软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(Suppl.(1)):108−119

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

无线传感器网络中近似加权聚集算法*

郑 旭+, 李建中

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Approximate Aggregation Algorithm for Weighted Data in Wireless Sensor Networks

ZHENG Xu+, LI Jian-Zhong

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: zhengxuhit@gmail.com

Zheng X, Li JZ. Approximate aggregation algorithm for weighted data in wireless sensor networks. *Journal of Software*, 2012,23(Suppl.(1)):108−119 (in Chinese). http://www.jos.org.cn/1000-9825/12012.htm

Abstract: In wireless sensor networks, a weighted aggregation is an important method for the users to obtain the information when monitoring the environment. This method enforces the objectivity of the aggregation result by assigning different weights to different nodes or sensed data. On the other hand, the WSNs are both energy constraint and unstable, so it is better to do the approximate data aggregation if one can ensure the error-bound is tolerable for the users. A group-based sampling algorithm for approximate weighted aggregation is proposed. The theoretical analysis demonstrates that the proposed algorithm can reach arbitrary precision. Furthermore, the proposed algorithm is scalable, and it can adapt to large-scale dynamic sensor networks, and support the modification of the precision during the processing of a query. Experimental results show the correctness of the proposed algorithm and demonstrates the high performance of the proposed algorithm by comparing it with previous algorithms.

Key words: wireless sensor network; data aggregation; weighted computation; sampling algorithm; approximate query

摘 要: 在无线传感器网络中,加权聚集操作是用户获取检测对象信息的重要手段,这一操作通过赋予各个节点或感知数据不同的权值来确保聚集结果更加真实地反映监测对象.另一方面,考虑到能量的限制、网络的不稳定性,如果能保证误差在用户容忍的范围内,近似加权聚集更加适用于传感器网络.针对感知数据的近似加权聚集问题,提出了一种基于分组抽样的(ε,δ)-近似算法,理论证明算法可以达到任意的精度要求.同时,提出的算法具有良好的扩展性,可以适用于大规模、动态变化的传感器网络,并且支持查询过程中的精度调整.仿真实验验证了算法的正确性,并且通过和已有算法比较证明了所提出算法的高效性.

关键词: 无线传感器网络;数据聚集;加权计算;抽样算法;近似查询

传感器节点具有感知、计算、通信的能力,可以收集周围环境的信息[1],而大量节点组成的无线传感器网络,可以收集并返回一片监测区域的信息,实现对监测区域的监控,例如军事侦察、交通流量监测、自然环境监测

* 基金项目: 国家自然科学基金(61033015, 60933001, 61190115)
收稿时间: 2012-05-05; 定稿时间: 2012-08-17

等[2].

在许多应用中,相对于单个节点的感知数据,所有节点感知数据的聚集结果具有更重要的意义[3],例如自然环境监测中监测区域的平均温度[4]、生态环境监测中监测区域出现的动物总数[5].已经有大量的工作考虑了如何在传感器网络中实现精确或近似的 COUNT,SUM,AVG,MAX/MIN[9],Top-k[11,12]等聚集操作.精确的聚集查询多数都基于聚集树进行,通过网内聚集减少总的通信开销;而针对近似聚集的研究,大都基于时空相关性[6−8]或抽样统计理论[9,10].然而,这些聚集结果有时并不足以反映监测区域的真实情况,例如下面的情况:

平均温度是环境监测中一个具有重要意义的参数,假设一个节点不均匀分布的无线传感器网络,包含 A,B 两个面积相等的区域,A 区域的节点数为 B 区域的两倍.现在要计算监测区域的平均温度,若此时 A 区域节点的感知数据平均值为 20°C,B 区域为 10°C,则计算出的平均温度大约为 16.7°C,然而真实的平均温度应当接近 15°C.

针对上述情况,如果给每个感知数据一定的权值后进行数据聚集,能得到更加真实的监测环境信息,即进行加权的数据聚集[13],例如令 B 区域节点的权值为 2,A 区域节点的权值为 1,则计算出的平均温度接近 15°C.加权数据聚集被应用在空间数据聚集[14,15]、事件监测[16]等问题中,保证聚集查询返回更加真实的环境信息.文献[14]比较了 3 种常见的权值定义方法,通过理论和实验分析证明了维诺图是最优方案.然而执行具体聚集查询时,算法需要所有节点参加计算,会大大缩短网络的寿命.文献[16]定义在事件监测中每一个节点的权值为其后继节点个数,在执行聚集查询时同样需要所有的节点都参加计算,不适用于能量有限的传感器网络.

基于时间相关性[7]和空间相关性[6]的数据聚集方法可以应用于加权聚集查询,这些算法通过建立各自的模型为每个节点设置一个值域窗口,汇聚节点通过模型预测节点感知数据,计算出近似的聚集结果,只有当节点的新感知数据超出窗口时才需要向汇聚节点传送.这类方法的优点在于利用了传感器网络固有的时空相关性,减小了发送感知数据的节点个数,避免了相似数据传输引起的不必要的能量开销;缺点在于当精度要求变化时通常需要额外的能量开销来调整或者重建模型,并且在网络规模扩大时维护预测模型的成本不断增加,不适用于大规模的传感网.同时,在误差减小到一定值以后,这些模型的效果会迅速降低.

文献[9]提出了可以满足任意(ε,δ)-近似要求的聚集查询处理算法,汇聚节点在收到用户的查询请求后,根据精度要求和网络规模等参数计算出抽样概率,每个节点按照抽样概率确定是否参加聚集计算,汇聚节点根据抽样到的数据计算出近似的聚集结果.这类方法的优点在于执行过程简单,每个节点只需要记录抽样概率,可以适用于大规模的传感器网络.同时,当查询精度要求变化时能够迅速适应,不需要额外的时间和能量来调整或重建预测模型.然而这种方法只考虑了普通的数据聚集,在加权数据聚集中,不同权值的节点,其感知数据对误差会有不同的影响[17],利用这一点可以进一步减小聚集计算所需的节点规模,然而文献[9]中的算法没有考虑权值对于计算结果准确性的影响,当处理加权数据聚集查询时不能够进一步减小每个节点的抽样概率.

