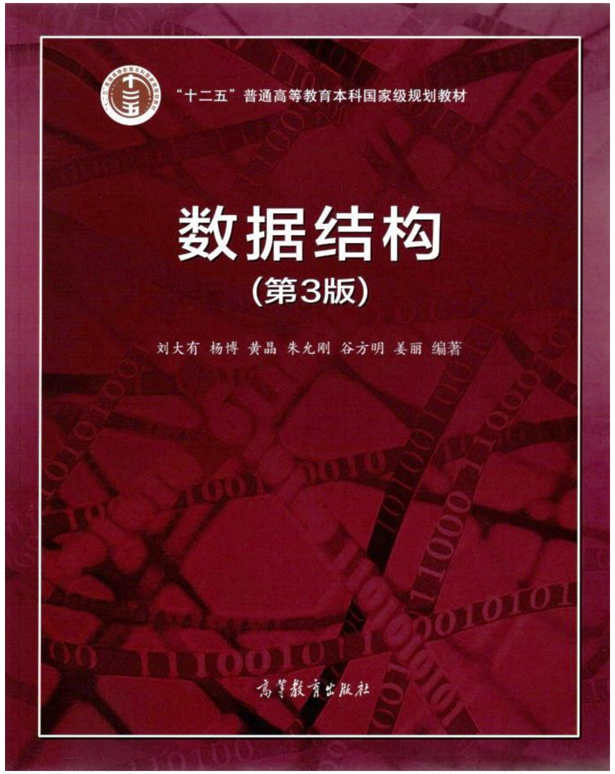


树和二叉树的应用

- 压缩与哈夫曼树
- 表达式树
- 并查集

数据之法
结构之美
算法之道

zhuyungang@jlu.edu.cn

建议大家争取在大学四年中积累编写10万行代码的经验。我们必须明白的是：好程序员是写出来的。

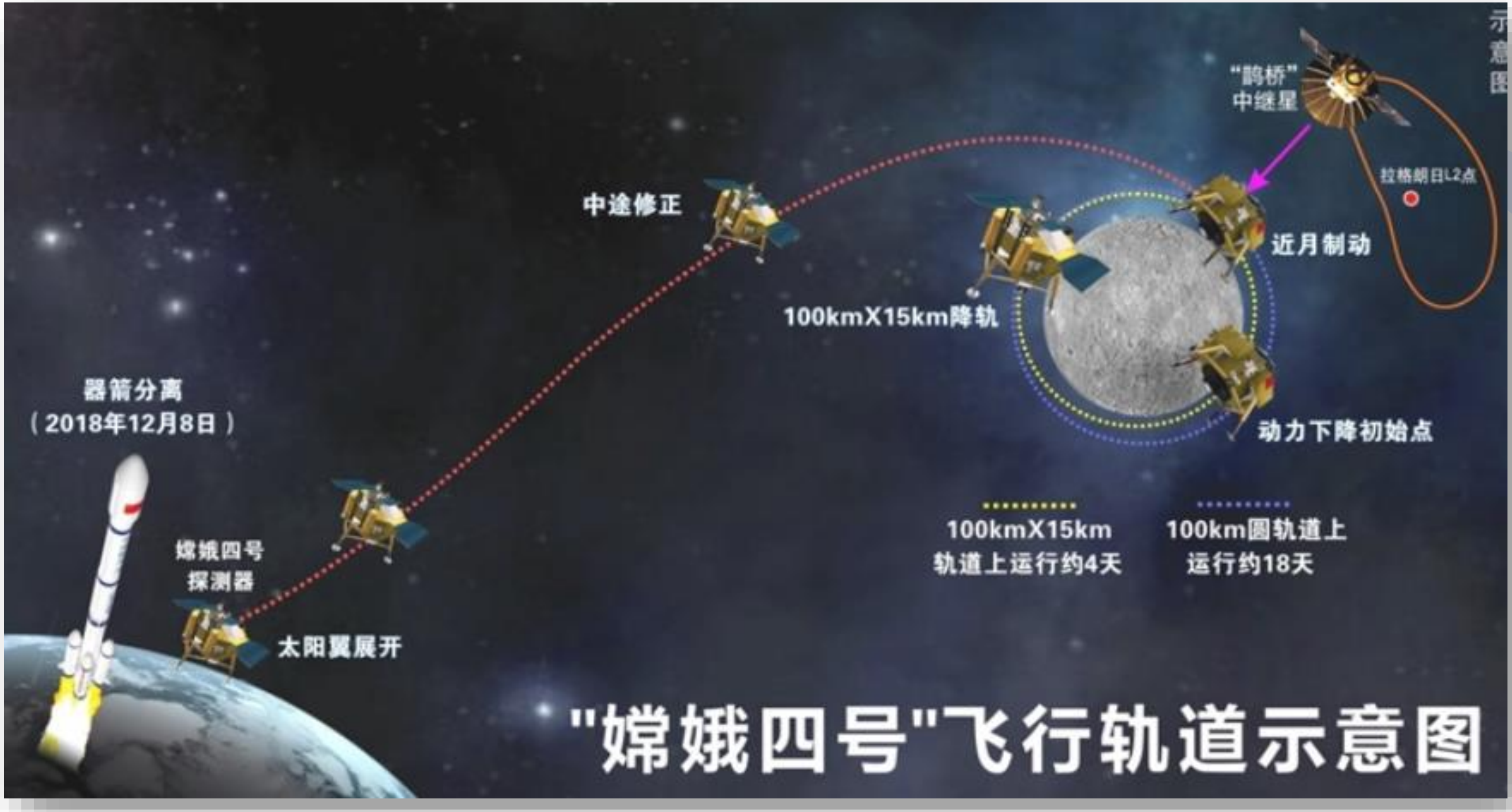
——李开复

原微软全球副总裁、微软亚洲研究院院长

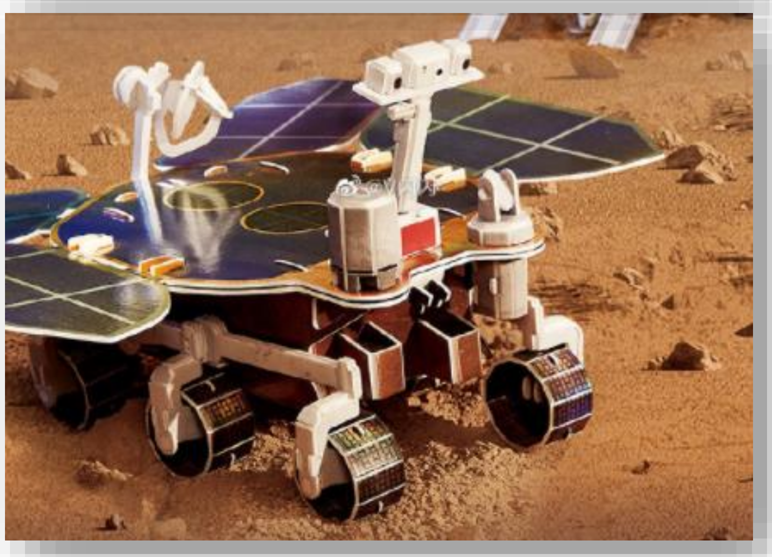
原Google全球副总裁兼大中华区总裁

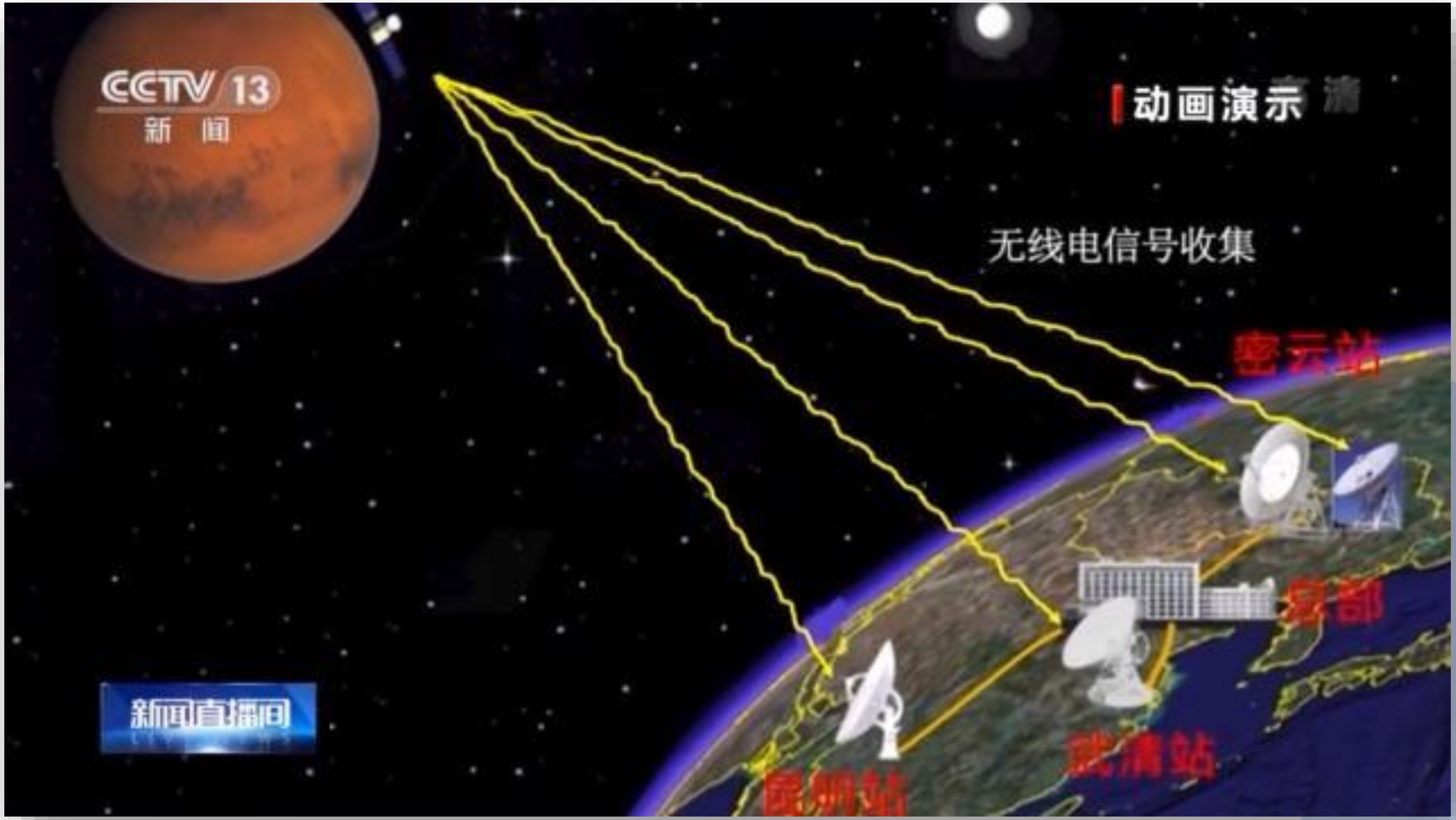
来源：《程序员》杂志，2006年第4期

嫦娥月球探测器，与地球传输图像，需进行数据压缩



天问一号火星探测器与祝融号火星车，与地球传输图像，需进行数据压缩





数据压缩

- **数据压缩**是计算机科学中的重要技术。
- 数据压缩过程称为**编码**，即将文件中的每个字符均转换为一个唯一的二进制位串。
- 数据解压过程称为**解码**，即将二进制位串转换为对应的字符。

信息编码

假设有一个文本文件仅包含4种字符：*A*、*B*、*C*、*D*，且文件中有11个*A*，4个*B*，3个*C*，2个*D*。

若采用等长编码，因为$\lceil \log_2 4 \rceil = 2$，所以每个字符都至少由一个2位的二进制数表示。于是文件所需存储空间（文件的总编码长度）为：

$$(11 + 4 + 3 + 2) \times 2 = 40 \text{ bit} = 5 \text{ Byte}$$

还有更好的方案么？

信息编码

假设有一个文本文件仅包含4种字符：*A*、*B*、*C*、*D*，且文件中有11个*A*，4个*B*，3个*C*，2个*D*。

若采用等长编码，因为$\lceil \log_2 4 \rceil = 2$，所以每个字符都至少由一个2位的二进制数表示。于是**文件所需存储空间（文件的总编码长度）**为：

