Journal of Computer Applications
计算机应用,2014,34(2):313 -317
ISSN 1001-9081
CODEN JYIIDU
2014-02-10
http://www.joca.cn

文章编号:1001-9081(2014)02-0313-05
doi:10.11772/j.issn.1001-9081.2014.02.0313

无线传感器网络不重复记录求和近似算法

刘彩苹[1],蔡玉武[1*],毛建旭[2],蔡玉文[3]
(1.湖南大学 信息科学与工程学院,长沙 410082; 2.湖南大学 电气与信息工程学院,长沙 410082;
3.中南民族大学 生物医学工程学院,武汉 430000)
(*通信作者电子邮箱 1848886571@qq.com)

摘 要:针对现有的求和算法基本上都是对副本敏感的算法,提出一种对副本不敏感的求和近似算法 FM-S。网络中各节点由 FM-S 和服从二项分布的随机数样本对节点记录进行哈希转换以填充一个长度为 L 的二进制求和序列,并且每个节点会把生成的序列转发给路由树中的父亲节点,根节点将接收到全网的求和序列,最终根据此序列可计算出网络中不重复记录求和的近似值。实验结果显示该算法是一种分布式、低功耗、容错性高、扩展性和健壮性强的聚集查询算法。

关键词:无线传感器网络;分布式算法;求和查询;近似算法;聚集查询
中图分类号:TP212.9 **文献标志码**:A

Approximate summation algorithm of distinct records for wireless sensor network

LIU Caiping[1], CAI Yuwu[1*], MAO Jianxu[2], CAI Yuwen[3]
(1. *School of Information Science and Engineering, Hunan University, Changsha Hunan* 410082, *China*;
2. *College of Electrical and Information Engineering, Hunan University, Changsha Hunan* 410082, *China*;
3. *College of Biomedical Engineering, South-Central University for Nationalities, Wuhan Hubei* 430000, *China*)

Abstract: Since the existing summation aggregation algorithms are almost duplicate-sensitive, an approximate algorithm Flajolet-Martin SUM (FM-S) of distinct summation query for Wireless Sensor Network (WSN) was proposed. In FM-S, each node in WSN combined the FM-S algorithm and the random number sample of binomial distribution to do hash conversion so as to fill a summation sequence of length L, and each node forwarded the generated sequence to the father node in routing tree. Then the root node received the summation sequence of whole network. Finally, according to the sequence of root node, the approximation summation value of distinct records in sensor networks could be obtained. The experimental results show that the distributed algorithm is of low power consumption, high fault tolerance, robustness and scalability.

Key words: Wireless Sensor Network (WSN); distributed algorithm; summation query; approximate algorithm; aggregate algorithm

0 引言

目前,无线传感器网络广泛应用于交通管制[1]、医疗卫生[2]和建筑物健康监测[3]等领域。聚集运算操作经常在无线传感器网络的数据查询处理中使用,SUMmation(SUM)聚集查询是将传感器网络中所有节点采集的数据进行求和。SUM 聚集对重复数据是敏感的,大量冗余数据的存在会严重影响最终的聚集结果,但是在很多的情况下,并不需要知道传感器网络 SUM 聚集的准确值,可以用近似值来代替准确值。本文的 FM-S(Flajolet-Martin SUM)算法就是一种近似 SUM 聚集算法,该算法有效地解决了 SUM 聚集对重复数据敏感的问题,能够提高网络的整体性能。

为了减少网络的通信开销,延长网络的寿命,在实际中会采用分布式聚集查询。现在大部分分布式聚集查询的核心思想是 Madden 等[4]提出的 Tiny Aggregation (TAG)算法,及 TinyDB[5]数据库系统。算法思想是将聚集查询通过类 SQL 语句下发到整个传感器网络,并由路由算法生成一棵代表全网节点的路由树。在整个聚集过程中,节点将会把接收到来自子节点的数据集和自己生成的数据集合并生成一个新的数据集,并把新的数据集发送给父节点,最终汇聚节点会得到一个包含了全网所有数据的聚集值。

