

汉字字符集及其编码 技术的标准化现状

○ 中国标准化与信息分类编码研究所 刘碧松

1. 概述

汉字信息处理技术的每一个发展进程均得益于一系列相关标准的支撑。仅以汉字编码为例:GB 2312的问世,保证了各种DOS中文平台及各类文书处理、情报检索、电子排版、机器翻译等应用软件的正常发展,实现了汉字信息资源的交换、共享和相关设备的连网;GB 13000.1的问世,将保证更为广泛的、全球性的世界语言文字信息处理技术的发展。本文结合我国汉字信息处理技术的发展情况,就汉字字符集及其编码技术的标准化工作作一总结,并对某些尚无标准可循但又极其重要的方面提出建议性的方案供读者参考。

2. 七位编码字符集

GB 1988《信息处理 信息交换用七位编码字符集》是根据国际标准ISO 646制定的。它规定了一个由七位位组表示的128个字符组成的字符集及其编码表示,主要适用于数据处理系统与有关设备之间以及数据通信系统内的信息交换。

在已编码的128个字符中,按字符位置和功能划分,可将它们分为下列三个部分:

(1)由0列和1列的32个字符组成的控制字符集,用作C0集。

(2)由2列(不包括位于2/0的“间隔”字符)至7列(不包括位于7/15的“抹掉”字符)的94个字符组成的图形字符集,用作G0集。

(3)“间隔”字符和“抹掉”字符。其中“间隔”字符既可视作图形字符,表示一个没有可视图形的图形符号;亦可视为一个控制字符,具有格式控制字符和信息分隔控制字符的功能。“抹掉”字符则为控制字符,它主要用于消除错误的或不要的字符。

3. 八位编码字符集

GB 11383《信息处理 信息交换用八位代码结构和编码规则》规定了八位编码字符集的代码结构和编码规则,它等同采用了ISO

4873,是在七位代码体系的基础上增加一位构成的。

在八位代码结构中,涉及的元素主要有C0集、“间隔”字符、G0集、“抹掉”字符、C1集、G1集、G2集、G3集。其中C0集(00/14和00/15除外)、“间隔”字符、G0集、“抹掉”字符与GB 1988完全相同。其他元素的定义如下:

C1集:由位组为08/00至09/15表示的最多32个控制字符组成的字符集,这些字符应为C0集以外的控制字符。

G1集:由位组为10/01至15/14或10/00至15/15表示的94个或96个图形字符组成的字符集。该集所含字符应为G0集以外的图形字符。

G2、G3集:分别为由94个或96个图形字符组成的字符集。其中G2集中的字符应为G0集和G1集以外的图形字符,G3集中的字符应为G0集、G1集和G2集以外的图形字符。

为便于处理,GB 11383规定了八位代码的三个嵌套等级。在等级1中,只包括C0集、“间隔”字符、G0集、“抹掉”字符、C1集和G1集,不得使用移位功能,且假定将G0集和G1集永久地分别调入第02至07列和第10至15列。等级2包括等级1,并增加了G2集和G3集。

4. 七位和八位代码的扩充方法

GB 1988和GB 11383规定了七位和八位编码字符集,基本上满足了小字符量或中字符量的语言文字信息处理的需求。但是,对于字符数量达数千乃至数万的文种而言,其编码空间远远不够使用。为解决这一矛盾,同时又保持与原有的七位或八位代码体系的兼容,我国于1980年根据国际标准ISO 2022制定了国家标准GB 2311《信息处理 七位和八位编码字符集 代码扩充技术》,为七位或八位编码体系的代码扩充奠定了重要的基础。

在该标准中,代码的扩充方法分为两类:一类为替代法,它利用类似GB 1988结构的代码把其中的某些字符用另一些字符来代替,使其

满足另外的一些应用需求。另一类为增大字符总表法,相对复杂。

在七位代码系中,可用移位功能扩充图形字符集,还可用转义序列扩充代码。转义序列主要提供非传输控制用的单个控制功能或控制功能集。

八位代码系是在七位代码的每一个位组上增加一位得到的,与 GB 11383规定的八位代码具有相同的结构。为了使八位代码扩充技术与七位代码扩充技术所用的技术兼容,在八位代码中应严格的按与七位代码中相同的方式使用转义字符来构成转义序列,并使其含意保持不变。