$$(11 + 4 + 3 + 2) \times 2 = 40 \text{ bit} = 5 \text{ Byte}$$

字符被出现的频率不同，可否借助这一信息，设计不等长编码，使文件总长度更短？

如何才能压缩总编码长度？

假设有一个文件中有100个*A*，1个*B*，1个*C*，1个*D*。

编码策略1：

A、*B*、*C*、*D*都用2个二进制位表示。

总编码长度：$103 \times 2 = 206$ bit

编码策略2：

*A*用1个二进制位表示；*B*、*C*、*D*用5个二进制位表示。

总编码长度：$100 + 15 = 115$ bit

采用不等长编码，希望：

① 熵编码：文件中出现频率高的字符的编码长度尽可能短。

解码过程不能出现歧义性

字符	编码
A	10
B	01
C	1001

歧义：1001 = *C* 还是 *AB*?

原因：*A*的编码是 *C* 的前缀。

采用**不等长**编码，希望：

② **前缀码（无前缀冲突编码，Prefix-Free Codes）**：字符集中任何字符的编码都不是其它字符的编码的前缀。

怎样的前缀码才能使文件的总编码长度最短？

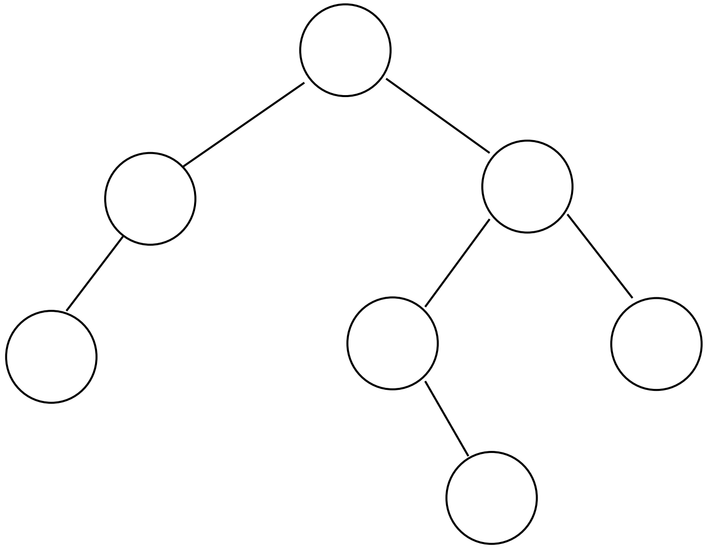
最优编码问题描述：

设组成文件的字符集$A=\{a_1,a_2,\ldots,a_n\}$，其中a_i出现的次数为c_i ，a_i的编码长度为l_i。设计一个前缀码方案，使文件的总编码长度最小：

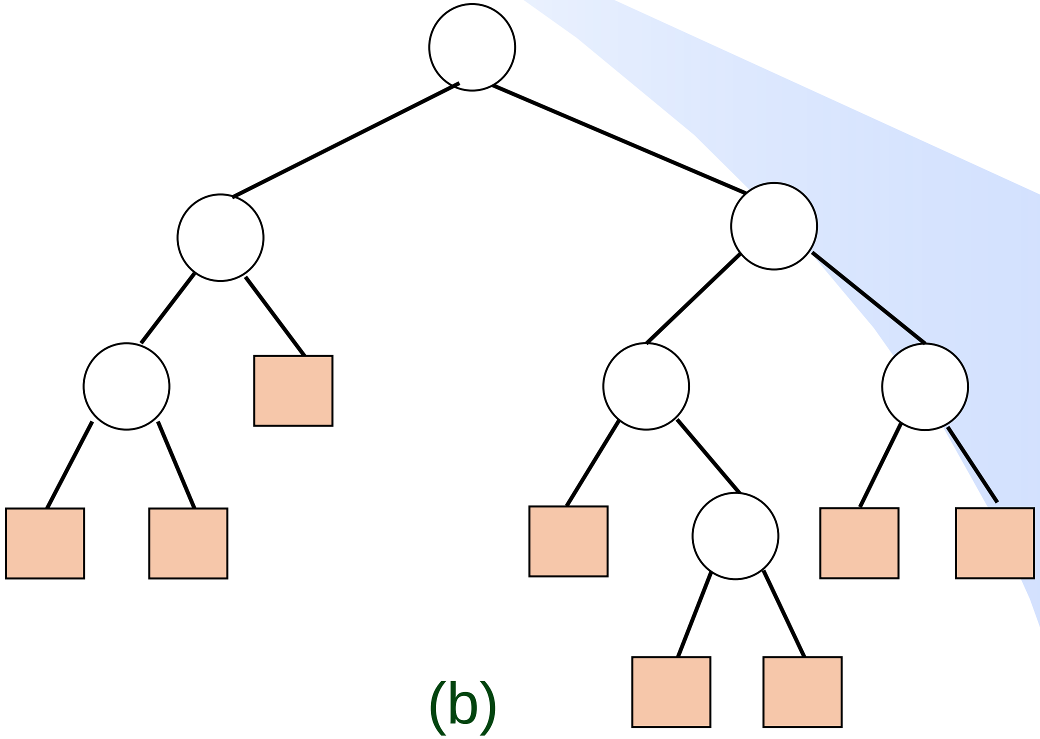
$$\min \sum_{i=1}^{n} c_i \cdot l_i$$

扩充二叉树

定义 在二叉树中空指针的位置，都增加特殊的结点（空叶结点），由此生成的二叉树称为扩充二叉树。



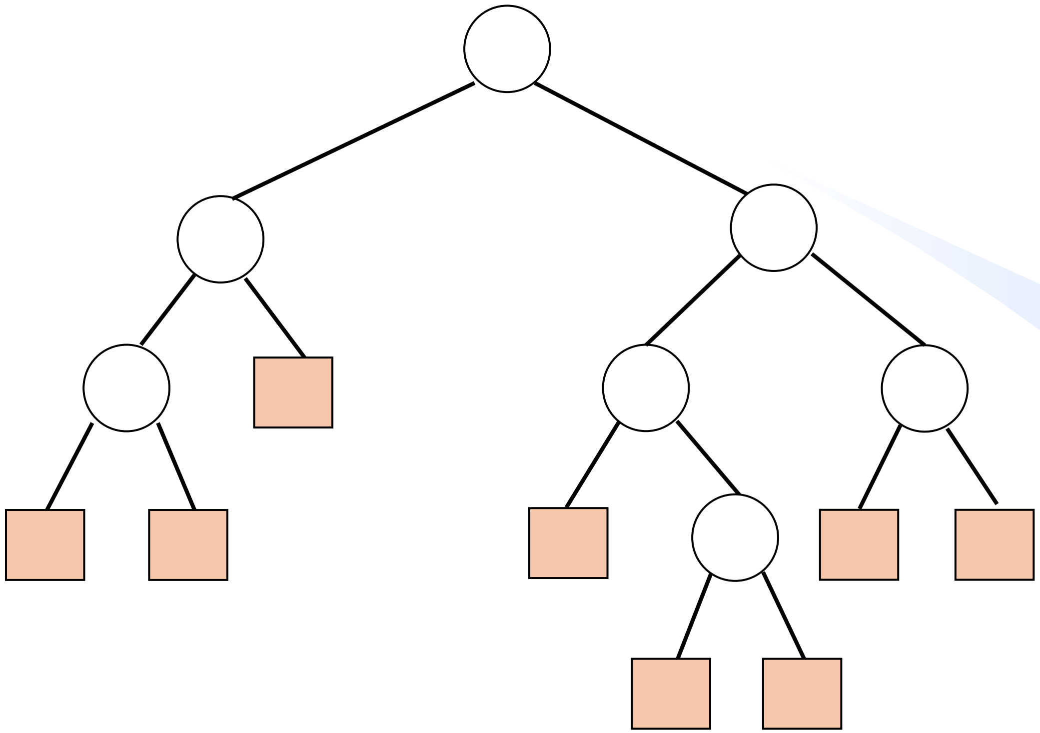
(a)



