传感器网络中节点能量有效均衡的 Top-k 查询技术

宋保利[①] 郑吉平*[①②] 王海翔[①]
[①](南京航空航天大学计算机科学与技术学院 南京 210016)
[②](南京大学计算机软件新技术国家重点实验室 南京 210093)

摘 要：无线传感器网络中 top-k 查询处理的节点能量高效以及实现各节点的能量消耗均衡，可以有效延长网络的生命周期。该文提出一种基于采样技术和节点空间相关性，来实现节点的能量均衡和高效的查询处理算法，称为能量均衡采样(ε, δ)近似 top-k 算法 EBSTopk(ε, δ)。首先对传感器网络进行分区处理，利用区域内两两节点间的空间相关性对其建立线性回归预测模型和高斯预测模型；然后根据用户给定的相对误差界 ε 和置信水平 $1-\delta$ 建立节点高相关性预测准则；最后根据上述预测模型和准则，提出基于反复随机采样的能量均衡算法 EBSTopk(ε, δ)-LR 和 EBSTopk(ε, δ)-MG。实验表明，所提出的 EBSTopk(ε, δ)算法减少了无线传感器网络中的全局能量消耗，且在多次 top-k 查询后各节点的能量消耗达到均衡。

关键词：无线传感器网络；能量均衡；近似 top-k 查询；反复随机采样
中图分类号： TP393 **文献标识码**： A **文章编号**：1009-5896(2014)06-1478-07
DOI: 10.3724/SP.J.1146.2013.01163

Energy-efficient and Balanced Top-k Query Techniques in Sensor Networks

Song Bao-li[①] Zheng Ji-ping[①②] Wang Hai-xiang[①]
[①](*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016, *China*)
[②](*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093, *China*)

Abstract: Energy efficiency and balance of sensor nodes in processing top-k queries can prolong the lifetime of wireless sensor networks. In this paper, an Energy-efficient and Balanced query Sampling Top-k algorithm named EBSTopk(ε, δ) is proposed, which is based on the sampling techniques and the spatial correlations among sensor nodes. First, the sensor network is partitioned into several regions. Next, the linear regression prediction model and Gaussian prediction model are constructed based on the spatial correlations of pairwise sensor nodes. Then, the criteria of high spatial correlation is established due to the given relative error bound ε and the confidence level $1-\delta$. Finally, according to the predicting models and criteria above, two energy balanced algorithms named EBSTopk(ε, δ)-LR and EBSTopk(ε, δ)-MG are proposed, which are based on iterative random sampling technique. Experimental results show that, the proposed EBSTopk(ε, δ) algorithms not only reduce the global energy consumption in wireless sensor networks, but also achieve balanced energy consumption among all sensor nodes after continuous processing top-k queries.

Key words: Wireless sensor networks; Energy balance; Approximate top-k query; Iterative random sampling

1 引言

Top-k 查询[1,2]作为不确定数据和关系数据库中广泛使用的一种聚集查询技术，在无线传感器网络中有着广泛应用。传感器节点由电池供电，能量有限。因此，节点能量高效以及能量均衡在无线传感器网络 top-k 查询中具有重要的现实价值和研究意义。

对于集中式环境中的 top-k 查询，已经提出许多基于排序和修剪技术的方法减小计算复杂度以提高效率。然而，这些方法并不适应于分布式环境[3]。对于传感器网络环境下的 top-k 查询，人们已进行了一些研究工作。Madden 等人[4]提出聚集技术 TAG(Tiny AGregation)，但该方法存在大量冗余数据的传输。Wu 等人[5]提出基于滑动窗口的 top-k 查询算法 FILA(FILter-based monitoring Approach)，

2013-07-31 收到，2013-10-16 改回
国家 973 计划项目(2014CB744900)，教育部博士点基金(20103218110017)，航空科学基金(20115552030)，江苏高校优势学科建设工程，南京航空航天大学青年科技创新基金(NN2012102, NS2013089)和南京航空航天大学研究生开放基金(KFJJ120222)资助课题
*通信作者：郑吉平 zhjcs@nuaa.edu.cn

该算法基于过滤思想，为每一个节点设置一个过滤窗口，减少了冗余数据的传输，但会产生很大的窗口更新代价。Mai 等人[6]提出通过预测方法减小窗口更新代价的 DAFM(Distributed Adaptive Filter based Monitoring)算法，但节点读数在时间上的线性自回归预测法具有局限性。Chen 等人[7]提出分位数过滤算法 QF(Quantile Filter)，该算法可以避免大量冗余数据的上传，但每个节点与父节点都要进行通信，传感器网络的无线通信次数仍然很大。在一些实际应用中，近似的 top-k 查询结果往往可以满足用户的需求。对于近似 top-k 查询处理的研究，Ye 等人[8] 提出传感器网络环境下的概率 top-k 查询，很大程度上减少了数据的传输，节省了能量，但传感器网络中数据是分布式存储，计算代价较高，且中间节点能量消耗过快，不能实现传感器网络能量的均衡使用。Silberstein 等人[9]提出基于采样的 top-k查询优化算法 PROSPECTORLP+LF(用 LP+LF 表示)，该算法可以有效地减少网络中的全局能量消耗，但会导致频繁落在 top-k 结果集中的节点能量消耗过快而失效，不能实现节点能量的均衡使用。此外，上述两种近似 top-k 查询处理算法不能满足用户的精度要求。