5. 信息交换用汉字编码字符集 基本集

按照 GB 2311规定的代码扩充方法,我国于1980年制定了第一个汉字信息处理技术的基础性标准 GB 2312《信息交换用汉字编码字符集 基本集》。该标准规定了汉字信息交换用的基本图形字符及其代码表示,适用于一般汉字处理系统之间和汉字通信系统的信息交换。

GB 2312采用了双字节的编码方式,每个字节使用了七位代码表中2至7列(位置2/0和7/15除外)的位组,因而形成了总计有94个区、每区有94个位的编码空间。在该空间中,第1至15区为拼音文字及符号区,收入了202个一般符号、60个序号、22个数字、52个拉丁字母、169个日文假名、48个希腊字母、66个俄文字母、26个汉语拼音符号和37个汉语语音字母,总计682个图形字符。第16至94区为汉字字符区,收入了6763个汉字字符,其余的空间均作为进一步标准化的区域。

按照我国1975年进行的字频统计,汉字字符覆盖率可达99.99%,因而可以基本满足我国绝大多数汉字信息处理系统的需求。

在综合考虑汉字频度的高低、构词能力的强弱、实际用途的大小等要素的基础上,将汉字分为常用字集和次常用字集。常用字集的总数为3755个,位于第16区至55区。次常用字集的总数为3008个,位于第56区至87区。为便于对代码表中的汉字进行检索,常用字集的排列是按汉语拼音字母顺序排列的,同音字按笔形横、竖、撇、点、折为序,若起笔相同则按第二笔,并依此类推。次常用字集则按部首排列。

6. 信息交换用汉字编码字符集 辅助集

GB 2312中的6763个汉字已基本满足了我国汉字信息处理的应用需求。但是,在某些特定的文字应用环境中,如辞书编纂、文献检索等,需要用到更多的、包括繁体字在内的汉字。因此,1984年我国按照简体与繁体对应编码进行汉字扩充的原则提出了制定另外五个信息交换用汉字编码字符集的计划,并分别将其命名为辅助集、第二辅助集、第三辅助集、第四辅助集和第五辅助集。已颁布的基本集和第二及第四辅助集为简化字集,辅助集、第三和第五辅助集为繁体字集。而且,基本集与辅助集、第二辅助集与第三辅助集、第四辅助集与第五辅助集分别具有简繁体字一一对应的关系。目前,这五个辅助集已制定完成,它们的编码表示均采用双字节方式。每个代码表分为94个区,每区94个位,其中前15个区保留未用(不包括辅助集),后79个区为汉字区。

7. 通用多八位编码字符集

80年代以来,多文种信息处理技术引起了国内外的广泛注意。如何把离散的、只适用于某一地域或时域的各民族字符集集合为一个完整的、全球共识的通用编码字符集,成为国际社会急待解决的重要课题。为此,从1983年起,国际标准化组织 ISO/IEC/JTC1便开展了相关研究,并于1993年底形成国际标准 ISO/IEC 10646《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》。同年,与此对应的国家标准 GB 13000.1也通过国家标准审查。该标准以全新的编码格局,定义了世界范围内用于书面语言的各文种的主要图形字符,适用于世界书面语言信息的表示、传输、交换、处理、存储、输入及显印,是一跨国度、跨语种、跨行业的全新编码字符集。

(1) UCS 的总体结构及编码特点

UCS 是一个由四维变量组、平面、行和位构成的实体。它最多允许有128个组,每个组由256个平面组成,每个平面由256个行组成,每个行又由256个位组成。七位和八位编码体系中的控制字符位置也可用于图形字符编码,因而形成的编码空间高达 $128 \times 256 \times 256 \times 256$ 。在该编码空间中,00组中的00平面叫作基本多文种平面(BMP),它包括字母文字、音节文字及表意文字中通常使用的字符及各种符号与数字。后续的平面称为辅助平面,它用于放置 BMP 中尚未安排的图形字符。00组中平面八位值为E0到FF的32个平面为专用平面,组八位值为60至7F的32个组为专用组。对于专用组和专用

平面中的字符内容,该标准未作任何形式的约束,但除此以外的、尚未在该标准中安排字符的代码位置不得用于任何其他目的,应留作将来进一步标准化使用。

UCS 对字符的编码采用了两种形式。一种为正则形式,由四个八位表示,并相应地指定字符所处的组、平面、行、和位。另一种为双八位的 BMP 形式,由两个八位表示,基本多文种平面可当作双八位编码字符集使用。

