ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.8, August 2010, pp.1936−1953
doi: 10.3724/SP.J.1001.2010.03641

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

无线传感器网络中(ε,δ)-近似聚集算法*

程思瑶+, 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)

(ε,δ)-Approximate Aggregation Algorithm in Wireless Sensor Networks

CHENG Si-Yao+, LI Jian-Zhong

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

+ Corresponding author: E-mail: csyhit@126.com

Cheng SY, Li JZ. (ε,δ)-Approximate aggregation algorithm in wireless sensor networks. *Journal of Software*, 2010,21(8):1936−1953. http://www.jos.org.cn/1000-9825/3641.htm

Abstract: This paper proposes an approximate aggregation algorithm based on Bernoulli sampling to satisfy the requirement of arbitrary precision in wireless sensor networks (WSN). Besides, two sample data adaptive algorithms are also provided. One is to adapt the sample to the varying precision requirement. The other is to adapt the sample to the varying sensed data in networks. Theoretical analysis and experimental results show that the proposed algorithms have good performance in terms of accuracy and energy cost.

Key words: wireless sensor network; approximate aggregation; Bernoulli sampling

摘 要: 提出了一种基于Bernoulli抽样的近似聚集算法,以满足无线传感器网络(简称WSN)中用户给定的任意精度需求.同时,还提出了两种样本数据的自适应算法,分别用于处理用户的精确度需求以及网络中的感知数据发生变化的情况.理论分析及实验结果表明,所提出的算法在近似结果的精确度、能量开销等方面均优于已有的近似聚集算法.

关键词: 传感器网络;近似聚集;Bernoulli 抽样

中图法分类号: TP393 **文献标识码**: A

随着通信技术、嵌入式计算技术和传感技术的飞速发展和日益成熟,无线传感器网络在国防军事、国家安全、环境监测、交通管理、医疗卫生、制造业、反恐抗灾等领域具有广泛的应用前景.在这些应用中,感知数据的聚集(如求和等)是一种常用的重要操作.近年来,人们在聚集算法方面开展了很多研究工作.早期的研究工作[1−5]主要集中在如何获取精确的聚集值方面,故需要网络中所有节点的感知数据均参与聚集运算,能量消耗很大.同时,由于节点的感知数据常伴有噪声,这些方法仍很难得到100%的精确聚集结果.

实际上,很多 WSN 应用并不需要精确的聚集结果,使用近似聚集结果就能进行决策分析等工作[6−8].于是,

* Supported by the National Natural Science Foundation of China under Grant Nos.60533110, 60703012 (国家自然科学基金); the National Basic Research Program of China under Grant No.2006CB303000 (国家重点基础研究发展计划(973)); the Program for New Century Excellent Talents in University of China under Grant No.NCET-05-0333 (新世纪优秀人才支持计划); the NSFC/RGC Joint Research Scheme under Grant No.60831160525 (NSFC/RGC 联合资助项目)

Received 2009-02-09; Accepted 2009-04-29

为了节省能量开销,人们开展了近似聚集算法的研究.基于 Sketch 的近似聚集算法[6,9]首先被提出来.这类算法通过传送压缩后的感知数据来达到节能的目的,但这类算法仍需要所有感知数据都参与聚集计算过程中,其数据通信量和能量开销依然很大.随后,人们又提出了另外两种近似聚集算法:基于时间相关性的近似聚集算法[7,10,11]和基于空间相关性的近似聚集算法[8,12,13].这些算法在能量开销等方面有了很大的进步,但是这些算法均具有固定的误差界,并且该误差界很难调节.对于基于时间相关性的算法来说,调整算法的误差界需要由 Sink 根据节点感知数据变化程度,为每个节点重新分配过滤区间.为了完成上述工作,Sink 节点需要了解整个网络感知数据的变化趋势,这对于大规模传感器网络来说将很难实现.对于基于空间相关性的算法来说,调整算法的误差界需要调整每个节点的局部数据预测模型以及 Sink 节点的全局预测模型,必将引入大量的计算和数据通信工作,消耗大量的计算资源和能量,这对于能量有限的大规模传感器网络来说也很难实施.

在实际应用中,一个 WSN 拥有大量用户,不同用户或不同聚集操作对近似聚集结果的精确度的要求也不同.以水质监测 WSN 为例,用户 A 需要根据传感器节点监测数据来判断当前水质是否满足饮用标准,而用户 B 则需要确定当前水质具体属于哪类水.根据中国地表水环境的质量标准,可饮用水分为 3 个等级.显然,用户 B 的精确度要求必然比用户 A 的精确度要求高.而已有的近似聚集方法具有固定的误差界,故可能只适于处理用户 A 的查询而不适于处理用户 B 的查询.综上所述,我们正面临一个挑战:如何设计一个能量有效的近似聚集算法,使其能够直接有效地完成具有任意精确度要求的聚集计算.当然,这种算法的时间复杂性、通信复杂性和能量复杂性随着精度的提高而增加.针对该问题,本文基于 Bernoulli 抽样技术,提出了一种近似聚集算法,称为(ε,δ)-近似聚集算法,其中,$\varepsilon(\varepsilon\geqslant 0)$表示相对误差上限,$\delta(1\geqslant\delta\geqslant 0)$表示失误概率上限.即对于任意小的$\varepsilon$和$\delta$,该算法产生的近似聚集结果与真实值的相对误差大于等于ε的概率小于等于δ.当$\varepsilon=0,\delta=0$ 时,该算法将给出精确的聚集结果.

对于给定的ε和δ,(ε,δ)-近似聚集算法主要通过 3 步来完成聚集计算:第 1 步,算法将根据ε和δ确定抽样概率的大小;第 2 步,算法将分布式地在网内抽取样本数据;第 3 步,算法使用样本数据,依据数学原理计算不同类型的近似聚集结果,包括近似聚集和、近似均值、近似无重复计数等,使得近似聚集结果与真实值的相对误差大于等于ε的概率小于等于δ.

本文的主要贡献如下:

(1) 提出了根据用户给定的ε和δ来确定抽样概率的数学原理和相应的抽样概率确定方法.该方法保证了近似聚集结果与真实值的相对误差大于等于ε的概率小于等于δ.同时,由该方法确定的抽样概率是优化的.

(2) 提出了低能耗的分布式抽样方法.

(3) 给出了估计近似聚集和、近似均值、近似无重复计数的数学方法,根据这些数学方法提出了基于 Bernoulli 抽样的(ε,δ)-近似聚集算法,同时证明了该算法的正确性及低能耗性.

(4) 给出了两种近似聚集结果维护算法,分别适用于ε和δ发生变化及节点感知数据发生变化的情况.

本文第 1 节综述相关工作.第 2 节给出问题的形式化定义.第 3 节讨论(ε,δ)-近似聚集方法的数学基础.第 4 节介绍 WSN 中的分布式 Bernoulli 抽样算法.第 5 节介绍基于 Bernoulli 抽样的(ε,δ)-近似聚集算法,并给出两种近似聚集结果维护算法.第 6 节通过实验分析(ε,δ)-近似聚集算法的性能.第 7 节给出结论.

1 相关工作

基于抽样技术的近似聚集方法已经在传统数据库系统、数据流系统和 P2P 数据管理系统中得到了应用,很多基于抽样技术的近似聚集算法已经被提出来了[14−19].但是,这些算法都不适合于 WSN.

文献[14,15]介绍了一种传统数据库系统中基于顺序抽样(sequential sampling)的集中式近似聚集方法.它的主要思想是,在传统数据库中顺序地读入元组,同时根据已读入样本数据的方差来判断利用这些样本数据是否能够获得满足用户精确度需求的近似聚集结果.如果能,则停止读入元组,并利用已有的样本数据完成计算;如果不能,则继续重复进行读入元组及判断的操作.这种方法的集中式特性不适合 WSN 这种特殊的大规模分布式

系统.此外,这种方法需要进行多次抽样才能完成近似聚集值的计算过程.多次在 WSN 中抽样将消耗大量的能量,是 WSN 的大忌.

文献[16]讨论了如何在数据流中利用 Bernoulli 抽样进行近似聚集运算,但是该方法仍然是集中式方法,难以在 WSN 中实现.

文献[17,18]介绍了 P2P 数据管理系统中基于抽样的近似聚集算法.该算法是一种分布式算法,并且在 P2P 数据管理系统中具有较高的执行效率.该算法采用了随机游走的方式来完成抽样过程.对于 WSN 来说,在网络中进行随机游走抽样将需要大量的通信,消耗大量的能量.因而,随机游走的抽样方法不适合 WSN.此外,P2P 数据库中的数据较为稳定,不需要对以往获得的样本数据进行动态的维护,而对于传感器网络来说,节点的感知数据是不断变化的,需要对样本数据进行动态的维护.综上,P2P 数据管理系统中基于抽样的近似聚集算法也不适合 WSN.

文献[19]研究了如何在数据仓库中利用 Sampling Synopses 计算近似无重复计数的问题,提出了相应的近似聚集算法.由于该算法也是一种集中式算法,所以也不适用于 WSN.

2 问题的定义

不失一般性,设传感器网络有 N 个节点,从 1 到 N 编号.整个传感器网络划分为若干个互不相交的簇.

令 s_{ti} 表示在时刻 t 节点 i 所感知的数据,集合 $S_t=\{s_{t1},s_{t2},\ldots,s_{tN}\}$ 表示在时刻 t 网络中所有感知数据的集合,$Dis(S_t)$表示由 S_t 中所有不同值构成的集合.由于传感器节点的感知数据都是有界的,则 S_t 是个有界集合,设其上、下界为分别 $\sup(S_t)$和 $\inf(S_t)$.

传感器网络感知数据的聚集是定义在集合 S_t 上的操作,包括聚集和 $Sum(S_t)=\sum_{i=1}^{N}s_{ti}$、均值 $Avg(S_t)=\frac{1}{N}\sum_{i=1}^{N}s_{ti}$、无重复计数 $Dcount(S_t)=|Dis(S_t)|$等.本文将研究这些聚集值的近似计算算法.

定义 1((ε,δ)-近似值). 对于任意的$\varepsilon(\varepsilon\geqslant 0)$及$\delta(1\geqslant\delta\geqslant 0)$,我们称近似聚集值 $\hat{I}_t$ 为 I_t的(ε,δ)-近似值,当且仅当 $\hat{I}_t$ 满足以下关系: $\Pr(|(\hat{I}_t-I_t)/I_t|\geqslant\varepsilon)\leqslant\delta$.其中,$\Pr(X)$是 X 成立的概率.

定义 2((ε,δ)-近似聚集). 设 I_t 为传感器网络在时刻 t 的精确聚集值,如 $Sum(S_t)$,$Avg(S_t)$或 $Dcount(S_t)$等. $\hat{I}_t$ 称为(ε,δ)-近似聚集值当且仅当 $\hat{I}_t$ 为 I_t的(ε,δ)-近似值.

