维普资讯 http://www.cqvip.com

用于空间数据系统的无损数据压缩技术

李维忠　贾小铁　徐志立
（内蒙古工业大学信息工程学院　呼和浩特010062）

Lossless Data Compression for the Space Data System

LI Wei-Zhong　JIA Xiao-Tie　XU Zhi-Li
(Information Engineering School, Inner Mongolia Polytechnic University, Hohhot, China. 010062)

Abstract　This paper gives a brief and clear description to the Lossless Data Compression Recommendation for Space Data System Standards, and presents its own opinions for some unintelligible algorithm detail in the recommendation.
Keywords　Lossless data compression, Space data system, CCSDS

1. 引言

在航天科技领域中，由航天器获得的有关图像或非图像的科学数据有必要经过压缩处理后再予以传输。国际空间数据系统咨询委员会(CCSDS)认为这种必要性至少体现在三个方面：节俭传输信道的带宽；降低缓冲及存储设备的需求；以及缩短在给定速率下的传输时间。

根据 Shannon 的信息学理论，信源数据中包含的信息量可以通过计算其所具有的熵值来得到精确的量化，整条信息的熵值对应于完整表述该信息所需的最小数据量。通常情况下信源数据中存在大量的编码冗余，通过适当的编码变换手段剔除源码中的数据冗余即可得到有效的数据压缩。数据压缩技术分为有损压缩与无损压缩两种。有损压缩技术会在剔除冗余的同时丢失信息中的部分成分，在一些容许细节失真的场合这种技术往往因其可以获得较高的压缩比率而得到应用，比如常见的 MPEG、JPEG 等。但对于象卫星遥感遥测这类贵重的科学数据来说，显然应该采用无损的数据压缩技术。而无损压缩技术所能达到的压缩效果应该尽可能地接近源码的熵值。

常见的无损数据压缩技术有 Huffman 编码、算术编码、LZW 编码及其各自的变异版本。CCSDS 在对包括上述在内的一些算法进行了广泛的研究并比较了各自针对一组测试图像[5]所表现出的压缩性能后，选择了 Rice 算法作为指导各成员国在开发空间数据系统时所应遵循的推荐标准。该标准于1997年以文[1]的形式正式发布。我国是 CCSDS 的观察员国，在空间开发领域参与国际合作或吸收先进的国际经验无疑具有重要的意义。

但是，仅仅根据 CCSDS 的文[1,2]并不足以完整地理解并实现建议书所推荐的压缩方案。为了具备必要的灵活性以适应技术的发展，建议书指出了一些要素的可选性质，但同时也忽略了一些实现方案时所需了解的算法与技术细节。笔者在研究分析后撰写本文，目的就在于力求简明清晰地对 CCSDS 的建议方案给出一个易于理解的阐述。

2. Rice 压缩算法的信源编码原理

同 Huffman 算法一样，Rice 算法也采用了可变长度的符号码字(表1)作为基本的压缩手段。然而 FS 码更为简单的规律性使其在编码与解码操作时无需通常的查表过程。

表1　Rice 算法的基本序列码编码规则

预处理后的样本值 δ	基本序列码(FS)
0	1
1	01
2	001
…	…
2^n-1	2^n 个0后跟个1

Rice 信源编码器由预处理器与自适应熵编码器两部分组成(图1)。

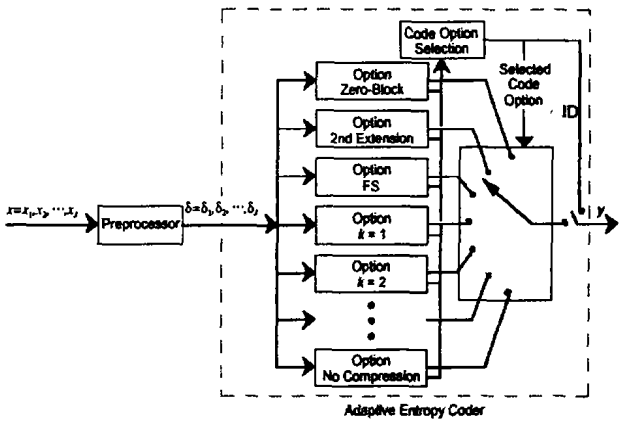

图1　带有预处理器的自适应熵编码器

2.1　预处理器

信源数据中的每个单位数据(如文件中的一个字节或图像中的一个像素)称之为一个样本，每个样本均为 n 个比特。信源数据在输入预处理器前先要被分割为长度为 J 的样本块以提高编码器对熵值变化的适应性。CCSDS 建议书规定 J=8 或16，而8≤n≤32。

