

## 基于概率补偿的无哈夫曼树变长压缩编码

杨多星, 刘蕴红

(大连理工大学 电气工程学院, 辽宁 大连 116023)

**摘 要:** 现在广泛使用的压缩编码方法都要通过哈夫曼树来实现, 这样围绕着哈夫曼树就存在着许多运算过程. 为了化简编码过程, 提出了一种无需哈夫曼树就能实现的变长最佳编码方法, 通过一个概率补偿的过程, 可以直接得到所有信源的最佳码长. 知道码长和概率后也无需通过哈夫曼树就可以确定最后的编码, 并且可以证明结果满足变长最佳编码定理和前缀编码. 经测试, 该方法可以快速有效得到变长最佳编码, 并简化了变长编码的运算存储过程.

**关键词:** 哈夫曼树; 变长最佳编码; 概率补偿; 前缀编码

**中图分类号:** TP301.6

**文献标识码:** A

**文章编号:** 1000-7180(2011)06-0051-03

## Variable-length Compressed Encoding Without Huffman Tree Based on the Probability Compensation

YANG Duo-xing, LIU Yun-hong

(Department of Electrical and Electronics Engineering, Dalian University of Technology, Dalian 116023, China)

**Abstract:** Nowadays most of the widely used compressions encoding methods are implemented by using the Huffman tree. There are many operational processes around the Huffman tree. One variable-length optimal encoding method is proposed to simplify the encoding process. The optimal code length can be get from the probability compensation process. The final coding can be determined without the Huffman tree after knowing the code length and probability, and the result meets the challenge of the variable-length optimal coding theorem and the prefix code. After testing, the method is proved to be able to get the variable-length optimal coding quickly and effectively. The calculation and stored procedures are simplified.

**Key words:** Huffman tree; variable-length optimal coding; probability compensation; prefix code

### 1 引言

哈夫曼编码是 1950 年由哈夫曼教授研制开发的, 根据字符出现的概率来构造平均长度最短的编码, 是一种变长编码. 它借助了数据结构当中的树型结构, 在哈夫曼算法的支持下构造出一棵最优二叉树, 我们把这类树命名为哈夫曼树. 哈夫曼编码是哈夫曼树的一个应用, 是一种最优的前缀编码技术. 因此, 准确地说, 哈夫曼编码是在哈夫曼树的基础上构造出来的一种编码形式, 它的本身有着非常广泛的应用. 现在大部分的文件、图片、视频压缩都是使用这种方法<sup>[1-3]</sup>.

也有人提出过各种改进的哈夫曼编码方法, 但这些方法基本都离不开哈夫曼树. 既然存在树, 必然存在树的创建、遍历以及存储过程. 那么能不能省略这些过程, 也就是不需要建立哈夫曼树, 依然能够得到最佳编码, 本文提供了一种方案.

### 2 传统哈夫曼编码方法

例 1: 对下列一组字符进行哈夫曼编码  
e、d、a、e、d、b、d、c、d、e、b、c、e、b、a、d  
统计各数据出现概率并由小到大排列:  
a—0.125, c—0.125, b—0.187 5  
e—0.25, d—0.312 5.

创建如图 1 所示的哈夫曼树。

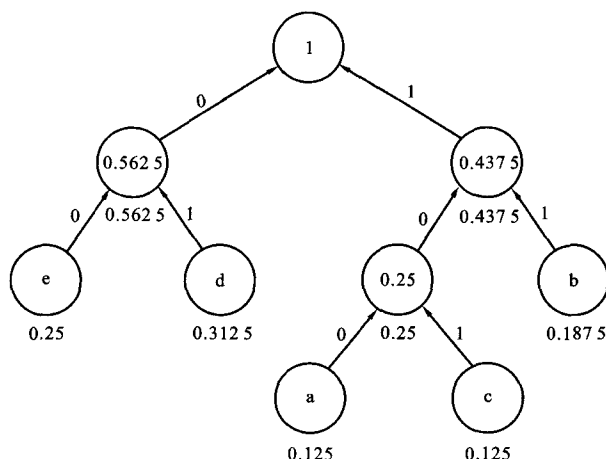


图 1 哈夫曼树

分配码字:

a:100,c:101,b:11,e:00,d:01.

