无线传感器网络(ε, δ)−近似 Top-k 查询处理算法

毕冉，李建中，程思瑶
(哈尔滨工业大学 计算机科学与技术学院，黑龙江 哈尔滨 150001)

摘 要：提出一种适合于任意数据分布的基于抽样的(ε, δ)–近似 Top-k 查询处理算法。其中，$\varepsilon \geqslant 0$ 和 $0 \leqslant \delta < 1$ 分别是相对误差界和失败概率界。理论分析表明，对于任意$\varepsilon \geqslant 0$ 和 $0 \leqslant \delta < 1$，该算法返回的查询结果的相对误差界大于 $\varepsilon/(1+\varepsilon)$的概率小于$\delta$。于是，该算法可以达到任意精度。同时，还给出了支持近似 Top-k 查询的优化的抽样算法，并通过节点上的数据过滤技术来减少通信能量的消耗。理论分析和仿真结果表明，提出的算法能量消耗低并且计算复杂度低。

关键词：近似 Top-k 查询；抽样算法；无线传感器网络
中图分类号：TP393.01 **文献标识码**：A **文章编号**：1000-436X(2011)08-0045-10

(ε, δ)-approximate Top-k query processing algorithm in wireless sensor networks

BI Ran, LI Jian-zhong, CHENG Si-yao
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: A sampling based approximate Top-k algorithm was proposed that is adaptive for any data distribution. $\varepsilon \geqslant 0$ and $0 \leqslant \delta < 1$ are respectively relative error bound and failure probability bound. The theoretical analysis demonstrates that for any $\varepsilon \geqslant 0$ and $0 \leqslant \delta < 1$ the probability that the relative error bound of the results returned by this algorithm is larger than $\varepsilon/(1+\varepsilon)$ is less than δ. So the proposed algorithm can reach arbitrary precision. Furthermore, an optimal sampling algorithm was proposed that supported the approximate Top-k query, and through the technique of data filtering the energy consumption of communication was reduced. Theoretical analysis and simulation show that the proposed algorithm is efficient and consumes little energy.

Key words: approximate Top-k query; sampling algorithm; wireless sensor networks

1 引言

无线传感器网络的主要功能是收集并返回传感器节点所监测区域的信息。用户通过向无线传感器网络提出各种感知数据查询，实现对无线传感器网络所监测环境的监控。

为了有效地监控环境、生态变化等，感知数据的 Top-k 查询是一种在传感器网络应用中经常使用的查询。Top-k 查询返回某一时间段内网络中 k 个最大(或最小)的感知值及相应的位置信息，使得用户能够获得其最感兴趣的数据，有助于用户获得监测区域内的异常信息并做出相应的分析决策。例如，在基于无线传感器网络的环境监测应用中，Top-k 查询结果可以使得用户了解哪些区域污染最

收稿日期：2010-06-17；**修回日期**：2010-11-12
基金项目：国家自然科学基金重点资助项目（61033015）；国家自然科学基金资助项目（60933001, 60831160525, 60703012）
Foundation Items: The Key Program of the National Natural Science Foundation of China (61033015); The National Natural Science Foundation of China (60933001, 60831160525, 60703012)

为严重，并做出相应的决策。无线传感器网络的Top-k 查询处理问题已经引起了人们广泛的关注，提出了一些Top-k 查询处理算法[1~4]。然而，这些算法主要集中在如何给出精确的Top-k 查询结果，能量消耗很大，降低了无线传感器网络的寿命。

在很多实际应用中，人们并不需要精确的Top-k 查询结果，仅需要满足一定精度要求的近似查询结果[2,5,6]。因此，为了减少通信能量消耗，需要研究处理Top-k 查询的近似算法。针对无线传感器网络的近似Top-k 查询处理，文献[5]提出的算法仅考虑了如何最小化丢失的查询结果的数量，没有考虑用户的精度要求，不能给出任意精度的查询结果。当改变误差界时，文献[2]提出的算法需要重新设置阈值或松弛条件，计算代价较高；不能实时地为用户返回近似结果，不能满足用户的任意精度需求。

在实际应用中，一个无线传感器网络往往具有多个用户，查询结果的精度要求各不相同。同一个用户也经常要求具有不同精度的查询结果。下面给出2个实例。

例 1 在水质监测应用中，用户A想知道PH值最高的k 个监测位置；用户B想知道水质级别最高的k 个监测位置。根据中国地表水水质监测标准，相同的水质级别可能具有不同的PH值，因而A的精度标准高于B。

例 2 在噪声监测应用中，用户A想知道分贝值最高的k 个监测位置，用户B想知道噪声等级最高的k 个位置。根据国家环境噪声标准，相同等级噪声的分贝值可能是不同的，因而A的精度标准高于B。

上述2个实例说明，不具有任意精度或至少多种精度的近似Top-k 查询处理算法是不能满足实际应用要求的。文献[5]提出的近似算法显然不能满足用户的任意或多精度要求。研究具有多精度或任意精度的Top-k 查询处理算法是十分必要的。为此，本文研究具有任意精度的Top-k 查询处理算法。

