传感器网络中一种基于一元线性回归模型的空时数据压缩算法

王雷春 马传香
(湖北大学数学与计算机科学学院 武汉 430062)

摘 要：针对传感器网络中节点采样数据的空间和时间冗余特点以及节能要求，该文提出了一种基于一元线性回归模型的空时数据压缩算法 ODLRST。ODLRST 先在每个节点内进行消除时间冗余的数据压缩，再在节点汇集处对来自不同节点的数据消除空间冗余以进一步压缩数据。仿真实验证明，ODLRST 能够极大地减少节点发送的数据量和网络中的通信流量，节省并平衡网络中的能量消耗。

关键词：传感器网络；线性回归；空时相关；数据压缩
中图分类号：TP393 **文献标识码**： A **文章编号**：1009-5896(2010)03-0755-04
DOI: 10.3724/SP.J.1146.2009.00704

A One-dimensional Linear Regression Model Based Spatial and Temporal Data Compression Algorithm for Wireless Sensor Networks

Wang Lei-chun Ma Chuan-xiang
(*School of Mathematics & Computer Science*, *Hubei University*, *Wuhan* 430062, *China*)

Abstract: Considering spatial and temporal redundancy of data and demand of saving energy in Wireless Sensor Networks (WSN), a One-Dimensional Linear Regression model based Spatial and Temporal(ODLRST) data compression algorithm, is proposed. By eliminating temporal redundancy of data in single node and spatial redundancy of data among nodes respectively in WSN, ODLRST greatly compresses these data. Simulation results show that ODLRST can reduce data size sent by nodes and network traffic in WSN, and save and balance energy consumption in the network.

Key words: WSN; Linear regression; Spatial and temporal correlation; Data compression

1 引言

在传感器网络中，节点在电池能量、处理能力、存储容量以及通信带宽等方面的资源严重受限；同时，节点密集布置使得采样数据之间存在着极大的空间和时间冗余。因此，在传感器网络的信息收集过程中，采用节点单独传送数据到汇聚节点的方法并不合适。一方面，这将浪费网络中有限的通信带宽和能量；另一方面，它也将降低信息收集的效率。

目前，已经有许多文献对传感器网络中的数据压缩问题进行了研究。刘向宇等[1]提出的 CODST 方法通过采用多项式函数对传输的多个数据点进行近似拟合压缩并传输，并在 Sink 处还原数据，从而达到减少数据传输量和节省网络带宽的目的。Puthenpurayil 等[2]介绍了一种基于计算和通信成本能量预算的数据压缩算法。基于传感器节点的位置信息，文献[3]给出了一种基于区间小波变换的混合熵数据压缩方法。文献[4]给出了一种基于环模型的分布式时-空小波数据压缩算法，同时挖掘传感器网络中数据的时间和空间相关性。基于簇结构，罗武胜等提出了一种基于 LBT 的多节点协同图像压缩算法(MCIC)[5]，即采用低复杂度、高压缩效能的 LBT 图像压缩算法，通过多个中继节点协作，共同完成图像的压缩编码和转发任务。Wang 等[6]提出了一种基于空时编码技术的多率压缩和存储方案以减少数据传输量和减轻数据拥塞。上述工作对传感器网络中数据压缩研究起到了一定的推动作用，但大多数压缩算法比较复杂，对于处理和存储能力有限、能量严重受限的传感器节点来说并不适用。

考虑传感器网络中节点采样数据具有的空间和时间相关特性，以及传感器网络对数据压缩算法的简单、易实现等要求，本文提出了一种基于一元线性回归模型的空时数据压缩算法 ODLRST。ODLRST 采用在单个节点和汇集节点处分别消除节点采样数据时间和空间冗余的方法对网络中传输的数据进行压缩，以减少节点发送的数据量和网络中的通信流量，并延长网络生命。

2 时间序列相关理论

2.1 时间序列定义

传感器网络中的时间序列可定义如下：

2009-05-12 收到，2009-09-25 改回
国家自然科学基金(60603069)资助课题
通信作者：王春雷 wlc2345702@163.com

定义 1 传感器网络中的时间序列传感器网络中节点按时间顺序采样的、具有一定时间间隔的一系列数据的集合，记为 $S=((t_1,d_1),(t_2,d_2),\cdots,(t_n,d_n))$。其中 (t_i,d_i) 表示在 t_i 时刻时间序列的值为 d_i，n 为时间序列的长度，即时间序列中数据的个数。在不引起混淆的情况下，可以简化表示为 $S=(d_1,d_2,\cdots,d_n)$。

