

气象资料和历史天气图图像 数据压缩方法研究

花灿华 郭发辉

(国家气象中心)

摘要 文章较详细地介绍数据文件和图像数据压缩的一般原理,论述对存档的气象资料数据文件使用的压缩方法,探讨适用于历史天气图图像数据压缩的技术方案。

关键词 压缩方法 气象资料 图像数据

当今,由于信息量剧增,存储媒介的昂贵,传输效率的要求以及数据保密等诸多因素,使数据压缩技术得到迅猛的发展,并越来越广泛地应用于各种硬软件的界面上。对于每天需要搜集、处理、传输和存储庞大数据、图像数据的气象部门来说,数据压缩技术正在得到高度的重视和广泛的应用。

1 数据压缩技术的一般原理

数据压缩技术是一项信息处理的技术,它广泛地应用于通信、语音、图像处理、信息存储等领域。所谓压缩技术就是在不丢失信息或在一定质量损失容限的基础上,通过改变数据表示及存储格式,简单地说,就是启用有效的编码来表示数据,使得数据存储时能够减少空间,在通讯传送时能够缩短时间,减少传输带宽,与此同时,还提供了数据保密的附加作用。

数据压缩技术都是以信源为其研究对象的。因而对于不同领域的信源,其压缩方法也是多种多样的,但就其整个压缩技术而言,其压缩方法可分为两个主要类型,即逻辑压缩和物理压缩。逻辑压缩主要是从存储或数据表示格式上考虑,以牺牲原始数据表示的直观性去换取数据规模的减少^[1]。例如日期的表示方式,“1981 APRIL 01”需要存储空间16个字节(byte)(月份按最长英文单词所占

长度),如以数字表示,年、月、日各2个字节共6个字节就够了,若以二进制位(bit)表示,则为(00001 0100 1010001)₂,只需两个字节。由此可以看出,根据数据特点,采用不同的表示方式,就可以大大节省存储空间。物理压缩则主要是从数据中组成元素的重复或冗余上考虑,一般用于数据被作为独立的有区别的项目编码并且字符和字符组出现概率有明显不同的地方。对出现频率较高的字符或字符串给予较短的编码,出现频率较低的则给予较长的编码,这样就可以有效地减少存储空间。

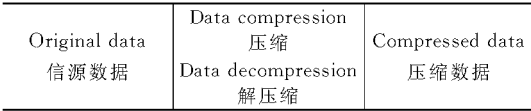
评价一项压缩方法时,总是以其压缩率、失真度和复杂度这三个指标来衡量。压缩率表示压缩后存储空间的节省率,其计算方法如下:

$$\text{压缩率} = \frac{\text{原始数据总长度}}{\text{压缩数据总长度}}$$

失真度是原始数据失真的量度,允许有一定的精度损失,换取高的压缩效率。复杂度就是压缩方法的复杂程度,即编码和解码算法的复杂性。从数据压缩的目的来说,希望是高压缩比,低失真度和简单的算法,但在实际应用中,这三个指标是相互制约的。因此,针对某种压缩方法,必须从实际出发,对这三个指标进行综合分析,力求达到最佳。

不管采用何种压缩方法,其压缩过程都

是：由读取符号串(symbolical string)和把它转换为码字(code)这两者组成，即对信源数据集的数据，经过一种编码技术变换到空间更小的压缩数据集之中，通过恢复(解压缩)又可从压缩数据集的数据还原到信源数据集之中。其基本数据压缩框图如下：



基本数据压缩框图

2 常用的数据压缩方法

数据压缩方法很多，下面只介绍其中几种。

2.1 空白压缩

空白(Null)或空格(Blank)压缩是最早被使用的一种数据压缩技术。空白压缩就是对原始数据流进行扫描，从中寻找重复的空白或空格，如果找到这样的序列，就用有序字符对替换空白序列，其中第一个字符标明空白压缩的出现，第二个字符指出空白序列的字符个数，这种过程叫做空白压缩编码过程。格式如下所示：

压缩指示字符	空白或空格个数
--------	---------

共2个字节

例如：原始数据流为 XYZbbbbQRX

压缩数据流为 XYZSc5QRX

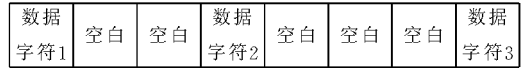
其中 Sc 为压缩指示符，b 为空格。

当空白压缩数据流被传送时，接收设备或程序负责寻找指示空白压缩的那个特殊字符。一旦找到，接收设备或程序则认为其下一个字符是被压缩的空白个数。基于这个信息可以把压缩数据恢复为原始数据，这种过程称为空白压缩解码过程。

