一种新的数据无损压缩编码方法

蔡 明 乔文孝* 鞠晓东 车小花 卢俊强 贾安学
(中国石油大学油气资源与探测国家重点实验室 北京 102249)
(北京市地球探测与信息技术重点实验室 北京 102249)

摘 要：为了降低数据存储和传输的成本，对数据进行压缩处理是一种有效的手段。该文针对具有较小均方值特征的整型数据序列提出了一种新的可用于数据无损压缩的位重组标记编码方法。该方法首先对整型数据序列进行位重组处理，以提高部分数据出现的概率；然后根据数据流中局部数据的概率分布特点自适应地选择合适的编码方式对数据流进行编码。运用实际具有较小均方值特征的整型数据序列对该文方法和其它几种无损压缩方法进行了压缩解压测试，并对比分析了各种压缩算法的压缩效果。测试结果表明，新方法可以实现数据的无损压缩与解压，且其压缩效果优于LZW编码，经典的算术编码，通用的WinRAR软件和专业音频数据压缩软件FLAC的压缩效果，具有良好的应用前景。

关键词：数据传输；编码；无损压缩；整型数据；位重组；标记
中图分类号：TN919.6+4 **文献标识码**：A **文章编号**：1009-5896(2014)04-1008-05
DOI: 10.3724/SP.J.1146.2013.00863

A New Coding Method for Lossless Data Compression

Cai Ming Qiao Wen-xiao Ju Xiao-dong Che Xiao-hua Lu Jun-qiang Jia An-xue
(*State Key Laboratory of Petroleum Resources and Prospecting*, *China University of Petroleum*, *Beijing* 102249, *China*)
(*Earth Explorer and Information Technology Laboratory*, *Beijing* 102249, *China*)

Abstract: Data compression is an effective measure to save the costs of data transmission and storage. A new and effective bit-recombination mark coding method that can be used to lossless data compression is proposed for the integer data sequence which has a small mean squared value. In the new method, the bit-recombination process is firstly applied to the integer data sequence to increase the occurrence probabilities of some data; then, the correct coding format is adaptively selected to encode the data stream according to the occurrence probability distribution characteristics of local data. Integer data sequences that have small mean squared values are applied to test the proposed method with several other lossless compression methods, and the compression effects are compared and analyzed. Test results show that, the integer data sequences can be compressed and decompressed losslessly by the proposed method. Moreover, the compression effect of the proposed method is superior to that of the classical arithmetic coding method, the LZW method, the universal WinRAR software, and the professional audio data compression software FLAC. The experimental results demonstrate the proposed method has a good application prospect.

Key words: Data transmission; Coding; Lossless compression; Integer; Bit-recombination; Mark

1 引言

数据压缩是一种消除原始数据之间的冗余性，并通过特殊的编码方式将原始数据文件转化为另一个占用存储空间更小的数据文件的技术[1−5]。数据压缩技术在过去20年里得到了快速的发展[5,6]。目前，它已广泛应用于数字通信、数字存储、计算机、数字出版及智能控制等众多领域[6−12]。编码是所有数据压缩方法的关键组成部分，且不同的编码方法对不同类型的数据序列有效[2]。如果采用一种专门为图像或音频数据设计的压缩程序(或编码方式)对文本文件进行压缩，则压缩后的文件大小可能大于甚至远大于原始数据文件。因此，针对不同类型的数据文件，选择或设计合适的编码方式是压缩成功的关键。

2013-06-19收到，2013-10-25改回
国家自然科学基金(11204380, 11374371, 11134011, 61102102)，国家油气重大科技专项(2011ZX05020-009)，中国石油天然气集团公司项目(2011A-3903, 2011B-4001)和中国石油科技创新基金(2013D-5006-0304)资助课题
*通信作者：乔文孝 qiaowx@cup.edu.cn

