无线传感器网络中中位数查询算法研究

吴中博[1,2,3],张 辉[1,2],陈 红[1,2]
(1.中国人民大学信息学院,北京 100872;2.中国人民大学数据工程与知识工程教育部重点实验室,北京 100872;
3.襄樊学院数学与计算机科学学院,湖北襄樊 441053)

摘 要: 低廉的价格和恶劣的环境会导致传感器节点采样数据中存在误差和异常数据,所以有时候需要通过中位数查询来反映整个监测区域的平均水平.本文首先提出了基于等高直方图的中位数查询算法 HMA,然后我们对其进行了扩展,提出了结合直方图与过滤器的 HFMA 算法,每个采样周期中只需要收集落在过滤器当中的数据并聚集数据的影响因子,基站根据收集的数据和影响因子聚集值计算出中位数.实验表明 HFMA 算法优于 NAIVE 算法和 HMA 算法,可以有效的节省能量开销,提高网络生命周期.

关键词: 传感器网络; 中位数查询; 过滤器; 影响因子
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2010) 2A-133-05

Median Query Research in Wireless Sensor Networks

WU Zhong-bo[1,2,3], ZHANG Hui[1,2], CHEN Hong[1,2]
(1. *Information School, Renmin University of China, Beijing* 100872, *China*;
2. *Key Laboratory of Data Engineering and Knowledge Engineering, Ministry of Education, Beijing* 100872, *China*;
3. *School of Mathematics and Computer Science, Xiangfan University, Xiangfan, Hubei* 441053, *China*)

Abstract: Poor quality and harsh condition can result in faulty and outlier data in sampling data of sensor nodes. So we need median query to reflect average level of monitoring region. First, we put forward HMA algorithm. Second, we extend HMA algorithm and put forward HFMA algorithm. In HFMA, We only need collect data inside filter and aggregate influence coefficient during sampling period. Base station can compute median result according to the sample data inside filter and influence coefficient aggregation value. Experimental results have shown that HFMA outperforms Naive algorithm and HMA algorithm and can prolong the lifetime of sensor network.

Key words: sensor network; median query; filter; influence coefficient

1 引言

传感器的电源能量极其有限,网络中的传感器由于电源能量的原因经常失效或废弃,如何在网络工作过程中节省能源,最大化网络的生命周期,是传感器网络研究面临的重要挑战[1,2].

由于传感器网络中节点采集数据的误差和监测区域中异常数据的存在,常常会导致平均值(AVG 查询)并不能很好的反映整个监测区域的平均水平.在这种情况下,就需要通过中位数(Median)查询来反映整个监测区域的平均水平.

在这篇文章中,我们致力于解决无线传感器网络中连续的、精确的中位数查询.首先我们提出了基于等高直方图的中位数查询算法 HMA(Equal-height Histogram based Median Query Algorithm),然后我们对其进行了扩展,提出了结合直方图与过滤器的 HFMA 算法(Histogram and Filter based Median Query Algorithm).就我们所知,这是第一次在无线传感器网络中研究连续的、精确的 Median 查询处理问题.

2 相关工作

目前已经开发出一些无线传感器网络数据管理系统用于支持传感器网络中的查询处理,例如 TinyDB[3], Cougar[4]等,但这些系统并不支持 Median 查询.文献[5]研究了数据仓库中 Quotient cube 上的 Median 查询计算.文中作者提出了一种 addset 数据结构,并在此基础上提出了基于滑动窗口的增量维护的 Median 计算技术.这种集中式的算法不适合传感器网络这种分布式的结构.

收稿日期:2009-05-01;修回日期:2009-11-12
基金项目:国家自然科学基金(No.60603046,No.60673138);国家 863 高技研究发计划(No.2008AA01Z120);教育部新世纪优秀人才支持计划

文献[6]利用传感器网络内部的数据分布,通过一些规则自动的将数据组成大小不一的桶,然后在 SINK 节点和网内节点上都维护各自独立的 q—digest(QUANTILE digest),在此基础上进行聚集查询、MEDIAN 查询、频繁值查询和区域查询.该方法研究的是近似的 MEDIAN 查询,而非精确 MEDIAN 计算,而且也不适用于普通的 QUANTILE 查询.

