基于数据流和网络编码的无线传感器网络数据聚合算法

封慧英 周 良 丁秋林
(南京航空航天大学计算机科学与技术学院 南京 210016)

摘 要 为了减少分簇的无线传感器网络(WSN)中数据包传输的数量,并使传感器网络的能量效率最大化,提出了一种节能的自适应数据聚合算法。在该算法中,源节点凭借其存储和计算能力,利用数据流技术减少数据包的传输量;当数据从源节点传输到簇头时,簇头根据控制信息选择一组节点作为编码节点,当数据相关性低于某阈值时,该组节点对数据包进行网络编码,若数据相关性高于某阈值,该组节点则会成为聚合节点进行数据聚合,网络编码和数据聚合可以减少簇头冗余流量,提高能量效率。实验结果显示,使用该算法后,数据包交付率有所提高,能量消耗显著减少。

关键词 数据流,网络编码,无线传感器网络,数据聚合
中图法分类号 TP212 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.5.027

Data Stream and Network Coding-based Data Aggregation Algorithm in Wireless Sensor Networks

FENG Hui-ying ZHOU Liang DING Qiu-lin
(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract An energy-efficient adaptive data aggregation algorithm was developed to reduce the number of packets transmitted in clustering wireless sensor networks(WSN), which also maximizes the efficiency of the sensor networks energy. With the ability of storage and calculation, the source nodes use the data stream technology when sensing data in this algorithm, which leads to the reduction of data transmission. When data are transmitted from source node to cluster head, a set of nodes are selected as network coders by cluster head according to the control information. If the data correlation value is lower than a specific threshold, network coding will be performed by these nodes between the packets. However, the network coder nodes will act as aggregation points if data correlation is higher than that threshold. Network coding and data aggregation can reduce the additional energy consumption in cluster head. Experimental results show that the packet delivery rate is increasing and the energy consumption is significantly decreasing after the algorithm is implemented.

Keywords Data streams, Network coding, WSN, Data aggregation

1 引言

无线传感器网络由大量自治传感器节点以自组织的方式构建,并完成特定的任务。大多数WSN的能量是有限的,传感器不能充电或更换。数据融合技术是针对WSN的高能耗提出的,可以减少大量不必要的网络监控数据传输。传统的数据聚合技术可以减少数据相关性较高的传感器网络流量,但无法保证数据通信的可靠性。在稀疏型WSN中,节点之间相距较远,数据相关性较低,传统数据聚合方法对于能效提升效果不明显;而且,对于数据相关性大小随时间变化的传感器数据,除了普通的数据聚合方法之外,需要额外的高能效机制进行通讯。

国内外的学者对于提高无线传感器网络的性能做了大量的研究和实践工作。Zechinelli等人在充分考虑了传感节点的内部存储RAM之后,根据传感数据的时间相关性提出了一种数据流融合模型[1]。该模型避免了传感器直接将采集数据发往上级节点,利用传感节点的RAM临时保留历史数据。数据流融合模型通过优化发送包大小来减少网络数据包的传输量,但是没有充分考虑数据的空间相似性。HeinZelman等人提出的LEACH协议是一种低功耗自适应聚类路由算法,其基本思想是将网络划分为不同的簇,引入随机选择簇头和轮换簇头以达到能量消耗均衡[2]。但是,LEACH协议没有考虑节点剩余能量的限制条件,并且存在簇头分布不均的问题。ADUC算法是在分析LEACH优缺点的基础上,提出的一种自适应均匀分簇的数据融合算法[3]。该算法保证了簇结构均匀分布,节点负载均衡,但是增加了簇首节点计算的复杂度和计算量。Mhatre和Rosenberg设计了一种分簇式WSN以进行数据采集,在节点和簇头之间使用单跳和多跳混合模式进行通信[4]。此模式比单一通信模式效率有所提高,但是没有考虑能量消耗问题。Bandyopadhyay和Coyle设计了一种传感器网络:传感器在监控区域内均匀分布,整个区域被分成几个簇,每一个簇均有一个簇头。传感器节点使用多跳方法与簇头进行通信。簇头聚合数据并将数据转发到汇聚节点[5]。但是由于簇头和汇聚节点附近的数据冗余量大,能量

到稿日期:2014-06-04 返修日期:2014-09-29 本文受江苏省产学研联合创新资金项目(SBY201320423)资助。
封慧英(1990-),女,硕士生,主要研究方向为信息系统及集成、物联网工程等,E-mail:fenghuiying222@126.com;周 良(1966-),男,博士,副教授,硕士生导师,主要研究方向为信息系统、知识工程;丁秋林(1936-),男,博士,教授,博士生导师,主要研究方向为信息系统、企业信息化。