本文首先提出了一种(ε, δ)-近似Top-k 查询，其语义为对于用户给定的任意$\varepsilon \geqslant 0$ 和 $0 \leqslant \delta < 1$，查询返回结果的误差大于 $\varepsilon/(1+\varepsilon)$的概率小于$\delta$，其中，$\varepsilon \geqslant 0$ 和 $0 \leqslant \delta < 1$ 分别为相对误差界和失败概率界。如果 $\varepsilon = 0$ 且 $\delta = 0$，则查询返回精确的 Top-k 结果。

然后，本文提出了一种基于抽样的处理(ε, δ)-近似Top-k 查询的近似算法。算法的主要思想如下：首先，根据用户指定的ε和δ确定样本S的大小；其次，根据样本大小对传感器网络进行优化均匀地随机抽样；最后，利用样本数据计算近似Top-k 结果。本文的贡献如下。

1) 提出了(ε, δ)-近似 Top-k 查询，并给出了其严格的语义定义；

2) 提出了基于抽样的(ε, δ)-近似 Top-k 查询处理的近似算法，证明了其结果可以任意精确；

3) 以最小化样本集合大小为目标，提出了有效支持(ε, δ)-近似Top-k 查询处理算法的优化抽样方法；

4) 提出了通过节点上的数据过滤技术来减少通信能量消耗的方法；

5) 通过理论分析和实验，验证了(ε, δ)-近似Top-k 查询处理算法通信能量消耗低和算法的高效性。

本文之后的章节安排如下：第2节介绍相关工作；第3节给出(ε, δ)-Top-k 查询处理问题的形式化定义；第4节介绍算法的数学基础，证明了算法中需要的数学结果；第5节介绍优化的抽样算法，给出算法的理论分析和性能分析；第6节通过实验验证了(ε, δ)-Top-k 查询处理算法的有效性；第7节是结束语。

2 相关工作

针对数据库的近似Top-k 查询处理，文献[6]基于TA算法的思想，提出 θ-近似 Top-k 查询处理问题。算法返回满足用户精度要求的近似结果，其中θ 为相对误差界。文献[7]和文献[8]返回带有概率保证的近似 Top-k 结果。由于传感器网络中的数据是分布式存储的，读取代价较高，因此算法不适合应用于传感器网络。

一些工作致力于分布式近似 Top-k 查询算法的研究[2,3,9,10]。在这些算法中，一个对象的打分值是局部打分值的加权和。其中，每个数据源提供若干对象在某个属性上的排序。对象的局部打分值分别来自不同节点。在传感器网络中，一个传感器节点为监测对象提供完整的观测数据，故上述算法的应用环境与本文讨论的传感器网络有很大的不同，该算法亦不适合应用于传感器网络中。在P2P环境下，文献[4]返回带有固定概率保证的近似Top-k 查询结果；然而当k 值调整时这种方法不够灵活，尤其当k 值较大时往往需要重新设置阈值。

由于传感器感知物理环境固有的物理误差或者随机误差，使得很多研究工作致力于传感器网络

的近似数据管理技术[11~13]。文献[11]提出的算法可以为用户递增地精炼事前收集的近似数据，使得误差任意小。文献[14]提出带有误差界保证的近似聚集算法，更进一步地，笔者在文献[12]和文献[13]中提出(ε, δ)-近似聚集算法。对用户给定的任意(ε, δ)算法返回近似聚集结果的相对误差界大于 ε 的概率小于 δ。

3　问题定义

令 n 为传感器网络的节点数量。传感器网络分为 L 个不相交的簇，记为 $C_1, C_2,\cdots,C_L$。簇内节点与簇头节点的平均通信跳数为 h，簇头节点与 sink 节点的平均通信跳数为 h'。令 I 为传感器节点标号集合，即 $I=\{1, 2,\cdots, n\}$。t 时刻各传感器节点的感知数据集合记为 X_t, $X_t=\{d(1), d(2),\cdots,d(n)\}$, $d(i)$表示标号为 i 的节点感知的数据，简记为 d_i。为了叙述算法简练，采用上述网络拓扑结构，但是笔者提出的理论保证及算法适用于一般的信道模型和网络协议下的传感器网络。

定义 1　X_t的 Top-k 集合记为 $Top(k, X_t)$，为 I 的子集，$Top(k, X_t)$满足以下条件：

1) $Top(k, X_t)\subset I$；

2) $\left|Top(k, X_t)\right|=k$；

3) $\forall i\in Top(k, X_t), \forall j\notin Top(k, X_t)$ 满足 $d_i\geqslant d_j$。

定义 2　对 $\forall\varepsilon>0$，X_t的 ε-近似$Top(k, X_t)$简记为 ε-$Top(k)$，其中，ε-$Top(k)$ 满足以下条件：

1) ε-$Top(k)\subset I$；

2) $|\varepsilon$-$Top(k)|=k$；

3) 不妨设$Top(k, X_t)=\{i_1, i_2,\cdots, i_k\}$，其中，$d(i_1)\geqslant d(i_2)\geqslant\cdots\geqslant d(i_k)$，$\varepsilon$-$Top(k)=\{j_1, j_2,\cdots, j_k\}$且 $d(j_1)\geqslant d(j_2)\geqslant\cdots\geqslant d(j_k)$。那么 $\max\left\{\dfrac{d(i_f)}{d(j_f)}\middle|1\leqslant f\leqslant k\right\}\leqslant 1+\varepsilon$。

