信息技术与系统

基于 Huffman 编码的区域控制器记录数据压缩算法的研究

王福源[1]，雷成健[1]，任建新[2]
（1. 湖南中车时代通信信号有限公司，湖南 长沙 410005；2. 湖南铁路科技职业技术学院，湖南 株洲 412001）

摘 要：基于通信的列车控制（CBTC）系统中区域控制器 (ZC) 子系统为全天候连续工作设备，处于整个系统数据交互的中心，实际工作中，ZC 系统日志最高可产生达每天 10 GB 的数据量，给存储和转储工作带来较大压力。为此，文章基于 ZC 记录数据的特点，提出了一种专门针对此类数据的压缩算法，通过数据压缩以减少数据存储空间。研究结果表明，该压缩算法对 ZC 记录数据的压缩率在 30% 左右，有效提高了系统存储能力，减轻了数据转储的工作量。

关键词：列车运行控制系统；区域控制器；Huffman；数据压缩
中图分类号：U284.55 **文献标识码**：A **文章编号**：2096–5427(2020)03–0089–04
doi:10.13889/j.issn.2096–5427.2020.03.018

Research on Compression Algorithm of Zone Controller Record Data Based on Huffman Coding

WANG Fuyuan[1], LEI Chengjian[1], REN Jianxin[2]
(1.Hunan CRRC Times Signal & Communication Co., Ltd., Changsha, Hunan 410005, China;
2.Hunan Vocational College of Railway Technology, Zhuzhou, Hunan 412001, China)

Abstract: Zone controller(ZC) subsystem in the communication based train control(CBTC) system is a full-time continuously working equipment, which is in the center of data interaction of the whole system. ZC system log can generate up to 10 GB data per day in operation that brings great pressure on storage and dumping. In this case, according to the characteristics of ZC record data, a software compression algorithm was proposed to compress the stored data, which aimed to reduce the space of data storage. The results show that the compression algorithm can compress data by 30%. Storage capacity of the system is effectively improved, and the workload of data dumping is also reduced.

Keywords: CBTC system; zone controller; Huffman; data compression

0 引言

区域控制器（zone controller， ZC）系统是负责城轨列车安全运行的关键信号设备之一。在城市轨道交通信号系统的应用中，ZC 系统根据列车所汇报的位置信息以及联锁所排列的进路和轨旁设备提供的占用 / 空闲信息，实时地为通信列车计算移动授权，从而确保列车之间的安全追踪，是基于通信的列车控制（communication based train control，CBTC）系统中列车自动保护（automatic train protection, ATP）的轨旁部分[1]。ZC 系统为全天候连续工作设备，工作期间实时、不间断地发送和接收各种数据信息并计算存储，其承担的信息量大、数据多[2]。在实际工作中，ZC 系统日志最高可产生的数据量达每天 10 GB，对存储和转储工作带来较大压力。此前，针对此类数据压缩处理方法的研究较少，为此，本文提出了一种软件压缩算法对其进行

收稿日期：2019–10–29
作者简介：王福源（1990—），男，工程师，主要从事城市轨道交通全自动运行系统设计、ZC/DMS 子系统研发工作。
基金项目：湖南创新型省份建设专项（2019GK4015）

处理[3]。

1 数据压缩

数据压缩是一个减小数据存储空间的过程，是信息论的最重要成果之一，其利用数学工具采用多种方法来管理和处理信息[4]。按照压缩精度划分，数据压缩一般有无损压缩和有损压缩两种。在有损压缩算法中，可以接受一定的损失，用以换取更大的压缩比。在某些应用中，如图像处理和音频处理，一定的损失是可以接受的，因为这种损失会受到严格控制，不会影响播放效果。而 ZC 系统日志数据需采用无损压缩，以保证解压缩时准确地还原原始数据。无损压缩数据算法，典型的有算术编码、RLE 算法、LZ 算法和 Huffman 算法。其中算术编码运算复杂、速度慢，主要用于图像处理；LZ 系列算法需要构建索引字符串字典，字典通常会很大；RLE 算法主要是对数据块进行压缩运算；而 Huffman 编码是通过构造 Huffman 树来进行编码的，适用于文件处理，特别是对字符数据的编码。ZC 日志数据具有标志位多、状态位多及重复率高[5-6]的特点，故本研究选用针对数据文件处理的基于 Huffman 编码的无损压缩算法。