(b)

二叉树及其对应的扩充二叉树

扩充二叉树



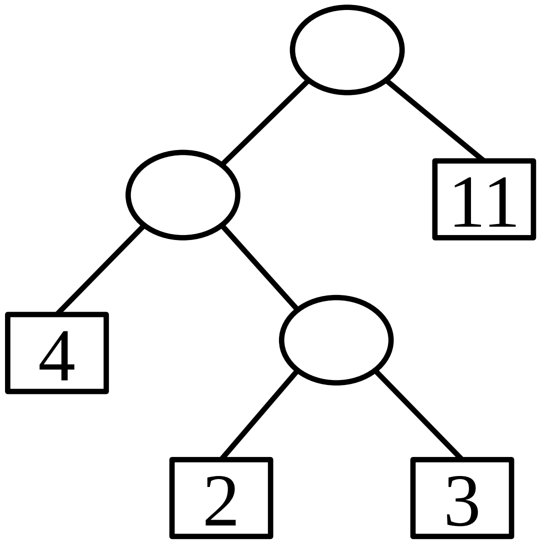
- 称圆形结点为<u>内结点</u>，方形结点为<u>外结点</u>。
- 每个内结点都有2个孩子，每个外结点没有孩子。
- 规定空二叉树的扩充二叉树是只有一个外结点。

加权路径长度（*Weighted Path Length, WPL*）

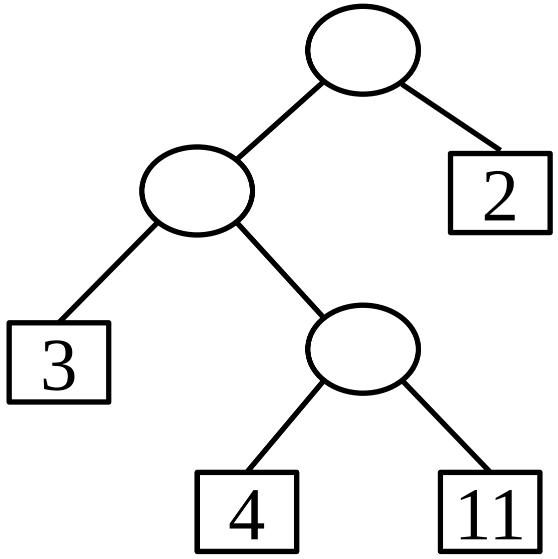
设扩充二叉树有n个外结点，为每个外结点赋予一个实数，称为该结点的权值，第i个外结点的权值为w_i，深度为L_i，则**加权路径长度**定义为：

$$WPL=\sum_{i=1}^{n} w_i L_i$$

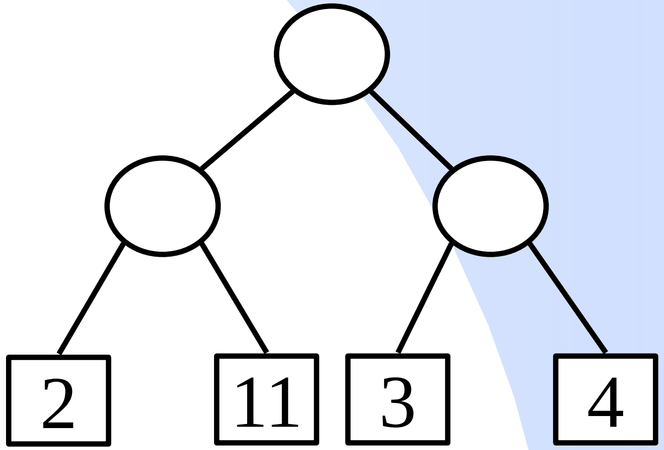
n个带权外结点构成的所有扩充二叉树中，WPL值最小者称为**最优二叉树**。



4×2+2×3+3×3+11×1=34



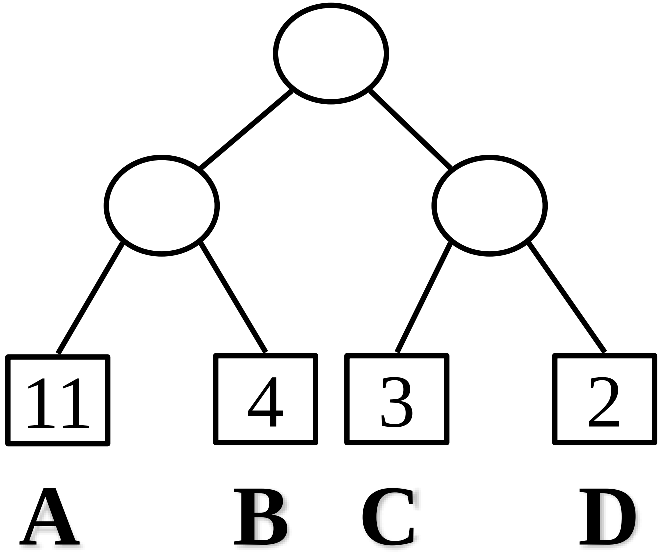
3×2+4×3+11×3+2×1=53



2×2+11×2+3×2+4×2=40

- 一种文件编码方案可以映射为一棵扩充二叉树。

文件编码	扩充二叉树
字符a_i ←----→	外结点i
字符的出现次数c_i ←----→	外结点的权值 w_i
字符的编码长度l_i ←----→	外结点的深度 L_i
文件总编码长度 ←----→	扩充二叉树的 WPL值