求解(ε,δ)-近似聚集值 $\hat{I}_t$ 的问题称为(ε,δ)-近似聚集计算问题.该问题可如下定义:

输入:

1. 一个具有 N 个节点的无线传感器网络 WSN;
2. 存储在 WSN 中的时刻 t 的感知数据集合 $S_t=\{s_{t1},s_{t2},\ldots,s_{tN}\}$;
3. 相对误差上限$\varepsilon(\varepsilon\geqslant 0)$和失误概率上限$\delta(1\geqslant\delta\geqslant 0)$.

输出:满足定义 2 的(ε,δ)-近似聚集值.

本文采用基于 Bernoulli 抽样的分布式计算方法来求解(ε,δ)-近似聚集计算问题.该算法由如下 3 步构成:

步骤 1. Sink 节点根据用户给定的ε和δ确定所需的抽样概率的大小.

步骤 2. Sink 根据所确定的抽样概率在网内进行分布式抽样,并在网络中分布式地存储样本数据信息.

步骤 3. 根据样本数据,完成(ε,δ)-近似聚集的计算,获得(ε,δ)-近似聚集结果.

此外,在无线传感器网络工作过程中,如果传感器节点的感知数据或用户的查询条件发生了变化,则需对 Sink 和网内存储的近似聚集结果进行动态维护.

为了方便阅读.表 1 列出了本文中常用符号的说明.

Table 1 Symbols

表 1 符号

N	The number of nodes in the network
C_l	The id of cluster l
s_{ti}	The sensed data of node i at t
$S_t=\{s_{t1},s_{t2}\ldots,s_{tN}\}$	The sensed data set of the whole network at t
$Dis(S_t)$	The distinct values set of S_t
$s_{tv}^{(d)}$	A element of $Dis(S_t)$
y_{tv}	The times of $s_{tv}^{(d)}$ appearing in S_t
$\inf(S_t)$	The lower bound of S_t
$\sup(S_t)$	The upper bound of S_t
$Sum(S_t)$	The exact sum result of S_t
$Avg(S_t)$	The exact average result of S_t
$Dcount(S_t)$	The exact distinct-count result of S_t
q	The sampling probability
$B^{(q)}$	The sample data set with sampling probability q

3 (ε,δ)-近似聚集算法的数学基础

时刻 t,WSN 中的 Bernoulli 抽样就是从感知数据集合 S_t 中随机地选择一个子集合 $B^{(q)}$,满足对于$\forall s_{ti}\in S_t$均有 $P(s_{ti}\in B^{(q)})=q$,并且对于$\forall i,j(1\leqslant i\neq j\leqslant N)$,事件 $s_{ti}\in B^{(q)}$与事件 $s_{tj}\in B^{(q)}$相互独立.其中,q 表示抽样概率.

3.1 近似聚集和的计算模型

设 S_t 表示 WSN 在时刻 t 的感知数据集合,$B^{(q)}$表示 S_t 的抽样概率为 q 的 Bernoulli 样本数据集合,则基于 Bernoulli 抽样的近似聚集和的计算模型为

$$\widehat{Sum(S_t)}_B^{(q)}=\frac{1}{q}\sum\nolimits_{s_{ti}\in B^{(q)}}s_{ti}.$$

其中,$\widehat{Sum(S_t)}_B^{(q)}$ 的上角标 q 表示抽样概率,下角标 B 表示 Bernoulli 抽样.

下面我们将证明 $\widehat{Sum(S_t)}_B^{(q)}$ 是精确聚集和 $Sum(S_t)$的无偏估计.

定义 3(无偏估计). $\hat{I}_t$ 称为 I_t 的无偏估计,当且仅当 $\hat{I}_t$ 的数学期望等于 I_t,即 $E(\hat{I}_t)=I_t$.

定理 1. $\widehat{Sum(S_t)}_B^{(q)}$ 的数学期望 $E(\widehat{Sum(S_t)}_B^{(q)})$ 满足 $E(\widehat{Sum(S_t)}_B^{(q)})=Sum(S_t)$,方差 $Var(\widehat{Sum(S_t)}_B^{(q)})$ 满足:

$$Var(\widehat{Sum(S_t)}_B^{(q)})\leqslant \sup(S_t)\times Sum(S_t)\times(1-q)/q.$$

证明:设 $Dis(S_t)$是由 S_t 中所有不同值构成的集合,$s_{tv}^{(d)}$ $(1\leqslant v\leqslant|Dis(S_t)|)$表示集合 $Dis(S_t)$中的一个元素,y_{tv},x_{tv}分别表示数值 $s_{tv}^{(d)}$ 在集合 S_t 及 $B^{(q)}$中出现的次数.

精确聚集和 $Sum(S_t)$可按如下方式进行计算: $Sum(S_t)=\sum_{i=1}^{N}s_{ti}=\sum_{s_{tv}^{(d)}\in Dis(S_t)}s_{tv}^{(d)}y_{tv}$.

由于 $B^{(q)}$为 S_t 的一个抽样概率为 q 的 Bernoulli 样本数据集合,x_{tv} 表示数值 $s_{tv}^{(d)}$ 在集合 $B^{(q)}$中出现的次数(如果 $s_{tv}^{(d)}\notin B^{(q)}$, 则 $x_{tv}=0$),因而 $\widehat{Sum(S_t)}_B^{(q)}=\frac{1}{q}\sum_{s_{ti}\in B^{(q)}}s_{ti}=\sum_{s_{tv}^{(d)}\in Dis(S_t)}s_{tv}^{(d)}\frac{x_{tv}}{q}$.

根据 Bernoulli 抽样的性质,$x_{tv}$$(1\leqslant v\leqslant|Dis(S_t)|)$可视为一个随机变量.它服从参数为 y_{tv},q 的二项分布,即 $E(x_{tv})=qy_{tv}$,$Var(x_{tv})=(1-q)qy_{tv}$,并且对于$\forall v,w(1\leqslant v\neq w\leqslant|Dis(S_t)|)$,$x_{tv}$ 与 x_{tw} 相互独立.于是,

$$E(\widehat{Sum(S_t)}_B^{(q)})=E\left(\sum\nolimits_{s_{tv}^{(d)}\in D(S_t)}s_{tv}^{(d)}\frac{x_{tv}}{q}\right)=\sum\nolimits_{s_{tv}^{(d)}\in D(S_t)}s_{tv}^{(d)}\frac{E(x_{tv})}{q}=\sum\nolimits_{s_{tv}^{(d)}\in D(S_t)}s_{tv}^{(d)}y_{tv}=Sum(S_t),$$

$$Var(\widehat{Sum(S_t)}_B^{(q)}) = \sum_{s_{tv}^{(d)} \in Dis(S_t)} (s_{tv}^{(d)})^2 \frac{Var(x_{tv})}{q^2} = \sum_{s_{tv}^{(d)} \in Dis(S_t)} (s_{tv}^{(d)})^2 \frac{y_{tv}(1-q)}{q}$$
$$\leqslant \sup(S_t)\frac{1-q}{q} \sum_{s_{tv}^{(d)} \in D(S_t)} s_{tv}^{(d)} y_{tv} = \sup(S_t)\frac{1-q}{q} Sum(S_t).$$ □

根据定理 1, $\widehat{Sum(S_t)}_B^{(q)}$ 是精确聚集和 $Sum(S_t)$的无偏估计.因而,应用中心极限定理可知,当样本容量大于 30 时, $\widehat{Sum(S_t)}_B^{(q)}$ 将近似服从参数为 $Sum(S_t)$和 $Var(\widehat{Sum(S_t)}_u^{(m)})$ 的正态分布[20].由于无线传感器网络的规模很大,即使按很低的精度要求对其进行抽样,所需的样本容量也将远大于 30,即 $\widehat{Sum(S_t)}_B^{(q)}$ 近似服从正态分布的条件是成立的.于是,我们可以通过下面的定理 2 来确定求(ε,δ)-近似聚集和所需的抽样概率的大小.

定理 2. 如果 Bernoulli 抽样的抽样概率满足如下公式:

$$q \geqslant \frac{1}{N \times \frac{\inf(S_t)}{\sup(S_t)} \times \frac{\varepsilon^2}{\phi_{\delta/2}^2} + 1} \tag{1}$$

则 $\widehat{Sum(S_t)}_B^{(q)}$ 为 $Sum(S_t)$的(ε,δ)-近似值,其中,$\phi_{\delta/2}$是标准正态分布的上侧$\delta/2$ 分位数.

证明:由公式(1)可得, $N \times \inf(S_t) \geqslant \frac{\phi_{\delta/2}^2}{\varepsilon^2}\sup(S_t)\frac{1-q}{q}$,由于 $Sum(S_t) = \sum_{i=1}^{N} s_{ti} \geqslant \inf(S_t) \times N$,故

$$Sum(S_t) \geqslant \frac{\phi_{\delta/2}^2}{\varepsilon^2}\sup(S_t)\frac{1-q}{q} \tag{2}$$

根据定理 1, $Var(\widehat{Sum(S_t)}_B^{(q)}) \leqslant \sup(S_t)\frac{1-q}{q} Sum(S_t)$, 应用公式(2),有 $Sum(S_t)^2 \geqslant \frac{\phi_{\delta/2}^2}{\varepsilon^2} Var(\widehat{Sum(S_t)}_B^{(q)})$, 即

$$\phi_{\delta/2} \times \sqrt{Var(\widehat{Sum(S_t)}_u^{(m)})} \leqslant \varepsilon \times Sum(S_t) \tag{3}$$

由于 $\widehat{Sum(S_t)}_B^{(q)}$ 近似服从正态分布,且其数学期望为 $Sum(S_t)$,于是,

$$\Pr\left(\left|\widehat{Sum(S_t)}_B^{(q)} - Sum(S_t)\right| \geqslant \phi_{\delta/2}\sqrt{Var(\widehat{Sum(S_t)}_B^{(q)})}\right) \leqslant \delta \tag{4}$$

根据公式(3)、公式(4),有 $\Pr(|(\widehat{Sum(S_t)}_B^{(q)} - Sum(S_t)/Sum(S_t))| \geqslant \varepsilon) \leqslant \delta$.

即 $\widehat{Sum(S_t)}_B^{(q)}$ 为 $Sum(S_t)$的(ε,δ)-近似值. □

3.2 近似均值的计算模型

设 S_t 表示 WSN 在时刻 t 的感知数据集合,$B^{(q)}$表示 S_t 的抽样概率为 q 的 Bernoulli 样本数据集合,则基于 Bernoulli 抽样的近似均值的计算模型为 $\widehat{Avg(S_t)}_B^{(q)} = \frac{1}{qN}\sum_{s_{ti} \in B^{(q)}} s_{ti}$.