原数据串编码为:00 01 100 00 01 11 01 101 01 00 11 101 00 11 100 01. 使用哈夫曼编码后,码长共 36 位. 可以看出,整个过程最关键的就是创建哈夫曼树<sup>[4]</sup>,下面介绍不使用哈夫曼树的编码方法.

### 3 概率比较确定码长

#### 3.1 标准情况

设需要处理的数据为  $2^n$  个,  $n$  为自然数, 输入信源  $x_i$  出现的概率为  $P_i$ ,  $P_i$  满足  $1/2^n$ , 即都是 2 的整数次幂的倒数, 那么  $x_i$  的码长可以按表 1 由  $P_i$  来直接确定.

表 1 概率码长对应表

| $P_i$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{8}$ | ... | $\frac{1}{2^n}$ | ... |
|-------|---------------|---------------|---------------|-----|-----------------|-----|
| 码长    | 1             | 2             | 3             | ... | $n$             | ... |

例 2: 为 16 字符的数据 f、a、c、f、e、f、f、e、e、f、f、d、e、f、b、f 编码

各字符  $P_i$ :

$$P_a = \frac{1}{16}, P_b = \frac{1}{16}, P_c = \frac{1}{16}, P_d = \frac{1}{16}, P_e = \frac{1}{4},$$

$$P_f = \frac{1}{2}.$$

码长:

a—4, b—4, c—4, d—4, e—2, f—1.

各字符出现的概率和  $\sum P_i = 1$ , 不缺少内容, 所以码长确定完成. 使用哈夫曼树进行验证, 得到的码长相同. 当然, 大多数情况下,  $P_i$  不太可能正好是  $1/2^n$ , 如例 1. 此时, 需要人为令  $P_i$  满足  $1/2^n$ .

#### 3.2 概率补偿

$$\text{如果 } P_i \text{ 满足 } \frac{1}{2^n} < P_i < \frac{1}{2^{n-1}} \quad (1)$$

则令修改后的新概率  $P'_i = 1/2^n$ .  $1/2^n$  表示  $P_i$  此时对应的新概率, 而  $1/2^{n-1}$  表示其对应的上一等级的新概率. 例 1 中各字符出现的概率为

$$d=0.3125, e=0.25, b=0.1875,$$

$$c=0.125, a=0.125.$$

则对应的  $P'_i$  为

$$P'_d = \frac{1}{4}, P'_e = \frac{1}{4}, P'_b = \frac{1}{8},$$

$$P'_c = \frac{1}{8}, P'_a = \frac{1}{8}.$$

因为  $\sum P_i < 1$ , 所以需要进行概率补偿. 令  $E = 1 - \sum P'_i$ ,  $E$  表示换成新概率后的概率损失, 不取绝对值. 当  $P_i$  距它对应的上一等级概率足够近时, 可以对它进行概率补偿. 这个距离可以用  $2P'_i - P_i$  来表示. 补偿流程如图 2 所示.

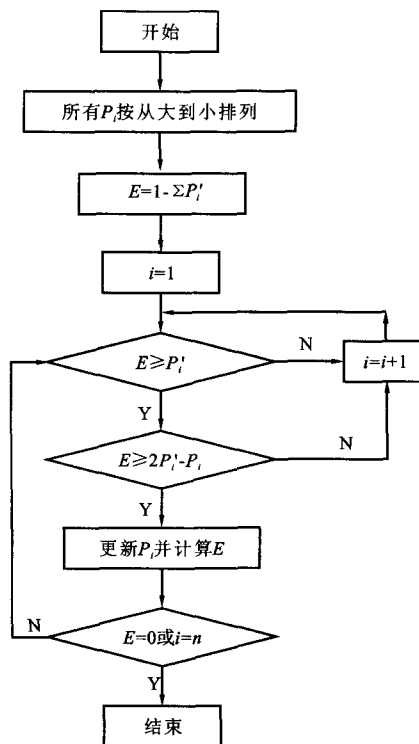


图 2 概率补偿流程图

(1) 所有  $P_i$  按从大到小排列并计算  $E = 1 - \sum P'_i$ ;

(2) 令  $i = 1$ ;

(3) 当  $E \geq P'_i$  时, 执行(4), 否则,  $i + 1$  重复(3);

(4) 如果  $E \geq 2P'_i - P_i$ , 更新  $P'_i$ , 令  $P'_i -$

$1/2^{n-1}$ ,即概率乘以 2 提高一个级别,当前字符的码长缩短一位;如果  $E < 2P'_i - P_i$ ,不更新  $P'_i$ ,并令  $i + 1$ ;

(5) 重新计算  $E$ ,如果  $E = 0$  或全部信源都处理过了,结束概率补偿. 否则,回到(3).