本文为均方值(定义均方值为各样值平方的算术平均值)较小的整型数据序列(或均值为 0，方差较小的整型数据序列)设计了一种新的编码方式——位重组标记编码，该编码方式是一种无损编码方式。对于满足均方值较小的整型数据序列，可直接用位重组标记编码进行压缩处理；对于不满足该条件的数据序列，可先利用相关的信号处理方法进行处理，使之满足条件后再利用位重组标记编码进行压缩处理；即位重组标记编码既可作为一种独立的数据压缩方法，也可与其它信号处理方法结合组成新的数据压缩方法，可应用于多种类型数据的压缩。利用具有较小均方值特征的整型数据序列对算术编码，LZW 编码，WinRAR，专业音频数据压缩软件 FLAC(Free Lossless Audio Codec)[13]和本文的方法进行了压缩解压测试，并对比分析了上述方法的压缩效果。

2　位重组标记编码

位重组标记编码主要包括位重组和标记编码两个组成部分，是为具有较小均方值特征的整型数据序列设计的一种有效的无损编码方式。图像数据、音频数据和其它波形数据经过预测处理得到的预测误差或经过小波变换处理得到的小波系数以及具有均值为零，方差较小的正态分布特征的实验数据等数据序列的均方值都比较小，可利用位重组标记编码进行压缩处理，且理论上可以取得较好的压缩效果。测试表明，位重组标记编码对具有较小均方值特征的整型数据序列的压缩效果优于算术编码，LZW 编码，WinRAR 软件和专业音频数据压缩软件 FLAC 的压缩效果。

2.1 标记编码

标记编码的基本思想源于文献[14]中的一种 Index-Data 记录格式的启发。Index-Data 记录格式是一种比较初级的标记记录格式，只有在数据 0 出现的频次占绝对优势时才会取得较好的压缩效果，否则可能导致压缩后数据文件大小大于甚至远大于原始数据文件大小。本文提出了一种更为有效的标记编码方式，可以根据数据的分布特点自适应地进行标记编码。

标记编码的核心思想是利用较短的码字(也称为替代码字，指需要用较少二进制位数表示的符号)代替数据流中出现频率最高的数据，其它数据正常编码(即不做任何改变，原样记录需要编码的数据，且每个数据认为是一个码字，称为非替代码字)，并在每个非替代码字前给出一个特殊的位标记。除了替代码字之外，其它码字均具有相同的长度。给出特殊的位标记的目的是为了在数据解压时能够正确地区分非替代码字和替代码字，以保证解码的正常进行。

本文的标记编码采用了双重标记，可根据局部数据的概率分布特点选择合理的编码方式，具有自适应性。标记编码的具体实现过程主要分 3 步完成。第 1 步，将输入数据流分隔成多个数据块，前面的数据块具有相同的大小，最后一个数据块可以稍小一些；第 2 步，统计数据块中各个数据出现的概率，并记录出现概率最大的数据及其对应的概率；第 3 步，根据第 2 步记录的最大概率值大小选择标记编码或正常编码(原样输出输入的各个数据)对该数据块进行编码；编码方式选择的准则是：当最大概率值大于设定的阈值时选择标记编码方式，否则选择正常编码方式。其中，第 3 步是本文标记编码方式的关键和难点。第 1 步中确定的数据块大小对编码效果也会有一定的影响，但影响不太大；一般而言，对于相同类型的数据序列，最佳数据块大小也是相近的，可以通过经验确定。

编码过程是对输入数据流中各数据块依次进行的，直到最后一个数据块被编码。首先，根据将被编码的数据块选择的编码方式，将其对应的编码方式标记符(此为一重标记)记录到输出文件中；然后，按选定的编码方式对该数据块进行编码；若选择标记编码方式进行编码，则先原样输出数据块中出现概率最大的数据，然后依次对数据块中的各数据进行编码，当编码过程中遇到出现概率最大的数据时，则用选定的较短的替代码字代替，其它数据原样输出并在前面给出一个特殊的位标记(此为二重标记)；若选择正常编码方式进行编码，则依次原样输出数据块中各数据即可。如此循环，即可完成对整个输入数据文件的编码，编码过程如图 1 所示。

