

基于自适应间隔游程编码的空间数据实时无损压缩方法

姚丹丹¹, 陈 军², 卢涵宇³, 苗 放^{1,4}

(1 成都理工大学 地球物理学院, 四川 成都 610059; 2 成都信息工程学院 资源环境学院, 四川 成都 610225;
3 贵州大学 电子信息学院, 贵州 贵阳 550025; 4 成都理工大学 空间信息技术研究所, 四川 成都 610059)

摘 要: 在综合考虑空间数据的存储特征和算法复杂度的基础上, 提出了一种基于自适应间隔游程编码的空间数据实时无损压缩方法. 该算法改进了传统游程编码的规则, 并采用自适应间隔编码技术, 使空间数据的实时 TCP 传输成为可能.

关键词: 自适应; 游程编码; 无损压缩; 空间数据

中图分类号: P208

文献标识码: A

文章编号: 1000-7180(2014)07-0121-04

Real-time Compression Algorithm for Spatial Data Based on Self-Adaptive Interval Run Length Coding

YAO Dan-dan¹, CHEN Jun², LU Han-yu³, MIAO Fang^{1,4}

(1 College of Geophysics, Chengdu University of Technology, Chengdu 610059, China;
2 College of Resources and Environment, Chengdu University of Information Technology, Chengdu 610225, China;
3 College of Electronic and Information, Guizhou University, Guiyang 550025, China;
4 Institute of Spatial Information Technology, Chengdu University of Technology, Chengdu 610059, China)

Abstract: In this paper, a real-time compression algorithm for spatial data was proposed to improve the quality of spatial information service. The algorithm includes two key points. Firstly, the coding rules of RLE were developed. Only if the number of repeated byte reaches to 3 or more, the stream segment is allowed to code with the two bytes, one is the repeated number, another is the repeated byte. Otherwise, the stream segment is output directly after the number of uncoded bytes. Secondly, self-adaptive interval run length coding was put forward. The intervals of 24, 16 and 4 are used to test the compression ratio of the origin spatial data stream. The interval of the highest compression ratio is used as the final coding strategy. It was proved that the algorithm of this paper can be used to the spatial data transmission and make real-time transmission of spatial data possible.

Key words: adaptive; run length coding; lossless compression; spatial data

1 引言

空间数据尤其是栅格数据为海量数据, 数据存储和移动成为空间信息服务的瓶颈. 为降低空间数据在网络中传输量, 数据压缩成为关键技术之一. 为保证数据在传输过程中压缩的数据正常还原, 需使用无损压缩编码方案. 在实时性要求不高的空间数据共享应用领域中, 事先对整个数据压缩, 网络传输

压缩后的文档, 客户端下载后再解压. 由于实时性要求不高, 可以使用高压缩率的算法压缩空间数据, 无需考虑压缩与解压的时间性能. 在空间信息服务中, 服务器与客户端的 TCP 数据流的实时性传输是评价空间信息服务性能的一个重要指标. 由于实时性的要求, 对空间数据以文档的方式整体压缩传输已经不能满足需求, 需要研究基于空间数据实时的、具有较高压缩率的算法.

收稿日期: 2013-09-26; 修回日期: 2013-11-05

基金项目: 国家自然科学基金资助项目(61071121)

目前,无损压缩算法主要有熵编码法和字典编码法等^[1-7].作为一种无损压缩编码技术,熵编码法将数据流中每个字母用一段不同长度的比特(Bit)代替.霍夫曼、算数编码、香农-范诺编码等压缩方式均属该类;字典编码利用信息源符号之间的相关性,为数据流中的信息建立字典.根据字典,重复信息使用编码代替,从而达到压缩数据的目的.RLE、LZ77/78、LZSS、LZW等均属于字典编码.霍夫曼编码作为一种经典的压缩算法,算法复杂度主要集中在概率统计和概率树的生成,当信息中各符号出现概率差别不大时,霍夫曼编码不能很好地压缩,甚至可能得到更大体积的压缩数据;LZ系列的算法对多数信息具有较高的压缩率,但算法复杂度集中在符号串的比较上;算数编码的算法复杂度则集中于浮点数的计算上.