基于上述原因,在本文中,我们提出了可以满足任意精度要求的近似加权数据聚集算法——分组均匀抽样算法,我们的算法不需要单个节点的权值信息,适用于大规模网络中连续的聚集查询.本文的主要贡献如下:

1) 给出了分组均匀抽样用于计算加权数据近似聚集的定义和结果的无偏性,证明了分组均匀抽样比普通均匀抽样具有更好的近似效果,给出了保证聚集结果满足(ε,δ)近似的抽样概率的计算方法.

2) 给出了计算加权和、加权平均值的算法,算法能够适应节点的权值发生变化的情况,并且能够在查询执行过程中动态调整精度要求.

3) 通过分析和实验,验证了算法的正确性,通过和已有算法的比较证明了算法的高效性.

本文第 1 节给出近似加权聚集计算问题的形式化定义.第 2 节给出分组抽样算法正确性和高效性的相关证明.第 3 节介绍分组抽样算法的详细执行过程,给出算法的相关性能分析和证明.第 4 节通过实验验证分组抽样算法的正确性、可扩展性和高效性.第 5 节为结束语.

1 问题定义

1.1 网络模型

假设一个由 N_t+1 个节点组成的传感器网络,各节点的编号集合为 N={0,1,2,…,N_t},其中 0 号节点为汇聚节点,汇聚节点不产生感知数据.存在一棵以汇聚节点为根的最小生成树,所有节点均可以通过最小生成树上的路径将数据传输到汇聚节点,汇聚节点也可以通过最小生成树(广播树)将查询请求分发到所有节点,假设所有节点均以广播的方式发送数据.节点保持时钟同步,每隔 T_0 时间产生新的感知数据.在 t 时刻,网络中各点的感知数据集合为 S_t={$S_{t,1}$,$S_{t,2}$,…,$S_{t,Nt}$},简记为 S_t={S_1,S_2,…,S_{Nt}}.本文提出的方法适用于一般条件下的传感器网络,上述的假设可以使算法的表述更加简明.

1.2 相关定义

定义 1. 在 t 时刻,网络中各点的感知数据集合为 S_t={$S_{t,1}$,$S_{t,2}$,…,$S_{t,Nt}$},则 S_t 的加权和记为 $Sum(S_t)=\sum_{i=1}^{Nt}V_i\times S_{t,i}$, 其中,$V$={$V_1$,$V_2$,…,$V_{Nt}$}为各个节点在当前聚集查询中的权值.

定义 2. 在 t 时刻,网络中各点的感知数据集合为 S_t={$S_{t,1}$,$S_{t,2}$,…,$S_{t,Nt}$},则 S_t 的加权平均值记为 $Avg(S_t)=\frac{Sum(S_t)}{Sum(V)}=\frac{\sum_{i=1}^{Nt}V_i\times S_{t,i}}{\sum_{i=1}^{Nt}V_i}$,其中,$\sum_{i=1}^{Nt}V_i$ 为各个节点在当前聚集查询中权值的和.通常权值是不变的,或者少量节点的权值变化对权值和的影响可以忽略,因此权值的和可以视为一个常数.

定义 3. 随机变量 $\overline{I_t}$ 是 I_t 的一个估计,则 $\overline{I_t}$ 称为 I_t 的一个无偏估计,如果 $E(\overline{I_t})=I_t$, $E(\overline{I_t})$ 是估计 $\overline{I_t}$ 的期望.

定义 4. $\overline{I_t}$ 称为 I_t 的一个(ε,δ)-近似估计,如果 $\Pr\left(\left|\frac{E(\overline{I_t})-I_t}{I_t}\right|\geqslant\varepsilon\right)\leqslant\delta$, $\Pr(X)$表示事件 X 发生的概率.

定义 5. 传感器网络中的分组均匀抽样是按如下过程执行的抽样:

(1) 每个节点根据分组条件确定自己所在的分组;

(2) 每个分组内的节点以相同的概率 P 进行伯努利实验,实验成功则参加抽样,不同分组之间实验成功的概率可以不同,所有参加抽样节点的感知数据通过网内计算,得到对准确聚集结果的一个估计.

注意到,当所有分组的抽样概率 P 均相等时,分组均匀抽样即为针对所有节点的整体(全局)均匀抽样[9],本文余下的内容将以权值大小作为分组条件进行讨论.下面给出以权值大小作为分组条件,通过分组均匀抽样估算加权和的公式和分组方法:

$$\overline{Sum(S_t)}=\sum_{j=1}^{k}\left(\frac{1}{q_j}\sum_{V_i\in G_j}V_i\times S_{t,i}\times X_i\right).$$

公式中,随机 X_i 表示节点 i 是否参加抽样,参加为 1,否则为 0,对于第 j 组(1≤j≤k)的节点 i,X_i (1≤i≤|G_j|),$\Pr(X_i=1)=q_j$,$\Pr(X_i=0)=1-q_j$,并且所有同组或不同组中的 X_i 均相互独立,$E(X_i)=1\times q_j+0\times(1-q_j)=q_j$,$Var(X_i)=q_j\times(1-q_j)$,$q_j$ 为第 j 组的抽样概率.

具体分组方法为:设最大、最小的权值分别为 $V_{\max}$,$V_{\min}$,将节点根据权值大小分为 k 组 G_1,G_2,…,G_k,并且对 $\forall i$,1≤i≤N_t,$\forall j$,1≤j≤k,节点 $n_i\in G_j$ 当且仅当 $V_{\min}+(V_{\max}-V_{\min})\times(j-1)/k\leqslant V_i<V_{\min}+(V_{\max}-V_{\min})\times j/k$,每组的节点个数分别为|$G_1$|,|$G_2$|,…,|$G_k$|,显然,|$G_1$|+|$G_2$|+…+|$G_k$|=$N_t$.

1.3 问题定义

计算感知数据(ε,δ)-近似加权和、加权平均值的问题定义如下:

输入:

S_t={$S_{t,1}$,$S_{t,2}$,…,$S_{t,Nt}$},各个节点在 t 时刻的感知数据;

Agg=*Sum* or *Avg*,聚集操作的类别;

$V=\{V_1,V_2,\ldots,V_{Nt}\}$,各个节点在聚集计算中的权值,初始时由各个节点本地储存;

$\varepsilon(\varepsilon\geqslant 0),\delta(0\leqslant\delta\leqslant 1)$,用户的精度要求;

T,查询持续时间,包含 *m* 个采样周期.

