

文章编号:1001-9081(2015)12-3341-03

doi:10.11772/j.issn.1001-9081.2015.12.3341

## 基于属性相关性的无线传感网络缺失值估计方法

许可\*, 雷建军

(重庆邮电大学 计算机科学与技术学院, 重庆 400065)

(\*通信作者电子邮箱 xuke@cqupt.edu.cn)

**摘要:**针对无线传感器网络(WSN)中感知数据易缺失问题,提出了一种基于感知数据属性相关性的缺失值估计方法。该方法采用多元线性回归模型,对属性相关的感知数据的缺失值进行估计;同时,为提高算法估计的鲁棒性,提出了基于感知数据属性的数据交织传送策略。仿真结果表明,所提出的估计方法能有效估计无线传感器网络中的缺失值,相比基于时空相关性的线性插值模型(LM)算法和传统的最近邻插值(NNI)算法具有更高的精度和稳定性。

**关键词:**无线传感器网络;属性相关性;缺失值;数据交织;鲁棒性

**中图分类号:** TP393.03    **文献标志码:**A

### Estimating algorithm for missing values based on attribute correlation in wireless sensor network

XU Ke\*, LEI Jianjun

(School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

**Abstract:** The missing of the sensing data is inevitable due to the inherent characteristic of Wireless Sensor Network (WSN), which affects various applications significantly. To solve the problem, an estimation algorithm for missing values based on attribute correlation of the sensing data was proposed. The multiple regression model was adopted to estimate missing values of attribute-correlated sensing data. Meanwhile, a data interleaved transmitting strategy was proposed to improve the robustness of the algorithm. The simulation results show that the proposed algorithm can estimate the missing values and is more accurate and reliable than some algorithms based on temporal and spatial correlation such as Linear interpolation Model (LM) algorithm and the traditional Nearest Neighbor Interpolation (NNI) algorithm.

**Key words:** Wireless Sensor Network (WSN); attribute correlation; missing value; data interleaving; robustness

### 0 引言

近年来,无线传感器网络(Wireless Sensor Network, WSN)迅猛发展,被广泛应用于环境感知<sup>[1]</sup>、生态监控<sup>[2]</sup>和应急方案<sup>[3]</sup>等领域。无线传感器网络由分布于特定区域的具有一定计算、存储和通信能力传感器节点构成,每个节点持续地观测、报告物理现象或特定事件以满足用户需求。但无线传感器网络的自身固有特点决定了感知数据的缺失问题不可避免,如传感器节点断电、较低的发射功率、硬件故障和恶劣的环境条件等均会导致感知数据缺失或损坏。

为获取缺失的感知数据,采用简单的重传机制可能导致较大的时间延迟,并消耗额外的能量和带宽。此外,当发生硬件故障时,在传感器节点转发也并不可行。相对于重传机制,汇聚节点具有更充足的能量供应和计算能力,因此,在汇聚节点对必要的数据进行恢复更为高效和可行。文献[4]提出一种以牺牲缺失值的估计精度为代价的数据估计模型,从而节省能量。文献[5]采用图方法估计任意兴趣区域的数据,但该方法关注的是如何最小化需要访问的传感器节点数量,而非感知数据的估计误差。文献[6~7]采用数据挖掘技术,研究了感知数据流上的缺失值估计问题。算法通过对流数据进行关联规则计算,找到多个数据源节点的频繁模式,并利用该频繁模式来估计缺失值。然而,文献[6~7]中提出的窗口关

联规则挖掘(Window Association Rule Mining, WARM)算法和基于闭频繁项集的关联规则挖掘(Association Rule Mining based on Closed frequent itemsets, CARM)具有极大的局限性,无法被广泛应用。文献[8~11]对无线传感器网络丢失数据估计方法进行了较深入的研究,主要都是基于感知数据的时空相关性建立缺失值估计模型。文献[12~14]使用压缩感知技术或稀疏表达方法采集较少的感知数据,然后在汇聚节点进行全局数据重建,这些方法能够有效降低网络流量和节点能耗,但不能保证重建数据的精度。文献[15~16]提出最小二乘支持向量机(Least Squares Support Vector Machine, LSSVM)模型和粒子群优化的LSSVM模型估计缺失值,同样基于感知数据的时空相关特性。面向移动传感器网络应用,文献[17]提出一种基于数据挖掘技术的缺失值处理方法,其本质仍然是使用感知数据的时空相关特性。