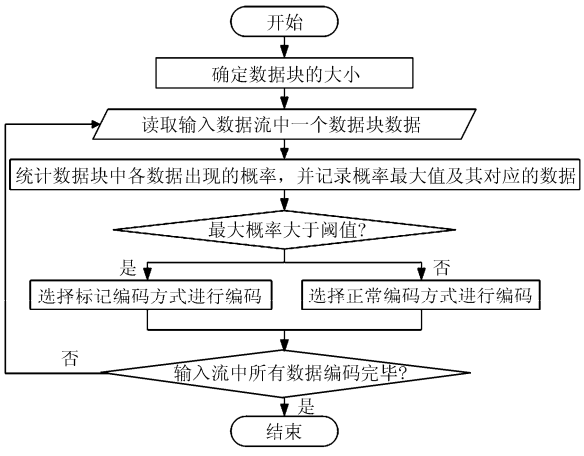


图 1 标记编码流程

将输入数据流分块进行编码的好处是可以自适应地根据数据流中局部数据的分布特点选择合适的编码方式进行编码。然而，普通的标记编码方式(不对数据流进行分割，对整个数据流采用统一的标记编码方式进行编码)则没有这种自适应性。利用普通标记编码方式进行编码时，对于非替代数据(出现频率非最高的数据)，除了原样输出该数据外，还需要给出特殊的标记符号，而过多的标记符号也将额外地占据大量的存储空间，不利于提高数据压缩的效果。因此，当数据块中没有数据出现的频率占有绝对优势时，若仍采用标记编码方式进行编码，则需要给出大量的标记符号，这可能导致压缩后数据文件比原始数据文件占据的存储空间还大，达不到数据压缩的目的；而此时若选择正常编码方式进行编码，则可以避免这种不利情况的发生。

本文的标记编码方式可以根据各数据块中数据的概率分布特点自适应地选择合适的编码方式进行编码。当数据块中有数据出现的概率高于设定的阈值时选择标记编码方式进行编码，使压缩后数据小于或远小于压缩前数据占据的存储空间；否则选择正常编码方式进行编码，使压缩前后数据文件大小相同；这样就保证了每个数据块压缩后的数据比压缩前的数据占据的存储空间小或相等，从而获得整体上的压缩效果。因此，本文的标记编码方式一般都会取得一定的压缩效果，绝对不会出现压缩后数据文件大小远大于压缩前数据文件大小的情况。

2.2 位重组

位重组的目的是为了提高某一数据(一般情况下主要是数据 0)出现的概率，是为标记编码服务的。位重组的基本思想来源于洗牌(Shuffle)原理[15,16]。本文使用的位重组处理方法可描述为：首先依次提取数据流中各数据的最高位，然后依次提取各数据的次高位，如此循环，直至最后依次提取各数据的最低位，之后将这些提取的数据位按顺序组合成新的数据；对于 16 位无符号整型数而言，首选按顺序依次提取数据流中各数据的第 16 位，第 15 位，…，第 1 位，然后将这些提取到的数据位按顺序组合成新的数据(每 16 位组合成 1 个数据)即完成了位重组处理。

具有较小均方值特征的整型数据序列中各数据的高位部分基本都是由 0 组成的，对这类数据进行位重组处理理论上可以大大提高某些数据(主要是数据 0)出现的概率。实际测试表明，对具有较小均方值的整型数据序列进行位重组处理确实可以明显提高某些数据出现的概率，有利于提高最终的压缩效果。

3 基于位重组标记编码的数据压缩与解压方法

3.1 数据压缩方法

由于具有较小均方值特征的整型数据序列一般既包括负整数，又包括非负整数，即为有符号整型数据序列。所以在采用位重组标记编码对这类数据序列进行压缩之前，有必要先对数据序列进行数据类型转换处理，即将有符号整型数据序列转换为无符号整型数据序列。做这一处理有两点好处：(1)可以避免编码和解码过程中对符号位进行特殊的处理；(2)可以提高位重组标记编码的处理效果，进而提高最终的压缩效果。由于负整数的符号位为 1，而非负整数的符号位为 0，所以有符号整型数据序列的符号位由 0 和 1 组成，且其出现的概率可能相近，这不利于提高位重组标记编码的效果。而转换为无符号整型数据序列之后再进行位重组处理则可以避免这一问题，因为序列中无符号整型数的最高位主要由 0 组成，且一般情况下全为 0。