本文提出一种基于采样技术和节点间空间相关性，实现节点能量高效且均衡的 top-k 查询处理技术。首先对传感器网络进行区域划分，利用区域内两两节点间的空间相关性对其建立预测模型。然后根据相对误差界 ε 和置信水平 $1-\delta$，制定节点高相关性预测准则。最后，根据上述预测准则提出基于反复随机采样的算法 EBSTopk(ε, δ) (Energy Balance Sampling Topk(ε, δ))。在 EBSTopk(ε, δ) 算法中每次只采样区域中的一个节点，然后利用节点高相关性预测准则，根据该节点感知的数据对其它节点的值进行估计，依此进行，直至整个区域内所有节点的值被直接采集或间接估计。EBSTopk(ε, δ) 算法很大程度上减少了传感器网络中节点的数据采集和无线通信。实验表明，在满足用户查询精度要求的前提下，本文提出的 EBSTopk(ε, δ) 算法不仅可以降低网络中的全局能量消耗，而且在多次近似 top-k 查询后各节点能量消耗相对均衡。

2　传感器网络中近似查询 Topk(ε, δ)

2.1 问题定义

定义 Topk(ε, δ) 查询 无线传感器网络中有 n 个节点，对所有节点标号，组成的集合记为 I, I=$\{1,2,\cdots,n\}$，标号为 i 的节点感知的数据记为 $X(i)$，经过近似查询得到节点 i 的值记为 $X'(i)$。传感器网络中的精确 top-k 结果集记为 $\{X(i_1), X(i_2), \cdots, X(i_k)\}$，其中 $X(i_1) \geq X(i_2) \geq \cdots \geq X(i_k)$，$i_1, i_2, \cdots, i_k \in I$；近似的 top-$k$ 结果集记为 $\{X'(j_1), X'(j_2), \cdots, X'(j_k)\}$，其中 $X'(j_1) \geq X'(j_2) \geq \cdots \geq X'(j_k)$, $j_1, j_2, \cdots, j_k \in I$。若 $\forall f, 1 \leq f \leq k$，满足

$$P\left(\frac{|X'(j_f) - X(i_f)|}{X(i_f)} > \varepsilon\right) < \delta \qquad (1)$$

则称该查询为 Topk(ε, δ) 查询。以上定义表示：最终得到的近似 top-k 查询结果集中每个值与精确 top-k 结果集中对应位置值的相对误差大于 ε 的概率小于 δ。

2.2 查询精度分析

为分析查询精度，引入定理 1 和定理 2。

定理 1　设集合 I=$\{i_1, i_2, \cdots, i_n\}$=$\{j_1, j_2, \cdots, j_n\}$, 集合 A=$\{a(i_1), a(i_2), \cdots, a(i_n)\}$，其中 $a(i_1) \geq a(i_2) \geq \cdots \geq a(i_n)$；集合 B=$\{a'(j_1), a'(j_2), \cdots, a'(j_n)\}$，其中 $a'(j_1) \geq a'(j_2) \geq \cdots \geq a'(j_n)$。若满足 $\forall h$, $1 \leq h \leq n$, $a'(j_h)(1-\varepsilon) \leq a(j_h) \leq a'(j_h)\cdot(1+\varepsilon)$，其中 $0<\varepsilon \ll 1$。则有：$a'(j_h)(1-\varepsilon) \leq a(i_h) \leq a'(j_h)(1+\varepsilon)$。

证明　假设 $a(i_h) > a'(j_h)(1+\varepsilon)$，则 $a(i_1) \geq a(i_2) \geq \cdots \geq a(i_h) > a'(j_h)(1+\varepsilon)$，集合 A 中至少有 h 个数大于 $a'(j_h)(1+\varepsilon)$。而 $\forall m, h \leq m \leq n$，有 $a(j_m) \leq a'(j_m)(1+\varepsilon) \leq a'(j_h)(1+\varepsilon)$，则集合 A 中至少有 $a(j_h), a(j_{h+1}), \cdots, a(j_n)$ 共 $n-h+1$ 个数小于等于 $a'(j_h)(1+\varepsilon)$，即集合 A 中至多有 h–1 个数大于 $a'(j_h)(1+\varepsilon)$。前后矛盾，所以假设不成立，则 $a(i_h) \leq a'(j_h)(1+\varepsilon)$。同理，可得 $a(i_h) \geq a'(j_h)(1+\varepsilon)$。

