

基于多参数的数据压缩算法

高怀远, 陈英豪

(上海大学 自动化系, 上海 200072)

摘要:通过对 Huffman 编码方法的研究,文中提出了一种基于多参数的数据无损压缩算法。基于原始数据集的元素个数统计,对原始数据集进行多次的合并,使合并后所得到的新数据集满足 Huffman 最佳编码要求,由此生成规模较小的数据合并对应表,并将数据编码分为一元即时码(前缀)和区分码(后缀)两个部分。数据多次合并的不同起始点为文中无损压缩方法的多参数,利用这些参数结合编码前缀及后缀即可唯一表示原始数据,去除了编码表。解码时无需逐位匹配即可复原原始数据。与传统方法相比,文中构造的基于多参数的数据无损压缩方法,编码结构简单,运算开销小,编解码效率较高。

关键词:无损压缩;元素合并;1 元即时码;区分码

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2014)09-0041-04

doi:10.3969/j.issn.1673-629X.2014.09.009

A Lossless Compression Algorithm Based on Multi-parameter

GAO Huai-yuan, CHEN Ying-hao

(Automation Department of Shanghai University, Shanghai 200072, China)

Abstract: According to the study and analysis of Huffman coding method, propose a kind of lossless compression algorithm which is based on multi-parameter. Through sort and statistical for the number of original data, then merge them to meet the requirement of best Huffman encoding, thereby generating a data merging table which occupies less space, and encode the original data which is divided to one-unit code (prefix code) and distinction code (suffix code). The start point of the data merging is the multi-parameter in this research. The original data can be determined by using these parameter. There is no need to bit by bit matching or generating encoding table when decoding. Compared with the original method, the lossless compression algorithm which is based on multi-parameter has simple coding structure and operating. It has higher efficiency in both coding and decoding.

Key words: lossless compression; element merging; one-unit code; distinction code

0 引言

如今科学技术日新月异,通讯技术、互联网技术越来越发达,这对于数据的传输也有着越来越高的要求^[1]。高清图像,视频,以及音乐等资源占有大量的数据,这对于传输带宽来说是一个巨大的挑战,因此对于数据压缩也有着更高的要求。好的压缩算法能够有效地提高通讯效率^[2]。

数据压缩大体上可分为两类,有损压缩和无损压缩^[3]。

有损压缩是让数据在压缩的时候损失一部分数据。虽然可以达到较高的压缩比,但是在解码时无法还原原有的数据。而且经过多次压缩的数据会与原始数据有越来越大的差异,使其精度大为降低,这对于数

据精度要求高的领域是不可取的^[4-8]。

无损压缩可将数据的数据量有效减少,同时保证数据的精确程度,并且能够完全还原原始数据。数据精度的还原非常重要,特别是在医学成像、天文观测、风洞数据的采集、金融等领域^[9]。

Huffman 编码是常见的无损压缩编码,它是一种变长无失真信源编码方法。首先将信源符号按照概率从大到小的顺序排列,然后给两个概率最小的信源符号各分配一个码位 0 和 1,将这两个信源合并成一个新的符号,并用这两个最小概率之和作为新符号的概率,结果就得到了比原信源缩减了一个信源符号的新信源,再将新信源按照概率从大到小顺序排列并重复上述步骤,最终缩减的信源只剩下两个符号,然后从最

收稿日期:2013-10-27

修回日期:2014-01-26

网络出版时间:2014-05-21

基金项目:国家自然科学基金资助项目(71201097)

作者简介:高怀远(1988-),男,硕士研究生,研究方向为嵌入式开发和编码算法;导师:秦霆镐,教授,研究方向为嵌入式系统开发、图像处理。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140525.1242.014.html>