输出:

汇聚节点返回(ε,δ)-近似的 $Sum(S_t)$或者 $Avg(S_t)$的估计值 $\{\overline{I_1},\overline{I_2},\ldots,\overline{I_m}\}$.

权值的具体决定方法不在本文的研究范围之内,这是由于针对不同应用的加权聚集查询,权值的确定方法各不相同,本文主要考察在已经确定了各节点权值的情况下如何高效地完成加权聚集查询.

本文主要使用的符号及其描述见表 1.

Table 1 List of symbols used in this paper

表 1 本文使用的符号及描述

符号	描述
Nt	节点总数(不包含汇聚节点)
$S_{t,i}$	i 节点 t 时刻的感知数据
V_i	i 节点针对当前查询的权值
$V_{\max},V_{\min}$	最大、最小权值
$Sup(S_i),Inf(S_i)$	感知数据上、下界
$G_i,i=1,\ldots,k$	第 i 组节点组成的集合
$\lvert G_i\rvert,i=1,\ldots,k$	第 i 组节点个数
$q_i,i=1,\ldots,k$	第 i 组节点抽样概率
q_0	整体均匀抽样概率
$Sum(S_t),Avg(S_t)$	t 时刻的加权和、加权平均值
$\overline{Sum(S_t)}$	t 时刻加权和的估计
$\overline{Avg(S_t)}$	t 时刻加权平均值的估计
$Sum(V_i)$	所有节点权值的和

2 数学基础

2.1 估值无偏性证明

定理 1. 设分组均匀抽样得到的加权和、加权平均值的估计值分别为 $\overline{Sum(S_t)}$ 和 $\overline{Avg(S_t)}$,则它们是聚集结果的无偏估计,即 $E(\overline{Sum(S_t)})=Sum(S_t),\ E(\overline{Avg(S_t)})=Avg(S_t)$.

证明:

由定义,

$$E(\overline{Sum(S_t)})=E\sum_{j=1}^{k}\left(\frac{1}{q_j}\sum_{V_i\in G_j}V_i\times S_{t,i}\times X_i\right),$$

由于所有的 X_i 均相互独立,有

$$E(\overline{Sum(S_t)})=\sum_{j=1}^{k}\left(\frac{1}{q_j}\sum_{V_i\in G_j}V_i\times S_{t,i}\times E(X_i)\right).$$

由于

$$E(X_i)=1\times q_j+0\times(1-q_j)=q_j,$$

其中,j 是节点 i 所在的分组,所以,

$$E(\overline{Sum(S_t)})=\sum_{j=1}^{k}\sum_{V_i\in G_j}V_i\times S_{t,i}=Sum(S_t).$$

同理,

$$E(\overline{Avg(S_t)})=E\left(\frac{\overline{Sum(S_t)}}{Sum(V)}\right)=\frac{Sum(S_t)}{Sum(V)}=Avg(S_t).$$

□

2.2 抽样方案有效性证明

整体均匀抽样是特殊情况下的分组均匀抽样,即所有分组有相同的抽样概率,然而针对加权数据聚集,整体均匀抽样并非最有效(给定精度要求,所需样本规模最小)的抽样方法.由文献[17]可知,假设 $\overline{I_1}$ 和 $\overline{I_2}$ 都是对某个总体参数 I_t 的无偏估计,并且 $\overline{I_1}$ 的方差小于 $\overline{I_2}$ 的方差,记为 $Var(\overline{I_1})<Var(\overline{I_2})$,则 $\overline{I_1}$ 是一个更有效的估计.下面的定理证明了分组均匀抽样处理加权数据聚集时更加有效.

定理 2. 若以权值大小作为分组条件,则存在一个 $q_1,q_2,\ldots,q_k$ 不全相等的抽样,$q_1|G_1|+q_2|G_2|+\ldots+q_k|G_k|=q_0N_t$,并且 $Var(\overline{Sum(S_t)}_{q_1,q_2,\ldots,q_k})\leqslant Var(\overline{Sum(S_t)}_{q_0})$.

证明:由方差定义,

$$Var(\overline{Sum(S_t)})=Var\left(\sum_{j=1}^{k}\left(\frac{1}{q_j}\sum_{V_i\in G_j}V_i\times S_{t,i}\times X_i\right)\right)=\sum_{j=1}^{k}\left(\frac{1}{q_j^2}\sum_{V_i\in G_j}V_i^2\times S_{t,i}^2\times Var(X_i)\right).$$

令:

(1) $q_2=q_3=\ldots=q_{k-1}=q_0$,

(2) $q_1|G_1|+q_k|G_k|=q_0(|G_1|+|G_k|)$,

设 $q_1=h\times q_k$,$|G_k|=m\times|G_1|$,其中,$h\neq1$,$m>0$,$q_1\leqslant1$,$q_k\leqslant1$;

因为 $q_1|G_1|+q_k|G_k|=q_0(|G_1|+|G_k|)$,有 $q_0=(q_1|G_1|+q_k|G_k|)/(|G_1|+|G_k|)$,

又 $V_{\max}/V_{\min}>1$,所以有 $V_{\max}=a\times V_{\min}$,$a>1$,则由方差定义,

$$\begin{aligned}
&Var(\overline{Sum(S_t)}_{q_0})-Var(\overline{Sum(S_t)}_{q_1,q_2,\ldots,q_k})\\
&=\frac{1-q_0}{q_0}\left(\sum_{V_i\in G_1}V_i^2\times S_{t,i}^2+\sum_{V_i\in G_k}V_i^2\times S_{t,i}^2\right)-\frac{1-h\times q_k}{h\times q_k}\sum_{V_i\in G_1}V_i^2\times S_{t,i}^2-\frac{1-q_k}{q_k}\sum_{V_i\in G_k}V_i^2\times S_{t,i}^2\\
&=\left(\frac{m+1-q_k\times m+h\times q_k}{q_k\times m+h\times q_k}-\frac{1-h\times q_k}{h\times q_k}\right)\sum_{V_i\in G_1}V_i^2\times S_{t,i}^2+\left(\frac{m\times+1-q_k\times m+h\times q_k}{q_k\times m+h\times q_k}-\frac{1-q_k}{q_k}\right)\sum_{V_i\in G_k}V_i^2\times S_{t,i}^2\\
&=\frac{(h-1)\left(m\sum_{V_i\in G_1}V_i^2\times S_{t,i}^2-h\sum_{V_i\in G_k}V_i^2\times S_{t,i}^2\right)}{(m+h)\times h\times q_k}.
\end{aligned}$$