数据类型转换可以通过一种可逆的一一映射来实现，即将第 n 个负整数(如 $-n$)映射到第 n 个奇数(如 $(2n-1)$)，将第 m 个正整数映射到第 m 个偶数($2m$)。具体转换式为[17,18]

$$y=\begin{cases}2x, & x\geq 0\\ 2|x|-1, & x<0\end{cases} \tag{1}$$

其中 x 表示有符号整型数，y 表示转换后的无符号整型数，$|x|$ 表示 x 的绝对值。

基于上述讨论容易得出本文数据压缩的实现方法。本文的数据压缩方法主要包括 3 个部分：(1)对具有较小均方值特征的整型数据序列进行数据类型转换处理，将有符号整型数转换为无符号整型数；(2)对无符号整型数据序列做位重组处理；(3)标记编码，实现压缩。数据压缩流程如图 2 下半部分所示。

3.2 数据解压方法

数据解压是数据压缩的逆过程。数据解压方法同压缩方法一样也包括 3 个部分：(1)对压缩数据进行标记解码，这是标记编码的逆过程且根据标记编码的原理和实现方法容易得到标记解码的实现方法；(2)对标记解码后的数据做位恢复处理，这是与位重组相对应的反处理，即将各数据位恢复到位重组前的原始位置上；(3)进行数据类型转换；解压过程中的数据类型转换是将无符号整型数转换为有符号整型数，这与压缩过程中的数据类型转换正好相反，由式(1)很容易得出无符号整型数到有符号整型数的转换公式。完成以上 3 步处理之后就重构出了原始数据序列，解压过程完毕。数据解压流程如图 2 上半部分所示。

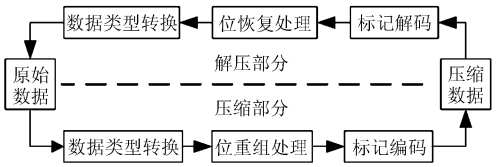


图 2 数据压缩/解压流程

根据数据压缩各处理部分的原理及实现方法容易得到对应的数据解压各处理部分的实现方法。由于压缩各部分处理都是完全可逆的，所以整个压缩处理方法也是完全可逆的，即通过压缩数据文件可以完全无失真地重构出原始数据序列。对具有较小均方值特征的整型数据序列进行的压缩解压测试也证明了这一点。

4　应用效果与分析

根据上文介绍的压缩解压方法，本节在 VC++平台上编写了相应的压缩和解压程序。运用编写的程序和其它几种无损压缩方法(或编码方法)对实际的具有相对较小均方值的整型数据序列进行了压缩解压测试，并对比分析了它们的压缩效果。

测试数据是由实际的声波测井波形数据经过去相关处理后得到的。我们从实际的声波测井资料中随机地抽取了 12 道波形，并利用信号处理方法对波形数据进行去相关处理得到了 12 道均方值相对较小的测试数据序列，每道测试数据文件大小均为 21968 Byte。测试数据序列的均方值明显小于原始声波测井数据序列的均方值，其波形如图 3 所示。利用这 12 道测试数据序列对算术编码，LZW 编码，WinRAR 软件，专业音频数据压缩软件 FLAC 和本文方法进行了压缩解压测试。测试结果表明，这几种压缩方法都是无损的，且各种压缩方法对不同道测试数据的压缩因子(原始数据文件大小与压缩数据文件大小之比)如表 1 所示。