后一级的缩减信源开始依照原先的编码路径向后返回就可以得到 Huffman 编码。Huffman 编码使得拥有越高概率出现的信源符号拥有越短的编码长度,这使得整个信源拥有较短的平均码长^[10]。但是 Huffman 编码存在工作量大、编解码时间较长的缺陷,给实际应用带来困难。一些常用的亚优编码方法,如 B2 码以及移位码等,虽然编解码效率得到提高,但编码形式固定、单一,对信源的统计有特殊要求,压缩效率较低^[11-13]。

针对上述问题,文中提出一种基于多参数的数据无损压缩方法。从一种具有特殊条件的数据统计集出发,将具有任意统计特性的原始数据进行有限次合并,使合并后新的数据集的统计特性满足上述特殊条件的数据统计集,按数据个数由大到小给各数据分配 1 元即时码作为编码前缀,对合并数据用自然位区分码加以后缀,从而实现数据的无损压缩。数据存储时仅用数据合并码表替代规模较大的数据编码表或个数统计表,解码时无需生成编码表,仅由编码前缀与后缀(需要时)即可复原原始数据,解码效率优于 Huffman 等传统方法^[14]。

1 多参数编码方法

对文中提出的基于多参数的无损压缩编码方法分为三步。

(1) 统计原始元素集个数,并按个数从大到小重新排列元素集;

(2) 合并经步骤一处理后的元素集,使其满足元素集中每一元素个数大于等于后续全体元素的个数之和;

(3) 对合并后元素集中各元素从小到大分配 1 元即时码作为编码前缀,被合并数据分配对应的区分码作为后缀。

1.1 编码基本原理

记原始数据块中独立元素的元素数组为 $D = \{a_0, a_1, \dots, a_m\}$,统计原始数据块中各独立元素的个数,并按个数从大到小重新排列数组 D ,数值依次赋予 $0, 1, \dots, m$,相应满足单调减的元素个数数组记 $N = \{n_0, n_1, \dots, n_m\}$ 。

假定数组 N 满足

$$n_i \geq \sum_{j=i+1}^m n_j \quad (i = 0, 1, \dots, m-1) \quad (1)$$

依据 Huffman 二叉树构成方法,将出现概率最小的两个信号源合并,其中两个信号源中概率最小的分配一位 1,次小的分配一位 0。即 m 分配一位 1, $m-1$ 分配一位 0,合并后的新元素记为 $m-1$,对应 $m-1$ 的元素个数为 $n_m + n_{m-1}$,由于元素个数集合 N 满足式

(1),所以 $n_m + n_{m-1} \leq n_{m-2}$,则元素 $m-1$ 继续分配一位 1, $m-2$ 分配一位 0,以此类推。最终元素 0 仅分配一位 0。由此得到下述各元素的 1 元即时码 0, 10, 110, 1110, ……

1 元即时码每一结束位均为 0,解码时只需计算到 0 结束位的编码单元个数,就可恢复对应的十进制数,无需传统编码方法解码时逐一匹配过程,极大地提高了解码效率。

1.2 编码生成步骤

一般而言,任一原始数据块中独立元素的个数统计数组经从大到小排序后,不能严格满足条件(1)。为此,可按照排序后的独立元素个数统计表对独立元素数组进行元素合并,由此使新的独立元素个数统计数组满足条件(1),具体做法如下:

(1) 独立元素个数统计及排序。

设独立元素数组为

$$D^* = \{a_0, a_1, \dots, a_m\} \quad (2)$$

对应的个数统计数组为

$$N^* = \{n_0, n_1, \dots, n_m\} \quad (3)$$

从大到小重新排列统计数组 N^* ,得下述个数统计数组

$$N = \{p_0, p_1, \dots, p_m\} \quad (4)$$

其中, $p_i \in \{n_0, n_1, \dots, n_m\}$ 。

数组 D^* 按相应顺序重新排列,分别赋予数值 $0, 1, \dots, m$,即得

$$D = \{0, 1, \dots, m\} \quad (5)$$

(2) 元素的合并。

如果数组(4)满足条件(1)转步骤(3),如果不满足条件(1),则进行合并,设 $p_q < p_{q+1} + \dots + p_m$,将数组 D^* 中元素 $q, q+1, \dots, m$ 按式(6)计算。

