链滴

字符集编码的意义?

作者：Alexs

原文链接：https://ld246.com/article/1518098539880

来源网站：链滴

何为编码，在维基百科上，查到的信息更好理解一点，在此贴上。
编码是信息从一种形式或格式转换为另一种形式的过程。解码，是编码的逆过程。
扩展定义
对于特定的上下文，编码有一些更具体的意义。
编码（Encoding）在认知上是解释传入的刺激的一种基本知觉的过程。技术上来说，这是一个复杂、多阶段的转换过程，从较为客观的感觉输入（例如光、声）到主观上有意义的体验。
字符编码（Character encoding）是一套法则，使用该法则能够对自然语言的字符的一个集合（如字表或音节表），与其他东西的一个集合（如号码或电脉冲）进行配对。
文字编码（Text encoding）使用一种标记语言来标记一篇文字的结构和其他特征，以方便计算机进处理。
语义编码（Semantics encoding），以正式语言乙对正式语言甲进行语义编码，即是使用语言乙表语言甲所有的词汇（如程序或说明）的一种方法。
电子编码（Electronic encoding）是将一个信号转换成为一个代码，这种代码是被优化过的以利于输或存储。转换工作通常由一个编解码器完成。
加密（Encryption）是为了保密而对信息进行转换的过程。
转码（Transcoding）是将已经编码的信息从一种格式转换到另一种格式的过程。
解码（Decoding）是编码的相反，亦即把编码过的信息恢复成原来样式。
而 UTF-8 和 GBK 则是字符编码集，将人熟悉的信息一一通过字符集比对上机器熟悉的二进制，像是找了个第三方，通过第三方代理来交流。在这中间出现了很多问题，所以相应的也有许多解决方。于是相应的字符编码集就被发明出来了。
ASCII（American Standard Code for Information Interchange，美国信息交换标准代码）是于拉丁字母的一套电脑编码系统。它主要用于显示现代英语，而其扩展版本 EASCII 则可以部分支持他西欧语言，并等同于国际标准 ISO/IEC 646。ASCII 码最初使用指定的 7 位二进制数组合来表示 128 种可能的字符，后来扩展到 8 位二进制以表示更多的字符，然而最多只能表示 256 中，还是太少。界上语言太多了，未收录的字符太多。为了满足种种需求各种各样的编码集被研发出来，可是没有一通用的标准，编码解码就显得很蠢了。一个游戏，在中国得支持中国的汉字编码，在美国得支持英文码，在拉丁国家需要支持拉丁编码。ISO （国际标谁化组织）的国际组织决定着手解决这个问题。他采用的方法很简单：废了所有的地区性编码方案，重新搞一个包括了地球上所有文化、所有字母和符的编码！他们打算叫它”Universal Multiple-Octet Coded Character Set”，简称 UCS, 俗称 “UN CODE”。
UNICODE 开始制订时，计算机的存储器容量极大地发展了，空间再也不成为问题了。于是 ISO 直接规定必须用两个字节，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，NICODE 包持其原编码不变，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部重新统一编码。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，因此这种大气方案在保存英文文本时会多浪费一倍的空间。
但是，UNICODE 在制订时没有考虑与任何一种现有的编码方案保持兼容，这使得 GBK 与 UNICODE 在汉字的内码编排上完全是不一样的，没有一种简单的算术方法可以把文本内容从 UNICODE 编码和一种编码进行转换，这种转换必须通过查表来进行。UNICODE 是用两个字节来表示为一个字符，他共可以组合出 65535 不同的字符，这大概已经可以覆盖世界上所有文化的符号。
UNICODE 来到时，一起到来的还有计算机网络的兴起，UNICODE 如何在网络上传输也是一个必须虑的问题，于是面向传输的众多 UTF（UCS Transfer Format）标准出现了，顾名思义，UTF8 就是次 8 个位传输数据，而 UTF16 就是每次 16 个位，只不过为了传输时的可靠性，从 UNICODE 到 UTF 时并不是直接的对应，而是要过一些算法和规则来转换。
简单的总结一下：
● 中国人民通过对 ASCII 编码的中文扩充改造，产生了 GB2312 编码，可以表示 6000 多个常汉字。
● 汉字实在是太多了，包括繁体和各种字符，于是产生了 GBK 编码，它包括了 GB2312 中的编码，时扩充了很多。
● 中国是个多民族国家，各个民族几乎都有自己独立的语言系统，为了表示那些字符，继续把 GBK 码扩充为 GB18030 编码。
● 每个国家都像中国一样，把自己的语言编码，于是出现了各种各样的编码，如果你不安装相应的编，就无法解释相应编码想表达的内容。
● 终于，有个叫 ISO 的组织看不下去了。他们一起创造了一种编码 UNICODE，这种编码非常大，到可以容纳世界上任何一个文字和标志。所以只要电脑上有 UNICODE 这种编码系统，无论是全球哪

文字，只需要保存文件的时候，保存成 UNICODE 编码就可以被其他电脑正常解释。
● UNICODE 在网络传输中，出现了两个标准 UTF-8 和 UTF-16，分别每次传输 8 个位和 16 个位。
于是就会有人产生疑问，UTF-8 既然能保存那么多文字、符号，为什么国内还有这么多使用 GBK 等码的人？因为 UTF-8 等编码体积比较大，占电脑空间比较多，如果面向的使用人群绝大部分都是中人，用 GBK 等编码也可以。但是目前的电脑来看，硬盘都是白菜价，电脑性能也已经足够无视这点能的消耗了。所以推荐所有的网页使用统一编码：UTF-8。
参考资料：
http://blog.csdn.net/dk0520/article/details/70157426
维基百科编码词条，ASCII 词条，utf-8 词条。