定理 3. $\widehat{Avg(S_t)}_B^{(q)}$ 的数学期望 $E(\widehat{Avg(S_t)}_B^{(q)})$ 满足 $E(\widehat{Avg(S_t)}_B^{(q)}) = Avg(S_t)$, 方差 $Var(\widehat{Avg(S_t)}_B^{(q)})$ 满足:

$$Var(\widehat{Avg(S_t)}_B^{(q)}) \leqslant \sup(S_t)\frac{1-q}{Nq} Avg(S_t).$$

证明:根据定理 1,有 $E(\widehat{Sum(S_t)}_B^{(q)}) = Sum(S_t)$,从而

$$E(\widehat{Avg(S_t)}_B^{(q)}) = E\left(\frac{1}{qN}\sum_{s_{ti} \in B^{(q)}} s_{ti}\right) = \frac{1}{N}E(\widehat{Sum(S_t)}_B^{(q)}) = \frac{1}{N}Sum(S_t) = Avg(S_t).$$

根据 $Var(\widehat{Sum(S_t)}_B^{(q)}) \leqslant \sup(S_t)\frac{1-q}{q} Sum(S_t)$, 有

$$Var(\widehat{Avg(S_t)}_B^{(q)}) = Var\left(\frac{1}{qN}\sum_{s_{ti}\in B^{(q)}} s_{ti}\right) = \frac{1}{N^2}Var(\widehat{Sum(S_t)}_B^{(q)}) = \sup(S_t)\frac{1-q}{Nq}Avg(S_t). \quad \square$$

根据定理 3 及中心极限定理,可以确定计算(ε,δ)-近似均值所需的抽样概率的大小.

定理 4. 如果 Bernoulli 抽样的抽样概率 q 满足 $q \geqslant \dfrac{1}{N\times\dfrac{\inf(S_t)}{\sup(S_t)}\times\dfrac{\varepsilon^2}{\phi_{\delta/2}^2}+1}$,则 $\widehat{Avg(S_t)}_B^{(q)}$ 为 $Avg(S_t)$的(ε,δ)-近似值.其中,$\phi_{\delta/2}$ 是标准正态分布的上侧$\delta/2$ 分位数.

定理 4 的证明与定理 2 的证明类似,在此不再赘述.

3.3 近似无重复计数的计算模型

设 S_t 表示 WSN 在时刻 t 的感知数据集合,$B^{(q)}$表示 S_t 的抽样概率为 q 的 Bernoulli 样本数据集合,$Dis(S_t)$表示 S_t 中所有不同值构成的集合,则构造近似无重复计数计算模型的基本思想是:利用 $B^{(q)}$来构建集合 $B_{dis}^{(q)}$,使得 $B_{dis}^{(q)}$ 是集合 $Dis(S_t)$的抽样概率为 q 的 Bernoulli 样本数据集合.进而,近似无重复计数 $\widehat{Dcount(S_t)}_B^{(q)}$ 等于 $|B_{dis}^{(q)}|/q$.

由 $B^{(q)}$构建 $B_{dis}^{(q)}$ 的方法如下:

(1) $B_{dis}^{(q)} = \varnothing$ 初始为空.

(2) 对于 $\forall s_{tv}^{(d)} \in B^{(q)}$,令 x_{tv} 为 $s_{tv}^{(d)}$ 在 $B^{(q)}$中出现的次数,

i. 如果 x_{tv}>1,则 $s_{tv}^{(d)}$ 将以概率 q 包含于集合 $B_{dis}^{(q)}$ 之中,即 $\Pr(s_{tv}^{(d)} \in B_{dis}^{(q)} \mid x_v > 1) = q$;

ii. 如果 x_{tv}=1,令 i 表示抽样过程中返回感知值 $s_{tv}^{(d)}$ 的节点编号,如果在 WSN 中存在着某个节点 j 满足 s_{tj}= $s_{tv}^{(d)}$ 且 j>i,则 $s_{tv}^{(d)}$ 将以概率 q 包含于集合 $B_{dis}^{(q)}$ 之中;否则,$s_{tv}^{(d)}$ 以概率 1 包含于集合 $B_{dis}^{(q)}$ 之中,即

$$P(s_{tv}^{(d)} \in B_{dis}^{(q)} \mid x_{tv}=1 \wedge s_{tv}^{(d)} = s_{ti} \wedge s_{ti} \in B^{(q)} \wedge (\exists j \mid s_{tj} = s_{tv}^{(d)}, j > i)) = q,$$

$$P(s_{tv}^{(d)} \in B_{dis}^{(q)} \mid x_{tv}=1 \wedge s_{tv}^{(d)} = s_{ti} \wedge s_{ti} \in B^{(q)} \wedge (\neg\exists j \mid s_{tj} = s_{tv}^{(d)}, j > i)) = 1.$$

定理 5. $B_{dis}^{(q)}$ 为集合 $Dis(S_t)$的抽样概率为 q 的 Bernoulli 样本数据集合.

证明:显然,$B_{dis}^{(q)} \subseteq Dis(S_t)$. 从而,我们只需证明对于 $\forall s_{tv}^{(d)} \in Dis(S_t)$,均有 $P(s_{tv}^{(d)} \in B_{dis}^{(q)}) = q$ 即可.

根据全概率公式,有公式(5):

$$\begin{aligned}\Pr(s_{tv}^{(d)} \in B_{dis}^{(q)}) = {} & \Pr(x_{tv}=1 \wedge s_{tv}^{(d)} = s_{ti} \wedge s_{ti} \in B^{(q)} \wedge (\neg\exists j \mid s_{tj} = s_{tv}^{(d)}, j > i)) + \\ & q\Pr(x_{tv}=1 \wedge s_{tv}^{(d)} = s_{ti} \wedge s_{ti} \in B^{(q)} \wedge (\exists j \mid s_{tj} = s_{tv}^{(d)}, j > i)) + \\ & qP(x_{tv}>1) + 0 \times \Pr(s_{tv}^{(d)} \notin B^{(q)})\end{aligned} \tag{5}$$

根据 Bernoulli 抽样的性质,对于$\forall v(1\leqslant v\leqslant |Dis(S_t)|)$,$x_{tv}$ 服从参数为 y_{tv},q 的二项分布,其中,y_{tv} 为 $s_{tv}^{(d)}$ 在 S_t 中出现的次数.于是,

$$\Pr(x_{tv}>1) = 1 - \Pr(x_{tv}=0) - \Pr(x_{tv}=1) = 1 - (1-q)^{y_{tv}} - y_{tv}q(1-q)^{y_{tv}-1} \tag{6}$$

同时,对于$\forall i,j(1\leqslant i\neq j\leqslant N)$,事件 $s_{ti}\in B^{(q)}$与事件 $s_{tj}\in B^{(q)}$相互独立,从而,

$$\Pr(x_{tv}=1 \wedge s_{tv}^{(d)} = s_{ti} \wedge s_{ti} \in B^{(q)} \wedge (\neg\exists j \mid s_{tj} = s_{tv}^{(d)}, j > i)) = q(1-q)^{y_{tv}-1} \tag{7}$$

$$\Pr(x_{tv}=1 \wedge s_{tv}^{(d)} = s_{ti} \wedge s_{ti} \in B^{(q)} \wedge (\exists j \mid s_{tj} = s_{tv}^{(d)}, j > i)) = (y_{tv}-1)q(1-q)^{y_{tv}-1} \tag{8}$$

将公式(6)~公式(8)代入到公式(5)中,有

$$P(s_{tv}^{(d)} \in B_{dis}^{(q)}) = q(1-q)^{y_{tv}-1} + q\times(y_{tv}-1)q(1-q)^{y_{tv}-1} + q\times(1-(1-q)^{y_{tv}} - y_{tv}q(1-q)^{y_{tv}-1}) = q.$$

即 $B_{dis}^{(q)}$ 为集合 $Dis(S_t)$的抽样概率为 q 的 Bernoulli 样本数据集合. $\square$

定理 6. $\widehat{Dcount(S_t)}_B^{(q)}$ ($\widehat{Dcount(S_t)}_B^{(q)} = |B_{dis}^{(q)}|/q$) 的数学期望 $E(\widehat{Dcount(S_t)}_B^{(q)})$ 满足:

$$E(\widehat{Dcount(S_t)}_B^{(q)}) = Dcount(S_t).$$

方差 $Var(\widehat{Dcoun(S_t)}_B^{(q)})$ 满足 $Var(\widehat{Dcoun(S_t)}_B^{(q)}) = (1-q)Dcoun(S_t)/q$.

证明:根据定理 5 可知, $B_{dis}^{(q)}$ 是集合 $Dis(S_t)$的抽样概率 q 的 Bernoulli 样本数据集合.从而,根据 Bernoulli 抽样的性质有:$|B_{dis}^{(q)}|$ 将服从参数为$|Dis(S_t)|$,q 的二项分布,即 $|B_{dis}^{(q)}|$ 的数学期望及方差分别满足公式(9)、公式(10).

$$E(|B_{dis}^{(q)}|) = |Dis(S_t)| \times q = Dcount(S_t) \times q \tag{9}$$

$$Var(|B_{dis}^{(q)}|) = |Dis(S_t)| \times q \times (1-q) = Dcount(S_t) \times q \times (1-q) \tag{10}$$

由于 $\widehat{Dcount(S_t)}_B^{(q)} = |B_{dis}^{(q)}|/q$,从而,根据公式(9)有

$$E(\widehat{Dcount(S_t)}_B^{(q)}) = E(|B_{dis}^{(q)}|/q) = E(|B_{dis}^{(q)}|)/q = Dcount(S_t).$$

根据公式(10)有 $Var(\widehat{Dcount(S_t)}_B^{(q)}) = Var(|B_{dis}^{(q)}|/q) = Var(|B_{dis}^{(q)}|)/q^2 = Dcount(S_t)(1-q)/q$. □

根据定理 6 及中心极限定理,可以确定计算(ε,δ)-近似无重复计数所需的抽样概率的大小.

定理 7. 如果 Bernoulli 抽样的抽样概率满足如下公式 $q \geqslant \dfrac{1}{\inf(Dcount(S_t))\varepsilon^2/\phi_{\delta/2}^2 + 1}$, 则 $\widehat{Dcount(S_t)}_B^{(q)}$ 为 $Dcount(S_t)$的(ε,δ)-近似值.其中,$\phi_{\delta/2}$ 是标准正态分布的上侧$\delta/2$ 分位数,$\inf(Dcount(S_t))$为精确无重复计数 $Dount(S_t)$的下限值.

定理 7 的证明与定理 2 的证明类似,在此不再赘述.

4 基于簇的 Bernoulli 抽样算法

4.1 抽样概率的确定

根据第 3 节中的定理 2、定理 4 和定理 7,对于任意的$\varepsilon \geqslant 0$,$\delta \geqslant 0$,在实际应用中,只需将抽样概率设置为 $\dfrac{\phi_{\delta/2}^2}{\varepsilon^2 \min\left(N \times \dfrac{\inf(S_t)}{\sup(S_t)}, \inf(Dcount(S_t))\right) + \phi_{\delta/2}^2}$,即能保证 $\widehat{Sum(S_t)}_B^{(q)}$,$\widehat{Avg(S_t)}_B^{(q)}$ 及 $\widehat{Dcount(S_t)}_B^{(q)}$ 均是精确聚集值的(ε,δ)-近似值.其中,$\inf(S_t)$,$\sup(S_t)$分别是感知数据的上、下限,$\inf(Dcount(S_t))$表示无重复计数 $Dount(S_t)$的下限值.