若不存在 f 使得 $\dfrac{d(i_f)}{d(j_f)}\geqslant 1+\varepsilon$ 则 $d(j_f)\geqslant\dfrac{d(i_f)}{1+\varepsilon}$，故 $d(i_f)-d(j_f)\leqslant\dfrac{\varepsilon d(i_f)}{1+\varepsilon}$，$\dfrac{d(i_f)-d(j_f)}{d(i_f)}\leqslant\dfrac{\varepsilon}{1+\varepsilon}$。由 ε-$Top(k)$的定义知，近似结果的相对误差小于等于 $\dfrac{\varepsilon}{1+\varepsilon}$。

定义 3　对于感知数据集合 X_t，不妨设 $d(i_1)\geqslant d(i_2)\geqslant\cdots\geqslant d(i_n)$，定义 $d(i_1)$的 ε-近似节点 ID 集合。$ID(i_1,\varepsilon)=\{j\,|\,(1+\varepsilon)\times d(j)\geqslant d(i_1)\}$，类似地，$ID(i_2,\varepsilon)=\{j\,|\,(1+\varepsilon)\times d(j)\geqslant d(i_2)\}$，$\cdots$，$ID(i_n,\varepsilon)=\mathrm{I}$。而且易见 $ID(i_1,\varepsilon)\subset ID(i_2,\varepsilon)\subset\cdots\subset ID(i_n,\varepsilon)$，$1\leqslant|ID(i_1,\varepsilon)|\leqslant|ID(i_2,\varepsilon)|\leqslant\cdots\leqslant|ID(i_n,\varepsilon)|=n$。

定义 4　对于感知数据集合 X_t，不妨设 $d(i_1)\geqslant d(i_2)\geqslant\cdots\geqslant d(i_n)$ 定义 $d(i_1)$ 的 ε - 近似频率 $P_{X_t}^{\varepsilon}(i_1)=\dfrac{|ID(i_1,\varepsilon)|}{n}$，简记为 $P_X^{\varepsilon}(i_1)=\dfrac{|ID(i_1,\varepsilon)|}{n}$, 同理有 $P_X^{\varepsilon}(i_2)=\dfrac{|ID(i_2,\varepsilon)|}{n}$ 等。给定 P^* 的值，满足 $P_X^{\varepsilon}(i_1)=P^*$ 的感知数据集合统一记为 $X(P^*,n)$。

定义 5　针对样本 S，其中样本 S 是节点标号集合 I 的子集，定义 $P_S^{\varepsilon}=\dfrac{|\{j\,|((1+\varepsilon)\times d_j\geqslant d_{i_h})\wedge(j\in S)\}|}{|S|}$。其中，感知数据集合 X_t 满足 $d(i_1)\geqslant d(i_2)\geqslant\cdots\geqslant d(i_h)\geqslant\cdots\geqslant d(i_n)$。

定义 6　对于用户给定的 $\forall\varepsilon>0, 0<\delta<1$, X_t 的 (ε,δ) -近似 $Top(k, X_t)$，通常简记为 (ε,δ) -$Top(k)$，其中，(ε, δ)满足以下条件：

1) (ε, δ)-Top-k $\subset I$;

2) |(ε, δ)-Top-k| = k。

不妨设 $Top(k, X_t)=\{i_1,i_2,\cdots,i_k\}$，其中，$d(i_1)\geqslant d(i_2)\geqslant\cdots\geqslant d(i_k)$。对于给定的 $\varepsilon'(0<\varepsilon'<1)$, 若随机样本满足 $1\leqslant h\leqslant k$ 时，$P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')P_X^{\varepsilon}(i_h)$ 的概率小于 δ，即 $\min\left(\Pr\left\{P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')P_X^{\varepsilon}(i_h)\right\}\right)\leqslant\delta$，则称该随机样本 S 输出的前 k 个最大(小)值为(ε, δ)-Top-k。

不妨设用户提交了一个(0.2,0.15)-Top(3)查询，样本S输出的前3个最大值为(0.2,0.15)-Top(3)结果，则 $Pr\left\{P_S^{0.2}(i_2)\geqslant(1-\varepsilon')P_X^{0.2}(i_2)\right\}\geqslant 0.85$。若 $\varepsilon'=0.1$, $P_X^{0.2}(i_2)=0.1$, $|S|=20$，则 $|S|\times P_S^{0.2}(i_2)\geqslant|S|\times(1-\varepsilon')P_X^{0.2}(i_2)=0.9\times 0.1\times 20=1.8$ 的概率为 0.85，又因 $|S|\times P_S^{0.2}(i_2)$ 为整数，所以 $|S|\times P_S^{0.2}(i_2)\geqslant 2$ 的概率为 0.85。于是样本输出的近似结果中至少存在 h_1 和 h_2，使得这 2 个节点的观测值 $d(h_1)$、$d(h_2)$ 满足 $(1+\varepsilon)d(h_1)\geqslant d_{i_2}$，$(1+\varepsilon)d(h_2)\geqslant d_{i_2}$ 的概率至少为 0.85。故样本输出的结果非常接近 ε-Top(3)的结果。

