

文章编号:1001-9081(2006)06-1273-06

无线传感器网络中数据汇聚技术的研究

张建明^{1,2}, 宋迎清², 周四望¹, 欧阳竟成¹

(1. 湖南大学 计算机与通信学院, 湖南 长沙 410082; 2. 湖南城市学院 计算机科学系, 湖南 益阳 413049)

(eway_chang@hotmail.com)

摘要: 无线传感器网络能量和计算资源严重受限, 数据汇聚技术是减少能耗、消除数据冗余、增加从源节点到基站的有用信息流、延长网络寿命的重要方法。数据汇聚可以集成在路由协议中, 也可以实现与路由协议紧密交互的独立的协议(或技术)。首先介绍了无线传感器网络中数据汇聚协议的背景; 然后分析和综述了主要的数据汇聚技术, 包括汇聚路由、聚集函数、数据挖掘; 最后提出了这个快速发展领域的研究方向。

关键词: 数据汇聚; 以数据为中心的路由; 聚集函数; 数据挖掘; 无线传感器网络

中图分类号: TP393.04 **文献标识码:**A

Survey on data aggregation techniques in wireless sensor networks

ZHANG Jian-ming^{1,2}, SONG Ying-qing², ZHOU Si-wang¹, OUYANG Jing-cheng¹

(1. School of Computer and Communication, Hunan University, Changsha Hunan 410082, China;

2. Department of Computer Science, Hunan City University, Yiyang Hunan 413049, China)

Abstract: Aiming the severe energy and computing resource constraints of wireless sensor networks, data aggregation is an important approach that assists it in decreasing the energy consumption, eliminating data redundancy, increasing the energy efficiency of the useful information flow from the source to the sink and extending the network lifetime. Data aggregation can be incorporated with routing protocols or can be implemented as individual protocols/techniques that interact closely with routing protocols. Firstly, the backgrounds of aggregation were introduced. Then the main aggregation techniques were analyzed and reviewed, including aggregation routings, aggregate functions and data mining. Finally, the future research directions of data aggregation in wireless sensor network were pointed out.

Key words: data aggregation; data-centric routing; aggregate function; data mining; WSN(Wireless Sensor Networks)

0 引言

微机电系统(Micro-Electro-Mechanism System, MEMS)的迅速发展奠定了设计和实现片上系统(SoC)的基础, 使将传感器、数据处理单元以及通信模块集成到一块集成电路设备中成为可能。这些微型传感器通过自组织方式就构成了无线传感器网络(WSN)。

传感器网络能够协作地实时监测、感知和采集网络分布区域内的各种环境或监测对象的数据, 并对这些数据进行处理, 获得详尽而准确的信息, 传送到需要这些信息的用户。借助于节点内置的形式多样的传感器测量所在周边环境中的热、红外、声纳、雷达和地震波等信号, 从而探测包括温度、湿度、噪声、光强度、压力、土壤成分、移动物体的大小、速度和方向等众多物理现象^[1]。由于具有体积小、价格低廉以及彼此之间可以在近距离内进行无线通信等良好特性, 无线传感器网络在国防军事、反恐抗灾、智能家居、环境监测、地震与气候预测、交通管理、医疗卫生、制造业等许多方面都具有广泛的应用前景。

1 WSN 数据汇聚的背景

收稿日期:2005-12-19; 修订日期:2006-02-20

基金项目:湖南省自然科学基金资助项目(03JJY3098); 湖南省教育厅科研项目(03A011, 05C776); 湖南城市学院科技计划项目(20057306)

作者简介:张建明(1976-), 男, 湖南益阳人, 讲师, 博士研究生, 主要研究方向:机器学习、无线传感器网络; 宋迎清(1966-), 男, 湖南沅江人, 教授, 博士, 主要研究方向:统计学习理论、数字图像处理; 周四望(1971-), 男, 湖南岳阳人, 讲师, 博士研究生, 主要研究方向:无线传感器网络、小波分析; 欧阳竟成(1967-), 男, 湖南岳阳人, 讲师, 博士研究生, 主要研究方向:P2P。

1.1 传感器节点的软硬件结构

传感器节点是一种微型化的嵌入式系统, 硬件部分一般由数据采集、数据处理、无线通信和电源四大基本功能部件组成; 在某些应用中可能还包含节点定位系统、自身移动系统、自供电系统等。

节点通过数据采集部件收集原始数据, 经过数据处理部件处理后, 再进行数据传输, 由此减少通信开销。由于传感器节点既要采集相关的信息又要转发其他节点的数据, 还可能需要控制执行器(Actuator), 为了获得多任务下的实时性, 需要一种适合这类硬件资源和能量有限的廉价硬件平台的超微型嵌入式操作系统; 此外还要具有高度自适应能力的通信协议栈和相关的查询处理、减小通信流量的数据汇聚算法和对原始传感器信号进行简单处理与分析的应用层软件, 所有这些软件构成了传感器节点的软件系统。

1.2 WSN 的通信体系结构

传感器网络是一种新型无线网络, 通信方式和要求都与传统单跳(Hop)的蜂窝、多跳的移动 Ad Hoc 网络(Mobile Ad hoc Network, MANET)不同, 因此对传感器网络中物理、链路、网络、传输以及应用等各层协议和算法都需要进行开创性的研究工作。WSN 在向实用化发展的过程中还有很多理论和

工程方面的问题需要解决,在各个层面以及跨层设计上都有很大的研究和发展空间^[2]。

传感器网络节点所监测的对象往往具有移动性质,因此和 MANET 一样,网络中没有骨干网络,要求支持节点间的多跳通信和路由。但传感器网络中的数据传输和 MANET 具有明显区别^[3]:1) 传感器网络中的通信主要是传感器节点向 Sink 传输数据的一种单向通信,类似一种逆向的组播通信;2) 传感器网络中许多节点监测的可能是同一个对象,因此通信数据具有冗余性质;3) 传感器网络的动态性主要表现在监测对象和汇聚节点,传感器节点本身的移动并不频繁;4) 传感网络中节点数目很大,没有统一的标识,一般不是按 IP 地址而是按数据属性寻址;5) 在许多应用环境下无法进行能量补给,节点易失效,能量消耗是非常重要的性能指标。由于以上这些主要原因,MANET 中端到端的路由协议不适应于传感器网络,并且 MANET 中没有大量流数据要传输。

