8859-1 字符集

作用：扩展 ASCII，表示西欧、希腊语等。

位数：8 位

范围：从 00 到 FF，兼容 ASCII 字符集。

➢ GB2312 字符集

作用：国家简体中文字符集，兼容 ASCII。

位数：使用 2 个字节表示，能表示 7445 个符号，包括 6763 个汉字，几乎覆盖所有高频率汉字。

范围：高字节从 A1 到 F7，低字节从 A1 到 FE。将高字节和低字节分别加上 0XA0 即可得到编码。

➢ BIG5 字符集

作用：统一繁体字编码。

位数：使用 2 个字节表示，表示 13053 个汉字。

范围：高字节从 A1 到 F9，低字节从 40 到 7E，A1 到 FE。

➢ GBK 字符集

作用：它是 GB2312 的扩展，加入对繁体字的支持，兼容 GB2312。

位数：使用 2 个字节表示，可表示 21886 个字符。

范围：高字节从 81 到 FE，低字节从 40 到 FE。

- GB18030 字符集

作用：它解决了中文、日文、朝鲜语等的编码，兼容 GBK。

位数：它采用变字节表示(1 ASCII，2，4 字节)。可表示 27484 个文字。

范围：1 字节从 00 到 7F; 2 字节高字节从 81 到 FE，低字节从 40 到 7E 和 80 到 FE；4 字节第一三字节从 81 到 FE，第二四字节从 30 到 39。

UCS 字符集

- 作用：国际标准 ISO 10646 定义了通用字符集(Universal Character Set)。它是与 UNICODE 同类的组织，UCS-2 和 UNICODE 兼容。

位数：它有 UCS-2 和 UCS-4 两种格式，分别是 2 字节和 4 字节。

范围：目前，UCS-4 只是在 UCS-2 前面加了 0×0000。

- UNICODE 字符集

作用：为世界 650 种语言进行统一编码，兼容 ISO-8859-1。

位数：UNICODE 字符集有多个编码方式，分别是 UTF-8，UTF-16 和 UTF-32。

■ 按所表示的文字分类

语言	字符集	正式名称
英语、西欧语	ASCII，ISO-8859-1	MBCS 多字节
简体中文	GB2312	MBCS 多字节
繁体中文	BIG5	MBCS 多字节
简繁中文	GBK	MBCS 多字节

中文、日文及朝鲜语	GB18030	MBCS 多字节
各国语言	UNICODE，UCS	DBCS 宽字节

两种编码方式

- UTF-8：采用变长字节(1 ASCII, 2 希腊字母, 3 汉字, 4 平面符号) 表示，网络传输即使错了一个字节，不影响其他字节，而双字节只要一个错了，其他也错了。UTF-8 最多可用到 6 个字节。
- UTF-16：采用 2 字节，Unicode 中不同部分的字符都同样基于现有的标准。这是为了便于转换。从 0×0000 到 0×007F 是 ASCII 字符，从 0×0080 到 0×00FF 是 ISO-8859-1 对 ASCII 的扩展。希腊字母表使用从 0×0370 到 0×03FF 的代码，斯拉夫语使用从 0×0400 到 0×04FF 的代码，美国使用从 0×0530 到 0×058F 的代码，希伯来语使用从 0×0590 到 0×05FF 的代 码。中国、日本和韩国的象形文字（总称为 CJK）占用了从 0×3000 到 0×9FFF 的代码；由于 0×00 在 c 语言及操作系统文件名等中有特殊意义，故很多情况下需要 UTF-8 编码保存文本，去掉这个 0×00。

■ 优缺点：

- UTF-8、UTF-16 和 UTF-32 都可以表示有效编码空间 (U+000000-U+10FFFF) 内的所有 Unicode 字符。
- 使用 UTF-8 编码时 ASCII 字符只占 1 个字节，存储效率比较高，适用于拉丁字符较多的场合以节省空间。对于大多数非拉丁字符（如中文和日文）来说，UTF-16 所需存储空间最小，每个字符只占 2 个字节。
- Windows NT 内核是 Unicode（UTF-16），采用 UTF-16 编码在调用系统 API 时无需转换，处理速度也比较快。采用 UTF-16 和 UTF-32 会有 Big Endian 和 Little Endian

之分，而 UTF-8 则没有字节顺序问题，所以 UTF-8 适合传输和通信。

- UTF-32 采用 4 字节编码，一方面处理速度比较快，但另一方面也浪费了大量空间，影响输速度，因而很少使用。