1.1 Huffman 编码

Huffman 编码是基于最小冗余编码的数据压缩算法。最小冗余编码指的是，如果能统计出一组数据中所有符号的出现频率，就用一种特定的方法来表示符号，以减少存储数据所需的空间。对出现频率高的符号，编码使用尽可能少的位来表示；对出现频率低的符号，编码使用尽可能多的位来表示[7]。在数据中，所表示的符号不一定要占用一个字节，经过重新编码，可以是任意大小的数据[8]。

1.2 熵和最小冗余

要使用 Huffman 编码，需了解熵的概念。熵的作用是量化数据信息。信息的概念很抽象，此前人们无法精确指出信息量到底有多少，直到 1948 年，数学家香农定义了“信息熵”，信息的量化问题因此得以解决。他首次利用数学语言阐明了概率与信息冗余度的关系：任何数据信息都有冗余，冗余的大小与数据信息中每个符号出现的概率有关[9]。

任何数据集合的信息量都是一定的，即所谓熵。对于一组数据而言，熵就是这组数据中每个符号熵的总和。具体而言，熵有如下定义：

$s_z=-\lg P_z$

式中：P_Z——符号 z 在数据集中出现的概率。

如果确切知道 z 出现了多少次，就知道 z 出现的频率[10]。例如，如果一个 40 个符号的数据集合中 z 出现了 10 次，即出现的概率为 1/4，于是 z 的熵 $s_z=-\lg(1/4)=2$ 位。由此可知，如果在数据集中用来表示符号 z 使用了超过两位的二进制数，这将是一种浪费。目前一般都用一个字节来表示一个符号，在这种情况下使用合适的数据压缩算法可以很大程度地减小数据的容量，即通过重新编码，可以实现使用较少的位对出现频率高的符号编码，用较多的位对出现频率低的符号编码。

2 压缩算法的实现

2.1 数据压缩过程解析

数据压缩包括两个过程：一是压缩或编码数据，使数据容量减小；二是解压缩或解码数据，使数据还原到本身的状态[11]。

（1）压缩部分

压缩部分采取单输入单输出结构，输入为需要压缩的数据地址，输出为压缩后的数据地址。压缩部分按照字节长度对数据进行压缩，压缩字符数最大为 256 个，其按照每包最大 1 400 个字节进行压缩，且能持续压缩，直至目标文件全部执行完毕。

（2）解压缩部分

解压缩部分采取单输入单输出结构，输入为需要解压缩的数据地址，输出为解压缩后的数据地址。解压缩部分按照字节长度对数据进行解压缩，解压缩字符数最大为 256 个，其按照每包最大 1 400 个字节进行解压缩，且能持续解压缩，直至目标文件全部执行完毕。

2.2 算法的具体实现

ZC 日志记录软件中的 Huffman 压缩算法采用分包压缩和循环压缩的方法。将数据包按字节输入，每包数据最大有 256 个字符；压缩后重新按字节组成数据包输出，具体的 Huffman 数据压缩软件流程如图 1 所示[12-13]。

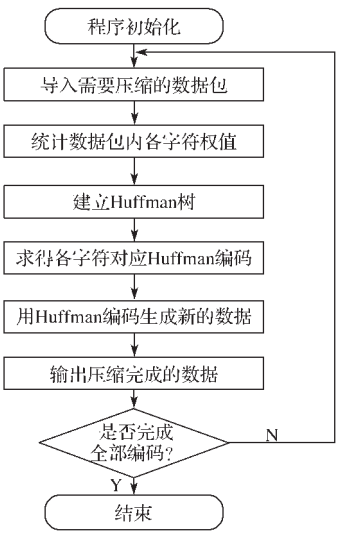

图 1 Huffman 数据压缩软件流程

Fig. 1 Flowchart of Huffman data compression software

解压缩算法同样采用分包解压缩和循环解压缩的方

法。将需要解压缩的数据包按字节输入，组成新的字符串后根据 Huffman 编码解压缩；解压缩完成后依然按字节组成数据包输出[14]。具体的 Huffman 数据解压缩软件流程如图 2 所示。