$x=x_1,x_2,\cdots,x_J$ 经预处理器变换为适合自适应熵编码器处理的 $\delta=\delta_1,\delta_2,\cdots,\delta_J$，这里的每个 δ_i 仍有 n 个比特，即0 $\leqslant\delta_i\leqslant(2^n-1)$。预处理过程其实就是对数据进行解相关处理后再将结果变换为具有更佳概率分布的正整数的过程，过程中的这两个步骤分别称之为预测和误差映射。

所谓预测，就是求当前数据值 x_i 减去预期值 $\hat{x}_i$ 后的差值，也即预测误差。CCSDS 建议书[1]仅介绍了预测技术中的单元延迟(Unit-Delay)法，在该方法中预期值 $\hat{x}_i=x_{i-1}$。

李维忠　副教授，硕导，研究方向：计算机网络。贾小铁　教授，硕导，研究方向：自动控制。徐志立　研究生，讲师，研究方向：计算机网络。

维普资讯 http://www.cqvip.com

表面上看，两个 n 比特数的差将产生一个动态范围在[$-2^n+1, 2^n-1$]间的 n+1比特的数值。然而对于每次具体的预测来说，其预测误差 Δ_i 只会有2^n 个可能的值，我们用 x_{min}和 x_{max}分别表示任一输入样本 x_i 可能的最小值和最大值，显然 Δ_i必是区间[$x_{min}-\hat{x}_i, x_{max}-\hat{x}_i$]中的一个值。由于我们希望当 $|\Delta_i|>|\Delta_j|$时其概率分布呈现 $p_i<p_j$ 的性质，因此用函数

$$\delta_i=\begin{cases}2\Delta_i & 0\leqslant\Delta_i\leqslant\theta\\ 2|\Delta_i|-1 & -\theta\leqslant\Delta_i<0\\ \theta+|\Delta_i| & otherwise\end{cases}$$

（其中 $\theta=\min(\hat{x}_i-x_{min}, x_{max}-\hat{x}_i)$；对于 n 比特的正整数 $x_{min}=0, x_{max}=2^n-1$）

对 Δ_i 进行映射处理，得到 n 比特的正值作为自适应熵编码器的输入。

CCSDS 推荐了上述的预处理方案，但同时也允许采用其它的预测和映射方法或者在必要时将预处理器旁路。

2.2 自适应熵编码器

Rice 自适应熵编码器具有若干个处理选项(图1)，编码器根据每个输入样本块 δ 的情况对其作出最优的选择。选项标识码(参见文[1])要插放在该块输出数据的前部，以备解码器解压时识别。

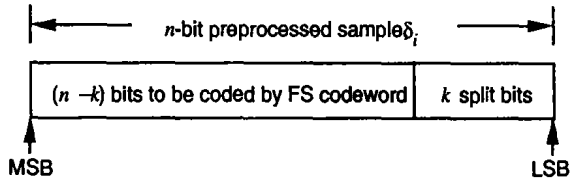

图2 k 位样本分解示意图

样本分解是 Rice 算法的核心概念之一。由于经预处理后的样本 δ_i中信息比较集中地分布在其低位部分，因此将样本分割为高低两份，对熵值较低的高位部分采用 FS 编码，对熵值较高的低位部分采用 k 位原码，由此就实现了数据压缩。显然，k 值的选取是否得当直接关系到压缩效率，来自文[2]的图表(图3)给出了恰当选取 k 值的理论依据。

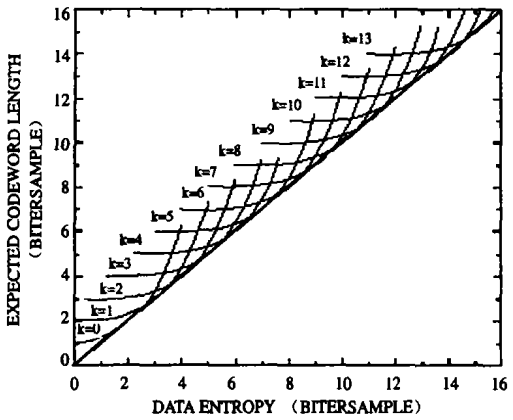

图3 对于不同 k 值的性能曲线

不难看出，编码器中的 FS 选项其实就是 k=0时的特例。

Rice 算法中的二次拓展(Second Extension)方法专门用作处理那些熵值小于2又大于0的样本块。由于此时每个样本的值很小，故可以将相邻两个样本值合并，将每个 J 样本块变换为 J/2个 n 比特的值。用于合并的变换函数为：