$$\text{文件编码长度}=\sum_{i=1}^{n} c_i \cdot l_i$$

$$WPL=\sum_{i=1}^{n} w_i \cdot L_i$$

最优编码问题

给定n种字符和每种字符出现的次数

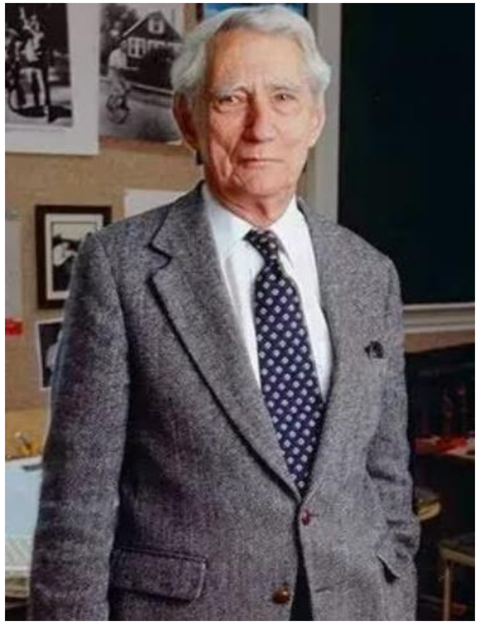
构建一种总编码长度最短的编码方案（最优编码方案）

构造最优二叉树问题

给定n个外结点和每个结点的权值

构建一棵WPL值最小的扩充二叉树（最优二叉树）

Shannon-Fano算法



Claude Shannon
（1916-2001）
麻省理工学院 教授
美国科学院院士
美国工程院院士
信息论之父

Robert Fano
（1917-2016）
麻省理工学院 教授
美国科学院院士
美国工程院院士

Shannon-Fano算法

将字符按频率递减排序

重复做：将字符集切分成两部分，使两部分频率之和尽可能相等

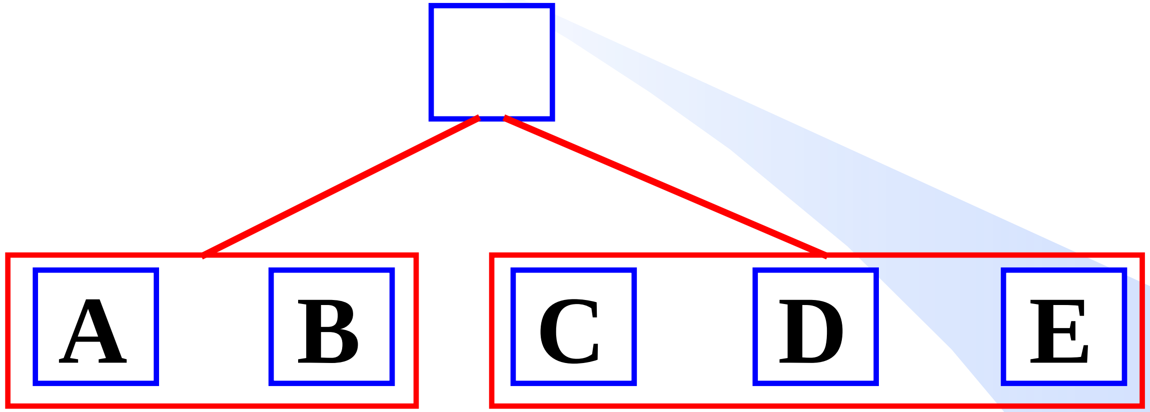
字符	出现次数
A	15
B	7
C	6
D	6
E	5

A B C D E

Shannon-Fano算法

重复做：将字符集分成两部分，使两部分频率之和尽可能相等

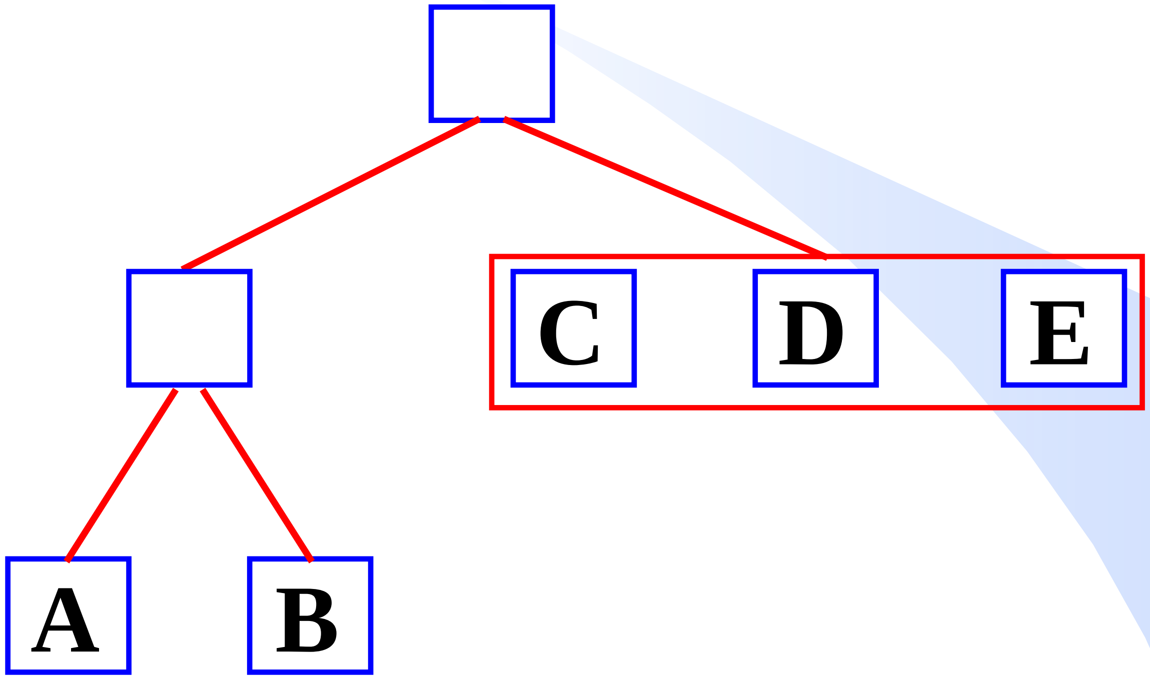
字符	出现次数
A	15
B	7
C	6
D	6
E	5



Shannon-Fano算法

重复做：将字符集分成两部分，使两部分频率之和尽可能相等

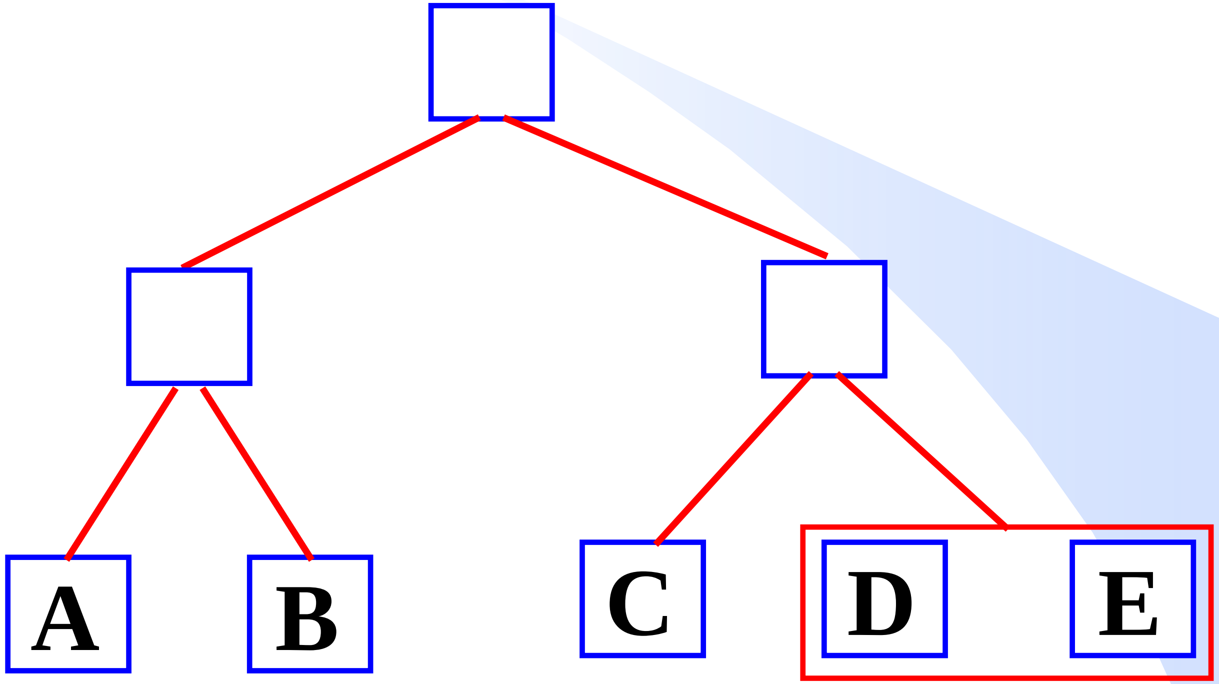
字符	出现次数
A	15
B	7
C	6
D	6
E	5



Shannon-Fano算法

重复做：将字符集分成两部分，使两部分频率之和尽可能相等

字符	出现次数
A	15
B	7
C	6
D	6
E	5

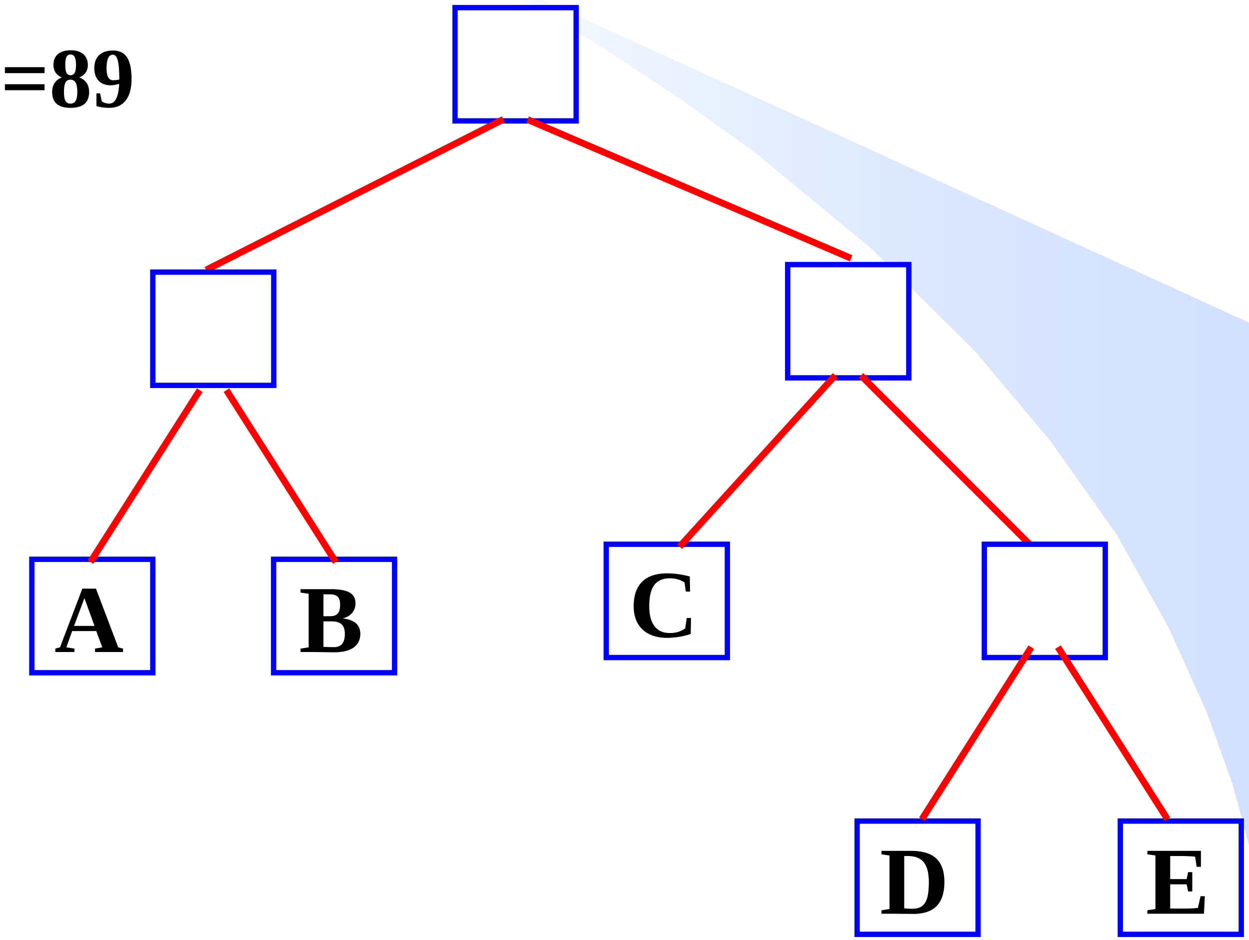


Shannon-Fano算法

自顶向下建树，并非最优解

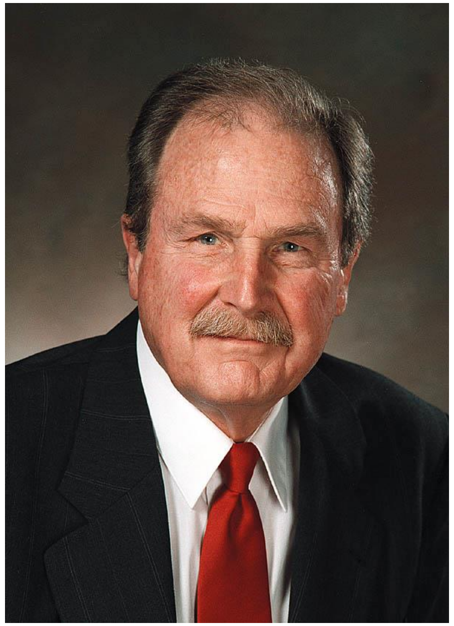
WPL=(15+7+6)*2+(6+5)*3=89

字符	出现次数
A	15
B	7
C	6
D	6
E	5

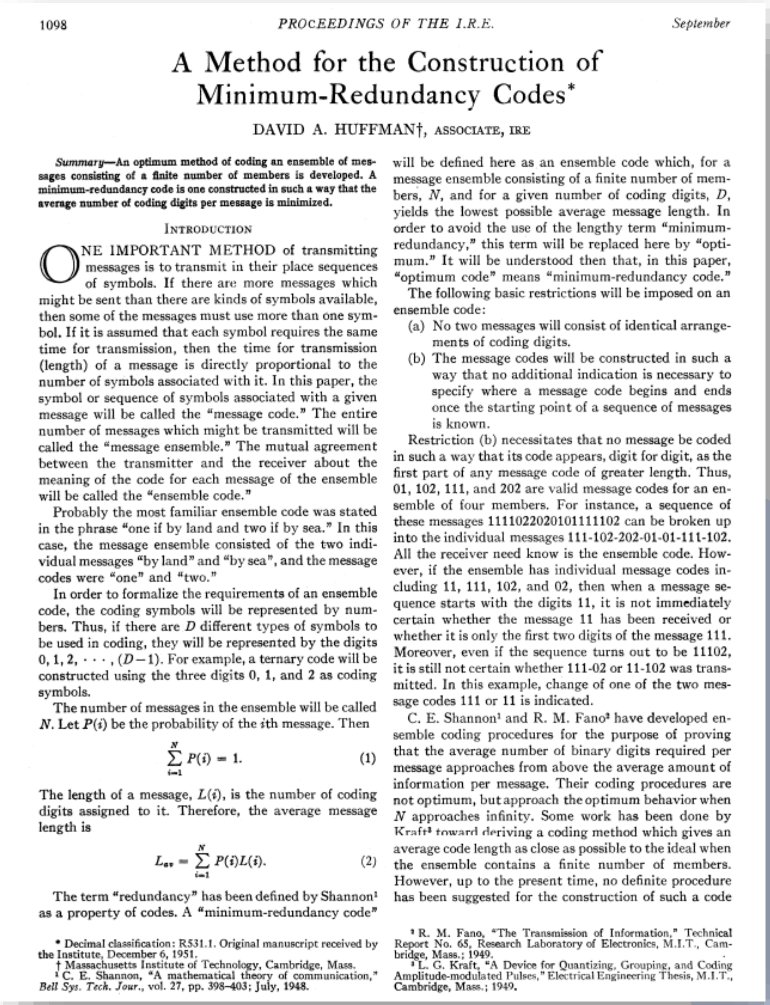


Huffman算法

Robert Fano
（1917-2016）
麻省理工学院 教授
美国科学院院士
美国工程院院士



David Huffman
（1925-1999）
麻省理工学院 博士
加州大学圣克鲁兹分校 教授



A Method for the Construction of Minimum-Redundancy Codes*

DAVID A. HUFFMAN†, ASSOCIATE, IRE

***Summary*—An optimum method of coding an ensemble of messages consisting of a finite number of members is developed. A minimum-redundancy code is one constructed in such a way that the average number of coding digits per message is minimized.**

Introduction

ONE IMPORTANT METHOD of transmitting messages is to transmit in their place sequences of symbols. If there are more messages which might be sent than there are kinds of symbols available, then some of the messages must use more than one symbol. If it is assumed that each symbol requires the same time for transmission, then the time for transmission (length) of a message is directly proportional to the number of symbols associated with it. In this paper, the symbol or sequence of symbols associated with a given message will be called the "message code." The entire number of messages which might be transmitted will be called the "message ensemble." The mutual agreement between the transmitter and the receiver about the meaning of the code for each message of the ensemble will be called the "ensemble code."

Probably the most familiar ensemble code was stated in the phrase "one if by land and two if by sea." In this case, the message ensemble consisted of the two individual messages "by land" and "by sea", and the message codes were "one" and "two."

In order to formalize the requirements of an ensemble code, the coding symbols will be represented by numbers. Thus, if there are D different types of symbols to be used in coding, they will be represented by the digits $0, 1, 2, \cdots, (D-1)$. For example, a ternary code will be constructed using the three digits 0, 1, and 2 as coding symbols.

The number of messages in the ensemble will be called N. Let $P(i)$ be the probability of the ith message. Then

$$\sum_{i=1}^{N} P(i) = 1. \tag{1}$$