令 $h<1$,若能够假设 G_1 和 G_k 中的节点随机分布在网络中,则有 $m\sum_{V_i\in G_1}S_{t,i}^2\approx\sum_{V_i\in G_k}S_{t,i}^2$,此时,若

$$h>\frac{m\sum_{V_i\in G_1}V_i^2\times S_{t,i}^2}{\sum_{V_i\in G_k}V_i^2\times S_{t,i}^2}\Rightarrow h\geqslant\frac{m\sum_{V_i\in G_1}V_{\min}^2\times S_{t,i}^2}{a^2\sum_{V_i\in G_k}V_{\min}^2\times S_{t,i}^2}=\frac{m\sum_{V_i\in G_1}S_{t,i}^2}{a^2\sum_{V_i\in G_k}S_{t,i}^2}.$$

因为 $m\sum_{V_i\in G_1}V_i^2\times S_{t,i}^2-h\sum_{V_i\in G_k}V_i^2\times S_{t,i}^2<0$,所以 $\frac{1}{a^2}\leqslant h<1$ 时结论成立,此时,$q_1\neq q_k$.

若上述假设不成立,则

$$h>\frac{m\sum_{V_i\in G_1}V_i^2\times S_{t,i}^2}{\sum_{V_i\in G_k}V_i^2\times S_{t,i}^2}\Rightarrow h\geqslant\frac{m\sum_{V_i\in G_1}sup(S_i)^2}{a^2m\sum_{V_i\in G_1}inf(S_i)^2}.$$

当 $V_{\max}/V_{\min}\geqslant sup(S_i)/inf(S_i)$ 时,$q_i'/q_j'=\sqrt{\frac{|G_j|\sum_{V_i\in G_1}V_i^2S_{t,i}^2}{|G_i|\sum_{V_i\in G_j}V_i^2S_{t,i}^2}}$,所以 $\frac{sup(s_i)^2}{a^2\times inf(s_i)^2}\leqslant h<1$ 时结论成立,此时,$q_1\neq q_k$.

同理,当 $V_{\max}/V_{\min} < sup(S_i)/inf(S_i)$ 时,令 $h>1$,有 $1<h\leqslant \frac{sup(s_i)^2}{a^2 inf(s_i)^2}$ 时结论成立,此时,$q_1\neq q_k$. □

综上所述,存在一个不全相等的抽样概率组合 $q_1,q_2,\ldots,q_k, q_1|G_1|+q_2|G_2|+\ldots+q_k|G_k|=q_0N_t$,并且 $Var(\overline{Sum(S_t)}) < Var(\overline{Sum(S_t)}_{q_0})$.

上述定理说明,在样本规模的期望值相同的情况下,分组均匀抽样通常可以给出比全局均匀抽样更优化的估计,进而说明对于相同的精度要求,分组均匀抽样可以用更小的样本规模实现.以空间数据聚集为例,空间数据聚集主要应用于节点分布不均匀的传感器网络,每个节点的权值为其维诺图的面积[14],由于节点分布不均匀,不同节点维诺图的面积存在差异,因此一定存在一个比整体均匀抽样更加高效的分组均匀抽样方案.

3 分组抽样算法

3.1 分组抽样概率确定方法

下面的引理证明在总样本容量的期望值一定的情况下,如何获得最有效的加权和估计:

引理 1. 设 $q_1,q_2,\ldots,q_k$ 是分组均匀抽样中各组的抽样概率,并且 $q_1|G_1|+q_2|G_2|+\ldots+q_k|G_k|=q_0N_t$,则对于任意的 $q_1,q_2,\ldots,q_k$, $Var(\overline{Sum(S_t)})_{q_1,q_2,\ldots,q_k} \geqslant Var(\overline{Sum(S_t)})_{q'_1,q'_2,\ldots,q'_k}$,其中,$q'_1|G_1|+q'_2|G_2|+\ldots+q'_k|G_k|=q_0N_t$,并且对任意的 q'_i 和 q'_j,有

$$q'_i/q'_j=\sqrt{\frac{|G_j|\sum_{V_i\in G_1}V_i^2S_{t,i}^2}{|G_i|\sum_{V_i\in G_j}V_i^2S_{t,i}^2}}.$$

证明:

假设存在一组抽样概率 $q_1,q_2,\ldots,q_k$,对其中存在的任意 q_i 和 q_j,

$$q_i/q_j\neq\sqrt{\frac{|G_j|\sum_{V_i\in G_1}V_i^2S_{t,i}^2}{|G_i|\sum_{V_i\in G_j}V_i^2S_{t,i}^2}}.$$

令

$$Var(\overline{Sum(S_t)})_{q_i,q_j}=\left(\frac{1}{q_i}-1\right)\sum_{V_i\in G_i}V_i^2S_{t,i}^2+\left(\frac{1}{q_j}-1\right)\sum_{V_j\in G_j}V_i^2S_{t,i}^2,$$

则存在一组 q'_i 和 q'_j,满足 $q'_i/q'_j=\sqrt{\frac{|G_j|\sum_{V_i\in G_1}V_i^2S_{t,i}^2}{|G_i|\sum_{V_i\in G_j}V_i^2S_{t,i}^2}}$,并且 $q_i|G_i|+q_j|G_j|=q'_i|G_i|+q'_j|G_j|$.此时,

$$Var(\overline{Sum(S_t)})_{q_i,q_j}-Var(\overline{Sum(S_t)})_{q'_i,q'_j}=\frac{1}{q_i}\sum_{V_i\in G_i}V_i^2S_{t,i}^2+\frac{1}{q_j}\sum_{V_i\in G_j}V_i^2S_{t,i}^2-\frac{1}{q'_i}\sum_{V_i\in G_i}V_i^2S_{t,i}^2-\frac{1}{q'_j}\sum_{V_i\in G_j}V_i^2S_{t,i}^2.$$