以上文献提出的缺失值估计算法和模型,对处理平稳变化和非平稳变化的感知数据的缺失值均能取得较好的估计效果。但几乎都是基于数据的时空相关性,需要获取相邻节点的数据,往往需要消耗大量的通信带宽;同时,在很多应用中并不能准确获得每个节点的位置信息,节点之间无法建立邻居关系,从而导致无法对节点的缺失值进行估计或对时空相关性不高的节点估计精度较低。本文根据集成型传感器感知数据属性的相关性,提出一种基于属性相关性的缺失值估计

收稿日期:2015-05-04;修回日期:2015-07-16。基金项目:人力资源和社会保障部留学人员科技活动择优资助项目(F201404);重庆市基础与前沿研究项目(cstc2013jcyjA40023);教育部留学回国人员科研启动基金资助项目(F201503);重庆邮电大学青年科学项目(A2012-90)。

作者简介:许可(1982-),男,重庆人,讲师,硕士,主要研究方向:无线传感器网络、嵌入式系统;雷建军(1976-),男,湖北武汉人,副教授,博士,主要研究方向:无线传感器网络、无线互联网。

方法,在传输前对数据进行交织,并给出提高算法鲁棒性的数据交织策略。实验证明,相对于线性插值模型(Linear interpolation Model, LM)<sup>[11]</sup>和最近邻插值(Nearest Neighbor Interpolation, NNI)<sup>[13]</sup>算法,本文方法能更精确地估计无线传感器网络中的缺失值。

## 1 基于属性相关性的缺失值估计算法

无线传感器网络中在一个特定的监测区域通常会布置很多传感器节点,并且单一节点通常集成多类型传感器,如温度、湿度和光照度。多类型传感器的集成成为数据处理带来了新的特性。各种类型传感器采集的感知数据往往存在某些关联,即属性相关性。通过对 Intel 伯克利实验室采集的真实温度和湿度数据进行分析,发现温度、湿度和关照度等不同属性间存在极大的相关性。其中温度和湿度相关性如图 1 所示,通过挖掘并利用这些相关性能够恢复缺失值。

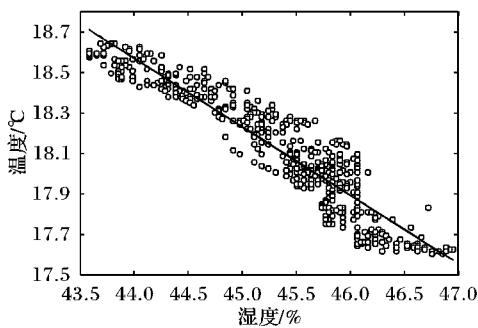


图 1 温度与湿度相关性

对大量真实实验数据研究表明,节点内部感知属性数据的相关性可以采用多元回归模型进行描述。对于任意时刻,多类传感器集成的节点,不同属性感知数据之间的关系可表示为:

$$v_j = \alpha_0 + \alpha_1 \cdot v_1 + \cdots + \alpha_i \cdot v_i \quad (1)$$

其中: $v_j$  和  $v_i$  分别为各类传感器在同一时刻的不同感知属性数据;  $\alpha_i$  为对应于属性  $v_i$  的相关系数。

由多元回归模型可知,选定样本数据对相关系数  $\alpha_i$  进行回归计算,可获得相关系数估计值  $\hat{\alpha}_i$ 。当某一传感器数据缺失时,可以用  $\hat{\alpha}_i$  替换  $\alpha_i$ ,进而通过多元回归模型对缺失值  $\hat{v}_j$  进行估计。

相关系数的估计量可以表示为:

$$(\hat{\alpha}_0, \hat{\alpha}_1, \dots, \hat{\alpha}_i)^T = (X^T X)^{-1} (X^T Y) \quad (2)$$

其中  $Y = (v_{j1}, v_{j2}, \dots, v_{jn})^T$  为属性  $v_j$  的最近时刻连续  $n$  个感知数据。相关属性  $v_i$  的  $n$  个历史数据为:

$$X = \begin{bmatrix} 1 & v_{11} & v_{21} & \cdots & v_{i1} \\ 1 & v_{12} & v_{22} & \cdots & v_{i2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & v_{1n} & v_{2n} & \cdots & v_{in} \end{bmatrix} \quad (3)$$