2.2 时间序列的拟合回归线

传感器网络中的时间序列可看作是一个以采样时间点 t 为自变量、以时间点采样数据值 d 为因变量的函数。如果时间点的采样数据以线性规律分布在一条直线的周围，则该时间序列可用一条直线段进行拟合。当该直线段是通过一元线性回归模型方法计算出来的，则称为时间序列的一元线性拟合回归线，在本文中简称为时间序列的拟合回归线(如图 1 所示)。

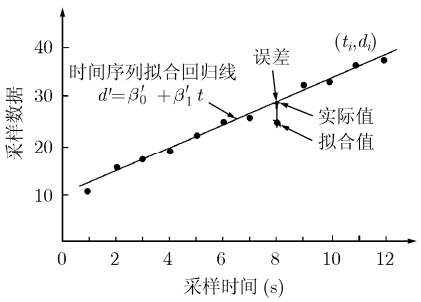

图 1 时间序列拟合回归线

考虑回归函数是 t 的线性函数的情况，因而

$$d=\beta_0+\beta_1 t+\varepsilon,\ \varepsilon\sim(0,\sigma^2) \tag{1}$$

对上述参数进行最小二乘法进行估计，可得 β_0，β_1 的估计值 β_0'，β_1' 分别为

$$\beta_1'=\frac{\left[\sum_{i=1}^{n}t_i y_d-\frac{1}{n}\left(\sum_{i=1}^{n}t_i\right)\left(\sum_{i=1}^{n}d_i\right)\right]}{\left[\sum_{i=1}^{n}t_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}t_i\right)^2\right]},$$

$$\beta_0'=\frac{1}{n}\sum_{i=1}^{n}d_i-\left(\frac{1}{n}\sum_{i=1}^{n}t_i\right)\beta_1'$$

于是，可得回归方程为

$$d=\beta_0'+\beta_1' t \tag{2}$$

2.3 时间序列特征表示

定义 2 时间序列拟合回归线特征设一个时间序列 $S=\{(t_i,d_i)\}(i=1,2,\cdots,n)$，其拟合回归线 L 的特征可表示为：$LF=(S_1,S_2,S_3,S_4)$。其中 S_1,S_2,S_3,S_4 是 L 的 4 个特征系数，$S_1=\sum_{i=1}^{n}t_i$，$S_2=\sum_{i=1}^{n}d_i$，$S_3=\sum_{i=1}^{n}t_i^2$，$S_4=\sum_{i=1}^{n}t_i d_i$，$n$ 为时间序列中数据的个数。

定义 3 时间序列特征设一个时间序列 $S=\{(t_i,d_i)\}(i=1,2,\cdots,n)$，则其特征可表示为

$$SF=(S,LF)=(d_1,d_2,\cdots,d_n,S_1,S_2,S_3,S_4)$$

时间序列性质 1 设时间序列 $S=\{(t_i,d_i)\}(i=1,2,\cdots,n)$ 的拟合回归线 L 的特征可表示为 $LF=(S_1,S_2,S_3,S_4)$，(t_{n+1},d_{n+1}) 为下一个时间点采样数据，则时间序列 $S'=\{(t_i,d_i)\}(i=1,2,\cdots,n+1)$ 的拟合回归线 L' 特征和时间序列特征 SF' 分别为

$$LF'=(S_1+t_{n+1},S_2+d_{n+1},S_3+t_{n+1}^2,S_4+t_{n+1}d_{n+1})$$

$$SF'=(S,d_{n+1},S_1+t_{n+1},S_2+d_{n+1},S_3+t_{n+1}^2,S_4+t_{n+1}d_{n+1})$$

证明从略。

3 基于一元线性回归模型的空时数据压缩算法

3.1 单个节点的数据压缩

在传感器网络中，节点周期性对环境中的属性进行采样，数据量非常巨大。为了加快数据处理速度、减少计算负荷和节省能量，本文提出了一种加速时间序列数据添加的判定方法。

判定方法 1 给定一个时间序列 $S=\{(t_i,d_i)\}(i=1,2,\cdots,n)$ 及其拟合回归线 $L(t)$，设下一个时间点采样数据值为 d_{n+1}。如果时间序列 S 拟合回归线在 $t=n+1$ 上的拟合值 d_{n+1}' 和 d_{n+1} 之间的差值小于给定的极限值 ε，即 $|d_{n+1}-d_{n+1}'|<\varepsilon$，则该时间点的采样数据可以加入时间序列 $S=\{(t_i,d_i)\}(i=1,2,\cdots,n)$，成为一个新的时间序列 $S=\{(t_i,d_i)\}(i=1,2,\cdots,n+1)$。