令 $k_i=\sqrt{\sum_{V_i\in G_i}V_i^2S_{t,i}^2}$, $k_j=\sqrt{\sum_{V_i\in Gj}V_i^2S_{t,i}^2}$, $m_i=|G_i|$, $m_j=|G_j|$,代入上式,得

$$\begin{aligned}Var(\overline{Sum(S_t)})_{q_i,q_j}-Var(\overline{Sum(S_t)})_{q'_i,q'_j}&=\frac{(\sqrt{m_i}k_i)^2}{m_iq_i}+\frac{(\sqrt{m_j}k_j)^2}{m_jq_j}-\frac{(\sqrt{m_i}k_i+\sqrt{m_j}k_j)^2}{m_iq_i+m_jq_j}\\&=\frac{(m_jq_j\sqrt{m_i}k_i-m_iq_i\sqrt{m_j}k_j)^2}{m_iq_im_jq_j(m_iq_i+m_jq_j)}\geqslant 0.\end{aligned}$$

所以 q'_i 和 q'_j 是最优化的抽样概率.又对

$$q_i / q_j = \sqrt{\frac{|G_j| \sum_{V_i \in G_1} V_i^2 S_{t,i}^2}{|G_i| \sum_{V_i \in G_j} V_i^2 S_{t,i}^2}}, q_i / q_k = \sqrt{\frac{|G_k| \sum_{V_i \in G_1} V_i^2 S_{t,i}^2}{|G_i| \sum_{V_i \in G_k} V_i^2 S_{t,i}^2}},$$

有

$$q_j / q_k = \sqrt{\frac{|G_k| \sum_{V_i \in G_j} V_i^2 S_{t,i}^2}{|G_j| \sum_{V_i \in G_k} V_i^2 S_{t,i}^2}},$$

故存在一组最优的抽样概率 $q_1,q_2,...,q_k$. □

由上述定理可知,在分组条件确定的情况下,当给定误差大小时,存在一组抽样概率 $q_1,q_2,...,q_k$ 在保证样本总数的期望值最小的同时达到误差的要求.

由于最优抽样概率随着感知数据变化,需要使用全部的感知数据来计算,丧失了抽样方法的意义,故这里采用近似的比值 $q_i/q_1=V_{\max i}/V_{\max 1}, i=1,2,\ldots,k$,这一近似比易于计算,同时也接近于最优的比值.下面的定理给出根据用户的(ε,δ)近似要求确定各分组抽样概率的方法,其中各个节点按照权值大小进行分组.

定理 3. 当 $q_i / q_1 = V_{\max i} / V_{\max 1}, i = 1,2,...,k$ 且 $q_1 \geqslant \frac{Sup(S_t) \times V_{\max 1} \times \phi_{\delta/2}^2}{N_t \times \overline{V} \times Inf(S_t) \times \varepsilon^2 + Sup(S_t) \times V_{\max 1} \times \phi_{\delta/2}^2}$ 时, $\overline{Sum(S_t)}_{q_1,q_2,...,q_k}$ 是对 $Sum(S_t)$ 的一个(ε,δ)近似,其中 $V_{\max i}=V_{\min}+i\times(V_{\max}-V_{\min})/k$, $\phi_{\delta/2}^2$ 是标准正态分布的$\delta/2$ 临界值, $\overline{V} = Sum(V)/N_t$.

证明:由已知, $Var(\overline{Sum(S_t)}) = \sum_{j=1}^{k} \frac{1-q_j}{q_j} \sum_{V_i \in G_j} V_i^2 S_{t,i}^2$,又 $q_i / q_1 = V_{\max i} / V_{\max 1}, i = 1,2,...,k$,所以,

$$\begin{aligned} Var(\overline{Sum(S_t)}) &= \sum_{j=1}^{k} \left(\frac{V_{\max 1} - V_{\max j} q_1}{V_{\max j} q_1} \sum_{V_i \in G_j} V_i^2 S_{t,i}^2 \right) \\ &\leqslant Sup(S_t) \sum_{j=1}^{k} \left(\frac{V_{\max 1} - V_{\max j} q_1}{V_{\max j} q_1} V_{\max j} \sum_{V_i \in G_j} V_i S_{t,i} \right) \\ &\leqslant Sup(S_t) \times \frac{V_{\max 1} Sum(S_t) - q_1 \sum_{j=1}^{k} \left(V_{\max j} \sum_{V_i \in G_j} V_i S_{t,i} \right)}{q_1} \\ &\leqslant Sup(S_t) \frac{V_{\max 1} Sum(S_t) - q_1 V_{\max 1} Sum(S_t)}{q_1}. \end{aligned}$$

当 $q_1 \geqslant \frac{Sup(S_t) \times V_{\max 1} \times \phi_{\delta/2}^2}{N_t \times \overline{V} \times Inf(S_t) \times \varepsilon^2 + Sup(S_t) \times V_{\max 1} \times \phi_{\delta/2}^2}$ 时,有

$$\frac{\phi_{\delta/2}^2 (Sup(S_t) \times V_{\max 1} - Sup(S_t) \times q_1 V_{\max 1})}{q_1} \leqslant N_t \times \overline{V} \times Inf(S_t) \times \varepsilon^2 \leqslant Sum(S_t) \times \varepsilon^2.$$

所以,

$$\begin{aligned} \phi_{\delta/2}^2 Var(\overline{Sum(S_t)}) &\leqslant \frac{\phi_{\delta/2}^2 (Sup(S_t) \times V_{\max 1} - Sup(S_t) \times q_1 V_{\max 1})}{q_1} \times Sum(S_t) \\ &\leqslant N_t \times \overline{V} \times Inf(S_t) \times \varepsilon^2 \times Sum(S_t) \\ &\leqslant Sum(S_t)^2 \times \varepsilon^2. \end{aligned}$$

