GB

中国国家标准汇编

2010年　修订-14

中国质检出版社
中国标准出版社

中 国 国 家 标 准 汇 编

2010 年修订-14

中国标准出版社　编

中国质检出版社
中国标准出版社
北　京

图书在版编目（CIP）数据

中国国家标准汇编：2010年修订．14/中国标准出版社编．—北京：中国标准出版社，2011

ISBN 978-7-5066-6509-4

Ⅰ．①中… Ⅱ．①中… Ⅲ．①国家标准-汇编-中国-2009 Ⅳ．①T-652.1

中国版本图书馆CIP数据核字(2011)第187725号

中国质检出版社
中国标准出版社 出版发行

北京市朝阳区和平里西街甲2号(100013)
北京市西城区三里河北街16号(100045)

网址：www.spc.net.cn
总编室：(010)64275323 发行中心：(010)51780235
读者服务部：(010)68523946

中国标准出版社秦皇岛印刷厂印刷
各地新华书店经销

*

开本 880×1230 1/16 印张 92.5 字数 2 800 千字
2011年12月第一版 2011年12月第一次印刷

*

定价 390.00 元

出 版 说 明

1.《中国国家标准汇编》是一部大型综合性国家标准全集。自1983年起，按国家标准顺序号以精装本、平装本两种装帧形式陆续分册汇编出版。它在一定程度上反映了我国建国以来标准化事业发展的基本情况和主要成就，是各级标准化管理机构，工矿企事业单位，农林牧副渔系统，科研、设计、教学等部门必不可少的工具书。

2.《中国国家标准汇编》收入我国每年正式发布的全部国家标准，分为“制定”卷和“修订”卷两种编辑版本。

“制定”卷收入上一年度我国发布的、新制定的国家标准，顺延前年度标准编号分成若干分册，封面和书脊上注明“20××年制定”字样及分册号，分册号一直连续。各分册中的标准是按照标准编号顺序连续排列的，如有标准顺序号缺号的，除特殊情况注明外，暂为空号。

“修订”卷收入上一年度我国发布的、修订的国家标准，视篇幅分设若干分册，但与“制定”卷分册号无关联，仅在封面和书脊上注明“20××年修订-1，-2，-3，……”字样。“修订”卷各分册中的标准，仍按标准编号顺序排列（但不连续）；如有遗漏的，均在当年最后一分册中补齐。需提请读者注意的是，个别非顺延前年度标准编号的新制定的国家标准没有收入在“制定”卷中，而是收入在“修订”卷中。

读者配套购买《中国国家标准汇编》“制定”卷和“修订”卷则可收齐上一年度我国制定和修订的全部国家标准。

3.由于读者需求的变化，自1996年起，《中国国家标准汇编》仅出版精装本。

4.2010年我国制修订国家标准共2846项。本分册为“2010年修订-14”，收入新制修订的国家标准1项。

中国标准出版社

2011年8月

目　　录

ICS 35.040
L 71

中华人民共和国国家标准

GB 13000—2010/ISO/IEC 10646:2003
代替 GB 13000.1—1993

信息技术 通用多八位编码字符集(UCS)

Information technology—Universal multiple-octet coded character set (UCS)

(ISO/IEC 10646:2003,IDT)

2011-01-10 发布　　2011-11-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

前　言

本标准的体系结构与基本多文种平面部分是强制性的，其余为推荐性的。

本标准等同采用国际标准 ISO/IEC 10646:2003《信息技术　通用多八位编码字符集(UCS)》(英文版)。为便于读者理解，本标准在第 1 章增加了脚注，在附录 S(资料性附录)的最后增加了“S.4 对‘CJK 汉字认同和排序规则’的补充说明”。