$$s = \text{Round}\left(\frac{t+q}{2}\right) \quad (6)$$

其中, $t \in \{q, q+1, \dots, m\}$; Round 为取整函数。

于是元素 q 与元素 $q+1$ 合并为新元素 q ,用 0 和 1 作为元素 q 和 $q+1$ 的区分码;元素 $q+2$ 与元素 $q+3$ 合并为新元素 $q+1$,同样用 0 和 1 作为这两个元素的区分码,后续元素按同样方式合并。经上述合并处理后,原始个数统计数组合并为

$$N^{(1)} = \{p_0, \dots, p_q + p_{q+1}, \dots\} \quad (7)$$

记 $N^{(1)} = \{p_0^{(1)}, \dots, p_q^{(1)}, p_{q+1}^{(1)}, \dots\}$ 。其中, $p_i^{(1)} = p_i$ ($i = 0, 1, \dots, q-1$); $p_q^{(1)} = p_q + p_{q+1}$, $p_{q+1}^{(1)} = p_{q+2} + p_{q+3}$ ……

对应的元素数组记为 $D^{(1)}$ 。如果个数统计数组 $N^{(1)}$ 满足条件(1),则转步骤(3),若不满足,则从不满足条件(1)元素开始,进行元素的第二次合并,由此得到元素数组 $D^{(2)}$ 和对应的个数统计数组 $N^{(2)}$,若数组 $N^{(2)}$ 依然不满足条件(1),则可继续合并,经过有限次

(设为 v 次)合并后所得个数统计数组 $N^{(v)}$ 必定满足条件(1),转步骤(3)进行元素编码。其中,参与 v 次合并的元素有 v 位的区分码。

文献[1]介绍了一种基于三参数的编码压缩方法,其主要思想是依据元素个数统计数组的整体合并后,获得一种满足条件(1)的新元素个数统计数组,每次合并时,无论元素个数统计数组中开头部分是否满足条件(1),均从第一个元素开始合并,做法较为简单,但势必引起概率较大元素编码位的增加。

(3)编码前缀与编码后缀。

设经过步骤(2)合并处理后的元素数组为 $D^{(v)} = \{0,1,\cdots,L\}$,按1元即时码依次给元素数组中各元素分配的编码前缀为0,10,110,1110, \cdots ,11 \cdots 10,11 \cdots 1。

其中,元素 $i(0 \leq i \leq L)$ 由 i 个1和一个0组成,编码长度为 $i+1$ 。最后一个元素 L 的编码前缀为 L 个1组成。如果元素 i 不是经合并后生成的新元素,则该元素的编码仅为1元即时码组成的前缀编码,后缀编码为空。如果元素 i 是经 v 次元素合并后生成的新元素,则该元素的后缀码为 v 位二进制自然码组成。

如元素10是经2次合并后生成的新元素,则其对应的元素分别是10,11,12,13。这四个元素的区分码分别为00,01,10和11。解码时通过元素10所对应的1元即时码,结合元素合并参数表即可分别复原数据10,11,12和13,再对照原始数据的排序码表即可复原原始数据。

2 元素合并参数表与数据编码

元素合并参数表由多个合并起始元素组成,如果合并次数为 v ,则合并参数表有 v 个起始元素。比如合并次数为 v ,则元素合并参数表如表1所示。

表1 元素合并参数表(1)

合并次数	起始值	起始值	起始值	起始值
v	e_1	e_2	\cdots	e_v

表1中 e_1 表示第一次合并元素起始值, e_2 表示第二次元素合并起始值, e_v 表示第 v 次元素合并起始值。

元素编码时,按元素合并后该元素有无经过合并确定其编码前缀与后缀。比如,某经过二次元素合并后的元素数组为 $D^{(2)} = \{0,1,2,3,4,5\}$,对应的合并参数表如表2所示。