在现实环境中,传感器节点的部署非常复杂,节点随时都有可能因为电量耗尽而失效,很容易导致传感器网络拓扑结构的变化;同时节点在发送数据时由于无线信道存在环境干扰、包冲突和信噪比低从而导致丢包和通信连接失败,就会破坏生成的路由树,从而会严重影响最终的聚集值[6];如果选择数据重发则会消耗更多的能量,缩短网络的寿命[7],同时也会造成数据的冗余,最终会严重影响对副本敏感的聚集,例如 COUNT 和 SUM。

如何在环境异常复杂,通信能力、计算能力和能量非常有

收稿日期:2013-07-18;**修回日期**:2013-10-11。 **基金项目**:国家自然科学基金资助项目(61072121);湖南省自然科学基金资助项目(12JJ2035);湖南大学青年教师成长计划项目(531107040287)。

作者简介:刘彩苹(1978 -),女,湖南邵阳人,讲师,主要研究方向:无线传感器网络、数据挖掘; 蔡玉武(1990 -),男,湖南娄底人,硕士研究生,主要研究方向:无线传感器网络、数据挖掘; 毛建旭(1974 -),男,江西安义人,副教授,主要研究方向:无线传感器网络、数字图像处理、模式识别; 蔡玉文(1988 -),男,湖南娄底人,硕士研究生,主要研究方向:生物医学工程医学传感器。

限的无线传感器网络中进行包括SUM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。

Liu等[8]提出了基于节点历史数据代表路由树的近似查询算法。算法的主要思想是通过节点间的相关度构建代表路由树进行近似聚集查询。由于查询时只要遍历代表路由树,因此可以减少通信开销,延长网络寿命;但结果容易受网络拓扑结构变化的影响。

Considine等[9]提出了基于Sketch的近似聚集查询算法。算法的核心思想是采用Sketch技术对节点的原始记录数据进行压缩传输,并在网内进行聚集。与其他算法相比,该算法传输的是压缩数据;但是,该算法需要全网的原始数据参与聚集运算,因此通信开销比较大。

Xin等[10]提出了一种基于时间相关性的近似聚集算法。主要思想是为节点设置一个筛选范围,节点数据不在筛选范围内时才将数据传输给汇聚节点,否则汇聚节点采用历史数据计算近似聚集结果。能有效应用于数据连续的聚集;但是算法误差较大。

Su等[11]提出了一种基于空间相关性的近似聚集算法。算法核心思想是网络中各节点建立代表模型,当某一个节点的数据不能被其他节点代表时,才将数据传输给汇聚节点。算法处理Snapshot查询时效率较高;但数据流连续的查询效率较低。

目前,还没有一种有效的算法既能够提高传感器网络的性能,又能解决SUM聚集对副本敏感的问题。本文在FM(Flajolet-Martin)估计计数算法的基础上提出一种对副本不敏感的聚集求和近似算法,记为FM-S。该算法在FM估计计数算法的理论基础上总结出了一种生成二进制求和序列的高效算法,根据该序列可以近似求出传感器网络的SUM聚集值,同时,FM-S算法引入了多路复用(Multi-Path Routing)技术和重复不敏感策略(Duplicate Insensitive Sketches)。FM-S算法在进行SUM聚集操作时对重复数据不敏感,可以得到求和的近似值。例如,节点某一时刻的记录值集合为 $M = \{x_1, x_2, x_3, x_4, x_5\}$,$x_i = (p_i, c_i, t_i)$,其中:$p_i$ 表示传感器节点,c_i 表示观测值,t_i 表示时间;只有当节点记录值的 p_i, c_i, t_i 参数都相同时才算是节点重复的感知数据。假设 x_1 和 x_3,x_2 和 x_4 是重复记录,简单来说,如果使用TAG算法计算 $SUM = c_1 + c_2 + c_3 + c_4 + c_5$,而使用FM-S算法近似计算 $SUM \approx c_1 + c_2 + c_5$,与TAG算法相比,FM-S算法既提高了结果的准确性,又提高了网络的性能。

1 不重复记录求和查询近似算法FM-S

不重复记录求和近似算法FM-S分成以下三个步骤:

步骤1 开始时基站节点采用洪泛的方式将SUM查询传发到传感器网络中所有节点,并且将所有传感器节点按照路由算法生成一棵路由树。

步骤2 各节点把从子节点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 $M = \{x_1, x_2, x_3, \cdots\}$,对新数据集中的每一条记录执行分布式FM-S算法生成一个长度为 L 的求和序列 $FM\text{-}S(M)[0,1,\cdots,L-1]$,每个节点再将多个求和序列组织成一个二维序列发送给上一层节点。

步骤3 重复执行步骤2,根节点会得到一个可用于代表全网数据集结构的 $FM\text{-}S(M)$ 二维序列,最终由这个 N 行 $FM\text{-}S(M)$ 二维序列估算出全网的不重复记录的SUM值。

1.1 分布式FM-S算法

FM近似计数查询算法最先由Flajolet和Martin在文献[12]中提出,FM算法最大的好处是不需要多次扫描数据库或数据流,只需要一次扫描就可以快速估计出数据库或数据流中不重复记录的个数,并且算法只要在一块远小于数据规模的内存空间里不断更新一个代表数据集的结构——概要数据结构,并且随时都能够根据这个结构快速获得近似查询的结果。

引理1[12] 当 $i < \text{lb}\, n - 2\, \text{lb lb}\, n$($n$ 为不重复记录的个数)时,$FM(M)[i]$ 被设置为1的概率几乎为100%;当 $i \geqslant \frac{3}{2}\, \text{lb}\, n$ 时,$FM(M)[i]$ 被设置为0的概率几乎为100%。

引理1在文献[12]给出了详细的证明,它是FM-S算法近似计算不重复记录求和的理论依据。

分布式FM-S算法所涉及的参数定义如下:

$p_1, p_2, \cdots, p_n$ 代表 n 个任意相连的无线传感器节点。

M 为节点 p_i 在一段时间内生成的数据集,$M = \{x_1, x_2, x_3, \cdots\}$,$x_i = (p_i, c_i, t_i)$。其中:$p_i$ 表示传感器节点,c_i 表示观测值,t_i 表示时间,只有节点记录值的 p_i, c_i, t_i 参数都相同时才说明是同一个传感器节点在同一时刻生成的数据,即为重复数据。

$FM\text{-}S(M)$ 表示数据集 M 按FM-S算法生成长度为 L 的二进制求和序列,记为 $FM\text{-}S(M)[0,1,\cdots,L-1]$。

SUM 表示通过运算得到的不重复记录求和的近似值。

N 表示整个网络中所有的数据子集的并集 M 中不重复记录的个数。

Y 表示随机产生符合 $B(c_i, 2^{-\delta})$ 二项分布的随机样本,其中 $\delta = \lfloor \text{lb}\, c_i - 2\, \text{lb lb}\, c_i \rfloor$。

Hash(x) 表示对样本中的每一个值 x 进行Hash变换的函数,经过Hash变换后会得到一个长度为 L 的二进制序列 y。

$bit(y,k)$ 表示二进制序列 y 中的第 k 位。

$p(y)$ 表示二进制序列 y 中最右边的1所处的位置,其公式定义为:

$$p(y) = \min\{k \mid bit(y,k) = 1\} \tag{1}$$

FM-S算法的特点就是能用一小块远小于数据集数据范围的内存空间表示数据集,用 $FM\text{-}S(M)$ 表示将数据集 M 进行FM-S运算后得到的长度为 L 的二进制序列 $y(L \ll M)$,表示为 $FM\text{-}S(M)[0,\cdots,L-1]$,其公式定义如下:

$$FM\text{-}S(M)[i] = 1, \text{iff} \exists x \in M, p(\text{Hash}(x)) = i \tag{2}$$

在上述理论的基础上,不重复记录求和的定义如下:

