计算机科学2000Vol. 27№ 11

动态哈夫曼算法在电力线计算机网络数据压缩中的应用*)

Applications of Dynamic Huffman Code Algorithms in the Data Compression of Power-Line Computer Network

黄荣辉 周明天 曾家智
(电子科技大学 成都610054)

Abstract This thesis is to analyze the characteristics of data packets in power-line computer network and to discuss a data compression method of present study in abroad. Briefly describing the different Huffman code algorithms, it presents the data compression results by testingg the data packets in power-line computer network. The result shows that it is better to use the Advanced Dynamic Huffman Code method in power-line computer network. Finally, the methods of improving operation in engineering are proposed.

Keywords Computer, Network, Power-line, Data-compression, Algorithm

数据压缩的方法有许多种，从数据是否能完全恢复来看，包括有损压缩和无损压缩；从压缩方法来看，有行程编码、哈夫曼编码、字典编码、算术编码等。网络数据包的压缩必须是无损压缩。无损压缩主要有行程编码、哈夫曼编码、字典编码等方法。对于电力线介质的计算机网络，目前在国内未见有采用数据压缩方法，国外一些研究也没有采用较好的编码压缩方法。本文将首先阐明电力线网络数据包的特性，然后讨论目前国外研究的电力线计算机通讯采用的压缩方法及其缺点。提出将动态哈夫曼方法用于电力线计算机网络，并用两种哈夫曼方法对电力线计算机网络的数据包进行了压缩实验，表明高级动态哈夫曼方法具有较高的压缩比并适应电力线网络短数据包的特点。

一、电力线介质数据包特性

由于电力线的基本用途是用于电能供给，因此当作为通讯介质时，它与其他专用的通讯介质有着完全不同的信道特性。通讯信道的阻抗、信号衰减和噪声是设计通信信号方式、错误控制码和通讯协议的基础。对于电力线介质数据包来讲，噪声干扰是影响数据包结果的主要因数。电力线上存在着大量的噪声源：交直流电机、可控硅设施、电视机等设备，此外，质量低劣的电力线漏电和放电也产生许多噪声。电力线上的噪声可归纳为如下四种：

(1)与电网频率同步的谐波噪声。这类噪声由电网中各种可控硅开关部件产生。

(2)具有光滑频谱的噪声。这类噪声主要由各种交直流电机产生，这种噪声可看成白噪声。

(3)脉冲噪声。这类噪声主要来源于照明电器、热工仪表及其它可能产生脉冲干扰的开关。

(4)非同步的周期性噪声。这是一种与电网频率成线性非相关的噪声。

研究表明，电力线上的噪声干扰平均时间间隔约8-10毫秒，鉴于电力线介质的衰减特性和阻抗特性，数据传输速率为150Kbps。因此，数据包的最大长度似乎应在1200Bits和1500Bits之间，但是，如果考虑到数据包的启动传输时间并非总是在干扰的边缘，数据包的最大长度将要小得多。显然，在电力线网络中，包长度越短，数据包被干扰而重传的可能性就越小。但包越短，系统的传输效率就越低。数据包过长或过短都会降低传输效率。设噪声干扰间隔时间为m，数据包时间长度为n，网络中各站点启动传输的时间是均匀分布的，则数据包正确传送的概率为(m-n)/m。传送一帧的平均时间为：

$n*(m-n)/m$

不难得出，$n_{MAX}=m/2$。

由此可见，数据包时间长度的最大值与噪声干扰时间间隔有关。若m=8毫秒，则$n_{MAX}=4$毫秒。若数据传输率为150Kbps，则最大数据包长度应不超过614Bits。