证毕

定理 2　已知若事件 A 发生，事件 B 一定发生，且 $P(A) > 0$, $P\{B\} > 0$，则 $P(B) \geq P(A)$。

证明　由条件得 $P(B \mid A) = 1$，又由 $P(B \mid A) = P(AB)/P(A)$，可得：$P(AB) = P(A)$。由 $P(A \mid B) \leq 1$ 且 $P(A \mid B) = P(AB)/P(B)$，得：$P(A)/P(B) \leq 1$，即 $P(B) \geq P(A)$。　　证毕

设某一时刻传感器网络中各节点真实的数据集为 $\{X(i_1), X(i_2), \cdots, X(i_n)\}$，其中 $X(i_1) \geq X(i_2) \geq \cdots \geq X(i_n)$。对于用户给定的参数 ε, δ，通过近似查询得到所有节点的值，组成的数据集为 $\{X'(j_1), X'(j_2), \cdots, X'(j_n)\}$，其中 $X'(j_1) \geq X'(j_2) \geq \cdots \geq X'(j_n)$，且对于 $\forall h, 1 \leq h \leq n$，满足式(2)：

$$P\left(X'(j_h)(1-\varepsilon) < X(j_h) < X'(j_h)(1+\varepsilon)\right) \geq 1-\delta \quad (2)$$

由定理 1 可知，当 $X'(j_h)(1-\varepsilon) \leq X(j_h) \leq X'(j_h)\cdot(1+\varepsilon)$ 时，必有 $X'(j_h)(1-\varepsilon) \leq X(i_h) \leq X'(j_h)(1+\varepsilon)$，

再由定理 2 得

$$P\{X'(j_h)(1-\varepsilon)<X(i_h)<X'(j_h)(1+\varepsilon)\}$$
$$\geq P\{X'(j_h)(1-\varepsilon)<X(j_h)<X'(j_h)(1+\varepsilon)\}$$
$$\geq 1-\delta \tag{3}$$

进一步得

$$P\left(-\frac{\varepsilon}{1+\varepsilon}\leq\frac{X'(j_h)-X(i_h)}{X(i_h)}\leq\frac{\varepsilon}{1-\varepsilon}\right)\geq 1-\delta \tag{4}$$

当 ε 足够小时，可以认为

$$P\left(-\varepsilon\leq\frac{X'(j_h)-X(i_h)}{X(i_h)}\leq\varepsilon\right)\geq 1-\delta \tag{5}$$

由此可得满足式(2)的近似查询所得到的 top-k 结果集即为 Topk(ε, δ) 查询所得到的结果集。

3 EBSTopk(ε, δ)算法

3.1 节点读数相关性建模

由于传感器节点能量受限且无法进行远距离的传输，因此，所部署的传感器网络中节点的密度较高。而空间位置相近的节点所感知的数据分布比较相似，它们之间存在着高度的空间相关性[10−12]。

假设存在着空间相关性的两个传感器节点分别记为 S_i, S_j，通过历史数据集训练得到它们之间的空间相关性模型。在基站保存了最近 s 个时刻所有节点感知的数据。假设节点 S_i 和 S_j 的历史数据组成数据集为 H_D, H_D 有如下形式：$HD=\{(x_i^{(l)}, x_j^{(l)}); l=1,2,\cdots,s\}$，其中 $x_i^{(l)}$ 和 $x_j^{(l)}$ 分别表示节点 S_i 和 S_j 在第 l 个时刻感知的数据，s 为训练集中样本的个数。通常，空间相关性模型可以定义为

$$x_j=h(x_i)+N(0,\sigma_{ij}^2) \tag{6}$$

节点 S_i 在 t 时刻新感知的数据为 $x_i(t)$时，S_j 感知数据的估计值 $x_j^*(t)=h(x_i(t))$，真实值 $x_j(t)$服从高斯分布：

$$x_j(t)\sim N(x_j^*(t),\sigma_{ij}^2) \tag{7}$$

图 1 中实心点表示真实的数据点$(x_i^{(l)}, x_j^{(l)})$，函数 $h(*)$表示通过历史数据学习得到的回归预测模型。图中高斯模型即为式(7)所示，即：真实值 $x_j(t)$ 服从均值为 $x_j^*(t)$，方差为 σ_{ij}^2 的高斯分布。

通常由于传感器网络的应用环境不同，传感器节点所感知的数据分布也会不同。可以根据不同的数据分布特点，发现其空间相关性，建立不同的预测模型。本文以线性回归预测模型和多元高斯预测模型为例。