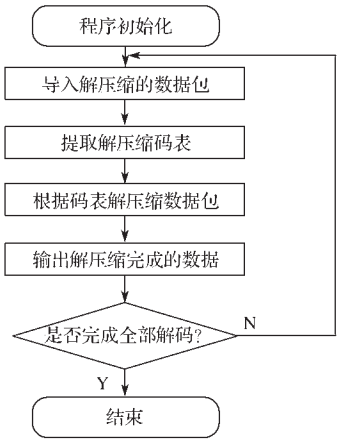

图 2　Huffman 数据解压缩软件流程

Fig. 2　Flowchart of Huffman data decompression software

ZC 日志记录软件中的 Huffman 压缩算法部分代码截图如图 3 所示。压缩过程中，对相关模块采用必要的模块化封装，以保证该软件算法的独立性。

```
class Huffman
{
public:
    Huffman();
    ~Huffman();
    void EnCode(Msg *inMsg,Msg *outMsg,Clist codelist[]);
    void DeCode(Msg *inMsg,Msg *outMsg,Clist codelist[]);
    void SaveText(QString filename);
    void ReadCode(QString filename);
    void SaveCode(QString filename);
    void TextCharsWeight(Msg *inMsg);  //统计原文中各字符权值
    void BuildCharMap();
    QString code;//编码
    QString text;//原文

private:
    HuffNode huffnode;  //哈弗曼树
    CharMap chars;//字符表
    int n;//字符数
    quint16 tLength;
    quint8 buff[C_MAX_BUFF_LEN];
    int Convert(CharMap inChar);
};
#endif // HUFFMAN_H
```

图 3　部分代码截图

Fig. 3　Partial code screenshot

3　应用结果及分析

根据 Huffman 压缩算法的特性可知，被压缩数据包内字符数越少，字符重复率越高，压缩率越高[15]。软件编写完成之后，进行了多次调试与实验，以验证软件可用性和实际的压缩效果。下面设定 3 组实验进行验证。

（1）实验一

数据集中字符的频率和种类都能对数据压缩效果产生影响。本实验是为了测试字符频率的变化对 Huffman 压缩算法的影响。首先，设置一个包含 1 400 个字节的数据包，设定其中各字符权值相等；其次，从 2 个字符到 256 个字符依次实验，将数据包压缩之后再进行解压缩；确认压缩数据无误，将压缩后数据包字节数记录下来并统计压缩前后数据字节变化。实验结果如图 4 所示。可以看出，在各字符权值相等的情况下，压缩变比随着字符数增多而趋于减小，说明在各字符权值相同的情况下，Huffman 编码依然能够对数据进行有效的压缩。根据 Huffman 编码理论，字符重复率越高，Huffman 编码压缩效果越好。