3 问题定义

在由 N 个传感器节点构成的无线传感器网路中,每个采样期间每个传感器节点都会产生一个感应数据,结果每个采样期间整个无线传感器网络内节点的感应数据就构成了一个含有 N 个感应数据的数据集合 Q.定义 1 给出了中位数的形式化定义:

定义 1(中位数):设 Q 是一个含有 N 个值的数据集合,对 Q 进行排序后得到一个有序数据集合 P, $P = \{P_1, P_2, P_3, \cdots, P_N\}$.用 $MPlace$ 表示中位数位置,用 $MValue$ 表示中位数的值, $MPlace = \dfrac{N+1}{2}$.

当 N 是奇数时, $MValue = P_{MPlace} = P_{(N+1)/2}$;

当 N 是偶数时, $MValue = \dfrac{P_{N/2} + P_{N/2+1}}{2}$.

4 HFMA 算法

在这一节我们首先描述 HMA 算法,然后基于 HMA 算法进行改进提出 HFMA 算法,最后给出相关分析.

4.1 HMA 算法

HMA 算法的基本思想是利用直方图可以确定中位数落在哪个区间,然后只用收集区间内的数据,从而减少需要传输的数据.

HMA 算法分为两个阶段,初始化阶段和连续查询阶段.

初始化阶段基站收集所有节点的数据,排序后建立等高直方图.等高直方图建立好后,基站将等高直方图的区间下发至每个节点.

连续查询阶段分为 3 步.第一步是直方图建立阶段,从叶子节点开始按照区间划分建立局部直方图.中间节点合并自己孩子节点的局部直方图并传给父节点.基站在收到全局直方图后,据此计算中位数落在哪个区间中.第二步是数据采集阶段.基站通知该区间的节点传输数据.落在区间中的数据将数据传送至基站处.第三步是计算阶段,基站根据全局直方图和收到的数据计算出中位数并将结果返回给用户.

对于一次查询,执行一次就可以了.对于连续查询,首先完成初始化阶段,然后重复执行连续查询阶段即可.

4.2 HFMA 算法

我们发现 HMA 算法连续查询阶段中,当前轮次的数据采集阶段中的数据采集区间和上一轮次的数据采集区间在很多时候是相同的.这是因为数据采集的时间相关性造成的.如果我们在每一轮次中不再去建立直方图,而是直接用上一轮次的数据采集区间来采集数据将取得更高的效率.但是这样会带来新的问题,虽然有了采集区间内的数据,但是由于不知道区间外的数据分布情况,并不知道区间内的哪一个数据会成为中位数.于是我们提出影响因子这一概念,用来对数据采集区间外的数据对中位数造成的影响进行评估.

下面描述 HFMA 算法.首先执行等高直方图一次查询算法.但是在第二次,不去建立等高直方图,而是直接用第一次的数据采集区间.

为了描述方便,把一次查询中的数据采集区间左右两个端点的感应数据值分别记作 F_{low} 和 F_{high},则该区间可以表示为一个大小固定的过滤器 $[F_{low}, F_{high}]$.它会把每个传感器节点的感应数据空间分成左区间 $(-\infty, F_{high})$,中区间 $[F_{low}, F_{high}]$ 和右区间 $(F_{high}, +\infty)$.每个节点根据过滤器确定并记录下自己的原始采样区间.

连续查询阶段分为两个阶段.数据采集阶段和计算阶段.

数据采集阶段中节点采样后根据之前收到的过滤器确定自己的当前采样区间,然后由自己的原始采样区间和当前采样区间确定自己的影响因子(当前采样对最终结果的影响,见表 1).当采样数据落在中区间 $[F_{low}, F_{high}]$ 内时,该节点将采样数据和影响因子一同发送给父节点;当采样数据落在左区间和右区间时,则该节点只向父节点发送影响因子,而不用发送采样数据.

定义 3(中间聚集结果):传感器节点 V_i 的中间聚集结果用 A_i 表示, $A_i = \{S_i, \mu_i\}$.其中 S_i 是其子树上所有落在 $[F_{low}, F_{high}]$ 中的采样数据集合, μ_i 是子树上所有数据的影响因子聚集结果.

表 1 影响因子计算方法

情况	原始采样区间	当前采样区间	传输标志位	影响因子
1	左	左	0	0
2	左	中	1	1/2
3	左	右	0	1
4	中	左	0	-1/2
5	中	中	1	0
6	中	右	0	1/2
7	右	左	0	-1
8	右	中	1	-1/2
9	右	右	0	0

定义 4(HFMA 聚集):节点 V_i 在收到其孩子节点 $\{V_j, V_K, \cdots, V_n\}$ 的中间聚集结果后,分别予以合并以形成自己的中间聚集结果.其中 $S_i = S_j \cup S_k \cup \cdots \cup S_n$, μ_i

$= \mu_j + \mu_k + \cdots + \mu_n$.