3.1.1 线性回归预测模型 假设传感器节点感知的数据存在着很强的线性关系，可以对其建立线性回归预测模型。

$$x_j=\varphi_{ij}+\psi_{ij}x_i+N(0,\sigma_{ij}^2) \tag{8}$$

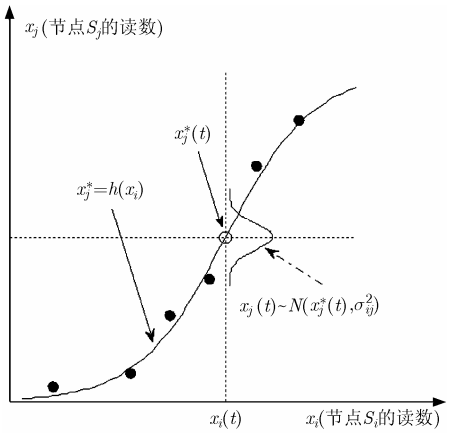


图 1 回归函数以及高斯分布

通常根据最小二乘法对参数 φ_{ij}, ψ_{ij} 的值进行估计。在已知 $x_i=x_i(t)$的条件下，由式(8)可得

$$x_j\mid x_i=x_i(t)\sim N(\varphi_{ij}+\psi_{ij}x_i(t),\sigma_{ij}^2) \tag{9}$$

当已知节点 S_i 感知的数据为 $x_i(t)$时，节点 S_j 感知数据的估计值 $x_j^*(t)=\varphi_{ij}+\psi_{ij}\ x_i(t)$。

3.1.2 多元高斯预测模型 在一段时间内，传感器节点感知的数据围绕一个数值上下波动，通常假定符合高斯分布[13]。位置靠近的两个传感器节点 S_i, S_j 所感知的数据服从高斯分布 $x_i\sim N(\mu_i,\sigma_i^2)$，$x_j\sim N(\mu_j,\sigma_j^2)$，且构成 2 维随机向量 $\boldsymbol{x}_{ij}$, $\boldsymbol{x}_{ij}\sim N(\boldsymbol{\mu}_{ij},\boldsymbol{\Sigma}_{ij})$。可以通过历史数据集 H_D 对其参数进行最大似然估计。在已知 $x_i=x_i(t)$的条件下，由式(8)可得

$$x_j|x_i=x_i(t)\sim N\left(\mu_j+\frac{r_{ij}\sigma_j}{\sigma_i}(x_i(t)-\mu_i),\sigma_j^2(1-r_{ij}^2)\right) \tag{10}$$

在实际传感器网络环境中，节点读数在不同时刻不会只服从单一高斯分布。以一天的温度值为例：早晨温度通常逐渐升高，而到了晚上温度逐渐下降。因此，可以将一天分成多个时间段分别建立高斯模型。

3.2 节点高相关性预测准则

当节点 S_i, S_j 读数具有很强的相关性时，方差 σ_{ij}^2 会比较小。当通过已知节点 S_i 感知的数据估计节点 S_j 的值时，就能以较高的置信水平估计出一个较小的置信区间。下面针对用户近似查询给定的相对误差界 ε 和置信水平 $1-\delta$ 制定节点高相关性预测准则，其中 $0\leq\varepsilon\ll 1, 0\leq\varepsilon\ll 1$。由式(7)可得

$$P\{x_j^*(t)-z_{\delta/2}\sigma_{ij}<x_j(t)<x_j^*(t)+z_{\delta/2}\sigma_{ij}\}=1-\delta \tag{11}$$

当置信度足够大时，可以认为节点的真实值 $x_j(t)$ 就落在这个置信区间里，即

$$x_j(t)\in(x_j^*(t)-z_{\delta/2}\sigma_{ij}, x_j^*(t)+z_{\delta/2}\sigma_{ij}) \tag{12}$$

要求真实值 $x_j(t)$与预测值 $x_j^*(t)$ 的相对误差小于 ε，即

$$\frac{\left|x_j(t)-x_j^*(t)\right|}{x_j(t)}<\varepsilon \tag{13}$$

由式(12)可得$|x_j(t)-x_j^*(t)|<z_{\delta/2}\sigma_{ij}$，且$x_j(t)>x_j^*(t)-z_{\delta/2}\sigma_{ij}$，为了使得式(13)始终成立，$|x_j(t)-x_j^*(t)|$取最大值，$x_j(t)$取最小值后仍满足它们的比值小于$\varepsilon$，即要求

$$\frac{z_{\delta/2}\sigma_{ij}}{x_j^*(t)-z_{\delta/2}\sigma_{ij}}<\varepsilon \tag{14}$$