若有 $P_S^{\varepsilon}(i_h)>0$，则知样本 S 中必然存在 j 使得 $(1+\varepsilon)\times d(j)\geqslant d(i_h)$。若不然，则不存在这样的 j 使得 $(1+\varepsilon)\times d(j)\geqslant d(i_h)$，那么 $P_S(i_h)=0$。这与前提条件矛盾。于是 $P_S^{\varepsilon}(i_h)>0, h=1,2,\cdots,k$，是随机样本能够返回 ε-$Top(k)$的必要条件。

不妨设 $Top(k, X_t)=\{i_1,i_2,\cdots,i_k\}$，其中，$d(i_1)\geqslant d(i_2)\geqslant\cdots\geqslant d(i_k)$，随机样本 S 输出的前 k 个最大值 $\{d(j_1),\cdots,d(j_k)\}$ 为 (ε,δ)-$Top(k)$。若 $P_S^{\varepsilon}(i_h)>$

$(1-\varepsilon')P_X^{\varepsilon}(i_h)$，其中，$h=1,2,\cdots,k$，则必然有 $P_S^{\varepsilon}(i_h)>0,\ h=1,2,\cdots,k$。于是$(\varepsilon,\delta)$-*Top-k* 为带有概率保证的 ε-*Top-k* 近似 *Top*(*k*) 结果。即对 $\forall l,1\leqslant l\leqslant k$，令 $h=\min\{\ (1-\varepsilon')P_X^{\varepsilon}(i_l)\,|S|,l\ \}$ 则 $(1+\varepsilon)\times d_{j_h}\geqslant d_{i_l}$ 成立的概率为 1–δ，即误差界大于 $\varepsilon/(1+\varepsilon)$的概率小于$\delta$。

4 数学基础

定理 1 为了返回ε-近似 Top-*k* 结果，所需的样本容量的数学期望为 $\frac{1}{P_X^{\varepsilon}(i_1)}+\frac{1}{P_X^{\varepsilon}(i_2)}+\cdots+\frac{1}{P_X^{\varepsilon}(i_k)}$

证明 不妨设 $Top(k,X_t)=\{i_1,i_2,\cdots,i_k\}$ 且 $d(i_1)\geqslant d(i_2)\geqslant\cdots\geqslant d(i_k)$。对 $\forall j\in I$，若 $(1+\varepsilon)\times d(j)\geqslant d(i_h)$，在该证明中将 $d(j)$ 统一记为 $d_{\varepsilon-Top\_h}$。记 $Y_{h_{j-1}}$ 为从抽到 $d_{\varepsilon-Top\_h_{j-1}}$ 之后到抽到 $d_{\varepsilon-Top\_h_j}$ 所需的样本容量，则 $X=\sum Y_{h_1}+Y_{h_2}+\cdots+Y_{h_k}$，$h_1,h_2,\cdots,h_k$ 为 $1,2,\cdots,k$ 的一个排列。显然对一切 $h_j,j=1,2,\cdots,k$，Y_{h_j} 服从 $Pr(h_j)=P_X^{\varepsilon}(h_j)$ 的几何分布，即 $j=1,2,\cdots,k$ 时 $Pr\{Y_{h_j}=m\}=(1-P_X^{\varepsilon}(h_j))^{m-1}P_X^{\varepsilon}(h_j)$。从而 $\mathrm{E}(Y_{h_j})=\frac{1}{P_X^{\varepsilon}(h_j)}$。1~*k* 的一个排列为 $h_1,h_2,\cdots,h_k$ 的概率为 $(k!)^{-1}$。于是 $\mathrm{E}(X)=\sum_{1,2,\cdots,k\text{全排列}}\frac{1}{k!}\mathrm{E}(Y_{h_1}+Y_{h_2}+\cdots+Y_{h_k})=\frac{1}{k!}\sum\mathrm{E}\left(Y_{h_1}+Y_{h_2}+\cdots+Y_{h_k}\right)$，$1,2,\cdots,k$ 全排列。所以

$$\begin{aligned}\mathrm{E}(X)&=\sum\mathrm{E}(Y_{h_1})+\mathrm{E}(Y_{h_2})+\cdots+\mathrm{E}(Y_{h_k})\\&=\frac{1}{P_X^{\varepsilon}(i_1)}+\frac{1}{P_X^{\varepsilon}(i_2)}+\cdots+\frac{1}{P_X^{\varepsilon}(i_k)}\end{aligned}$$

由于很难获得 $P_X^{\varepsilon}(i_1),P_X^{\varepsilon}(i_2),\cdots,P_X^{\varepsilon}(i_k)$ 的精确值，甚至是近似值，尤其是当 ε、k 变化时，很难估计样本容量的期望。因此本文主要研究$(\varepsilon,\ \delta)$-Top-*k* 查询处理算法。

定理 2 (切尔诺夫界) 设 $X_1,\cdots,X_n$ 是独立的泊松实验满足 $Pr(X_i)=p_i$，设 $X=\sum_{i=1}^{n}X_i,\mu=\mathrm{E}[X]$。那么下面的切尔诺夫界成立：