本标准代替 GB 13000.1—1993《信息技术　通用多八位编码字符集(UCS)　第一部分：体系结构与基本多文种平面》。

本次修订对 1993 年版的主要变动如下：

a) 朝鲜文及其补充从基本多文种平面的 3400～4DFF 移至基本多文种平面的 AC00～D7FF(原 O 区，即第一版的保留区)，空出的代码位置分配给新增的 CJK 统一汉字扩充 A；

b) 增加收录多种我国少数民族文字及其他文字、字符，如藏文、蒙古文、彝文等；

c) 新增用于 UTF-16 的 S 区(代理区)，代码位置是基本多文种平面的 D800～DFFF(原 O 区，即第一版的保留区)，并有专门的附录对其进行说明；

d) 增加了辅助平面，包括 00 组 01 平面(文字和符号辅助多文种平面)、00 组 02 平面(辅助表意文字平面，用于 CJK 统一汉字扩充 B 和 CJK 兼容汉字补充)和 00 组 0E 平面(辅助特殊用途平面)。

本标准的附录 A、附录 B、附录 C 和附录 D 是规范性附录，附录 E、附录 F、附录 G、附录 H、附录 J、附录 K、附录 L、附录 M、附录 N、附录 P、附录 Q、附录 R、附录 S、附录 T 和附录 U 是资料性附录。

本标准由中华人民共和国工业和信息化部提出。

本标准由全国信息技术标准化技术委员会(SAC/TC 28)归口。

本标准起草单位：中国电子信息产业发展研究院、中国电子技术标准化研究所、教育部语言文字应用研究所、中国科学院软件研究所、北京北大方正电子有限公司。

本标准主要起草人：张轴材、陈壮、王晓明、吴健、尹江红、何正安。

本标准于 1993 年首次发布，本次为第一次修订。

引　言

本标准规定了通用多八位编码字符集(UCS)。它适用于世界上各种语言(文字)的书面形式以及附加符号的表示、传输、交换、处理、存储、输入及显现。

通过对多文种文本的编码的一致性的定义,本标准使得数据的国际交换成为可能。信息技术产业获得了数据的稳定性,更强的全球可互操作性和数据可交换性。转化为本国家标准的国际标准 ISO/IEC 10646 已经在新的互联网协议中广泛采用,并被当今的操作系统和计算机语言所实现。本标准收纳了 95 000 多个世界上多种文字的字符。

本标准包含了一些电子化资料,它们适用于使用机读格式数据的用户。这些资料由下列可打印文件组成:

——CJKU_SR. txt

——CJKC_SR. txt

——Allnames. txt

——HangulX. txt

——HangulSy. txt

用户可向本标准归口单位(中国电子技术标准化研究所)索取上述电子化资料。

地址:北京市东城区安定门东大街 1 号(北京市 1101 信箱)

邮编:100007

电话:84043004

信息技术　通用多八位编码字符集(UCS)

1　范围

本标准规定了通用多八位编码字符集(UCS)。

本标准适用于世界上各种语言(文字)的书面形式以及附加符号的表示、传输、交换、处理、存储、输入及显现。

本标准：

——规定了 GB 13000 的体系结构；

——定义了 GB 13000 中使用的术语；

——描述了本编码字符集的总体结构；

——规定了 UCS 的基本多文种平面(BMP)；

——规定了 UCS 的若干个辅助平面：辅助多文种平面(SMP)、辅助表意文字平面(SIP)，以及辅助特殊用途平面(SSP)；

——定义了一个图形字符集，用于世界各种语言的手写和书面形式；

——规定了 BMP、SMP、SIP、SSP 的图形字符的名称及编码表示；

——规定了 UCS 的肆八位(32 位)正则形式：UCS-4；

——规定了 UCS 的双八位(16 位)BMP 形式：UCS-2；

——规定了控制功能的编码表示；

——规定了未来对编码字符集进行增补的管理办法。

UCS 是一种与 GB/T 2311 规定的编码体系不同的编码体系。本标准 16.2 规定了从 GB/T 2311 中指明 UCS 的方法。

本标准中任意一个图形字符，无论是在 BMP 平面还是在辅助平面，只分配唯一的一个码位[1)]。

注：Unicode Standard 的 4.0 版包括了与本标准等同的字符集、名称和编码表示。为了便于实现，它还提供了字符属性、处理算法以及定义的细节。

2　符合性

2.1　总则

专用字符一旦按本标准规定的方法使用，这些专用字符本身就不再受下列符合性要求的约束。

2.2　信息交换的符合性

如在编码信息内用于交换的编码字符数据元素符合以下条件，则符合本标准：

a)　在该编码字符数据元素内的全部图形字符的编码表示都符合第 6 章和第 7 章，及选自第 13 章或附录 C 或附录 D 的一种已标识的形式，并且还符合选自第 14 章的一种已标识的实现级别；

