


# 非数值数据的编码表示

南京大学
计算机科学与技术系
袁春风
email: cfyuan@nju.edu.cn
2015.6

# 逻辑数据的编码表示

- 计算机中何时会用到逻辑数据？
  - 表示逻辑（关系）表达式中的逻辑值：真 /假

  例如，对于关系表达式：(x>0) 并且 （y<=0)

  “x>0”、“y<=0”、“(x>0) 并且(y<=0)”都是逻辑值
- 表示
  - 用一位表示 。N位二进制数（位串）可表示N个逻辑数据
- 运算
  - 按位进行。如，按位与 / 按位或 / 逻辑左移 / 逻辑右移 等
- 识别
  - 逻辑数据和数值数据在形式上并无差别，也是一串0/1序列，计算机靠指令来识别。

# 西文字符的编码表示

- 特点
  - 是一种拼音文字，用有限几个字母可拼写出所有单词
  - 只需对有限个字母和数学符号、标点符号等辅助字符编码
  - 所有字符总数不超过256个，使用7或8个二进位可表示
- 表示（常用编码为7位ASCII码）
  - 十进制数字：0/1/2.../9
  - 英文字母：A/B/.../Z/a/b/.../z
  - 专用符号：+/-/%/*/&/.....
  - 控制字符（不可打印或显示）

必须熟悉对应的ASCII码！

- 操作
  - 字符串操作，如:传送/比较　等

# ASCII码表

| $b_6b_5b_4b_3b_2b_1b_0$ | $b_6b_5b_4$ =000 | $b_6b_5b_4$ =001 | $b_6b_5b_4$ =010 | $b_6b_5b_4$ =011 | $b_6b_5b_4$ =100 | $b_6b_5b_4$ =101 | $b_6b_5b_4$ =110 | $b_6b_5b_4$ =111 |
|---|---|---|---|---|---|---|---|---|
| $b_3b_2b_1b_0$=0000 | NUL | DLE | SP | 0 | @ | P | ` | p |
| $b_3b_2b_1b_0$=0001 | SOH | DC1 | ! | 1 | A | Q | a | q |
| $b_3b_2b_1b_0$=0010 | STX | DC2 | “ | 2 | B | R | b | r |
| $b_3b_2b_1b_0$=0011 | ETX | DC3 | # | 3 | C | S | c | s |
| $b_3b_2b_1b_0$=0100 | EOT | DC4 | $ | 4 | D | T | d | t |
| $b_3b_2b_1b_0$=0101 | ENQ | NAK | % | 5 | E | U | e | u |
| $b_3b_2b_1b_0$=0110 | ACK | SYN | & | 6 | F | V | f | v |
| $b_3b_2b_1b_0$=0111 | BEL | ETB | ‘ | 7 | G | W | g | w |
| $b_3b_2b_1b_0$=1000 | BS | CAN | ( | 8 | H | X | h | x |
| $b_3b_2b_1b_0$=1001 | HT | EM | ) | 9 | I | Y | i | y |
| $b_3b_2b_1b_0$=1010 | LF | SUB | * | : | J | Z | j | z |
| $b_3b_2b_1b_0$=1011 | VT | ESC | + | ; | K | [ | k | { |
| $b_3b_2b_1b_0$=1100 | FF | FS | , | < | L | \ | l | \| |
| $b_3b_2b_1b_0$=1101 | CR | GS | - | = | M | ] | m | } |
| $b_3b_2b_1b_0$=1110 | SO | RS | . | > | N | ^ | n | ~ |
| $b_3b_2b_1b_0$=1111 | SI | US | / | ? | O | _ | o | DEL |

# 汉字及国际字符的编码表示

- **汉字特点**
  - –汉字是表意文字，一个字就是一个方块图形。
  - –汉字数量巨大，总数超过6万字，给汉字在计算机内部的表示、汉字的传输与交换、汉字的输入和输出等带来了一系列问题。
- **编码形式**
  - –有以下几种汉字代码：
  - · 输入码：对汉字用相应按键进行编码表示，用于输入
  - · 内码：用于在系统中进行存储、查找、传送等处理
  - · 字模点阵或轮廓描述: 描述汉字字模点阵或轮廓，用于显示/打印

问题：西文字符有没有输入码？有没有内码？

有没有字模点阵或轮廓描述？

# GB2312-80字符集

- 由三部分组成
  - ① 字母、数字和各种符号，包括英文、俄文、日文平假名与片假名、罗马字母、汉语拼音等共687个
  - ② 一级常用汉字，共3755个，按汉语拼音排列
  - ③ 二级常用汉字，共3008个 ，不太常用，按偏旁部首排列
- 汉字的区位码
  - 码表由94行、94列组成，行号为区号，列号为位号，各占7位
  - 指出汉字在码表中的位置，共14位，区号在左、位号在右
- 汉字的国标码
  - 每个汉字的区号和位号各自加上32（20H），得到其“国标码”
  - 国标码中区号和位号各占7位。在计算机内部，为方便处理与存储，前面添一个0，构成一个字节

# 汉字内码

- 至少需2个字节才能表示一个汉字内码。为什么？
  - –由汉字的总数（超过6万字）决定！ $2^{16}=65536$
- 可在GB2312国标码的基础上产生汉字内码
  - –为与ASCII码区别，将国标码的两个字节的第一位置“1”后得到一种汉字内码（可以有不同的编码方案）

例：汉字“大”在码表中位于第20行、第83列。因此区位码为0010100 1010011，在区、位码上各加32得到两个字节编码，即00110100 01110011B=3473H。前面的34H和字符“4”的ACSII码相同，后面的73H和字符“s”的ACSII码相同，但是，将每个字节的最高位各设为“1”后，就得到其内码：B4F3H (10110100 11110011B)，因而不会和ASCII码混淆。

# 多媒体信息的表示

- 图形、图像、音频、视频等信息在机器内部也用0和1表示
  - 图形用构建图形的直线或曲线的坐标点及控制点来描述，而这些坐标点或控制点则用数值数据描述
  - 图像用构成图像的点（像素）的亮度、颜色或灰度等信息来描述，这些亮度或颜色等值则用数值数据描述
  - 音频信息通过对模拟声音进行采样、量化（用二进制编码）来获得，因此量化后得到的是一个数值数据序列（随时间变化）
  - 视频信息描述的是随时间变化的图像（每一幅图像称为一帧）
  - 音乐信息（MIDI）通过对演奏的乐器、乐谱等相关的各类信息用0和1进行编码来描述
  - .....

多媒体信息用一个复杂的数据结构来描述，其中的基本数据或者是数值数据，或者是用0/1编码的非数值数据