由中心极限定理,$\overline{Sum(S_{t,G_j})}=\frac{1}{q_j}\sum_{V_i\in G_j}V_i\times S_{t,i}\times X_i$ 满足以 $E(\overline{Sum(S_{t,G_j})})=Sum(S_{t,G_j})$ 为期望、$Var(\overline{Sum(S_{t,G_j})})=\frac{1-q_j}{q_j}\sum_{V_i\in G_j}V_i^2S_{t,i}^2$ 为方差的正态分布,又正态分布的线性加之和仍然是正态分布的,因此,$\overline{Sum(S_t)}_{q_1,q_2,\ldots,q_k}$ 是满足以 $E(\overline{Sum(S_t)})=Sum(S_t)$ 为期望、$Var(\overline{Sum(S_t)})=\sum_{j=1}^{k}\frac{1-q_j}{q_j}\sum_{V_i\in G_j}V_i^2S_{t,i}^2$ 为方差的正态分布,则有 $\Pr(|\overline{Sum(S_t)}-Sum(S_t)|\geqslant\phi_{\delta/2}\sqrt{Var\overline{Sum(S_t)}})\leqslant\delta$ 即 $\Pr\left(\left|\frac{\overline{Sum(S_t)}-Sum(S_t)}{Sum(S_t)}\right|\geqslant\varepsilon\right)\leqslant\delta$, 此时 $\overline{Sum(S_t)}_{q_1,q_2,\ldots,q_k}$ 是对 $Sum(S_t)$的一个(ε,δ)近似.□

本定理给出的抽样概率同样保证(ε,δ)近似的 $Avg(S_t)$估计.

3.2 分组抽样算法

算法分为预处理阶段和查询执行阶段.在预处理阶段,算法收集节点权值的相关聚集信息;在查询执行阶段,汇聚节点根据收集到的信息和用户的精度要求计算出各分组的抽样概率,而后将抽样概率分发到整个网络.在每次抽样时,各个节点根据所在分组的抽样概率决定是否参加此次抽样,参加抽样的节点将数据沿聚集树上传,中间节点将从下层节点收集到的感知数据进行聚集,形成部分聚集结果后转发,最终汇聚节点返回抽样所得的估计值.

3.2.1 预处理阶段

在预处理阶段,算法首先需要得到各节点针对当前查询的最大和最小权值 $V_{\max},V_{\min}$ 以及所有节点权值之和,主要方法有两种,第 1 种为通过查询获得,由于针对某一种查询,节点的权值通常是固定的,故可以查询后将权值存储在汇聚节点,普通的 Min,Max,Sum 查询可以完成这一工作;第 2 种方法为根据经验估计相应的上、下界,例如空间数据聚集查询中,最大权值的上界可以用节点的最大可覆盖面积表示,而权值之和可以用网络覆盖的总面积表示.在网络规模较大的情况下,通过估计得到权值的信息可以节省查询带来的能量开销,但是若估计值距离真实最大、最小权值的偏差过大,则会降低分组均匀抽样的效率.

3.2.2 查询执行阶段

在查询执行阶段,汇聚节点首先根据 $q_1=\frac{Sup(S_t)\times V_{\max 1}\times\phi_{\delta/2}^2}{N_t\times\bar{V}\times Inf(S_t)\times\varepsilon^2+Sup(S_t)\times V_{\max 1}\times\phi_{\delta/2}^2}$ 以及 $q_i/q_j=V_{\max i}/V_{\max 1}$, $i=1,2,\ldots,k$ 计算出各个分组的抽样概率,而后将抽样概率广播到网络中的所有节点,广播的内容为 q_1 和最大、最小权值 $V_{\max},V_{\min}$,每个收到广播的传感器节点 m 根据自身的权值 V_m 计算出相应的抽样概率 $p_m=q_j$, $j=\lfloor(V_m-V_{\min})/(V_{\max}-V_{\min})\rfloor$,其中,$q_j=q_1\times\frac{(k-j)V_{\min}+jV_{\max}}{(k-1)V_{\min}+V_{\max}}$.查询的具体执行过程如下:

每隔时间 t,网络中的传感器节点产生新的感知数据后:

1) 每个节点生成一个[0,1]之间的随机数,任意生成的随机数小于或等于 p_m 的节点 m,其感知数据参加本次聚集计算;

2) 参加计算的叶子节点向上发送部分聚集结果 $\overline{Sum(S_t)}_j=V_j\times S_{t,j}/q_m$;

3) 中间节点在收到所有子节点的数据后(若某个子节点及其后继节点均未参加抽样,则不发送数据),计算出部分聚集结果:

$$\overline{Sum(S_t)}_j=V_j\times S_{t,j}/q_{\mathrm{m}}+\overline{Sum(S_t)}_{j_1}+\ldots+\overline{Sum(S_t)}_{j_m},$$

向上传递;

4) 若中间节点的感知数据不参加聚集计算则计算出部分聚集结果:

$$\overline{Sum(S_t)}_j=\overline{Sum(S_t)}_{j_1}+\ldots+\overline{Sum(S_t)}_{j_m},$$

向上传递;

5) sink 节点得到本次聚集结果的估计;

$$\overline{Sum(S_t)}=\overline{Sum(S_t)}_{j_1}+...+\overline{Sum(S_t)}_{j_m} \text{ 或 } \overline{Avg(S_t)}=\frac{\overline{Sum(S_t)}_{j_1}+...+\overline{Sum(S_t)}_{j_m}}{Sum(V_i)},$$

加入输出队列 $\{\overline{I_1},\overline{I_2},...\}$;

6) 当查询执行结束后,返回 $\{\overline{I_1},\overline{I_2},...,\overline{I_m}\}$.

3.3 算法性能分析

设传感器网络中建立的聚集树的深度为 h,树中每个非叶子节点的平均子节点个数为 k,在每次抽样时,每个节点生成随机数的代价为 $O(1)$,计算部分聚集结果的代价为 $O(1)$,所有节点都需要生成随机数,而节点需要参与部分聚集结果计算的概率则与它在聚集树中的层次有关,假设汇聚节点为 0 层,最深的叶节点为 h 层,则对于 a 层的节点 i,需要计算部分聚集结果的概率约为 $1-(1-\overline{q})^{k^{h+1-a}}$,其中 $\overline{q}=\frac{q_1|G_1|+...+q_k|G_k|}{N_t}$,又第 a 层的节点个数约为 k^a,由于 $(1-\overline{q})^{k^{h+1-a}}$ 随着 a 的减小接近于 0,故这里不考虑不同层之间的相关性,总的需要计算部分聚集结果的节点个数为 $\sum_{a=1}^{h}k^a(1-(1-\overline{q})^{k^{h+1-a}})=O(N_t\overline{q})$,计算的代价为 $O(N_t+N_t\overline{q})=O\left(\frac{N_t\varepsilon^2+\phi^2_{\delta/2}}{\varepsilon^2}\right)$.同理,若用发送数据的节点个数来描述通信代价,则通信代价为 $O(N_t\overline{q})=O\left(\frac{\phi^2_{\delta/2}}{\varepsilon^2}\right)=O\left(\frac{1}{\varepsilon^2}\ln\frac{1}{\delta}\right)$.

