关于 Huffman 编码的一个注记

林嘉宇[1,2],刘 荧[1]
(1. 国防科技大学电子科学与工程学院,湖南长沙 410073;2. 电子科技大学光纤国家重点实验室,四川成都 610054)

摘 要: Huffman 编码是无损压缩中的重要方法,在数据压缩、音频编码、图像编码中得到广泛的应用. 除了压缩效率以外,作为变长码的 Huffman 编码,还有其他的判断其编码优劣的准则,例如码方差、抗误码的能力等. 本文讨论 Huffman 编码后的码流中 0、1 码元(二进制情况下)出现的概率问题. 研究结果表明,通常的经典 Huffman 编码的 0、1 码元出现的概率差最大,在出现概率均衡准则下的性能最劣. 文章进行了严格的数学建模,并给出了一种算法,可以使编码后码流中 0、1 码元的分布概率(趋向)均等;并且,算法可在原 Huffman 编码中结合进行,所增加的计算量很小. 文章最后进行了实验验证.

关键词: Huffman 编码; 0、1 码元概率
中图分类号: TN911.21 **文献标识码**: A **文章编号**: 0372-2112 (2003) 04-0602-03

A Note on Huffman Coding

LIN Jia-yu[1,2], LIU Ying[1]
(1. *National University of Defence Technology, Changsha, Hunan* 410073, *China*;
2. *National Key Lab of Broadband Optical Transmission and Communication Networks, UESTC, Chengdu, Sichuan* 610054, *China*)

Abstract: Huffman codes are most prevalent in practical operation. Other than compressing efficiency, there are other criteria, such as codewords' MSE, channel bit-error resilience, etc. This paper discusses the problem of the probabilities of 0, 1 bits in codeword stream after Huffman coding. Mathematical modeling and analysis is given, which shows that accustomed Huffman codes are the worst according to this criterion. The optimal case is derived, and a sub-optimal algorithm is constructed to make the probabilities of 0, 1 bits equal. The algorithm can be included in the original Huffman coding, and the complexity added is small. Experiments are given in the end, which attest our ideas.

Key words: Huffman codes; probabilities of 0、1 bits

1 引言

Huffman 编码是无损压缩中的重要方法. 自 1952 年 Huffman 发表其相关论文[1]以来,Huffman 编码在数据压缩、音频编码、图像编码中得到广泛的应用,例如,MPEG1 音频标准的 LayerIII[2](即著名的 MP3 音频压缩建议)、H.263 视频编码标准[3]中都使用 Huffman 编码,作为其码流上信道之前的无损符号编码方式.

Huffman 编码的关键在于其码本的建立,即,给出信源符号所对应的变长码字. 码本的建立包含先后两个步骤. 首先,根据信源符号的概率分布构建 Huffman 码树;然后,对 Huffman 码树分配码字元素.

Huffman 码是紧致码(compact codes),在无损符号编码中具有最高的压缩效率. 但除了压缩效率之外,对于 Huffman 码,还有其他判断其编码优劣的准则,例如码方差[4]、抗误码的能力[5]、码字自身的同步能力[6]、便于快速搜索且减少所需内存[7]等;基于这些判断准则的优的 Huffman 码,近几年来仍受到研究者们的重视.

收稿日期:2001-08-02;修回日期:2002-07-12
基金项目:宽带光纤传输与通信系统技术国家重点实验室开放基金

本文讨论 Huffman 编码后的码流中 0、1 码元(二进制情况下)的出现概率问题. 一般情况下,码流中的 0、1 码元越平均,对码流的信道传输越好. 例如[8],码流中 0、1 码元的平均出现,可使数字传输系统对各种数字信息具有透明性,可改善位定时恢复的质量、使信号频谱弥散而保持恒稳(数字基带码的谱特性与信源分布概率相关)、改善帧同步及时域均衡等子系统的性能.

但是,一般的 Huffman 编码并未考虑码流中 0、1 码元的出现概率问题,本文将讨论这一问题. 以下的研究结果将表明,通常的经典 Huffman 编码的 0、1 码元出现的概率差最大,在此准则下的性能最劣. 本文给出了一种算法,可以使得编码后码流中 0、1 码元的分布概率(趋向)均等;并且,算法可在原 Huffman 编码中结合进行,所增加的计算量很小. 文章最后进行了实验验证.

