

文章编号:1001-9081(2010)05-1153-03

无线传感器网络中位数查询抽样算法研究

刘彩苹¹, 李仁发¹, 付彬¹, 毛建频²

(1. 湖南大学 计算机与通信学院, 长沙 410082; 2. 江西抚州职业技术学院 信息工程系, 江西 抚州 344000)

(liucaiping314@yahoo.com.cn)

摘要:提出一种基于无线传感器网络的中位数查询抽样算法 SAMQ。在 SAMQ 中, 网络中各节点将分布式产生各自的样本集, 然后将样本集聚集传递后汇集到根节点形成全网的样本集, 最后使用这个远小于全网数据集规模的、可用于代表全网数据集结构的样本集, 迅速获得中位数查询的近似结果, 从而无需将各传感器节点的所有数据都传输至根节点, 同时采用了共享无线通道的方式进行通信, 减少了网络数据丢包。理论分析和实验结果显示该算法功耗低、误差较小, 能有效地延长网络的生命周期。

关键词:无线传感器网络; 中位数查询; 抽样算法; 聚集算法

中图分类号: TP212.9 **文献标志码:** A

Research on sampling algorithm for median query based on wireless sensor network

LIU Cai-ping¹, LI Ren-fa¹, FU Bin¹, MAO Jian-pin²

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

2. Department of Information Engineering, Fuzhou Vocational and Technical College of Jiangxi, Fuzhou Jiangxi 344000, China)

Abstract: A Sampling Algorithm for Median Query (SAMQ) based on Wireless Sensor Network (WSN) was proposed. In SAMQ, each node in WSN created a fresh sample summarizing its own observed values and the received values from children nodes, and then broadcasted its subsequence to the parents. Finally, these samples were combined to a single sample in the root node whose data structure was far smaller than the size of the whole data set. The approximate value for median query could be obtained from the sample quickly. A shared channel was used to reduce packet loss. Analytical and experimental results show that the proposed algorithm has the advantages of low power consumption, small error range, and is able to significantly prolong network lifetime.

Key words: Wireless Sensor Network (WSN); median query; sampling algorithm; aggregate algorithm

0 引言

中位数 (Median) 表示一组数据按照大小的顺序排列时, 处于中间位置的数值, 即中位数将数据分成两部分, 一半大于该数值, 一半小于该数值。在实际应用中, 平均数是最典型也是最常用的统计量, 目的是确定一组数据的均衡点。但是平均数 (Average) 受极端值的影响很大, 个别的极端值会直接影响平均数的变化, 不如中位数稳定。因此如果存在极端值或分布偏离比较大时, 常使用中位数描述数据的均衡点。在无线传感器网络 (Wireless Sensor Network, WSN) 环境下, 有很多情况下需要使用中位数查询。例如: 传感器节点很容易发生故障, 导致得到极端的测量值, 在这种情况下采用中位数查询是很有必要的。

Median 查询是传感器网络聚集查询处理的关键运算之一。如何在通信状况不稳定、能量有限、空间有限、计算能力有限的无线传感器网络中进行包括 Median 查询在内的聚集查询是一个全新的挑战, 目前已经有一些这方面的研究成果。

加州大学伯克利分校研究了传感器网络的数据查询技术, 研制了一个数据库系统 TinyDB^[1], 该系统采用了 Tiny Aggregation (TAG)^[2] 机制实现数据聚集查询。在该系统中实

现了 Median 查询, 记为 TAGMD, 算法过程如下: 首先将查询下发到网络中各传感器节点, 同时将各节点组织成一个路由树; 然后将各传感器的记录沿着路由上传至根节点; 最后将所有的记录都聚集在根节点, 最终计算出 Median 值。在没有网络丢包的理想状态, TAGMD 算法能得到准确的 Median 值, 但在通信状况不稳定的无线传感器环境下, TAGMD 算法会有大量丢包产生, 导致结果误差极大。其实不仅仅只有 TAGMD 聚集查询效率差, TAG 中其他聚集查询, 包括 Count 和 Sum, 在有网络丢包的现实环境下运行, 误差都很大。针对 TAG 机制的缺点, 有部分学者提出了一些适用于动态拓扑环境的改进聚集算法, Kamra 等人^[3] 提出一个适应网络拓扑环境不断变化的 CountTorrent 系统; Chen 等人^[4] 提出了一个适应拓扑环境动态变化的分布式算法; Manjhi 等人^[5] 提出了采用 Tributaries-Deltas 方法来减少 TAG 系统在现实网络环境下的丢包率。还有一些学者提出采用近似算法以减少通信量, 如 Roy 等人^[6] 提出了一个在传感器网络中估算 Median 值的方法; Nath 等人^[7] 提出了一个统计不重复记录值的 ODI 系统; Considine 等人^[8] 提出用 Sketches 方法来估计 Count 和 Sum 值等。