1) 对任意$\varepsilon>0$，

$$\Pr(X\geqslant(1+\varepsilon)\mu)\leqslant\left(\frac{\mathrm{e}^{\varepsilon}}{(1+\varepsilon)^{(1+\varepsilon)}}\right)^{\mu}$$

2) 对任意$0<\varepsilon<1$，

$$\Pr(X\leqslant(1-\varepsilon)\mu)\leqslant\left(\frac{\mathrm{e}^{-\varepsilon}}{(1-\varepsilon)^{(1-\varepsilon)}}\right)^{\mu}$$

定理证明见文献[15]。

定理 3 令 *X* 为一个任意数据分布的感知数据集合，不妨设 $Top(k,X)=\{i_1,\cdots,i_k\}$ 则 d_{i_h} 的 ε 近似频率 $P_X^{\varepsilon}(i_h)=\frac{|ID(i_h,\varepsilon)|}{n}$，$P_S^{\varepsilon}(i_h)=\frac{|\{j|\left((1+\varepsilon)d_j\geqslant d_{i_h}\right)\wedge\left(j\in S\right)\}|}{|S|}$ 那么对任意一个随机样本 *S* 有下式成立：

对任意$\varepsilon'>0$，有

$$\Pr\{P_S^{\varepsilon}(i_h)\geqslant(1+\varepsilon')P_X^{\varepsilon}(i_h)\}\leqslant\left(\frac{\mathrm{e}^{\varepsilon'}}{(1+\varepsilon')^{(1+\varepsilon')}}\right)^{|S|P_X^{\varepsilon}(i_h)}\tag{1}$$

对任意 $0<\varepsilon'<1$，有

$$\Pr\{P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')P_X^{\varepsilon}(i_h)\}\leqslant\left(\frac{\mathrm{e}^{-\varepsilon'}}{(1-\varepsilon')^{(1-\varepsilon')}}\right)^{|S|P_X^{\varepsilon}(i_h)}\tag{2}$$

证明 利用切尔诺夫界证明这个定理。考虑节点标号集合$\{v_1,\cdots,v_{m(i_h)}\}$使得$(1+\varepsilon)\times d(v_j)\geqslant d(i_h)$，$1\leqslant j\leqslant m(i_h)$。令 X_j 为一个指示变量，如果节点 v_j 的感知数据被抽样算法抽到则 X_j=1。对于均匀抽样算法，X_j 是独立同分布的随机变量 $\Pr\{X_j=1\}=\frac{m(i_h)}{n}$。即对任意数据分布的感知数据，对感知数据的抽样满足独立的泊松实验要求。令 $X^S(i_h)=\sum_{j=1}^{m(i_h)}X_j$，因随机变量 $X^S(i_h)$ 的期望为 $|S|\times P_X(i_h)$，由定理 2 切尔诺夫界可以推出 $\Pr\{X^S(i_h)\geqslant(1+\varepsilon')|S|P_X(i_h)\}\leqslant\left(\frac{\mathrm{e}^{\varepsilon'}}{(1+\varepsilon')^{(1+\varepsilon')}}\right)^{|S|P_X^{\varepsilon}(i_h)}$，$\Pr\left\{\frac{X^S(i_h)}{|S|}\geqslant\frac{(1+\varepsilon')|S|P_X(i_h)}{|S|}\right\}\leqslant\left(\frac{\mathrm{e}^{\varepsilon'}}{(1+\varepsilon')^{(1+\varepsilon')}}\right)^{|S|P_X^{\varepsilon}(i_h)}$ 则知式(1)成立，同理可知式(2)也成立。

由定理 3 可知随着样本数的增加，$P_S^{\varepsilon}(i_h)$ 发生偏离其期望 ε' 倍的概率越小；当样本数较小时 $P_S^{\varepsilon}(i_h)$ 发生偏离其期望 ε' 倍的概率较大。事实上，均匀随机地抽样可以获得全局数据的部分信息。若 $P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')P_X^{\varepsilon}(i_h)$ 且样本数 $|S|$ 不够充分大，使

得$|S|\times P_S^{\varepsilon}(i_h)<1$成立，当$1\leqslant h\leqslant k$时样本 S 输出的近似 Top-k 必然不满足 $\max\{\frac{d_{i_f}}{d_{j_f}}|1\leqslant f\leqslant k\}\leqslant 1+\varepsilon$，故相对误差大于$\frac{\varepsilon}{1+\varepsilon}$。

定理 4　令$f(\varepsilon')=\frac{\mathrm{e}^{-\varepsilon'}}{(1-\varepsilon')^{(1-\varepsilon')}}$，当$|S|\geqslant\frac{\log_{f(\varepsilon')}\delta}{P_X^{\varepsilon}(i_h)}$时，$\Pr\{P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')P_X^{\varepsilon}(i_h)\}\leqslant\delta,0<\varepsilon'<1$