WSN 的通信体系结构主要有平面式和层次式两种。平面式的 WSN 由许多传感器节点和一个基站(Base Station)组成。基站也被称为 Sink。传感器节点一般靠近监测对象部署,其收集的数据经过处理后,通过无线网络传输到 Sink 或基站,再通过卫星或有线网传输到数据处理中心。这是一种多跳无网络基础设施的网络。外界环境的不确定性经常要求布置成百上千的传感器协同工作,这时可采用层次式。传感器节点按簇(Cluster)分组。每个簇里面的传感器节点根据采取的策略轮流当簇头,簇头直接与基站通信,而簇内其他全部节点只能直接与簇头通信。

1.3 延长 WSN 寿命的根本解决方案

无线传感器网络中的能量资源有限,而前人研究表明^[4]数据传输消耗了总能量的 70%,必须合理利用能量,努力延长 WSN 工作寿命。原始监测数据中有大量冗余信息,通过数据汇聚来减少数据传输量是减少能耗最重要的技术之一。

2 WSN 数据汇聚协议分类

WSN 的通信带宽、节点的计算和存储能力、电池能量都是有限的;在某些应用场景中还存在数据隐私问题;监控应用通常需要快速响应。因此,将数据收集起来后集中式处理是困难和不可扩展的。在大规模稠密部署的传感器网络中,部署的传感器之间不超过某个距离,则位置上相对接近的传感器的采集数据中存在着空间相关。另外,WSN 监测的环境物理量一般变化较慢,则同一个传感器在不同时刻的采集数据中存在着时间相关。如何消除监测数据中的时空相关性是目前数据汇聚的主要目的。

传统网络采用基于地址(Address-Centric, AC)的路由协议,即寻找可设定地址的端节点对之间的最短路由。而 WSN 希望通过在网(in-network)合并冗余数据,寻找从多个源节点到单一目标节点间的路由,在数据传输过程中实现数据汇聚。在给定有限能量和计算资源约束的无线传感器网络中,数据汇聚(Data Aggregation)是帮助提高带宽和能量效率的重要范型。

数据汇聚可以集成在路由协议中,被称为以数据为中心的(Data-Centric, DC)路由协议。此时,汇聚与监测数据的特征与表示形式有关,需要跨协议层理解数据的含义,并且汇聚时可能导致丢失的信息过多。因与应用层相关,故称之为应用相关的数据汇聚(Application Dependent Data Aggregation, ADDA)^[5],如图 1(a)。这类方式是数据汇聚技术的主流。

数据汇聚也可以实现与路由协议紧密交互的独立的协议。AIDA(Application Independent Data Aggregation)^[5]是自适应应用无关的数据汇聚方案,如图 1(b)。AIDA 将汇聚方法实现为网络层和数据链路层之间的一个模块,不需要修改已有的 MAC 层和网络层的协议;并且能够与其他协议层的汇聚技术共存,如图 1(c)。AIDA 通过减少数据封装头部的开销以及 MAC 层的发送冲突来节能。AIDA 不是要最大化网络寿命,而是提出了一种基于网络负载反馈的调度模式,当网络拥塞较轻时不进行或进行低程度的汇聚,在网络负载较重、MAC 层发生碰撞时进行较高程度的汇聚。

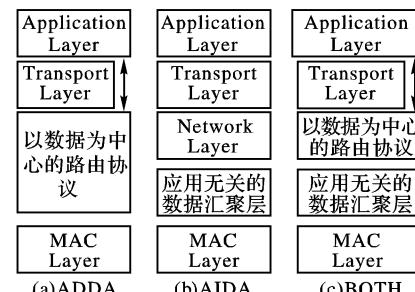


图 1 数据汇聚在 WSN 协议栈中的位置

3 以数据为中心的路由协议

数据汇聚树可以视为逆向组播树。设 k 个有冗余数据要发送的传感器源节点为 S_1, \dots, S_k , 执行汇聚的目标节点为 D 。网络对应的图 $G = (V, E)$ 由全部传感器节点组成, E 为可以直连通信的节点对。设数据汇聚树中任意节点要发送的数据量都为 1。已经证明^[3], 在 DC 协议中, 单位数据的最优传送次数等于 G 中节点集 (S_1, \dots, S_k, D) 的最小 Steiner 树的边数。这是 NP 难的。最优汇聚可作为评价其他汇聚技术的基准。

3.1 信息协商传感协议(SPIN)

SPIN(Sensor Protocols for Information via Negotiation)^[6] 主要目标是通过使用节点间的协商机制和资源自适应机制,解决洪泛(Flooding)法的不足。传感器节点在传送数据之前彼此进行协商,协商制度可确保传输有用数据。

节点间通过发送元数据(即描述传感器节点的采集数据的数据)而不是采集的整个数据进行协商。由于元数据大小远小于采集的数据,所以传输元数据消耗的能量相对较少。为避免盲目使用资源,所有传感器节点必须监控各自的能量变化情况。SPIN 有三种类型数据包:1) ADV。源节点在传输所有数据之前,先向邻节点广播描述这些数据的元数据通告。2) REQ。邻节点如果对这些数据感兴趣,就向源节点请求传输数据。3) DATA。源节点再向这些邻居节点发送数据。

3.2 定向传播路由(DD)