2 新码的构造

2.1 术语

对如图 1 所示的 Huffman 码树,我们称节点 A、B 及 F 组

成一个树丫,用虚线框表示.图中共有四个树丫,我们按其在 Huffman 码树建立时出现的先后次序,编号为树丫 1、2、3 及 4.1 号树丫中,F 节点称父节点,A、B 节点称子节点,父节点的概率为子节点的概率之和.父节点和子节点的连接称为树枝,树枝的子节点对应的概率称为枝条重.图中,在树丫中位于上方的树枝称为上树枝,下方的树枝称为下树枝.我们按标准 Huffman 编码过程规定,上树枝重不小于下树枝重,并且,为减少 Huffman 码的码方差,如果两枝条重相同,则多节点合并所成的节点应位于上树枝[4].另外,根据惯例,我们称待编码信源符号的节点 A 至 E 为叶节点,4 号树丫的父节点 I 为根节点,根节点所对应的概率为 1.

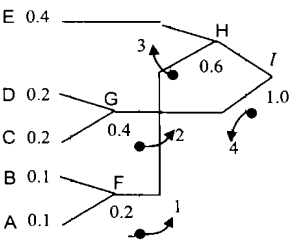

图 1　*Huffman* 码树及其相关术语

2.2　关于新码的数学分析

假设待编码信源符号的个数为 M,各信源符号对应的出现概率为 p_i, $\sum_{i=1}^{M} p_i = 1$,所编码字的码长为 l_i.设 Huffman 码字的平均码长为 $\bar{l}$,有:

$$\bar{l} = \sum_{i=1}^{M} p_i l_i \tag{1}$$

再设长为 l_i 的码字中,‘1’元的个数为 $l_i^{(1)}$,‘0’元的个数为 $l_i^{(0)}$,于是上式变为,

$$\bar{l} = \sum_{i=1}^{M} p_i l_i^{(1)} + \sum_{i=1}^{M} p_i l_i^{(0)} \tag{2}$$

定义码字中‘1’元的平均数为 $\bar{l}^{(1)}$,称为‘1’元的平均长度,‘0’元的平均数为 $\bar{l}(0)$,称为‘0’元的平均长度.我们先考虑‘1’元的平均长度问题.

可以认为每产生一个树丫,都意味着两个(子)节点合并为一个(父)节点,参加以后的码树建立.M 个初始(叶)节点,剩下一个(根)节点,需要经过 $M-1$ 次的树丫产生过程.于是,树丫的个数为 $M-1$.

这样,‘1’元的平均长度以树丫为出发点,有:

$$\bar{l}^{(1)} = \sum_{i=1}^{M} p_i l_i^{(1)} = \sum_{i=1}^{M} p_i \left(\sum_{j=1}^{M-1} T_{ij}^{(1)} \right) \tag{3}$$

其中,在叶节点 i 至根节点的连接中,如果经过了第 j 个树丫,并且所分配的码元为 1,则 $T_{ij}^{(1)} = 1$,其他情况则 $T_{ij}^{(1)} = 0$.进而,式(3)变为:

$$\bar{l}^{(1)} = \sum_{j=1}^{M-1} \sum_{i=1}^{M} \left(T_{ij}^{(1)} p_i \right) \tag{4}$$

式(4)表示,从树丫 j 的角度考虑,哪些叶节点至根节点的连接会经过此树丫 j,并被分配了码元‘1’?显然,树丫 j 的分配了码元‘1’的枝条的重,就是 $\sum_{i=1}^{M} \left(T_{ij}^{(1)} p_i \right)$,记为 $p_{B_j}^{(1)}$.于是,式(4)变为:

$$\bar{l}^{(1)} = \sum_{j=1}^{M-1} p_{B_j}^{(1)} \tag{5}$$

关于‘0’元的分析及相应符号的定义可以类比进行,我们不再赘述.

这样,码元‘1’元的平均长度 $\bar{l}^{(0)}$ 同‘0’元的平均长度 $\bar{l}^{(0)}$ 之差为:

$$| \Delta \bar{l}^{(1-0)} | = \left| \sum_{j=1}^{M-1} \left(p_{B_j}^{(1)} - p_{B_j}^{(0)} \right) \right| \tag{6}$$

其中,$p_{B_j}^{(0)}$ 为 $\sum_{i=1}^{M} \left(T_{ij}^{(0)} p_i \right)$.由我们的目标,应让 $| \Delta \bar{l}^{(1-0)} |$ 尽量小.