在实际应用中,属性之间的相关系数可能随时间缓慢变化,导致较大的估计偏差。因此,算法中通过周期性更新回归模型系数,可避免随时间变化导致的估计累计偏差。

## 2 数据交织传送策略

传统的数据传送策略将某一节点同一时刻产生的不同属性数据同时进行传输,一旦数据丢失则意味着该时刻所有属性感知数据缺失,因而无法使用基于属性相关性的缺失值估计方法进行恢复。因此,数据传输前可以采用交织传送策略,

发送节点将缓存的一段时间内的感知数据元组不同属性值进行交叉并分包传输,可有效提高系统的鲁棒性。

如感知节点集成 3 类传感器且各传感器采用同一频率采集数据,则节点内部数据元组的存储方式可表示为:

$$D(Location, Timestamp, Attribute_A, Attribute_B, Attribute_C)$$

其中  $Attribute_A, Attribute_B, Attribute_C$  分别表示 3 种传感器测量的属性值。传感器节点在 3 个连续的时间点采集到的数据分别为:

$$D1(Location, Timestamp1, Attribute_A1, Attribute_B1, Attribute_C1)$$

$$D2(Location, Timestamp2, Attribute_A2, Attribute_B2, Attribute_C2)$$

$$D3(Location, Timestamp3, Attribute_A3, Attribute_B3, Attribute_C3)$$

通过交织算法后传感器节点获得传输数据分别为:

$$DT1(Location, Timestamp1, Attribute_A1, Attribute_B2, Attribute_C3)$$

$$DT2(Location, Timestamp2, Attribute_A2, Attribute_B3, Attribute_C1)$$

$$DT3(Location, Timestamp3, Attribute_A3, Attribute_B1, Attribute_C2)$$

交织算法如下:

输入:本地节点缓存数据;

输出:交织后数据。

- 1) For  $i = 1 : m$
- 2) For  $j = 1 : k$
- 3) If  $i + j <= m$
- 4)      $ii = i + j;$
- 5) Else
- 6)      $ii = (i + j) \% m;$
- 7) Endif
- 8)      $entry(i, j) = entry(ii, j);$
- 9) Endfor
- 10) Endfor

该算法中,  $m$  和  $k$  分别为交织元组数和属性个数,算法的时间复杂度为  $O(mk)$ 。在实际应用中,考虑到实时性要求,缓存的元组数据很少且单个节点集成的传感器数量有限,因此该算法具有较高的执行效率。

通过该交织算法,汇聚节点接收到  $m$  个数据包的任何一个或多个数据包,均可使线性回归模型恢复缺失值,避免了对丢失数据的重传要求。

## 3 实验和仿真

为了对算法性能进行评估,本文利用 Intel 伯克利实验室采集的真实传感器数据集(<http://db.cs.mit.edu/labdata/labdata.html>)进行 Matlab 仿真实验,实验场景如图 2 所示。该数据集包含 54 个传感器节点在 36 天内采集的传感数据。其中,节点每隔 30 s 进行一次采样,每条记录包括温度、湿度、关照度和节点电压 4 个属性数据。由于原始的感知数据集中含有缺失值,在实验过程中首先从原始的数据集合中挑选含有较少缺失值的节点的一段数据,并将缺失值替换为邻近时刻感知数据的加权平均值,进而形成一个完整的测试数据集。将本文算法与经典的基于时空相关性的 LM 算法<sup>[11]</sup>和传统的最近邻插值法 NNI<sup>[13]</sup>进行比较。由于 LM 算法需要使用具有较强空间相关性的近邻节点感知数据,因此所有算法均使

用 10 号节点的一段数据作为测试数据,选取 4、6、7、8 和 9 号节点作为 LM 算法的近邻节点。仿真中,随机地将测试数据集中的已知湿度数据标记成缺失值,然后使用 3 种算法对缺失值进行恢复,评价各种算法的估计精度。在本文算法中使用温度数据对缺失的湿度数据进行恢复。