负荷较高，Bandyopadhyay 和 Coyle 没有考虑靠近汇聚节点的能量负荷问题。网络编码的概念是 Ahlswede 等人提出的[6]，其思想是允许中间转发节点在发送报文前对报文进行组合，从而减少带宽使用，提高节点接收报文的速率。但此网络编码算法的提出是应用在有线网中。研究无线自组网中如何应用网络编码最著名的工作之一是 Katti 等人提出的一种基于异或操作的简单实用的网络编码算法 COPE[7]，该算法针对的是单播路由问题。实验结果表明，COPE 算法显著地提高了网络吞吐量。但是，Katti 等人并没有对算法采用的最优化网络编码条件进行过理论分析，而且网络性能的提高依赖于网络拥塞程度和传输协议等因素。李姗姗[8]等人将网络编码技术与多路径数据传输相结合，减小了链路失效带来的影响，在有效保证传输可靠性的同时，进一步减小能耗，延长了网络的生命周期。但是由于传感器网络数据冗余的特点，对所有数据编码是没有必要的，该算法没有考虑数据之间的相关性。文献[9]利用距离构造传感器之间的支持度，进而提出了一种基于支持度矩阵的各传感器数据一致度计算方法。文献[10]通过计算数据一致度，以聚类分析来判断数据整体分布状况，从而确定聚类的阈值。

本文利用了源数据聚合技术和分簇 WSN 的思想，在充分考虑数据相关性的基础上，提出一种基于数据流和网络编码的无线传感器网络数据聚合算法(Data stream and network coding-based data aggregation algorithm in Wireless Sensor Networks，SNC-DA)。算法研究主要从数据传输及簇头额外的能量消耗的减少两方面展开，在源节点感知数据时利用节点的计算和存储能力以及数据流技术减少数据传输。在簇内，根据数据包的数据相关等级，利用网络编码和数据聚合技术减少簇头额外的能量消耗，以提高算法对无线传感网络的包交付率和能量效率的改善。

2 SNC-DA 算法设计

2.1 相关定义

WSN 分布在一个非常大的地理区域中，并被分成几个面积很大的圆盘形的簇。每个簇覆盖一定面积的监控区域，每个簇有一个簇头(C_{hi})，簇中的传感器节点使用多跳路径与簇头进行通信。传感器通过切换工作/休眠状态来节省能量。

定义 1(源节点，编码/聚合节点，中继节点)　一个传感器均匀分布的网络部署区域 A，包含了 K 个半径为 r_i 的簇 C_i。C_i 中的源节点 s 可以感知数据并将数据传输到簇头节点 C_{hi}。半径为 b_i 的编码区 G_i 有中继节点 N_r 和编码/聚合节点 N_c，中继节点 N_r 仅简单地再次传输已收到的数据。源节点到簇头的多跳路径上的数据在传输到簇头之前，节点 N_c 会对其进行编码或聚合(见图 1 和图 2)。节点 N_r 和 N_c 是由簇头根据控制信息决定的。

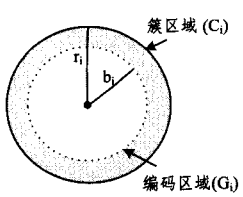

图 1　簇区域和编码区域示意图

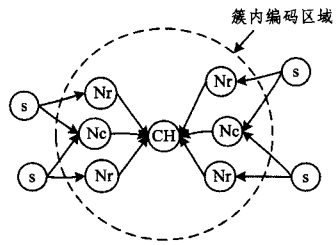

图 2　簇内编码区域示意图

定义 2(占空比)　传感器可以转换工作/休眠状态以保存能量。占空比 μ 是指传感器节点在工作状态的时间与总时间比值的平均值。

定义 3(瓶颈区)　在汇聚节点 Sink 周围半径为 R 的区域形成了瓶颈区 B(见图 3)。半径 R 的最小值为传感器的最大传输范围，所以数据传输一定会经过瓶颈区。本文设一般的路由跳跃长度比簇头与汇聚节点周围的瓶颈区半径要小得多。

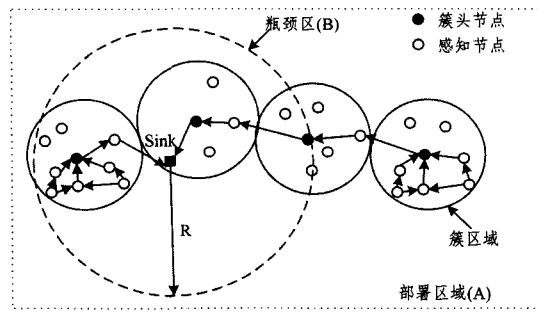

图 3　基于簇的 WSN 数据传输图

簇头在编码区 G_i(见图 1)中选择编码节点和中继节点。簇内编码区域如图 2 所示，簇头 CH 将编码区中的一部分节点选为编码节点，其余节点选为中继节点。编码区内的传感器节点 N_r 和 N_c 利用单跳方法将数据传输到簇头。N_r 仅简单转发数据，N_c 根据接收数据包的相关性大小有选择地进行数据聚合或网络编码，以减少传输流量和簇头的额外负载。另外，在本文中，编码区域之外的传感器节点产生的数据利用多跳路径经路由转发到簇头以提高数据可靠性。

