电力线窄带通信报文压缩算法研究

刘 萌[1],丁香乾[2],侯军伟[1],王 锐[1]
(1.中国海洋大学 计算机科学系,山东 青岛 266100;
2.中国海洋大学 信息工程中心,山东 青岛 266071)

摘 要:电力线载波通信报文传输易受噪声干扰从而影响传输效率,扩频技术的采用避免了信号的衰减但同时也降低了通信速率,因此提出通过对报文数据的压缩来提高通信速率、降低误码率。然而由于电力线载波通信报文短小,不易从统计模型角度进行压缩,因此探索了通用的顺序压缩算法LZ77在电力线载波通信报文压缩中的应用。

关键词:电力线通信;数据压缩;LZ77

中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2010)16-0065-04

Compression algorithms of packets in narrowband power line communication

LIU Meng[1], DING Xiang Qian[2], HOU Jun Wei[1], WANG Rui[1]
(1.Department of Computer Science, Ocean University of China, Qingdao 266100, China;
2.Center of Information Engineering, Ocean University of China, Qingdao 266071, China)

Abstract: Power Line Communication (PLC) packet transmission is vulnerable by the noise that pervades everywhere on power line. The use of spread spectrum technology can avoid signal attenuation but at the same time it reduces the transmission rate. Compressing the packet before transmitting is a way to increase the transmission rate and to lower the error rate. However, as PLC packets message is short, it is difficult to be compressed from the perspective of the statistical model. In this paper explore the application of a universal compression algorithm—LZ77 in PLC packet compression.

Key words: power line communication; data compression; LZ77

电力线载波通信是利用已有的电力线路进行数据传输的一种通信方式,无需专门架设通信基础设施并且具有相当广泛的网络分布,因此,电力线载波通信是一种非常经济的通信方式。然而,由于电力线载波通信存在着一些技术难题,如传输信道间歇噪声大、阻抗随负载变化大、信号衰减大等问题[1],使得目前电力线载波通信仅在自动抄表系统(AMRS)的应用上得到了比较好的发展。

自动抄表系统要求能够稳定准确地抄到每个表,然而由于电力网络的分布电容、分布电感、负载性质、负载阻抗值、噪声等都是动态的,而非恒定的,然而一个设计定型的系统产品,其调制/解调制式、工作频率、发送功率、信道参数、通信效果等通常都是不变的,这就导致了抄表系统不能保证在各种环境下都可以可靠地运行,因此产生了一系列技术难题[2]。

使用扩频通信技术来避免电力线的强干扰、强衰减等缺陷,然而扩频导致了通信速率的大大降低,这使得窄带通信的传输速率只是宽带的几百分之一,这在一定程度上限制了扩频通信的广泛应用[3]。

将数据压缩技术引入到电力抄表系统可以提高通信速率、降低误码率,从而使电力线载波抄表系统更加稳定。

1 数据压缩原理

数据压缩实际上也是一种编码,如果压缩是有效的,那么编码后的数据比原始数据占用的存储空间小。数据压缩根据信息论的基本概念分为无损压缩和有损压缩,本文讨论的是无损压缩。数据之所以能被压缩是因为它存在某种规律或者结构,从信息论角度来看就是

数据中存在冗余信息，而数据压缩就是要去除数据中的冗余信息。

关于数据压缩有很多算法，针对不同特点的数据选择不同的压缩算法从而达到最优的压缩效果。LZ77 是一种通用的顺序数据压缩算法，它不需要知道数据本身的一些特性，对于任何数据都可以进行压缩[4]，思路简单，自从 J.Ziv 和 A.Lempel 于 1977 年提出该算法之后很快得到了广泛应用。

1.1 通用压缩算法 LZ77

LZ77 通过引入滑动窗口（sliding-window），在字符流上顺序滑动 sliding-window，从而实现字符流的压缩。以图 1 中数据为例，LZ77 算法将从左至右滑动 sliding-window 对其进行压缩表示，sliding-window 分为两个部分：search buffer（搜索缓冲区，大小为 7，编号从 0 开始）和 look-ahead buffer（向前查找缓冲区，大小为 5），A="abcbbacde"是滑动窗口滑过的字符串，B="bbadeaa…"是等待被压缩的字符串。当前即将被压缩的位置为 B 中的第一个字符—b，算法将在 search buffer 里面搜索 B 中从该位置向后的最长匹配并将其用一个三元组（position，length，next symbol）简略表示（其中 position 表示被压缩字符串在 search buffer 里面最长匹配的起始位置，length 表示该匹配的长度，next symbol 表示 look-ahead buffer 中第一个不被匹配的字符）。那么，当前从 b 开始的"bbad"将被压缩表示为（1，3，d），然后滑动窗口向右滑过 4 个字符，下一次压缩从 e 处开始，如图 2 所示。