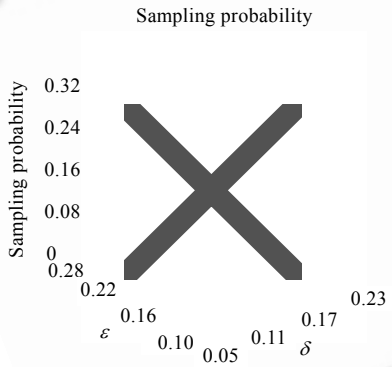


Fig.1 Relationship among sampling probability, ε and δ

图 1 抽样概率与ε,δ的关系

由于 WSN 中的感知数据是地理位置相关的,因而可利用这一特性对 $\inf(Dcount(S_t))$的大小进行估计.例如,对于一个监测温度的 WSN 来说,其覆盖的区域的面积为 A,而该地区温度不发生变化的区域的面积至多为 a,此时,$\inf(Dcount(S_t))$可取值为 A/a.并且,由于感知数据集合 S_t 是非空的,故无重复计数 $Dcount(S_t)$满足 $Dcount(S_t) \geqslant 1$,进而 $\inf(Dcount(S_t)) \geqslant 1$.因此,在不具有任何相关知识的情况下,$\inf(Dcount(S_t))$可取值为 1.

对于规模较大的传感器网络,上述抽样概率是比较小的.对于一个拥有 5 000 个节点的监测温度的传感器网络来说,其中环境温度的上、下限分别为 100°C,10°C,无重复计数的下限为 500,完成(ε,δ)-近似聚集计算所需的抽样概率与ε,δ的关系由图 1 给出.由图 1 可知,当 ε=0.15,δ=0.1 时,抽样概率为 0.19,即仅需要利用 19%的感知数据,就能保证近似聚集值与精确聚集值的相对误差小于 0.15 的概率大于 0.9.因而,利用 Bernoulli 抽样进行近似聚集运算也可以有效地降低节点的访问量,从而达到节能的目的.

以上讨论仅针对 $\sup(S_t)$与 $\inf(S_t)$均大于 0 的情况.当 $\sup(S_t)$与 $\inf(S_t)$均小于 0 时,抽样概率可设置为

$$\frac{1}{\frac{\varepsilon^2}{\phi^2_{\delta/2}}\min\left(N\times\frac{\sup(S_t)}{\inf(S_t)},\inf(Dcount(S_t))\right)+1};$$

当 $\sup(S_t)$大于等于 0 而 $\inf(S_t)$小于等于 0 时,抽样概率可设置为

$$\frac{1}{\frac{\varepsilon^2}{\phi^2_{\delta/2}}\min\left(N\times\frac{\sup(S_t)-\inf(S_t)+\theta}{\theta},\inf(Dcount(S_t))\right)+1},$$

其中,θ为某正数.

4.2 BSC抽样算法

一旦抽样概率确定了,Sink 节点就将在网内发起 Bernoulli 抽样以获得样本数据的聚集信息.我们用三元组$(Sum(q),S(q),D(q))$表示样本数据的聚集信息,其中:$Sum(q)$是所有样本数据的聚集和;$S(q)$是一个集合,满足对于$\forall s_{ti}\in S(q)$,感知值 s_{ti}在样本数据集合中仅出现一次,并且在网内不存在节点 j 使得 $s_{ti}=s_{tj}$且$j>i$;$D(q)$是那些不属于集合 $S(q)$的样本数据的二进制 sketch[21].

本节将介绍一种基于簇的 Bernoulli 抽样算法,称为 BSC(Bernoulli sampling based on clusters)算法,以使 Sink 节点能够获得样本数据的聚集信息$(Sum(q),S(q),D(q))$.

为了便于各个簇传送样本聚集信息,我们在网内构建一棵以 Sink 为根、包含所有簇头节点的生成树.簇头之间可能无法一跳到达,所以该生成树中将包括一些非簇头节点.生成树的构建与维护可参考文献[22].

BSC 算法由图 2 给出,它包括以下 4 个步骤:

(1) Sink 节点将抽样概率 q 沿生成树发送到各个簇.

(2) 对于任意簇 C_l,簇头节点将在本簇内进行抽样,以获取本簇内的样本聚集信息$(Sum_l(q),S_l(q),D_l(q))$,其中:$Sum_l(q)$是本簇内的样本数据的聚集和;$S_l(q)$是一个集合,满足对于$\forall s_{ti}\in S_l(q)$,感知值 s_{ti}在本簇内的样本数据中仅出现 1 次,并且在簇内不存在节点 j 使得 $s_{ti}=s_{tj}$且$j>i$;$D_l(q)$是本簇内那些不属于集合 $S_l(q)$的样本数据的一个二进制 sketch[21].

(3) 各个簇的样本聚集信息将沿生成树传送至 Sink 节点,并在传送过程中进行网内聚集.

(4) Sink 节点计算三元组$(Sum(q),S(q),D(q))$,并对该三元组加以保存.

BSC Algorithm.
Input: q, $S_t=\{s_{t1},s_{t2},\ldots,s_{tN}\}$.
Output: A triple of sample aggregation result $(Sum(q),S(q),D(q))$.
1. Sink sends q to every cluster by spanning tree protocol
2. **For** each cluster in the network
3. ***Sampling-In-Cluster***(q)
//Sampling the sensed data inside the cluster
4. $(Sum(q),S(q),D(q))$=***Send-Samples-Up***()
//Transmit the sample information to the sink
5. **If** $S(q)\neq\varnothing$
6. **For** each item s_{ti} in set $S(q)$
7. Sink queries node i's neighbor clusters
8. **If** there exist node j satisfies $s_{ti}=s_{tj}$ and $j>i$
9. Delete s_{ti} from S
10. $D(q)=D(q)$ OR ***FM_sketch***(s_{ti})
11. Return and store $(Sum(q),S(q),D(q))$
Binary digit ***FM_sketch***(*Real number*): See Ref.[21]

Fig.2 BSC Algorithm

图 2 BSC 算法

对于任意簇 $C_l(1\leqslant l\leqslant k)$,簇内样本数据的抽取算法(*Sampling-In-Cluster*())由图 3 给出,主要包括以下步骤:

(1) 簇头节点将抽样概率 q 在簇内广播.

(2) 簇内地成员节点将按概率 q 来决定是否将自身的感知值发送至簇头节点.

(3) 簇头节点将对本簇内的样本数据进行收集,累加求和可以得到 $Sum_l(q)$.

(4) 对于仅在样本数据中出现 1 次的感知值 s_{ti},簇头节点将在本簇内查询是否存在节点 j 满足 $s_{ti}=s_{tj}$ 且 $j>i$. 若存在,则不将 s_{ti} 加入到集合 $S_l^{(B)}$ 中;否则,将其加入到集合 $S_l(q)$中.

(5) 簇头节点利用 Sketch 技术,可将每个不属于 $S_l(q)$的样本数据映射为一个二进制串,将这些二进制串进行或操作即可获得 $D_l(q)$.

(6) 簇头节点将三元组($Sum_l(q)$,$D_l(q)$,$S_l(q)$)作为本簇内的样本数据聚集信息加以保存.

Sampling-In-Cluster Algorithm.
Input: q, the sensed data in Cluster C_l.
Output: A triple of partial sample aggregation result ($Sum_l(q)$,$S_l(q)$,$D_l(q)$).
1. The cluster head broadcasts q in its cluster
2. **For** each member node i_r in C_l
3. Generate a random number p in range [0,1]
4. **If** $p<q$
5. Send its sensed value s_{ti_r} to the cluster-head
6. The cluster-head collects the sample values in its cluster: $B_l(q)=\{s_{ti_1},s_{ti_2},...,s_{ti_{m_l}}\}$
7. $Sum_l(q)=s_{ti_1}+s_{ti_2}+...+s_{ti_{m_l}}$
8. $D_l(q)=0$
9. $S_l(q)=\varnothing$
10. **For** each item s_{ti_v} in set $B_l(q)$
11. $x_{tv}=|\{s_{ti_w}\mid s_{ti_w}=s_{ti_v}\wedge s_{ti_w}\in B_l(q)\}|$
12. **If** $x_{tv}>1$
13. $D_l(q)=D_l(q)$ OR ***FM_sketch*** (s_{ti_v})
14. **Else**
15. The cluster head broadcast s_{ti_v} in its cluster
16. **If** there exist node j satisfies $s_{tj}=s_{ti_v}$ and $j>i_v$
17. $D_l(q)=D_l(q)$ OR ***FM_sketch*** (s_{ti_v})
18. **Else**
19. $S_l(q)=S_l(q)\cup\{s_{ti_v}\}$
20. The cluster-head stores ($Sum_l(q)$,$S_l(q)$,$D_l(q)$)
21. Return ($Sum_l(q)$,$S_l(q)$,$D_l(q)$)
Binary digit ***FM_sketch***(*Integer*): See Ref.[21]

Fig.3 Sampling-In-Cluster algorithm

图 3 Sampling-In-Cluster 算法

样本聚集信息的传递及网内聚集算法(send-sample-up 算法)由图 4 给出.该算法包含以下 4 步:

(1) 对于生成树中非簇头节点 p,令 $Sum_p(q)=0$,$S_p(q)=\varnothing$,$D_p(q)=0$.

(2) 对于生成树中的任意叶子节点 i,将其保存的三元组($Sum_i(q)$,$D_i(q)$,$S_i(q)$)发送至其父亲节点.

(3) 对于生成树中的任意中间节点 j,它将负责从其儿子节点处收集样本信息三元组,并将所收集的三元组与其自身保存的三元组进行聚集,具体的聚集方法如图 4 所示.节点 j 将新的聚集结果加以保存,并把此结果发送至其父亲节点.

(4) 生成树中的根节点,即 Sink 节点,将采用类似方法对来自儿子节点的样本信息三元组进行聚集,并将聚集结果($Sum(q)$,$S(q)$,$D(q)$)返回.

BSC 算法的计算复杂度和通信复杂度与 Sample-In-Cluster 算法和 Send-Sample-Up 算法的复杂度有关.

对于任意簇 C_l,在执行 Sample-In-Cluster 算法时的计算复杂度和通信复杂度与簇内的样本容量有关.根据 Bernoulli 抽样的性质,簇 C_l 的样本容量的数学期望为 $q\times n_l$,其中,n_l 表示簇 C_l 中的节点数目,因而对于簇 C_l,执行 Sample-In-Cluster 算法时的平均计算复杂度和通信复杂度均为 $O(q\times n_l)$.

Send-Sample-Up 算法的计算复杂度和通信复杂度将取决于生成树中节点的数目.设 k 表示网络中包含簇的个数,α 表示两个相邻簇之间的簇头节点的平均距离,则生成树中的平均节点数目为 αk,故 Send-Sample-Up 算法的平均计算复杂度和通信复杂度均为 $O(\alpha k)$.