The length of a message, $L(i)$, is the number of coding digits assigned to it. Therefore, the average message length is

$$L_{av} = \sum_{i=1}^{N} P(i)L(i). \tag{2}$$

The term "redundancy" has been defined by Shannon[1] as a property of codes. A "minimum-redundancy code" will be defined here as an ensemble code which, for a message ensemble consisting of a finite number of members, N, and for a given number of coding digits, D, yields the lowest possible average message length. In order to avoid the use of the lengthy term "minimum-redundancy," this term will be replaced here by "optimum." It will be understood then that, in this paper, "optimum code" means "minimum-redundancy code."

The following basic restrictions will be imposed on an ensemble code:

(a) No two messages will consist of identical arrangements of coding digits.
(b) The message codes will be constructed in such a way that no additional indication is necessary to specify where a message code begins and ends once the starting point of a sequence of messages is known.

Restriction (b) necessitates that no message be coded in such a way that its code appears, digit for digit, as the first part of any message code of greater length. Thus, 01, 102, 111, and 202 are valid message codes for an ensemble of four members. For instance, a sequence of these messages 1111022020101111102 can be broken up into the individual messages 111-102-202-01-01-111-102. All the receiver need know is the ensemble code. However, if the ensemble has individual message codes including 11, 111, 102, and 02, then when a message sequence starts with the digits 11, it is not immediately certain whether the message 11 has been received or whether it is only the first two digits of the message 111. Moreover, even if the sequence turns out to be 11102, it is still not certain whether 111-02 or 11-102 was transmitted. In this example, change of one of the two message codes 111 or 11 is indicated.

C. E. Shannon[1] and R. M. Fano[2] have developed ensemble coding procedures for the purpose of proving that the average number of binary digits required per message approaches from above the average amount of information per message. Their coding procedures are not optimum, but approach the optimum behavior when N approaches infinity. Some work has been done by Kraft[3] toward deriving a coding method which gives an average code length as close as possible to the ideal when the ensemble contains a finite number of members. However, up to the present time, no definite procedure has been suggested for the construction of such a code

* Decimal classification: R531.1. Original manuscript received by the Institute, December 6, 1951.
† Massachusetts Institute of Technology, Cambridge, Mass.
[1] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Jour.*, vol. 27, pp. 398–403; July, 1948.
[2] R. M. Fano, "The Transmission of Information," Technical Report No. 65, Research Laboratory of Electronics, M.I.T., Cambridge, Mass.; 1949.
[3] L. G. Kraft, "A Device for Quantizing, Grouping, and Coding Amplitude-modulated Pulses," Electrical Engineering Thesis, M.I.T., Cambridge, Mass.; 1949.