2.2 SNC-DA 算法

本文提出的 SNC-DA 算法总体流程如图 4 所示。

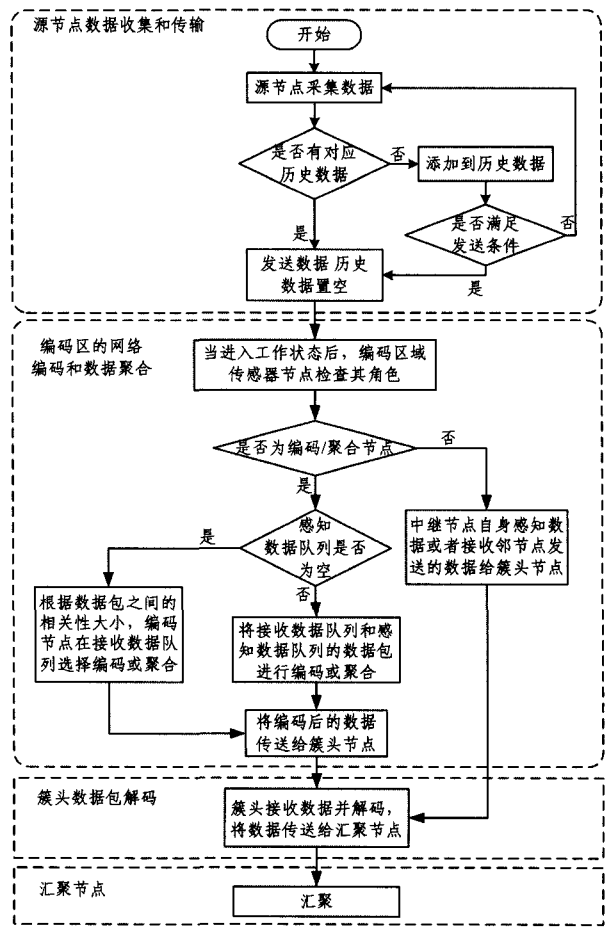

图 4　SNC-DA 算法总体流程

SNC-DA 算法的基本思想是基于分簇，簇头将经过算法处理的数据传送给汇聚节点。首先，源节点利用节点的存储和计算能力，减少数据包传输的数量，最大化传感器网络的能量效率。另外，当数据从源节点传输到簇头时，编码节点对数据包进行编码或聚合，以减少冗余流量，提高能量效率。数据量减少后继续从簇头传输到汇聚节点。如图 2 所示，根据接收包中数据相关性的大小，编码节点选择进行数据聚合还是网络编码。

2.2.1 源节点数据收集和传输

如图 4 中源节点数据收集和传输部分所示，传感器源节点感知到数据后没有立即发送最新的数据到上级节点，而是暂时将数据存储在自己的 RAM 历史位置上，用于临时存储的 RAM 通常包含 1 到 n 的历史样本以及一个计数器。每次插入一个样本，该计数器计数增加，计数器主要用于快速估计数据包的大小。传感器在感知到新数据之前会保持睡眠模式(sleep(d))，传感器感知到数据包(ReceivedPacketExceptionR)后会检查历史样本，如果历史记录中有储存样本(H. getSize())，那么这些样本将会传送到上级节点(H. send())，采样计数器会通过写入数值 0 的方式将历史记录清空(H. reset())。传感器感知到历史存储无记录的新数据(NewDataAvailableException N)时，会将其添加到历史记录中(H. addSample())。当以下条件满足时，传输数据包并清空历史记录：1)历史记录状态为满($S=M_{HIST}$)；2)存储的数据达到最优数据包长度或者最大数据包的有效载荷($S=M_{PAYLOAD}$)；3)信道状态恶化，为防止信道状态继续恶化，需要立即传输数据包。表 1 为 SNC-DA 算法中源数据聚合算法处理过程。

表 1 源数据聚合算法

```
while(true){
    try{sleep(d);}
    catch (ReceivedPacketException R){
      if(H. getSize()>0){H. send();H. reset();}
    }//end catch
    catch(NewDataAvailableException N){
    H. addSample();
    S=H. getSize();
    O=computerOptimalSize(Chan. estimateBER());
    if(S=MHIST or S=MPAYLOAD or O<=S){
      H. send();H. reset();
    }//end if
    }//end catch
  }// end while
```

在保证低网络丢包率的情况下，源数据聚合技术通过减少发送到编码区的数据包来提高能源效率。

2.2.2 编码区的网络编码与数据聚合

网络编码技术[11]是指网络中的节点可以对从邻居节点获取的信息适当地进行编码。以下列出了线性网络编码的广义编码和解码方法[12]。

进行线性网络编码的编码包被视为有限域 $GF(2^s)$ 中的元素。一个节点需要传输网络编码数据时，会选择一组系数 $q=(q_1,q_2,\cdots,q_n)$，q 被称为 $GF(2^s)$ 中的编码向量。一组数据包 $G_i(i=1,2,3,\cdots,n)$ 通过线性编码变成单输出包，通过以下线性组合方程可以计算输出编码包：