Send-Sample-Up Algorithm.
Output: ($Sum(q)$,$S(q)$,$D(q)$).
1. **For** each node j belongs to Spanning Tree
2. **If** j is not a cluster-head
3. $Sum_j(q)=0$; $D_j(q)=0$; $S_j(q)=\varnothing$
4. **If** j is a leaf node
5. Send ($Sum_j(q)$,$D_j(q)$,$S_j(q)$) to its parent
6. **Else**
7. Receive $\{(Sum_{j_u}, D_{j_u}, S_{j_u}) \mid 1 \leqslant u \leqslant r\}$ from its sons.
8. $Sum_j(q)=Sum_j(q)+ Sum_{j_1}(q)+...+Sum_{j_r}(q)$
9. $D_j(q)=D_j(q)$ OR $D_{j_1}(q)$ OR ... OR $D_{j_r}(q)$
10. $S_{j_{r+1}}(q)=S_j(q)$
11. $SS=\bigcup_{1\leqslant p<q\leqslant r+1}(S_{j_p}\cap S_{j_q})$
12. **For** each item s_{ti} in set SS
13. $D_j(q)=D_j(q)$ OR ***FM_sketch***(s_{ti})
14. $S_j(q)=\left(\bigcup_{1\leqslant p\leqslant r+1}S_{j_p}\right)\Big/SS$
15. **For** each value s_{tr} in $S_j(q)$
16. **If** $D_j(q)$ OR ***FM_sketch***(s_{tr})=$D_j(q)$
17. Delete s_{tr} from $S_j(q)$
18. **If** j is the sink
19. $Sum(q)=Sum_j(q)$; $D(q)=D_j(q)$; $S(q)=S_j(q)$
20. Return ($Sum(q)$,$D(q)$,$S(q)$)
21. **Else**
22. Node j store ($Sum_j(q)$,$S_j(q)$,$D_j(q)$)
23. Send ($Sum_j(q)$,$S_j(q)$,$D_j(q)$) to its parent
Binary digit ***FM_sketch***(*Integer*): See Ref.[21]

Fig.4 Send-Sample-Up algorithm

图 4 Send-Sample-Up 算法

根据上述分析,BSC 算法的平均计算、通信复杂度为 $O(q\times N+\alpha k)$.设 j_1 为执行一条指令的能量开销,j_2 为发送及接收一个数据包的能量开销,故 BSC 算法的平均能量开销为 $O((j_1+j_2)\times(q\times N+\alpha k))$.根据第 4.1 节的讨论,$q=\dfrac{\phi^2_{\delta/2}}{\varepsilon^2\min\left(N\times\dfrac{\inf(S_t)}{\sup(S_t)},\inf(Dcount(S_t))\right)+\phi^2_{\delta/2}}$,所以 BSC 算法的计算复杂度、通信复杂度及能量开销分别为

$$O((\phi^2_{\delta/2}/\varepsilon^2+\alpha k))=O\left(\frac{1}{\varepsilon^2}\ln\left(\frac{1}{\delta}\right)+\alpha k\right),\ O\left(\frac{1}{\varepsilon^2}\ln\left(\frac{1}{\delta}\right)+\alpha k\right),\ O\left((j_1+j_2)\times\left(\frac{1}{\varepsilon^2}\ln\left(\frac{1}{\delta}\right)+\alpha k\right)\right).$$

5 基于 Bernoulli 抽样的(ε,δ)-近似聚集算法

5.1 基于Bernoulli抽样的(ε,δ)-近似聚集算法

根据第 3 节介绍的数学理论及 BSC 算法,(ε,δ)-近似聚集算法由图 5 给出.该算法将包括以下 5 步:

(1) 利用第 4.1 节介绍的抽样概率的确定方法,Sink 节点根据给定的ε,δ计算出所需的抽样概率.

(2) 调用 BSC 算法,获得整个网络的样本数据聚集三元组($Sum(q)$,$S(q)$,$D(q)$).

(3) 利用 Sketch 技术及二进制数 $D(q)$,可计算网内样本数据的无重复计数 $Dcount(B^{(q)})$.

(4) 计算(ε,δ)-近似聚集和 $\widehat{Sum(S_t)}_B^{(q)}=\dfrac{1}{q}Sum(q)$、(ε,δ)-近似均值 $\widehat{Avg(S_t)}_B^{(q)}=\dfrac{1}{N\times q}Sum(q)$、(ε,δ)-近似无重复计数 $\widehat{Dcount(S_t)}_B^{(q)}=Dcount(B^{(q)})+\dfrac{1}{q}|S(q)|$.其中,$|S(q)|$表示集合 $S(q)$的大小.

(5) 返回(ε,δ)-近似聚集结果 $\widehat{Sum(S_t)}_B^{(q)}$,$\widehat{Avg(S_t)}_B^{(q)}$,$\widehat{Dcount(S_t)}_B^{(q)}$,Sink 节点对上述近似聚集结果及抽样概率加以保存.

由图 5 可知,(ε,δ)-近似聚集算法的计算复杂度和通信复杂度取决于 BSC 算法,并且该算法的能量开销也由

BSC 算法的能量开销决定.因而,(ε,δ)-近似聚集算法的计算复杂度、通信复杂度均为 $O\left(\frac{1}{\varepsilon^2}\ln\left(\frac{1}{\delta}\right)+\alpha k\right)$,$(\varepsilon,\delta)$-近似聚集算法的能量开销为 $O\left((j_1+j_2)\times\left(\frac{1}{\varepsilon^2}\ln\left(\frac{1}{\delta}\right)+\alpha k\right)\right)$.

(ε,δ)-Approximate Aggregation Algorithm.
Input: ε, δ, $S_t=\{s_{t1},s_{t2},\ldots,s_{tN}\}$.
Output: $\widehat{Sum(S_t)}_B^{(q)}$, $\widehat{Avg(S_t)}_B^{(q)}$, $\widehat{Dcount(S_t)}_B^{(q)}$.

1. $q=\dfrac{\phi_{\delta/2}^2}{\varepsilon^2\min\left(N\times\dfrac{\inf(S_t)}{\sup(S_t)},\inf(Dcount(S_t))\right)+\phi_{\delta/2}^2}$
2. Call $BSC(q,S_t)$, get $(Sum(q),S(q),D(q))$
3. $\widehat{Sum(S_t)}_B^{(q)}=Sum(q)/q$
4. $\widehat{Avg(S_t)}_B^{(q)}=\widehat{Sum(S_t)}_B^{(q)}/N$
5. $Dcount(B^{(q)})=\boldsymbol{Sketch}(D(q))$
6. $\widehat{Dcount(S_t)}_B^{(q)}=Dcount(B^{(q)})+\dfrac{1}{q}|S(q)|$
7. Sink stores q, $\widehat{Sum(S_t)}_B^{(q)}$ and $\widehat{Dcount(S_t)}_B^{(q)}$
8. Return $\widehat{Sum(S_t)}_B^{(q)}$, $\widehat{Avg(S_t)}_B^{(q)}$ and $\widehat{Dcount(S_t)}_B^{(q)}$

Integer ***Sketch***(*Binary digit*): See Ref.[21]

Fig.5 (ε,δ)-Approximate aggregation algorithm
图 5 (ε,δ)-近似聚集算法

5.2 多查询的处理

一个 WSN 拥有大量的用户,并且每个用户的精确度要求不尽相同.如果对于每个聚集查询均应用第 5.1 节中介绍的算法,则将消耗大量能量.因而,本节将给出一种节能的算法,以处理具有不同精度要求的多查询请求.

该算法的主要思想是尽量使用 Sink 保存的历史样本信息来处理多查询请求,其中,Sink 保存的样本信息将包括抽样概率、样本数据聚集三元组、近似聚集值等.该方法的描述如下:

首先,如果 Sink 节点以前未处理过聚集查询,则 Sink 利用第 5.1 节中介绍的(ε,δ)-近似聚集算法来计算(ε,δ)-近似聚集值.Sink 节点将抽样概率、样本数据聚集三元组、近似聚集值加以保存.

其次,如果 Sink 节点以前处理过聚集查询,则 Sink 节点将保存有历史样本信息.此时,Sink 节点根据新的ε,δ计算出新的抽样概率 q_{new}.将 q_{new} 与其保存的 q_{old} 相比较,如果 $q_{new}\leqslant q_{old}$,则无须访问网络,只需将 Sink 保存的近似聚集结果返回;否则,Sink 将调用 BSC 算法,在网络中进行抽样概率为$(q_{new}-q_{old})/(1-q_{old})$的 Bernoulli 抽样.Sink 节点利用抽样所得的新的样本数据与历史样本信息可计算出新的近似聚集结果.

最后,Sink 节点将更新其保存的样本信息.

下面的定理 8 证明了上述方法的正确性.

定理 8. 设 $B^{(q_{old})}$,$B^{((q_{new}-q_{old})/(1-q_{old}))}$ 分别是感知数据集合 S_t 的抽样概率为 q_{old} 和$(q_{new}-q_{old})/(1-q_{old})$的 Bernoulli 样本数据集合,并且两次抽样过程是相互独立的,那么 $B^{(q_{old})}\cup B^{((q_{new}-q_{old})/(1-q_{old}))}$ 是 S_t 的一个抽样概率为 q_{new} 的 Bernoulli 样本集合.其中,$q_{new}>q_{old}$.

证明:我们只需证明:对于$\forall s_{ti}\in S_t$,均有 $\Pr(s_{ti}\in B^{(q_{old})}\cup B^{((q_{new}-q_{old})/(1-q_{old}))})=q_{new}$ 即可.

对于$\forall s_{ti}\in S_t$,均有等式(11)成立:

$$\begin{aligned}\Pr(s_{ti}\in B^{(q_{old})}\cup B^{((q_{new}-q_{old})/(1-q_{old}))}) &= \Pr((s_{ti}\in B^{(q_{old})})\cup(s_{ti}\in B^{((q_{new}-q_{old})/(1-q_{old}))}))\\ &= \Pr(s_{ti}\in B^{(q_{old})})+\Pr(s_{ti}\in B^{((q_{new}-q_{old})/(1-q_{old}))})-\\ &\quad \Pr(s_{ti}\in B^{(q_{old})}\cap B^{((q_{new}-q_{old})/(1-q_{old}))})\end{aligned}\tag{11}$$

由于两次抽样过程是相互独立的,有随机事件 $s_{ti}\in B^{(q_{old})}$ 与 $s_{ti}\in B^{((q_{new}-q_{old})/(1-q_{old}))}$ 是相互独立的,于是

$$\Pr(s_{ti}\in B^{(q_{old})}\cap B^{((q_{new}-q_{old})/(1-q_{old}))})=\Pr(s_{ti}\in B^{(q_{old})})\Pr(s_{ti}\in B^{((q_{new}-q_{old})/(1-q_{old}))}).$$

同时,由于 $B^{(q_{old})},B^{((q_{new}-q_{old})/(1-q_{old}))}$ 分别是感知数据集合 S_t 的抽样概率为 q_{old} 和$(q_{new}-q_{old})/(1-q_{old})$的 Bernoulli 样本数据集合,故对于$\forall s_{ti}\in S_t,\Pr(s_{ti}\in B^{(q_{old})})=q,\Pr(s_{ti}\in B^{((q_{new}-q_{old})/(1-q_{old}))})=(q_{new}-q_{old})/(1-q_{old})$.将上述结论代入到公式(11)中,有 $\Pr(s_{ti}\in B^{(q_{old})}\cup B^{((q_{new}-q_{old})/(1-q_{old}))})=q_{old}+\frac{q_{new}-q_{old}}{1-q_{old}}-q_{old}\frac{q_{new}-q_{old}}{1-q_{old}}=q_{new}$. □

5.3 近似聚集结果的维护方法

如果 Sink 节点存储的近似聚集结果一直有效,则当用户的精度要求不发生变化时,我们无须访问网络,仅将 Sink 节点存储的近似聚集值返回给用户即可.但是,传感器节点的感知数据是不断变化的,可能导致 Sink 节点存储的近似聚集结果失效.因而,需要在一定范围内触发抽样过程,以更新失效的近似聚集结果.