本文提出的算法是基于 TinyDB 系统的, 还有一些基于其

收稿日期: 2009-11-18; 修回日期: 2010-01-24。

基金项目: 国家自然科学基金资助项目 (60673061); 高等学校博士点基金资助项目 (20070532048); 湖南省自然科学基金资助项目 (07JJ6135)。

作者简介: 刘彩苹 (1978-), 女, 湖南邵阳人, 讲师, 博士研究生, 主要研究方向: 无线传感器网络、数据挖掘; 李仁发 (1957-), 男, 湖南郴州人, 教授, 博士生导师, 博士, 主要研究方向: 嵌入式计算、无线网络、网络与数字媒体; 付彬 (1978-), 女, 广东英德人, 讲师, 博士研究生, 主要研究方向: 无线传感器网络、无线网络; 毛建频 (1969-), 女, 江西安义人, 讲师, 主要研究方向: 无线传感器网络。

他系统的聚集算法,如康奈尔大学设计了一个 Cougar 系统^[9-10],研究者定义了一个适用于 long-running 查询的模型,提出了节省能源的计算聚集函数的树构造算法,并通过实验证明了无线通信机制对聚集运算的性能有很大的影响^[11-12]。Zhao 等人^[13]提出了一个适用于监控环境下的聚集系统。Vinh 等人^[14]提出了一种基于 Gossip 算法的聚集运算技术, Gossip 算法也称为 Epidemic(流行病)算法,目前在分布式系统研究中作为一个健壮和具有扩展性的信息传播方法得到了普遍应用。

同时,为了减少与 TAG 类似的数据聚集系统的通信量和带宽的损耗, Deshpande 等人^[15]提出用数据压缩的方法减少通信量; Deligiannakis 等人^[16]提出了一种可控误差范围的近似算法。

本文提出一种基于 WSN 的中位数查询抽样算法 (Sampling Algorithm for Median Query, SAMQ)。与 TAGMD 算法不同, SAMQ 无需将各节点生成的数据集都传输到根节点,而只需要每一个节点传递一个样本集,然后将所有样本集经过聚集传递后汇集到根节点形成全网的样本集,最后使用这个远小于全网数据集规模的、可用于代表全网数据集结构的样本集,迅速获得中位数查询的近似结果。理论分析和实验结果显示该算法功耗少、误差较小,能有效延长网络的生命。

1 Median 值查询抽样算法

1.1 抽样算法

经典的随机抽样算法简单易实施,但无法保证生成一个不包含重复数据的样本集,如果对一个含有较多重复值的数据集使用随机抽样算法生成样本集,该样本集中会有很多重复数据,因此得到 Median 值会有较大的误差。而为减少丢包,传感器网络在传输数据时通常会采用一些特殊技术,例如 Churn Cache 技术^[2]、Snooping 技术^[2]等,这些技术在减少网络丢包的同时,也为网络增加了大量的冗余包,因此这种环境下不太适合使用随机抽样算法。本文对随机抽样算法进行了改进,使样本集在数据聚集传输过程中将重复的数据删除,使新算法能得到一个不包含重复数据的样本集,减少算法的误差率。

本文算法将从所有节点生成的数据中抽取 $K\%$ 个不重复数据作为全网的样本集,然后将样本集排序得到 Median 值,算法步骤如下所示。

步骤 1 汇聚节点将 Median 查询下发到网络中各传感器节点,同时将各节点组织成一个路由树。

步骤 2 在每个节点生成一条数据 $\langle val_u, r_u \rangle$, 其中 val_u

为传感器的测量值, r_u 为随机生成的数值, 范围在 $[0, 255]$ 。节点将该数据沿路由树上传给上一层节点。如果节点生成了多条数据, 则按 r_u 排序, 输出 $K\%$ 个 r_u 值最大的数据作为样本集 S 传给上一层节点。