在电力线介质的计算机网络中，为提高数据传输效率，也必须采用数据压缩方法。美国一种电力线Mo-

*)本课题得到四川省重点实验室资助。

dem 采用了一种简单的数据压缩方法，其基本压缩过程是这样的：首先是两个字节的压缩标志字节，其后是压缩或没有压缩的实际数据。标志共两字节16位，与后面的数据是一一对应的。如果标志位为0，表示相应的数据字节是未压缩的；如果标志位为1，表示对应的连续两个字节是压缩字节，其中前一个字节表示相对位置，后一个字节表示重复字节数量。这个压缩方法有如下几个缺点：1)它是面向字符的压缩方法，至少要两个字节一样才有压缩效果，2)前后的相关性限制在256字节内，3)压缩比难以提高，这种方法的优点是速度比较快，算法简单。

为克服上述问题，我们对电力线网络数据包使用了一种称之为"动态哈夫曼编码"的压缩方法，这是近年来不断发展起来的一种压缩算法。1985年 D. E Knuth 提出了动态哈夫曼编码，1987年 J. S. Vitter 对其进行了优化的算法设计。1997年，C.-H. Kuo、M.-K. Tsay 和 C.-C. Lu 又对该算法做出了改进，我们称之为"高级动态哈夫曼"算法。

二、动态哈夫曼编码

普通的哈夫曼算法，需先扫描整个数据包，统计各个字符出现的频率，然后按照得到的频率构造哈夫曼树，再根据得到的哈夫曼树对数据中的各字符进行编码。根据信息论理论，对同样信息，这样得到的编码长度最小。我们把它称为"静态哈夫曼"算法。这个算法有几个缺点：①需要对数据进行两次扫描；②需要将哈夫曼树的信息随压缩数据传输，对短数据压缩很不利；③不能解决字符串冗余。动态哈夫曼算法解决了①②两个缺点，而高级动态哈夫曼算法则部分解决了③的缺点。

动态哈夫曼算法的基本思想是，用前面 n-1个字符的频率来编码第 n 个出现的字符。由于第 n 个字符编码之后，第 n 个字符对应的字母频率改变了，必须调整哈夫曼树才能使之符合哈夫曼树的定义，解决这个问题又可从两个方面入手：1)第 n 个字符的字母在哈夫曼树中还没有出现；2)第 n 个字符的字母已出现在哈夫曼树中。

根据哈夫曼树的定义，一个具有 p 个叶子节点的二叉树是哈夫曼树的充要条件是：

(1)每个叶子节点有非负的重量 $w_1,\cdots,w_p$，每个内部节点的重量是其孩子重量之和；

(2)节点按非负的重量的大小顺序来编号，所以对于 $1\leqslant j\leqslant p-1$，节点 $2j-1$ 和 $2j$ 是兄弟，它们的共同双亲具有更高的编号。

节点的编号对应哈夫曼算法中节点合并的顺序，如节点1和节点2首先合并，然后是3和4，再是5和6，等等。

假设已经有 $\eta_t=\{a_1,a_2,\cdots,a_t\}$ 共 t 个字符已经处理，下一个字符 a_{t+1} 将按 η_t 的哈夫曼树编码和解码，现在的主要问题是如何快速修改这棵树，以得到一棵 η_{t+1} 的哈夫曼树，我们先看如何解决问题2)，如图1-a，这时 t=32，a_{t+1}="b"，只是简单地在 a_{t+1} 所在的字母节点及其所有祖先节点的重量加1显然是不够的，因为其结果不是哈夫曼树，将违背上述的充要条件(节点4大于节点5的重量)。

处理办法可以用一个两阶段的过程来描述。首先，将 η_t 的哈夫曼树作如下调整：从 a_{t+1} 所在的字母的节点开始，以此节点作为当前节点，连同子树一起与同重量的最高序号的节点进行交换，然后以此节点的双亲作为当前节点，重复这个过程，直到树根。这样得到的树仍然是哈夫曼树，如图1-b。接着，从 a_{t+1} 所在字母节点开始，对它及其祖先的重量加1，就得到 η_{t+1} 的哈夫曼树。如图1-c。