(1)叶子节点 V_l 的处理

叶子节点 V_l 首先判断自己的当前采样区间,然后根据原始采样区间和当前采样区间来确定影响因子,并将影响因子结果放入 μ_l;如果当前采样区间为中(采样数据落在区间$[F_{low}, F_{high}]$内),则将采样数据放入到 S_l,否则 S_l 为空;将 A_l 上传给父节点.

(2)中间节点 V_i 的处理

(a)V_i 在收到孩子节点的中间聚集结果后,执行 HFMA 聚集;

(b)V_i 重复执行(a),直到所有孩子都处理完毕,最终形成自己的 A_i;

(c)中间节点判断自己的当前采样区间,然后根据原始采样区间和当前采样区间来确定影响因子,并将影响因子结果与 μ_i 合并;如果当前采样区间为中,则将采样数据放入到 S_i;

(d)将 A_i 上传给自己的父节点.

最终,基站收集到所有落在中区间$[F_{low}, F_{high}]$内的感应数据(设为 M 个)以及所有影响因子的聚集值,将这 M 个数据从小到大排序,然后根据定义(1)计算出 *MPlace*,最后根据公式 1 计算出 *MValue*.

$$MPlace = \frac{M+1}{2} + \mu \tag{1}$$

下面我们对 HFMA 算法进行一个简单的分析.由于相邻采样期间的多个感应数据的变化是具有累积性的,即可以通过多次单个感应数据的变化而得到.因此我们只要证明相邻采样期间只有单个感应数据变化时 HAMA 算法的正确性,同时也就证明了相邻采样期间多个感应数据变化时 HAMA 算法的正确性.

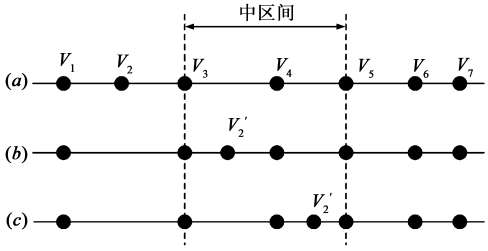


图1　采样数据变化图

从表 1 可以看出我们将采样数据的变化分为了 9 种情况,情况 1、5、9 分别表示采样区间没有发生变化,所以影响因子为 0.下面我们以情况 2 为例进行分析,情况 2 中原本在左区间的采样数据变到中区间,该数据可以落在 *MValue* 前,也可以落在 *MValue* 后.如图 1(a)表示原始采样情况,可以看到 V_4 为中位数,中区间为$[V_3, V_5]$.假设 V_2 发生改变,从左区间变到中区间且在 V_4 前,如图 1(b)所示.我们可以看到中位数实际上并不会发生改变,还是 V_4,但是由于中区间中多了一个节点且在 V_4 前,如果直接对过滤器中的数据求中位数,*MPlace* 会从 V_4 向左偏移 1/2,所以影响因子应该为 1/2.假设 V_2 发生改变,从左区间变到中区间且在 V_4 后,如图 1(c)所示.我们可以看到中位数应该是 V_4 向右偏移 1 位,由于中区间中多了一个节点且在 V_4 后,如果直接对过滤器中的数据求中位数,*MPlace* 会从 V_4 向右偏移 1/2,所以影响因子也应该为 1/2.其他情况也是类似的,在此就不一一证明了.

4.3　HMA 中等高直方图区间的设置

在这一节我们考虑等高直方图区间的设置.从算法中我们很容易看出,节点能量消耗分为两部分,构建直方图耗费的能量和数据采集阶段耗费的能量.

假设传感器网络中有 S 个节点,等高直方图有 K 个数据区间,则将会有$\frac{S}{K}$个节点的采样数据会落在每个数据区间.假设节点消息包除包头、包尾和检验外的可用容量为 M.

