地震数据的无损压缩存储

孙 毅,韩坤亮
(上海大学 计算机工程与科学学院,上海 200072)

摘 要:为解决地震数据的无损压缩存储问题,研究了多种地震数据标准以及整型和浮点型数据的无损压缩算法。在此基础上,提出了地震数据的无损压缩存储方法。该方法通过分析多种地震标准的内部结构特点,有针对性地对整型和浮点型数据分别进行无损压缩。在保证了地震数据文件中各参数、标识等信息不被破坏的条件下,最大限度地减小了地震数据文件的大小。试验证明,本方法不仅对整型数据而且对于浮点型数据都可以做到无损的数据压缩与还原。

关键词:地震数据;无损压缩;整型压缩;浮点型压缩;提升小波变换
中图分类号:TP311.1 文献标识码:A 文章编号:1673-629X(2011)08-0177-04

Lossless Compression of Seismic Data

SUN Yi,HAN Kun-liang
(School of Computer Engineering and Science, Shanghai University,Shanghai 200072,China)

Abstract: To solve the problem of seismic data lossless compression effectively,many seismic data standards, lossless integer and float-point encoding are studied. Based on these, a design method of lossless compression of seismic data is presented. By analyzing the structure of the seismic standards, the integer data or float-point data is compressed separately in the method. The messages of seismic files are saved and the volume of the files is decreased as possible in the method. Finally, the feasibility of the method is validated by practical application.

Key words: seismic data; lossless compression; integer compression; float-point compression; lifted wavelet

0 引 言

随着传感器等硬件设备的不断发展以及地理信息系统(Geographic Information System,GIS)的发展,为了分析得到更精确的地质信息,人们所需要的地震数据的信息量越来越大。同时,由于存储设备以及传输速度的发展较处理器的发展相对滞后,对海量的数据进行压缩存储就很有必要。于是地震数据的压缩存储较多地被关注起来。

针对地震数据的压缩存储研究层出不穷。从起初的分形方法到子带编码,再到当前主流的小波变换等变换方法结合量化、编码等算法进行压缩。传统的地震数据压缩存在两个弊端:为片面提高压缩比,大多研究采用有损压缩[1,2]。对于不断提高的地震数据分析,这就造成无谓的信息的丢失,降低了地震数据分析的准确性;无视地震道数据的类型,正如文献[1]中所述的地震数据的压缩,是将地震数据认为是浮点型数据进行压缩存储,通过损失一些浮点型数据的精度,达到较高的压缩比,而文献[3]中提升小波是对整型数据进行处理。所以这样的地震数据压缩即使是无损压缩,也只针对整型数据,并不涉及浮点型数据。但地震数据的存储标准中通常既支持浮点型同样也支持整型数据存储。

1 地震数据标准

对地震数据进行压缩,首先,需要了解地震数据格式。文中通过对几种常见的地震数据标准的分析来总结其结构特点。

1.1 SEG-2 标准

勘探地球物理家学会(The Society of Exploration Geophysicists,SEG)在1990年发表地震数据文件格式标准。SEG-2文件格式共由三部分组成:文件描述块(File Descriptor Block,FDB)、多个地震道数据描述块(Track Descriptor Block,TDB)和多个数据块(Data)。其中,FDB位于整个文件的最开头,保存了所有记录道的公共信息;FDB的后面交替的跟着多个TDB和数据块,每一个TDB中保存一道地震记录的有关信息;每一个TDB后面都跟着一个数据块,保存该道地震的波

收稿日期:2011-01-04;修回日期:2011-04-25
基金项目:上海市重点学科建设项目(J50103);上海市科委重点实验室基金(09DZ2272600)
作者简介:孙 毅(1984-),男,硕士研究生,研究方向为数据可视化。