定义1 给定一个数据集 $M = \{x_1, x_2, x_3, \cdots\}$,其中 $x_i = (p_i, c_i, t_i)$,不重复记录求和可以记为:

$$SUM = \sum_{j=1}^{N} \sum_{distinct((p_i,c_i,t_i) \in M)} c_i \tag{3}$$

其中 c_i 严格保持唯一性,当 c_i 的值较小时,可以使用FM计数序列的方法进行近似求和,即将 c_i 的值转化为计数 c_i 个非重复记录。例如:假如非重复记录 c_i 的值为5,则此时可以转化

成5条非重复记录$(p_i,c_i,t_1,1),\cdots,(p_i,c_i,t_i,5)$,利用分布式FM非重复计数算法填充FM-S序列,然后根据FM-S序列求出的非重复记录数就是c_i的值。

当c_i的值比较小时,可以用估计计数算法来计算SUM的值,并且它与估计计数算法具有一样的准确性和时间复杂性$O(c_i)$;但是当c_i的值很大时,这种方法的使用效果并不是很理想。因此,必须采用一种高性能的算法来处理c_i值很大时的情况。

为了充分利用FM算法思想来计算非重复求和,可以生成一个二进制求和序列(Summation Sketches),然后利用FM-S估计方法求出c_i的值。可以采取一种比较有效的方法来生成求和序列,方法具体分为两步:1)根据引理1,当$i < \text{lb}\ n - 2\ \text{lb lb}\ n$时,$FM\text{-}S(M)[i]$被设置为1的概率几乎为100%;因此首先设置求和序列的前$\delta = \lfloor \text{lb}\ c_i - 2\text{lb lb}\ c_i \rfloor$位全部为1。2)通过生成符合要求分布的随机样本,并且对样本中的每个值采用Hash变换来填充求和序列余下的$L-\delta$位。最终,可以生成一个长度为L的二进制求和序列$FM\text{-}S(M)[0,\cdots,L-1]$。

由FM-S算法可知,一个记录值x_i的二进制求和序列第z位是1必须满足$\forall_{0\leqslant j<z}(h(x_i,j)=0)$,同时也等价于关系表达式$\min\{j \mid h(x_i,j)=1\} \geqslant z$,并且这种情况出现的概率为$2^{-z}$。因此,记录值$c_i$的求和序列中第$\delta$位后出现1的概率是一个服从参数为$c_i$和$2^{-\delta}$的二项分布(Binomial Distribution)。首先,获取一个服从$B(c_i,2^{-\delta})$二项分布的随机样本Y,并且该样本Y中的每一个值进行Hash转换后使1出现的位置在记录值c_i求和序列第δ位后,然后将样本Y中的所有值进行Hash变换后来填充求和序列余下的$L-\delta$位。

本文算法中将$FM\text{-}S(M)[0,1,\cdots,L-1]$中最左边的0所在的位置记为大小为$\text{lb}\ n$的标志性位,可以近似计算传感器网络中非重复记录$c_i$的值。

标志位的公式定义如下:

$$R = leftmost\{i \mid FM(M)[i]=0\} \tag{4}$$

引理2[12]　r为运行FM-S算法得到的标记位,r的期望值$E(r) \approx \text{lb}(\varphi c_i)$,其中$\varphi = 0.775\,351$,$r$的方差$\sigma(r) \approx 1.12$。

从引理2可以知道,使用φ纠错性标志位能够获取c_i相对准确的估计值$SUM(i)$,其计算表达式如下所示:

$$SUM(i) = (1/\varphi)2^r \tag{5}$$

根据上文的分析可知,在节点p_i运行分布式FM-S算法的伪代码如下所示:

```
Program SumInsert( M_i )
Input: M_i is Multiset of data in p_i
Output: FM-S[n][0,…,L-1]
Do while not eof( M_i )
{
  c ← getElement(M_i)
  d ← select_prefix(c)
  /* 设置FM-S序列的前 d = ⌊lb c - 2 lb lb c⌋ 位全部为1 */
  j ← 0
  while j < d do
    j ← j + 1
    FM-S[i][j] ← 1
  end while
  Y[a] ← select_sample(seed = (x,c),c,1/2^d)
  /* 随机生成符合 B(c,1/2^d) 二项分布的样本 Y,大小为 a */
  for k ← 0 to a
    j ← d
    /* 对样本中每一个随机变量进行Hash生成二进制序列,并且设置的二进制位置在第 d 位之后 */
    while Hash(x,c,Y[k],j) = 0 do
    j ← j + 1
    end while
    FM-S[i][j] ← 1
  end for
}
return FM-S[n][0,…,L-1]
```