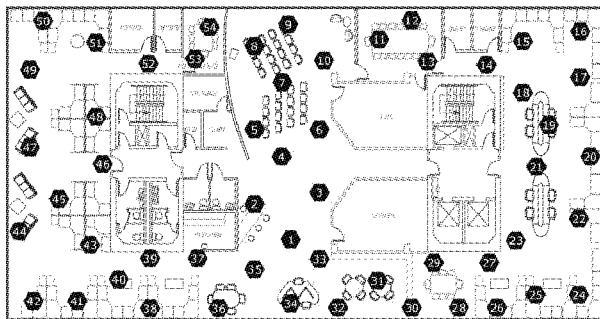


图 2 传感器节点布置

评估时使用均方根误差(Root Mean Square Error, RMSE)对算法的精度进行评估:

$$RMSE = \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / n} \quad (4)$$

其中:  $y_i$  是真实的测量数据,  $\hat{y}_i$  为通过算法获取的缺失估计值,  $n$  表示缺失值的总数。

感知数据的缺失间隔时间是影响算法性能的主要因素,为仿真数据缺失时间间隔对算法性能的影响,本文测试了时间间隔为 1~30 min 的算法性能,结果如图 3 所示。在不同数据缺失间隔时间下,NNI 与 LM 算法均有较小误差,缺失时间间隔没有对其算法性能产生较大影响,因为其精度更多依赖于感知数据的空间相关性。本文算法在数据缺失间隔很小时,算法性能受到较大影响,但随着数据缺失时间间隔增大,误差迅速减小,并略低于 NNI 和 LM 算法。主要原因是当间隔太小时,本文算法回归模型无法获取足够数据对参数进行准确计算。当间隔时间超过 15 min 后,湿度的缺失值估计精度并不会进一步提高,说明该算法在样本容量达到 30 后,回归模型的预测结果已基本稳定。该实验同时表明本文算法对数据缺失不太频繁的应用进行缺失值估计能获得更高的精度。

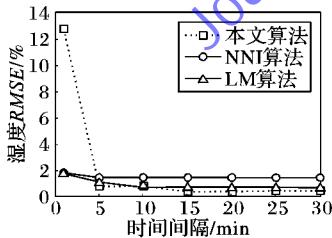


图 3 不同缺失时间间隔算法性能比较

影响算法性能的另一个因素为连续缺失值数量。仿真中,测试了感知数据连续缺失个数为 1~30 时各算法的性能,仿真结果如图 4 所示。从图 4 可看出,各算法误差均随连续缺失值数量增加而增大,这是因为三种算法均需要缺失值邻近时刻的非缺失感知数据。当缺失值与其邻近的非缺失感知数据的时间间隔增大,使得缺失值与其邻近的非缺失感知数据之间的时间相关性降低,从而导致 NNI 和 LM 算法的估计误差增大。同样,相关度的降低也会导致本文算法的回归参数计算不够准确,但该算法具有自适应的参数调节能力,因此无论感知数据的连续缺失值如何变化,本文算法估计性能总具有最好的估计结果和稳定性。

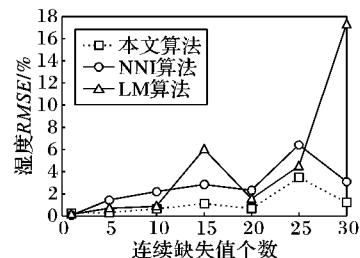


图 4 不同连续缺失值数量算法性能比较

#### 4 结语

感知数据的缺失问题是无线传感器网络的固有问题。为了减少缺失值对无线传感器网络应用的影响,本文提出基于属性相关性的缺失值估计方法,利用单一节点集成的多类型传感器采集的不同属性数据的关联性对缺失值进行估计;同时为提高该方法的鲁棒性,提出使用数据交织传送策略,并给出了交织算法。通过在真实传感器采集的数据集上利用温度属性值对缺失湿度属性值进行估计的仿真实验证明,该方法具有较高的精确度和稳定性。