证明　由定理 4 的式(2)可知$\Pr\{P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')\cdot P_X^{\varepsilon}(i_h)\}\leqslant\left(\frac{\mathrm{e}^{-\varepsilon'}}{(1-\varepsilon')^{(1-\varepsilon')}}\right)^{|S|P_X^{\varepsilon}(i_h)}$。已知$|S|\geqslant\frac{\log_{f(\varepsilon')}\delta}{P_X^{\varepsilon}(i_h)}$且$\left(\frac{\mathrm{e}^{-\varepsilon'}}{(1-\varepsilon')^{(1-\varepsilon')}}\right)<1$，则下式成立，$\Pr\{P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')\cdot P_X^{\varepsilon}(i_h)\}\leqslant\left(\frac{\mathrm{e}^{-\varepsilon'}}{(1-\varepsilon')^{(1-\varepsilon')}}\right)^{\log_{f(\varepsilon')}\delta}$，即$\Pr\{P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')\cdot P_X^{\varepsilon}(i_h)\}\leqslant\delta$。

推论　当随机样本满足$|S|\geqslant\frac{\log_{f(\varepsilon')}\delta}{P_X^{\varepsilon}(i_1)}$时，可知下式成立$\Pr\{P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')P_X^{\varepsilon}(i_h)\}\leqslant\delta$，$1\leqslant h\leqslant k$。

证明　因$\frac{\log_{f(\varepsilon')}\delta}{P_X^{\varepsilon}(i_h)}$是$P_X^{\varepsilon}(i_h)$的递减函数，当$|S|\geqslant\frac{\log_{f(\varepsilon')}\delta}{P_X^{\varepsilon}(i_1)}$，$1\leqslant h\leqslant k$时，$\Pr\{P_S^{\varepsilon}(i_h)\leqslant(1-\varepsilon')\cdot P_X^{\varepsilon}(i_h)\}\leqslant\delta$是显然的。

若有$P_S^{\varepsilon}(i_h)\geqslant(1-\varepsilon')P_X^{\varepsilon}(i_h)$，则知样本 S 中必然存在 j 使得$(1+\varepsilon)\times d_j\geqslant d_{i_h}$。若不然，则不存在这样的 j 使得$(1+\varepsilon)\times d_j\geqslant d_{i_h}$，那么$P_S(i_h)=0$。这与前提条件矛盾。

综上，当$|S|\geqslant\frac{\log_{f(\varepsilon')}\delta}{P_X^{\varepsilon}(i_1)}$时，随机样本输出的近似 Top-$k$ 结果满足$P_S(i_h)\geqslant(1-\varepsilon')P_X(i_h)$的概率大于等于$1-\delta$，其中，$1\leqslant h\leqslant k$，$0<\varepsilon'<1$。

可以优化地选取ε'，在$|S|\times(1-\varepsilon')P_X(i_1)\geqslant1$的条件下，使得$\frac{\log_{f(\varepsilon')}\delta}{P_X(i_1)}$取到最小值。

5　抽样算法

5.1　简单抽样算法

首先根据用户给定的(ε, δ)和历史信息估计的$\hat{P}_{X_{t-1}}^{\varepsilon}(i_1)$确定样本的大小。不妨设为|S|，根据|S|和传感器节点标号集合 $I=\{1,2,\cdots,n\}$，t 时刻的简单抽样算法步骤如下：

1) sink 节点随机独立地产生|S|个 1~n 之间的自然数，$S=\{s_1,s_2,\cdots,s_{|S|}\}$。

2) sink 节点将这|S|个 ID 值广播至传感器网络。

3) 若传感器节点的 ID 值属于 S，则该节点将感知数据传送到 sink 节点。

但是广播|S|个随机数并将感知数据传送到 sink 节点的通信代价较大，因此笔者提出基于历史信息和数据过滤技术的优化算法。优化的主要思想是：1) 考虑实际的应用背景和感知数据的相关性。抽样算法返回的 t–1 时刻(ε, δ)-*Top*(*k*)的感知位置在 t 时刻感知的数据成为(ε, δ)-*Top*(*k*)结果的可能性大一些，所以希望得到这些节点的观测值；2) 通过簇头节点和簇内节点的数据过滤技术来减少通信代价。

5.2　优化的均匀抽样算法

输入：

1) 具有 n 个节点并已经分为 L 个簇的传感器网络；

2) t 时刻传感器网络的感知数据集合 $X_t=\{d_1,d_2,\cdots,d_n\}$；

3) 样本容量|S|。

输出：

(ε, δ)-*Top*(*k*)候选集和 c_j，且$1\leqslant j\leqslant L$。

5.3　Sink 节点的优化抽样算法步骤

1) 首先 sink 节点根据用户给定的(ε, δ)、算法估计的$\hat{P}_{X_{t-1}}^{\varepsilon}(i_1)$确定样本的大小，不妨设为|S|。

2) 将 t–1 时刻(ε, δ)–*Top*(*k*)结果的节点标号按照簇进行划分，并为每个簇记录属于该簇的节点数量 m_j。即 m_j 表示 t–1 时刻的(ε, δ)-近似 *Top*(*k*)结果中有 m_j 个节点来自 C_j 簇。

3) sink 节点随机独立地产生$|S|-k$个 1~L 之间的自然数，并记为 $Y_1,Y_2,\cdots,Y_{|S|-k}$。对任意 Y_i $(1\leqslant i\leqslant|S|-k)$。$Pr(Y_i=j)=\frac{|C_j|-m_j}{n-k}$，其中，$1\leqslant j\leqslant L$，$|C_j|$为 C_j 簇的大小。对任意 $j(1\leqslant j\leqslant L)$，sink 节点计算产生的随机数中等于 j 的个数并记为 s_j。