由于传感器网络主要目的是收集传感数据,因此 DD 路由方案(Directed Diffusion)^[7] 以数据为中心考虑路由,其突出特点是引入梯度来描述网络中间节点对该方向继续搜索获得匹配数据的可能性。DD 中传感器网络的节点不以地址作为标识 ID,而是以节点可以提供的监测数据作为寻址依据,用几种约定的属性对其采集的数据命名。

Sink 在网络中广播以属性组合构成的消息询问它所感兴趣的监测数据,这种消息简称为兴趣(Interest)。每个传感器节点在收到兴趣消息后保存在各自的 Cache 中。每个兴趣项(Interest Entry)包含一个时间戳域和若干个梯度域(Gradient

Field,按成本最小化和能量自适应原则引导数据扩散的方向)。当一个兴趣消息传遍整个网络后,与这种兴趣匹配的节点(即兴趣所在区域的传感器节点,称为源节点)到Sink之间的梯度就建立起来了。一旦源节点采集到兴趣所需的数据,响应这种查询(称为事件),并沿该兴趣的梯度路径回送数据给Sink。因为节点分布的稠密性,存在有多个节点匹配兴趣,部分节点可能向Sink回送同样的监测数据;为了节约节点的能量,提高节点的生存期限,这种数据可以在传输路径上先进行汇聚,去掉冗余数据,然后再传输给Sink。

SPIN 和 DD 都是平面式路由协议,由于要求所有传感器节点均具有路由功能,易于导致节点能量消耗过大而失效,从而使得网络拓扑经常发生变化,路由性能较低。为了尽可能地延长节点的生存期限,提高网络的稳定性与路由性能,提出了许多层次式路由。在层次路由中只有一部分节点负责路由信息的转发,因此可以延长大部分节点生存期。层次路由协议存在的主要问题是选举簇头需要额外的开销。

3.3 低能量适应成簇层次路由(LEACH)

LEACH(Low-Energy Adaptive Clustering Hierarchy)^[8]是一种基于簇的层次路由协议。TEEN、PEGASIS 等层次路由协议基本上是对 LEACH 的改进。

LEACH 协议分两个阶段进行操作,即簇建立阶段(Setup Phase)和就绪阶段(Ready Phase),两个阶段持续时间之和总称为一轮(Round)。在簇建立阶段,LEACH 协议随机选择一个传感器节点作为簇头,随机性保证簇头与基站之间数据传输的高能耗成本均匀的分摊到各个传感器节点。在簇头选定好后,各个簇头节点对网络中所有节点进行广播宣布自己为簇头。非簇头传感器节点收到广播数据包,根据接收到的各个簇头节点广播的信号强度,加入信号强度最大的簇头节点所在的簇,向其发送成为其成员的数据包。簇形成后,簇内所辖的节点以 TDMA 的方式分时向簇头传输数据,数据经簇头融合和压缩后,将整合的数据传送给 Sink,由此减少大多数节点的能量消耗。就绪阶段持续了一段时间后,WSN 进入下一轮工作周期。

3.4 分布式数据汇聚层次路由算法(DDCH)

DDCH(Distributed Data-Centric Clustering Hierarchical Routing Algorithm)算法^[9]利用由具有局部最大剩余能量的连通支配集节点所构成的“能量核”来进行路由。首先,网络的邻接节点之间周期性的交换信息,构造一个能量核;然后,当源节点向 Sink 发送数据时,各源节点先在能量核中寻找一条到 Sink 的局部最短路径;然后再将数据汇集至该路径中与源节点相邻的能量核中的节点;最后,能量核中的该节点在其收集到的所有数据进行融合后,沿找到的局部最短路径进行传送。

3.5 基于平衡汇聚树的路由协议(BATR)

BATR 协议(Balanced Aggregation Tree Routing)^[10]利用平衡树来进行路由。每个节点的孩子节点个数与其附近分布的传感器节点个数成正比,从而每个节点上的能量耗费可以得到很好的平衡,延长了网络的寿命。

4 数据汇聚技术的发展方向

4.1 安全的以数据为中心的汇聚路由

具有数据汇聚(融合)功能的路由协议(或称以数据为中心的路由协议),如何构造汇聚树,仍然是目前研究的热点。

同时,簇头接收传感器节点采集的数据,通过检查数据内

容来进行汇聚,减少冗余。这样,这些中间节点必须理解采集数据的含义,不能在传感器节点和基站之间采用加密的方法实现端到端的安全,这就带来了安全隐患。路由协议还要考虑敏感数据汇聚时的隐私安全问题,汇聚过程中与密钥分发、密码体制、签名认证等安全技术紧密结合。由于资源受限,一般采用私钥体制。

4.1.1 安全的数据汇聚与验证协议(SecureDAV)

在按簇组织的 WSN 中,只要簇中妥协节点数不超过 t ,文献[11]提出了一种安全的数据汇聚与验证机制,包括两步:簇密钥建立协议(CKE)和安全的数据汇聚与验证协议(SecureDAV)。

由于传感器节点部署的不确定性,在节点中预先内置对偶密钥并不很合适。由于椭圆曲线密码体制(Elliptic Curve Cryptosystems, ECC)密钥长度较小,计算速度较快,减少了能量、存储、带宽的消耗,CKE 密钥管理采用了 ECC 体制。CKE 给每个簇生成一个秘密的簇密钥;采用 (t, n) 门限秘密共享体制,每个传感器节点只知道这个簇密钥的一部分(称为簇密钥共享),簇密钥对每个节点而言是保密的;簇密钥用于节点认证。这个秘密的簇密钥的公钥对基站和簇内所有节点是公开的。最初,每个传感器节点预先保存了 ECC 域参数和基站的 EC 公钥。部署后,每个传感器将计算自己的 EC 公钥/私钥对,并把自己的公钥向簇内所有节点广播;这个 EC 密钥对用于同其他节点进行安全通信。

