多媒体技术基础:第四次作业

PB20061372 朱云沁　Nov. 1, 2023

38. 名词解释: UTF-8, UTF-16, UTF-32.

UTF 即 Unicode 转换格式. Unicode 为世界上所有字符都分配了一个唯一的数字编号, 这个编号范围从 0x000000 到 0x10FFFF, 但没有规定这个编号如何存储. UTF-8, UTF-16, UTF-32 等编码格式用于将 Unicode 字符映射到二进制数据.

- UTF-8: 一种可变长度编码, 它使用 1 到 4 个字节来表示一个字符. 当字符是 ASCII 字符时, 只需要使用 1 个字节, 这使得它在存储 ASCII 字符时非常高效.
  (1) 对于单字节的符号, 字节的第一位设为 0, 后面的 7 位为这个符号的 Unicode 码, 因此对于英文字母, UTF-8 编码和 ASCII 码是相同的.
  (2) 对于 n 字节的符号($n > 1$), 第一个字节的前 n 位都设为 1, 第 $n+1$ 位设为 0, 后面字节的前两位一律设为 10, 剩下的位用来存储字符的 Unicode 码.
- UTF-16: 一种可变长度编码, 它使用 2 或 4 个字节来表示一个字符. UTF-16 不像 UTF-8 那样高效地存储 ASCII 字符, 但是对于非 ASCII 字符, UTF-16 可以比 UTF-8 更加紧凑.
  (1) 编号在 U+0000 到 U+FFFF 的字符, 直接用两个字节表示.
  (2) 编号在 U+10000 到 U+10FFFF 的字符, 用四个字节表示.
- UTF-32: 一种固定长度编码, 它使用 4 个字节来表示一个字符. UTF-32 可以表示所有的 Unicode 字符, 但是由于其固定长度, 它的存储效率较低.

39. 某符号的 Unicode 数字编号为 0x4E2D, 写出 UTF-8 编号后的 16 进制结果.

0x4E2D = 0b0100111000101101 → 0b111001001011100010101101 = 0xE4B8AD.

40. 已知信源 $X:\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$, 各信源符号的概率依次为 $P(X):\{0.2,0.19,0.18,0.17,0.15,0.1,0.01\}$. 求霍夫曼编码, 并计算编码效率.

构建霍夫曼树的过程如下:

1. 节点: $\{x_7\},\{x_6\},\{x_5\},\{x_4\},\{x_3\},\{x_2\},\{x_1\}$.
   概率: $0.01 < 0.1 < 0.15 < 0.17 < 0.18 < 0.19 < 0.2$.
2. 节点: $\{x_7,x_6\},\{x_5\},\{x_4\},\{x_3\},\{x_2\},\{x_1\}$.
   概率: $0.11 < 0.15 < 0.17 < 0.18 < 0.19 < 0.2$.
3. 节点: $\{x_4\},\{x_3\},\{x_2\},\{x_1\},\{x_5,x_6,x_7\}$.
   概率: $0.17 < 0.18 < 0.19 < 0.2 < 0.26$.
4. 节点: $\{x_2\},\{x_1\},\{x_5,x_6,x_7\},\{x_3,x_4\}$.
   概率: $0.19 < 0.2 < 0.26 < 0.35$.
5. 节点: $\{x_5,x_6,x_7\},\{x_3,x_4\},\{x_1,x_2\}$.
   概率: $0.26 < 0.35 < 0.39$.
6. 节点: $\{x_1,x_2\},\{x_3,x_4,x_5,x_6,x_7\}$.
   概率: $0.39 < 0.61$.
7. 节点: $\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$.
   概率: 1.

霍夫曼编码如下:

$$x_1:01,\quad x_2:00,\quad x_3:111,\quad x_4:110,\quad x_5:101,\quad x_6:1001,\quad x_7:1000$$

编码效率, 即平均码长为:

$$(0.2\times 2)+(0.19\times 2)+(0.18\times 3)+(0.17\times 3)+(0.15\times 3)+(0.1\times 4)+(0.01\times 4)=2.72$$

41. 对一个具有符号集 $B=\{b_1,b_2\}=\{0,1\}$, 设信源产生 2 个符号的概率分别为 $P(b_1)=0.2, P(b_2)=0.8$. 对二进制数 1001 进行算术编码 (结果用十进制数表示).

区间划分: $[0,0.2)$ 对应符号 b_1, $[0.2,1)$ 对应符号 b_2. 对于二进制数 1001,

- 第一位是 1，因此选取区间 $[0.2,1)$;
- 第二位是 0，因此选取区间 $[0.2,0.36)$;
- 第三位是 0，因此选取区间 $[0.2,0.232)$;
- 第四位是 1，因此选取区间 $[0.2064,0.232)$.

结果为 0.2064.

42. 对信息 00002033001110000600110111 1 进行 RLE 编码.

401210232031401620211041.