2.2 位映像(Bit mapping)压缩方法

这种压缩技术在被压缩的数据中存在着大量的特殊类型如数字或特殊字符如空格，而且这种特殊字符不一定连续存在时，显得比较有效。其压缩过程为(1)把原始数据按8

个字符(character)分组；(2)给每一组一个位映像字节，该字节8个位分别与该组8个字节对应，对应字节为空白置0，否则置1；(3)去掉该组中所有空白或空格即得压缩数据。例如：原始数据流：



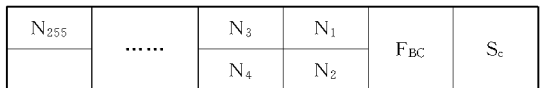
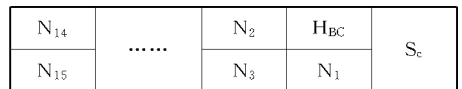
压缩后数据：



此例中的原始数据流如果采用空白压缩只能从8个字节压缩到7个字节，而用位映像压缩则可压缩到4个字节，即3个字节的数据和一个位映像字节。

2.3 半字节成组(half-byte packing)压缩方法

这种压缩方法主要是对字符集之中数字连续出现的相同部分(如 EBCDIC 码表示数字的前4个比特位均为“1111”)进行压缩。因此，在压缩数据中应有标志压缩开始的标识符 Sc，此外，还应有压缩字符串的长度，其长度值可以用半字长，也可以用一个字节，从而分为半字节计数的半字节成组和整字节计数半字节成组方法，两者的差别仅是计数范围不同。为把数据压缩为以半字节为单位，可以考虑采用下面图示的格式。



← 数据流

其中，H_{BC}：半字节计数 ≤ 15；

F_{BC}：整字节计数 ≤ 255；

S_c：半字节成组压缩的标识字符；

N₁, N₂,：一组以半字节为单位的压缩后数字。

最后,根据上图可得到编码方式如下:

$$S_1=0; \quad S_2=10;$$

$$S_3=110; \quad S_4=111$$

3 气象资料常见的几种压缩方法

对气象数据进行有效的交换和存贮,对气象业务、科研和服务都是必不可少的。采用压缩方法存贮气象数据,其目的就是为了更好地满足这些要求。目前,国内外气象资料常用的压缩方法一般有如下几种:

3.1 半字长存贮压缩方法

半字长存贮压缩方法是一种常用的存贮压缩方法,它的每个数据占半个字长(word)(二字节),可表示-32768到32767范围内的任一个数值,而这个数的范围可以满足一般气象数据的表示。英国气象局海表水温网格点资料就是采用这种压缩方法。

3.2 偏差值存贮压缩方法

这也是一种常用的存贮压缩方法,对于常规气象资料来说,大部分气象要素值都可用10个比特位表示。其压缩公式为: $X = (Y - R)/2^S + D$,式中 X 为压缩值, Y 为原值, R 为基值, S 为比例因子, D 为偏差值。在实际应用中,基值 R 一般取平均值,比例因子取决于数据整数化后的有效位数。偏差值的确定是为了保证 X 值为正数, X 取正整数。由 X 的最大可能值,决定 X 存储的比特位。美国 NCAR 采用这种方法,对北半球逐日逐次海平面气压网格点、全球热带网格点、南半球逐日逐次海面气压、500hPa、300hPa 高度网格点资料都采用这种压缩方法。

3.3 选择比例因子的压缩方法

这也是一种常用存贮压缩方法。其压缩公式为: $X = (Y - R)/2^S$,式中, X 为压缩值, R 为所有原值数据中的最小值, S 为比例因子。澳大利亚南半球网格点资料就是利用这种压缩方法。