一旦在簇头执行了 CKE 协议后,每个传感器节点将拥有簇密钥共享。传感器利用簇密钥共享,采用椭圆曲线数字签名算法 ECDSA,将采集数据生成部分签名。簇头收集来自簇内各成员的部分签名,组合后形成一个完全签名。簇头将汇聚结果和完全签名一起送到基站。基站使用公钥来验证这个完全签名。由于攻击者不知道完整的簇密钥,就不能生成完全签名;同时,若簇内共谋的妥协节点少于 t 个时,基站也不会接受这个有问题的汇聚结果。通过门限签名模式确保了汇聚结果的真实性。

通过采用 Merkle 哈希树确保了汇聚结果的完整性。传感器节点将采集到数据的加密形式及其 Hash 值传送到簇头。簇头基于接收到的 Hash 值建立 Merkle 树。当基站收到带签名的加密汇聚结果时,使用公钥来验证这个签名;通过反复询问簇头每个独立的采集数据,来验证完整性。

Mahimkar 等人的这个方法采用公钥密码体制,计算量存储量要求大,硬件要求高;特别是抗节点复制能力差,并可能导致 DoS 攻击。

4.1.2 能量有效的基于模式码的安全数据汇聚(ESPDA)

ESPDA 协议(Energy-efficient and Secure Pattern-based Data Aggregation)^[12]是按簇组织的 WSN 中,能量有效的、安全的数据汇聚协议。ESPDA 采用模式码,本质上也可以看成是一种元数据协商机制。

各个传感器节点执行模式生成算法,从感应到的原始数据生成模式码(Pattern Code),然后将模式码发送给簇头。簇头根据模式码来标识和分类原始数据,具有相同模式码的原始数据其实际值相近,视为冗余数据。簇头使用模式比较算法,确定模式码的被选集合,请求被选中的传感器以加密形式发送实际数据。具有相同模式码的不同传感器节点集合只需传送其中之一的原始数据到簇头。

簇头根据模式码选择要接收的数据,即实现了汇聚。簇头不需要知道发给基站的数据的含义,不必将传感器的原始数据

解密。这使得在传感器节点和基站之间建立了安全的端到端的连接。簇头和传感器节点之间不需要密钥的分发、广播。

每个传感器节点有一个唯一的内置密钥。基站在每期会话开始都生成一个会话密钥并广播。传感器节点利用内置密钥和接收到的会话密钥计算一个节点相关密钥(NSSK),用于本期会话中数据的加密解密。基站知道所有传感器节点的内置密钥,从而可以计算 NSSK 来解密。传感器节点和基站之间数据传输中采用 Blowfish 加密算法加密。通过使用 NOVSF 跳块技术,在不同会话期间,随机改变数据块到 NOVSF 时间槽的映射方式,提高了安全性。

例:设给定区域有 5 个传感器节点,每个传感器都可以同时测量温度(d1)、压力(d2)、湿度(d3)。如表 1 所示,设所有

测量数据的范围在 0~100 之间。

表 1 鉴定值查找表表

阈值	范围	鉴定值
30	0~30	5
50	31~50	3
70	51~70	7
80	71~80	8
90	81~90	1
95	91~95	4
100	96~100	6

在任意时刻传感器节点感应到的数据 $D = (d_1, d_2, d_3)$ 。模式生成算法利用表 1 计算模式码,如表 2。

表 2 模式码生成表

传感器 iD(d1, d2, d3)	传感器 1(56, 92, 70)	传感器 2(70, 25, 25)	传感器 3(58, 93, 69)	传感器 4(68, 28, 30)	传感器 5(63, 24, 26)
d1 鉴定值	7	7	7	7	7
d2 鉴定值	4	5	4	5	5
d3 鉈定值	7	5	7	5	5
模式码	747	755	747	755	755

传感器 1 和传感器 3 采集的数据是冗余的;类似的,传感器 2、4、5 采集的数据也是冗余的。基于时间戳,对每个冗余数据集,簇头只要选择传感器 1、4 来接收数据。

Cam 等人的工作没有考虑数据的认证,可能导致 Stealthy 攻击。并且 ESPDA 协议中只有一层汇聚节点,这样网络规模受限。

4.1.3 基于参考数据的安全数据汇聚协议(SRDA)

SRDA 协议(Secure Reference-based Data Aggregation Protocol)^[13] 是一种安全的数据汇聚协议。通过比较原始数据与参考(Reference)数据,确定差异(Difference)数据;传感器节点传送差异数据而不是原始数据,从而减少数据传送量。

2002 年 Eschenauer 和 Gligor 基于密钥预置机制,提出了一种基于概率密钥预置(Probabilistic Key Predistribution)模型的传感器网络对偶密钥建立算法^[14]。其主要思想是每个传感器节点在部署之前从密钥池(Key Pool)中随机选择一组密钥得到一个密钥环(Key Ring),使得任意两个传感器节点在一定概率上具有至少一个共同密钥。在此基础上,SRDA 提出了一种基于传感器放置位置估计的密钥分发体制。每个传感器节点预置了包含多个密码的密钥环。采用具有可调参数改变加密强度的 RC6 算法,加密强度随着数据包被传输到更高层次的簇头而提高。

4.2 聚集函数及其安全计算

传感器节点在空闲时经常是关闭的,只有当接到任务指令或监测对象出现时传感器节点才开始工作并产生传感数据。其数据具有阵发性、数据到达持续性、数据量的大小有不可预知性、数据量随任务的变化而变化等,由此可知传感器网络中的数据具有流数据的特点,因而可以用流数据的处理方法来管理传感器网络中的数据。许多数据库研究人员从流数据管理的角度研究了 WSN 的数据收集、存储和管理,已经取得较大进展。

WSN 中,观察者感兴趣的是由 WSN 监测到的时空相关的事件,而不是传感器本身或者大量无关的观测数据。观测者会经常提出与事件相关的时空查询。聚集函数(Aggregate Function)主要包括 COUNT(计数)、SUM(求和)、AVG(平均