形数据,道数据块中存放的是一连串相同类型的地震数据,数据格式可以是 IBM 浮点型、IEEE 浮点型、整型、长整型等。

1.2 SEG-Y、SEG-D、SEED 标准

其他许多地震数据标准例如 SEG-Y、SEG-D 以及 SEED 标准等,在地震数据的存储中都发挥着一定作用。其具体格式详见参考文献[4~6]。

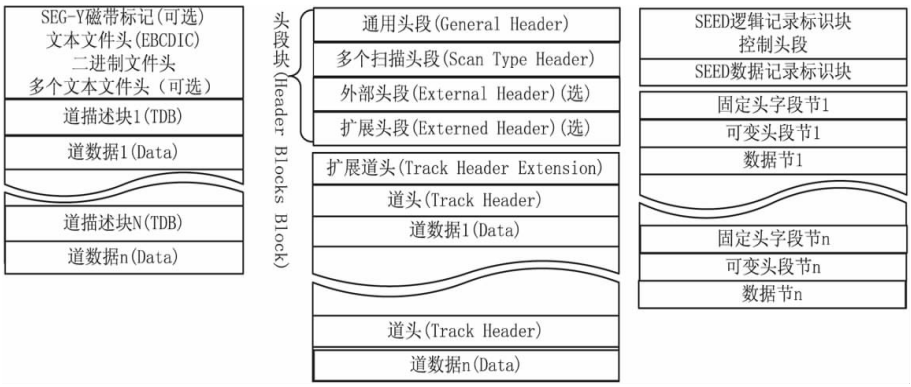

图 1 SEG-Y 文件结构(左) SEG-D 文件结构(中) SEED 文件结构(右)

SEG-D 与 SEG-Y 格式同样是 SEG 委员会指定的标准磁带数据格式;SEED(Standard for the Exchange of Earthquake Data,地震数据交换标准格式)是美国地质调查局(U. S. Geological Survey,USGS)提出的,并被国际数字地震台网联合会(International Federation of Digital Seismograph Networks,FDSN)定为国际数字地震数据交换的标准格式。

通过对几种数据标准的分析可以得出如下结论:

1)以往的地震数据存储标准少有对数据进行压缩存储(例如 SEG-2,SEG-Y,SEG-D)。这无疑造成了存储介质的不必要浪费以及传输过程的不必要开销。只是 SEED 标准中,采用了 STEIM1 与 STEIM2 算法,对数据进行了一定的压缩。但是 SEED 标准依然存在一定的缺陷。首先,SEEIM1 与 STEIM2 均是采用差分压缩的算法,两种算法的差分值只是简单的相邻点数值的差,当存储剧烈的地震数据时,相邻数据依然存在较大的差值;其次,当采用较小的单位存放数据的时候,相邻的数据的差值也会很大,压缩效果不理想。另外,STEIM2 算法中,尽管效率上比 STEIM1 要高一些,但是,对于大于 24 位的数据压缩后,无法实现无损还原[7]。最后,两种压缩算法,都只针对整型,对于浮点型的数据不能采用该算法。

2)地震数据存储都遵循“文件头+道头+道数据+……+第 n 个道头+第 n 个道数据”的模式。由于文件头部分以及每一道中的道头相对都很小,这样可以有针对性的只对道数据进行压缩,压缩后的数据存放地方与地震标准中存放地方相一致,对文件中参数不进行压缩,这样就可以保留文件以及地震道的原有信息,同时可以存储一些关于道数据压缩的信息(例如压缩算法标识等),便于数据解压缩时读取。

3)各种标准中的地震数据,既可以是整型也可以是浮点型数据。对于此,合理的做法应该是,针对不同的数据类型分别进行有针对性的压缩编码。压缩前的原数据类型以及压缩编码的选择等信息可以储存到文件头参数中,以方便解压缩时读取使用。

2 道数据的压缩

文中将地震数据按照整型和浮点型分别进行不同的压缩存储。对于整型数据采用提升小波变换与 ECMA-335 整数编码的方式进行压缩;对于浮点数,为降低工程复杂度,创造性地将浮点数无损转化为整数,然后与整型数据进行同样的压缩编码。

2.1 整型数据的提升小波变换

信号有局部相关性,某一点的信号值可以通过其相邻的信号值经过适当的预测算子预测出来,尤其是地震的波形数据,前后数据间存在着极大的相关性,又因为提升小波变换具有计算简单、无损还原等特点,文中将采用提升小波变换去掉数据间的相关性。

提升小波理论分为三个环节[8,9]:

