目录

[1.引言 2](#_Toc10445)

[2. Unicode字符集 3](#_top" \o "Unicode字符集)

[来历 3](#_Toc27921)

[产生原因 3](#_Toc21908)

[解决的问题 3](#_Toc13770)

[1.1 UTF8 3](#_Toc15062)

[字节数 3](#_Toc5657)

[描述 4](#_Toc20399)

[兼容的字符集 4](#_Toc20305)

[1.2 UTF16 4](#_Toc8658)

[字节数 4](#_Toc5657)

[描述 4](#_Toc20399)

[兼容的字符集 4](#_Toc20305)

[3.GB2312 5](#_Toc5359)

[来历 5](#_Toc25625)

[产生原因 5](#_Toc23310)

[解决的问题 5](#_Toc28418)

[字节数 5](#_Toc5657)

[描述 5](#_Toc20399)

[兼容的字符集 5](#_Toc20305)

[4. big5 6](#_Toc24688)

[来历 6](#_Toc32115)

[产生原因 6](#_Toc29689)

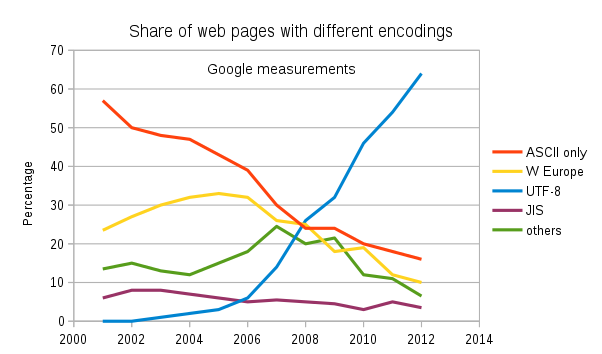
[解决的问题 6](#_Toc11174)

[字节数 6](#_Toc26387)

[描述 6](#_Toc12517)

5.参考文献 [7](#_Toc12517)

1. 引言

本文的目的在于区分和辨别几种涵盖中文字符集的主流编码集，以及它们的来历，产生原因，兼容性和支持的语言等等详细信息。其中，GB2312和big5两种字符集均在Unicode之前出现的，旨在打破各地区编码不统一或者没有相应文字的编码；而Unicode出现后，首先UTF-8能完美兼容ASCⅡ码，并且英文字符只占一个字节，中文字符占三个字节，这样既能实现与旧程序的融合，还能节省不同语言所需要的空间，UTF-16则是所有字符均用2个字节表示，没有UTF-8所具有的兼容性和节省空间，因此UTF-8更受欢迎。

这张来自谷歌统计的网站使用编码所占百分比中我们可以看出，UTF-8的确成为了各大网站所用的编码风向标。

1. Unicode字符集

来历：Unicode还未被创造出之前，各个国家都像中国这样搞出一套自己的编码标准，结果互相之间谁也不懂谁的编码，谁也不支持别人的编码，连大陆和台湾这样只相隔了150海里，使用着同一种语言的兄弟地区，也分别采用了不同的 DBCS 编码方案。正在这时，一个叫 ISO （国际标谁化组织）的国际组织决定着手解决这个问题。他们采用的方法很简单：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、所有字母和符号的编码。

产生原因：大部分国家都形成了适用于本国文字的编码规范。这样造成了国际上计算机语言不能互通的局面，需要一种可以表示所有语言的编码规范。Unicode诞生了，1990年开始设计，1994年正式发布。Unicode可以表示1114112个字符。

解决的问题：解决了国际上计算机语言不能互通的局面。

1.1 UTF8

字节数：UTF-8是一种使用1到8字节能够编码Unicode中所有1112064个有效码点的字符编码。

● 占1个字节的有：英文等ASCⅡ中的127个字符；

● 占2个字节的有：带附加符号的拉丁文、希腊文、西里尔字母、亚美尼亚语、希伯来文、阿拉伯文、叙利亚文；

● 占3个字节的有：含有21000多个汉字；

● 占4个字节的有：中日韩超大字符集里的汉字，有5万多个

描述：UTF-8的好处是兼容性，很好地兼容了ASCⅡ,并且支持全世界国家所需要用的字符，是国际编码，通用性强。如果外国人浏览中国网站，网站用的是UTF-8编码，那么外国人的英文IE浏览器上也能显示中文，不需要另外下载中文语言支持包。

兼容的字符集：UNICODE支持欧洲、非洲、中东、亚洲（包括统一标准的东亚象形汉字和韩国表音文字）。但是，UNICODE并没有提供对诸如Braille,Cherokee, Ethiopic, Khmer, Mongolian, Hmong, Tai Lu, Tai Mau文字的支持。同时它也不支持如Ahom, Akkadian, Aramaic, BabylonianCuneiform, Balti, Brahmi, Etruscan, Hittite, Javanese, Numidian, Old Persian Cuneiform, Syrian之类的古老文字。