单个节点消除数据时间冗余的实现算法描述如表 1 所示。

表1

算法输入：时间序列数据 $S=\{(t_i,d_i)\}(i=1,2,\cdots,n)$。

算法输出：不同时间段拟合回归线的 β_0 和 β_1。

算法步骤：

步骤1 时间序列初始化。

Initiate $SF=\{d_1,d_2,S_1,S_2,S_3,S_4\}$;

步骤2 时间序列添加成员。

(1)if ($N_i\geq N_m$) goto 7; //如果 $S(t)$ 的成员数达到规定的数据个数阈值。

else next; //转到下一步。

(2)calculate $L(t)$ of $S=S(t)=(t=1,2,\cdots,i+1)$; //计算 $S=S(t)=(t=1,2,\cdots,i+1)$ 拟合回归线。

(3)calculate $S'(t)$ of $L(t)(t=1,2,\cdots,i+1)$;

(4)if ($|d_{n+1}-d'_{n+1}|>\varepsilon$) goto 步骤 3; //不满足判定方法1。
else next; //转到下一步

(5)update $SF=\{d_1,d_2,\cdots,d_{i+1},S_1+t_{i+1},S_2+d_{i+1},S_3+t_{i+1}^2,S_4+d_{i+1}^2\}$; //更新时间序列。

(6) goto 11;

(7)calculate d'_{n+1} of $L(t)(t=1,2,\cdots,i)$;

(8)calculate $\Delta d_{i+1}=|d'_{n+1}-d_{i+1}|$;

(9)if ($\Delta d_{i+1}>dx_m$) goto 步骤 3; //超过了规定的误差范围。
else goto 步骤 2;

(10)update $SF=\{d_1,d_2,\cdots,d_{i+1},S_1,S_2,S_3,S_4\}$; //更新时间序列.

(11)if ($\Delta t=T$) goto 步骤 3; //达到了规定的发送时间。
else goto 步骤 2;

步骤3　算法结束。

(1)calculate β_0 and β_1 of $L(t)$;

(2)the algorithm ends。

3.2 汇集节点的数据压缩

定义 4　空间噪音点假定空间相关性较强的不同节点在同一时刻对同一区域采样数据集为 $S(d_{\text{Node}\_i})$ $(i=1,2,\cdots,n)$ (n 是节点的个数)，给定一个阈值 Δd_{sc_th}，如果某个节点 Node_i 在该时刻的采样值 d_i 满足下面的条件：

$$|d_i-\text{Ave}(S(d_{\text{Node}\_i}))|>\Delta d_{sc\_\text{th}}$$

则认为节点在该时刻的采样数据值是空间噪音点。其中 $\text{Ave}(S(d_{\text{Node}\_i}))$ 是空间相关性较强的不同节点在同一时刻采样数据集中所有数据的平均值，$\text{Ave}(S(d_{\text{Node}\_i}))=\frac{1}{n}\sum_{i=1}^{n}d_i$ $(i=1,2,\cdots,n)$。

算法原理　先利用具有空间相关性节点在同一时间对同一区域采样数据相近的规律，根据定义 4 去除不同节点在同一时刻采样数据的空间噪音；然后求出在该时间点不同节点采样数据的平均值；最后，对得到的数据采用与单个节点相似的办法求出该周期内数据的拟合回归线集合 $\{L_i\}$ $(i=1,2,\cdots,m)$ $(m\geq 1$，是该周期内拟合回归线的数目)。集合 $\{L_i\}$ 中拟合回归线的起始和终止时间点及相应的回归系数就是最后要发送的压缩数据。

汇集节点消除数据空间冗余的实现算法描述如表 2 所示。

表2

算法输入：时间序列数据 $S=\{(t_i,d_i)\}(i=1,2,\cdots,n)$。

算法输出：不同时间段拟合回归线的 β_0 和 β_1。

算法步骤：

步骤1　汇集节点接收来自不同节点的数据。

receive $L_j(t)$ from Nodes;