值)等。

文献[15]给出了 AVG 的预测计算方法。采用多元线性回归方法,给出了数据流上 AVG 函数值的预测模型和查询处理算法;当预测失败的次数大于预设的阈值时,给出了预测模型的自动调整策略,以降低预测误差。

4.2.1 安全信息聚集(SIA)

用户给出一个时空查询,WSN 返回每个传感器节点收集的原始数据是没必要的和低效的。SIA(Secure Information Aggregation)^[16]指明了大规模无线传感器网络中安全的信息汇聚框架。SIA 首次提出了几个测量数据聚集函数的安全计算方法,如求均值、求最大值、求最小值、求和等。

WSN 中有些传感器节点被称为汇聚器(Aggregator),它们从其他传感器收集原始数据、进行本地处理、响应远程用户的汇聚查询(Query)请求。远离监测现场,提出查询的计算机被称为 Home 服务器。

SIA 采用 aggregate-commit-prove 三步,保证了数据汇聚的安全。在汇聚阶段,通过让汇聚器理解数据含义,可对接收到的数据进行本地运算,只传送运算或查询结果给 Home 服务器。在提交阶段,汇聚器将从传感器收集到的数据集采用 Merkle 哈希树构造一个证明汇聚正确性的提交标志(Commitment),也发给执行查询的 Home 服务器。在证明阶段,Home 服务器和汇聚器执行有效的交互证明,验证汇聚结果的正确性。

4.2.2 基于 Q-Digest 树的数据聚集

文献[17]提出了一种新的数据汇聚结构 Q-Digest 树的构造和合并的算法。利用 Q-Digest 树,使得传感器网络可以响应更多种查询,包括分位数(如中位数 median)、出现频率最高的数据值(如多数值 consensus)、数据分布直方图以及范围。这些查询使用中值、占多数的值等来做近似。这里,一个传感器把从其他传感器接收到的数据汇聚成一个固定大小的消息。

4.3 分布式流数据挖掘

虽然使用聚集函数可节省能量,但是丢失了数据中大量的原始结构,只提供了粗糙的统计量,掩饰了令人感兴趣的局

部变化。为了获得数据不同粒度的表示, 可以尝试采用数据挖掘的方法。

美国工业与应用数学学会(SIAM)在 SDM 2005 大会中第一次组织了关于 WSN 中的数据挖掘的 Workshop; 而这在传感器网络的研究人员中被称为基于模型的数据汇聚(Model-based Data Aggregation)。

将 WSN 产生的数据传送到中心服务器, 再在服务器上采用传统的技术进行数据挖掘有明显的缺陷, 大量数据的传送消耗了宝贵的网络能量。因此, 目前传感器网络中的数据挖掘研究主要集中在网内数据挖掘, 利用传感器节点有限的计算和存储能力减少网络通信量, 从而节省网络能量。传感器节点执行本地算法而非集中式算法, 分布式的并行汇聚, 从而提高可扩展性。

利用传统的信号处理、数据挖掘、模式分类、机器学习技术, 如主成分分析、独立成分分析、回归分析、聚类分析、神经网络、贝叶斯网络、小波分析等, 在各个传感器节点上或具有更强处理能力的专门节点上对 WSN 数据进行融合、压缩、建模、预测, 从而减少数据的传送量。

以上都是将传感器采集的数据直接或在传感器本地消除数据中的时间相关后, 传送到簇头或(和)基站, 在簇头或(和)基站再进行去除冗余、数据挖掘。4.3.3 是真正意义上的分布式算法, 其回归系数是分布式存储和分布式计算的; 每个传感器节点可以回答用户对该节点本地区域的查询。

4.3.1 主成分分析(PCA)

收集式主成分分析 CPCCA(Collective PCA)^[18] 是一种分布式的 PCA。CPCA 适合异构数据环境, 即不同节点上的数据关系模式可以不同。主要步骤如下: 1) 在每个节点本地感应到的数据集上执行 PCA 算法; 选择较大的几个特征值, 并计算对应的特征向量(称为支配特征向量、主元); 将本地数据集沿支配特征向量方向作投影。2) 将各节点投影数据的部分采样、本地支配特征向量, 传送到中心站点。3) 中心站点利用来自其他各个节点的支配特征向量和投影数据, 重构近似的全局数据。4) 在重构出的全局数据集上再执行 PCA; 计算全局的支配特征向量并广播。采用 CPCCA 算法, 文献[18]提出了一种基于全局主元的分布式聚类算法。

Bonamente 提出了一种两层的体系结构^[19], 下层由多个簇头节点收集各自簇内传感器的数据, 并在簇头上分别执行 PCA 算法的顺序形式 PAST 算法; 上层基站收集下层各簇头传送给的压缩了的数据, 执行懒惰学习算法(Lazy learning)获取预测模型。

4.3.2 基于 k-均值的有效数据汇聚(KMBDA)

KMBDA(k-Means-Based Efficient Data Aggregation Protocol)^[20] 是一个分布式算法, 在每个传感器节点和簇头上分别独立执行。WSN 按簇组织, 要避免将带冗余的数据直接从簇内各传感器节点传送到簇头。各传感器节点利用感应到的数据, 应用 k-均值算法生成包含 k 个均值的数据集合。这些均值集合被传送到簇头。簇头利用从簇内各个传感器接收到的均值数据集合, 再应用 k-均值算法生成了合成数据集。然后, 合成数据集被传送到基站。

本协议中, 每个传感器操作按轮划分为两个阶段: 1) 感应阶段。传感器收集数据的时间段。只要传感器收集到了充分的数据量, 它就进入下一个阶段。2) k-均值阶段。在感应阶段收集到的数据集上执行 k-均值算法。结果得到了一个具有 k 个数据项的简化集合($k \ll n$), 它给出了传感器感应到的

n 个数据项的一个好的表示。

在簇头上, 最初均值集为 0。当簇头计算出均值集后, 向本簇成员和基站广播此均值集。接下来, 设置定时器, 簇头开始接收来自簇内传感器节点的均值。定时器超时后, 簇头又开始执行 k-均值算法计算均值。