1.2 UTF16

字节数：所有字符均用2个字节表示

描述：UTF-16是[Unicode](https://baike.baidu.com/item/Unicode" \t "https://baike.baidu.com/item/UTF-16/_blank)的其中一个使用方式。在Unicode[基本多文种平面](https://baike.baidu.com/item/%E5%9F%BA%E6%9C%AC%E5%A4%9A%E6%96%87%E7%A7%8D%E5%B9%B3%E9%9D%A2" \t "https://baike.baidu.com/item/UTF-16/_blank)定义的[字符](https://baike.baidu.com/item/%E5%AD%97%E7%AC%A6" \t "https://baike.baidu.com/item/UTF-16/_blank)（无论是[拉丁字母](https://baike.baidu.com/item/%E6%8B%89%E4%B8%81%E5%AD%97%E6%AF%8D" \t "https://baike.baidu.com/item/UTF-16/_blank)、汉字或其他文字或符号），一律使用2[字节](https://baike.baidu.com/item/%E5%AD%97%E8%8A%82" \t "https://baike.baidu.com/item/UTF-16/_blank)储存。而在辅助平面定义的字符，会以代理对的形式，以两个2字节的值来储存。

UTF-16比起[UTF-8](https://baike.baidu.com/item/UTF-8" \t "https://baike.baidu.com/item/UTF-16/_blank)，好处在于大部分字符都以固定长度的字节 (2字节) 储存，但UTF-16 却无法兼容于[ASCII](https://baike.baidu.com/item/ASCII" \t "https://baike.baidu.com/item/UTF-16/_blank)编码。

兼容的字符集：与UTF-8一致

3.GB2312

来历：最开始计算机诞生后，1967年美国设计了一种涵盖英语世界文字的编码规范，后来经过改进，并逐步成为美国文字编码的统一标准，即ASCCII码。ASCII码使用一个字节（8位）中的后7位，最高位做奇偶校验，用于校验网络传输中数据的完整性。这样ASCII码可以表示128个字符，包括数字、大小写字母、控制字符、和美国人常用的字符。后来传到欧洲，ASCII码不能满足需求，如法语会在字母的上方加声调。这样法语将做奇偶校验的那一位用于表示字符，这样可以表示256个字符。但这种规则在俄语、波兰语等语言中就不适用了，这些语言就采用两个字节（16位）编码，可以表示65536个字符。

       在20世纪80年代计算机进入中国，首先需将计算机语言本土化，我国计算机工作者经过努力，1981年制定了汉字的编码标准，即GB2312（\*\*），涵盖了绝大多数的常用汉字。该标准采用两个字节编码，故最多只能表示65536个汉字，而汉字有十万多个。这就是GB2312的局限所在。

产生原因：没有可以显示汉字的字符集，于是设计出这套简体中文的字符集编码方案。

解决的问题：使得计算机能够显示汉字

字节数：汉字占2个字节，英文字母占1个字节

描述：在使用过程中，很快便发现了，有些汉字没法打印出来的问题，虽然GB2312中没用到的码位可以填充一部分，但是后来仍然不够用，就将127之后的编码往后扩展，扩展后的字符集叫做GBK，在往后的使用中还加入了少数民族的文字，就成了GB18030。

兼容的字符集：ASCⅡ,简化汉字及一般符号、序号、数字、拉丁字母、日文假名、希腊字母、俄文字母、汉语拼音符号、汉语注音字母，共7445个图形字符。

4.big5

来历：在台湾、香港与澳门地区，使用的是繁体中文字符集。而1980年发布的GB2312面向简体中文字符集，并不支持繁体汉字。在这些使用繁体中文字符集的地区，一度出现过很多不同厂商提出的字符集编码，这些编码彼此互不兼容，造成了信息交流的困难。为统一繁体字符集编码，1984年，台湾五大厂商宏碁、神通、佳佳、零壹以及大众一同制定了一种繁体中文编码方案，因其来源被称为五大码，英文写作Big5，后来按英文翻译回汉字后，普遍被称为大五码。

产生原因：1980年问世的GB2312仅仅只支持中文，而港澳台地区使用的繁体中文没有对应的字符集。

解决的问题： 解决了没有繁体字符集编码，统一了其他地区不同的繁体字符集编码。

字节数：中文字符占2个字节

描述：大五码是一种繁体中文汉字字符集，其中繁体汉字13053个，808个标点符号、希腊字母及特殊符号。大五码的编码码表直接针对存储而设计，每个字符统一使用两个字节存储表示。第1字节范围81H－FEH，避开了同ASCII码的冲突，第2字节范围是40H－7EH和A1H－FEH。因为Big5的字符编码范围同GB2312字符的存储码范围存在冲突，所以在同一正文不能对两种字符集的字符同时支持。尽管Big5码内包含一万多个字符，但是没有考虑社会上流通的人名、地名用字、方言用字、化学及生物科等用字，没有包含日文平假名及片假名字母。

兼容的字符集：支持繁体中文

1. 参考文献

**ITeye社区（http://www.iteye.com/）**

**百度百科（[https://baike.baidu.com/](http://www.baidu.com/link?url=E4T9l9RE7JqLRbkSw7JivlmHWbveUNI6T4065ugauELW1jvMjhlrm9ITBHtLG-iK" \t "https://www.baidu.com/_blank)）**

**CSDN（http://www.csdn.net/）**

**知乎（http://www.zhihu.com/）**

**百度知道（[https://zhidao.baidu.com/](http://www.baidu.com/link?url=E4T9l9RE7JqLRbkSw7JivlmHWbveUNI6T4065ugauELW1jvMjhlrm9ITBHtLG-iK" \t "https://www.baidu.com/_blank)）**

**维基百科（https://en.wikipedia.org/）**