1)分割。把原始数据 $X(n)$ 分解成两部分,奇数列和偶数列,即 $xe(n)=x(2n)$,$xo(n)=x(2n+1)$。

2)预测。又称为对偶提升,利用 $xe(n)$ 预测 $xo(n)$,即 $d(n)=xo(n)-p[xe(n)]$。式中,p 为预测算子,$d(n)$ 表示预测误差。在提升的理论中,预测误差 $d(n)$ 也称为小波系数(对应高频分量),当预测值 $p[xe(n)]$ 越接近 $xo(n)$,预测误差 $d(n)$ 就越小[10]。

3)更新。又称为原始提升,利用 $d(n)$ 更新 $xe(n)$,即 $c(n)=xe(n)+u[d(n)]$,式中,u 为更新算子。在提升理论中,$c(n)$ 也称尺度系数(对应低频分量),是原始数据的一个粗糙近似。

小波提升是一个完全可逆的过程,其正反变换结构对称,算子符号相反,由此可以实现精确重构。提升方案可以实现原位计算和整数提升,并且中间结果是交织排列的。选用 w5/3 小波,则变换公式以及原理如公式(1)、(2)和图 2 所示[11]:

$$c(2n+1)=x(2n+1)-\left[\frac{x(2n)+x(2n+2)}{2}\right] \quad (1)$$

$$d(2n)=x(2n)+\left[\frac{c(2n-1)+c(2n+1)+2}{4}\right] \quad (2)$$

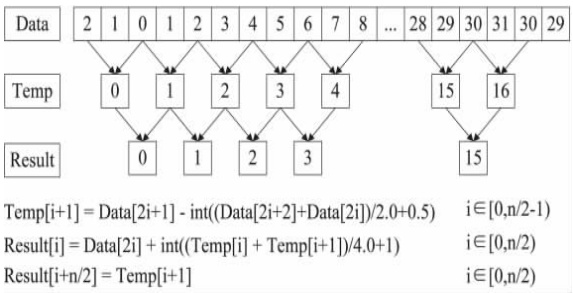

图 2　提升小波变换原理图

2.2　整数压缩编码

根据提升小波变换,可以看出,处理后的每一道整型数据,除了第一个数字之外,其他的都是细节系数,而这些细节系数大多是都很小且重复性很大。基于此,采用改进的行程编码与 ECMA-335 中采用的整数压缩算法混合编码方法。

2.2.1　改进的行程编码

行程编码是将连续出现的字符采用"字符+记号(S)+出现次数"的形式来代替并实现压缩的。对于重复出现 3 次或 3 次以下的字符不采用该编码进行压缩,否则采用该编码。

这里需要说明的是,"记号"用 63(二进制 111111)来表示。同时需要注意几个规定: 连续的字符数不能超过 62,否则应当断开,后面的再使用 3 个字节存储; 文件中 ASCII 码值为 63 的数值,采用两个 63 来表示。

2.2.2　ECMA-335 整数编码

ECMA-335 标准即公共语言架构(Common Language Infrastructure,CLI)。该标准用于将不同的高级语言编写的程序可以在不同的系统环境下能够正确运行。其中描述了一个整数压缩算法,该算法概括的说是将整数的取值范围划分为几个区段,而整数数值根据其所在的区段不同,放置在 1、2 或 4 个字节中。对于带符号整数,将符号位循环左移到最后一位,再按照表 1 列出的区段取得最终占用的字节数。

表 1　ECMA-335 标准中整型压缩区段

区段	字节数	掩码	二进制形式
[00000000h, 00000 07Fh]	1	80h	0NNNNNNN
[00000080h, 00003F FFh]	2	C0h	10NNNNNN NNNNNNNN
[00004000h, 1FFFF FFFh]	4	E0h	110NNNNN NNNNNNNN NNNNNNNN NNNNNNNN

