中国矿业大学

第四章 贪心算法

§4.4 哈夫曼编码

韩丽霞

计算机科学与技术学院
School Of Computer Science And Technology, CUMT

学习要点

01 前缀码的概念

02 贪心算法求解最优前缀码

问题

设某信源产生a、b、c、d、e和f 6种符号，其频率见下表。

字符	定长码	频率
a	000	45
b	001	13
c	010	12
d	011	16
e	100	9
f	101	5

000 001 000 ……

a b

定长码译码简单

问题目标

给定编码字符集C及任一字符c的出现频率$f(c)$，

定义编码方案的平均码长

$$B(T)=\sum_{c\in C} f(c)d_T(c)$$

码长

目标：找到使平均码长达到最小的编码方案。

定长码

字符	a	b	c	d	e	f
频率(千次)	45	13	12	16	9	5
定长码	000	001	010	011	100	101

定长码平均码长：

$$(45 + 13 + 12 + 16 + 9 + 5) \times 3 = 300$$

最优编码？

背景

- 1951年，哈夫曼在MIT信息论课程中，需完成学期报告：寻找最有效的二进制编码。
- 1952年，根据香农（Shannon）在1948年和范若（Fano）在1949年阐述的编码思想提出了一种**不定长编码**的方法，也称哈夫曼（Huffman）编码。

定长码

字符	a	b	c	d	e	f
频率	45	13	12	16	9	5
变长码	0	1	10	11	100	101

0 1 0 1 1 1 0 1

a

译码正确性？

分析

字符	a	b	c	d	e	f
变长码	0	1	10	11	100	101

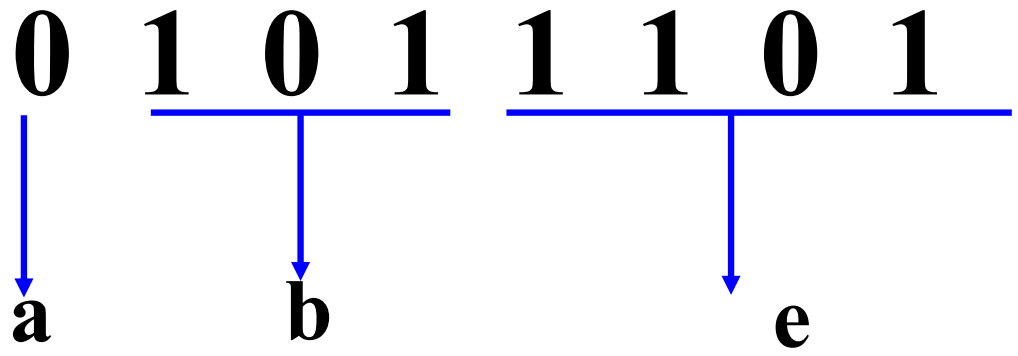
0 1 0 1 1 1 0 1

1)0不是其余编码的前缀　　2)1、10是部分编码的前缀

前缀码：对每一个字符规定一个0,1串作为其代码，并要求任一字符的代码都不是其它字符代码的前缀。

变长码

字符	a	b	c	d	e	f
频率	45	13	12	16	9	5
变长码	0	101	100	111	1101	1100



平均码长

字符	a	b	c	d	e	f
频率	45	13	12	16	9	5
变长码	0	101	100	111	1101	1100

定长码：300

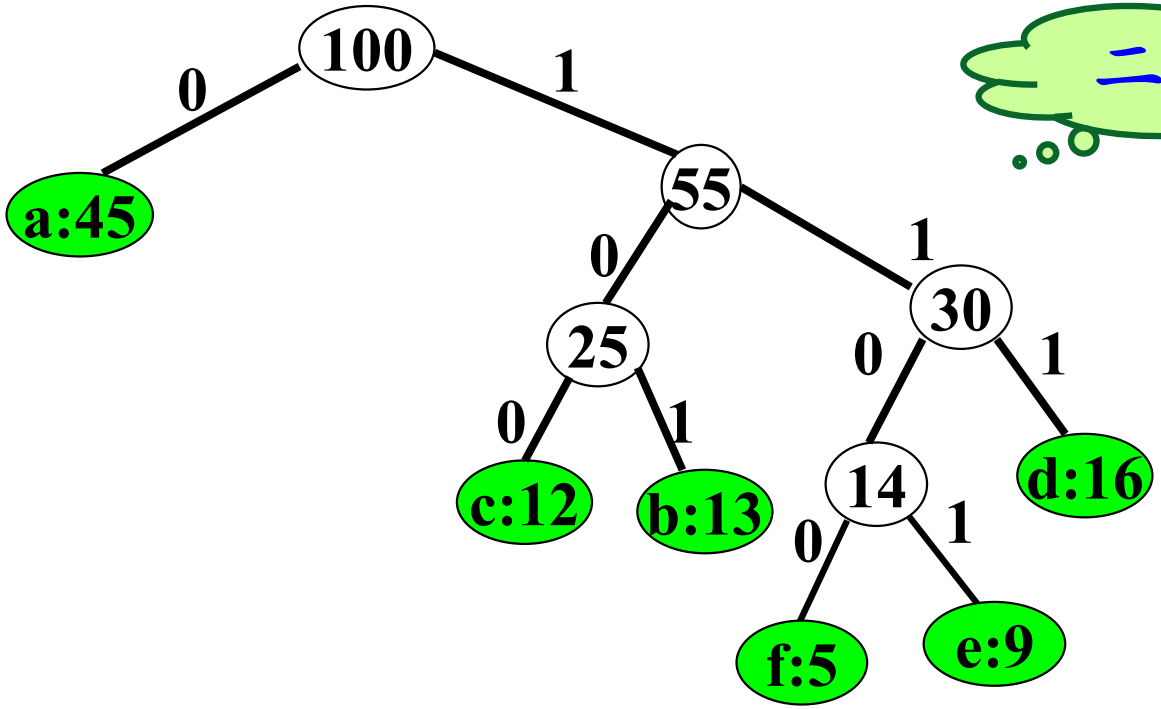
变长码：

25%

$$45\times1+13\times3+12\times3+16\times3+9\times4+5\times4=224$$

思考

字符	a	b	c	d	e	f
频率	45	13	12	16	9	5
变长码	0	101	100	111	1101	1100



二元完全树

二元完全树中，树根到树叶的路径为什么是前缀码？

编码问题

1)分析出定长码的弊端，考虑变长码的编码方式

2)考虑变长码译码的正确性，前缀码的定义

3) 如何寻找前缀码，建立二元完全树

树叶

哈夫曼编码

给定编码字符集C及任一字符c的出现频率$f(c)$。

C的一个前缀码方案对应于一棵二元完全树T。字符c在树中的深度记为$d_T(c)$。该编码方案的平均码长

$$B(T) = \sum_{c \in C} f(c) d_T(c)$$

目标：找到使平均码长达到最小的前缀码编码方案 。

小结

01 哈夫曼编码问题

02 前缀码

03 思考：贪心算法如何求解哈夫曼编码问题