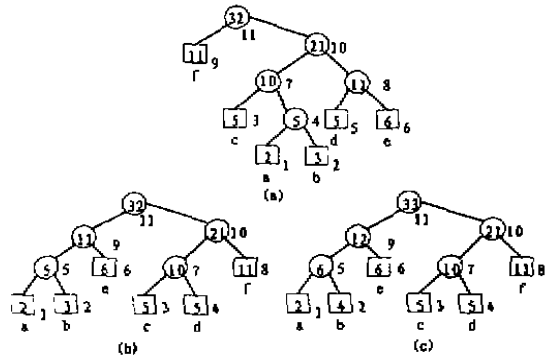

图1　在哈夫曼树中插入一个已存在的节点

对于问题1)，我们用一个0-节点代替树中还没有的字母，如果 a_{t+1} 这个字符的字母还没有出现在哈夫曼树中，我们用0-节点对 a_{t+1} 进行编码，后跟一些多余的字符以区别那些还没有编码的字母。在调整哈夫曼树方面，如图2，先将0-节点作为双亲，再创建两个叶子为它的孩子，其中左孩子继续作为0-节点，右孩子为新的 a_{t+1} 这个字符的字母节点。然后同解决问题2)的方法一样，对整个哈夫曼树进行更新。

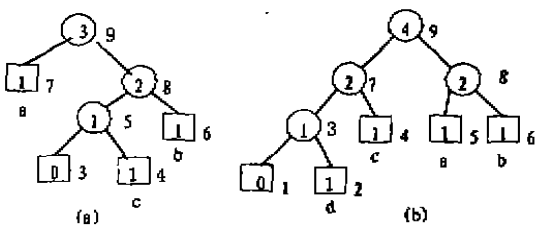

图2　在哈夫曼树中插入一个不存在的节点

下面是更新哈夫曼树的完整算法：

```
procedure Update;
begin
q:=leaf node corresponding to a(i+1);
if(q is the 0-node)and(k<n-1)then
begin
      Replace q by a parent 0-node with two leaf 0-node chil-
          dren.numbered in the order left child.right child.
          parent;
      q:=right child just created
end;
if q is sibling of a 0-node then
begin
      Interchange q with the highest numbered leaf of the
      same weight;
      Increment q's weight by 1;
      q:=parent of 1 q
end;
while q is not the root of the Huffman tree do
begin {Main loop}
      Interchange q with the highest numbered node of the
      same weight;
      {q is now the highest numbered node of its weight}
      Increment q's weight by 1;
      q:=parent of q
    end;
  end;
```

三、高级动态哈夫曼编码

动态哈夫曼编码虽然解决了静态哈夫曼编码的两次扫描和传输哈夫曼树的缺点，但同静态哈夫曼编码一样，并没有对冗余的字符串信息进行处理。对动态哈夫曼编码的一种改进是增加了对重复字符串的编码处理。其基本思想是：增加了一个字符串重复查找器，对每一个欲处理的字符，首先看它是否属于重复字符串中的一个。如果存在一个重复的字符串，称之处于重复模式，则对这个重复的字符串的最近出现的位置和长度编码；否则称其处于正常模式，按正常的动态哈夫曼编码进行。这需要在字母集中增加一个表示重复模式的字母，以其为前缀的编码说明随后的编码是按重复模式进行的编码。问题是如何高效地查找到重复的字符串？按照文[3]所提供的办法，其具体过程是这样的：设输入序列 $x=x_1x_2\cdots x_N$，按照如下规则产生一个输出序列 $y_1y_2\cdots$：

$$y_t=\begin{cases} t;t=1\text{或 }x_{t-1}x_t\text{ 是第一次出现，} \\ \min\{k\mid x_{t-k-1}x_{t-k}=x_{t-1}x_t\};\text{否则。}\end{cases}$$