根据上述过程,例 1 补偿后的概率为

$$P'_d = \frac{1}{4}, P'_e = \frac{1}{4},$$
$$P'_b = \frac{1}{4}, P'_c = \frac{1}{8}, P'_a = \frac{1}{8}$$

概率补偿后,就转化成了标准情况,并且一定满足  $\sum P'_i = 1$ . 再根据表 1,得到各字符的码长分别为 d:2 位,e:2 位,b:3 位,c:3 位,a:3 位.

4 码字分配

确定各字符码长后可按下面的规则直接分配码字(图 3).

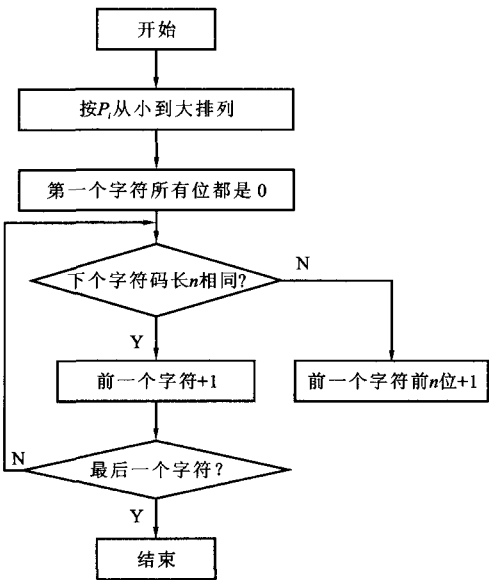


图 3 码字分配的流程图

- (1) 将所有字符按  $P_i$  从小到大排列,第一个字符( $P_i$  最小)所有位都为 0;
  - (2) 把码长相同的字符作为一组,同一组内的每个字符码字依次加 1;
  - (3) 每组最后一个字符的码字,取前  $n$  位( $n$  是下一组的码长)再加 1,就得到下一组第一个字符的码字.
- 得到的结果最后一个字符( $P_i$  最大)所有位必然都为 1.

举例说明,在例 1 中,码长为 3 的字符有两个,依次是 a,c,码字依次为 000、001. 码长为 2 的字符有

三个,依次是 b,e,d,码字依次为 01、10、11. 所以最后的码字分配结果是:

$$a:000,c:001,b:01,e:10,d:11.$$

在例 2 中,码长为 4 的字符有四个,依次是 a,b,c,d,码字依次为 0000、0001、0010、0011. 码长为 2 的字符有一个,是 e,码字为 01. 码长为 1 的字符有一个,是 f,码字为 1. 所以最后的码字分配结果是:

$$a:0000,b:0001,c:0010,d:0011,e:01,f:1.$$

可以看出,它们都是前缀编码.

5 有效性证明

使用哈夫曼树主要有两个用处,一是能保证得到的码长是最佳,二是能保证分配的码字都是前缀编码,那么只要新方法也能满足码长最佳和前缀编码即可.

5.1 变长最佳编码证明

根据变长最佳编码定理:在变长编码中,对出现概率大的信息符号赋予短码字,而对于出现概率小的信息符号赋予长码字. 如果码字长度严格按照所对应符号出现概率大小逆序排列,则编码结果平均码字长度一定小于任何其他排列方式.

证明:如果  $P_m > P_n$ ,按照式(1) 修改后,  $P'_m \geq P'_n$ . 概率补偿按照从大概率到小概率的顺序进行操作,只有当  $P'_m$  处理完后才会处理  $P'_n$ ,如果  $P'_m$  没有提高一个级别,  $P'_n$  必然也不会提高一个级别. 所以当处理结束后的,  $P'_m$  依然大于等于  $P'_n$ . 由表 1 可知,  $x_m$  的码长一定小于等于  $x_n$  的码长.

码字长度严格按照所对应符号出现概率大小逆序排列,满足变长最佳编码定理,所以新方法是最佳编码.

5.2 前缀编码证明

如果不是前缀编码,那么码长相对短的码字将会与某些码长相对长的码字的开头部分相同,在本文的方法中,就是指某一个字符的码字与它之前的某一组的码字开头部分相同.