例如,当数据集M中c的记录值为16384时,使用FM-S算法生成FM-S二进制求和序列值的过程如下:

1)由$c=16384$知$d=\lfloor \text{lb}\ c - 2\ \text{lb lb}\ c \rfloor = 6$,于是设置FM-S序列的前6位均为1;

2)产生一个随机数样本Y,它满足二项分布$B(16384,1/2^6)$,开始会产生16384个满足均匀分布$U(0,1)$的随机数$r_1,r_2,\cdots,r_{16384}$;然后统计$r_i(i=1,2,\cdots,16384)$中使得$r_i < 1/2^6$的个数n_i,重复循环得到:$n_1,n_2,\cdots,n_k$即为所求随机数列,本例重复了20次得到一个大小为20的样本Y,且$Y=\{128,320,412,1440,512,1240,2786,124,256,526,1024,3084,784,2662,4096,1320,3544,2048,1632,128\}$(1.2节给出了具体算法),对样本中每一个值进行Hash(x)变换生成长度为16位的二进制序列,最终生成的FM-S序列为1111 1111 1111 0100,在这个FM-S序列中,最左边的0位于第13位,由式(5)可以计算得到聚集求和的近似值$SUM(i)=10566$。

对数据集$M=\{x_1,x_2,x_3,\cdots\}$中每一个记录值分别生成FM-S序列,最终根节点会收到一个二维求和序列$FM\text{-}S[N][0,\cdots,L-1]$,由式(5)可知,即可求出$SUM=\sum_{i=0}^{N-1} SUM(i)$。

例如,假设数据集$M=\{(1,1024,12:30),(2,16384,12:40),(3,4096,12:50),(4,32768,12:50),(2,16384,12:40),(4,32768,12:50)\}$是某一个传感器节点$P$在一段时间内生成的数据集,其中有重复记录值16384和32768。

对节点P运行FM-S算法会输出一个4行16列的二维数组,如下所示:

1111 1111 0110 0000
1111 1111 1111 0100
1111 1111 1101 1000
1111 1111 1111 1010

由式(5)可知,

$$SUM = \sum_{i=0}^{3} SUM(i) = \sum_{i=0}^{3} (1/\varphi)2^{R_i} =$$
$$660 + 10565 + 2641 + 21125 = 34991$$

最终求出非重复记录的近似和为34991,实际和值为54272,而使用TAG计算得到SUM值为103424。

定理 1 一个记录 $x_i = (p_i, c_i, t_i)$ 生成求和序列的时间复杂度为 $O(\mathrm{lb}^2 c_i)$。

证明 α_i 表示要生成求和序列余下 $L - \delta_i$ 位二项分布样本的个数，生成求和序列前面 δ_i 位的时间复杂度为 $O(\delta_i)$，α_i 个样本值 Hash 转换生成求和序列的时间复杂度为 $O(\alpha_i)$，生成 x_i 求和序列总的时间复杂度为 $O(\delta_i + f(\alpha_i) + \alpha_i)$，其中 $f(\alpha_i)$ 表示选取样本大小为 α_i 的时间。因此，时间取决于 α_i 和用于选取 α_i 的方法，由二项分布期望公式知 α_i 的期望值如下：

$E(\alpha_i) = c_i \times 2^{-\delta}, \delta = \lfloor \mathrm{lb}\ c_i - 2\ \mathrm{lb}\ \mathrm{lb}\ c_i \rfloor$，即

$\mathrm{lb}^2 c_i \leqslant E(\alpha_i) < 2\ \mathrm{lb}^2 c_i$

所以 FM-S 算法改进后生成一个记录的求和序列的时间复杂度为 $O(\mathrm{lb}^2 c_i)$。

1.2 产生二项随机数

设随机变量 X 的分布律为 $P\{X = x_i\} = p(i = 1,2,\cdots)$，令 $CDF^{(0)} = 0, CDF^{(1)} = p_1, \cdots, CDF^{(n)} = \sum_{i=1}^{n} p_i (n = 1,2,\cdots)$。