如果 $y_{t-1}=y_t$，我们就认为 x_t 处于重复模式，否则，则是正常模式。因此，如果 $y_t=y_{t+1}=\cdots=y_{t+\delta}(\delta\geqslant 1)$，并且 $y_{t-1}\neq y_t$，$y_{t+\delta}\neq y_{t+\delta+1}$，则其对应的字符串 $x_tx_{t+1}\cdots x_{t+\delta}$ 形成一个长为 $\delta+1$ 的重复模式。为了增加一次匹配的长度，可设 $a=y_{t-1}$，如果 $y_t<y_{t-1}$，按前面的算法，x_t 将不属于重复模式，然后我们 x_tx_{t-1} 与 $x_{t-a}x_{t-a-1}$，如果 $xt_{a-1}\neq x_{t-a},x_{t-a-1}$，那么 x_t 就不属于重复模式，若 x_{t-1} 是重复模式的话，则重复模式到此结束，否则，我们令 $y_t=a$，即 x_t 处于重复模式，因为 x_{t-1} 和 x_t 有同样的输出值，这就得到了更长的字符串。

四、动态哈夫曼数据压缩方法的压缩结果

为了检查高级动态哈夫曼编码对电力线网络数据包的压缩效率，我们从电力线介质计算机网络上截获了一些实际的数据包来进行测试，测试方法是，对同一个数据包，应用 DHUF（动态哈夫曼）、AHUF（高级动态哈夫曼）两种方法进行压缩，列出压缩的结果大小和压缩比。如表1，从中我们可以发现，高级动态哈夫曼编码具有良好的压缩效果。另一方面也可以看出，网上数据包的可压缩性的变化也很大。

表1

Pkt size	AHUF		DHUF	
(buffer)	14k		6k	
108	64	0.778	104	0.963
106	85	0.802	111	1.047
216	114	0.528	162	0.750
212	155	0.731	174	0.821
308	222	0.721	251	0.815
304	213	0.701	268	0.882
440	243	0.552	298	0.667
436	126	0.289	283	0.649
532	148	0.278	212	0.398
532	118	0.222	170	0.320
640	198	0.309	396	0.619
630	649	1.030	829	1.316
744	201	0.270	307	0.413
748	406	0.543	488	0.652
816	369	0.452	484	0.593
832	180	0.216	480	0.577
932	81	0.087	188	0.202
915	368	0.402	606	0.662
976	682	0.699	911	0.933
976	971	0.995	1157	1.185

由于网络上传输的数据差异很大，可压缩性变化也很大。虽然高级动态哈夫曼编码的压缩效率很好，但也存在负压缩情况。因此有些数据包是不能压缩的，同时，很短的数据包也没有压缩的必要。我们在电力线网络数据包的压缩过程中充分考虑了这些因素，具体方法如下：

(1)小于60字节的数据包不压缩，存在负压缩情况时不压缩。

(2)压缩的数据从电力线网络数据包头后第三个字节开始，即包头后的两个类型字节不属于压缩范围。

(3)区分一个数据包压缩与否的方法：用电力线网络数据包头的源地址域的 Src3和 Src2两个字节作标志，如果 Src3和 Src2的值是零，则后面的数据没有压缩；否则，这两个字节表示压缩前的数据长度，可作为解压时的参考值。

结论 电力线介质计算机网络必须采用无损压缩。国外研究尚未采用较好的压缩方法，国内未见有在电力线计算机网络中采用数据压缩方法。通过分析，电力线介质计算机网络必须采用短包结构，动态哈夫曼压缩方法优于国外目前采用的编码压缩方法，适用于短包结构的电力线介质计算机网络。 （下转第48页）

CorbaScript 解释器相当于 CORBA 系统的一个 Shell，可以用来简化用户对 CORBA 系统的操作，目前，基于 CORBA 标准的分布平台系统一般都有200到300个 IDL 接口，对于一般用户来说，完全掌握是比较困难的。而为 CORBA 系统加上 CorbaScript 解释器，用户就可以象使用一般操作系统一样来对 CORBA 系统进行操作，如输入 mkdir，rm 等简单命令即可，不再需要掌握 CORBA 系统中那些繁琐的对象操作。CorbaScript 解释器与 ORB 的类型无关，因此一个 CorbaScript 解释器可以在任何的 ORB 上运行。CorbaScript 解释器的实现基于 CORBA 的标准组件（Component），其中：解释器用接口池存根（Interface Repository Stub）来访问接口池中 IDL 的信息；通过 ORB 接口用来对对象引用等进行操作；由 Script 命令构造的请求则通过 DII 激活服务对象。