#### 参考文献:

- [1] SHEPHERD R, BEIRNE S, LAU K T, et al. Monitoring chemical plumes in an environmental sensing chamber with a wireless chemical sensor network [J]. Sensors and Actuators B: Chemical, 2007, 121(1): 142–149.
- [2] SZEWCZYK R, OSTERWEIL E, POLASTRE J, et al. Habitat monitoring with sensor networks [J]. Communications of the ACM, 2004, 47(6): 34–40.
- [3] LORINCZ K, MALAN D J, FULFORD-JONES T R F, et al. Sensor networks for emergency response: challenges and opportunities [J]. IEEE Pervasive Computing, 2004, 3(4): 16–23.
- [4] LI Y, AI C, DESHMUKH W P, et al. Data estimation in sensor networks using physical and statistical methodologies [C]// Proceedings of the 28th International Conference on Distributed Computing Systems. Piscataway: IEEE, 2008: 538–545.
- [5] ZHANG H, MOURA J M F, KROGH B. Estimation in sensor networks: a graph approach [C]// Proceedings of the 4th International Symposium on Information Processing in Sensor Networks. Piscataway: IEEE, 2005: 27.
- [6] HALATCHEV M, GRUENWALD L. Estimating missing values in related sensor data streams [C]// Proceedings of the Eleventh International Conference on Management of Data. New York: ACM, 2005: 83–94.
- [7] JIANG N, GRUENWALD L. Estimating missing data in data streams [C]// Proceedings of the 12th International Conference on Database Systems for Advanced Applications. Berlin: Springer, 2007: 981–987.
- [8] PAN L, LI J. K-nearest neighbor based missing data estimation algorithm in wireless sensor networks [J]. Wireless Sensor Network, 2010, 2(2): 115.
- [9] PAN L, GAO H, LI J, et al. CIAM: An adaptive 2-in-1 missing data estimation algorithm in wireless sensor networks [C]// Proceedings of the 19th IEEE International Conference on Networks. Piscataway: IEEE, 2013: 1–6.
- [10] PAN L, GAO H, GAO H, et al. A spatial correlation based adaptive missing data estimation algorithm in wireless sensor networks [J]. International Journal of Wireless Information Networks, 2014, 21(4): 280–289.

(下转第 3347 页)

求得的增量监测点数量会减少,进而使得新增监测点集与原网络监测点集合并后达到有效监测网络流量的同时,并没有额外增加监测点的数量。总的来说,在网络结构扩充之后,利用本文算法得到的新增网络监测点数量与网络中原有监测点的数量之和,与对整个网络重新布置监测点的数量大致相同,这样只需要对新增网络布置新的网络监测点。

表 3 网络交点数变化的监测点选取结果

$n_{\text{inter}}$	$S_E$	$S + S_E$	$S'$
5	17	94	95
25	17	94	94
50	18	95	95
75	16	93	95

表 4 交点集中监测点数变化的监测点选取数据

$n_{\text{inter}}$	$S_E$	$S + S_E$	$S'$
10	21	98	100
14	19	96	97
16	15	92	94
20	13	90	90

### 3 结语

本文提出一种增量网络监测点的增量选取算法,该算法主要针对网络扩充后,原有网络中布置的监测点不易重新布局时,如何在原有监测点集的基础上增加部分节点从而得到全部网络弱顶点覆盖的监测点集。实验证明增量算法得到的全网监测点数与在全新的网络中重新计算得到的监测点数相差不大,可有效应用于实际的网络监测点部署。

事实上,网络监测点的选择不仅仅与网络的流量覆盖相关,需要考虑的因素很多,例如节点的设备类型、数据的处理能力等。进一步的研究将集中在受限模型中如何进行最优监测点集的选取问题。