表2 元素合并参数表(2)

合并次数	起始值	起始值
2	2	2

由合并参数表可知,元素合并前的元素数组为 $D = \{0,1,\cdots,14(15)\}$ 。

第一次合并后,2与3合并为新元素2,区分码为0和1;4与5合并为新元素3,区分码为0和1; \cdots ;14与15合并为新元素8,区分码为0和1。由此得到一次合并元素数组为 $D^{(1)} = \{0,1,2,3,4,5,6,7,8\}$ 。

第二次合并后,数组 $D^{(1)}$ 中3与4合并为新元素3,区分码为0和1;5与6合并为新元素4,区分码为0和1;7与8合并为新元素5,区分码为0和1。由此得到二次合并数组 $D^{(2)}$ 。

数组 D 中各元素编码为

0→0,1→10
2→110(0),3→110(1)
4→1110(00),5→1110(01),6→1110(10),7→1110(11), \cdots ,
12→111110(00),13→111110(01),
14→111110(01),15→111110(01)
圆括号内的编码为区分码,即编码后缀。

3 数据编码压缩实验

文中以三幅256级灰度,大小为512×515的图像为例,见图1~图3,按文献[1]的三参数方法及文中所述的多参数方法进行了编码实验,结果见表3。

表3 实验结果比较

方法	bpp		
	图片1	图片2	图片3
信源熵	5.07	3.68	7.12
文献[1]	5.15	3.74	7.76
文中方法	5.11	3.71	7.35



图1 测试图片(1)

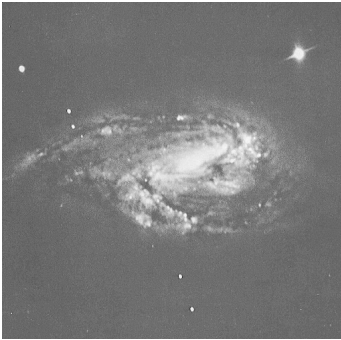


图2 测试图片(2)



图 3 测试图片(3)

4 结束语

文中利用了概率较大元素编码较短、概率较小元素编码较长的基本原理,利用 Huffman 的编码方法的一种特殊情形导出了一种多参数的数据编码方法,该方法对文献[1]所述三参数方法进行了改进,压缩效率进一步提高。文中用一个规模较小的元素参数合并表替代了传统方法中规模较大的数据编码表,并将数据编码分为十分简单的编码前缀和自然二进制数表示的编码后缀两部分,解码时通过编码前缀及元素合并参数表即可得到编码后缀的位数,并由此获得对应数据的区分码,根据区分码及其编码前缀得到对应的十进制数据,然后利用元素排序表即可复原原始数据。文中提供的多参数方法具有编解码过程简单和高效、压缩效率较高等特点。

参考文献:

- [1] 高 健,饶 珩,孙瑞鹏. 基于 3-参数变长编码的图像无损压缩算法[J]. 自动化学报,2013,39(8):1289-1294.
- [2] 陈 运. 信息论与编码[M]. 北京:电子工业出版社,2007.


(上接第 40 页)

- ing for space communications. Rhodes Island: IEEE, 2008: 1-6.
- [2] 雍明远,梁 俊,袁小刚. 宽带移动卫星通信信道模型研究[J]. 通信技术,2009,42(1):65-67.
- [3] 关庆阳. 低轨宽带卫星移动通信系统 OFDM 传输技术研究[D]. 哈尔滨:哈尔滨工业大学,2011.
- [4] 王文博,郑 侃. 宽带无线通信 OFDM 技术[M]. 第 2 版. 北京:人民邮电出版社,2007.
- [5] 杨明川. 卫星移动信道衰落特性模拟研究[D]. 哈尔滨:哈尔滨工业大学,2010.
- [6] 陈明举. OFDM 系统中 MMSE 与 LS 信道估计算法的比较研究[J]. 四川理工学院学报(自然科学版),2009,22(2): 91-93.
- [7] 季 伟. OFDM 系统的信道估计算法研究[D]. 成都:电子科技大学,2009.
- [8] Coleri S, Ergen M, Puri A, et al. Channel estimation tech-