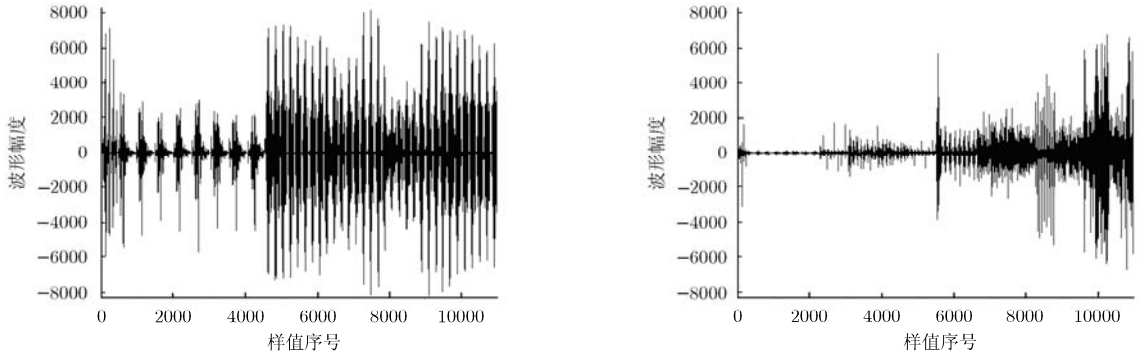
从表 1 可以看出，FLAC 软件基本没有压缩能力，且对于有些测试道数据，其压缩后数据文件大小反而超过了原始数据文件大小；LZW 编码方法有一定的压缩效果，但效果不佳；相比而言，算术编码，WinRAR 软件和本文方法均取得了相对较好的压缩效果。另外，无论从单个测试道数据的压缩效果，还是最后各种压缩方法的平均压缩因子的对比情况来看，本文方法的压缩效果均明显优于其它压缩方法的压缩效果。

5　结论

针对具有较小均方值特征的整型数据序列，本文提出了一种新的有效的可用于无损数据压缩的位重组标记编码方法，并在 VC++平台上实现了相应的压缩解压程序。利用实际具有较小均方值特征的整型数据序列对本文方法和其它几种无损压缩方法进行了压缩解压测试，并对比分析了各种压缩算法

表 1 各种压缩方法对不同道测试数据压缩因子统计表

道号	FLAC 软件	LZW 编码	算术编码	WinRAR 软件	本文方法
1	1.00	1.12	1.32	1.41	1.56
2	1.02	1.14	1.34	1.44	1.60
3	0.99	1.09	1.29	1.28	1.54
4	1.00	1.11	1.30	1.28	1.56
5	1.01	1.12	1.31	1.41	1.57
6	1.01	1.12	1.32	1.41	1.58
7	1.03	1.13	1.33	1.43	1.61
8	1.03	1.15	1.34	1.43	1.62
9	1.02	1.15	1.34	1.44	1.62
10	0.99	1.11	1.30	1.25	1.54
11	1.00	1.10	1.29	1.38	1.56
12	1.01	1.12	1.32	1.42	1.58
平均压缩因子	1.01	1.12	1.32	1.38	1.58



(a)原始声波测井波形　　(b)去相关后的数据波形

图 3 某道测试数据波形及其对应的原始声波测井波形图

的压缩效果。本研究主要得出如下结论：(1)本文数据压缩方法可以实现整型数据序列的无损压缩与解压，压缩解压速度快，效果稳定； (2)对具有较小均方值特征的整型数据序列，位重组标记编码是一种较好的无损压缩方法，且一般情况下均方值越小，压缩效果越好；(3)本文方法对满足应用条件的数据序列具有较高的压缩因子，压缩效果优于经典的算术编码，LZW 编码，通用的 WinRAR 软件和专业音频数据压缩软件 FLAC 的压缩效果；(4)对于具有较小均方值特征的整型数据序列，可直接利用位重组标记编码进行压缩处理；对于不满足该条件的数据序列，可先利用相关的信号处理方法进行处理，使之满足条件后再利用位重组标记编码进行压缩处理；因此，位重组标记编码可应用于多种类型数据的压缩，应用范围广泛。

本文方法是针对具有较小均方值特征的整型数据序列设计的，只有对满足此条件的数据序列才会取得良好的压缩效果，且一般情况下数据序列的均方值越小，压缩效果越好。另外，在必要的情况下，可进行进一步的研究，将该方法发展为适用于平均有效位数较少的浮点型数据序列的位重组标记编码方法。

参 考 文 献

[1] Salomon D. Data Compression: The Complete Reference[M]. London: Springer-Verlag, 2007: 1–952.

[2] Salomon D and Motta G. Handbook of Data Compression[M]. London: Springer-Verlag, 2009: 1–1198.

[3] Li X. Compression of well-logging data in wavelet space[C]. 1996 SEG Annual Meeting, Denver, Colorado, 1996: 1615–1618.