$$Y=\sum_{i=1}^{n}q_iG_i,q_i\in GF(2^s) \tag{1}$$

接收节点利用这些系数对编码包进行解码。在接收节点，编码向量 q 与网络中的编码数据一起接收。假设一个节点成功接收了一组数据 $(q^1,Y^1),\cdots,(q^m,Y^m)$。接收节点必须对下面的线性组合方程 ($m$ 个方程、n 个未知数)求解：

$$Y^j=\sum_{i=1}^{n}q_i^jG_i,j=1,\cdots,m \tag{2}$$

式中，未知项 G_i 包含了转发到网络的原始包。本文利用的 XOR 网络编码[7]是线性网络编码的一种特殊的形式。网络中传输的所有包都是 $GF(2)=\{0,1\}$ 中的元素。

假设有两个不同的数据包，从不同的源节点传输到编码节点，它们之间的相关系数为 $\rho_{v_iu_i}$。如果 ρ 大于阈值 γ，那么编码器节点将自身感知数据和接收到的数据进行聚合；如果 ρ 小于阈值，那么节点先对数据包进行编码(XOR 编码方法)，再进行传送。阈值 γ 是根据传感器网络的部署情况设定的。在密集型网络中，传感器节点之间的距离较近，稀疏型网络传感器节点之间的距离较远。本文通过传感器数据的一致度来描述传感器网络的部署情况，一致度则是根据传感器之间的距离来确定。借鉴文献[10]中计算传感器数据一致度的方法，本文定义了两传感器数据一致度 K 随距离 dif 的变化关系。理论上可以把一致度的下降过程分成多段线段，线段越多越能够精确地描述。在实际测试中，使用二段法来描述一致度的下降情况已基本能够满足使用要求。两传感器数据的一致度随距离的变化关系如图 5 所示。

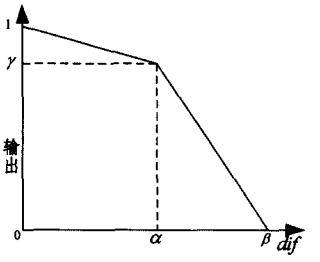

图 5 两传感器数据的一致度随距离的变化关系

表 2 网络编码算法

```
输入：簇中编码节点集 S 和中继节点集 W。Nc 接收到的源节点感知数据
      后，满足发送条件的情况下发送的数据包 h(μ)。
输出：转交给簇头节点的聚合或编码后的数据包。
1. 节点 Nc 从源节点 h1,h2,…,hf 接收数据 hi(u)存在队列 RecvQueue(u)
   中，每一个 Nc 将自己感知的数据存在队列 SensQueue(v)。
2. 如果队列 RecvQueue(u)和 SensQueue(v)非空
3.   取数据包 h(ui)∈RecvQueue(u)和 g(vi)∈SensQueue(v) 并计算相关
     因子 ρvi ui
4.     如果 ρvi ui≥γ 并且均是未编码包
5.     执行 g(Oi)=g(vi)+h(ui)(1−ρvi ui)
6.   否则
7.     执行 g(Oi)=g(vi)⊕ h(ui)
8.   结束
9.   节点 Nc 将 g(Oi)传送给簇头节点
10. 否则如果 SensQueue(v)为空
11.   取数据包 h(ui)和 h(uj)∈RecvQueue(u)并计算相关因子 ρui uj
12.     如果 ρui uj≥γ 并且均是未编码包
13.     执行 g(Oi)=h(ui)+h(uj)(1−ρui uj)
14.   否则
15.     执行 g(Oi)=h(ui)⊕ h(uj)
16.   结束
17.   节点 Nc 将 g(Oi)传送给簇头节点
18. 结束
    重复步骤 2—18，直到队列 RecvQueue(u)和 SensQueue(v)均为空
```

图 5 中，dif 代表两传感器之间的距离，在 0 到 α 之间，两

传感器距离较近，数据一致度下降得慢；在 α 到 β 之间，两传感器距离较远，数据一致度下降得快。α 可以作为区分密集型网络和稀疏型网络的分界点，本文阈值 γ 就是取 dif 为 α 时，K 的值。结合图 5 和测量的实际情况可知，密集型网络 γ 值较高，稀疏型网络 γ 值较低。表 2 为 SNC-DA 算法中网络编码的处理过程。