- [3] 冯 希. 几种图像无损压缩与编码方法的比较研究[D]. 北京:中国科学院研究生院,2008.
- [4] 王学武,石跃祥. 对图像灰度级分段的压缩编码[J]. 计算机工程与设计,2006,27(2):222-223.
- [5] 籍俊伟. 无损图像压缩技术的研究与应用[D]. 北京:北京化工大学,2004.
- [6] 张晓咏,熊承义,胡开云,等. 基于灰度纹理信息的图像压缩感知编码与重构[J]. 计算机技术与发展,2013,23(1): 47-50.
- [7] 马媛媛,杨 峰,信 科,等. 基于 DCT 的 JPEG 图像压缩的研究[J]. 计算机技术与发展,2011,21(8):133-136.
- [8] 周晓燕,王继成. 静止图像压缩标准 JPEG 和 JPEG2000 的多尺度模式[J]. 计算机技术与发展,2007,17(1):12-14.
- [9] 吴凤辉,郑郁正. 小波变换在图像压缩中的应用[J]. 成都大学学报(自然科学版),2008,27(3):216-218.
- [10] 刘立波. 一种用于图像压缩的积分小波变换算法[J]. 宁夏大学学报(自然科学版),2003,24(4):353-355.
- [11] Stabno M, Wrembel R. RLH: bitmap compression technique based on run-length and Huffman encoding[J]. Information Systems, 2009, 34(4-5): 400-414.
- [12] Yang Enhui, Wang Longji. Joint optimization of run-length coding, Huffman coding, and quantization table with complete baseline JPEG decoder compatibility[J]. IEEE Transactions on Image Processing, 2009, 18(1): 63-74.
- [13] Papadonikolakis M E, Kakarountas A P, Goutis C E. Efficient high-performance implementation of JPEG-LS encoder[J]. Journal of Real-time Image Processing, 2008, 3(4): 303-310.
- [14] Kavousianos X, Kalligeros E, Nikolos D. Optimal selective Huffman coding for test-data compression[J]. IEEE Transactions on Computers, 2007, 56(8): 1146-1152.

- niques based on pilot arrangement in OFDM systems[J]. IEEE Transactions on Broadcasting, 2002, 48(3): 223-229.
- [9] Edfors O, Sandell M, van de Beek J J, et al. OFDM channel estimation by singular value decomposition[J]. IEEE Trans on Communications, 1998, 46(7): 931-939.
- [10] 樊同亮. OFDM 系统的信道估计和信号均衡技术的研究[D]. 重庆:重庆大学,2012.
- [11] 尹长川,罗 涛,乐光新. 多载波宽带无线通信技术[M]. 北京:北京邮电大学出版社,2004.
- [12] 仲伟志. 宽带卫星移动通信小波包分复用传输关键技术研究[D]. 哈尔滨:哈尔滨工业大学,2010.
- [13] Patzold M. 移动衰落信道[M]. 陈 伟,译. 北京:电子工业出版社,2009.
- [14] Ozdemir M K, Arslan H. Channel estimation for wireless OFDM systems[J]. IEEE Communications Surveys & Tutorials, 2007, 9(2): 18-48.

基于多参数的数据压缩算法

作者: [高怀远](#), [陈英豪](#), [GAO Huai-yuan](#), [CHEN Ying-hao](#)
作者单位: [上海大学 自动化系, 上海, 200072](#)
刊名: [计算机技术与发展](#) 
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2014(9)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjtz201409009.aspx