b)　在该编码字符数据元素内所表示的全部图形字符都来源于一个已标识的子集(见第 12 章)；

c)　在该编码字符数据元素内的全部控制功能的编码表示都符合第 15 章。

符合性声明必须标识出所采用的形式、所采用的实现级别以及所采用的以汇集清单和(或)字符清

1)　实际上，上述“一字一码”的原则在本标准的下列情形中，作了特别处理：

a)　不同文种的同形字符是分别编码的，比如拉丁 A，希里尔文 A 和希腊文 A，就是分别编码的；

b)　CJK 同一源字集中的同形字符，可能在 CJK 统一汉字编码区和 CJK 兼容汉字编码区中分别编码；

c)　CJK 统一汉字中可做部首的汉字，可能在 Kangxi Radical 或 CJK Radical Supplement 中有同形的编码部首。

单给出的子集。

2.3 设备的符合性

如果一台设备符合下列 a)项的要求,且符合 b)项及 c)项二者之一或其全部要求,则称该设备符合本标准。

注 1:"设备"这一术语(在 4.18 中)被定义为信息处理装备中的部件,它可以传送和(或)接收在编码字符数据元素内的编码信息。设备可以指常规意义上的输入/输出设备,也可指应用程序或网关功能等进程。

符合性声明必须标识出一个含有在下面 a)项中规定的描述的文档,并且必须标识出所采用的形式、所采用的实现级别、所采用的(以汇集清单和/或字符)清单给出的子集以及依据第 15 章所采用的控制功能。

a) 设备描述:符合本标准的设备应是一种描述的对象。所谓描述,就是像下列 b)和 c)项所规定的那样,标识用户给设备提供字符的手段,和(或)用户接受到这些字符后的辨识方式。

b) 始发设备:始发设备必须允许用户提供来自所采用的子集中的任意字符,并且能够依据所采用的形式及实现级别传送编码字符数据元素内的这些字符的编码表示。

c) 接收设备:接收设备必须能够依据所采用的形式及实现级别,接收并解释编码字符数据元素内的任何字符的编码表示,并且必须使得来自所采用的子集中的任何相应字符以用户能识别的方式提供给用户。

对于采用的子集中没有的任何相应字符,应以某种方式向用户提示,但不必区分这些字符。

注 2:可通过两种方式向用户提示:用同一个字符来表示所采用的子集中不具备的字符;或者,对某类用户合适时,提供一种能鉴别的有声信号或可视信号。

注 3:关于具有再传输能力的接收设备,可参见附录 J。

3 规范性引用文件

下列文件中的条款通过本标准的引用而成为本标准的条款。凡是注日期的引用文件,其随后所有的修改单(不包括勘误的内容)或修订版均不适用于本标准,然而,鼓励根据本标准达成协议的各方研究是否可使用这些文件的最新版本。凡是不注日期的引用文件,其最新版本适用于本标准。

GB/T 2311—2000 信息技术 字符代码结构与扩充技术(idt ISO/IEC 2022:1994)

GB/T 5261—1994 信息技术 七位和八位编码字符集用的控制功能(eqv ISO/IEC 6429:1988)

Unicode Standard Annex,UAX#9,The Unicode Bidirectional Algorithm,Version 4.0.0,2003-04-17

Unicode Standard Annex,UAX#15,Unicode Normalization Forms,Version 4.0.0,2003-04-17