定理 2 若 $F(x)$ 是任意随机变量 X 的累计分布函数（Cumulative Distribution Function, CDF），则 $Y = F(x)$ IID U(0,1)，且与 X 的分布特性无关，其中 IID 表示独立同分布（Independently and Identically Distribute）。

证明 令 $Y = F(x)$，$F(x)$ 是 X 的 CDF；Y 也是一个随机变量，令 $G(y)$ 为 Y 的 CDF。

$$G(y) = P(Y \leqslant y) = P(F(x) \leqslant y) = P(x \leqslant F^{-1}(y)) = F(F^{-1}(y)) = y$$

其中 $F(x)$ 具有单调非降特性。

即：$G(y) = y$；Y 是在[0, 1]区间上具有均匀分布特性随机变量的 CDF，因此，$Y = F(x)$ IID U(0,1)；显然 $G(y) = y$ 与 X 的分布特性无关。

若随机变量 $R \sim U(0,1)$，也就是说服从均匀分布，并且有

$CDF\{CDF^{(n-1)} < R \leqslant CDF^{(n)}\} = CDF^{(n)} - CDF^{(n-1)} = p_n;\ n = 1,2,\cdots$

令 $\{CDF^{(n-1)} < R \leqslant CDF^{(n)}\} = \{X = x_n\}$，则有 $CDF\{X = x_n\} = p_n (n = 1,2,\cdots)$。

产生 X 随机数的算法步骤如下：

1）产生一个(0,1) 区间上均匀分布随机数 r(RND)；

2）若 $CDF^{(n-1)} < r \leqslant CDF^{(n)}$，则令 X 取值为 x_n。

产生随机数算法的伪代码如下所示：

```
unif = a random number
oldcdf = 0
newcdf = 0
m = 1
DO
  newcdf = oldcdf + f( v( m) )
  if( oldcdf < unif < newcdf)
     draw = v( m)
     exit do loop
  else
    m = m + 1
    oldcdf = newcdf
  end if
END DO
print draw
```

如果随机变量 X 服从二项分布 B(n,p)，并且分布律满足的条件为 $P\{X = k\} = C_n^k p^k (1-p)^{n-k} (k = 1,2,\cdots,n)$，其中 $0 < p < 1$，随机变量 X 是在 n 次独立贝努里试验中事件 A 发生的总次数，记 $p = P(A)$。

产生满足二项分布 B(n,p) 随机数列的算法步骤如下：

1）产生 n 个满足均匀分布 U(0,1) 的随机数 $r_1, r_2, \cdots, r_n$；

2）统计随机数 $r_i (i = 1,2,\cdots,n)$ 中满足条件 $r_i \leqslant p$ 时 r_i 的个数记为 n_i。

算法重复进行多次可以产生满足要求的随机数序列 $n_1, n_2, \cdots, n_k$。

2 仿真实验及结果分析

2.1 仿真实验

仿真实验是在无线传感器网络仿真程序 TAG[4] 上进行的。在实验中，节点的数目由网络直径大小决定，节点个数是网络直径大小的平方，通过改变网络直径和丢包率等参数来进行实验。在本文算法中，每个节点要传送一个 10 行 16 列 FM-S 二位数组，每个节点要传送的无压缩数据 40 字节。本文算法采用了文献[12]中提到的压缩技术进行压缩，能将数据压缩至原来的 1/4 左右，显著减少了网络的传输量。