$$\gamma=(\delta_i+\delta_{i+1})(\delta_i+\delta_{i+1}+1)/2+\delta_{i+1}$$

对于某些熵值很高不宜压缩的样本块应当选择无压缩选项。此时在原样本块数据前插放无压缩标识符后直接作为输出。由于标识符引起的开销，输出数据在尺寸上会有轻微的增大。

表2 全零块 FS 编码

全零块的数目	FS 码字
1	1
2	01
3	001
4	0001
ROS	00001
5	000001
·	·
63	000…0001

Rice 算法还有一项有效的压缩手段是零值块处理。如果一个或几个连续的样本块为全0值时，FS 码字不再象用其它选项时那样用来表示单个样本的值，而是用来表示连续的全零块数。用来表示零值块数的 FS 码参见表2。若连续的零值块数大于 64 个时要按每段 64 块依次进行分段，“ROS (Remainder-Of-Segment)用来表示剩余段由5个或更多的零值块组成”。关于 ROS，有关文献均仅给出了上述这句令人费解的说法，文[2]的有关图示也颇难自圆其说。笔者认为，ROS 的合理用法应该是仅用来表示当前的数据是一个含有64个块的零值段，与该段是否还有后续的零值块无关。

在有预处理器的情况下，自适应熵编码器所处理的数据是映射后的样本 x_i 间差值。为了解压时能够根据差值重构出原始数据，就需要在压缩后的输出数据中包含一个原始初值作为参照样本，参照样本以 n 比特的原始形式插放在输出数据中压缩选项标识符与数据编码之间。插入参照样本的间隔不得大于256个样本块(如 J=16时是4096个样本)。在含有参照样本的块输出中，编码的个数与选项有关：对于 k 位样本分解是 J−1个；对于二次拓展则是 J/2个(按 J−1个 δ 前插个0处理)。

表3 二次拓展解码逻辑

m	β	m_s
0	0	0
1,2	1	1
3,4,5	2	3
6…9	3	6
10…14	4	10
15…20	5	15
21…27	6	21
28…35	7	28

2.3 Rice 解压算法

对应于二次拓展的解压算法步骤如下：

① m=FS 码字中0的个数，

② 根据 m 值在右表中查出 β 和 m_s，

③ $\delta_{i+1}=m-m_s$，

④$\delta_i=\beta-\delta_{i+1}$

从 δ_i 到 Δ_i 的逆映射函数如下：

if $\delta_i\leqslant 2\theta$，

$$\Delta_i=\begin{cases}\delta_i/2 & \text{when } \delta_i \text{ is even}\\ -(\delta_i+1)/2 & \text{when } \delta_i \text{ is odd}\end{cases}$$

if $\delta_i>2\theta$，

维普资讯 http://www.cqvip.com

$$\Delta_i=\begin{cases}\delta_i-\theta & \text{when } \theta=\hat{x}_i-x_{min}\\ \theta-\delta_i & \text{when } \theta=x_{max}-\hat{x}_i\end{cases}$$

where $\theta=\text{minimum}(\hat{x}_i-x_{min},x_{max}-\hat{x}_i)$

3. 最佳 k 值的计算方法

文[3]给出了计算最佳 k 值的简便方法：

$k_{opt}\sim=|\log_2(A/J)|$ 其中 $A=somm|X_i-X_{pred}|$

由于 k 应是整数，故取 $k_{opt}=\min\{k'|(2^{\wedge}k')J>=A\}$，至此，只需一条 C 语句 for(k=0;(J≪k)<A;k++);便可计算出最佳 k 值。

文中对什么是 somm 运算没有给出说明。根据信息理论分析，笔者认为就是 Σ。

结束语 CCSDS 推荐的上述无损数据压缩方法可广泛应用于多种科学数据的压缩处理场合，对于通常的图像科学数据一般能够提供2倍左右以上压缩比率。文[2]给出了针对8幅空间光谱图像数据的压缩性能报告。尽管在个别事例中其压缩率低于 LZW 算法，但在考虑了空间无线分组通讯的特点及差错抑制方面的因素后，该算法还是具有明显的优点。

Rice 的合作者之一 Pen-Shu Yeh 实现的通用科学数据信源编码器产品可参见文[6]。荷兰的 Valentina D'Arrigo 也在 UNIX 系统上完成了该算法的软件实现[3,4]。较之其它压缩算法，Rice 算法尚少为人知，希望本文能对关心相关技术的读者有所裨益。