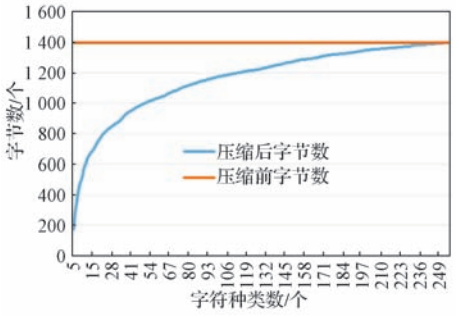

图 4　实验一压缩变比图

Fig. 4　Compression transformation diagram of the experiment 1

（2）实验二

为了测试字符种类的变化对 Huffman 压缩算法的影响，设置一个包含 1 400 个字节的数据包，其恒定含有 256 个字符 (0x00~0xFF)。设定每个字符重复率一样，进行压缩、解压缩实验；然后设定其重复率按比例上升，共进行 5 次实验，分别将数据包压缩之后再进行解压缩；确认压缩数据无误，再次将压缩后数据包字节数记录下来。其压缩变比如图 5 所示。可以发现，随着数据重复率比例的升高，Huffman 算法压缩后数据包越小，压缩效果越好；而实际 ZC 日志记录数据中，按照通信协议，有大量的重复数据，如 0x00/0x01/0xFF/ 0x55/0xAA 等，这说明 Huffman 压缩算法在本应用中压缩效果是确实可行的，且数据重复率越高、字符数越少，压缩的效果越好。

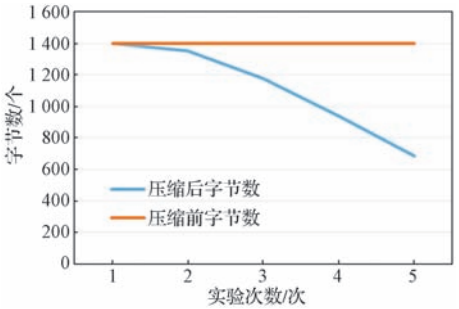

图 5　实验二压缩变比图

Fig. 5　Compression transformation diagram of the experiment 2

（3）实验三

经过多次验证分析之后，按照实际的 ZC 日志记录数据来进行分析。现以 2019 年 9 月某日长沙地铁 4 号线路的 ZC 日志记录为例，从中连续取 100 组 1 400 个字节的数据，经过 Huffman 编码压缩后，根据码表进行解压缩，并验证数据无误。该 100 组数据解压缩后的效果如图 6 所示。可以看出，每一组 1 400 个字节的数据经过压缩后的数据量在 200 到 600 个字节之间，充分证

明了基于 Huffman 编码的数据压缩算法对 ZC 日志记录数据的压缩是可行的。

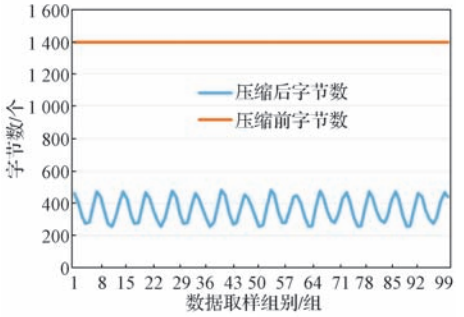

图 6 实验三压缩变比图

Fig. 6 Compression transformation diagram of the experiment 3

4 结语

随着城市轨道交通自动化程度的日益提高，不管是目前普遍运行的 GOA2 等级的线路，还是今后越来越多的 GOA3 等级或者 GOA4 级别线路，数据交互现象会越来越多，数据量会越来越大，数据压缩技术的研究也势在必行。本文针对实际 ZC 日志的数据压缩问题，提出了一种基于 Huffman 编码的数据压缩算法。从实验结果来看，其压缩效果好、效率高。目前该算法已被应用于实际的产品中。

当然，信息论领域有很多算法可以充分利用，而 Huffman 压缩算法是其中一种。今后将考虑对 ZC 日志纪录进行多重压缩，即在 Huffman 压缩算法的基础上增加其他的压缩算法，以更大程度地提高压缩率。另外，轨道交通产品正在向网络化、远程可视化方向发展，其数据的传输受到了网络带宽限制。本文所提供的数据压缩存储的思路，可以减轻网络传输的压力，为轨道交通中大数据的网络化应用提供参考。