步骤 3 在网络聚集过程中, 每个节点将接收到的数据集以及自己生成的数据集合并生成一个新的数据集, 在数据集集中的所有记录按 r_u 排序, 将重复的记录删除, 最后输出 $K\%$ 个 r_u 值最大的数据作为样本集 S 传给上一层节点。

步骤 4 最终汇聚节点按步骤 3 的方法得到了一个包含有全网所有数据中 $K\%$ 个数据的样本集, 该样本集中不含重复的数据。将样本集中的数据按 val_u 排序后, 得到了处于中间的 Median 值。

1.2 Median 值查询聚集过程

本文算法中的步骤 3 和 4 是将各节点生成的样本集聚集得到整个网络的样本集, 这里以一个连续查询为例说明 Median 值聚集过程。将一个聚集周期分成多个时间段, 在每一个时间段内, 收到聚集请求的节点将前一个时间段收到的样本集和本地生成的数据集合并并且抽样处理后得到一个样本集 S 。与随机抽样算法不同, 本文算法生成样本集是一个没有重复数据的样本集, 也就是说一个数据被复制成多份传播到网上, 也不会被多次抽样到样本集。因此, 本文算法可采用一些在随机抽样算法和 TAGMD 算法中都无法使用的技术用于减少丢包, 算法采用共享无线通道的方式进行通信, 数据包以广播的方式发送, 在通信范围内的所有节点都可以接收到这个数据包, 然而为减少通信量, 并不是所有的邻居节点都要接收数据包, 本算法仅要求节点的上一层邻居节点接收数据包。在聚集过程中, 每个节点会将收集到的多个样本集和自己生成的数据集合并并且抽样处理后得到一个样本集, 然后上传。这样, 第一个时间间隔, 根节点将发送聚集请求给其邻居节点; 第二个时间间隔, 根节点将收到与其距离一跳的节点发回的样本集; ……; 第 m 个时间间隔, 根节点将收到与其距离 $\lfloor m/2 \rfloor$ 跳的节点发回的样本集。设聚集请求到达叶节点和聚集结果返回到根节点的时间间隔为 t , 则经过 t 个时间间隔后, 根节点将生成第一个全网的样本集, 以后每一个时间间隔, 根节点都会生成一个全网的样本集。没有接收到聚集请求的节点监听到其他节点发送的样本集, 它也可向上传送样本集。图 1 描述了一个小型传感器网络中, 以流水线方式实现的样本聚集过程。图中两个节点之间实线表示这条路径是路由树中一部分, 而虚线表示上一层节点能依靠共享无线通信的方式监听到下一层节点发送的样本集, 然后上传给其上一层节点。

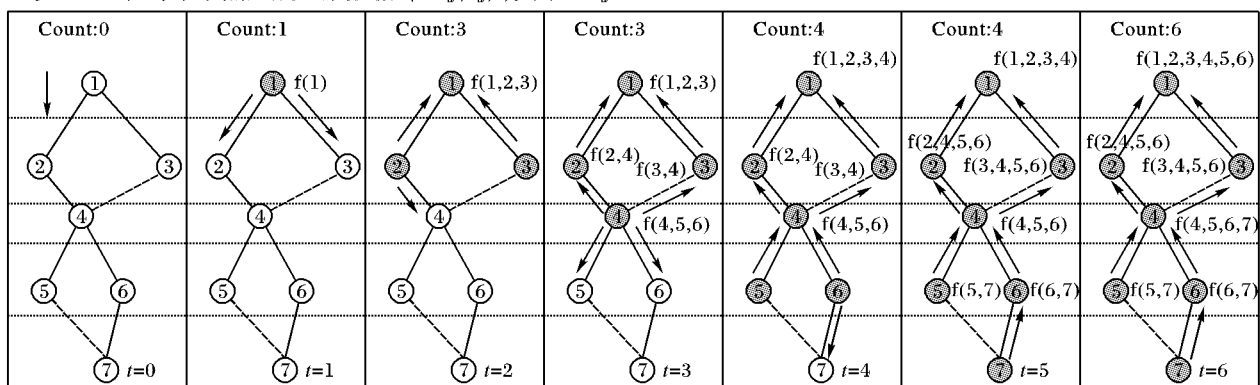


图1 流水线聚集过程

2 仿真实验及分析

2.1 仿真实验

本文在仿真程序 TAG^[2]上实现本文算法。实验模型由

多个节点组成, 每个节点代表一个无线传感器。在实验中, 节点数随着网络直径而变化, 节点数 = 网络直径 × 网络直径。除了选择的测试参数外, 其他参数都采用默认的设置: 网络直径为 30, 节点数为 900, 抽样比例 $K\%$ 为 10%。