Huffman算法

重复做：选择权值最小的两个结点生成新结点，新结点作为原结点的父亲，权值是原来两个结点权值之和。

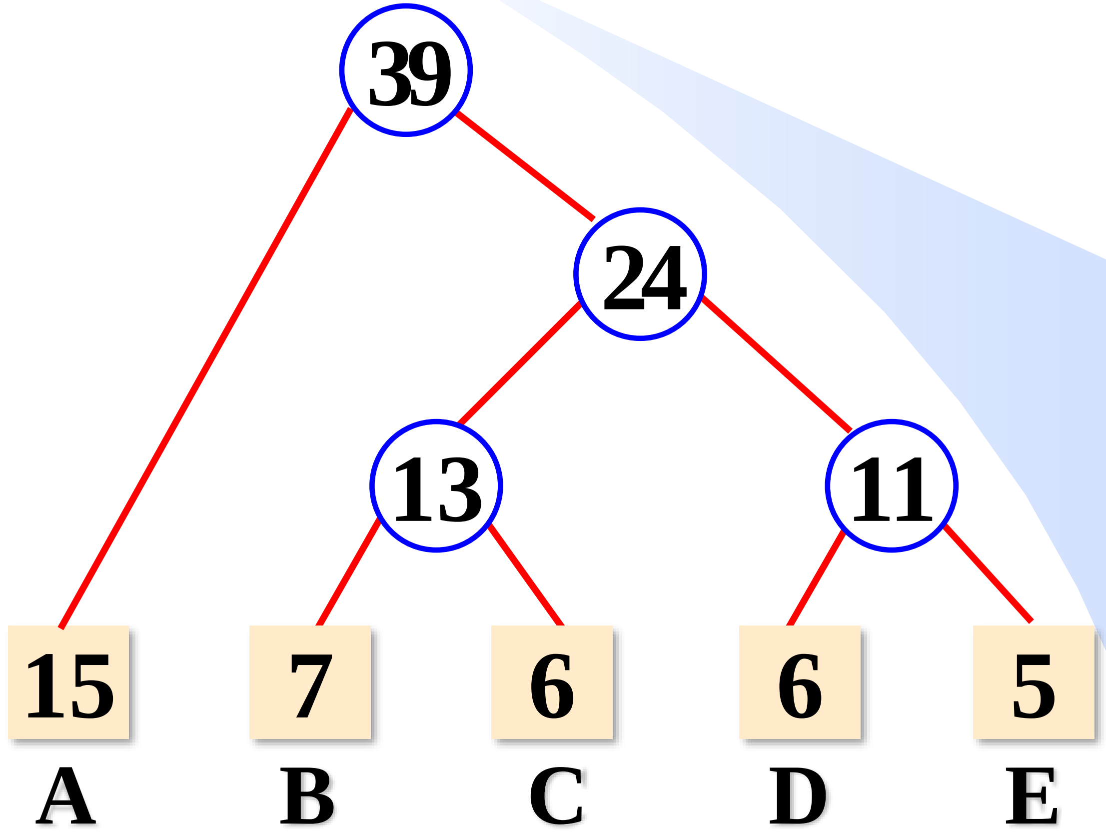
字符	出现次数
A	15
B	7
C	6
D	6
E	5

15	7	6	6	5
A	B	C	D	E

Huffman算法

重复做：选择权值最小的两个结点生成新结点，新结点作为原结点的父亲，权值是原来两个结点权值之和。

字符	出现次数
A	15
B	7
C	6
D	6
E	5

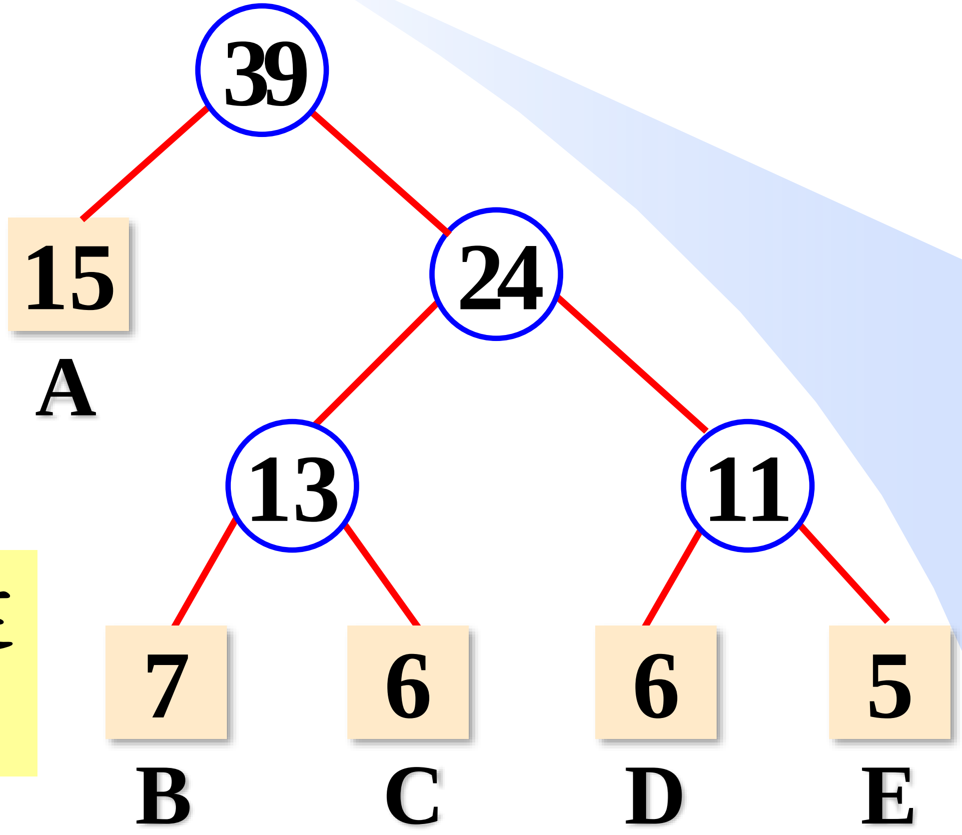


Huffman算法

重复做：选择权值最小的两个结点生成新结点，新结点作为原结点的父亲，权值是原来两个结点权值之和。

字符	出现次数
A	15
B	7
C	6
D	6
E	5

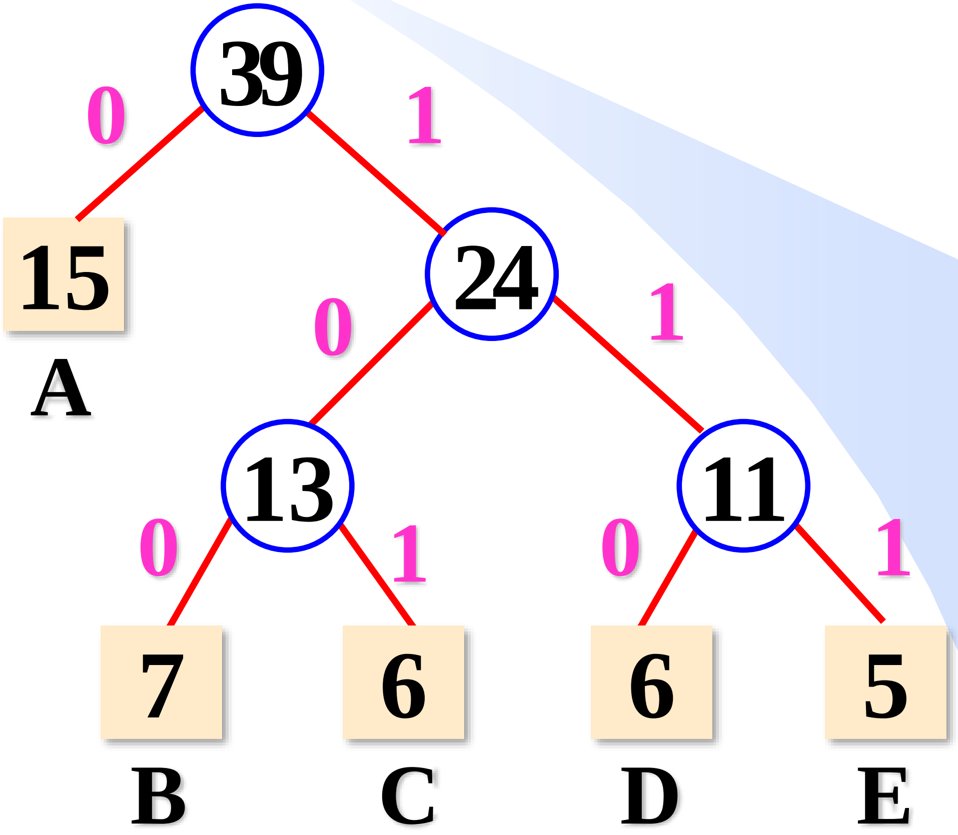
自下而上构建Huffman树



Huffman编码

每个左分支标记0，右分支标记1。把从根到叶的路径上的标号连接起来，作为该叶结点所代表的字符的编码。

字符	出现次数	编码
A	15	0
B	7	100
C	6	101
D	6	110
E	5	111



WPL=15*1+(7+6+6+5)*3=87

Huffman算法

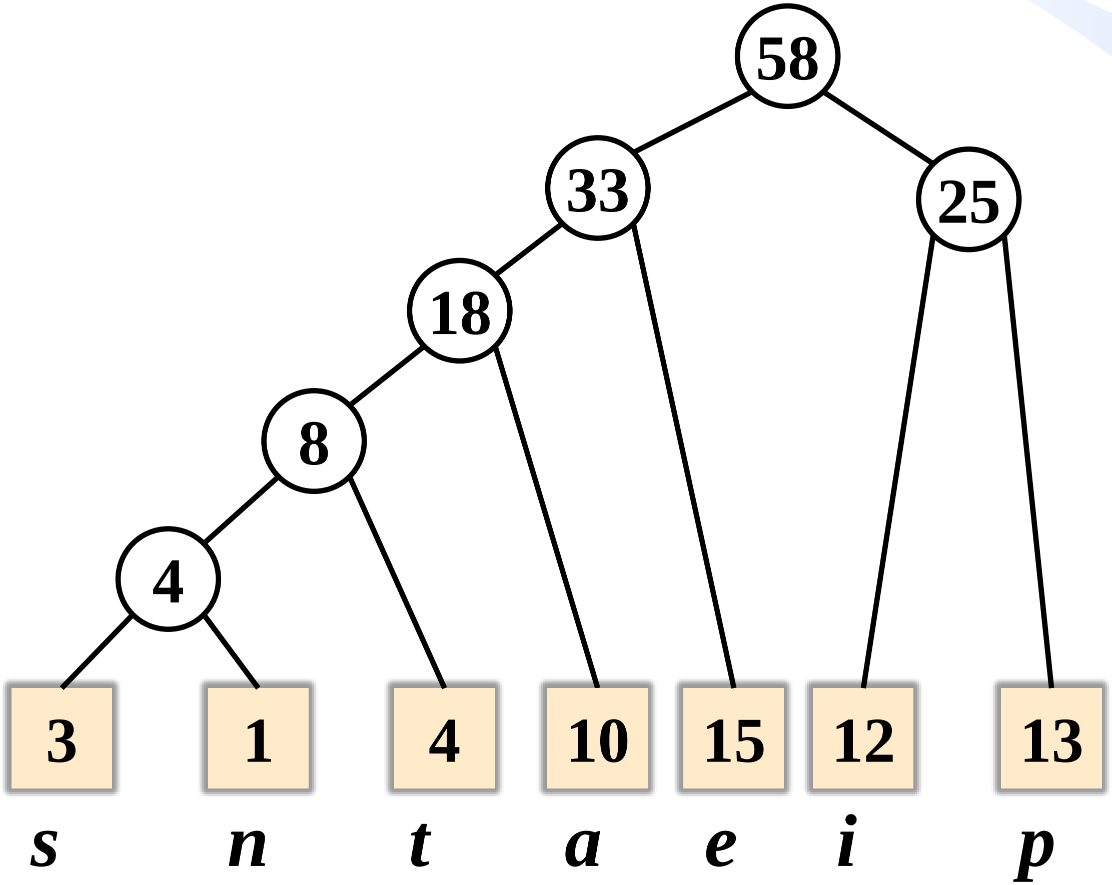
① 根据给定的n个权值$w_1, w_2, \ldots, w_n$构成n棵二叉树$T_1, T_2, \ldots, T_n$，每棵二叉树T_i中都只有一个结点，其权值为w_i；