游程编码原理简单,对数据源的一次遍历即可获得压缩结果,算法复杂度小,时间响应度高,适用于基于TCP数据流的数据实时压缩^[8-11].传统游程编码主要按字节为单位进行编码.当空间数据的存储单位为2字节、4字节或8字节时,传统的游程编码不能对其压缩.因此,对游程编码进行改进,提出适合于空间数据压缩的游程编码方法.

2 空间数据存储特征分析

空间数据分为矢量数据和栅格数据.矢量数据一般以坐标表示空间目标的位置.在二维矢量中,坐标记录格式为(X,Y);在三维矢量中,坐标记录格式为(X,Y,Z).在一般情况下,坐标的各分量使用双精度数值存储.点用一个坐标表示其位置;线通过坐标序列表示,按顺序将点坐标连接起来,构成空间中的曲线;在面条结构中,面的表示和线相同,使用坐标序列表示面的边界,坐标序列的起点和终点坐标值相等,构成了空间上封闭的环.复杂多边形由顺时针记录的外环和逆时针记录的一至多个内环构成.

栅格数据使用矩阵表示,矩阵中每一个元素为一个像元,像元值代表地面某一个方面的属性,如高程、降水、反射率等.根据需要,像元值可为字节、短整数、长整数、浮点数或双精度数值.

传统游程编码主要以字节为单位进行编码,当数据流中存在大量连续的相同字节时,游程编码可获得很好的压缩比.由于传统游程编码以位或字节为单位编码,当栅格像元以一个字节表示,并且相邻像元取值相同的概率大,该算法可能取得很好的压缩效果.但是,当空间数据以2字节、4字节或8字

节单元为单元存储时,以字节流考察空间数据,数据重复率降低.以一个具有5个像元的4字节整数数组为例,数组元素依次为

$$\{318, 315, 322, 316, 318\} \quad (1)$$

按字节为单元,上数组存储的值为

$$\{62, 1, 0, 0, 59, 1, 0, 0, 66, 1, 0, 0, 60, 1, 0, 0, 62, 1, 0, 0\} \quad (2)$$

由于上述字节流不存在连续的重复字节,传统游程编码方法只能得到更大体积的压缩数据流.为保证矢量坐标数据和所有类型的栅格数据均可得到理想的压缩效果,试图提出一种自适应游程编码算法,该算法分为两部分:游程编码的规则改进、自适应间隔编码.

3 游程编码的规则改进

传统游程编码对二进制数据流压缩时,使用两个字节表示一个压缩单元.第一个字节表示重复个数,第二个字节表示重复的字节.当数据流中存在大量连续的相同字节时,游程编码可获得很好的压缩比,理论上压缩率可达到1/128.但是,如果数据流中相邻字节均不相同,传统游程编码压缩文件的体积为原数据流的2倍.针对游程编码的缺点,前人提出了各种改进的游程编码.最简单的改进方法是当数据流中存在不重复的字节时不对数据作任何处理,直接将原始数据作为压缩后的数据^[4].

为进一步提高游程编码的压缩率,在RLE中设置两种编码规则,一是重复字节编码规则;二是不重复字节编码规则.

规则1:当重复字节大于等于3个时,使用2个字节表示编码结果.第一个字节取值为127+重复个数,第二个字节表示重复码;

规则2:对于不能使用规则一编码的片段,首先使用一个字节表示未编码的字节数,其后直接记录原始数据片段.

由于一个字节表示的数据范围为[0,255],规则1一次最多可对128个重复字节编码,多于128个重复字节必须拆分;在游程编码过程中,不符合重复字节编码规则的其余字节流按规则2编码,为避免与规则1冲突,规则2一次最多只能对129个字节编码,大于129个不重复字节的必须拆分.

考虑如下字节流:

$$25668777777 \quad (3)$$

根据规则1,字节7可进行游程编码,其余的按规则2输出,则得到的编码字节流为

5 2 5 66 8 133 7 (4)

改进 RLE 算法利用规则 1 对数据压缩,理论上最高的压缩率为 $2/128$;编码规则 2 则会增加压缩后的数据大小,在极端情况下,编码后的数据大小为原数据大小的 $130/129$,降低了 RLE 算法编码增大数据流体积的几率。