进一步求得

$$x_j^*(t)>\frac{z_{\delta/2}\sigma_{ij}(1+\varepsilon)}{\varepsilon} \tag{15}$$

若节点预测值$x_j^*(t)$满足式(15)，则预测值与真实值的相对误差小于ε，认为这两个节点满足节点高相关性预测准则；若不满足，则认为这两个节点相关性不够高，节点S_j感知的数据需要通过直接采集或者利用其它与其相关性更高的采样节点的读数预测得到。由式(15)可得

$$x_j^*(t)\varepsilon>\frac{x_j^*(t)\varepsilon}{1+\varepsilon}>z_{\delta/2}\sigma \tag{16}$$

由式(11)和式(16)可得

$$P\left\{x_j^*(t)-x_j^*(t)\varepsilon<x_j(t)<x_j^*(t)+x_j^*(t)\varepsilon\right\}\geq 1-\delta \tag{17}$$

结合式(2)和式(5)可得，通过节点高相关性预测准则直接采样和间接估计获得传感器网络所有节点值的近似查询即为 Topk(ε, δ)查询。

3.3 能量均衡的采样算法 EBSTopk(ε, δ)

将传感器网络划分成多个区域，位置较近的节点划分到同一区域。本文对如何更合理地进行区域划分不做深入研究，仅根据节点位置信息以 m 均值算法对传感器网络进行区域划分。以伯克利大学英特尔实验室无线传感器网络[14]为例，利用 m 均值算法把其划分成10个区域。表1所示为区域划分结果，其中用方框框起来的节点表示在对传感器网络进行 m 均值划分时各区域的初始质心节点。

表 1 伯克利大学英特尔实验室无线传感器网络区域划分结果

区域	节点编号
R_1	48, 49, 50, [51], 52
R_2	7, [8], 9, 10, 11, 53, 54
R_3	12, 13, [14], 15, 16
R_4	17, 18, [19], 20, 21
R_5	2, 3, [4], 5, 6
R_6	44, [45], 46, 47
R_7	38, 39, [40], 41, 42, 43
R_8	1, 33, 34, [35], 36, 37
R_9	28, 29, [30], 31, 32
R_{10}	22, 23, 24, [25], 26

基站通过历史数据对区域内两两节点间的空间相关性进行建模，求得模型参数，以及误差标准差σ。基站向每一个节点发送此节点与其所在区域内其它节点之间的模型参数以及误差标准差σ。区域内每一节点轮流担任区域头节点。首先，随机选取一个节点作为区域头节点，然后根据节点高相关性预测准则，选出与区域头节点满足高相关性预测准则的节点，这些节点的值可以通过区域头节点的值估计得到。进而，在区域内剩余节点再随机采样一个节点，估计与此节点满足高相关性预测准则的节点的值。以此类推，直至获得区域内所有传感器节点的值，其中包括直接采集值和间接估计值。区域头节点选择 k 个值上传给基站，不足 k 个值，所有值一起上传。最后基站把接收来的所有值进行排序，选择前 k 个值作为 top-k 结果集。基于上述思想，本文提出能量均衡的采样算法 EBSTopk(ε, δ)，具体算法如表 2 所示。

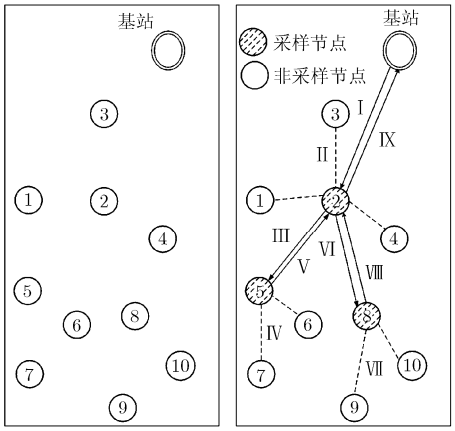
表 2 算法 EBSTopk(ε, δ) 伪代码

```
基站随机选择区域 R_i的头节点；
基站发送参数 ε , δ , k 至区域 R_i的头节点；
令 count(i)←1;
令 access(RegionHead(i))←1; /*标识已获得区域 R_i的头节点的
   读数*/
令 CurSNode←RegionHead(i); /*表示当前采样节点*/
while count(i)<n_i   /* n_i 表示区域 R_i的节点个数*/
   for NodeID ←1: n_i
      if access(NodeID)≠1
       if HighCorrelation(CurSNode, NodeID)=1 /*满足节点
       高相关性准则*/
            count(i)←count(i)+1;
            value(NodeID)←predict(NodeID);
            access(NodeID)←1;
       end if
      end if
      if count(i)<n_i
         CurSNode←sample(R_i); /*从剩余节点中选择一个采
         样节点*/

         count(i)←count(i)+1;
         access(CurSNode)←1;
      end if
   end for
  end while
区域头节点选择 k 个值上传给基站，不足 k 个值，所有值一起上
传
```