在传感器上, 在感应阶段感应和存储数据。当接收到来自簇头的均值集广播后, 转入 k-均值阶段。在感应阶段收集到的数据集上执行 k-均值算法。为了减少传输中的能耗, 传感器节点计算簇头发送过来的均值集与自己计算出的均值集的差, 以后传输的都是这个差集。然后回到感应阶段。

仿真表明, 由计算均值集所引入的与原始数据的偏差是非常低的; 随着数据冗余度的增加, 均值的最小集合可作为原始数据集的一种好的表示。但需要在两者之间折衷: 代表性均值数据的个数(k-均值算法的 k)、均值集与原始数据之间的偏差。

4.3.3 分布式回归分析(DR)

分布式回归分析(Distributed Regression)^[21] 是 WSN 的一种有效通用网内建模框架。通过将采集数据投影成低维表示, 可以精确的表示原始数据的结构, 同时有效的减少通信开销。具体是使用线性回归来完成投影, 即数据用基函数的带权线性组合来近似, 如公式(1)。

$$\hat{f}(t) = \sum_i w_i h_i(t) \approx f(t) \quad (1)$$

考虑数据中的本地相关, 用核回归来建模。此时测量值 f 是时间和节点位置的函数。显然, 传感器节点的位置可用一个二元(平面)或三元(空间)的向量 x 表示。用核函数来形式化定义区域(Region)。设 K_j 是一个将位置 x 映射成一个非负数的函数。区域 j 定义为 K_j 的支持(Support), 即使 $K_j(x) > 0$ 的位置的集合。选择核函数时, 要使其在区域边界平稳的减小为 0, 以保证最终的回归函数光滑。

如公式(2) 定义核 j 在位置 x 的规范化核权, 表示将位置 x 与区域 j 的关联度, 用所有 l 个区域在 x 的关联度之和规范化。

$$k_j(x) = \frac{K_j(x)}{\sum_{v=1}^l K_v(x)} \quad (2)$$

如果在每个区域 j 上定义基函数集 H_j , 则有公式(3)。将公式(3)的中括号之间的部分视为基函数, 则核回归实际是线性回归的一种特殊情况。

$$\begin{aligned} \hat{f}(x, t) &= \sum_{j=1}^l k_j(x) \sum_{h_i^j \in H^j} w_i^j h_i^j(x, t) \\ &= \sum_{j=1}^l \sum_{h_i^j \in H^j} w_i^j [k_j(x) h_i^j(x, t)] \end{aligned} \quad (3)$$

采用现有的查询分发技术, 在分发查询的同时一起分发核函数。采用分布式高斯消去法, 使用滑动窗口里的采集数据来拟合回归模型。传感器节点只需与核函数值大于 0 的相邻区域中的节点交换高斯消去法点积矩阵 A 和投影向量 b 的相应元素信息, 而不需与其他位置的节点交换信息。只与附近的几个区域有关, 回归系数是分布式存储和分布式计算的。

执行 DR 算法后, 每个传感器节点可以回答用户对该节点本地区域的查询; 或者各节点将 WSN 数据模型的回归系数传送到查询端, 供用户使用。

4.3.4 人工神经网络

Catterall 首次^[22] 将修改了的 SOM 神经网络引入到传感

器网络的数据处理中。

ART 网络不需假定聚类的类数。Kulakov 对 WSN 中的神经网络算法做了进一步的研究,提出了三种数据挖掘体系结构^[23]:一是在平面式的 WSN 中,簇头收集所有传感器的数据,然后在簇头上执行 Fuzzy ART 算法进行分类。二是在平面式的 WSN 中,各传感器节点执行 Fuzzy ART 算法对自己的采集数据分类,将初步分类结果传送到簇头;簇头收集各传感器的初步分类结果,执行 ART1 算法进行再分类。三是将二推广到了层次式 WSN 中。神经网络算法对数据进行聚类,减少维数,节省了通信耗能。

用 ANN 挖掘,鲁棒性好,对数据噪音不敏感。但是神经网络需要预训练。

5 结语

数据汇聚可以集成在路由协议中,也可以实现成与路由协议、安全技术紧密交互的独立的协议(或技术)。在普适计算时代,安全的面向数据的路由、聚集函数及其安全计算、分布式流数据挖掘是 WSN 数据汇聚研究的主要内容。具体来说,下一步研究计划的问题包括:

1) 部署传感器节点时,地理条件等现实因素的限制使得传感器节点不可能均匀分布成网格等那些规则形式;节点不均匀的分布使得数据的空间采样呈现不规则性。另外,规则的时间采样要求传感器时钟同步,数量众多的传感器节点使得进行精确的时间同步事实上是不可能的。现有的数据挖掘算法均未考虑采样数据的时空不规则性,必须在 WSN 数据挖掘算法中将其有效地处理。

2) 现在有少数人开始研究了数据的时空相关,而同一传感器节点的不同传感模块采集到的数据之间存在的多模相关必须处理。

3) 不同的传感器网络应用对网络中的数据有不同要求,某些应用中可能需要较高精度的数据以了解传感器网络中发生的事件的细节信息,而在另一些应用中,一些统计量信息就能满足系统需求。如何用小波变换^[24]实现这种多分辨分析是值得进一步研究的问题。

4) 传感器网络常常用来监测突发事件,这样,常常有某个传感器节点产生的数据和邻居节点相比产生大的偏差,或者说该节点的历史数据相比有大的偏差,这些数据往往是非常重要的^[25]。如何区分处理异常数据和噪声是值得进一步研究的问题。