参 考 文 献

1 Lossless Data Compression Recommendation for Space Data Systems Standards, CCSDS 121.0-B-1. Blue Book. Issue 1. Washington, ,D. C. ,CCSDS,May 1997
2 Lossless Data Compression. Report Concerning Space Data Systems Standards, CSDS 120.0-G-1. Green Book. Issue 1. Washington, ,D. C. ,CCSDS,May 1997
3 Lossless Data Compression: A Software Implementation for the Rice algorithm. Valentina D'Arrigo. WSP final report, The Netherlands, Aug. 1997
4 A Rice-based Lossless Data Compression System for Space Applications. Salvatore Coco, Valentina D'Arrigo, Domenico Giunta
5 http://www.ccsds.org/ccsds/p1_home.html
6 http://www.mrc.unm.edu/Uses/award/

(上接第98页)

规则随着管理对象的具体管理要求和行为而有相应的规则集合。可根据用户要求不同的 SLA 定制不同的 RS。

2)数据采集的非对称性，即II-Agent对各个设备代理可以根据具体的要求实施不相同的数据采集，而且通过局部网络采集的数据也可以是不同的。

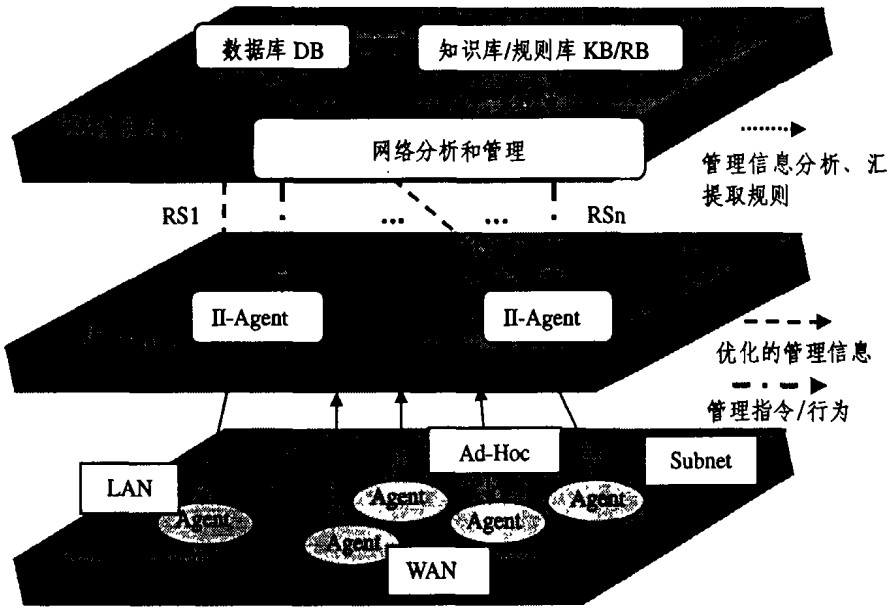

图3 基于 II-Agent 的 NGN 管理系统结构

3)II-Agent 具有相同的执行机制，即基于 II-Agent 的方法在网络中的组织结构、控制结构都是一致的，便于系统的扩展和调整。

系统的非对称性和一致的执行机制，针对 NGN，基于 II-Agent 的网络管理系统能够对 NGN 中所融合的不同网络系统进行统一管理、具体决策，统一的管理者对不同的被管对象使用不同的规则。

同时，II-Agent 结构大大降低了管理流量所占的带宽资源，其开放的接口使得系统具有良好的扩展性。此外，两个接口(管理协议接口 MPI 和管理规则接口 MRI)完成 II-Agent 平台与管理者和设备间的通信。MPI 和 MRI 均采用 M/A 结构，前者仅仅限于网络数据的采集及接口命令的交互；而后者提供给管理者与 II-Agent 之间的交互方式不同，使得管理行为和要求更加优化和智能化，即满足了复杂的 NGN 对网络管理的要求及对带宽资源的极大节省，避免了现行网络管理中，管理信息占据上层网络宝贵带宽的现象。

按照 II-Agent 的网络管理系统结构，下一步的工作将是具体构建完备的规则集，并加以描述及实现使其得以在 NGN 的网络管理系统中得到应用。

参 考 文 献

1 陶洋. 网络性能提升技术分析:[中科院博士后出站报告]. 2001.1
2 Case J D,et al. Simple Network Management Protocol. May 1990
3 Tennenhouse D L, Wetherall J. Towards an active network architecture. 2002
4 Gali A, Covaci S. Active Virtual Private Network Services on Demand. In:5th IFIP TC6 Intl. Symposium Interworking 2000 on next generation networks, Bergen, Norway, 2000
5 Kind A. Agent-Based Management. In: 7th Intl. Conf. on Intelligence in Services and networks, IS&N2000 Athens, Greece,2000