4 术语和定义

下列术语和定义适用于本标准。

4.1

基本多文种平面 basic multilingual plane;BMP

00 组的 00 平面。

4.2

块 block

具有共同特征的一组字符(诸如某种文字)在连续的编码区域中进行编码。块与块之间不互相重叠。块之中的一个或多个代码位置可能不安排字符。

4.3

正则形式 canonical form

规定本编码字符集中字符的形式:用 4 个八位表示每一个字符。

4.4

编码字符数据元素　CC-data-element;coded-character-data element

被交换信息的一个元素，由依据一个或多个已标识的编码字符集标准的字符的编码表示的序列组成。

4.5

字位　cell

在一行中可安排一个字符的位置。

4.6

字符　character

供组织、控制或表示数据用的元素汇集中的一个元素。

4.7

字符边界　character boundary

在八位流中，某一字符的编码表示中的最后一个八位与其下一个字符的编码表示中的第一个八位之间的分界。

4.8

编码字符　coded character

字符及其编码表示。

4.9

编码字符集　coded character set

一组无歧义的规则，用以建立一个字符集和该字符集中的字符及其编码表示之间的对应关系。

4.10

代码表　code table

示出一种代码中分配给各八位的诸字符的表。

4.11

汇集　collection

被编号、命名并且由其码位在一个或多个确定的范围内编码字符组成的一组编码字符。

注：如果一个确定的范围中包含没有分配字符的码位，在本标准将来的改动中一旦有附加的字符被分配那些位置，那么该汇集的代码表就会变化。但该汇集的编号和名称在本标准的以后版本中将保持不变。

4.12

组合用字符　combining character

本标准编码字符集中一个已标识的子集的一个字符，用于与其前导的非组合用图形字符相组合，或者与一个以非组合用字符为前导的组合用字符序列相组合(见 4.14)。

注：本标准的这一部分规定了包含组合用字符的若干个子集汇集。

4.13

兼容字符　compatibility character

主要为与现存编码字符集兼容而作为本标准的编码字符收入的图形字符。

4.14

复合序列　composite sequence

由一个非组合用字符后随一个或多个组合用字符所组成的图形字符的序列(见 4.12)。

注 1：用于复合序列的图形符号一般是由该序列中的每个字符的图形符号的组合而构成的。

注 2：复合序列不是字符，因此也不是本标准字汇中的结构要素。

4.15

控制功能　control function

影响数据的记录、处理、传输或解释的一种动作，其编码表示由一个或多个八位组成。

4.16

默认状态 default state

在未明显地指定状态时所假设的状态。

4.17

详细代码表 detailed code table

示出一个个字符的代码表，并且通常示出一行的局部。

4.18

设备 device

信息处理装备中的部件，能发送和/或接收编码字符数据元素内的编码信息。（它可以是常规意义上的输入/输出设备，也可以是诸如应用程序或网关功能那样的进程。）

4.19

固定汇集 fixed collection

在确定区域中每个码位都分配一个字符，并在本标准的以后版本中保持不变的汇集。

4.20

图形字符 graphic character

不同于控制功能的字符，通常具有书写、打印或显示的可视表示。

4.21

图形符号 graphic symbol

图形字符或复合序列的可视表示。

4.22

组 group

本编码字符集编码空间的一个子单位，具有 256×256×256 个字位。

4.23

高半区 high-half zone

为 UTF-16 的应用保留的单元（见附录 C）；与这些单元中的任一单元相对应的 RC 元素，都可能在 UTF-16 中用作表示一个非 BMP 平面中的字符的 RC 元素对中的第一个。

4.24

交换 interchange

采用通信手段或可交换的媒体把字符编码数据从一个用户传送到另一个用户。

4.25

交互运作 interworking

允许两个或两个以上采用不同编码字符集的系统能含义确切地交换字符编码数据的一种过程，其中可能涉及两种代码之间的转换。

4.26

GB 13000.1/ISO/IEC 10646-1

本标准较早的一个版本。它也被称作本标准的第 1 部分，包含了基本多文种平面（BMP）和整体体系结构的规格说明。

4.27

ISO/IEC 10646-2

本标准对应的国际标准 ISO/IEC 10646 曾经有过的一个组成部分。它也被称作 ISO/IEC 10646 的第 2 部分，包含了对辅助多文种平面（SMP）、辅助表意平面（SIP）和辅助特殊用途平面（SSP）的规格说明。ISO/IEC 10646-2 仅发布过第一版。

