几种常用无损数据压缩算法研究

郑翠芳
(中国工程物理研究院 计算机应用研究所,四川 绵阳 621900)

摘 要:随着网络承载的信息量的飞速增长,数据压缩必然会备受人们重视。数据压缩可分成两种类型,一种叫做无损压缩,另一种叫做有损压缩。文中主要介绍目前用得最多和技术最成熟的无损数据压缩技术,按照无损压缩方法采用的压缩技术的不同,从基于统计的压缩思想和基于字典的压缩思想两个方面对其中最具有代表性的无损数据压缩方法进行了详细的分类讨论和优缺点比较,并对基于字典压缩算法的一些成熟的改进算法进行了汇总介绍,便于对无损数据压缩技术感兴趣的同志学习参考。

关键词:数据压缩;无损数据压缩;统计压缩算法;字典压缩算法
中图分类号:TP311.5 **文献标识码**:A **文章编号**:1673-629X(2011)09-0073-04

Research of Several Common Lossless Data Compression Algorithms

ZHENG Cui-fang
(Institute of Computer Application, China Academy of Engineering Physics, Mianyang 621900, China)

Abstract: With the quick increase of network's information, data compression is paid more and more attention by people. Data compression can be divided into two types, one is called lossless compression, and the other is called loss compression. It takes lossless data compression as main line. According to different compression technology of lossless data compression, from two aspects of statistic and dictionary ideas, it introduces some representative lossless data compression approaches and analyzes these kinds of data compression algorithms' advantages and disadvantages. Gather some mature betterment algorithm based on dictionary compression algorithm together and introduce them; It is facilitate reference for people who is interest in lossless data compression technology.

Key words: data compression; lossless data compression; statistic compression algorithm; dictionary compression algorithm

0 引 言

随着信息化技术的飞速发展,各行各业都用计算机来处理信息,各种系统数据量越来越大,数据在时间和空间上日益增长,给信息存储特别是网络传输带来诸多的困难。为了节省信息的存储空间和提高信息的传输效率,必须对大量的实际数据进行有效的压缩。数据压缩作为解决海量信息存储和传输的支持技术受到人们的极大重视。

压缩算法分为无损压缩和有损压缩。相对于有损压缩来说,无损压缩的占用空间大,压缩比不高,但是它100%地保存了原始信息,没有任何信号丢失并且音质高,不受信号源的影响。而且随着限制无损格式的种种因素逐渐被消除(例如:硬盘容量的急剧增长而且价格越来越低廉),使得无损压缩格式具有广阔的应用前景。文中在查阅大量文献的前提下,对目前国内外的一些具有代表性的无损压缩算法进行详细的分类介绍,并对它们各自的优缺点进行归纳总结。由于篇幅的原因,文中并没有对每个算法的具体实现步骤进行描述,而只是引入了对每个算法介绍比较详细的中文文献信息。希望能为对无损压缩算法有兴趣的同志学习、查询提供方便。

收稿日期:2011-01-27;**修回日期**:2011-05-09
基金项目:中国工程物理研究院预先研究基金(09-0642)
作者简介:郑翠芳(1977-),女,硕士研究生,研究方向为软件开发。

1 无损数据压缩的分类

无损压缩技术即通常所说的通用压缩技术也称为信息保持编码、熵编码、无失真编码等,也就是根据一定方法对大量数据进行编码处理以达到信息压缩存储过程,在数据的压缩过程中不允许精度的损失,被压缩的数据应该能够通过解码恢复到压缩以前的原状态。主要用于文本文件、数据库、程序数据和特殊应用场合的图像数据(如指纹图像、医学图像等)的压缩。这类算法压缩率较低,一般为1/2~1/5。

通常压缩对象是文字或数字等要求精确的数据时,无损压缩是必然的选择。无损压缩从压缩模型上大体可以分为基于统计的压缩算法和基于字典的压缩算法。具体的分类图如图1所示。