在 Huffman 码本的建立过程中,构建 Huffman 码树时,各树丫的枝条重可以随树丫的生成而得到;我们就可以利用式(6),在分配码元时,适当调整上枝条及下枝条所对应的子节点分配‘1’元或‘0’元,以求最后形成的 Huffman 码,其‘1’、‘0’元的出现概率趋向均衡.而我们知道,在每个树丫中,‘1’元及‘0’元的分配,对上枝条及下枝条,虽然在通常的 Huffman 码中,是固定的;但是,其实是可以随机变化的,并不影响所生成的 Huffman 码的异字头性质.

定义各树丫的枝条重的差别为 d_j,

$$d_j = | p_{B_j}^{(+)} - p_{B_j}^{(-)} | , \quad 1 \leq j \leq M-1 \tag{7}$$

$p_{B_j}^{(+)}$ 及 $p_{B_j}^{(-)}$ 分别为树丫 j 的上枝条及下枝条重,根据 Huffman 码树的构建,有:

$$p_{B_j}^{(+)} \geq p_{B_j}^{(-)} \tag{8}$$

我们的目标为,求 $\{s_j\}$,s_j 为 1 或 -1,使得:

$$\min_{\{|s_j|=1\}} \left| \sum_{j=1}^{M-1} s_j d_j \right| \tag{9}$$

对通常的 Huffman 编码,每个树丫中,‘1’元及‘0’元的分配对上枝条及下枝条是固定,不妨设,

$$p_{B_j}^{(1)} = p_{B_j}^{(+)} , \quad p_{B_j}^{(0)} = p_{B_j}^{(-)} \tag{10}$$

即:

$$s_j = 1 , \quad 1 \leq j \leq M-1 \tag{11}$$

显然,此时式(6)最大,和式(9)的目标相悖,是码元分配的最劣方案.

而对式(9)的最优情况,可在 $\{s_j\}$ 的 2^{M-1} 种情况中搜索.当 M 较大时,全搜索不大可能.下面,我们构造一种局部最优的算法,一般即可达到实际的全局最优.

2.3　算法构造

从以上的分析可知,在 Huffman 码树构建完毕之后,进行码元分配时,可在每个树丫中进行‘1’元及‘0’元的适当调整,使得最终的 Huffman 码中,‘1’元及‘0’元的出现概率较为均衡.待优化目标函数为式(9).我们构造局部最优算法.

先将 $\{d_j\}$,$1 \leq j \leq M-1$ 按降序排列,记为 $\{e_j\}$,$M-1 \geq j \geq 1$,即:

$$e_1 \leq e_2 \leq \cdots \leq e_{M-1} \tag{12}$$

问题转为求:

$$\min_{\{|r_j|=1\}} \left| \sum_{j=1}^{M-1} r_j e_j \right| \tag{13}$$

不妨令 $r_{M-1} = 1$,则 $r_{M-1} e_{M-1} \geq 0$;试 $r_j = -1$,$j = M-2, \cdots$,直至:$\sum_{j=m_0+1}^{M-1} r_j e_j \geq 0$ 且 $\sum_{j=m_0}^{M-1} r_j e_j \leq 0$,由此确定 $r_j = -1$,$j = M-2, \cdots$,

m_0;然后再试 $r_j = 1, j = m_0 - 1, \cdots$,直至: $\sum_{j=m_1+1}^{M-1} r_j e_j \leq 0$ 且 $\sum_{j=m_1}^{M-1} r_j e_j \geq 0$,由此确定 $r_j = 1, j = m_0 - 1, \cdots, m_1$;如此以往,直至 $j = 1$. 我们得到相继分段为 1 及 - 1 的$\{r_j\}$,从而就求出了相对应的 $\{s_j\}$.

我们假设,当 s_j 为 1 时,依通常的 Huffman 码元分配,给码树 j 的上枝条分配码元'1',下枝条分配码元'0';那么,当 s_j 为 - 1 时,则反之,给码树 j 的上枝条分配码元'0',下枝条分配码元'1'.

很明显,此算法是非常简单的. 除了对$\{d_j\}$排序之外,算法仅仅进行 $m - 2$ 次的加法及比较. 算法可以嵌在经典的 Huffman 码元分配中进行.

3 实验

由式(2),码流中'1'元及'0'元出现的概率分别为:

$$p^{(1)} = \sum_{i=1}^{M} p_i l_i^{(1)} / \bar{l}, \quad p^{(0)} = \sum_{i=1}^{M} p_i l_i^{(0)} / \bar{l} \qquad (14)$$

对图 1 的例子,各树丫的枝条重之差为 $d_1 = 0, d_2 = 0, d_3 = 0.2, d_4 = 0.2$,根据我们的算法,可得到 $s_1 = 1, s_2 = 1, s_3 = -1, s_4 = 1$. 于是,我们得到的 Huffman 码字,在分配码元时,于树丫 3 处需将'1'、'0'倒位调节. 则我们所得到的 Huffman 码字,其 $p^{(1)}$和 $p^{(0)}$之差为 0;而通常的 Huffman 码字,其 $p^{(1)}$及 $p^{(0)}$分别为 59.1 %及 40.9 %,此性能最劣.