5) 现有的数据挖掘算法大多是在单个节点(传感器节点或簇头或基站)内执行,本质上是一种离线方式,即都是先要把要处理的数据全部收集到某个节点再挖掘。必须通过节点间交换信息,利用相关性,真正实现网内数据挖掘。

6) WSN 常常被划分成多个簇,如何根据各个簇的分布特点和数据的统计分布规律来自适应的选择不同的数据挖掘算法是值得进一步研究的问题。将整个传感器网络看成一个整体,如何找到一个全局最优的算法分布;如何根据采集数据的分类结果来重新分簇,选择最优的簇划分方法也需进一步研究。

参考文献:

- [1] 任丰原,黄海宁,林闯. 无线传感器网络[J]. 软件学报,2003,14(7):1282-1291.
- [2] 崔莉,鞠海玲,苗勇,等. 无线传感器网络研究进展[J]. 计算机

研究与发展,2005,42(1):163-174.

- [3] KRISHNAMACHARI B, ESTRIN D, WICKER S. Modeling Data-Centric Routing in Wireless Sensor Networks[A]. 2002 IEEE INFOCOM Proceedings [C]. New York, NY, USA: IEEE Communications Society, 2002. 42-49.
- [4] PERRIG A, SZEWCZYK R, et al. SPINS: Security protocols for sensor network[J]. Wireless Networks, 2002, 8(5):521-534.
- [5] BLUM TM, STANKOVIC J, ABDELZAHER T. Aida: Application independent data aggregation in wireless sensor networks [J]. ACM Transactions on Embedded Computing System, special issue on Dynamically Adaptable Embedded Systems, 2003.
- [6] HEINZELMAN WR, KULIK J, BALAKRISHNAN H. Adaptive protocols for information dissemination in wireless sensor networks [A]. 1999 5th ACM/IEEE Annual International Conference on Mobile Computing and Networking Proceedings [C]. Seattle, WA, USA: ACM, 1999. 174-185.
- [7] INTANAGONWIWAT C, GOVINDAN R, ESTRIN D, et al. Directed Diffusion for Wireless Sensor Networking [J]. IEEE/ACM Transactions on Networking, 2002, 11(1): 2-16.
- [8] HEINZELMAN W, KULIK J, BALAKRISHNAN H. Negotiation based protocols for disseminating information in wireless sensor networks [J]. ACM Wireless Networks, 2000(8):169-185.
- [9] 林亚平,王雷,陈宇,等. 传感器网络中一种分布式数据汇聚层次路由算法[J]. 电子学报,2004,32(11):1801-1805.
- [10] HYUN-SOOK K, KI-JUN H. A Power Efficient Routing Protocol Based on Balanced Tree in Wireless Sensor Networks [A]. 2005 1st International Conference on Distributed Frameworks for Multimedia Applications Proceedings [C]. USA: IEEE Computer Society, 2005.
- [11] MAHIMKAR A, RAPPAPORT TS. SecureDAV: A Secure Data Aggregation and Verification Protocol for Sensor Networks [A]. Proceedings of Globecom 2004 [C]. New York: IEEE Communications Society, 2004. 2175-2179.
- [12] CAM H, OZDEMIR S, NAIR P, et al. Espda: Energy efficient and secure pattern based data aggregation for wireless sensor networks [A]. Proceedings of the Second IEEE Conference on Sensors [C]. New York: IEEE Society Press, 2003. 732-736.
- [13] SANLI HO, OZDEMIR S, CAM H. Srda: Secure reference-based data aggregation protocol for wireless sensor networks [J]. Proceedings of IEEE VTC Fall 2004 Conference. New York: IEEE Society Press, 2004, (7):4650-4654.
- [14] ESCHENAUER L, GLICOR VD. A Key-management Scheme for Distributed Sensor Networks [A]. Proceedings of 9th ACM Conference on Computer and Communication Security [C]. New York: ACM Press, 2002. 41-47.
- [15] 李建中,郭龙江,张冬冬,等. 数据流上的预测聚集查询处理算法[J]. 软件学报,2005,16(7):1252-1261.
- [16] PRZYDATEK B, SONG D, PERRIG A. SIA: Secure Information Aggregation in Sensor Networks [A]. The First ACM Conference on Embedded Networked Systems (SenSys'03) [C]. New York: ACM Press, 2003. 255-265.
- [17] SHRIVASTAVA N, BURAGOHAIN C, AGRAWAL D, et al. Medians and beyond: New aggregation techniques for sensor networks [A]. ACM SenSys 2004 [C]. New York: ACM Press, 2004.
- [18] KARGUPTA H, HUANG W, SIVAKUMAR K, et al. Distributed Clustering Using Collective Principal Component Analysis [J]. Knowledge and Information Systems Journal, 2000, 3(4):422-448.

(下转第 1283 页)

定更新造成得时间延时为 80ms~160ms, 根据图 3 和图 4 所示模型, 当卫星移动路由器与地面核心 IP 网络的接入网关站发生切换时, 卫星移动路由器向其家乡代理进行注册和绑定更新造成的时间延时为 20ms~30ms, 这两种切换延时成周期性出现, 该周期与卫星过顶时间以及用户段发现卫星的时间有关。