图2考查了两种算法在一个聚集周期内节点发送数据的情况。从图2可知,SAMQ在一个聚集周期内所有节点发送的总字节数大约只有TAGMD算法的20%~30%,极大地减少了网络的通信量。

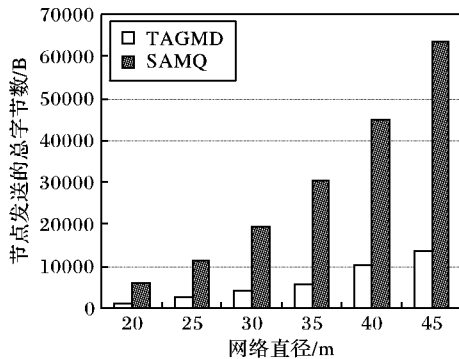


图2 一个聚集周期内所有节点发送的总字节数和网络直径的关系

TAG 仿真程序对现实中的无线传感器网络环境进行了模拟仿真,在这种环境下会有大量丢包产生,导致很多数据包无法送至根节点。以下每一个实验都是运行1000次的测试结果,因此实验数据是一个范围值,在图中我们给出了实验结果的范围以及均值。

图3考查了在理想网络环境即网络没有丢包的情况下,两种算法随着网络规模的增加得到中位数的情况。从图3可知,在理想网络环境下,TAGMD算法得到的实验结果范围非常小,而SAMQ实验结果范围较大,但随着网络规模的增加,实验结果范围在减少。

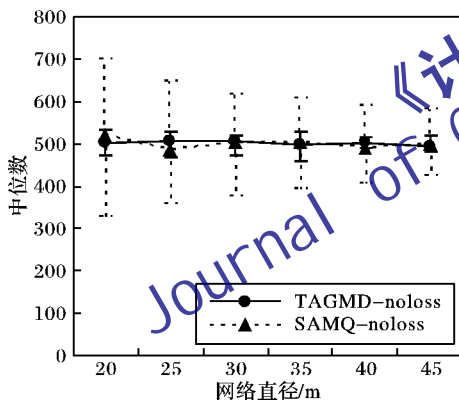


图3 在理想环境下两种算法得到中位数的情况

图4、5分别考查了在模拟现实网络环境下,TAGMD算法和SAMQ以及随机简单抽样(SampledMD)算法和SAMQ随着网络规模增加得到中位数的情况,其中SampledMD算法的抽样比例为10%。从图4、5可知,在模拟现实网络环境下,TAGMD算法和SampledMD算法得到的实验结果范围非常大,而SAMQ实验结果范围相对较小,而且随着网络规模的增加,三种算法实验结果范围的差距越来越明显。

图6考查了两种算法随着网络丢包率的增加得到中位数的情况。从图6可知,TAGMD算法得到的实验结果范围非常大,而SAMQ实验结果范围相对较小;同时随着网络丢包率的增加,两种算法的实验结果范围都在增加。

2.2 算法性能分析

与其他算法相比,本文算法有以下特性。

1) 功耗低。TAGMD算法是将所有数据全部上传至根节点,需要传送大量数据,而本文算法每个节点都只需要传送部分采样数据,需要传输的数据量远小于TAGMD算法,图2的实验结果很好地证明了这一点。

2) 误差较小。在有大量丢包产生的模拟现实环境下,TAGMD算法和SampledMD算法会有很大的误差;而本文算法采用共享无线通道的方式进行通信,即使网络中有丢包产生,仍然能保证将大部分节点的抽样数据信息传输至根节点,因此本文算法误差较小,图4~6的实验结果很好地证明了这一点。