如图 4 中编码区的网络编码与数据聚合部分所示，簇中编码节点集 S 和中继节点集 W 由簇头决定。用簇头发送的控制信息来选择编码区中的网络编码节点。传感器节点存在两个队列，即接收数据队列 $RecvQueue(u)$ 和感知数据队列 $SensQueue(v)$。当 $RecvQueue(u)$ 和 $SensQueue(v)$ 均非空时，取 $h(u_i) \in RecvQueue(u)$ 和 $g(v_i) \in SensQueue(v)$，如果 $SensQueue(v)$ 为空，取 $h(u_i)$ 和 $h(u_j) \in RecvQueue(u)$。$f(\rho, \gamma)$ 是一个二项函数，其根据数据相关度 $\rho_{v_i u_i}$ 计算两个数据包的显著化差异，并返回 false 和 true。如果两者的数据相关性比 γ 低，那么函数的值是 false，否则，函数的值是 true。编码器节点根据返回值决定进行编码还是只进行数据聚合。如果数据相关度非常高，趋近于 1，那么只进行数据聚合会节省更多的能量。然而，如果数据包之间的相关性极低，网络编码将会更加节能，并且有利于防止数据包在靠近簇头的位置链接失败。若接收到的两个包的相关系数为 1，编码器节点仅仅将数据包的一个单拷贝传输给簇头。当函数返回值为 true 时，两个数据包使用传统数据聚合方法，否则，进行网络编码。

2.2.3 簇头数据包解码

如图 4 中簇头数据包解码部分所示，簇头在编码区 G_i 的中继节点接收本地数据包，从网络编码器节点中接收编码数据包。簇头处理簇中所有的聚合数据，其中有一个包池来储存每个接收的本地数据包。簇头收取到包含 k 个本地数据包的编码包时，将依序恢复包池中的每一个数据包。簇头将 $k-1$ 个本地数据包与接收到的数据包进行异或运算来恢复丢失的包，相对于簇中的其他部分，靠近簇头节点的数据包丢失的可能性会更高。在靠近簇头的节点利用网络编码的方法会保证数据能够成功投递到簇头，并可以节约能量。

2.3 基于 SNC-DA 算法的 WSN 能量消耗

典型的传感器节点在感知、传输、接收数据和休眠状态下均会消耗能量。图 6 示出源节点和中继节点的状态转换情况。

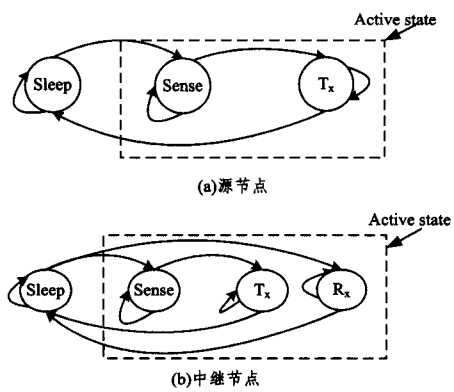

图 6 传感器节点状态转换

在 Bhardwaj 提出的能量消耗模型[13]中，源节点在时间 t 内消耗的能量包括 3 个部分：(1)感知数据和产生数据；(2)传输数据；(3)睡眠状态。$E_s = t[\mu(g\lambda e_s + E_{tx}) + (1-\mu)E_{sleep}]$，$g$ 是传感器在每次事件中产生的比特数，λ 是事件发生的平均概率，e_s 是感知一个比特所需要的能量，μ 是占空比，E_{sleep} 是传感器在睡眠状态下消耗的能量。距源节点距离为 d 的节点每比特消耗的能量为 $E_{tx} = R_d(\alpha_{11} + \alpha_2 d^n)$。$R_d$ 为数据传输率，n 为路径丢失指数，α_{11} 是在传输电路中每比特消耗的能量，α_2 是在传输放大器中消耗的能量。部署在 A 中的传感器总数为 N，工作状态下的传感器的总数为 l，假定它们之间服从二项分布 $[{}_N C_l \mu^l (1-\mu)^{N-l}]$。因此，在占空比 WSN 中长时间处于工作状态下的节点的个数是 $N\mu$。在时间 t 内中继节点的能量消耗包括(1)感知数据和产生数据；(2)传输数据；(3)接收数据；(4)睡眠状态。$E_R = t[\mu(g\lambda e_s + E_{txr}) + (1-\mu)E_{sleep}]$，收发器的能量消耗为 $E_{txr} = R_d(\alpha_{11} + \alpha_2 d^n + \alpha_{12}) = R_d(\alpha_1 + \alpha_2 d^n)$。$\alpha_{12}$ 是接收一个比特所需要的能量。本文假设在工作状态下，单独一个节点不会转换状态，并假设传感器节点保持在状态的可能性为 μ。

式(3)描述了分簇 WSN 的总能量消耗[14]。在区域 B 中，节点的密度大小为 $\frac{N}{A}$，能量消耗包括簇 C_i 内能量消耗的总和 $E_{DB_{cl}}$，簇头传送到汇聚节点的能量消耗为 $E_{DB,in}$、$E_{DB,out}$，另外，MAC 信道和路由协议也会消耗一部分能量，为 $\frac{\mu NB}{A} E_\Delta t$。

$$E_{DB} = \frac{B}{A} t(Ne_s\mu g\lambda - (N-K)\alpha_{12}\mu g\lambda + N(1-\mu)E_{sleep}) + \frac{\alpha_1 n\mu g\lambda t}{(n-1)d_m}\frac{B}{A}\sum_{i=1}^{K}\frac{n_i-1}{n_i}\int_0^{r_i}\int_0^{2\pi} x\beta \mathrm{SdQdS} + \frac{B}{A} \times \frac{\mu g\lambda t\alpha_1 n}{(n-1)d_m}\frac{M-1}{M}\frac{\rho+5}{8}\sum_{i=1}^{K} b_i(n_i-1) + \frac{n\alpha_1 R(A-B)}{(n-1)d_m A}\lambda t \sum_{i=1}^{k} g(O^i) + \frac{\mu NB}{A}E_\Delta t + \frac{\lambda t}{A}\sum_{i=1}^{k} g(O^i)\left(\frac{\alpha_1 n}{(n-1)d_m}\iint_B \Omega \mathrm{d}\sigma - \alpha_{12} B\right) \quad (3)$$