用 *COST*1 表示建立直方图的代价,则

$$COST1 = \left\lceil \frac{K}{M} \right\rceil * S$$

用 *COST*2 表示数据采集阶段的代价,由于落在每个数据区间中的节点数为$\frac{S}{K}$,这也意味着在数据采集阶段每一轮将收集约$\frac{S}{K}$个节点的采样数据:

$COST2 = Cost(\frac{S}{K})$每次查询的代价为:

$COST(HMA) = COST1 + COST2$

直观上看,直方图分得越细,则在数据采集阶段需要传输的数据就越少;但是每轮去构建直方图也是要耗费能量的,直方图分得越细,构建直方图耗费的能量就越多,所以需要找到一个平衡.

实验 3 对此进行了分析.

4.4　HFMA 中数据采集区间的更新策略

在 HFMA 算法中,由于我们每轮并不去建立直方图,而是直接收集落在过滤器$[F_{low}, F_{high}]$中的数据,当新一轮采样中的 Median 值 *MValue* 落在过滤器$[F_{low}, F_{high}]$之外时,就无法计算出 *MValue* 了.假设落在过滤器中的节点个数为 M,则当 $MPlace < 1$ 或者 $MPlace > M$ 时 HFMA 算法就失效了.当 HFMA 算法失效时我们就需要重新采集数据,并对过滤区间进行更新.

5　实验

在这一节我们对 HMA 算法和 HFMA 算法进行评估.采样、侦听和通讯是决定能量消耗的几个主要因素.由于 Naive、HMA、HFMA 以及文献[6]中的 q-digest 算法所采用的路由结构都是一样的,因而在侦听信道方

面耗费的能量是一样的,而且采样对所有算法来说都是必须的,所以我们主要考虑通讯的代价.我们用以下两个指标来衡量通讯代价,平均每轮发包数和网络生命周期.

平均每轮发包数:平均每个查询周期内的网络发包数.网络生命周期:从网络开始运行到有节点失效时的轮数.

实验采用 OMNet++进行设计,961 个节点分布在 100m×100m 的矩形区域中,节点的通讯半径为 10m,消息包大小为 29byte,直方图的区间设置为 11,查询运行周期为 1000 轮.以上为默认的实验设置,如果没有特别说明,实验都采用以上设置.

实验数据包含 2 个数据集,是采用模拟数据和真实数据混合生成的数据集.数据集 1 中节点的采样符合照度变化模型,数据集 2 中的节点采样符合温度变化模型.

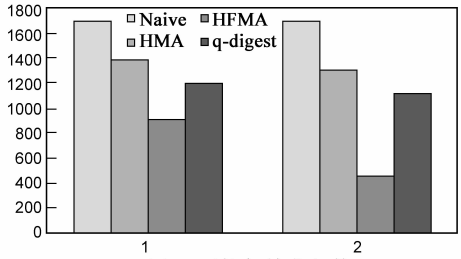


图2 平均每轮发包数

实验 1 比较了算法的平均每轮发包数,如图 3 所示,其中横坐标表示算法采用的数据集,纵坐标表示发包数.从图中可以看出 Naive 算法消耗的能量最多,HFMA 算法最少.HMA 算法每轮只需要收集落在中间直方图区间内的数据,相比较 Naive 算法收集的采样数据会少很多.HFMA 算法由于不必每轮都去建立直方图,所以取得比 HMA 算法更好的性能.对于两种不同的数据集,Naive 算法的性能没有任何变化,因为其总是要收集所有节点的数据;HMA 算法的性能变化也不大,因为建立直方图的代价是一样的,然后需要采集的数据个数也是大致相同的;HFMA 算法的性能变化比较大,在数据集 2 中的表现明显优于数据集 1,经过分析,我们发现由于温度的变化趋势比较小,过滤器有效的轮数会比较多,而照度的变化趋势比较大,导致过滤器的调整

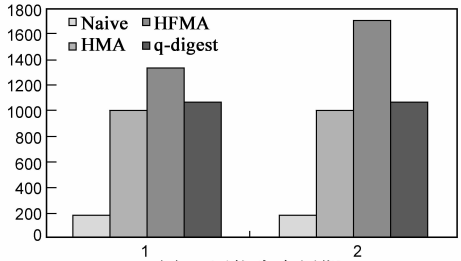


图3 网络生命周期

频率提高,因而会消费更多的能量.