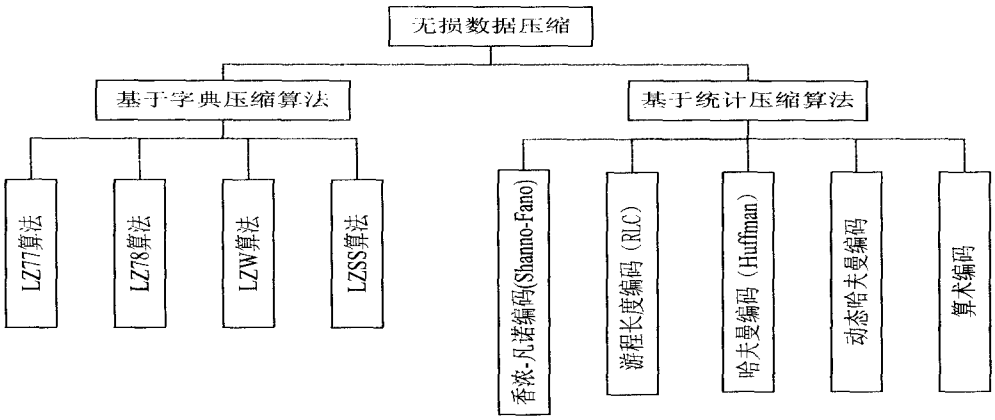

图1 常用无损数据压缩分类图

1.1 基于统计压缩算法

基于统计式压缩算法的起源较早,实质是统计字符的出现频率来对字符本身重新编码,属于熵编码类,与原始数据的排列次序无关而与其出现频率有关,主要的压缩算法有 Shanno-Fano 编码、游程长度编码(RLC)、哈夫曼编码和算术编码。

1.1.1 基于统计压缩算法简单介绍

按照压缩算法的产生时间分别介绍如下:

1)香浓-凡诺算法(Shanno-Fano 编码)由贝尔实验室的 Shannon 和 MIT 的 Robert Fano 开发。首先是 Shannon 在 1948 年给出了一种简单的编码方法—Shannon 编码[1],随后 Fano 在 1952 年又进一步提出了 Fano 编码。Shannon-Fano 编码的核心是构造二叉树,它是一种自顶向下的、非自适应的编码算法。

2)游程长度编码(Run-Length-Coding)[2]是针对一些文本数据的特点所设计的,主要是去除文本中的冗余字符或字节中的冗余位,从而达到减少数据文件所占的存储空间的目的。RLC 的压缩效能取决于整个数据流的重复字符出现次数、平均游程长度及所采用的编码结构。由于该算法是针对文件的某些特点所设计的,所以应用起来具有一定的局限性。为了数据压缩的通用性,一般很少单独采用该方法,主要与其它编码技术配合使用。

3)Huffman 编码[3]是 D. A. Huffman 在 1952 年发现的一种基于信号概率的数据压缩算法。此方法是根据数据信息中字符重复出现的概率生成的一种前缀编码方法。是目前用于压缩的最普遍方法之一。它的核心也是构造二叉树,但它的构造思想刚好与 Shanno-Fano 编码相反,是一种自下向上的、非自适应的编码算法。

4)动态哈夫曼[4,5]又叫自适应哈夫曼(Adaptive Huffman),是 1978 年 Gallager 在 Huffman 编码的基础上提出的一种改进算法。此算法取消了统计过程,一边压缩一边动态调整哈夫曼树,提高了速度,动态 Huffman 的运行时间与输入串长度成线性比,而存储空间的需求量不依赖输入串长度,是一个常数。

5)算术编码[6]也是一种根据字符出现几率的统计结果重新编码的压缩方案,是 Rissanen 在 1976 年提出的,思想和哈夫曼算法类似,是一种高效清除字串冗余的算法,打破了哈夫曼算法必须用整数来表示字符的限制。可以成功地逼近信息熵极限的编码方法。

1.1.2 基于统计压缩算法比较

基于统计的压缩算法各有所长,既有优势的一面,也有不利的一面。表1对每个基于统计的无损压缩算法进行了优缺点以及适用性的比较,方便读者在实际使用中选择合适的压缩算法。