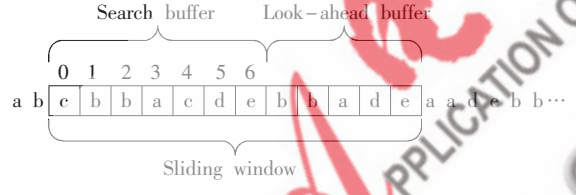

图 1 滑动窗口

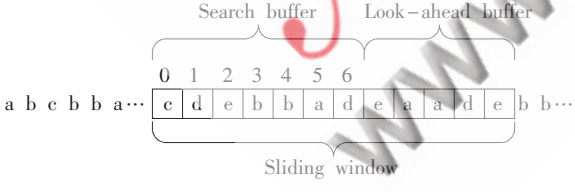

图 2 滑动窗口滑过 4 个字符

该算法简单易行，有较高的执行效率，然而很容易发现它存在的一些问题。于是接下来的几年里出现了很多 LZ77 的改进算法，如不限制窗口长度的 LZR 算法、引入 Huffman 编码的 LZH 算法以及改进 search buffer 数据结构和三元组表示的 LZSS 算法[5]等。

1.2 电力线通信报文压缩算法设计

针对电力线通信报文 W，以下压缩方案来自于 LZSS 压缩算法思想：

(1)首先将报文 W 从左至右，每 7 bit 组成一个字节，字节的最高位置 0，低 7 位来自 W；

(2)如果 $W[i\cdots j]=W[k\cdots l]$，$i<k$，则以两个字节 B_0B_1 替换 $W[k\cdots l]$。

其中：$B_1=i$；B_0 的最高位置 1，其余 7 位为二进制的 $j-i+1$ 数值表示。

例如，对于图 3(a)中的原始报文，附加标志位进行重组后形成报文如图 3(b)所示，进行压缩后的报文如图 3(c)，其中，B0 的最高位为 1 表示接下来两个字节代表了一个压缩表示，B0 的后 7 位等于 3 代表压缩了 3 个字节，$B_1=0$ 代表压缩匹配位置是从第 0 位开始的。

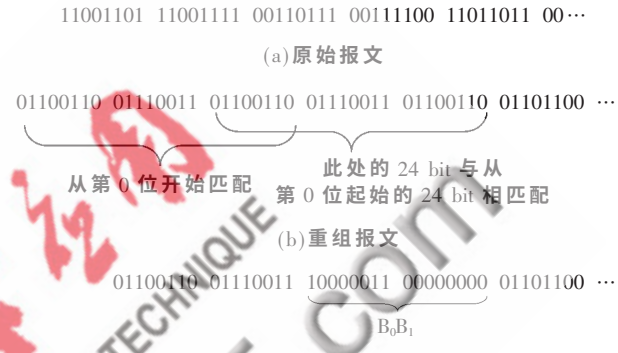

图 3 压缩实例 1

以上算法打破了字符串结构，在每个字节内附加一个标志位（flag）来标识该字节是否被压缩表示，这样大大降低了单个字符也用三元组表示而造成的浪费。

2 算法改进

利用电力线通信报文低速率、短报文（长度不超过 256 B）的特点，可以充分挖掘某种压缩算法（LZ77）的潜力。

2.1 压缩粒度

针对短报文，如果想尽可能地挖掘它的结构模式就要在更小的粒度级别上进行压缩。由于比特串只有 0 和 1 组成，重复串不限于起始结尾于字节，其更有可能出现重复的模式，因此相比较字节级别在位级的压缩应该更有效。如图 4 所示，字节级表示的字符流 B_1 和 B_2 中重复子串为"bb"，而在比特表示的字符流 b_1 和 b_2 中重复子串的长度达到 33 bit，超过了 2 B，扩展了可压缩的范围。

2.2 改进数据结构

由于电力线通信报文长度短，可压缩空间较小，以

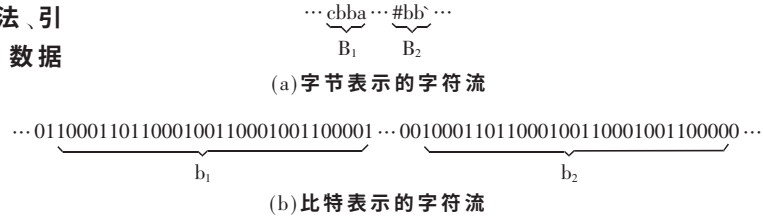

图 4 字节级压缩与位级压缩

上压缩算法可能会造成压缩后的报文比压缩前更长。对于大小为 n 个字节的报文来说，不管压缩与否，首先要附加 $n/7$ 个字节来标识每个字符是否压缩，因此，只有能够压缩大于 $n/7$ 个字节才能对报文进行压缩，否则该压缩将没有意义。