EBSTopk(ε, δ)算法中，函数 predict$(\bullet)$若采用的是线性回归预测模型，则称为 EBSTopk(ε, δ)-LR，若采用的是多元高斯预测模型，则称为 EBSTopk(ε, δ)-MG。针对 EBSTopk(ε, δ)算法，假

定某一区域中共有 10 个传感器节点，图 2(a)表示该区域在 EBSTopk(ε, δ)算法执行前的初始状态。算法执行时，基站首先随机选择其中一个节点作为区域头节点，假定随机选择的是 2 号节点，图 2(b)显示了 EBSTopk(ε, δ)算法执行过程。图中采样节点所能预测的节点用虚线与其连接。



(a)初始状态 (b)执行步骤

图 2 EBSTopk(ε, δ)算法执行过程

3.4 EBSTopk(ε, δ)算法能耗分析

节点发送 m 个字节的数据所消耗的总能量为 $\alpha_s+\beta_s m$，其中 α_s 表示每次建立通信连接发送方消耗的能量，β_s 表示发送每一字节的数据消耗的能量；接收 m 个字节的数据所消耗的总能量为 $\alpha_r+\beta_r m$，其中 α_r 表示每次建立通信连接接收方消耗的能量，β_r 表示接收每一字节的数据消耗的能量。对于 MICA2 型传感器节点 $\alpha_r \le 0.4\ \alpha_s$，$\beta_r \le 0.4\beta_s$ [15]，本文为了定量计算，取 α_r=$0.4\alpha_s$，β_r=$0.4\beta_s$。α_s，β_s，α_r 和 β_r 的取值如表 3 所示。

表 3 符号典型取值

符号	取值
α_s	0.645 mJ
β_s	0.0144 mJ/Byte
α_r	0.258 mJ
β_r	0.00576 mJ/Byte

每个区域的能量消耗分为两部分：采样节点(不包含区域头节点)以及区域头节点的能量消耗。区域 i 的总能量消耗记为 $E_t(i)$，采样节点的能量消耗记为 $E_s(i)$，区域头节点的能量消耗记为 $E_h(i)$，则

$$E_t(i) = E_s(i) + E_h(i) \tag{18}$$

$$E_s(i) = \alpha_s \times r_i + m \times (n_i - \eta_i - 1) \times \beta_s \tag{19}$$

$$E_h(i) = \left(\alpha_r \times r_i + m \times (n_i - \eta_i - 1) \times \beta_r\right) + \left(\alpha_s + m \times \mathrm{MIN}(k, n_i) \times \beta_s\right) \tag{20}$$

其中 n_i 表示区域 i 的节点个数，r_i 表示区域 i 的采样节点个数，η_i 表示区域 i 的头节点所能预测节点值的个数。由式(19)及式(20)可知，每一区域的能量消耗主要由采样次数决定。当采样次数较少时，区域能量消耗较小，而采样次数是由区域内节点间的相关性大小以及用户给定的查询精度要求决定的。用户对查询精度要求不是很苛刻时，采样次数较小，甚至每个区域只需要采样头节点读数即可；当用户对查询精度要求比较高时，采样次数较多，最极端的情况等效于 Naive$k$[5]方法。

4 实验测评

实验采用的数据集是 Intel Lab Data[14]，该数据集是由部署在伯克利大学英特尔实验室的 54 个传感器节点感知现实环境中的多个属性值得到。本文选择 2004 年 3 月 1 日的温度值作为实验数据，利用 Matlab 仿真实验来验证本文所提出算法的有效性。

图 3 为用户给定 ε，δ 值时，执行 EBSTopk(ε, δ)两种算法所需要的采样次数。采样次数和 k 值无关，在 ε，δ 的值都比较小的情况下，传感器网络的采样次数依然很小，甚至每个区域只需采样一次，并且 EBSTopk(ε, δ)-MG 采样次数少于 EBSTopk(ε, δ)-LR。可见，通过 EBSTopk(ε, δ)算法以很少的无线通信次数即可完成用户 top-k 查询，减少了传感器网络的能量消耗。

图 4 为在不同 k 值下 Naive$k$ 算法，Naive1 算法和本文中基于 EBSTopk(ε, δ)(其中 $\varepsilon = 0.02$, $\delta = 0.05$)的两种算法在 100 次 top-k 查询后所有节点所消耗总能量的对比图。从图 4 可以看到，本文提出的 EBSTopk(ε, δ)算法很大程度上减少了传感器网络的能量消耗，并且 EBSTopk(ε, δ)-MG 算法能量消耗要小于 EBSTopk(ε, δ)-LR 算法。