3 实验结果与分析

本文对 SNC-DA 算法和传统数据聚合算法进行仿真比较，分析了 SNC-DA 算法在包交付率(packet delivery ratio, PDR)、簇内能量效率和总能量效耗方面的改进，同时分析了数据相关因子、占空比和簇容量等因素对仿真结果的影响。实验环境为 MATLAB 网络仿真器 PROWLER，用于分析和仿真研究的参数设置如下：节点数(N)为 10^3，面积(A)为 200 * 200(m^2)，瓶颈区域(B)为 60m，路径损耗指数(n)为 2，α_{11} 为 0.937×10^{-6} J/bit，α_{12} 为 0.787×10^{-6} J/bit，E_{sleep} 为 30×10^{-6} J/s，每个工作节点的平均产包率为 960bits/秒。实验使用了简单 MAC 层协议：该协议中，传感器节点在传输包之前需要随机地等待一段时间，如果信道忙碌，还需要随机地等待一段退避的时间。WSN 中的节点不断尝试传输，直到其可以顺利进行。简单 MAC 协议比 802.11 MAC 等一些较复杂的协议节能，适于在 WSN 中使用。平均簇容量大约为 60，最终仿真结果为 20 次仿真结果的平均值。

(1)包交付率分析

图 7 为占空比不同时 SNC-DA 算法和传统数据聚合算法包交付率的变化。包交付率为成功交付的包的数量与发送的包总数之间的比值。当占空比不断提高时，发送节点和接

收节点之间处于工作状态节点的数量随之增加，造成了 PDR 的提高。当占空比较低时，发送节点产生的数据包可能无法到达目标节点，从而导致了 PDR 比较低。当占空比在 0.01 到 0.05 之间时，由于网络节点中的流量较低，SNC-DA 算法对网络能效的改善不明显。随着占空比不断增加，相对于传统数据聚合方法，SNC-DA 算法能显著提高网络的 PDR。

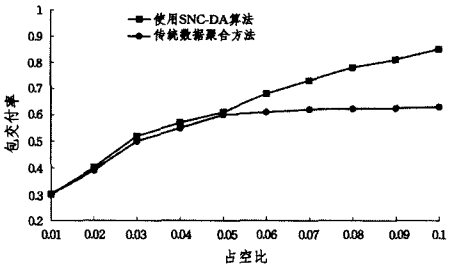

图 7　包交付率随占空比变化的比较

(2)簇内能量效率分析

图 8 显示了 SNC-DA 算法在减少簇内能量消耗方面的提高。簇内能量消耗的百分比由簇中所有节点的能量消耗和区域总能量消耗计算而得。与传统数据聚合方法相比，SNC-DA 算法的网络能量消耗明显降低。当占空比增加时，因簇中传输和接收数据包的次数增加，能量消耗也会增加。图 8 同时显示了数据相关性 ρ 不同时簇内能量消耗百分比的变化情况。当数据相关性提高时，簇中的能量消耗会下降。

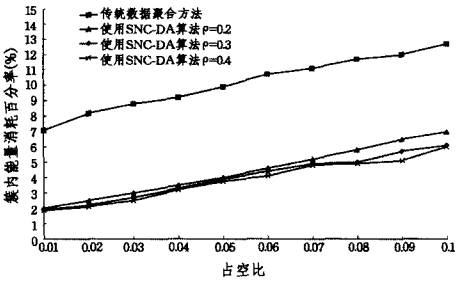

图 8　簇内能量消耗百分比随占空比变化的比较

(3)总能量消耗分析

如图 9 所示，相对于传统数据聚合方法，SNC-DA 算法减少了 WSN 的能量消耗。当数据包的数据相关性极低时，SNC-DA 算法对能效的改进效果非常显著。在稀疏型网络中，中继节点将数据包转发给簇头时，并不会得到相关系数高的包；在占空比密集型网络中，数据包到达中继节点时的相关性也较低。因此，一个具有数据聚合和网络编码功能的节点会帮助传感器网络提高能量效率。