[4] Schendel E R, Ye J, Shah N, *et al*.. ISOBAR preconditioner for effective and high-throughput lossless data compression [C]. 2012 IEEE 28th International Conference on Data Engineering, Washington, 2012: 138–149.

[5] Sayood K. Introduction to Data Compression[M]. Massachusetts: Morgan Kaufmann, 2006: 1–39.

[6] 吴家安. 数据压缩技术及应用[M]. 北京: 科学出版社, 2008: 1–282.

[7] Keymeulen D, Aranki N, Hopson B, *et al*.. GPU lossless hyperspectral data compression system for space applications [C]. 2012 Institute of Electrical and Electronics Engineers Aerospace Conference, Big Sky, MT, 2012: 1–9.

[8] Srisooksai T, Keamarungsi K, Lamsrichan P, *et al*.. Practical data compression in wireless sensor networks: a survey[J]. *Journal of Network and Computer Applications*, 2012, 35(1): 37–59.

[9] Hou Zhi-ling, Su Xian-yu, and Zhang Qi-can. Virtual structured-light coding for three-dimensional shape data compression[J]. *Optics and Lasers in Engineering*, 2012, 50(6): 844–849.

[10] 刘杰, 易茂祥, 朱勇. 采用字典词条衍生模式的测试数据压缩[J]. 电子与信息学报, 2012, 34(1): 232–235.
Liu Jie, Yi Mao-xiang, and Zhu Yong. Test data compression using entry derivative mode of dictionary[J]. *Journal of Electronics & Information Technology*, 2012, 34(1): 232–235.

[11] Chen F, Chandrakasan A P, and Stojanovic V M. Design and analysis of a hardware-efficient compressed sensing architecture for data compression in wireless sensors[J]. *IEEE Journal of Solid-State Circuits*, 2012, 47(3): 744–756.

[12] Salomon D. A Concise Introduction to Data Compression[M]. London: Springer-Verlag, 2008: 1:20.

[13] Coalson J. FLAC[OL]. http://flac.sourceforge.net/. 2011.

[14] 李传伟, 慕德俊, 李安宗, 等. 随钻声波测井数据实时压缩算法[J]. 西南石油大学学报(自然科学版), 2008, 30(5): 81–84.
Li Chuan-wei, Mu De-jun, Li An-zong, *et al*.. A real-time data compression algorithm for acoustic wave logging while drilling[J]. *Journal of Southwest Petroleum University* (*Science & Technology Edition*), 2008, 30(5): 81–84.

[15] 吴国清, 陈虹. 一种科学数据无损压缩方法[J]. 计算机工程与应用, 2006, (5): 172–175.
Wu Guo-qing and Chen Hong. A lossless compression scheme for scientific data from simulation[J]. *Computer Engineering and Application*, 2006, (5): 172–175.

[16] 刘杰, 徐三子. 用于编码压缩的测试位重组算法[J]. 计算机工程, 2010, 36(21): 19–21.
Liu Jie and Xu San-zi. Test-bit-rearrangement algorithm applied to code compression[J]. *Computer Engineering*, 2010, 36(21): 19–21.

[17] 胡学龙, 江新炼, 周琳, 等. 一种改进的无损压缩数字音频编码器[J]. 微电子学与计算机, 2003, (7): 23–25.
Hu Xue-long, Jiang Xin-lian, Zhou Lin, *et al*.. An improved lossless compression in digital audio coder[J]. *Microelectronics and Computer*, 2003, (7): 23–25.

[18] Robert F R. Some practical universal noiseless coding techniques, Part III[R]. Jet Propulsion Laboratory, California Institute of Technology, 1991.

蔡 明： 男，1986 年生，博士生，研究方向为声波测井及信号处理.

乔文孝： 男，1956 年生，教授，博士生导师，研究方向为声波测井及应用声学.

鞠晓东： 男，1953 年生，教授，博士生导师，研究方向为井孔地球物理及井下仪器.

车小花： 女，1976 年生，副研究员，硕士生导师，研究方向为声波测井及应用声学.

卢俊强： 男，1978 年生，博士(后)，讲师，研究方向为声波测井方法及仪器.

贾安学： 男，1985 年生，硕士，研究方向为声波测井.