算法改进了前人将重复字节数为 2 使用规则 1 编码的传统,可进一步提高数据的压缩率。这是因为,当重复字节等于 2 时,按规则 1 只能得到相同体积的压缩数据流,并且增加了不重复字节编码的数量。每增加一个不重复字节编码流,就会增加一个字节。如上面给出的原始字节流,如果将阈值改为 2,则规则 1 可对字节 6 和字节 7 编码,但剩余字节流从重复字节 6 位置上被强制拆分,得到的压缩数据流为

22 5 129 6 1 8 133 7 (5)

4 自适应间隔游程编码算法

矢量坐标使用双精度数值存储,栅格数据存在 2 字节、4 字节或 8 字节的像元类型。以 DEM 为例,每一个像元表示一个地点的高程,高程数据可能为 4 个字节的整数或浮点数,甚至是 8 个字节的精度小数,相邻区域的高程值差异一般不大,这为无损压缩提供了可能。但是,以字节为单元考察高程值,数值重复率被降低。假定(1)为 4 字节为单元存储的 DEM 数据片段,以字节为单元的数据流为(2)。考虑到 DEM 的像元存储单位为 4 字节,按 4 字节为单位分别对 $4n+1$ 、 $4n+2$ 、 $4n+3$ 和 $4n+4$ ($n \in \mathbb{N}, n \geq 0$)位置进行游程编码,则编码字节流分别为

$4n+1$ 位置: {5, 62, 59, 66, 60, 62} (6)

$4n+2$ 位置: {132, 1} (7)

$4n+3$ 位置: {132, 0} (8)

$4n+4$ 位置: {132, 0} (9)

将上述编码组合起来得到最终的编码串为

{5, 62, 59, 66, 60, 62, 132, 1, 132, 0, 132, 0} (10)

综合(2)和(10),得到间隔游程编码的数据压缩率为 $12/20$ 。

由于事先不知道空间数据的存储单位,使用自适应编码方案。结合矢量坐标值的分布规律和栅格数据流的特点,选择 4、16 和 24 三种间隔编码方案。首先以 24 位为间隔,将二进制流分成 24 等分按上述方法游程编码;然后按照每 16 位为间隔,按 16 等分编码;最后按照每 4 位为间隔,按 4 等分编码。为提高算法响应时间,不再按 8 位、2 位、1 位的间隔划

分。最终的编码方案选择压缩率最高的间隔编码,并在压缩数据流头部标明间隔信息。从编码方案来看,自适应游程编码算法复杂度仅为传统 RLE 的三倍。

5 算法评价

为证明算法的有效性,对两个空间数据(51.4 MB 的二维矢量坐标数据和 218.6 MB 的浮点型栅格数据)在相同网络环境下使用不同压缩方法进行 TCP 数据下载测试,一次测试包括数据压缩、数据传输、数据解压完整流程。从压缩率、平均网速、等效网速和 TCP 响应时间几个方面综合评定。表 1、2 中的 RLE2 表示本文算法。

表 1 二维矢量坐标数据不同压缩算法对比

	无压缩	RLE	HuffMan	Zlib	BZip2	RLE2
压缩率/%	100	99	74	65	51	58
平均网速/(Mb/s)	2.45	2.45	1.20	1.54	0.70	1.66
等效网速/(Mb/s)	2.45	2.45	1.61	2.34	1.35	2.86
TCP 响应时间/s	21	21	32	22	38	18

表 2 浮点型栅格数据的不同压缩算法对比

	无压缩	RLE	HuffMan	Zlib	BZip2	RLE2
压缩率/%	100	99	79	6	13	33
平均网速/(Mb/s)	2.38	2.38	1.25	0.17	0.19	1.00
等效网速/(Mb/s)	2.38	2.38	1.57	2.60	1.51	3.04
TCP 响应时间/s	92	92	139	84	145	72

从表 1 和表 2 可见,RLE2、ZLib 和 BZip2 具有较高的压缩率。由于本文算法复杂度小,同时具有较高的压缩率,在空间数据一次性完整下载过程中总耗时最少,提高了 TCP 的时间响应效率。

6 结束语

本文提出了一种基于自适应间隔游程编码的空间数据实时无损压缩方法。该算法依据游程编码算法复杂度低的特点,对传统游程编码进行改进,使游程编码较好地应用于空间数据的实时无损压缩。实验证明,本文算法运算速度快,压缩率较高,对于空间信息服务的快速响应具有重要的意义。另外,根据本文算法的特点,它还可以应用于数组数据的网络快速传输等方面。