3.2 集成方案的系统结构

系统的总体结构如图2所示，它由以下五部分组成：

·标准的 Web 环境：任何类型的 Web 浏览器都可以作为客户访问 CORBA 系统的界面；

·基于 ISAPI 的 Web 服务器扩展：Web 服务器利用 ISAPI 与 CorbaScript 脚本引擎通信；

·CorbaScript 脚本引擎：由 IIS 服务器安装、激活，对 CorbaScript 脚本文件解释执行；

·标准的 CORBA 环境：脚本引擎通过 DII 激活 CORBA 对象；

·应用服务对象：通过 IIS 服务器为客户提供服务的 CORBA 对象。

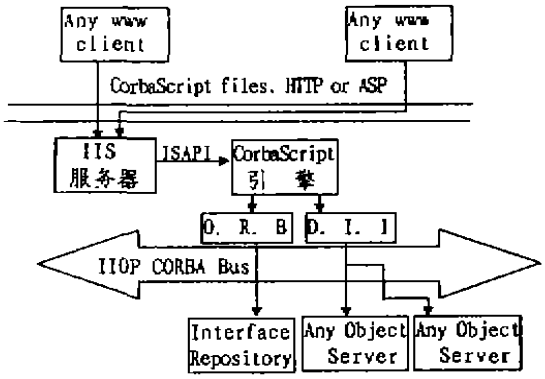

图2 系统结构图

结束语 Starbus 是我们开发的拥有自主版权的基于 CORBA 规范的分布计算平台。我们以此为系统平台，实现了 CorbaScript 解释器，并以此为基础，为 IIS 服务器构造了基于 ISAPI 扩展的 CorbaScript 脚本引擎，实现了 Starbus 与 IIS 服务器的集成。

参 考 文 献

1 王克宏. JAVA 语言 Applet 编程技术. 清华大学出版社，1997

2 朱玉山，等译. ISAPI 实用技术指南. 清华大学出版社，1998

3 CORBA Scription Language，OMG TC Document orbos/99-07-17. August 1999

4 Microsoft Internet Information Sever4.0使用大全. 微软软件使用大全系列丛书，人民邮电出版社，1998

5 Active Server Page 基础教材. Available at：http：//yft.yeah.net

（上接第38页）

参 考 文 献

1 周明天，汪文勇. TCP/IP 网络原理与技术. 清华大学出版社，1993

2 Welch T A. A technique for high-performance data compression. IEEE Comput ，1984，C-17(June)：8～19

3 Kuo C—H，Tsay M-K，Lu C-C. An Efficient Repetition Finder for Improving Dynamic Huffman Coding. IEEE Trans. Commun. ，1997，45(Nov.)：1363～1366

4 Vitter J S. Design and analysis of dynamic Huffman codes. J. Assoc. Comput. Mach. ，1987，34(Oct.)：825～845

5 Morgan H L. Attenuation of Communication Signals on Residential and Commercial Intrabuilding Power-Distribution Circuits. IEEE Transactions on Electromagnetic Compatibility，1986，EMC-28(4)

6 Roger M V. Noise on Residential Power Distribution Circuits. IEEE Transactions on Electromagnetic Compatibility，1984，EMC-28(4)

7 J. B. O'Neal jr. The Residential Power Circuit as a Communication Medium. IEEE Transactions on Consumer Electronics，1986，CE-32(3)

8 Rice B F. A multi-sequence spread spectrum system for powerline communications. In：1996 IEEE 4th Intl. Symposium on Spread Spectrum Techniques and Applications Proceedings. 1996. 809～815

9 Zhang zhengchuan. Lighting power supply line charactristics and performance an FSK discrimination detector. IEEE Teencon'93/Beijing

10 Computer Network 3rd Ed/Andrew S. Tanenbaum 1997