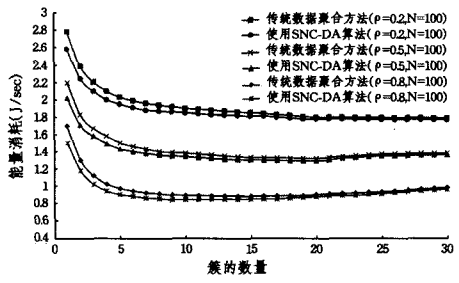

图 9　传统数据聚合算法与 SNC-DA 算法能量消耗对比

如图 10 所示，对于一个容量 K 和节点数 N 固定的簇，随着数据相关因子 ρ 的升高，能量消耗会降低。数据相关性越高，编码器节点转发的数据量就越少。图 10 同时也表明了能耗下降率与簇的容量有关，簇的容量变大会导致能耗下降速率变慢。在实际应用中，可以根据数据相关性大小来选择合适的簇容量。

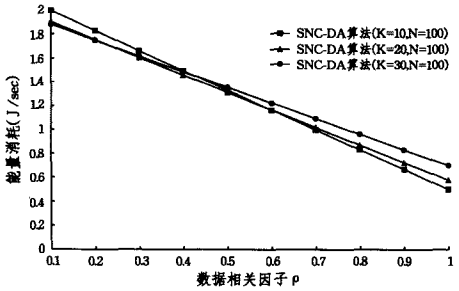

图 10　能量消耗与数据相关因子 ρ 的关系

结束语　本文通过对分簇的无线传感器网络的研究，提出了基于数据流和网络编码的无线传感器网络数据聚合算法。算法利用数据流技术减少数据包的传输量；簇头选择一组控制节点作为编码节点，根据数据相关性大小，该组节点对数据包进行网络编码或数据聚合，减少簇头冗余流量，最大化传感器网络的能量效率。结果显示，相对于传统数据聚合方法，SNC-DA 算法提高了簇中的能量效率和包交付率。本文没有考虑网络拓扑的动态性，如节点的加入、退出或移动可能对网络带来的影响，因此，利用 SNC-DA 算法提高动态网络能量效率的研究是我们后续工作的重点。

参 考 文 献

[1] Zechinelli-Martini J L, Bucciol P, Vargas-Solar G. Energy aware data aggregation in wireless sensor networks[C]//2011 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE). IEEE, 2011: 1-5

[2] Heinzelman W R, Chandrakasan A, Balakrishnan H. Energy-efficient communication protocol for wireless microsensor networks[C]//Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, 2000. IEEE, 2000, 2: 10

[3] 杨婷. 基于自适应动态均匀分簇的 WSN 数据融合算法[J]. 计算机科学, 2012, 39(3)

[4] Mhatre V, Rosenberg C. Design guidelines for wireless sensor networks: communication, clustering and aggregation[J]. Ad Hoc Networks, 2004, 2(1): 45-63

[5] Bandyopadhyay S, Coule E J. An energy efficient hierarchical clustering algorithm for wireless sensor networks[C]//Twenty-Second Annual Joint Conference of the IEEE Computer and Communications(IEEE INFOCOM'03). IEEE Societies, IEEE, 2003, 3: 1713-1723

[6] Ahlswede R, Cai N, Li S Y R, et al. Network information flow [J]. IEEE Transactions on Information Theory, 2000, 46(4): 1204-1216

[7] Katti S, Rahul H, Hu W, et al. XORs in the air: practical wireless network coding[J]. ACM SIGCOMM Computer Communication Review. ACM, 2006, 36(4): 243-254

[8] 李姗姗，廖湘科，朱培栋，等. 基于网络编码的无线传感网多路径传输方法[J]. 软件学报, 2008, 19(10): 2638-2647

[9] 张建业，王占磊，张鹏，等. 多传感器自主在线融合方法[J]. 计算

机应用,2011,31(10):2869-2871
[10] 黎亮,谭世海,师伟. 基于聚类的多传感器数据融合方法研究[J]. 计算机工程,2013,39(5):61-64,68
[11] 董赞强. 基于网络编码的数据通信技术研究[D]. 南京:南京邮电大学,2013
[12] Li S Y R, Yeung R W, Cai N. Linear network coding[J]. IEEE Transactions on Information Theory, 2003, 49(2): 371-381
[13] Bhardwaj M, Garnett, Chandrakasan A P. Upper bounds on the lifetime of sensor networks[C]// IEEE International Conference on Communications, 2001(ICC 2001). IEEE, 2001, 3: 785-790
[14] Rashmi R R, Soumya K G. Adaptive data aggregation and energy efficiency using network coding in a clustered wireless sensor network: An analytical approach[C] // Computer Communications. Volume 40, March 2014: 65-75

(上接第 118 页)

间可能有大量在非顺序的匹配符号[7],而 N-gram[4] 只能找到顺序匹配符号,从而会大幅降低聚类精度;SCS[7] 采用类似生物序列比对中常用的方法来找到匹配的子序列,然后构建符号子串和序列间的矩阵,并没有考虑序列长度的影响,且存在一个子串和多个不同子串同时匹配的情况,可解释性较差;更重要的是,SCS[7] 与 N-gram[4] 并没有严格考虑序列长度对序列间相似度的影响,导致序列间相似度量存在偏倚而影响聚类精度。