4) sink 节点根据历史信息计算阈值 $d_{\mathrm{Filter}(k)}$。$d_{\mathrm{Filter}(k)}=\frac{d_{\mathrm{Top}-k}^{t-1}+d_{\mathrm{Top}-k}^{t-2}+\cdots+d_{\mathrm{Top}-k}^{t-h}}{h(1+\varepsilon)}$，其中，$d_{\mathrm{Top}-k}^{t-1}$表示抽样算法返回 t–1 时刻的(ε, δ)-*Top*(*k*)结果，按降

序排在第 k 位的数据，h 可由用户指定。

5) sink 节点向 C_j 簇的簇头节点发送(m_j, s_j, $d_{\text{Filter}(k)}$)，$1 \leqslant j \leqslant L$。

5.4 簇内节点的优化抽样算法步骤

1) 不妨设 C_j 簇的节点 ID 集合为$\{i_{C_{j1}}, i_{C_{j2}}, \cdots, i_{C_{j|C_j|}}\}$。当 C_j 簇的簇头节点接收到 sink 节点发送的(m_j, s_j, $d_{\text{Filter}(k)}$)，簇头节点根据向 sink 节点发送的 $t-1$ 时刻按降序排列的数据及对应的节点 ID，记录前 m_j 个节点 ID，不妨设为 $M_j = \{i_{j_1}, i_{j_2}, \cdots, i_{j_{m_j}}\}$。

2) 簇头节点随机独立均匀地产生 s_j 个属于 $\{i_{C_{j1}}, i_{C_{j2}}, \cdots, i_{C_{j|C_j|}}\} \backslash M_j$ 的节点 ID，不妨设这个集合为 S_j。然后簇头节点将 $M_j \cup S$ 及 $d_{\text{Filter}(k)}$ 广播至该簇内。

3) 如果 $i \in M_j \cup S_j$ 且节点 i 的感知值 $d_i \geqslant d_{\text{Filter}(k)}$，则节点 i 将查询时刻的感知值 d_i 传送给簇头节点，否则节点 i 不向簇头节点传送数据。

5.5 簇头节点的优化抽样算法步骤

簇头节点可以将接收的全部数据传送到sink节点，可为下次抽样算法更精确地估计 $\hat{P}^{\varepsilon}_{X_t}(i_1)$。但是这使得通信代价太高了。为了减少通信消耗可进行下面的操作。

1) 首先，簇头节点将接收到的数据降序排序。不妨设 C_j 簇的簇头节点接收到 r_j 个数据且 $d(j_1) \geqslant d(j_2) \cdots \geqslant d(j_{r_j})$。

2) 令 $send_j^{t-1}$ 记为 $t-1$ 时刻 C_j 簇的簇头节点向 sink 节点传送的感知数据数量，$send_j^{t-2}$ 记为 $t-2$ 时刻 C_j 簇的簇头节点向 sink 节点传送的感知数据数量，…，那么 $\omega_j = \dfrac{\sum_{i=1}^{\lfloor h/2 \rfloor}\left(send_j^{t-j} - m_j\right)}{\lfloor h/2 \rfloor \times m_j}$，当 $\omega_j \leqslant 0$ 时，令 $\omega_j = 0$。

3) 其次簇头节点计算 $d(j_1), d(j_2), \cdots, d(j_{r_j})$ 中大于等于 $\dfrac{d(j_1)}{1+\varepsilon}$ 的个数，并记录下来令 $c_j = |\{\, j_h \mid d(j_h) \geqslant \dfrac{d(j_1)}{1+\varepsilon}, 1 \leqslant h \leqslant r_j \,\}|$，簇头节点向 sink 节点传送前 $(1+\omega_j)m_j$ 个数据和 c_j。即簇头节点向 sink 节点传送 $(d(j_1), d(j_2), \cdots, d(j_{(1+\omega_j)m_j}), c_j)$。

4) 若 $r_j \leqslant (1+\omega_j)m_j$，则簇头节点向 sink 节点传送 $(d(j_1), d(j_2), \cdots, d(j_{r_j}), c_j)$。若 $m_j = 0$，则簇头节点向 sink 节点传送 $(d_{j_1}, d_{j_2}, \cdots, d_{j_{tr_j}}, c_j)$，其中，$d(j_1) \geqslant d(j_2) \geqslant \cdots \geqslant d(j_{tr_j}) \geqslant (1+\varepsilon)d_{\text{Filter}(k)}$。$c_j$ 与上面定义相同。

不妨设(ε, δ)-*Top*(k)算法返回 $t-1$ 时刻的近似 Top-k 结果属于 C_j 簇的节点 ID 为 $\{e_1, e_2, \cdots, e_{m_j}\}$。虽然 sink 节点仅向 C_j 簇的簇头节点发送 m_j，没有发送节点 ID 值，但 $\{e_1, e_2, \cdots, e_{m_j}\}$ 必为簇头节点向 sink 节点发送的 $t-1$ 时刻按降序排列的前 m_j 个数据对应的节点标号。即 $M_j = \{i_{j_1}, i_{j_2}, \cdots, i_{j_{m_j}}\} = \{e_1, e_2, \cdots, e_{m_j}\}$，故 sink 节点仅需发送节点数量 m_j，簇头节点就得知该簇中属于(ε, δ)-*Top*(k)的节点标号值。