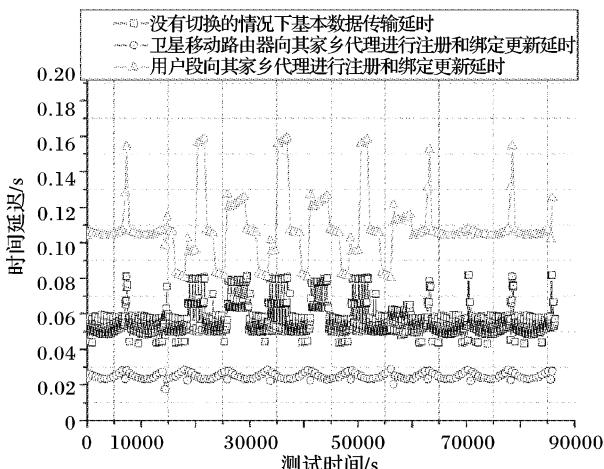


图 6 新型卫星星座通信网络中的星地切换延时

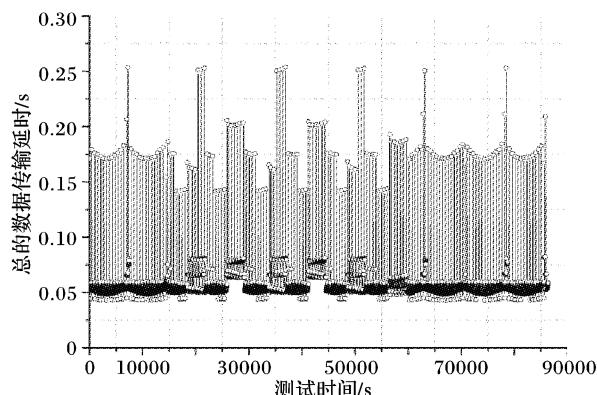


图 7 存在切换延时情况下的数据传输总延时

当将由于切换造成的延时和基本数据传输延时进行总延时统计分析时, 我们获得图 7 所示的仿真结果, 其总延时的变化范围在 43ms~260ms 之间。通过对仿真数据分析发现, 在 18 570s~21 720s, 25 900s~29 230s, 34 470s~37 060s, 41 240s~44 860s, 49 320~51 860s 时间间隔内, 时延显著增加主要是由于在中纬度区域维护越缝星际链路困难, 造成了时延大

幅增加。

4 结语

本文提出了一种基于移动式网络技术的未来卫星星座通信网络体系结构, 详细介绍了其三层的网络体系、基本切换模式以及相关的概念和组件, 并对新型卫星星座通信网络中切换造成的数据传输时延进行了分析。从 NS2 仿真实验的结果中得出, 当用户段发生跨星切换以及卫星移动路由器与地面核心 IP 网络的接入网关站发生切换时, 对数据传输造成瞬时的延迟增加, 影响了卫星星座通信网络的通信性能。提出有效的星地切换方案并实现移动路由器的快速注册和认证是进一步研究的方向。

参考文献:

- [1] IVANCIC W. Architecture Study of Space-Based Satellite Networks for NASA Missions [J]. IEEE Aerospace Conference 2003. Montana: IEEE, 2003. (3): 1179~1187.
- [2] IVANCIC W, STEWART D, BELL T, et al. Application of Mobile-IP to Space and Aeronautical Networks [J]. IEEE Aerospace Conference 2001. Montana: IEEE, 2001, (2): 1027~1033.
- [3] IVANCIC W, STEWART D, BELL T, et al. Application of Mobile Router to Military Communications [A]. MILCOM 2001 [C]. IEEE, 2001. 388~396.
- [4] Transformational Satellite Communications (TSAT) [EB/OL]. <http://www.capitolsource.northropgrumman.com/programs/tsat.html>, 2005.
- [5] Raytheon demonstrates software for T-SAT Ground Stations, EETimes [EB/OL]. <http://www.eetimes.com/news/latest/showArticle.jhtml?articleID=160501404>, 2005.
- [6] WOOD L, CLERGET A, ANDRIKOPOULOS I, et al. IP routing issues in satellite constellation networks [J]. International Journal of Satellite Communications. January/February 2001, 19 (1): 69~92.
- [7] DEVARAPALLI V, WAKIKAWA R, PETRESCU A. Network Mobility (NEMO) Basic Support Protocol [S]. IETF: RFC3963, January 2005.
- [8] IVANCIC W, PAULSEN P, WOOD L, et al. Secure, Network-Centric Operations of a Space-Based Asset: Cisco Router in Low Earth Orbit (CLEO) and Virtual Mission Operations Center (VMOC) [R]. NASA technical report, NASA/TM-2005-213556, Maryland: NASA, May 2005.

(上接第 1278 页)

- [19] BONTEMPI G, BORGNE Y-A. An adaptive modular approach to the mining of sensor networks data [A]. First International Workshop on Data Mining in Sensor Networks of 2005 SIAM International Conference on Data Mining [C]. USA: SIAM Press, 2005. 3~9.
- [20] DHAR S, KANNAN R, RAY L. K-Means Based Efficient Data Aggregation Protocol for Wireless Sensor Networks [EB/OL]. <http://bit.csc.lsu.edu/~rkannan/publications.html>, 2005~11~10.
- [21] GUESTRIN C, BODIK P, THIBAUX R, et al. Distributed Regression: an Efficient Framework for Modeling Sensor Network Data [A]. 3rd International Symposium on Information Processing in Sensor Networks (IPSN'04) [C]. New York: ACM Press, 2004. 1~10.
- [22] CATTERALL E, VAN LAERHOVEN K, STROHBACH M. Self-organization in Ad Hoc Sensor Networks: An Empirical Study [A].
- The 8th International Conference on Simulation and Synthesis of Living Systems [C]. 2002.
- [23] KULAKOV A, DAVCEV D. Data mining in wireless sensor networks based on artificial neural-networks algorithms [A]. First International Workshop on Data Mining in Sensor Networks of 2005 SDM [C]. USA: SIAM Press, 2005. 10~16.
- [24] ACIMOVIC J, CRISTESCU R, BEFERULL-LOZANO B. Efficient distributed multiresolution processing for data gathering in sensor networks [A]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [C]. New York: IEEE Society Press, 2005.
- [25] MA XL, YANG DQ, TANG SW, et al. Online Mining in Sensor Networks [A]. The IFIP International Conference on Network and Parallel Computing (NPC) [C]. LNCS 3222, 2004. 544~550.