"区段"列出了每个区段的最小值(含)和最大值(含); "字节数"列出了压缩后数值所占用的字节数; "掩码"列出用于判断数值区段的值: 如果压缩后的整数值占用 1 字节,则与掩码 80h 进行 &(按位与)操作后的结果为 0h; 如果压缩后的整数值占用 2 字节,则其首字节与掩码 C0h 进行 & 操作后的结果是 80h; 如果压缩后的整数值占用 4 字节,则其首字节与掩码 E0h 进行 & 操作后的结果是 C0h; "二进制形式"列出了压缩结果的二进制形式,其中的"1"和"0"都是固定值,而"N"则表示实际整数值的有效位。图 3 描述了该整型压缩算法压缩过程,解压缩算法是其逆过程,这里不再赘述。

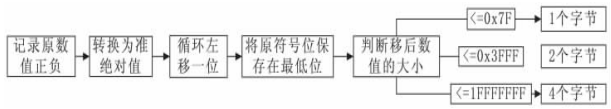

图 3　ECMA-335 标准压缩与解压缩流程图

2.3　浮点型数据的压缩编码

对于浮点型数据先分析其内部结构特点。目前通用计算机中使用的浮点数表示方法是 IEEE(Institute of Electrical and Electronics Engineers,电子电气工程师协会)在 1985 年制定的 IEEE754(IEEE Standard for Binary Floating-Point Arithmetic,ANSI/IEEE Std 754-1985)二进制浮点运算规范(下称"标准")。标准中规定了单精度、双精度以及双精度扩展等三种格式,这里以 32 位单精度浮点型为例详述。

标准中规定了单精度格式由三个字段组成: 23 位小数部分 M; 8 位移码指数 e; 以及一位符号位 Ms。

根据标准,一个单精度数值的大小为: $f = (-1)^{Ms} \bullet M \bullet B^{e}$。通过文献[12]中所做的工作,可以大致得到浮点型数据与整型数据一一映射的一个函数示意图,如图 4 所示。

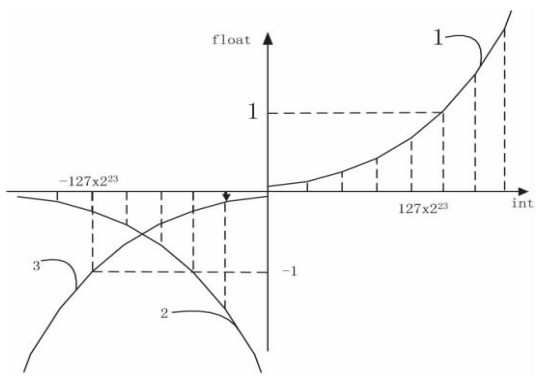

图 4　整型-浮点型映射示意图

图 4 中,线段 1 和线段 2 是不考虑整型补码表示的整型—浮点型映射函数示意图,由于标准中规定计算机中整型采用补码表示,所以在数值为负数的情况下,将对应的整数求补,然后容易得到线段 1 与线段 3 即是浮点型—整型映射的函数示意图。从图中得出,整数与对应二进制相同的浮点数的一一映射函数是在区间 [-max, +max] 的单向递增函数。因此可以断言,所谓浮点型数据就是改变了原来整型数据的间隔,在不同的区间乘以不同的系数 k。随着所表示的数字的增大,系数 k 增大。

综上,对浮点数地震数据的压缩,先将其映射成对

应的整型数值,然后采用前文所述的整型压缩算法进行压缩存储。然后在地震数据的文件头参中做相应的标记,在解码时按照所述算法,将整型数据转为相应的浮点数即可。

3 试验结果

试验分别采用多个浮点型以及整型数据存储的 SEG-2 标准地震格式数据。试验证明,该方法无论对于整型道数据还是浮点型道数据,都可以对压缩后的数据实现无损还原。通过对压缩前后的波形图进行比较,可以看出无论是整型数据还是浮点型数据在压缩前后波形都无明显变化。另外,通过对大量文件压缩比的比较可发现:无论整型还是浮点型数据,其压缩比随着数据波动范围的增大而增大,同时,随着波动强度的增加而适当地减小。

据试验所得数据中,最高压缩比可达 3.6 ∶ 1,最小压缩比可达 1.3 ∶ 1。这说明,该方法无论是对浮点型还是整型数据进行压缩时,对于波动整体数值较大的波形,或者,相邻点变化较慢的波形压缩明显;对于整体数据很小,而且相邻点变化较大的波形压缩效果较弱。

4 结束语

采用文中描述方法对地震数据进行压缩存储,将浮点型数据与整型数据区分对待,最大限度地提高了数据无损压缩的比例。无论整型数据还是浮点型数据,其解压缩的结果都是完全无损的,保证了数据的精确性。通过试验结果分析,可以得出压缩算法理想的适用范围。

参考文献:

[1] 冯占林, 张学工, 李衍达. 基于小波变换的地震勘探数据压缩的工程分析 [J]. 清华大学学报(自然科学版),2001,41(4):170-173.