图 5 为 LP+LF 算法，Naive$k$ 算法以及本文中基于 EBSTopk(ε, δ)(其中 $\varepsilon = 0.02$, $\delta = 0.05$)的两种算法执行 100 次 top-15 查询后各节点的能量消耗对比图。先执行 100 次 EBSTopk (ε, δ)+LR 算法，然后求得平均每次 top-15 查询所消耗的能量作为执行 LP+LF 算法时的能量消耗约束的参数，即让执行 100 次这两种算法后使得它们总的能量消耗基本一致，以此来观察各节点的能量消耗情况。LP+LF 算法使得各节点能量消耗严重失衡。从图中可以看到，本文提出的两种 EBSTopk(ε, δ)算法各节点的能量消耗相对均衡。

图6为Naive$k$算法，LP+LF算法以及本文提出的EBSTopk(ε, δ)(其中 $\varepsilon = 0.02$, $\delta = 0.05$)两种算法在不同 k 值下执行100次top-15查询后能量均衡程度的比较。能量均衡度量用 M_{EBL}(Metric of Energy Balance Level)表示，其定义如式(21)所示。

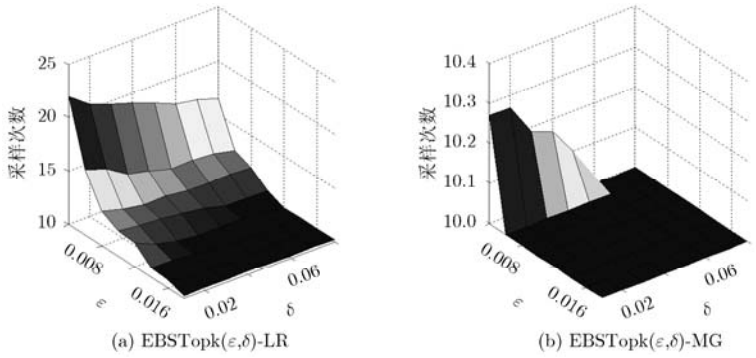


图3 不同精度要求下的采样次数

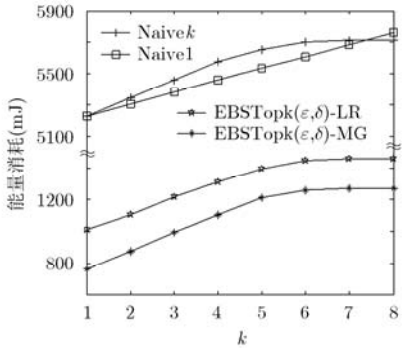


图4 总能量消耗对比

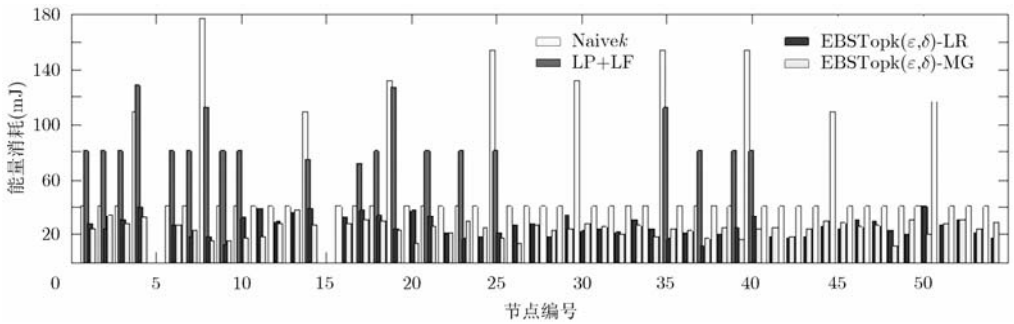


图5 各节点能量消耗对比

其中E(i)表示节点i消耗的能量。从图6可以看出，本文提出的两种EBSTopk (ε, δ) 算法的能量均衡水平明显优于LP+LF算法和Naive$k$算法。

$$M_{\mathrm{EBL}}=\sum_{i=1}^{n}\left(\mathrm{E}(i)-\sum_{i=1}^{n}\mathrm{E}(i)\Big/n\right)^2\Big/n \tag{21}$$

在用户给定ε，δ下，采用本文的 EBSTopk (ε, δ) 两种算法进行 100 次 top-15 查询，计算查询结果与精确值的平均相对误差。由图 7 可知，执行两种 EBSTopk (ε, δ) 算法，其平均相对误差远小于给定的相对误差界ε。

5　结束语

本文利用节点读数的空间相关性建立节点间的预测模型以及建立节点高相关性预测准则，提出实现节点能量有效均衡的近似 top-k 查询算法 EBSTopk (ε, δ)。理论和实验分析表明，本文提出的 EBSTopk (ε, δ) 算法在满足用户查询精度的前提下，不仅可以减少传感器网络中的无线通信，降低网络的全局能量消耗，而且在多次查询后，传感器网络各节点能量消耗相对均衡，从而延长了传感器网络的生命周期。