选取 TAG 系统中 SUM 算法（TAG-S）和本文的 FM-S 算法进行对比实验。

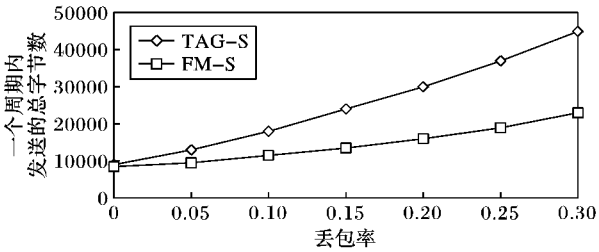

图 1 节点发送的总字节数和网络丢包率的关系

TAG 仿真程序模拟现实环境中存在丢包时各节点传输数据的情况。当有丢包发生时，TAG-S 算法会选择重发丢失的包，而 FM-S 采用了多路复用技术和重复不敏感策略，不需要重新发送也能获取比较准确的聚集数据，因此 FM-S 算法能够显著减少数据的发送量，延长网络的寿命。在这种环境中，会有大量的数据包无法发送至汇聚节点，必然会存在误差。其中，误差率的公式定义如下：

$$\text{误差率} = \left| \frac{x_{\text{realistic}} - x_{\text{correct}}}{x_{\text{correct}}} \right|$$

图 2 考察在模拟现实网络环境下的平均误差率与网络直径的关系。从图 2 可知，TAG-S 算法误差非常大，而本文 FM-S 算法可将误差范围控制在 25% ~35%。

图 3 考察在模拟现实网络环境下 TAG-S 和 FM-S 算法随着网络直径增加得到 SUM 聚集平均值的情况。从图 3 可知，在模拟现实网络环境下，TAG-S 算法得到的实验结果范围非常大，而 FM-S 实验结果范围相对较小，而且随着网络直径的增加，两种算法实验结果范围的差距越来越明显。

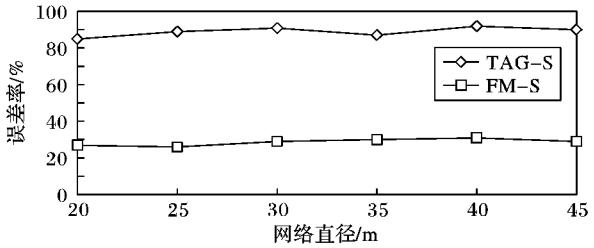

图 2　平均误差与网络直径的关系

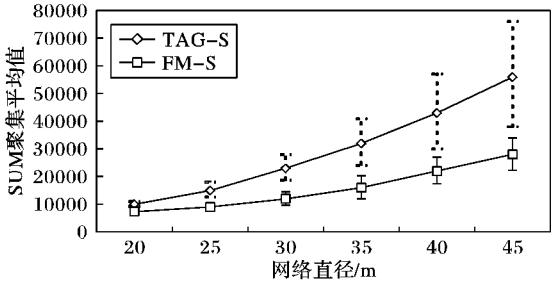

图 3　SUM 平均值随着网络直径增加的情况

图 4 考察两种算法随着网络丢包率增加得到 SUM 聚集平均值的情况。从图 4 可知,TAG-S 算法得到的实验结果范围非常大,而 FM-S 实验结果范围相对较小。

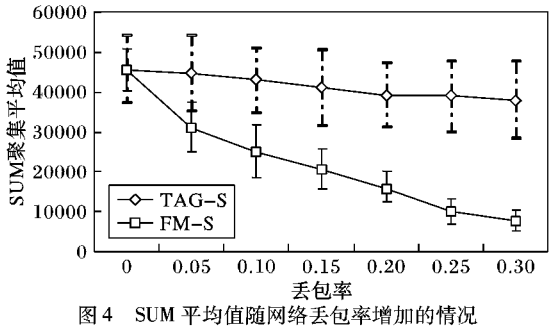

图 4　SUM 平均值随网络丢包率增加的情况

2.2　算法性能分析

与 TAG-S 算法相比,FM-S 算法具有以下特性:

1)功耗低。TAG-S 算法会将所有数据全部上传至汇聚节点,当有丢包发生时,还会选择重发,虽然在传输过程中会将重复的数据删除,但仍然需要传送大量数据;而本文算法每个节点都只需要传送一个经过压缩的 FM-S 二维数组,需要传输的数据量远少于 TAG-S 算法,从而降低了节点能量的消耗,延长了网络的寿命,图 1 的结果很好地说明了这一点。