(2)基本多文种平面

基本多文种平面被分为下列四个区:

①A—区:代码位置为0000至4DFF,用于安排字母文字、音节文字及各种符号。

②I—区:代码位置为4E00至9FFF,用于安排中国、日本和韩国(CJK)统一的表意文字。

③O—区:代码位置为A000至DFFF,供未来的标准化使用。

④R—区:代码位置为E000至FFFD,用于专用字符、变形显现字符(如阿拉伯文)以及为了与现行标准兼容而收入的字符。

目前,在 BMP 的 I 区中已收入20902个中国、日本和韩国的统一汉字。这些汉字源于中国大陆的基本集、辅助集、第三辅助集、第五辅助集、现代汉语通用字表、GB 8565《信息处理文本通信用编码字符集》、台湾省的 TCA CNS 11643《中国台北信息交换用汉字编码字符集》第一字面、第二字面和第十四字面、日本的 JIS X 0208《信息交换用日文编码图形字符集》、JIS X 0212《信息交换用日文编码图形字符集 辅助集》、韩国的 KS C 5601《信息交换用代码(朝鲜文和朝鲜汉字)》、KSC 5657《信息交换用韩国编码图形字符集 辅助集》等。它们都经 CJK 三方按共同认可的编码原则、认同原则、收字原则、排序规则等严格审查产生。

(3)UCS 提供的子集

考虑到实际上存在的不同应用需求,UCS 引入了两种图形字符子集的概念供用户使用。一种是有限子集,它是为使用 UCS 字符集的系统与使用现行字符集标准的系统之间交互运作而规定的,通常由 UCS 中的若干个单个的字符组成。另一种是选择子集,由 UCS 中定义的一个或若干个字符汇集组成。

(4)UCS 的交换格式

由于 UCS 是面向16位或32位编码的,不再按字节保留控制字符的位组,因而不能直接经

由对控制字符敏感的系统传送。为此,UCS 使用了一个交换格式 UTF-1。

8. 汉字内部码

汉字内部码是汉字在设备或系统内部处理用的汉字代码。

虽然汉字编码字符集大多是在 GB 1988和 GB 11383规定的七位和八位代码系的基础上,按 GB 2311的规定进行扩充的双字节、有序性的汉字集合,为一般汉字处理、汉字通信等系统之间的信息交换打下坚实的基础。但是,在现有软件中,直接将汉字的这类代码用于设备或系统内部处理,即把汉字交换码作为内部码使用,会导致过大的开发工作量,尤其是当它不加区别地与单字节字符混合时,将产生一系列的重大歧义。由于这个原因,汉字信息系统采用了若干类别的汉字内部码。

汉字内部码这种“多码并存”的现状,在单机单用户的场合下问题不大,但随着计算机应用水平的提高,从单机到网络、从单用户到多用户的发展,各种规模的汉字信息处理系统所建立的资源会不能共享,不同内部码格局的系统内部也会互不兼容。因此,汉字内部码的标准化也是极其重要的。

当前,我国尚无汉字内部码方面的国家标准,本文仅就 GB 2312和 GB 13000的现行内码格局作一推荐性说明。

(1)GB 2312的内码格局

用于表示字符的两个八位位组的高位均设为“1”,占用纵向由10/01至15/14、横向由10/01至15/14构成的方块,共8836个编码位置,并与 GB 2312的编码位置具有一一对应的映射关系。如果还需少量其他汉字,可采用动态再定义的方法收入最后一行。

(2)GB 13000的内码格局

基于 GB 13000的字符内码格局总体上采用8140—FEFE的矩形区域,剔除XX7F一条线,共23940个码位。汉字区占其中的21008个码位。GB 2312汉字区为B0A1—F7FE,总计6763个汉字,6768个码位,字序与 GB 2312一致。GB 13000中的其他 CJK 汉字,按 UCS 代码大小排列在由8140—A0FE(剔除XX7F)和AA40—FEA0(剔除XX7F)构成的两个矩形区域。用户自定义区由AAA1—AFFE、F8A1—FEFE和A140—A7A0(剔除XX7F)构成的三个矩形区域组成,共1894个码位。