参考文献:

- [1] Eengelson V, Fritzson D, Fritzson P. Lossless compression of high-volume numerical data from simulations[C]//Proceedings of the Data Compression Con-

- ference. [s. l.]:IEEE,2000: 574-586.
- [2] Larrauri J I. A new algorithm for lossless compression applied to two-dimensional static images [C]//Proceedings of the Proceedings of the 6th international conference on Communications and Information Technology, and Proceedings of the 3rd World conference on Education and Educational Technologies. [s. l.]: IEEE,2012;56-60.
- [3] Magli E. Multiband lossless compression ofhyperspectral images [J]. Geoscience and Remote Sensing, IEEE Transactions on, 2009, 47(4): 1168-1178.
- [4] 籍俊伟,林小竹. 图像编码中的 RLE 方法研究[J]. 石油化工高等学校学报,2004,17(1): 83-87.
- [5] 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,2006,42(5): 172-175.
- [6] 赵军. 图像的无损压缩[J]. 重庆邮电学院学报:自然科学版,2001,13(3):49-52.
- [7] 刘丽莉,付东翔,王亚刚. 一种图像自适应预测分组编码无损压缩方法[J]. 微电子学与计算机,2012,29(10): 148-151.
- [8] Akhter S, Haque M. ECG compression using run length encoding [J]. European signal processing (EU-SIPCO-2010),2010:1645-1649.
- [9] Hauck E L. Data compression using run length encoding and statistical encoding. US4626829A[P]. 1985-08-19.
- [10] Hinds S C, Fisher J L, D'amato D P. A document skew detection method using run-length encoding and the Hough transform[C]//Proceedings of the Pattern Recognition, 1990 Proceedings, 10th International Conference on. [s. l.]:IEEE,1990:464-468.
- [11] Messom C H, Sen Gupta G, Demidenko S N. Hough transform run length encoding for real-time image processing [J]. Instrumentation and Measurement, IEEE Transactions on, 2007, 56(3): 962-967.

作者简介:



姚丹丹 女,(1987-),博士研究生. 研究方向为空间信息技术、智能信息处理.

陈 军 男,(1979-),博士. 研究方向为遥感与 GIS 技术及应用.

卢涵宇 男,(1978-),博士. 研究方向为智能信息处理、3S 技术应用.

苗 放 男,(1958-),博士,教授,博士生导师. 研究方向为空间信息技术及应用.

(上接第 120 页)

到运行结果和主机上的运行结果保持一致. 基于此, 可以将主从机之间的互联互通方式进行推广, 并可将现有研究成果应用于远程控制.

参考文献:

- [1] 杨晖,胡永建,林志泉. 基于 Linux 和 S3C2410A 嵌入式图像传输系统设计[J]. 微计算机信息,2007,23(32):20-24.
- [2] 肖儿良,毛海军,鞠军平,等. 基于开源软件 MJPG_Streamer 的智能视频监控系统设计[J]. 微电子学与计算机,2013,30(006): 84-87.
- [3] 刘昌举,戴基智,龙再川,等. 基于嵌入式 Linux 和 S3C2410A 的远程 CCD 图像采集系统[J]. 半导体光电,2006,27(6):777-783.
- [4] 何剑锋,李祥,何月顺,等. 基于 S3C2410A 的嵌入式 LINUX 下 OLED 模块驱动设计[J]. 液晶与显示,2010,5(1):90-94.
- [5] Yun Sin-quan,Lu Qiang,Qian Pei-del. One implementation of Linux application based on Qt/Embedded[J]. Computer Application and Software, 2006, 23 (2): 105-107.
- [6] Liu C L, Layland J W. Scheduling algorithms for multiprogramming in a hard real-time environment [J]. Journal of the ACM,1973,20(1):44-61.
- [7] 杨继华,严国萍. 基于嵌入式 Linux 与 S3C2410A 平台的视频采集[J]. 单片机与嵌入式系统应用,2004,3(11):69-72.
- [8] Tian Lei. The protection of water resources environmental based on embedded linux system[J]. Water Resource and Environmental Protection,2011,3(2):2096-2099.

作者简介:



田 磊 男,(1980-),博士生,讲师. 研究方向为嵌入式系统、嵌入式 Linux 系统的图形研究.