需要注意到,我们的方法与单个节点的权值信息无关,因此可以适用于权值在查询过程中发生变化的情况,这是由于以下两个原因:(1) 权值通常在最大值和最小值之间变化,因此权值变化的节点只需要重新计算自身参加抽样的概率;(2) 对于大规模的传感器网络,单个节点的权值变化对于所有节点权值之和的影响很小.此时得到的聚集结果仍可以认为是满足(ε,δ)-近似的.

当查询的误差要求发生变化或者有多个查询到来时,算法只需要将新的抽样概率或者精度要求最高的查询对应的抽样概率广播到网络中即可,不需要其他能量开销.

4 性能分析与实验评价

本节将本文的算法与文献[9]中提出的采样算法进行比较,模拟实验中用到的感知数据为 Berkeley Intel 实验室的传感器网络测得的真实环境温度数据[18].

本节主要考察算法在 4 种权值分布情况下的表现:(1) 各个节点的权值均匀分布在最小、最大权值之间(均匀分布);(2) 各个节点的权值构成一个近似的正态分布,分布的均值为权值区间的中点(正态分布);(3) 大部分节点的权值较小,接近最小权值(低权值倾斜);(4) 大部分节点的权值较大,接近最大权值(高权值倾斜).每一种权值的分布都有相对的实际意义,以空间数据聚集为例,上述的权值分布情况分别对应于:节点随机非均匀分布的网络、节点随机均匀分布的网络、节点稠密的网络、节点稀疏的网络,并且假设最小权值 MIN*WEIGHT*=10,最大权值 MAX*WEIGHT*=100,最高温度 25°C,最低温度 15°C.

第 1 组实验用于观察算法在不同精度下的准确性,即算法是否能够满足(ε,δ)近似的要求.模拟采样次数为 10 000 次,考察抽样结果的平均相对误差和不满足(ε,δ)近似的抽样的次数.通过实验可以发现,本文所提出的算法不论是超出ε近似的样本数量还是聚集结果的平均误差,都能满足精度的要求,由图 1 和图 2 可知,在低权值倾斜的网络中,当ε=0.2,δ=0.2 时,不满足ε近似的抽样次数为 624,小于 10000×δ=2000 次,并且平均误差为 8.6%,小于用户要求的 0.2.权值在另外 3 种情况下的实验结果同样满足(ε,δ)近似要求.实验中其他参数为网络规模 N_t=4000,分组数 *GroupNumber*=3.

由图 3 可知,在全部 4 种权值分布中,当分组数增加时,样本规模均减小,但是当分组数大于 6 时,样本规模减小的幅度较缓,例如在权值正态分布的网络中,当分组数从 6 增加到 8 时,平均样本规模只减小了 8.另一方面,过大的分组数会在广播抽样概率的过程中引起额外的通信开销,并且过大的分组会缩小各个分组的样本规模,不

利于用中心极限定理进行参数估计.其他参数 N_t=4000(ε=0.1,δ=0.1).

第 3 组实验比较本文提出的方法和全局均匀抽样方法的效率,考察在不同网络规模下两种方法所需的样本容量.通过实验观察,分组抽样方法在各种情况下平均样本规模都小于全局的均匀抽样,特别是随着网络规模的增加.由图 4 可知,在均匀分布中,随着网络规模的增加,分组均匀抽样的样本规模分别从 581,620,646,662,678 减小到 496,511,522,528,533,而在图 5 中,分组抽样方法优化效果较小,但仍可以将样本规模从 492,519,537,549,562 减小到 488,504,509,514,521.需要注意,虽然样本规模随着网络规模的增加而增加,但抽样的概率却随着网络规模的增加而减小,这保证了算法可以适用于大规模的网络.实验中其他参数(ε=0.1,δ=0.1),*GroupNumber*=3.

第 4 组实验考察本文提出的方法和对比方法在不同精度要求下所需的样本规模.图 6 中,针对均匀分布和高权值倾斜两种情况,当δ=0.16 时,随着 ε 的变化,分组抽样的样本规模始终小于全局均匀抽样,图 7 中,当ε=0.16,ε变化时,分组抽样的样本规模同样小于全局均匀抽样.由此可见,分组抽样方法能够有效地适应不同的精度要求.实验中其他参数 N_t=4000,*GroupNumber*=3.