实验 2 比较了算法的网络生命周期.从图中我们可以看到,Naive 算法的生命周期最短,HFMA 具有最长的生命周期.对于 Naive 算法而言,靠近基站的节点由于需要传输子树上所有节点的采样数据,所以能量消耗会很快,当其失效时,网络生命周期也就结束了.HMA 算法的第一阶段是建立全局直方图,在第二阶段时,需要采集的数据已经大大减少,因而靠近基站的节点的负担会减少很多,从而也提高了网络生命周期.HMA 和 HFMA 在能量消耗方面更均衡,所以有更长的网络生命周期.

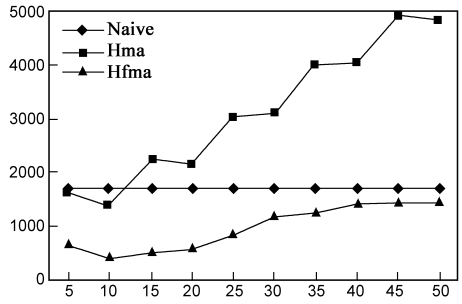


图4 直方图区间个数的影响

实验 3 比较了在不同的直方图区间情况下的算法性能,横坐标表示直方图的区间个数,纵坐标表示平均每轮发包数.从图中可以看出,HMA 算法随着直方图区间的增加平均发包数也增加了.这是由于消息包的容量是有限的,当直方图区间的个数增加时,导致构造全局直方图的能量消耗也大大增加.当然直方图区间个数也不是越小越好,因为区间小意味着落在数据采集区间中的数据多,将消耗更多的能量.HFMA 在直方图区间个数为 10 的时候达到最好效果,当直方图区间个数继续增多的时候,发送数据包呈增多趋势,这是由于当区间个数增多虽然导致了落在数据区间中的数据变少,但是过滤器失效速度却加快了,频繁调整过滤器导致浪费了更多的能量.

6 总结

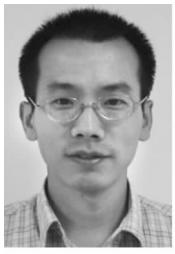
本文研究了无线传感器网络中连续的、精确的中位数查询.首先我们提出了基于等高直方图的中位数查询算法 HMA;然后我们对其进行了扩展,提出了结合直方图与过滤器的 HFMA 算法.实验表明 HFMA 算法可以有效的节省能量开销,提高网络生命周期.

参考文献:

[1] 李建中,李金宝,石胜飞.传感器网络及其数据管理的概念、问题与进展[J].软件学报,2003,14(10):1717-1727.
Li Jian-zhong, Li Jin-bao, Shi Sheng-fei. Concepts, issues and advance of sensor networks and data management of sensor net-

works[J]. Journal of Software, 2003, 14(10): 1717 – 1727. (in Chinese)
[2] 蔚赵春,周水庚,关佶红. 无线传感器网络中数据存储与访问研究进展[J]. 电子学报, 2008, 36(10): 2001 – 2010. Wei Zhao-chun, Zhou Shui-geng, Guan Ji-hone. Data Storage and Acce ss in Wirele ss Sensor Networks : A Survey[J]. Acta Electronica Sinica, 2008, 36(10): 2001 – 2010. (in Chinese)
[3] S. Madden, M J Franklin, J M Hellerstein, W Hong. Tinydb: an acquisitional query processing system for sensor networks[A]. In Proceedings of ACM TODS[C]. New York: ACM Press, 2005. 122 – 173.
[4] A. Demers, J. Gehrke, R. Rajaraman, N. Trigoni and Y. Yao. The Cougar Project: A Work-in-Progress Report[J]. Acm Sigmod Record, 2003, 32(4), 53 – 59.
[5] Cuiping Li, Gao Cong, KumHoe Tung, Shan Wang. Incremental Maintenance of Quotient Cube for Median[A]. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining[C]. New York: ACM Press, 2004. 226 – 235.
[6] Nisheeth Shrivastava, Chiranjeeb Buragohain, Divyakant Agrawal, Subhash Suri. MEDIANs and beyond: new aggregation techniques for sensor networks[A]. In proceedings of the 2nd International Conference on Embedded Networked Sensor Systems[C]. New York: ACM Press, 2004. 239 – 249.

作者简介:



吴中博　男,1980 年生于湖北襄樊,中国人民大学信息学院在读博士生,主要研究领域为数据库,数据仓库,传感器网络.
E-mail: rucwzb@163.com

张　辉　男,1985 年生于湖北随州,中国人民大学信息学院在读硕士生,主要研究领域为数据库,数据仓库,传感器网络.