步骤2　汇集节点根据节点的空间关系处理数据。

(1)calculate ΔC_{ij} between Node_i and Node_j; //计算节点之间的局部空间相关指标差值。

(2)if ($\Delta C_{ij}\leq\Delta C_m$) goto 步骤 3; //空间相关指标差值在规定的范围内。
else goto 步骤5;//空间相关指标差值超过规定的范围。

步骤3　消除数据的空间冗余。

(1)calculate d_i from different Nodes when $t=t_i$ according to $L_j(t)$;

(2)calculate $\overline{d}=\sum_{i=1}^{n}d_i$;

(3)calculate $\Delta d_i=|d_i-\overline{d}|$;

(4)if ($\Delta d_i>\Delta d_{\max}$) delete d_i; //大于规定的阈值，去除该数据。
else next; //转向下一步。

(5)repeat 2 to 4 until $\forall\Delta d_i\leq\Delta d_{\max}$;

步骤4　按照与消除时间冗余相同的方法生成拟合回归线。

(1)generate on-dimensional regression lines $\{L(t)\}$;

步骤5　发送数据。

(1)send $\{L(t)\}$ to next Node。

4　实验结果及分析

4.1 评价模型

在传感器网络中，数据压缩算法在追求高峰值信噪比(PSNR)的同时，又要简单、有效，以适应传感器节点存储容量小、计算能力弱和通信能力低的要求，节省能耗并减少网络延时。基于上述分析，构造如下模型评价算法性能(Data Compression Algorithm Performance，DCAP)：

$$\text{DCAP}=f(\text{EC},\text{PSNR})=\frac{\text{EC}}{\text{PSNR}} \tag{3}$$

其中 EC 表示能量消耗，其单位为 nJ；DCAP 是 EC 和 PSNR 的函数。

4.2 实验结果与分析

基于第 4.1 节建立的评价模型，评价了 ODLRST 算法的性能，并与非分布式压缩方法(Non-distributed Approach，NDA)及文献[7,8]提出的分布式小波压缩算法(Distributed Wavlet，DWT)进行了比较。实验数据取自热带大气海洋项目(TAO，http://www.pmel. noaa.gov/tao/)。实验所用数据集是 TAO 在多个地点不同深度共 100 个点的传感器，从 2009 年 3 月 8 日~2009 年 3 月 15 日每天的 12:00 采集到的海水温度数据。实验结果如图 2-图 4 所示。

随着网络中节点缓冲区容量的增加，各种算法的性能都有所改善，但当容量达到了一定大小时，性能趋于稳定(见图 2)。与非分布式压缩算法 NDA

及分布式小波压缩算法 DWT 相比，ODLRST 的压缩性能更好，这一方面是因为节点通过同时去除数据的时间和空间冗余减少了数据的传输量，另一方面是数据在传输前先在单个节点内消除了数据的时间冗余，减少了网络中的通信流量，节省了大量的能量，从而改善了算法的性能。

图 3 表明，随着节点总数的增加，各算法总的耗能与延时增大，分布式压缩算法 DWT 性能优于非分布式压缩算法 NAT，而 ODLRST 压缩算法增长幅度最小。这是由于随着节点总数的增加，导致需要传送到簇头的数据量增大，因而增加了网络耗能与延时。DWT 因为有效地降低了网络内部的数据量，因而更加有利于节省网络能量。ODLRST 能同时减少数据的时间和空间冗余，获得了比 DWT 更好的性能。

图 4 给出了节点间平均距离不同时算法的性能变化。在 NDA 压缩算法中，数据经本地节点压缩后直接传送到簇头，与节点间的平均距离没有关系，算法的性能基本保持不变。DWT 压缩算法通过相邻节点交换信息去除数据间的相关性，增加了额外的通信耗能和延时。节点间平均距离加大对算法 ODLRST 的性能影响很小，这是因为 ODLRST 先在单个节点内去除数据的时间相关性，再通过汇集节点去除空间相关性，这使得参与节点间数据交换的数据量减少，节省了网络耗能与延时，改善了算法的性能。