表1 优缺点比较列表

算法名称	优点	缺点	适用范围
游程长度编码	实现简单;压缩和还原速度快	呆板,适应性差;平均压缩率低	复杂度不高的原始点阵图像
Huffman 编码	简单而实用;编码译码时的唯一性	过于受被压缩文件大小的影响;速度较慢	通常用于压缩 GZIP, JPEG 的数据
动态哈夫曼	取消了统计过程,提高了速度	只能去除概率分布不均的冗余	对字符出现频率不均及大量数据的压缩
算术编码	压缩率最高	运算复杂,速度很慢	信源概率比较接近

1.2 基于字典压缩算法

在很长的一段时间内,基于统计的压缩算法占有很重要的地位,直到 1977 年由以色列科学家 Jacab Ziv 和 Abraham Lempel 撰写的两篇论文发表后才出现一种新的压缩算法——基于字典的压缩算法。

1.2.1 基于字典压缩算法简单介绍

字典编码方法是以类似查字典的方式进行编码,它的基本原理是以较长的字符串或经常出现的字母组

合构成字典中的各个词条,并用相对较短的数字或符号来表示的方法。是最为简单的压缩算法之一。压缩效果的好坏和重复数据的出现、字典的大小有关。主要包括:LZ77 算法、LZSS 算法、LZ78 算法、LZW 算法等几种基本算法。

1)LZ77 算法[7,8]是 1977 年两位以色列科学家 Jacob Ziv 和 Abraham Lempel 提出的一种不同其他压缩算法的基于字典的压缩算法,利用该算法进行数据压缩、解压缩的过程,就像一个窗口在原始数据中滑动过程,故也常称为基于滑动窗口的自适应的字典压缩方法。它是基于字符串匹配(或词典编码)的第一个具有实用价值的数据压缩算法。在单片机上实现起来较为理想。它的压缩效果好、速度快,但是压缩率相对较低,另外 LZ77 算法具有明显的空间局限性。

2)LZ78[9,10]压缩算法是 1978 年 Jacob Ziv 和 Abraham Lempel 提出的改进算法,LZ78 算法不同于 LZ77 算法,它放弃了窗口概念,采用树形结构构造字典和保存短语,从而确保文件中的内容均能反映到字典中。它比较适合处理具有一定区间重复性的情况。它继承了 LZ77 算法压缩效果好、速度快的优点,但它的编译码方法较复杂,实现起来比较困难。

3)LZSS 算法[11]是 1982 年 James Storer 和 Thomas Szymanski 为了改进 LZ77 的性能而提出的改进的实用算法,该算法采用二分搜索树,大大加快了压缩速度,解码时无须生成和维护树而更为迅速。该算法的压缩率较高,编译算法较简单。但不足之处是每次压缩都需要向前搜索到原文开头,对于较长的原文(因建立的二叉树过于庞大而降低了编码的效率)需要的时间是不可忍受的,另外无论匹配长度为多少,都分配相同的代码长度,这显然存在一定的冗余。

4)LZW 算法[10,12,13]是 1984 年 Welch 提出的基于 LZ78 算法的一个变种压缩算法。主要用于 GIF 格式的图像数据的压缩。该算法的压缩效果好、速度快且算法描述易于接受,是目前最常用和最有效的无损压缩算法。不足之处是:在压缩过程中给不同的代码字分配固定长度的整数,并且不考虑信息源的概率分布。所需存储空间与输入串长度成正比,并未真正做到最佳地为串表选择串式、最佳地分析输入数据,从而削弱了它的压缩性能。既不适合小文件的压缩也不适合太大文件的压缩,而且 LZW 码仅仅适合内容具有明显单词结构的文件,如. txt 文本文件和 C 语言源程序文件。

1.2.2　改进算法的归纳

字典压缩算法成为目前的主流无损数据压缩算法,得到了人们的极大关注,但在现实使用过程中,由于受到多种因素的限制使得每种压缩算法都存在这样或那样的不足,于是许多研究者就在字典压缩算法的基础上根据实际需要提出了很多的改进方法。所有的改进方法可大致分为以下三类:

1)对字典建立的改进。字典越大,代替的子串越多,但应用中字典容量则受一定限制,要权衡利弊选择合适的字典;

2)对字典更新的改进,一般分为抛弃整个字典或者抛弃字典中匹配率较小的节点两种方式;

3)变换代码长度,由于代码长度决定压缩率,代码长度越短,压缩率越高,为了对大、小文件都取得比较好的压缩效果,可以使用变换长度代码的方法。

表 2[14~20]对国内比较成熟的基于字典压缩算法的改进算法进行了简单罗列及描述。

表 2　基于字典压缩算法的改进算法列表

序号	改进算法	源于算法	描述
1	DTLZW	LZW 的改进	一种基于双串表的改进 LZW 算法,该算法的实质是通过引入双串表的机制来克服 LZ78 算法的缺陷
2	LZSSCH	LZSS 的改进	从编码方案、自适应索引扩大位和最大索引位长等方面修正 LZSS
3	LZWCH	LZW 的改进	从基本码集、更新策略和哈希函数方面修正 LZW
4	LZJH	LZ77+LZ78	此算法是一种基于代码库的自适应数据压缩算法,适用于分组数据网和循环分组业务
5	LZI	LZ78+LZ77	本算法在处理字节和单词重复方面具有良好的性能,对近期和远程数据都很敏感,但是实现起来比较复杂
6	DEFLATE	LZ77+Huffman	采用固定编码表的压缩算法和采用预设计滑动窗口与 Huffman 编码相结合的自适应方式字典编码压缩算法,在 HTTP 压缩中非常流行
7	HLZ	LZ78+LZ77	利用了 LZ78 和 LZ77 的互补特性,具有与基于 LZ78 和 LZ77 相似的计算复杂度和存储复杂度,但具有更好的全局与局部自适应性、更高的压缩效率

2　结束语

文中介绍的几种无损数据压缩算法都比较通用,

每个压缩算法都有自己的优点和缺点，都有自己的适用范围，不同算法的复杂性对空间的要求以及压缩率也不同。它们不仅仅依赖于压缩方法本身，也依赖于被压缩对象的特点。在具体的应用过程中，要根据实际情况的需要，有针对性地去选择、结合或改进一些算法，尽量发挥每一个无损压缩算法的优势，以得到比较理想的压缩数据。不过从综合特性看 LZW 算法是目前在无损压缩方面最常用的方法。但是对 LZW 算法而言其字典的更新速度和较高的压缩率之间的匹配问题还需要进一步的讨论分析。同时研究在保持高压缩比、提高压缩机解压缩速度的同时保持原始数据的完整性还是一个重要的研究课题。

参考文献：

[1] Shannon C E. A mathematical Theory of Communication [J]. The Bell System Technical Journal, 1948, 27(7): 379-423.

[2] 刘 冰. 游程长度编码算法的研究[J]. 天津理工学院学报, 2001, 17(4): 77-81.

[3] 张广学. 最优二叉树的生成及应用[J]. 现代电子技术, 2008, 273(10): 112-114.

[4] 袁 玫, 袁 文. 数据压缩技术及其应用[M]. 北京: 电子工业出版社, 1994.

[5] 游晓明, 陈传波, 刘 升. 数据压缩算法分析与改进[J]. 小型微型计算机系统, 1999, 20(8): 570-573.

[6] Rissanen J, Langdon G G. Universal modeling and coding[J]. IEEE Trans on Information Theory, 1981, 27(1): 12-23.

[7] Ziv J, Lempel A. A Universal Algorithm for Sequential Data Compression[J]. IEEE Transactions on Information Theory, 1977, 23(3): 337-343.

[8] 王忠义, 姜 丹. 关于 Lempel-Ziv77 压缩算法及其实现的研究[J]. 计算机研究与发展, 1996, 33(5): 329-340.