参考文献：

[1] 王卓然 , 贾学祥 . 我国城市轨道交通信号系统的发展方向 [J]. 交通世界 , 2019(12):158-159.
[2] 路向阳 , 吕浩炯 , 廖云 , 等 . 城市轨道交通全自动驾驶系统关键装备技术综述 [J]. 机车电传动 , 2018(2):1-6.
[3] 魏东冬 , 卢佩玲 , 郑长宗 , 等 . 基于互联互通的区域控制器安全通信计算机设计 [J]. 都市快轨交通 , 2017, 30(4):55-59,64.
[4] 刘粤 . 面向太阳全日面磁场图像的无损压缩算法及关键技术研究 [D]. 北京 : 北京交通大学 , 2018.
[5] 任颖 , 吕浩炯 , 宋瑞霞 , 等 .CBTC 系统中联锁与区域控制器的一体化设计 [J]. 机车电传动 , 2015(6):49-52.
[6] 李容 . 基于 SCADE 的 CBTC 区域控制器建模与验证 [D]. 成都 : 西南交通大学 ,2015.
[7] 张振 , 甄成刚 . 对数据压缩与解压技术的分析与研究 [J]. 信息系统工程 , 2019(7):152-153.
[8] 施鹏 , 李敏 , 于涛 , 等 . 基于 Huffman 编码的 XML 数据压缩方法 [J]. 北京化工大学学报 (自然科学版), 2013, 40(4):120-124.
[9] 李伟生 , 李域 , 王涛 . 一种不用建造 Huffman 树的高效 Huffman 编码算法 [J]. 中国图象图形学报 , 2005, 10(3):382-387.
[10] LOUDON K. 算法精解 :C 语言描述 [M] . 北京 : 机械工业出版社 , 2012.
[11] 刘海峰 , 刘澄澄 . 基于 VC++ 的无损压缩技术实现 [J]. 网络安全技术与应用 , 2019(6):38-41.
[12] 王防修 , 刘春红 . 一种哈夫曼编码的改进算法 [J]. 武汉轻工大学学报 , 2016, 35(1):88-91.
[13] 王晨曦 . 面向神经网络的无损压缩技术研究 [D]. 南京 : 南京大学 , 2019.
[14] 许子明 . 哈夫曼编码译码功能的简单实现 [J]. 科技风 , 2018(18):7.
[15] 苑思明 , 郑晗 , 李俊杰 . 基于哈夫曼树压缩的加密技术 [J]. 信息记录材料 , 2018, 19(6):57-58.

（上接第 76 页）并通过仿真验证了 DC-DC 变流器对储能电池组的充 – 放电双向运行控制的可行性以及 *SOC-I* 下垂控制的有效性。由于内外双下垂控制方法具有普遍适用性，该控制策略可推广应用至微电网等领域的多变流器协同控制。

参考文献：

[1] 徐龙堂，董晓妮 . 船舶共直流母线混合电力推进系统技术探讨 [J]. 渔业现代化 , 2017, 44(3):70-76.
[2] 杨光 , 牟照欣 , 吴迪 , 等 . 船舶直流组网电力推进技术发展优势 [J]. 舰船科学技术 , 2017, 39(7): 8-14.
[3] 庄绪州 , 张勤进 , 刘彦呈 . 基于负阻抗特性补偿的船舶 DC/DC 变换器控制策略 [J]. 电力系统及其自动化学报 , 2019, 31(5): 28-32.
[4] 庄伟 , 孙坚 , 王春杰 , 等 . 超级电容储能装置在混合动力型直流电推系统中应用与实践 [J]. 船电技术 , 2018, 38(7): 16-20.
[5] KANELLOS F D, TSEKOURAS G J, PROUSALIDIS J. Onboard DC grid employing smart grid technology: challenges, state of the art and future[J]. IET Proceedings, 2015(5):1-11.
[6] 杨惠 , 骆姗 , 孙向东 , 等 . 光伏储能双向 DC-DC 变换器的自抗扰控制方法研究 [J]. 太阳能学报 , 2018, 39(5): 1342-1350.
[7] 朱艳萍 , 阚志忠 , 梁梦娜 , 等 . 双极性 DC/DC 变换器及分布式储能功率控制 [J]. 电力电子技术 , 2019, 53(12): 39-42.
[8] 马中静 , 姜文材 . 一种应用于直流微电网的改进型下垂控制策略 [J]. 电力电子技术 , 2015, 49(9): 16-18.