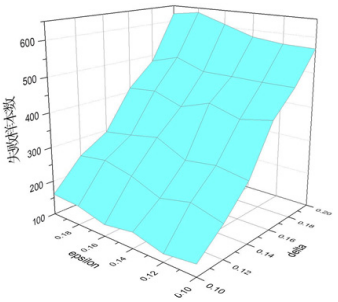

Fig.1 Impact of (ε,δ) on failure number

图 1 低权值倾斜下失败样本数与精确度的关系

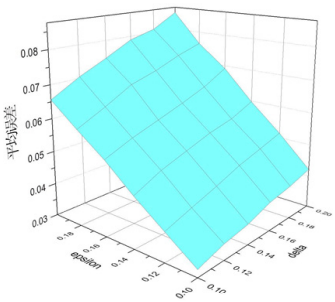

Fig.2 Impact of (ε,δ) on variance

图 2 低权值倾斜下平均误差与精确度的关系

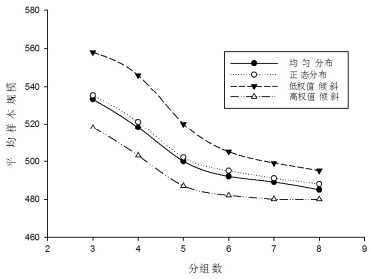

Fig.3 Impact of group number on scale

图 3 样本规模与分组数的关系

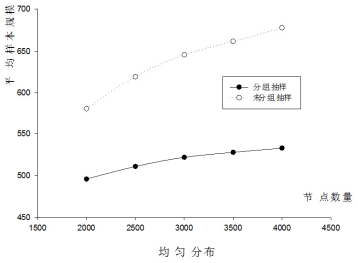

Fig.4 Network scale vs. sample scale

图 4 均匀分布下样本规模与网络规模的关系

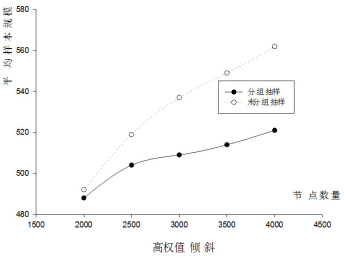

Fig.5 Network scale vs. sample scale

图 5 高权值倾斜下样本规模与网络规模的关系

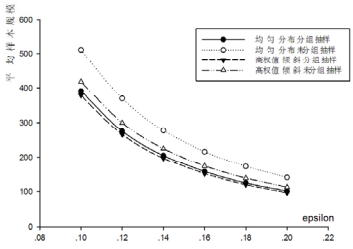

Fig.6 Impact of parameter ε on scale (δ=0.16)

图 6 δ=0.16 时样本规模与ε的关系

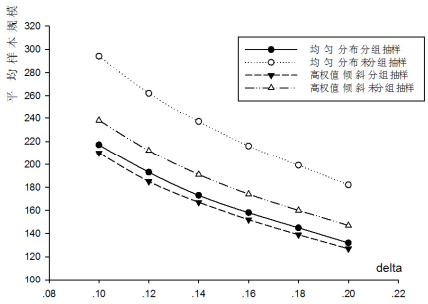

Fig.7 Impact of parameter δ on scale (ε=0.16)

图 7 ε=0.16 时样本规模与δ的关系

5 结束语

在传感器网络的实际应用中,简单的数据聚集有时不能反映物理环境的真实情况,需要对感知数据加权后再进行聚集,而现有的算法并没有关注这一问题.本文对传感器网络中的近似数据加权聚集问题进行了研究,提出了可以满足任意误差要求的抽样方法,并进行了相关的正确性证明和理论分析.通过模拟实验验证了算法的正确性、可扩展性和高效性,能够显著减小通信能量开销,延长网络寿命.

References:

[1] Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E. Wireless sensor networks: A survey. Computer Networks, 2002,38(4): 393−422.

[2] Sun LM, Li JZ, Chen Y, Zhu HS. Wireless Sensor Networks. Beijing: Tsinghua University Press, 2005. 14−16 (in Chinese).

[3] Madden S, Franklin MJ, Hellerstein JM. TAG: A tiny aggregation for ad-hoc sensor networks. In: Proc. of the ACM OSDI. Boston, Massachusetts, 2002. [doi: 10.1145/1060289.1060303]

[4] Polastre J, Szewczyk R, Mainwaring A, Culler D. Analysis of wireless sensor networks for habitat monitoring. Wireless Sensor Networks, 2004,399−423.

[5] Huang ZF, Yi K, Liu YH, Chen GH. Optimal sampling algorithms for frequency estimation in distributed data. In: Proc. of the IEEE INFOCOM. Shanghai, 2011. 1999−2005.

[6] Chu D, Deshpande A, Hellerstein JM, *et al*. Approximate data collection in sensor networks using probabilistic models. In: Proc. of the IEEE ICDE. Atlanta, 2006. 48−59.

[7] Deligiannakis A, Kotidis Y, Rossopoulos N. Processing approximate aggregation queries in wireless sensor networks. Information Systems, 2006,31:770−792.

[8] Olston C, Jiang J, Widom J. Adaptive filters for continuous queries over distributed data streams. In: Proc. of the ACM SIGMOD. San Diego, 2003. 563−574.

[9] Li JZ, Cheng SY. (ε,δ)-Approximate aggregation algorithms in dynamic sensor networks. IEEE Trans. on Parallel and Distributed Systems, 2012,23(3):385−396.

[10] Yu L, Li JZ, Cheng SY. Approximate continuous aggregation via time window based compression and sampling in WSNs. Wireless Sensor Networks, 2010,2:675−682.

[11] Sebastian M, Peter T, Gerhard W. KLEE: A framework for distributed top-*k* query algorithm. VLDB, 2005,648−659.

[12] Bi R, Li JZ, Cheng SY. (ε,δ)-Approximate top-*k* query processing algorithm in wireless sensor networks. Journal on Communications, 2011,32(8):45−54.

[13] Grossman J, Grossman M, Katz R. The First Systems of Weighted Differential and Integral Calculus. Rockport, Mass. : Archimedes Foundation, 1980.

[14] Sharifzadeh M, Shahabi C. Supporting spatial aggregation in sensor network database.In: Proc. of the 12th ACM Int'l Workshop on Geographic Information Systems. Washington, 2004. 166−175.

[15] Sharifzadeh M, Shahabi C. Utilizing Voronoi cells of location data streams for accurate computation of aggregate functions in sensor networks. Geoinformatica, 2006,10(1):9−36.

[16] Jagyasi BG, Chander D, Desai UB, *et al*. On robustness of adaptive weighted aggregation scheme for wireless sensor network. In: Proc. of the 10th Int'l Symp. on Wireless Personal Multimedia Communications. Jaipur, 2007. 472−476.

[17] Bernstein S, BernStein R, Daoji S. Elements of Statistics Ⅱ: Inferential Statistics. Beijing: Science Press, 2002. 214−216.

[18] Intel Lab. dataset. http://db.csail.mit.edu/labdata/labdata.htm

附中文参考文献:

[2] 孙利民,李建中,陈渝,朱红松.无线传感器网络.北京.清华大学出版社,2005.

[12] 毕冉,李建中,程思瑶.无线传感器网络(ε,δ)-近似 Top-k 查询处理算法.通信学报,2011,32(8):45−54.

郑旭(1987－),男,黑龙江哈尔滨人,博士生,主要研究领域为无线传感器网络.

李建中(1950－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为海量数据管理,无线传感器网络,CPS.