在此，再次改进压缩报文的数据结构来降低这种额外开销，使得对于未被压缩的字节不增加额外信息。为了实现这一目标，就要将标识位所表达的信息集中表示，这样在压缩后的报文开头用一个字节用来表示压缩信息：用一个字节表示压缩表示计数 c（0~255），用来表达该报文一共被压缩了几处（最多可压缩 255 处）。接下来的信息为压缩后的报文信息：压缩报文块 $k(i_k,j_k,l_k)$，其中 i_k：压缩位置；j_k：原始报文位置；l_k：匹配长度。每个压缩表示用 4 B 存储，其中前 11 位表示当前压缩位置 i_k，中间 11 位表示匹配原始报文位置 j_k，后 10 位表示匹配长度 l_k。再接下来的信息为不能进行压缩表示的余留报文数据，对于最后不足一个字节的报文补零凑够整字节。改进的数据结构如表 1 所示。

表 1 改进的数据结构

压缩信息	压缩表示				余留报文
压缩信息	压缩表示 1	压缩表示 2	…	压缩表示 k	未被压缩的报文
1 B(c)	4 B(i,j,l)	4 B	…	4 B	11010101011100011…

注：(1)压缩信息用一个字节表示压缩表示计数 c

(2)压缩表示 $k(i_k,j_k,l_k)$：用 4 个字节存储，其中前 11 位表示当前压缩位置 i_k，中间 11 位表示匹配原始报文位置 j_k，后 10 位表示匹配长度 l_k。

例如，对于图 5(a)中的原始报文数据，进行两处压缩后如图 5(b)所示（为了方便说明改进的数据结构，本例中使用的原始报文按照以上编码方式编码后形成的压缩后报文比原始报文更长，而在实际压缩算法设计时不会出现这种情况）。

改进后的数据结构与之前的数据结构在算法性能上的比较如表 2 所示。

对于第一种数据结构的额外开销压缩与未压缩报文均摊从而造成不必要的浪费，第二种数据结构的额外开销虽然均用在了被压缩的报文上没有造成浪费，但是这种结构缩小了能够进行压缩的范围，两种数据结构各有利弊，应根据实际情况权衡选择。

表 2 两种数据结构的比较(假设位级压缩)

	优点	缺点
第一种	重复子串>2 B 即可压缩表示	额外开销压缩与未压缩报文均摊，压缩字节数>$n/7$ B 时才有意义
第二种	额外开销全部用在压缩的报文上	重复子串>4 B 才可压缩表示

2.3 放松算法时间复杂度限制——最优化问题

在电力线通信中，由于报文传输速率低，用于传输的时间远远多于在本地处理报文所需要的时间，同样由于电力线通信报文属于短报文，算法输入规模小。基于以上两种报文特点，也可以放松对算法时间复杂度的限制，充分挖掘 LZ77 算法的潜力，尽量将报文压缩至最短，从而最大程度地提高报文的传输速率。

针对电力线通信短报文压缩，结合 LZ77 算法顺序滑动滑动窗口思想，很容易得出以下压缩思路，即从第一个位置开始对每个位置 i 求其最长重复子串，如果有价值则对其进行压缩表示，否则原样输出。这种思路简单易行，但是不能实现对报文的充分压缩，压缩效果不是最优。如果要挖掘 LZ77 算法在电力线通信报文压缩问题上的极限，那么要采用下面的压缩思路：

第一步：求出每个位置 i 为起点最长重复出现子串；

第二步：选择合适的重复子串压缩表示，以获取最大的压缩比。

其中，第一步的技术实现有两种方式：滑动窗口以及后缀树，滑动窗口结构简单易于实现，但是对于可变长度的滑动窗口来说，这种技术不利于进行重复子串的搜索，其复杂度为 $O(n^2)$。而如果用后缀树技术，在对重复子串的搜索上可以将时间复杂度降低到 $O(n)$，但是后缀树的构造比较复杂，并且将还会增加算法本身的复杂性，尤其在字节级别进行压缩时需要构造的后缀树将非常复杂。