证明:根据码字分配规则第 3 条:每组最后一个字符的码字,取前  $n$  位( $n$  是下一组的码长) 再加 1,就得到下一组第一个字符的码字. 很明显相邻两组间,后一组的所有码字都会比前一组的前  $n$  位码字至少大 1. 而不相邻的两组间,相差则更大,绝不可能与前面某一组的码字开头部分相同.

所以一定是前缀编码.

(下转第 57 页)

与实际情况更接近。

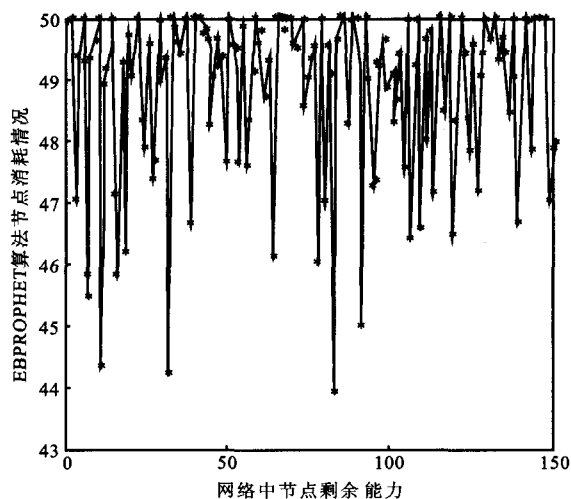


图3 EBPPOPHET算法中节点消耗能量情况

#### 参考文献:

- [1] 熊永平,孙利民,牛建伟,等. 机会网络[J]. 软件学报, 2009,20(1):124-137.
- [2] Pelusil, Passarella A, Contim, et al. Opportunistic net-working: Data forwarding in disconnected mobile Ad Hoc networks [J]. Communication Magazine, 2006,44 (11):134-141.
- [3] Becker VD. Epidemic routing for partially connected ad hoc networks. Technique Report, CS-2000-06 De-

partment of Computer Science[R]. Durham, NC; Duke University, 2000.

- [4] 李炯,汪文勇,潘家根. 无线传感器网络洪泛路由研究[J]. 计算机科学,2006(5).
- [5] 方敏,刘新华,旷海兰. 无线传感器网络中一种高能效低延时的泛洪算法研究[J]. 通信学报,2007(8).
- [6] Spyropoulos T, Psounis K, Raghavendra CS. Spray and focus: Efficient mobility-assisted routing for heterogeneous and correlated mobility[C]// Proc of the IEEE PerCom Workshop on Intermittently Connected Mobile Ad Hoc Networks. [s. l.]:IEEE, 2007.
- [7] Lindgren A, Doria A, Schelén O. Probabilistic routing in intermittently connected networks[J]. ACM Sigmobile Mobile Computing and Communications Review, 2003,7(3):19-20.
- [8] Broch J, Maltz DA, Johnson DB, et al. ACM/IEEE A performance comparison of multi-hop wireless ad hoc network routing protocols[C]// Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking. NY:ACM,1998.

#### 作者简介:

**袁俊玲** 女,(1985-),硕士.研究方向为机会网络、无线传感器网络.

**陈莉** 女,(1985-),硕士.研究方向为计算机网络、图像处理.

(上接第53页)

## 6 结束语

新的编码方法在储存时只要按  $P'_i$  从小到大的顺序依次记录每个字符,在记录每组第一个字符的码字及这一组的字符个数即可<sup>[5]</sup>.省略了哈夫曼树的创建、遍历、存储等过程,节省了运算步骤和存储空间.经多种实例检验,该方法实用有效.

#### 参考文献:

- [1] 王防修,周康. 通过哈夫曼编码实现文件的压缩与解压[J]. 武汉工业学院学报,2008,27(4):46-49.

[2] 胡波,王宏远. MPEG 编码算法的优化[J]. 微电子学与计算机,2008,25(4):201-204.

[3] 田端财,殷晓丽. 基于哈夫曼编码的图像压缩技术研究[J]. 科技资讯,2009(8):29-30.

[4] 康洪波. 静态哈夫曼编码的原理及应用[J]. 河北建筑工程学院学报,2009(1):125-126.

[5] 王群芳. 哈夫曼编码的另一种实现算法[J]. 安徽教育学院学报,2006,24(6):36-38.

#### 作者简介:

**杨多星** 男,(1985-),硕士研究生.研究方向为楼宇自动化.