4.28

低半区 low-half zone

为 UTF-16 的应用保留的一组单元(见附录 C);与这些单元中的任一单元相对应的 RC 元素,都可能在 UTF-16 中用作表示一个区别于 BMP 的平面中的字符的 RC 元素对中的第 2 个。

4.29

八位 octet

被看作一个单元的一个有序的八个位(比特)的序列。

4.30

平面 plane

一个组的一个子单位,具有 256×256 单元。

4.31

显现 presentation;to present

书写、打印或显示一个图形符号的过程。

4.32

显示形式 presentation form

在某些文字的显现中,表示一个字符的某种图形符号形式,这种形式依赖于该字符相对于其他字符的位置。

4.33

专用平面 private use plane

本编码字符集中的平面,其内容不由 GB 13000 规定(见第 10 章)。

4.34

RC 元素 RC-element

由对应本编码字符集编码空间中一个单元的四八位序列(正则形式)中的 R-octet 行八位和 C-octet 字位八位(见 6.2)组成的双八位序列。

4.35

字汇 repertoire

在一个编码字符集中描述的一个指定的字符汇集。

4.36

行 row

一个平面的一个子单位,具有 256 个字位。

4.37

文字 script

用于一种或多种语言书面形式的图形字符的一个汇集。

4.38

辅助平面 supplementary plane

除 00 组 00 平面之外的平面,容纳未安排在基本多文种平面中的那些字符的一个平面。

4.39

文字和符号辅助多文种平面 supplementary multilingual plane for scripts and symbols;SMP

00 组 01 平面。

4.40

辅助表意文字平面 supplementary ideographic plane;SIP

00 组 02 平面。

4.41

辅助特殊用途平面　supplementary special-purpose plane;SSP

00 组 0E 平面。

4.42

不成对的 RC 元素　Unpaired RC-element

一个编码字符数据元素中的这样一个 RC-element,或者是

——高半区中的一个 RC 元素,其后没有紧跟一个低半区中的 RC 元素,或者是

——低半区中的一个 RC 元素,其后没有紧随一个高半区中的 RC 元素。

4.43

用户　user

享用由设备提供的服务的个人或其他实体。(例如,若该"设备"是代码转换器或网关功能,则用户实体可以是诸如应用程序这样的进程。)

4.44

区　zone

代码表中字位的一个序列,由包含一个特定类别字符的一行或多行(整行或部分行)组成(举例见第 8 章)。

5　UCS 的总体结构

本章描述通用多八位编码字符集(以下简称"本编码字符集")的总体结构,并在图 1 和图 2 中加以说明。此结构的规范在后面的各章中给出。

在本标准中,任何一个八位的值均由从 00 到 FF 的十六进制记数法表示(见附录 K)。

本编码字符集的正则形式——它的表达方式——使用了一个被视为单一实体并由 128 个三维组构成的四维编码空间。

注 1:因此,在符合标准编码字符数据元素内,只要编码字符的正则形式中最高八位(octet)的第 8 位值为 0,这一位就可用于设备的内部处理。

每个组包含 256 个二维平面。每个平面包含 256 个一维行,每个行包含 256 个字位。一个字符在这个编码空间的一个字位上进行编码,否则,该字位声明未被使用。

在正则形式中,用 4 个八位来表示每一个字符,并相应地指定组、平面、行和字位。由于双八位不足以表示世界上所有的字符,而肆八位的表示符合现代处理系统的体系结构,所以正则形式由 32 位组成。

肆八位正则形式可用作肆八位编码字符集,此时称它为 UCS-4。

注 2:对于这种形式,术语"正则"的使用并不意味着任何限制,也不意味着这种形式优于其他可以用于表示 UCS 字符的变换格式,这些变换格式可能被遵从本标准的实现所选用。

本标准规定了下列平面中的图形字符和它们的代码表示:

——基本多文种平面(BMP,00 组 00 平面)。基本多文种平面可以用于标示为 UCS-2 的 16 位编码字符集。

——用于文字和符号的辅助多文种平面(SMP,00 组 01 平面)。

——辅助表意文字平面(SIP,00 组 02 平面)。

——辅助特殊用途平面(SSP,00 组 0E 平面)。

在将来可能定义附加的辅助平面,用以容纳更多的图形字符。

为专门用途保留的平面在第 10 章中详细说明。专用平面和区中的字位所对应字符不在本标准中说明。

每一个字符均按照其组八位、平面八位、行八位、字位八位安排在本编码字符集中。

为给出图形字符的子字汇,可使用编码空间中的子集。

在附录C中规定了UCS的一种变换格式(UTF-16),它可用于表示00组的16个辅助平面(从01平面到10平面)中的字符,除BMP(00平面)外,在某种形式上与双八位BMP形式兼容。

附录D规定了UCS的另一种变换格式(UTF-8),它可用在对按照GB/T 2311和GB/T 11383的8比特结构进行编码的,且对八位值敏感的控制字符的通信系统中传输文本数据。根据GB/T 11383,UTF-8还避免使用在广泛使用的文件处理系统的文件名字符串的解析中具有特殊意义的八位值。

6 基本结构及术语

6.1 结构

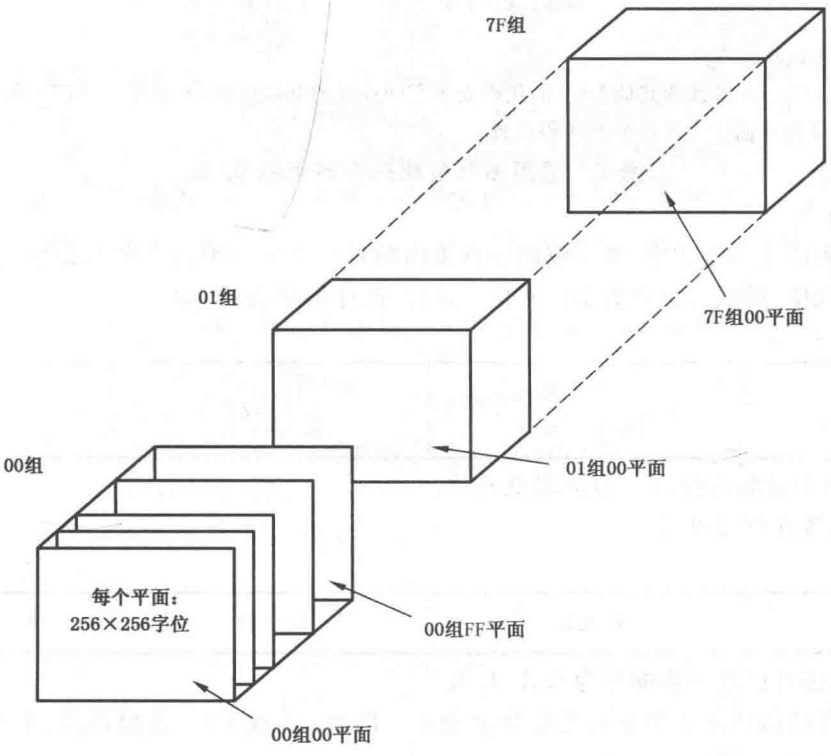
本标准规定的通用多八位编码字符集应被看作一个单一的实体。

整个编码字符集应被表达为包含128个组,其中每组有256个平面。每一平面应被视为含有256行字符,每行256个字位。在表示平面内容(如图2)的代码表中,水平轴应表示最低八位,越向左其值越小;而纵轴应表示较高八位,越向上其值越小。