② 选出权值最小的两个根结点合并成一棵二叉树：生成一个新结点作为这两个结点的父结点，新结点的权值为其两个孩子的权值之和。

③ 重复步骤②，直至只剩一棵二叉树为止，此二叉树便是哈夫曼树。

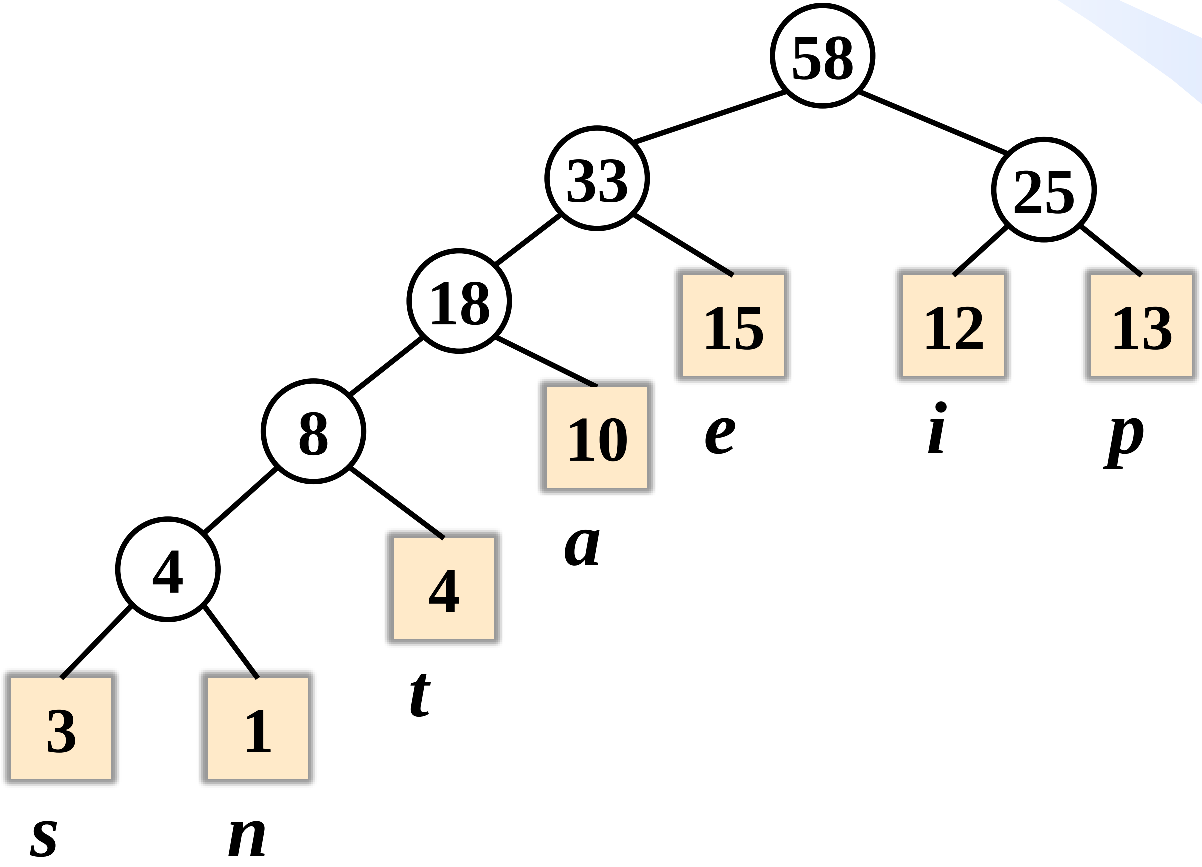
构建Huffman树示例

假设有一文件包含7种字符：*a*、*e*、*i*、*s*、*t*、*p*、*n*，且文件中有10个*a*, 15个*e*, 12个*i*, 3个*s*, 4个*t*, 13个*p*, 1个*n* .



构建Huffman树示例

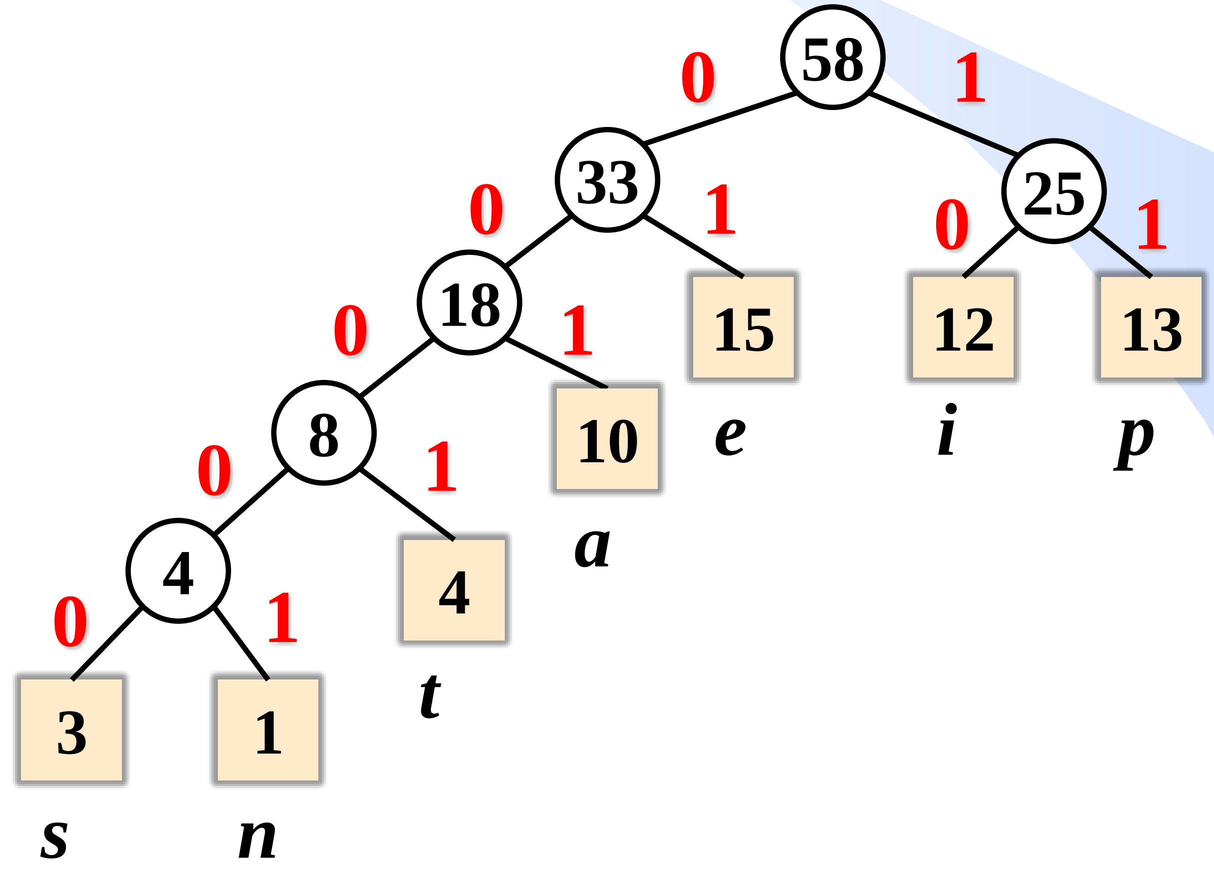
假设有一文件包含7种字符：*a*、*e*、*i*、*s*、*t*、*p*、*n*，且文件中有10个*a*, 15个*e*, 12个*i*, 3个*s*, 4个*t*, 13个*p*, 1个*n* .



Huffman编码

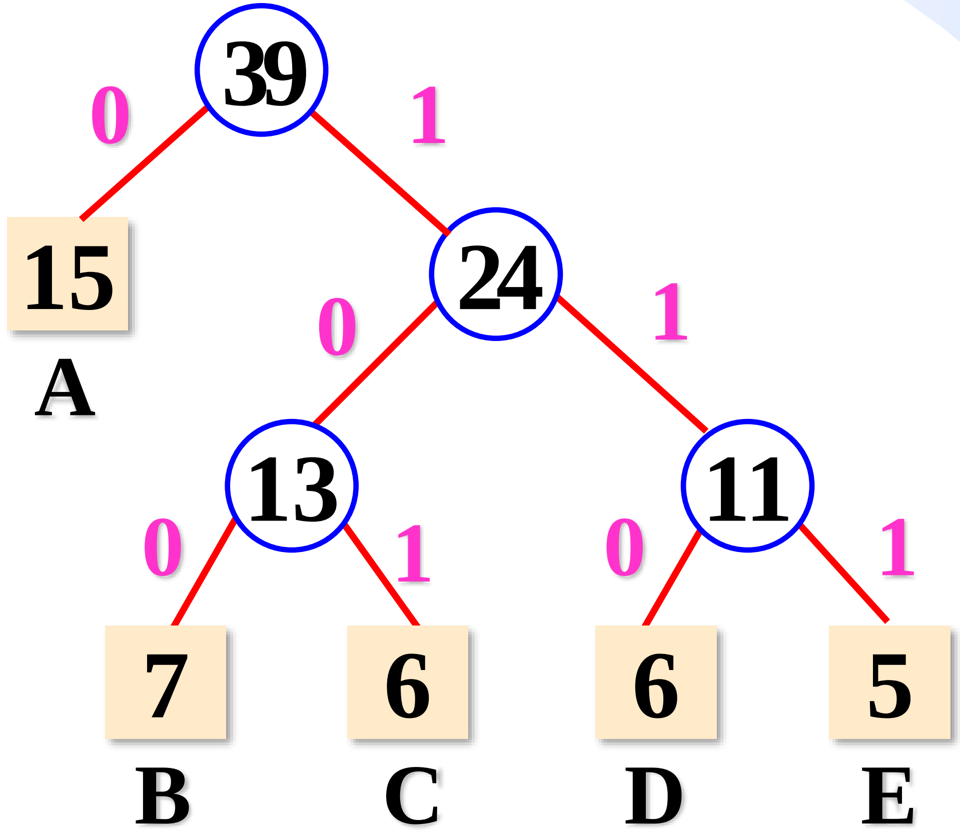
- 对哈夫曼树每个非叶结点的左分支标记0，右分支标记1。
- 把从根到叶的路径上的标号连接起来，作为该叶结点所代表的字符的编码。