最简单的重新抽样触发方法如下:簇头节点一旦发现本簇内的节点的感知数据发生了变化,就在本簇内重新抽样,获取新的样本聚集信息并传送至 Sink 节点;Sink 节点根据新的样本聚集信息重新计算(ε,δ)-近似聚集值.该方法虽然能够保证 Sink 节点所存储的近似聚集结果始终有效,但是将消耗大量能量.为了节省能量开销,本节将以维护近似聚集和为例,介绍一种随机算法.该算法保证了在任意时刻 Sink 节点所存储的近似聚集和有效的概率大于等于 $1-\delta/2$.该算法是基于下面的定理 9 构建的.

定理 9. 设 $S_t,S_{t'}$为时刻 t 及 t'整个网络感知数据的集合,$Sum_l(S_t),Sum(S_{t'})$分别表示时刻 t 及 t'的簇 C_l 内感知数据的聚集和,n_l 为簇 C_l 中节点的个数,$\widehat{Sum(S_t)}_B^{(q)}$ 是 $Sum(S_t)$的(ε,δ)-近似值,

$$\Delta=\min\left\{\frac{1}{1-\varepsilon}\left(1-\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}\right)-1,1-\frac{1}{1+\varepsilon}\left(1+\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}\right)\right\}.$$

如果下面的任意一个条件得到满足,则 $\widehat{Sum(S_t)}_B^{(q)}$ 也是 $Sum(S_{t'})$的(ε,δ)-近似值:

(1) $|Sum(S_{t'})-Sum(S_t)|\leqslant\Delta\times Sum(S_t)$.

(2) 对于任意簇 C_l,均有$|Sum(S_{t'})-Sum_l(S_t)|\leqslant\Delta\times Avg(S_t)\times n_l$.

(3) 对于任意节点 $i(1\leqslant i\leqslant N)$,均有$|s_{t'i}-s_{ti}|\leqslant\Delta\times Avg(S_t)$.

证明:设条件(1)成立,即$|Sum(S_{t'})-Sum(S_t)|\leqslant\Delta\times Sum(S_t)$,故$(1-\Delta)Sum(S_t)\leqslant Sum(S_{t'})\leqslant(1+\Delta)Sum(S_t)$.从而

$$\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_{t'})}-1\right)\leqslant-\varepsilon\right\}\leqslant\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{(1+\Delta)Sum(S_t)}-1\right)\leqslant-\varepsilon\right\}=\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_t)}-1\right)\leqslant(1-\varepsilon)(1+\Delta)-1\right\}\quad(12)$$

$$\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_{t'})}-1\right)\geqslant\varepsilon\right\}\leqslant\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{(1-\Delta)Sum(S_t)}-1\right)\geqslant\varepsilon\right\}=\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_t)}-1\right)\geqslant(1+\varepsilon)(1-\Delta)-1\right\}\quad(13)$$

根据已知 $\Delta\leqslant\frac{1}{1-\varepsilon}\left(1-\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}\right)-1$,即 $(1-\varepsilon)(1+\Delta)-1\leqslant-\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}$,从而根据公式(12)有

$$\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_{t'})}-1\right)\leqslant-\varepsilon\right\}\leqslant\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_t)}-1\right)\leqslant(1-\varepsilon)(1+\Delta)-1\right\}\leqslant\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_t)}-1\right)\leqslant-\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}\right\}\quad(14)$$

而由 $\Delta\leqslant1-\frac{1}{1+\varepsilon}\left(1+\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}\right)$,可得 $(1+\varepsilon)(1-\Delta)-1\geqslant\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}$,从而根据公式(13)有

$$\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_{t'})}-1\right)\geqslant\varepsilon\right\}\leqslant\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_t)}-1\right)\geqslant(1+\varepsilon)(1-\Delta)-1\right\}\leqslant\Pr\left\{\left(\frac{\widehat{Sum(S_t)}_B^{(q)}}{Sum(S_t)}-1\right)\geqslant\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}\right\}\quad(15)$$

故根据公式(14)和公式(15)有

$$\Pr\left\{\left|\frac{\widehat{Sum(S_t)}_B^{(q)}-Sum(S_{t'})}{Sum(S_{t'})}\right|\geqslant\varepsilon\right\}\leqslant\Pr\left\{\left|\frac{\widehat{Sum(S_t)}_B^{(q)}-Sum(S_t)}{Sum(S_t)}\right|\geqslant\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}\right\} \tag{16}$$

同时,由于 $\widehat{Sum(S_t)}_B^{(q)}$ 是 $Sum(S_t)$的(ε,δ)-近似值,根据定理 2, $q\geqslant\frac{\sup(S_t)\phi_{\delta/2}^2}{N\times\inf(S_t)\varepsilon^2+\sup(S_t)\phi_{\delta/2}^2}$, 于是,

$$\sup(S_t)\left(1-\frac{1}{q}\right)\leqslant N\times\inf(S_t)\frac{\varepsilon^2}{\phi_{\delta/2}^2}.$$

根据定理 1 有 $Var(\widehat{Sum(S_t)}_B^{(q)})\leqslant\sup(S_t)\left(\frac{1}{q}-1\right)Sum(S_t)$, 故 $Var(\widehat{Sum(S_t)}_B^{(q)})\leqslant\frac{\varepsilon^2}{\phi_{\delta/2}^2}Sum(S_t)^2\frac{\inf(S_t)}{Avg(S_t)}$.应用中心极限定理有

$$\Pr\left\{\left|\frac{\widehat{Sum(S_t)}_B^{(q)}-Sum(S_t)}{Sum(S_t)}\right|\geqslant\varepsilon\sqrt{\frac{\inf(S_t)}{Avg(S_t)}}\right\}\leqslant\delta \tag{17}$$

根据公式(16)和公式(17),有 $\Pr\left\{\left|\frac{\widehat{Sum(S_t)}_B^{(q)}-Sum(S_{t'})}{Sum(S_{t'})}\right|\geqslant\varepsilon\right\}\leqslant\delta$, 即 $\widehat{Sum(S_t)}_B^{(q)}$ 也是 $Sum(S_{t'})$的(ε,δ)-近似值.

设条件(2)成立,即对于任意簇 l,均有$|Sum_l(S_{t'})-Sum_l(S_t)|\leqslant\Delta\times Avg(S_t)\times n_l$,从而

$$|Sum(S_{t'})-Sum(S_t)|=\left|\sum_{C_l}(Sum_l(S_{t'})-Sum_l(S_t))\right|\leqslant\sum_{C_l}|Sum_l(S_{t'})-Sum_l(S_t)|\leqslant\Delta\times Avg(S_t)\times\sum_{l=1}^{k}n_l=\Delta\times Sum(S_t).$$

此时条件(1)成立,于是 $\widehat{Sum(S_t)}_u^{(m)}$ 是 $Sum(S_{t'})$的(ε,δ)-近似值.

设条件(3)成立,即任意节点 $i(1\leqslant i\leqslant N)$,均有$|s_{t'i}-s_{ti}|\leqslant\Delta\times Avg(S_t)$,从而

$$|Sum(S_{t'})-Sum(S_t)|=\left|\sum_{i=1}^{N}(s_{ti}-s_{t'i})\right|\leqslant\sum_{i=1}^{N}|s_{ti}-s_{t'i}|\leqslant\Delta\times Avg(S_t)\times N=\Delta\times Sum(S_t).$$

此时条件(1)成立,于是 $\widehat{Sum(S_t)}_u^{(m)}$ 是 $Sum(S_{t'})$的(ε,δ)-近似值. □

在定理 9 中,参数 $Avg(S_t)$表示时刻 t 所有感知数据的精确均值.然而在大规模传感器网络中,我们很难获得 $Avg(S_t)$.因而在实际应用中,我们可以利用 $\frac{\widehat{Avg(S_t)}_B^{(q)}}{1+\varepsilon}$ 代替 $Avg(S_t)$来构建重新抽样的触发条件,其中, $\widehat{Avg(S_t)}_B^{(q)}$ 是 $Avg(S_t)$的(ε,δ)-近似值.

根据(ε,δ)-近似值的计算,我们有 $\Pr\left(\frac{\widehat{Avg(S_t)}_B^{(q)}}{1+\varepsilon}\leqslant Avg(S_t)\right)=\Pr\left(\frac{\widehat{Avg(S_t)}_B^{(q)}}{1-\varepsilon}\geqslant Avg(S_t)\right)\geqslant 1-\frac{\delta}{2}$.

即利用 $\widehat{Avg(S_t)}_B^{(q)}/(1+\varepsilon)$ 代替 $Avg(S_t)$时,可以保证定理 9 为真的概率大于等于 $1-\delta/2$.根据上述分析,近似聚集和的维护算法可描述如下:

首先,对于网络中的任意节点 i,当其当前感知值 $s_{t'i}$ 与最近一次采样发生时刻的感知值 s_{ti} 之差满足 $|s_{it'}-s_{it}|\geqslant\frac{\Delta\widehat{Avg(S_t)}_B^{(q)}}{1+\varepsilon}$ 时,则向簇头节点发送差值 $s_{it'}-s_{it}$.

其次,对于网络中的任意簇 $C_l(1\leqslant l\leqslant k)$,簇头节点将计算其成员节点所汇报的数据的累加和,如果该和的绝对值大于 $\frac{\Delta n_l\widehat{Avg(S_t)}_B^{(q)}}{1+\varepsilon}$ (n_l 为簇 C_l 中的节点的数目),则簇头节点将会沿生成树将该和传送至 Sink 节点,并在传送过程中进行网内聚集.

再次,Sink 节点将会获得所有簇头节点的汇报数据的累加和.Sink 节点将判断该和的绝对值是否大于

$\Delta N\frac{\widehat{Avg(S_t)}_B^{(q)}}{1+\varepsilon}$.如果大于,则 Sink 节点将在汇报数据的簇中触发重新抽样过程,并由 Sink 节点计算新的近似聚集结果.