3.4 BUFR 和 GRIB 码

BUFR 和 GRIB 码都是世界气象组织(WMO)建议推广使用的具有较高压缩率的

二进制值表示代码。GRIB 码适用于表示数值天气预报系统加工分析和预报产品。BUFR 码适用于表示各种观测资料,具有简洁、灵活和自定义的优点^[4]。这两种代码都与计算机无关,它们的压缩率一般在50%以上。正是这种高压压缩率使得数据传输速率相对地提高,并减少了磁盘、磁带存储空间。BUFR 的自定义特性指的是一份 BUFR 报中不但包含了气象数据,还包含这些数据的完整描述:有关数据的物理含义、单位、精度、压缩方法及数据所占的比特位数等。这些信息即“数据描述”都包含在表中,这些表是 BUFR 文件的主要内容。由于具有自定义的特性,使 BUFR 码有很强的适应性。如果要发展一种新的观测或观测平台,则只要在 BUFR 文件中增加新的观测所涉及的要素的“数据描述”,不需要另创立一套电码来表示和传送这种新资料。

4 图像数据压缩方法

4.1 图像数据压缩原理

支持图像数据的可压缩的原理是基于图像信息的冗余性。若把图像看成是二值的文件数据,则图像的压缩与数据文件压缩一样可采用 Huffman 编码,算术编码和游程编码等方法。

4.2 图像数据压缩分类

从图像自身的形式可分为静止图像和非静止图像。静止图像也称纸质图像,非静止图像如动画图像、电视图像等,是在屏幕上或机器上的图像。从图像数据压缩解码后复原为原图像的结果看,图像数据的压缩可分为无损或信息保持型压缩编码和有损或信息非保持型压缩编码。无损压缩是指一个比特位都不能丢失的压缩技术,用于压缩数据库文件和保证压缩后能还原为与原来完全一致的图像。有损压缩则是允许一定的精度损失换取高的压缩效率,对一些不要求压缩解码后没有些微误差的图像是适用的。

4.3 图像数据压缩举例

任何文字、图像均可像数图传真那样处理,即通过一行一行的扫描,生成在任何给定的时间点上代表被扫描的一个小区域的亮或暗的信息源。所谓传真就是将数据流传送到接收站,然后利用驱动图像再生设备还原成原始数图。传真的结果的清晰度取决于扫描的细致程度。数图被扫描后的亮或暗的信息可以看成“1”或“0”的数据流,每一数据点就是一个“像元”,也称“像素”。对图像数据流的压缩有多种方法,下面介绍的相对差表示方法仅为其中的一种。

将一次完整的图像扫描结果保存在特定的存贮区域,与后继的扫描结果相比较,在相对压缩过程中,只记录与前一次扫描之间的差。如:

N 行0 0 0 0 0 0 1 1 1 0 0 0 1 1 0
 N+1行.....0 0 0 0 0 1 1 1 1 0 0 1 1 0 0
 Δ Δ Δ

当被扫描的文本包含的白像元远比黑像元多时,相对差表示法是最有利的数据压缩方法。这种方法具体又可分为以下两种:

(1) 位置标识法

相对差有不同记录方法,其中一种是记

40	6	80	20	175	4	350	31	480	8	930	14	1250	16	1310	5	1340	4
----	---	----	----	-----	---	-----	----	-----	---	-----	----	------	----	------	---	------	---

图1 位置变化及长度记录

40	6	40	20	95	4	175	31	130	8	450	14	300	16	60	5	30	4
----	---	----	----	----	---	-----	----	-----	---	-----	----	-----	----	----	---	----	---

初始位置 位 移

图2 位移记录

录变化位置,即位置标识;如有连续变化,则在位置之后紧接一个计数,指出改变序列的长度。表1分别列出像元变化的位置及其变化的长度,记录被传送的数据如图1所示。

被传送的数据为:

表1 像元变化位置及其变化长度

变化位置	长度
40	6
80	20
175	4
350	31
480	8
930	14
1250	16
1310	5
1340	4

(2) 位移标识法

这是一种基于位移概念的另一种表示改变位置的方法,即在计算相对变化后,不是像位置标识那样传送相对变化起始位置和连续变化的数字个数,而只是传第一个初始位置,而后传送相对于前一个序列变化位置的位移,如图2所示。

5 历史天气图图像数据压缩

5.1 地面、高空历史天气图数据压缩的必要性

地面、高空历史天气图是天气预报分析图经过整理、修订后出版的存档图像资料,是天气、气候研究与评价的重要资料。但由于这种

资料体积大、易损坏,保管和使用都有很大困难。如何妥善保存和应用这些资料,一直是气候资料工作者急需解决的问题。在这方面,对历史天气图缩微工作已经进行了十多年,其缩微胶片(卷)的拷贝、阅读和检索达到了解决历史天气图能长期保存、缩小体积和方便使用的目的。但缩微产品非数字化载体,不能直接应