### 参考文献:

- [1] LI X, QI F, YUAN Y, et al. Efficient method of station selection for passive monitoring in distributed network using information gain [C]// ISCC 2010: Proceedings of the 2010 IEEE Symposium on Computers and Communications. Washington, DC: IEEE Computer Society, 2010: 796 – 801.
- [2] ZHANG Y, ZHANG H, FANG B. A survey on Internet topology modeling [J]. Journal of Software, 2004, 15(8): 1220 – 1226. (张宇, 张宏莉, 方滨兴. Internet 拓扑建模综述[J]. 软件学报, 2004, 15(8): 1220 – 1226.)
- [3] HAN Q, PUNNEN A P. Strong and weak edges of a graph and linkages with the vertex cover problem [J]. Discrete Applied Mathematics, 2012, 160(3): 197 – 203.
- [4] LIU X, YIN J, TANG L, et al. Analysis of efficient monitoring method for the network flow [J]. Journal of Software, 2003, 14(2): 300 – 304. (刘湘辉, 殷建平, 唐乐乐, 等. 网络流量的有效测量方法分析[J]. 软件学报, 2003, 14(2): 300 – 304.)
- [5] BREBART Y, CHAN C-Y, CAROFALAKIS M, et al. Efficiently monitoring bandwidth and latency in IP network [C]// Proceedings of the 2001 IEEE INFOCOM. Piscataway: IEEE, 2001: 933 – 942.
- [6] TU J, GAO H, LAI W. Submodular function for the minimum weak vertex cover problem [J]. Journal of Beijing University of Chemical Technology: Natural Science, 2011, 38(1): 136 – 139. (涂建华, 高昊宇, 赖文华. 次模函数近似算法求最小弱顶点覆盖[J]. 北京化工大学学报: 自然科学版, 2011, 38(1): 136 – 139.)
- [7] ASHWIN A, OLAF M, MARTIN S. An incremental algorithm for the uncapacitated facility location problem [J]. Networks, 2015, 65(4): 306 – 311.
- [8] DELBOT F, LAFOREST C, PHAN R. New approximation algorithms for the vertex cover problem [C]// Proceedings of the 24th International Workshop on Combinatorial Algorithms, LNCS 8288. Berlin: Springer, 2013: 438 – 442.
- [9] JIANG H. Research on the key technologies of network performance optimization based on network traffic monitor and control [D]. Changsha: Hunan University, 2010: 21 – 31. (蒋红艳. 基于流量监控的网络性能优化关键技术研究[D]. 长沙: 湖南大学, 2010: 21 – 31.)
- [10] ZHOU M, YANG J, LIU H, et al. Modeling the complex Internet topology [J]. Journal of Software, 2009, 20(1): 109 – 123. (周苗, 杨家海, 刘洪波, 等. Internet 网络拓扑建模[J]. 软件学报, 2009, 20(1): 109 – 123.)
- [11] MAGONI D. Network topology analysis and Internet modelling with Nem [J]. International Journal of Computers and Applications, 2005, 27(4): 252 – 269.
- [12] HAERI S, TRAJKOVIC L. Deflection routing in complex networks [C]// Proceedings of the 2014 IEEE International Symposium on Circuits and Systems. Piscataway: IEEE, 2014: 2217 – 2220.

(上接第 3343 页)

- [11] PAN L, LI J, LUO J. A temporal and spatial correlation based missing values imputation algorithm in wireless sensor networks [J]. Chinese Journal of Computers, 2010, 33(1): 1 – 11. (潘立强, 李建中, 骆吉洲. 传感器网络中一种基于时-空相关性的缺失值估计算法[J]. 计算机学报, 2010, 33(1): 1 – 11.)
- [12] GUO D, LIU Z, QU X, et al. Sparsity-based online missing data recovery using overcomplete dictionary [J]. IEEE Sensors Journal, 2012, 12(7): 2485 – 2495.
- [13] KONG L, XIA M, LIU X-Y, et al. Data loss and reconstruction in sensor networks [C]// Proceedings of the 2013 IEEE INFOCOM. Piscataway: IEEE, 2013: 1654 – 1662.
- [14] MILOSEVIC B, YANG J, VERMA N, et al. Efficient energy management and data recovery in sensor networks using latent variables based tensor factorization [C]// Proceedings of the 16th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems. New York: ACM, 2013: 247 – 254.
- [15] GAO S, TANG Y, QU X. LSSVM based missing data imputation in nuclear power plant's environmental radiation monitor sensor network [C]// Proceedings of the 15th International Conference on Advanced Computational Intelligence. Piscataway: IEEE, 2012: 479 – 484.
- [16] GAO S, TANG Y, QU X. Particle swarm optimization least square support machine based missing data imputation algorithm in wireless sensor network for nuclear power plant's environmental radiation monitor [J]. Advanced Materials Research, 2013, 605/606/607: 2137 – 2144.
- [17] GRUENWALD L, SADIQ M S, SHUKLA R, et al. DEMS: a data mining based technique to handle missing data in mobile sensor network applications [C]// Proceedings of the 17th International Workshop on Data Management for Sensor Networks. New York: ACM, 2010: 26 – 32.