最后,Sink 计算 $\frac{\widehat{\Delta Avg(S_t)}_B^{(q)}}{1+\varepsilon}$,并将 $\frac{\widehat{\Delta Avg(S_t)}_B^{(q)}}{1+\varepsilon}$ 在网内广播以构成新的重新抽样的触发条件.

将上述算法稍加修改,即可用到近似均值、近似无重复计数的维护上,在此不再赘述.

6 实验结果

本节将通过实验来考察(ε,δ)-近似聚集算法的性能.在下面的实验中,我们利用 ns2 来模拟具有 5 000 个节点的传感器网络.网络中的节点随机地散布在一个 1000m×1000m 的矩形区域内.该区域被分割成 10×10 个网格,处于同一网格内的节点形成一个簇,任意簇均拥有一个簇头节点.

传感器节点的传输半径被设定为 50m.根据文献[23],在上述模拟网络中,节点每发送及接收 1 个字节的消息的能量消耗分别为 0.014 4mJ,0.005 7mJ.由于对于传感器节点来说,发送 1bit 数据的能量消耗相当于执行 1 000 条指令的能量消耗[23],故节点执行指令的能量消耗可忽略不计.

6.1 (ε,δ)-近似聚集算法的性能

第 1 组实验考察(ε,δ)-近似聚集算法所需的样本容量与用户的精度要求(即ε和δ)的关系.在这组实验中,ε变化的范围是 0.1~0.28,δ变化的范围是 0.05~0.23.对于每对ε,δ,我们计算(ε,δ)-近似聚集算法所需的样本容量.实验结果如图 6 所示.由图 6 可见,对于大规模传感器网络来说,(ε,δ)-近似聚集算法所需的样本容量是很小的.例如,对于一个规模为 5 000 的网络来说,当ε=0.15,δ=0.1 时,(ε,δ)-近似聚集算法所需的样本容量是 946,即(ε,δ)-近似聚集算法仅需利用 19%的感知数据,就能保证近似聚集值与精确聚集值的相对误差小于 0.15 的概率大于 0.9;当ε=0.19,δ=0.14 时,(ε,δ)-近似聚集算法所需的样本容量是 536.此时,该算法仅需利用 10%左右的感知数据,就能保证近似聚集值与精确聚集值的相对误差小于 0.19 的概率大于 0.86.上述结果表明,本文提出的(ε,δ)-近似聚集算法能够有效地降低节点的访问量及网络中的通信量,因而,该算法在查询处理过程中的能量消耗将很小.

第 2 组实验考察(ε,δ)-近似聚集算法的精度与ε,δ的关系.在这组实验中,ε变化的范围是 0.12~0.36,δ变化的范围是 0.04~0.36,对于每对ε,δ,我们计算(ε,δ)-近似聚集算法的精度.实验结果如图 7、图 8 所示.由图 7、图 8 可知,当ε≤0.35,δ≤0.2 时,(ε,δ)-近似聚集算法给出的近似聚集和、近似均值的相对误差小于 0.05,近似无重复计数的相对误差小于 0.1;当ε≤0.24,δ≤0.12 时,(ε,δ)-近似聚集算法给出的近似聚集和、近似均值、近似无重复计数已经十分接近真实聚集值.实验结果说明,(ε,δ)-近似聚集算法给出的近似聚集结果的精度很高,完全能够满足用户的任意精度要求.

第 3 组实验考察(ε,δ)-近似聚集算法的精度与抽样概率的关系.在这组实验中,抽样概率将由 0.06 增加到 0.18.对于每一个抽样概率,我们计算(ε,δ)-近似聚集算法的精度.实验结果如图 9 所示,(ε,δ)-近似聚集算法仅需少量的样本数据就能给出精度较高的近似聚集结果.例如,当抽样概率为 0.09 时,即(ε,δ)-近似聚集算法仅需使用 9%的感知数据,其所给出的近似聚集和、近似均值的相对误差将小于 0.05,近似无重复计数的相对误差将小于 0.06.上述实验结果进一步表明,(ε,δ)-近似聚集算法能够在较小的能量开销下给出精度较高的近似聚集结果.

第 4 组实验考察对于给定的ε和δ,网络规模对(ε,δ)-近似聚集算法所需的抽样概率的影响.在这组实验中,ε和δ 的取值为:ε=0.18,δ=0.05;ε=0.23,δ=0.1.我们考察网络规模由 3 000 变化到 10 000 时,(ε,δ)-近似聚集算法所需的抽样概率的大小.实验结果由图 10 给出.由该图可知,当网络规模变大时,(ε,δ)-近似聚集算法的抽样概率明显减小.该实验结果表明,网络规模越大,(ε,δ)-近似聚集算法的性能越好.

第 5 组实验考察网络规模、抽样概率对(ε,δ)-近似聚集算法的精度的影响.在这组实验中,网络规模由 1 000 增加到 5 000,抽样概率的变化范围为 0.1~0.5.对于每对网络规模、抽样概率值,我们计算(ε,δ)-近似聚集算法的精度,实验结果如图 11 所示.由图 1 可知:在中等规模的网络中,(ε,δ)-近似聚集算法仅需使用少于 40%的感知数

据,就能保证近似聚结果的相对误差小于 0.06;当网络规模变大时,(ε,δ)-近似聚集算法给出的近似结果的相对误差将明显减小.该实验结果同样说明了网络规模越大,(ε,δ)-近似聚集算法的性能越好,即(ε,δ)-近似聚集算法尤其适合于在大规模网络中进行近似聚集查询处理.

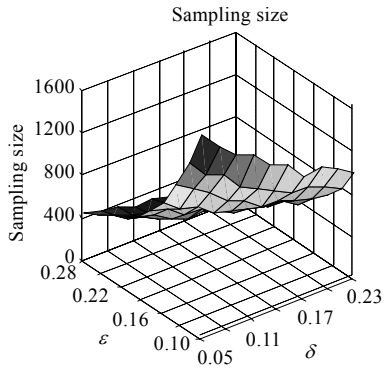


Fig.6 Sample size

图 6 样本容量

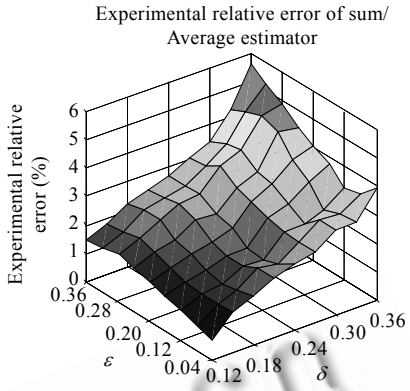


Fig.7 Accuracy of approximate sum (average)

图 7 近似聚集和(均值)的精度

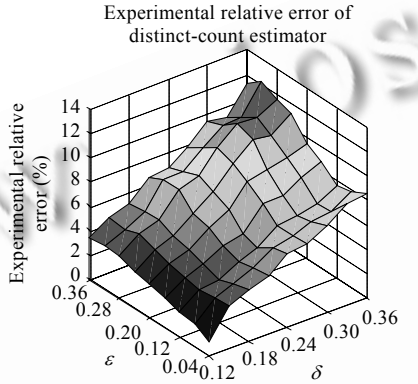


Fig.8 Accuracy of approximate distinct-count result

图 8 近似无重复计数结果的精度

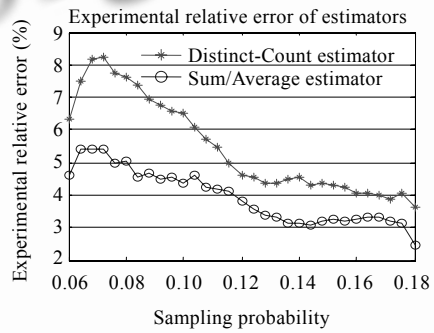


Fig.9 Relationship between algorithm accuracy and sampling probability

图 9 算法精度与抽样概率的关系

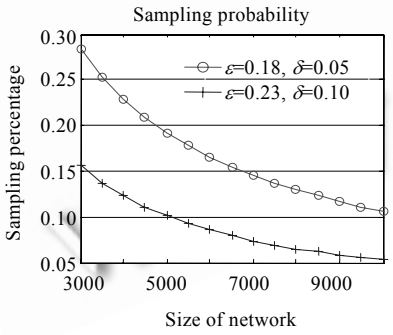


Fig.10 Relationship between sampling probability and network size

图 10 抽样概率与网络规模的关系

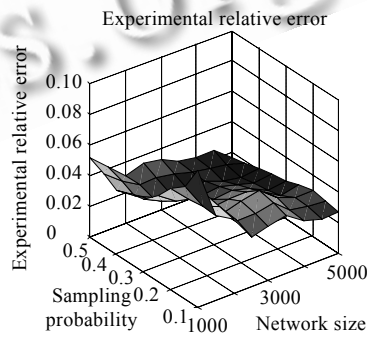


Fig.11 Accuracy affected by network size, sampling probability

图 11 算法精度受网络规模、抽样概率的影响

6.2 能量消耗比较

本节主要对(ε,δ)-近似聚集算法的能量开销与已有的近似聚集算法的能量开销进行对比.

第 1 组实验考察(ε,δ)-近似聚集算法和基于空间相关性的近似聚集算法在处理 snapshot 查询时的能量消耗.在处理 snapshot 查询时,基于空间相关性的近似聚集算法是目前性能最好的聚集算法.在这组实验中,用户给出的精度要求是近似聚集结果的相对误差分别小于 0.099,0.176,0.245,0.304,0.4 及 0.45,我们将考察两种算法达到上述精度的能量开销.由于基于空间相关性的近似聚集算法的误差界不能自动调节,我们将对其进行人为调节,并且不考虑调节过程中的能量消耗.

实验结果如图 12 所示.由图 12 可知,当所需满足的精度要求相同时,(ε,δ)-近似聚集算法的能量开销将远小于基于空间相关性的近似聚集算法的能量开销.原因如下:首先,对于基于时空相关性的近似聚集算法而言,虽然该算法避免了网络中的全部感知数据均参与到聚集运算中,但是其所需使用的感知数据的数目仍然比(ε,δ)-近似聚集算法所需的要多;其次,基于时空相关性的近似聚集算法需要网络中每个节点运行与 Sink 节点类似的推测模型,故这种算法虽然降低了节点与 Sink 的通信量,但无形中增加了许多网络内部的通信,故该类算法在网内执行仍需要耗费较多的能量.

第 2 组实验考察(ε,δ)-近似聚集算法和基于时间相关性的近似聚集算法在处理连续聚集查询时的能量开销.在处理连续查询时,基于时间相关性的近似聚集算法是目前性能最好的聚集算法.在这组实验中,连续查询将持续 30 个周期,每个周期模拟网络中的节点感知数据以 0.8 的概率增加一个小于 10 的随机数,以概率 0.2 减少一个小于 15 的随机数.用户给出的精度要求是近似聚集结果的相对误差分别小于 0.1,0.15,0.2,0.25,0.3,0.35.我们考察两种算法达到上述精度的能量开销.由于基于时间相关性的近似聚集算法的误差界不能自动调节,我们将人为地对其调节,并且不考虑调节过程中的能量消耗.