用于计算机。目前,国际国内对图形压缩的研究发展得很快,图形压缩方法的研究取得了可喜的成果。国际上有图像压缩方法标准化的专门组织,征集、审议并推荐图形压缩方法。因此,在计算机技术迅猛发展和体积小、容量大的存贮介质得到普遍应用的今天,图像数据压缩可以较容易地实现。从现在起着手进行天气图图像数据的压缩有着很强的实际意义和巨大的经济效益。

5.2 历史天气图图形数据压缩方案的选择

已出版的历史天气图可作为广播传真图对待,这样,通过专门的扫描仪(放片机)将产生计算机或工作站可以调用的图形文件进行压缩。

一种图形数据压缩方案的选择主要以图

形信息的压缩率高、无差错压缩、编解码时效高、减少开销等为原则。据此,我们选用 CCITT. T6的编码方案用于历史天气图的图形数据压缩。这一编码方案符合国际标准,压缩率较高,其使用的设备和技术“八五”期间在国家气象中心已得到较长时间的应用^[5]。

5.3 CCITT. T6编码原理

(1) 本方案是以 Huffman 编码为基础的二维无损压缩编码,采用二维逐线编码方法,即当前的正在编码的扫描线上的每一迁移像元的位置,是根据位于参考扫描线上的相应参考像元的位置来编码的。所谓迁移像元,就是在同一扫描线上,其颜色与前一像元的颜色不同的像元。对于参考扫描线上和正在编码扫描线上,共有五种类型的像元,如图3所示。

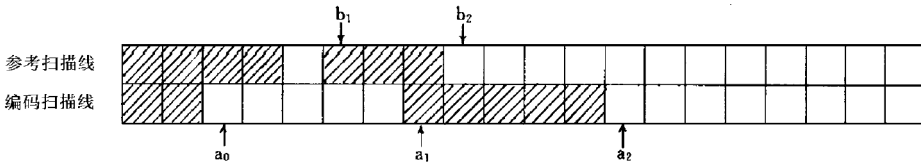


图3 迁移像元类型

表2 T6编码模式

模	需编码像元	符号		码字
通过	b_1, b_2	P		0001
水平	$a_0 a_1, a_1 a_2$	H		$001 + M(a_0 a_1) + M(a_1 a_2)$
垂直	a_1 正好在 b_1 之下	$a_1 b_1 = 0$	$V(0)$	1
		=1	$V_R(1)$	011
		=2	$V_R(2)$	000011
	a_1 在 b_1 的右面	=3	$V_R(3)$	0000011
		=1	$V_L(1)$	010
		=2	$V_L(2)$	000010
a_1 在 b_1 的左面	=3	$V_L(3)$	0000010	

从上图可以看出, a_0, a_1, a_2 在编码扫描线上,其中, a_0 是一个假设的白迁移像元,具体编码时,它的位置由前一个编码模式决定; a_1, a_2 分别为 a_0, a_1 右边的下一个迁移像元。

b_1, b_2 在参考扫描线上, b_1 位于 a_0 的右边且是与 a_0 颜色相反的第一个迁移像元; b_2 是位于 a_1 右边的下一个迁移像元。

(2) 编码模式

根据参考扫描线与正在编码扫描线上各迁移像元的相对位置的不同,T6编码模式有如下三种:(a) 通过模, b_2 位于 a_1 的左面;(b) 水平模, $|a_1 b_1| > 3$;(c) 垂直模, a_1 的位置应相对于 b_1 的位置进行编码。具体见表2。

(3) 编码方法

在编码过程中,首先要鉴别出正在编码的扫描线上每一迁移像元进行编码时所要使用的编码模式,从模式表中找出一个适当的码字进行编码。模式表中水平模 $M(\times \times)$ 码字从 T4 结尾码和组合基干码表查找(表格从略)。

参考文献

- 1 Held G. Data compression, techniques and applications. Hardware and Software Consideration. Wiley Heyden Ltd, 1983
- 2 Nelson M. 数据压缩技术原理与范例. 科学出版社, 1995
- 3 骆新, 陈睿. 数据压缩实用技术, 学苑出版社, 1993
- 4 国家气象中心. 实时气象资料数据库系统. 气象出版社, 1992
- 5 于福新. CCITT·T6建议编解码原理及算法实现. 台风、暴雨灾害性天气信息通信传输技术和数据处理技术的研究(第一分册). 气象出版社, 1996, P123—132