[2] 王国清. 小波变换理论及其在地震勘探数据压缩中的应用 [D]. 南京:南京理工大学,2005.

[3] 孙寰宇, 武文波, 杨志高,等. 基于 JPEG2000 的地震数据准无损压缩 [J]. 计算机工程与应用,2005,41(16):185-188.

[4] Society of Exploration Geophysicist SEG Y rev 1 Data Exchange Format [S]. 2003.

[5] SEG 委员会. SEGD Rev2, SEG Field Tape Standards [S]. 1996.

[6] 王秀文,姚立平,赖德伦,等. 地震数据交换标准 [J]. 地震地磁观测与研究,1994,2(15):1-42.

[7] 罗新恒,正 哲,王春明,等. SEED 数据压缩率的比较 [J]. 地震地磁观测与研究,2004, 24(4): 41-47,14-19.

[8] Swedens W. The lifting scheme: a new philosophy in biorthogonal wavelet constructions [J]. Proceedings of SPIE, 1995, 2569:68-79.

[9] 王文涛. 基于小波的图像压缩编码算法研究 [D]. 重庆:重庆大学,2005:22-24.

[10] Mandyam G, Magotra N, McCoy W. Lossless Seismic Data Compreesion using [M]. [s. l.]: Adaptive Linear Prediction, 1996.

[11] ISO/IEC JTC1/SC29/WG1. Coding of still pictures [S]. 2000-03-16/2005-04-11.

[12] 孙 毅. 浮点型数据的无损压缩 [C]//ICRCCS. 常州 [出版者不详],2010:24-27.

(上接第 176 页)

Architecture Straight from the Masters [M]. New York: Meghan Kiffer Press, 2004: 135-146.

[2] 谢正良, 赵建华, 李宣东,等. 一种基于 J2EE 平台的 MDA 模型转换技术 [J]. 计算机应用研究, 2005(3): 51-54.

[3] Martin F. Domain Specific Languages [M]. Toronto: Addison-Wesley Professional, 2010.

[4] 刘 辉,麻志毅,邶维忠. 元建模技术研究进展 [J]. 软件学报, 2008, 19(6) : 1317-1327.

[5] 周金根. MetaModelEngine:元模型引擎开发思路 [EB/OL]. [2010-10-12]. http://www.cnblogs.com/zhoujg/archive/2010/07/28/1786155.html.

[6] MetaCase. MetaEdit+Domain-Specific Modeling Environment [EB/OL]. [2010-09-10]. http://www.metacase.com/MetaEdit.html.

[7] Juha-Pekka T. MetaEdit+: integrated modeling and meta-modeling environment for domain-specific languages [C]// Companion to the 21st ACM SIGPLAN symposium on object-oriented programming systems, languages, and applications. New York: ACM, 2003: 690-691.

[8] 李思广, 林子禹, 胡 峰, 等. 基于 UML 的软件过程建模方法研究 [J]. 计算机工程与应用, 2003, 39(6): 76-78.

[9] 王 珊,萨师煊. 数据库系统概论 [M]. 第 4 版. 北京: 高等教育出版社, 2006.

[10] MetaCase. MetaEdit+4.5 Workbench User's Guide [EB/OL]. [2010-09-15]. http://www.metacase.com/support/45/manuals/mwb/Mw.html.

[11] 周彩兰, 孙 琳, 李素芬. 基于 JSP 的网络数据库连接技术 [J]. 计算机技术与发展, 2006, 16(4): 209-211.

[12] Sanna S. Domain-specific modeling language and code generator for developing repository-based Eclipse plug-ins [R]. Espoo: VTT Publications, 2008.