*s*的编码是：	00000
n 的编码是	00001
*t*的编码是：	0001
*a*的编码是：	001
*e*的编码是：	01
*i*的编码是：	10
*p*的编码是：	11



哈夫曼编码是否是“无前缀冲突编码”？

字符对应叶结点，任意一个叶结点不可能是其他叶结点的祖先，每个叶结点对应的编码不可能是其他叶结点对应的编码的前缀，故哈夫曼编码是无前缀冲突编码。



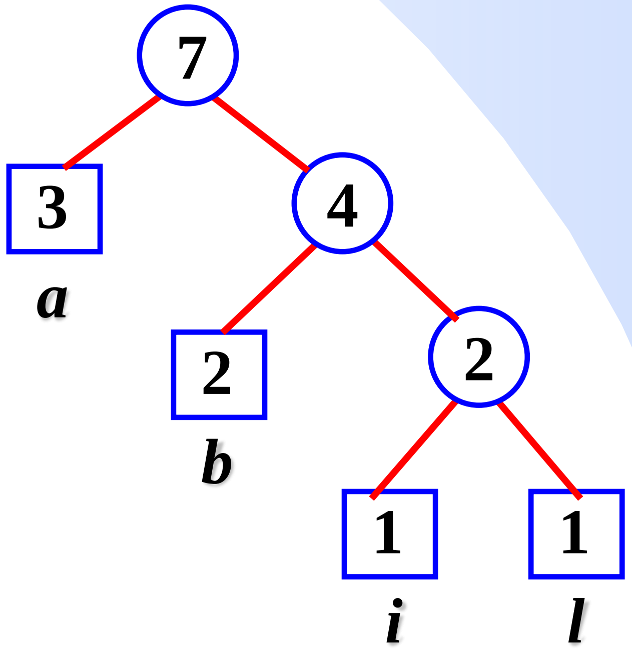
课堂练习

字符串"*alibaba*"的Huffman编码总长度有____位（bit）。【阿里笔试题】

A. 11　　B. 12　　C. 13　　D. 14

文本中有3个*a*，2个*b*，1个*i*，1个*l*

WPL=13



课下思考

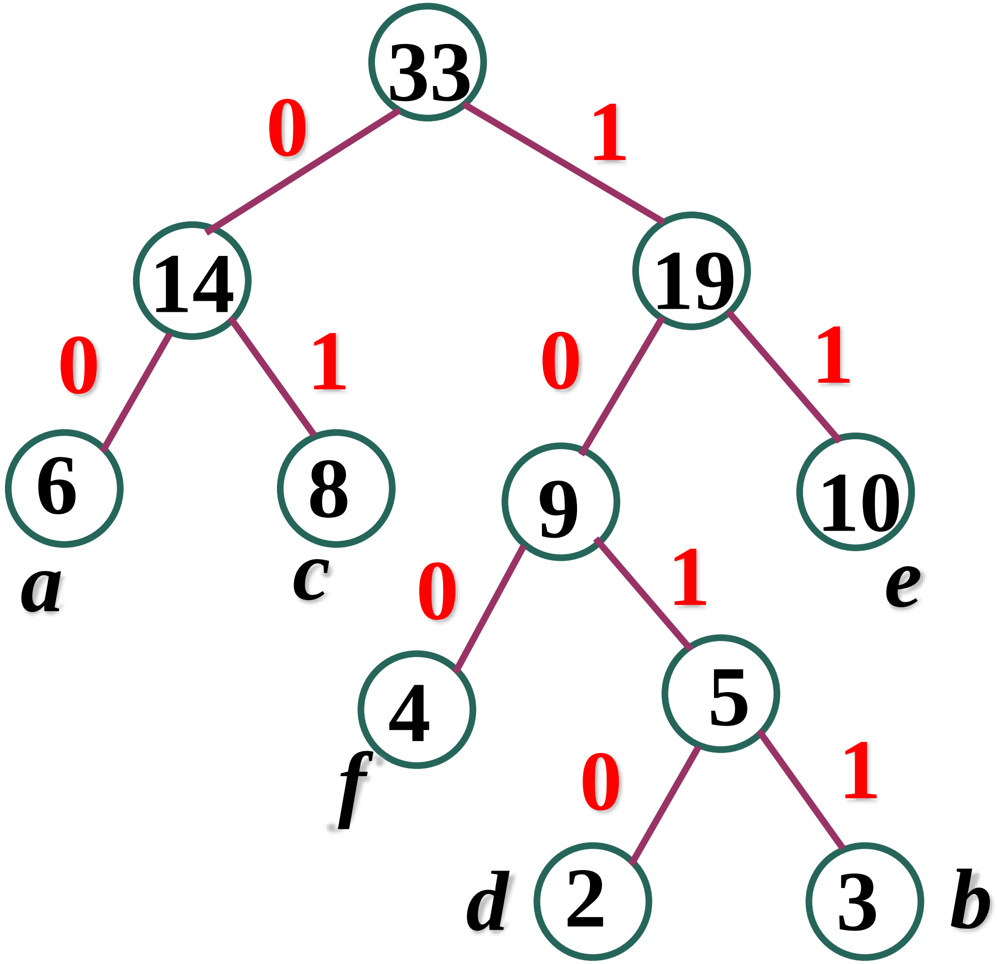
在有6个字符组成的字符集S中，各字符出现的频次分别为3、4、5、6、8、10，为S构造的哈夫曼树的加权路径长度WPL为_____。【2023年考研题全国卷】

A. 86　　B. 90　　C. 96　　D. 99

课下思考

已知字符集合是 { a, b, c, d, e, f }，各个字符出现的次数分别是6，3，8，2，10，4，则各字符对应的哈夫曼编码为__________

【2018年考研题全国卷】



编码

a (00)

b (1011)

c (01)

d (1010)

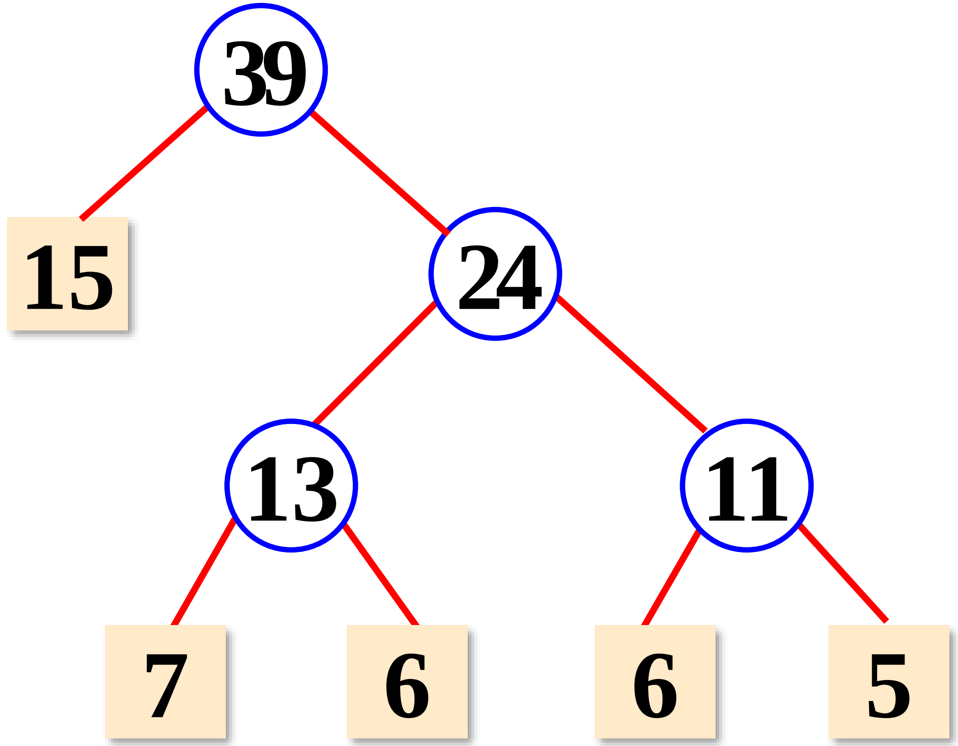
e (11)

f (100)

哈夫曼树——二子性

- 哈夫曼树中包含度为1的结点么?
- 哈夫曼树不包含度为1的结点。
- 若哈夫曼树n个叶结点，则必有n-1个非叶结点，一共2n-1个结点。

引理 $n_0=n_2+1$.



哈夫曼树——同权不同构

- 哈夫曼树、哈夫曼编码、最小编码长度唯一么?
- 哈夫曼树形态不唯一、编码不唯一。
- 对任意内结点而言，其左右子树互换后WPL不变，故最小编码长度唯一。