实验结果如图 13 所示.由图 13 可知,当处理相同精度要求的连续查询时,第 5.3 节介绍的随机算法所消耗的能量小于基于空间相关性的近似聚集算法.原因如下:首先,在初始化查询时,第 5.3 节介绍的随机算法的能量开销远小于基于时间相关性的近似聚集算法的能量开销;其次,当网络中的节点感知数据发生变化时,第 5.3 节介绍的随机算法的过滤能力很强,从而避免了传送那些无用的感知数据.综合考虑,在整个查询处理过程中,第 5.3 节介绍的随机算法所消耗的能量要小一些.

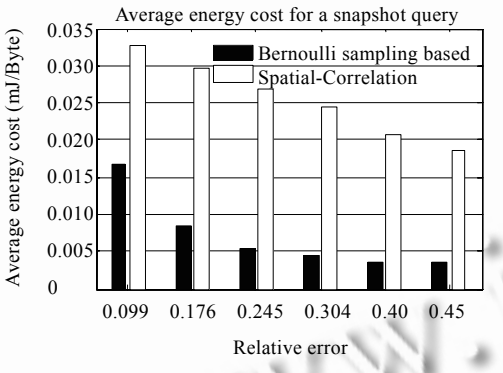


Fig.12 Average energy cost when processing snapshot queries

图 12 处理 Snapshot 查询的平均能量消耗

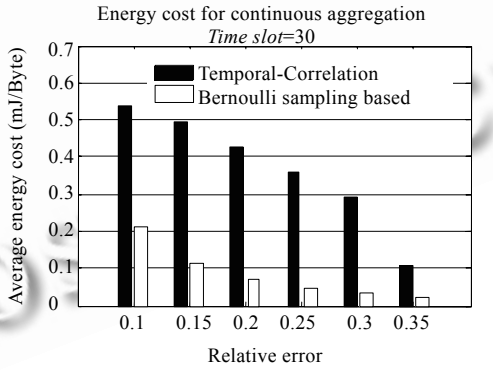


Fig.13 Average energy cost when processing continuous queries

图 13 处理连续查询的平均能量消耗

第 3 组实验考察近似聚集算法处理多查询时的能量开销.由于已有的近似聚集算法均具有固定的误差界,并且很难自动调整,故已有算法几乎不能处理具有不同精度要求的多查询,而本文在第 5.2 节介绍的算法却能处理这类查询.本组实验将主要考虑第 5.2 节介绍的算法的能量开销.在下面的实验中,我们首先令δ=0.1,并计算在ε由 0.5 降至 0.1 的过程中,第 5.2 节介绍的算法的能量开销,实验结果如图 14 所示;其次,令ε=0.1,并计算在δ

由 0.5 降至 0.1 的过程中,第 5.2 节介绍的算法的能量消耗,实验结果如图 15 所示.由图 14、图 15 可知,利用第 5.2 节中介绍的算法进行多查询处理的能量消耗小于 0.008mJ/Byte.将该结果与图 12 所示的实验结果进行比较可知,利用第 5.2 节中介绍的算法处理多查询将节约 50%的能量,因为第 5.2 节中介绍的算法在处理多查询的过程中,尽最大可能地利用了 Sink 节点保存的历史样本信息.

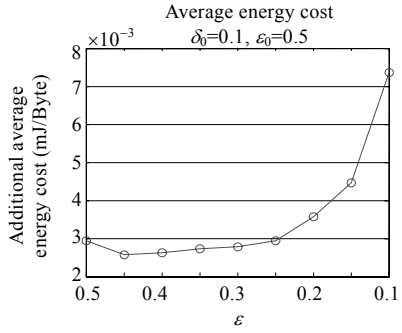


Fig.14 Average energy cost when ε is varying

图 14 ε变化时的平均能量消耗

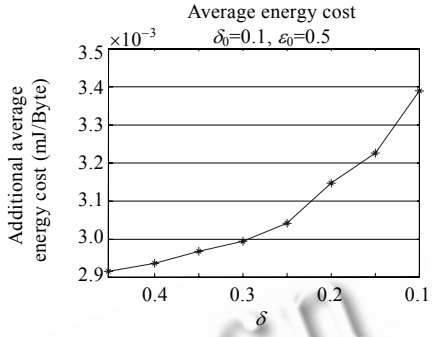


Fig.15 Average energy cost when δ is varying

图 15 δ变化时的平均能量消耗

7 结 论

本文提出了一种基于 Bernoulli 抽样的近似聚集算法,并证明了该算法能够有效地完成具有任意精确度需求的聚集计算.此外,本文还给出了两种自适应算法,分别用于处理用户的精确度需求、网络中的感知数据发生变化的情况.理论分析及实验结果表明,本文提出的算法可有效地处理传感器网络中的聚集查询请求,并在近似结果的精确度、能量开销等方面都优于已有的近似聚集算法.

References:

[1] Madden S, Szewczyk R, Frenklin MJ, Culler D. Supporting aggregate queries over ad-hoc wireless sensor networks. In: Franklin MJ, ed. Proc. of the 4th IEEE Workshop on Mobile Computing Systems and Applications. Washington: IEEE Computer Society Press, 2002. 49−58.

[2] Zhao J, Govindan R, Estrin D. Computing aggregates for monitoring wireless sensor networks. In: Kindberg T, ed. Proc. of the 1st IEEE Int'l Workshop on Sensor Network Protocols and Applications. Washington: IEEE Computer Society Press, 2003. 139−148.

[3] Madden S, Franklin MJ, Hellerstein JM, Hong W. The design of an acquisitional query processor for sensor networks. In: Franklin MJ, Moon B, Ailamaki A, eds. Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2003. 491−502.

[4] Zhao J, Govindan R. Understanding packet delivery performance in dense wireless sensor networks. In: Joseph A, Seitz J, Tobe Y, eds. Proc. of the ACM Conf. on Embedded Networked Sensor Systems. New York: ACM Press, 2003. 1−13.

[5] Madden S, Franklin MJ, Hellerstein JM, Hong W. TAG: A tiny aggregation service for ad-hoc sensor networks. In: Culler D, ed. Proc. of the 5th Symp. on Operating System Design and Implementation. New York: ACM Press, 2002. 131−146.

[6] Considine J, Li F, Kollios G, Byers J. Approximate aggregation techniques for sensor databases. In: Goldin D, ed. Proc. of the 20th Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2004. 449−460.

[7] Deligiannakis A, Kotidis Y, Rossopoulos N. Processing approximate aggregation queries in wireless senor networks. Information Systems, 2006,31(8):770−792. [doi: 10.1016/j.is.2005.02.001]

[8] Hartl G, Li BC. Infer: A Bayesian inference approach towards energy efficient data collection in dense sensor networks. In: Takizawa M, Papazoglou MP, Sinha P, eds. Proc. of the 25th IEEE Int'l Conf. on Distributed Computing Systems. Washington: IEEE Computer Society Press, 2005. 371−380.

[9] Nath S, Gibbons PB, Seshan S, Anderson ZR. Synopsis diffusion for robust aggregation in sensor networks. In: Goldin D, ed. Proc. of the ACM Conf. on Embedded Networked Sensor Systems. New York: ACM Press, 2004. 250−262.

[10] Deligiannakis A, Kotidis Y, Roussopoulos YN. Hierarchical in-network data aggregation with quality guarantees. In: Goldin D, ed. Proc. of the Int'l Conf. on Extending Database Technology. Washington: IEEE Computer Society Press, 2004. 658−675.

[11] Cormode G, Garofalakis MN, Muthukrishnan S, Rastogi R. Holistic aggregates in a networked world: distributed tracking of approximate quantiles. In: Ozcan F, ed. Proc. of the 2005 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2005. 25−36.

[12] Chu D, Deshpande A, Hellerstein JM, Hong W. Approximate data collection in sensor networks using probabilistic models. In: Barga RS, Zhou XF, eds. Proc. of the 22nd Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2006. 48−59.

[13] Silberstein A, Puggioni G, Gelfand A, Munagala K, Yang J. Suppression and failures in sensor networks: A Bayesian approach. In: Koch C, Gehrke J, Garofalakis MN, eds. Proc. of the 33rd Int'l Conf. on Very Larger Data Base. New York: ACM Press, 2007. 842−853.

[14] Haas PJ, Swami AN. Sequential sampling procedures for query size estimation. In: Stonebraker M, ed. Proc. of the 1992 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1992. 341−350.

[15] Larson P, Lehner W, Zhou JR, Zabback P. Cardinality estimation using sample views with quality assurance. In: Chan CY, Ooi BC, Zhou AY, eds. Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2007. 175−186.

[16] Cemulla R, Lehner W, Haas PJ. Maintaining Bernoulli sample over evolving multisets. In: Chan CY, Ooi BC, Zhou AY, eds. Proc. of the 26th ACM SIGMOD- SIGACT-SIGART Symp. on Principles of Database Systems. New York: ACM Press, 2007. 93−102.

[17] Benjamin A, Gautam D, Dimitrios G, Vana K. Approximating aggregation queries in peer-to-peer networks. In: Barga RS, Zhou XF, eds. Proc. of the 22nd Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2006. 642−654.

[18] Benjamin A, Lin S, Gunopulos D. Efficient data sampling in heterogeneous peer-to-peer networks. In: Korpeoqlu I, ed. Proc. of the 7th IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society Press, 2007. 28−31.

[19] Bayer K, Haas PJ, Reinwald B. On synopses for distinct-value estimation under multiset operations. In: Chan CY, Ooi BC, Zhou AY, eds. Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2007. 199−210.

[20] Bernstein S, Bernstein R, Wrote; Shi DJ, Trans. Elements of Statistics II: Inferential Statistics. Beijing: Science Press, 2002. 83−84 (in Chinese).

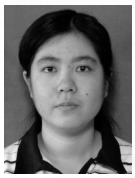
[21] Flajolet P, Martin GN. Probabilistic counting algorithms for data base applications. Journal of Computer and System Sciences, 1985,31(2):182−209. [doi: 10.1016/0022-0000(85)90041-8]

[22] Huang GY, Li XW, He J. Dynamic minimal spanning tree routing protocol for large wireless sensor networks.. In: Koch C, ed. Proc. of the 1st IEEE Conf. on Industrial Electronics and Applications. Washington: IEEE Computer Society Press, 2006. 1−5.

[23] Crossbrow Technology, Inc. MPR-Mote Processor Radio Board User's Manual. San Jose: Crossbrow Technology, Inc., 2003.

附中文参考文献:

[20] Bernstein S,Bernstein R,著;史道济,译.统计学原理——推断统计学(下册).北京:科学出版社,2002.83−84.



程思瑶(1982−),女,黑龙江佳木斯人,博士生,主要研究领域为传感器网络,对等计算.

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