第二步也有两种实现策略：

(1)选择相互独立子串，使压缩比最高，数学模型如下：

给定$[1\cdots n]$上的 m 个区间：$[i_1,j_1]\ [i_2,j_2]\cdots[i_m,j_m]$

求：选 k 个独立子区间：$[h_1,l_1]\ [h_2,l_2]\cdots[h_k,l_k]$

使得$(l_1-h_1+1-c)+(l_2-h_2+1-c)+\cdots+(l_k-h_k+1-c)$达到最大。

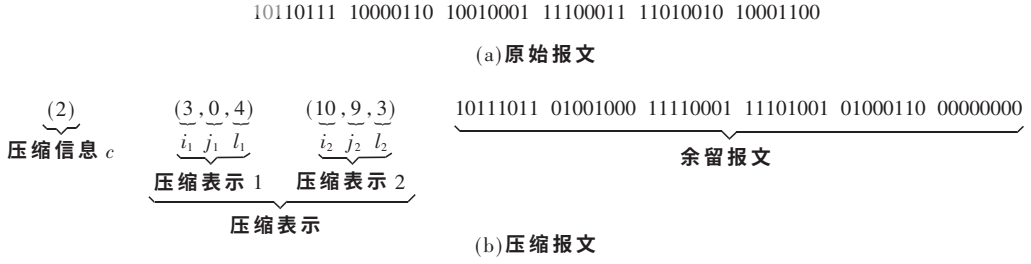

图 5 压缩实例 2

(2)在选择合适的重复子串时不限制子串的独立性，可以考虑对某些子串进行分解，数学模型如下：

给定$[1\cdots n]$上的 m 个区间：$[i_1,j_1]$ $[i_2,j_2]\cdots[i_m,j_m]$

求：选 k 个独立子区间：$[h_1,l_1]$ $[h_2,l_2]\cdots[h_k,l_k]$

满足$[h_i,l_i](1\leqslant i\leqslant k)$是给定某子区间的子集

使得$(l_1-h_1+1-c)+(l_2-h_2+1-c)+\cdots+(l_k-h_k+1-c)$达到最大。

注：c 为压缩表示大小

这种压缩思路打破了 LZ77 顺序压缩的思想，它不是随着滑动窗口的顺序滑动实时地进行数据压缩而是在标记了每个位置起始的最长重复子序列之后在这些子序列中选择一组最优的压缩组合，从而达到最大程度的压缩。同时这种思路的两种不同实现策略对报文的压缩程度也有不同，经过证明第二种策略较第一种策略对报文的压缩程度更大。

3 下一步工作

3.1 证明上述改进算法第二步的第二种策略是否是 NP 完全问题

下面从上述策略中的最优化问题导出如下判定问题：

Instance：$I=\{I_1,I_2,\cdots,I_m\}$为区间$[1\cdots n]$上的 m 个 interval 的集合，常数 k，c；

Question：是否存在$[1\cdots n]$上的独立 interval 集 $J=\{J_1,J_2,\cdots,J_r\}$

满足：

(1) $\forall\ 1\leqslant i\leqslant r$，$\exists\ 1\leqslant j\leqslant m$，$J_i\subseteq I_j$ 且 $J_i>c$

(2) $\sum_{1\leqslant i\leqslant r}|J_i|-rc\geqslant k$

其中：$|J_i|=b-a+1$，若 $J_i=[a,b]$

接下来的任务就是要证明(或否证)以上问题是 NP 完全的。如果它是一个 NP 完全问题，那么我们就要退而求次来寻求解决该问题的近似算法。

3.2 从信息论角度探索数据压缩的极限

从信息论角度探索数据压缩的极限，既然熵是消息包含信息量多少的度量，那么它就可以作为一个度量压缩算法对消息进行压缩的边界或者尺度，用来界定最多可以将消息压缩到什么程度。

参考文献

[1] 杨宗剑，冯娟.低压电力线载波抄表系统现状及发展[J].湖北电力，2008，32(5)：62-63，70.

[2] 林其田.低压电力线载波抄表系统[J].福建建设科技，2006(1)：52-54.

[3] 王学伟，张蕊.电力线载波 DS 扩频通信及数据压缩[J].中国住宅设施，2008(08)：50-53.

[4] ZIV J，LEMPEL A. A universal algorithm for sequential data compression[J]. IEEE Transactions on Information Theory，VOL.IT-23，NO.3，MAY 1977.

[5] NELSON M.数据压缩技术原理与范例[M].贾起东，译.北京，科学出版社，1995.

(收稿日期：2010-04-12)

作者简介：

刘萌，女，1986 年生，硕士研究生，主要研究方向：算法分析与设计。

丁香乾，男，1963 年生，教授，博士生导师，主要研究方向：计算智能、制造业信息化、物流技术。

侯军伟，男，1985 年生，硕士研究生，主要研究方向：网络流媒体。