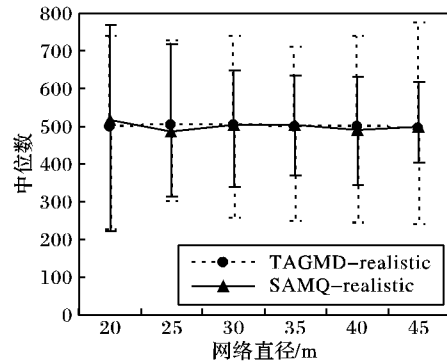


图4 模拟现实环境下TAGMD和SAMQ算法得到中位数的情况

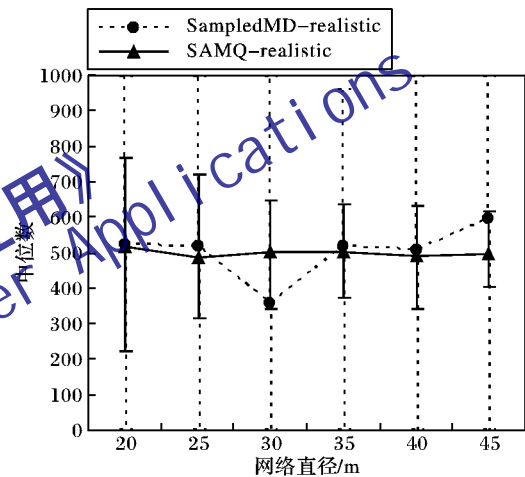


图5 模拟现实环境下SampledMD和SAMQ算法得到中位数情况

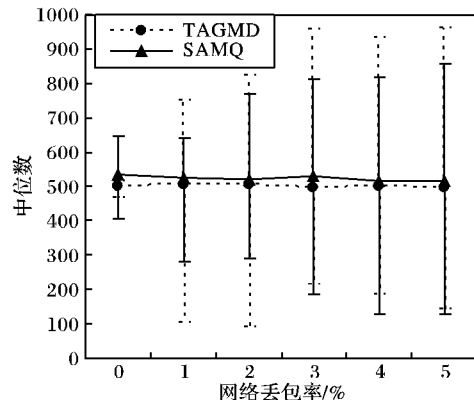


图6 随着网络丢包率的增加两种算法得到中位数的情况

3 结语

本文提出了一种基于无线传感器网络的中位数查询抽样算法。与TAGMD算法不同,本文算法不需要将网络中所有记录都传送到根节点,而是采用抽样的方式传送数据,每个节点最多传送 $K\%$ 个数据,较大地减少了网络的通信量。同时本文算法采用共享无线通道的方式进行通信,确保算法具有相对较小的误差。因此,本文算法非常适合用在像无线传感器网络这样动态的、能源有限的分布式数据流的查询系统中。

(下转第1190页)

率值加权和求概率丢包,使其减少了平均队列长度和丢包率,并减少了网络延迟及波动性,加强了网络的稳定性和鲁棒性。

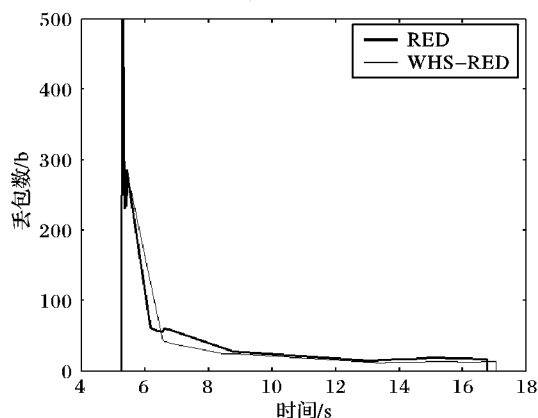


图5 丢包率比较

表1 丢包率数据对比

算法	发送包数	丢失包数	丢包率/%
RED	4899	349	7.12
WHS-RED	4899	295	6.02

4 结语

本文介绍了随机早期检测(RED)的原理和实现方式,分析了 RED 算法的优缺点,并针对其参数设置难和维持网络稳定性方面的问题,提出了一种基于加权和的改进 RED 算法。

(上接第 1155 页)

参考文献:

- [1] FLOYD S, JACOBSON V. Random early detection gateways for congestion avoidance [J]. IEEE/ACM Transactions on Networking, 1993, 1(4): 397-413.
- [2] FENG WU-CHANG, KANDLUR D D, SAHA D, et al. A self-configuring RED gateway [C]// INFOCOM '99: Proceedings of Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Washington, DC: IEEE Communications Society, 1999, 3: 1328-1346.
- [3] QUE D, CHEN ZHI-XIANG, CHEN BI. An improvement algorithm based on RED and its performance analysis [C]// ICSP 2008: 9th International Conference on Signal Processing. Washington, DC: IEEE Press, 2008: 2005-2008.
- [4] OTT T J, LAKESMAN T V, WONG L H. SRED: Stabilized RED [C]// INFOCOM '99: Proceedings of Eighteenth Annual Joint Conference on the IEEE and Computer and Communications Societies. Washington, DC: IEEE Communications Society, 1999, 3: 1346-1355.
- [5] 张明亮,叶澄清.一种基于速率的 RED 增强算法[J].浙江大学学报:工学版,2004,38(2):159-162.
- [6] BONNET P, GEHRKE J E, SESHADRI P. Towards sensor database systems [C]// Proceedings of the 2nd International Conference on Mobile Data Management, LNCS 1987. Berlin: Springer-Verlag, 2001: 3-14.
- [7] YAO Y, GEHRKE J. Query processing in sensor networks [C]// CIDR 2003: Proceedings of the First Biennial Conference on Innovative Data Systems Research. Washington, DC: IEEE Computer Society, 2003: 233-244.
- [8] INTANAGONWIWAT C, ESTRIN D, GOVINDAN R, et al. Impact of network density on data aggregation in wireless sensor networks [C]// ICDCS '02: Proceedings of the International Conference on 22th Distributed Computing Systems. Washington, DC: IEEE Computer Society, 2002: 457-458.
- [9] BIAN F, RANGWALA S, GOVINDAN R. Quasi-static centralized rate allocation for sensor networks [C]// SECON '07: Proceedings of the 4th Annual IEEE Communications Society Conference on Sensor, Mesh, and Ad-Hoc Communications and Networks. Washington, DC: IEEE Press, 2007: 361-370.
- [10] ZHAO J, GOVINDAN R, ESTRIN D. Computing aggregates for monitoring wireless sensor networks [C]// Proceedings of First IEEE Workshop on Sensor Network Protocols and Applications. Washington, DC: IEEE Press, 2003: 139-148.
- [11] VINH T-Q, TAKUMI M. A novel gossip-based sensing coverage algorithm for dense wireless sensor networks [J]. Computer Networks: The International Journal of Computer and Telecommunications Networking, 2009, 53(13): 2275-2287.
- [12] DESHPANDE A, GUESTRIN C, MADDEN S R, et al. Model-driven data acquisition in sensor networks [C]// Proceedings of the Thirtieth International Conference on Very Large Data Bases. [S.l.]: VLDB Endowment, 2004: 588-599.
- [13] DELIGIANNAKIS A, KOTIDIS Y, ROUSSOPOULOS N. Bandwidth-constrained queries in sensor networks [J]. The International Journal on Very Large Data Bases, 2007, 17(3): 443-467.
- [14] MADDEN S, FRANKLIN M J, HELLERSTEIN J M. TinyDB: An acquisitional query processing system for sensor networks [J]. ACM Transactions on Database Systems, 2005, 30(1): 122-173.
- [15] MADDEN S, FRANKLIN M J, HELLERSTEIN J M, et al. TAG: A tiny aggregation service for ad-hoc sensor networks [C]// OSDI '02: Proceedings of the Fifth Symposium on Operating Systems Design and Implementation. New York: ACM Press, 2002: 131-146.
- [16] KAMRA A, MISRA V, RUBENSTEIN D. CountTorrent: Ubiquitous access to query aggregates in dynamic and mobile sensor networks [C]// SenSys'07: Proceedings of the 5th International Conference on Embedded Networked Sensor Systems. Sydney, Australia: ACM Press, 2007: 43-57.
- [17] CHEN J-Y, PANDURANGAN G, XU DONGYAN. Robust computation of aggregates in wireless sensor networks: Distributed randomized algorithms and analysis [J]. IEEE Transactions on Parallel and Distributed Systems, 2006, 17(9): 987-1000.
- [18] MANJHI A, NATH S, GIBBONS P B. Tributaries and deltas: Efficient and robust aggregation in sensor network streams [C]// Proceedings of the 2005 ACM SIGMOD international conference on Management of Data. New York: ACM Press, 2005: 287-298.
- [19] ROY S, CONTI M, SETIA S. Securely computing an approximate median in wireless sensor networks [C]// Proceedings of the 4th International Conference on Security and Privacy in Communication Networks. New York: ACM press, 2008: 6.
- [20] NATH S, GIBBONS P B, SESHAN S, et al. Synopsis diffusion for robust aggregation in sensor networks [J]. ACM Transactions on Sensor Networks (TOSN), 2008, 4(2): 7.
- [21] CONSIDINE J, HADJIELEFTHARIOU M, LI FEIFEI, et al. Robust approximate aggregation in sensor data management systems [J]. ACM Transactions on Database Systems (TODS), 2009, 34(1): 6.