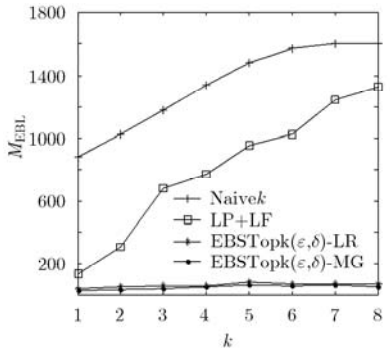


图 6 能量均衡水平对比

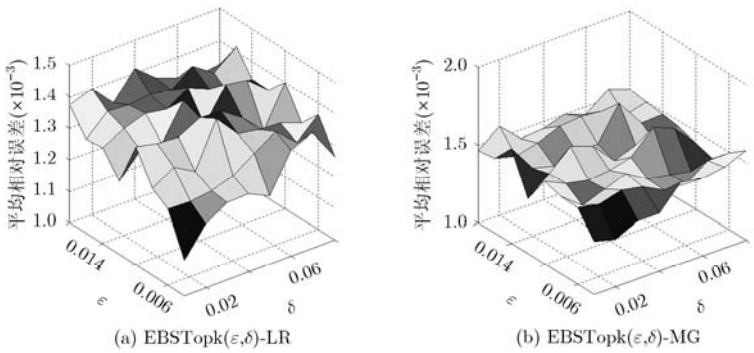


图 7 平均相对误差

参 考 文 献

[1]　李文风, 彭智勇, 李德毅. 不确定性top-K查询[J]. 软件学报, 2012, 23(6): 1542−1560.

Li Wen-feng, Peng Zhi-yong, and Li De-yi. Top-K query

processing techniques on uncertain data[J]. *Journal of Software*, 2012, 23(6): 1542–1560.

[2] Dylla M, Miliaraki I, and Theobald M. Top-k query processing in probabilistic databases with non-materialized views[C]. Proceedings of the 29th International Conference on Data Engineering, Brisbane, 2013: 122–133.

[3] Sun Yong-jiao, Yuan Ye, and Wang Guo-ren. Top-k query processing over uncertain data in distributed environments [J]. *World Wide Web*, 2012, 15(4): 429–446.

[4] Madden S, Franklin M, Hellerstein J, *et al*.. TAG: a Tiny AGgregation service for Ad-hoc sensor networks[J]. *ACM Special Interest Group on Operating Systems*, 2002, 35(SI): 131–146.

[5] Wu Min-ji, Xu Jian-liang, Tang Xue-yan, *et al*.. Top-k monitoring in wireless sensor networks[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(7): 962–976.

[6] Mai H, Lee Y, Lee K, *et al*.. Distributed adaptive top-k monitoring in wireless sensor networks[J]. *The Journal of Systems and Software*, 2011, 84(2): 314–327.

[7] Chen B, Liang W, Zhou R, *et al*.. Energy-efficient top-k query processing in wireless sensor networks[C]. Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, 2010: 329–338.

[8] Ye M, Lee W, Lee D, *et al*.. Distributed processing of probabilistic top-k queries in wireless sensor networks[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 25(1): 76–91.

[9] Silberstein A, Braynard R, Ellis C, *et al*.. A sampling-based approach to optimizing top-k queries in sensor networks[C]. Proceedings of the 22nd International Conference on Data Engineering, Atlanta, 2006: 68–78.

[10] Jindal A and Psounis K. Modeling spatially correlated data in sensor networks[J]. *ACM Transactions on Sensor Networks*, 2006, 2(4): 466–499.

[11] Wang C, Ma H, He Y, *et al*.. Adaptive approximate data collection for wireless sensor networks[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2012, 23(6): 1004–1016.

[12] Fateh B and Govindarasu M. Energy minimization by exploiting data redundancy in real-time wireless sensor networks[J]. *Ad Hoc Networks*, 2013, 11(6): 1715–1731.

[13] Deshpande A, Guestin C, and Madden S. Model-driven data acquisition in sensor networks[C]. Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, 2004: 588–599.

[14] Intel Berkeley Research. Laboratory. http://db. lcs.mit.edu /labdata/labdata.html. 2013. 2.

[15] Silberstein A, Braynard R, and Yang J. Constraint chaining: on energy-efficient continuous monitoring in sensor networks [C]. Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, Chicago, 2006: 157–168.

宋保利： 男，1987 年生，硕士生，研究方向为传感器数据管理.

郑吉平： 男，1979 年生，副教授，CCF 高级会员，研究方向为感知数据管理、粒子滤波、蒙特卡罗方法等.

王海翔： 女，1989 年生，硕士生，研究方向为信息物理融合、Skyline 计算.