SCNS 通过分析序列长度对相似度的影响,提出了规范因子的 3 要素,进而构建了规范因子函数。通过定理 1 和定理 2 证明可以看出这个函数满足规范因子所应具备的 3 个性质,实验表明,该规范因子有效地处理了序列长度对相似性度量带来的影响。

下面测试新方法的可伸缩性。为检验序列长度对相似性度量计算时间的影响,对含有 AGCT 4 个符号的合成的随机序列集进行测试,测试结果如图 2 所示。测试方法如下:选用 N-gram 与 SCS 作为对比算法;SCNS 中参数的 Γ_l 取为 12,N-gram[4] 中参数的 N 也取为 12;生成 6 个随机序列集 1—6,每个序列集包含 10 个样本,其中序列集 i 中每个随机序列的长度为 $1000 * i$;记录上述各个算法在每个数据集上运行 5 次的度量两样本间相似度花费的平均时间。因为这 3 个算法时间复杂度都为 $O(\alpha \cdot L_1 * L_2)$形式(其中 α 具体取值与具体算法有关),所以时间呈二次函数型增长。但从图 2 可以看出,SCNS 时间开销最低,且在实际应用中序列长度较短时,时间开销与序列长度呈线性关系。

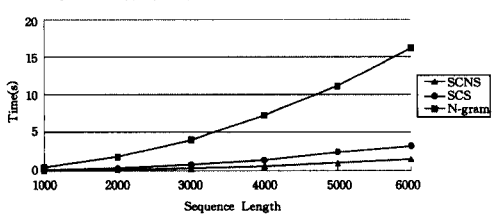

图 6 各算法相似性度量平均时间比较

结束语 现有的序列相似性度量算法没有有效地处理序列长度对相似度的影响,针对此问题,本文提出了新的相似性度量算法和新的聚类方法。在相似性度量算法中,首次提出了规范因子 3 个性质,并引入了满足此 3 要素的规范因子,新的相似性度量算法有效地减小了序列相似性度量中长度的影响。针对基于单链接的凝聚聚类算法的缺点,提出了基于图划分思想的聚类方法,该方法有效地克服了基于单链接的凝聚聚类易受噪声和孤立点影响的缺点,大大提高了聚类精度。实验表明,与现有序列聚类算法相比,SCNS 在多个领域的符号序列数据集上都有很好的聚类质量。下一步工作将研究 SCNS 参数选取策略和寻找提高其运行速度的优化算法。

参 考 文 献

[1] Xiong T, Wang S, Jiang Q, et al. A new Markov model for clustering categorical sequences[C] // Proceedings of the International Conference on Data Mining (ICDM). 2011: 854-863
[2] Dong Guo-zhu, Pei Jian. Sequence Data Mining[M]. New York: Springer-Verlag New York Inc., 2007: 1-65
[3] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61
[4] Kondrak G. N-gram similarity and distance [C]// String Processing and Information Retrieval. 2005: 115-126
[5] Ron D, Singer Y, Tishby N. The power of amnesia: Learning probabilistic automata with variable memory length[J]. Machine learning, 1996, 25(2/3): 117-149
[6] Kelil A, Wang S, Brzezinski R, et al. CLUSS: Clustering of protein sequences based on a new similarity measure[J]. BMC bioinformatics, 2007, 8(1): 286
[7] Kelil A, Wang S. SCS: A new similarity measure for categorical sequences[C]// International Conference on Data Mining. 2008: 343-352
[8] Alpaydin E. 机器学习导论[M]. 范明,等译. 北京:机械工业出版社,2009: 95-96
[9] Grossi R, Vitter J. Compressed suffix arrays and suffix trees with applications to text indexing and string matching[C] // Proc. of ACM STOC. 2000: 397-406
[10] Gusfield D. Algorithms on strings, trees, and sequences [J]. ACM SIGACT News, 1997, 28(4): 41-60
[11] Ukkonen E. On-line construction of suffix trees[J]. Algorithmica, 1995, 14(3): 249-260
[12] Yang J, Wang W. CLUSEQ: Efficient and effective sequence clustering[C] // International Conference on Data Engineering. IEEE, 2003: 101-112
[13] Hirschberg D S. Pattern Matching Algorithms [M]. Oxford Univ. Press, London, 1997: 123-142
[14] Karlin S, Ghandour G. Comparative statistics for DNA and protein sequences: single sequence analysis[J]. Proceedings of the National Academy of Sciences, 1985, 82(17): 5800-5804
[15] Melamed I D. Bitext maps and alignment via pattern recognition [J]. Computational Linguistics, 1999, 25(1): 107-130
[16] Wei D, Jiang Q, Wei Y, et al. A novel hierarchical clustering algorithm for gene sequences[J]. BMC bioinformatics, 2012, 13(1): 174
[17] Halkidi M, Batistakis Y, Vazrgiannis M. On clustering validation techniques[J]. Intelligent In formation Systems, 2001, 17(2/3): 107-145
[18] Larsen B, Aone C. Fast and effective text mining using linear-time document clustering[C] // Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 1999: 16-22