5 结论

本文提出了传感器网络中一种基于一元线性回归模型的空时数据压缩算法 ODLRST。与非分布式数据压缩算法NAT和分布式小波压缩算法DWT相比，该算法能够有效地同时去除传感器网络中节点内数据的时间冗余和节点间数据的空间冗余，节省、均衡网络内部的能量消耗，且容易实现。仿真实验证明了该算法的这些特点。

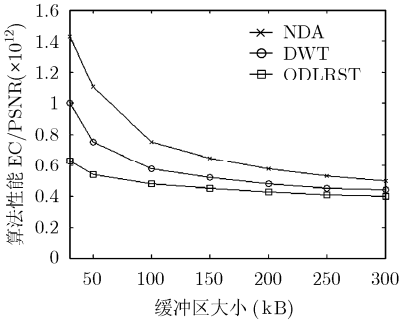

图 2 节点缓冲区大小不同时的算法性能

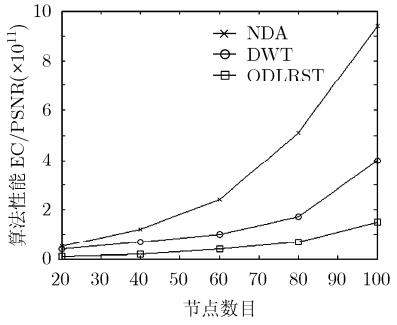

图 3 节点数目不同时的算法性能

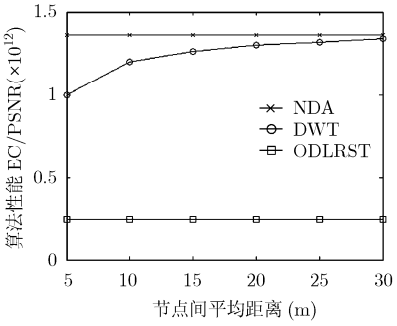

图 4 节点间平均距离不同时的算法性能

参 考 文 献

[1] 刘向宇, 王雅哲, 杨晓春. 面向传感器网络的流数据压缩技术[J]. 计算机科学, 2007, 34(2): 141–143.
Liu X Y, Wang Y Z, and Yang X C. Stream data compression algorithm in wireless sensor networks [J]. *Computer Sciences*, 2007, 34(2): 141–143.

[2] Puthenpurayil S and Ruirui G, Bhattacharyya S S. Energy-aware data compression for wireless sensor networks [C]. IEEE International Confererce on Acoustics, Speech and Signal Processing, 2007, 2: 15–20.

[3] 谢志军, 王雷, 林亚平. 传感器网络中基于数据压缩的汇聚算法[J]. 软件学报, 2006, 17(4): 860–867.
Xie Z J, Wang L, and Lin Y P. Data compression based aggregation algorithm for wireless sensor networks [J]. *Journal of Software*, 2006, 17(4): 860–867.

[4] 周四望, 林亚平, 张建明. 传感器网络中基于环模型的小波数据压缩算法[J]. 软件学报, 2007, 18(3): 669–680.
Zhou S W, Lin Y P, and Zhang J M. Ring based wavelet data compression algorithm for wireless sensor networks [J]. *Journal of Software*, 2007, 18(3): 669–680.

[5] 罗武胜, 鲁琴, 杜列波. 基于 LBT 的无线传感器网络多节点协同图像压缩算法[J]. 传感技术学报, 2008, 21(9): 1600–1604.
Luo W S, Lu Q, and Du L B. LBT based multi-node cooperative image compression algorithm for WMSNs [J]. *Chinese Journal of Sensors and Actuators*, 2008, 21(9): 1600–1604.

[6] Wang Y C, Hsieh Y Y, and Tseng Y C. Compression and storage schemes in a sensor network with spatial and temporal coding techniques [C]. Vehicular Technology Conference, VTC Spring 2008, IEEE, Singapore, 2008: 148–152.

[7] Acimovic J, Cristescu R, and Lozano B. Efficient distributed multi-resolution processing for data gathering in sensor networks [C]. Proc.of the Int'l Conf. on Acoustics, Speech, and Signal Processing, Piscataway: IEEE, 2005: 837–840.

[8] Cristescu R, Lozano B, and Vetterli M, *et al*.. On the interaction of data representation and routing in sensor networks [C]. Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing, Piscataway: IEEE, 2005: 1109–1112.

王雷春： 男，1974 年生，博士，研究方向为传感器网络、网络通信等.

马传香： 女，1971 年生，博士，副教授，研究方向为移动 Ad hoc 网络等.