我们再给出一个复杂一些的例子. 表 1 是带空格的英文字母表[6],我们给出了其概率、经典的 Huffman 码(Code 1)及我们所求的 Huffman 码(Code 2). 两个 Huffman 码的平均码长相同,为 4.1195;都尽量让码方差小(文献[6]中的 Huffman 码,其平均码长亦为 4.1195,但其码方差为 1.1316,构造码字时,未试图将码方差尽量减小),为 1.129. 但是,经典的 Huffman 码,其 $p^{(1)}$及 $p^{(0)}$分别为 54.732 %及 45.268 %,两者相差 9.464 %;而我们的 Huffman 码,$p^{(1)}$及 $p^{(0)}$都接近 50 %,两者仅差 0.0024 %. 我们的结果在此性能上比经典的 Huffman 码优了许多.

事实上,全局搜索的结果表明,0、1 元概率差最大的正是经典的 Huffman 码,而在以上的两个例子中,我们的结果和全局最优的结果相同.

表 1 带空格的英文字母表及其 Huffman 码

字母	space	*e*	*t*	*a*	*o*	*i*	*n*
概率	0.1859	0.1031	0.0796	0.0642	0.0632	0.0575	0.0574
Code 1	111	010	1101	1011	1001	0111	0110
Code 2	100	001	1010	1101	1110	0000	0001
字母	*s*	*r*	*h*	*l*	*d*	*u*	*c*
概率	0.0514	0.0484	0.0467	0.0321	0.0317	0.0228	0.0218
Code 1	0011	0010	0001	10101	10100	00001	00000
Code 2	0101	0100	0110	11000	11001	01111	01110
字母	*f*	*m*	*w*	*y*	*p*	*g*	*b*
概率	0.0208	0.0198	0.0175	0.0164	0.0152	0.0152	0.0127
Code 1	110011	110010	110001	100011	100010	100001	100000
Code 2	101101	101100	101111	111110	111111	111101	111100
字母	*v*	*k*	*x*	*j*	*q*	*z*	
概率	0.0083	0.0049	0.0013	0.0008	0.0008	0.0005	
Code 1	1100000	11000011	1100001011	1100001010	1100001001	1100001000	
Code 2	1011101	10111001	1011100001	1011100000	1011100010	1011100011	

4 结论

本文研究了编码后码流中 1、0 元出现概率趋向均衡的 Huffman 码. 就该问题建立了数学模型,进行了严格的数学分析. 结论表明,通常的 Huffman 码,1、0 元出现概率差是最大的. 文章构造了一种局部最优算法,可使得 1、0 元的概率(趋向)均等;并且,算法可在原 Huffman 编码中结合进行,所增加的计算量很小. 实验结果证实了我们的结论.

参考文献:

[1] Huffman D A. A method for the construction of minimum redundancy codes [J]. Proc IRE,1952,40(9):1098 - 1101.

[2] ISO/IEC Int'l Standard IS11172 - 3. Information Technology-Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5Mbit/s,Part3:Audio [S]. 1993.

[3] ITU-T Recommendation H.263. Video Coding for Low Bit Rate Communication[S]. 1998.

[4] 周荫清. 信息论基础[M]. 北京:北京航空航天大学出版社,1993.

[5] Takishima Y,Wada M,Murakami H. Reversible variable length codes [J]. IEEE Tr on Commu,1995,43(3):158 - 162.

[6] Escott A E,Perkins S. Binary Huffman equivalent codes with a short synchronizing codeword [J]. IEEE Tr on IT,1998 44(1):346 - 351.

[7] Hashemian R. Memory efficient and high-speed search Huffman coding [J]. IEEE Tr on Commun,1995,43(10):2576 - 2581.

[8] 曹志刚. 现代通信原理 [M]. 北京:清华大学出版社,1992.

作者简介:

林嘉宇 男,1973 年 2 月生于福建平潭,1998 年于国防科大获工学博士学位,研究领域为非线性信号处理、信源编码、信源信道联合编码、通信中的信号处理等.

刘 荧 女,1973 年 6 月生于湖南长沙,2000 年于国防科大获工学博士学位,研究领域为非线性信号处理、电磁场与微波等.