[9] Ziv J, Lempel A. Compression of Individual Sequences via Variable Rate Coding[J]. IEEE Transactions on Information Theory, 1978, 24(5): 530-536.

[10] 王 平. LZW 无损压缩算法的实现与研究[J]. 计算机工程, 2002, 28(7): 98-99.

[11] 王 平, 茅忠明. LZSS 文本压缩算法实现与研究[J]. 计算机工程, 2001, 27(8): 22-24.

[12] Welch. A Technique for High Performance Data Compression [J]. IEEE Computer, 1984, 17(6): 8-19.

[13] 许 霞, 马光思, 鱼 涛. LZW 无损压缩算法的研究与改进[J]. 计算机技术与发展, 2009, 19(4): 125-127.

[14] 吴宇新, 余松煌. 对 LZW 算法的改进及其在图像无损压缩中的应用[J]. 上海交通大学学报, 1998(9): 110-113.

[15] 华 强. 中文文本压缩的 LZSSCH 算法[J]. 中文信息学报, 1998, 12(1): 50-56.

[16] 华 强. 中西文文本压缩的 LZWCH 算法[J]. 计算机工程与应用, 1999, 35(3): 22-23.

[17] 裴文端, 吴 坚. 一种新的数据压缩算法[J]. 无线电通信技术, 2001, 27(3): 30-32.

[18] 卓 越, 杨长生, 宋广华. 一种基于自适应字典的通用无损压缩算法[J]. 计算机工程, 2001, 27(2): 149-151.

[19] 王 刚, 刘立柱. ZIP 文件压缩编码分析[J]. 微计算机信息, 2006, 22(5): 283-285.

[20] 杨长生, 宋广华, 卓 越. HLZ: 一种采用混合字典的自适应无损编码算法[J]. 浙江大学学报(工学版), 2002, 36(1): 40-43.

(上接第 72 页)

(2)对随机光场下采集到的立体像对进行极线校正，提高了本算法匹配的精度与效率。

(3)利用有限视差约束和顺序性约束，同样起到了降低匹配搜索速度与提高匹配精度的作用。

参考文献：

[1] 张素苓, 李竹林, 赵宗涛. 一种立体景象匹配技术及其应用[J]. 计算机技术与发展, 2010, 20(3): 221-224.

[2] Mattoccia S. A locally global approach to stereo correspondence [C]//3D Digital Imaging and Modeling (3DIM 2009). Kyoto, Japan: [s. n.], 2009: 1763-1770.

[3] Lhuillier M, Quan L. Match propagation for image-based modeling and rendering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(8): 1140-1146.

[4] 张令涛, 曲道奎, 徐 方. 一种基于图割的改进立体匹配算法[J]. 机器人, 2010, 32(1): 104-108.

[5] 时洪光, 张凤生, 郑春兰. 基于图像校正与灰度相关性的立体匹配算法研究[J]. 设计与研究, 2010, 37(8): 15-18.

[6] 周秀芝, 文贡坚, 王润生. 自适应窗口快速立体匹配[J]. 计算机学报, 2006, 29(3): 473-479.

[7] Tombari F, Mattoccia S, Stefano L D. Full search-equivalent pattern matching with incremental dissimilarity approximations [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(1): 129-141.

[8] 董 巍, 石光明. 改进的轮廓小波变换及其图像去噪应用研究[D]. 西安: 西安电子科技大学, 2007.

[9] 冯 鹏, 魏 彪, 潘英俊, 等. 一种循环平移的 Contourlet 变换去噪新方法[J]. 计算机仿真, 2006, 23(9): 116-118.

[10] 刘忠艳, 周 波, 车向前. 一种高效的图像匹配算法[J]. 计算机技术与发展, 2009, 19(4): 45-47.

[11] Bradski G, Kaebler A. Learning OpenCV[M]. 北京: 清华大学出版社, 2008.

[12] 陈胜勇, 刘 盛. 基于 OpenCV 的计算机视觉技术实现[M]. 北京: 科学出版社, 2008.