编码空间中每一轴线应由一个8位进行编码。在每一个8位中最高位应为第8位,最低位应为第1位。

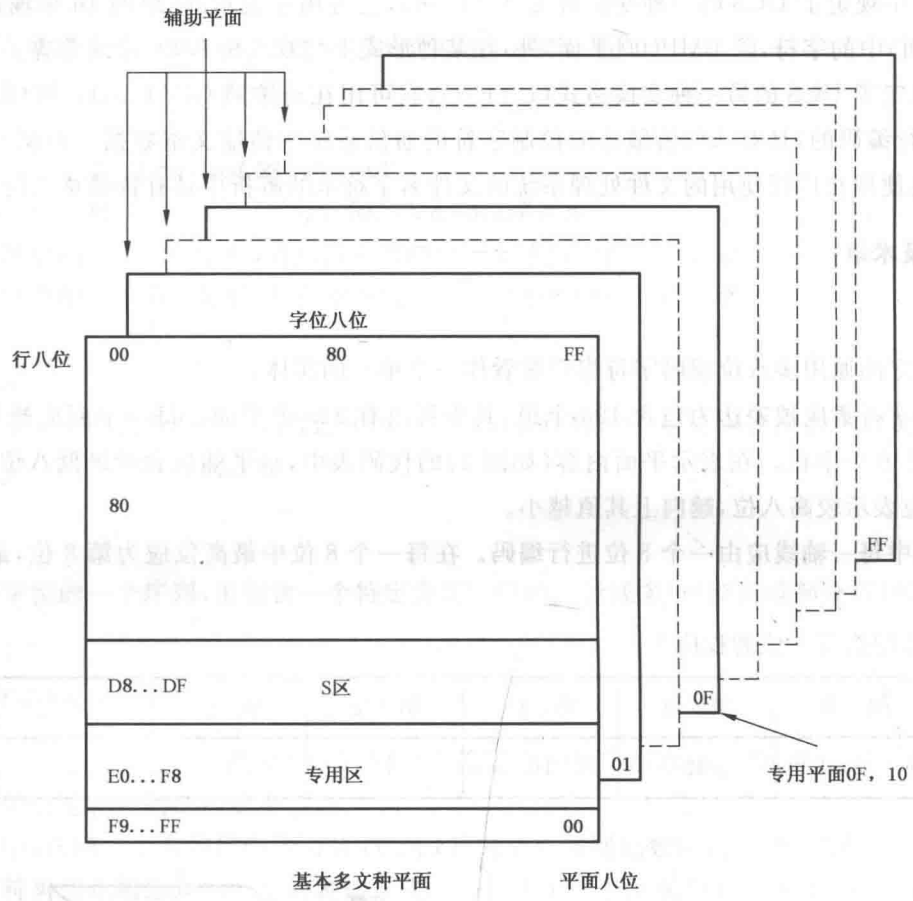
相应地,分配给每一位的权应为:

第8位	第7位	第6位	第5位	第4位	第3位	第2位	第1位
128	64	32	16	8	4	2	1



注:为了确保UTF-16形式和UCS的其他代码表示之间连续的互操作性,不会在00组11到FF平面中或在任何其他组的任何平面中为字符分配代码位置。

图1 通用多八位编码字符集的全部编码空间



注 1：S-区和专用区在第 8 章中规定。

注 2：为了确保 UTF-16 形式和其他 UCS 的代码表示之间连续的互操作性，不会在 00 组 11 到 FF 平面中或在任何其他组的任何平面中为字符分配代码位置。

图 2　通用多八位编码字符集的 00 组

6.2　字符的编码

在编码字符集的正则形式中，整个编码字符集内的每一个字符须由 4 个八位序列表示。该序列的最高八位应为组八位，最低八位应为字位八位。这样，此序列可表示为：

m. s.　　l. s.

组八位 (Group-octec)	平面八位 (Plane-octec)	行八位 (Row-octec)	字位八位 (Cell-octec)

其中，m. s. 意为最高八位，l. s. 意为最低八位。

为省略起见，各八位又可写为：

m. s.　　l. s.

G-八位	P-八位	R-八位	C-八位

在适当场合，还可以进一步缩写为 G、P、R 及 C。

任意一个八位的值应由 2 个十六进制数字表示。例如：31 或 FF。若想以组、平面、行及字位的值来标识单个字符，则应以下列形式表示：

0000 0030 表示 数字 0

0000 0041 表示 拉丁文大写字母 A

当引用一个确定平面内的字符时，前面的 4 个数字（表示组八位和平面八位）可以省略。例如：在