2)不受重复数据的影响。TAG-S 算法没有考虑重复数据的影响,如果有丢包发生,TAG-S 算法会选择重发,同一条数据会在网络中存在多份,将会严重影响到最终的聚集结果,而本文算法充分考虑到这个缺点,采用 FM-S 方法对副本不敏感的算法,确保最终结果的正确性。

3)误差较小。在存在丢包的模拟现实环境下,TAG-S 算法有很大的误差;而本文算法采用多路复用技术和重复不敏感策略,即使网络中有丢包发生,也不会重发,并且生成的二进制求和序列 FM-S 对重复数据不敏感,极大提高了网络的容错性,因此本文误差较小,图 3 和图 4 的结果很好地说明了这一点。

3　结语

本文提出了一种基于无线传感器网络的不重复记录求和近似算法。与其他 SUM 聚集算法不同,FM-S 算法对重复数据不敏感,采用重复不敏感策略来处理重复数据,并且引入了多路复用技术,确保了算法具有相对较小的误差,提高了网络的容错性,同时也减少了网络的通信量。因此,FM-S 算法可以应用于节点能量有限,又注重 SUM 查询效率,并且聚集结果不要求十分准确的无线传感器网络中。

参考文献:

[1] BARBAGLI B, BENCINI L, MAGRINI I, *et al.* A real-time traffic monitoring based on wireless sensor network technologies[C]// Proceedings of the 7th International Wireless Communications and Mobile Computing Conference. Washington, DC: IEEE Computer Society, 2011: 820 - 825.

[2] ZHANG F, DISANTO W, REN J, *et al.* A novel CPS system for evaluating a neuralmachine interface for artificial legs[C]// Proceedings of 2011 IEEE/ACM International Conference on Cyber-Physical Systems. Washington, DC: IEEE Computer Society, 2011: 67 - 76.

[3] BOCCA M, TOIVOLA J, ERIKSSON M, *et al.* Structural health monitoring in wireless sensor networks by the embedded Goertzel algorithm[C]// Proceedings of 2011 IEEE/ACM International Conference on Cyber-Physical Systems. Washington, DC: IEEE Computer Society, 2011: 206 - 214.

[4] MADDEN S, FRANKLIN M J, HELLERSTEIN J M, *et al.* TAG: a Tiny AGgregation service for Ad-Hoc sensor networks[C]// Proceedings of the 5th Symposium on Operating Systems Design and Implementation. New York: ACM, 2002: 131 - 146.

[5] THIAGARAJAN A, MADDEN S. Representing and querying regression models in a DBMS[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008: 284 - 292.

[6] HU W, YANG H, HUANG L . Multidimensional data reduction based on compressed sensing for sensor network[C]// Future Wireless Networks and Information Systems. Heidelberg: Springer, 2012: 721 - 728.

[7] SRIVASTAVA N. Challenges of next-generation wireless sensor networks and its impact on society[J]. Journal of Telecommunication, 2010, 1(1): 128 - 133.

[8] LIU Y, LIANG W. Approximate querying in wireless sensor networks[C]// Proceedings of the 3th International Conference on Pervasive Computing and Applications. Piscataway: IEEE, 2008: 145 - 145.

[9] CONSIDINE J, HADJIELEFTHERIOU M, LI F, *et al.* Robust approximate aggregation in sensor data management systems[J]. ACM Transactions on Database Systems, 2009, 34(1): 1 - 35.

[10] XIN J, WANG X J, CHEN L. *et al.* Energy-efficient evaluation of multiple skyline queries over a wireless sensor network[C]// Proceedings of the 14th International Conference on Database Systems for Advanced Applications. Heidelberg: Springer, 2009: 247 - 262.

[11] SU F, CHUNG Y, LEE C, *et al.* Efficient skyline query processing in wireless sensor networks[J]. Journal of Parallel and Distributed Computing, 2010, 70(6): 680 - 698.

[12] FLAJOLET P, MARTIN G. Probabilistic counting algorithms for data base applications[J]. Journal of Computer and System Sciences, 2001, 31(1): 134 - 143.