定理 5 不妨设 $t-1$ 时刻(ε, δ)-近似 *Top*(k)算法输出的节点 ID 集合为 $ID_{(\varepsilon,\delta)}^{t-1}$，则 t 时刻节点 i 观测的数据 $d_i\left(i \in I \backslash ID_{(\varepsilon,\delta)}^{t-1}\right)$ 被优化的抽样算法抽到的概率为 $\dfrac{1}{n-k}$。

证明 不妨设 $i \in C_j$ 且 $i \in I \backslash ID_{(\varepsilon,\delta)}^{t-1}$，由优化的抽样算法步骤知：算法必然返回 $t-1$ 时刻(ε, δ)-*Top*(k)查询结果的节点在 t 时刻的感知值。又因 sink 节点随机独立地产生 $|S|-k$ 个 $1 \sim L$ 之间的自然数，$\Pr(Y_i=j) = \dfrac{|C_j| - m_j}{n-k}$，$1 \leqslant j \leqslant L$，$|C_j|$ 为 C_j 的大小。由簇内节点的优化抽样算法步骤知，C_j 簇的簇头节点随机独立均匀地产生 s_j 个属于 $\{i_{C_{j1}}, i_{C_{j2}}, \cdots, i_{C_{j|C_j|}}\} \backslash M_j$ 的节点。因 sink 节点与簇头的操作步骤是独立的，于是 $\Pr\left(i \in S_j\right) = \dfrac{|C_j| - m_j}{n-k} \times \dfrac{1}{|C_j| - m_j} = \dfrac{1}{n-k}$。

定理 5 说明本文给出的优化抽样算法能均匀独立的生成随机样本。

5.6 优化的抽样算法性能分析

sink 节点将 $t-1$ 时刻的近似 Top-k 结果按簇划分的计算代价为 $O(k)$；生成$|S|-k$ 个随机数的计算代价为 $O(|S|-k)$；计算 $d_{\text{Filter}(k)}$的代价为 $O(1)$。故 sink 节点总的计算代价为 $O(|S|)$。sink 节点的通信代价为 $O(Lh)$，h 为 sink 节点传送一个数据到簇头节点的平均跳数。

C_j 簇的簇头节点生成 s_j 个随机数的计算代价为 $O(s_j)$，簇头节点将接收到的数据排序代价为 $O(r_j \ln r_j)$，计算 c_j 的代价为 $O(r_j)$，故簇头节点总的计算代价为 $O(r_j \ln r_j)$。簇头将 $d_{\text{Filter}(k)}$单播至对应的 (m_j+s_j)个节点所需的通信代价为 $O(m_j+s_j)$。簇头将(ε, δ)–*Top*(k)候选集和 c_j 传送到 sink 节点的通信代价为 $O(h((1+w_j) \times m_j+1))$，其中，$h$ 为簇头节点传送一个数据到 sink 节点的平均跳数。综上分析，网内簇头节点总的计算代价和通信代价的上界分别为 $O(|S|\ln|S|)$、

$O\left(|S|+h\sum_{j=1}^{L}\{(1+\omega_j)m_j+1\}\right)$。其他节点总的通信代价至多为 $O(|S|)$。

如果阈值 d_{Filter} 变化不大，为了节省能量不需要每运行一次算法就广播新的阈值。如果 $\frac{\left|d_{\text{Filter}(k)}^{\text{New}}-d_{\text{Filter}(k)}^{\text{Old}}\right|}{d_{\text{Filter}(k)}^{\text{New}}}\geqslant 1+\frac{1}{2}\varepsilon$，则更新 d_{Filter} 的阈值；否则保持原来的 d_{Filter} 阈值。

5.7 (ε, δ)-近似 Top-k 算法

1) sink 节点根据用户给定的(ε, δ)确定样本容量。

2) 运行优化的抽样算法。sink 节点接收各簇头节点的数据后，输出前 k 个最大值及对应的节点。

3) 不妨设 $\hat{d}_{i_1}=\max\{d_{j_1}\,|\,1\leqslant j\leqslant L\}$，令 $CID(\hat{d}_{i_1}(\varepsilon))=\{j\,|\,(1+2^{-\lambda}\varepsilon)d_{j_1}\geqslant\hat{d}_{i_1},1\leqslant j\leqslant L\}$，简记为 $CID(\varepsilon)$。那么 $\hat{P}_X(i_1)=\frac{\sum_{j\in CID(\varepsilon)}c_j}{|S|}$，对于任意数据的感知数据集合，sink 节点通过计算 $\hat{P}_X(i_1)$ 来估计 $P_X(i_1)$。λ 可由用户设定，或者用户根据返回的结果，在下一次抽样时更新 λ 的值。

(ε, δ)-近似 *Top*(*k*)算法除了 ε' 的计算代价，主要的计算代价和通信代价集中在优化的抽样算法。抽样算法的计算代价和通信代价为分别为 $O(|S|\ln|S|)$、$O\left(|S|+h\sum_{j=1}^{L}\{(1+\omega_j)m_j+1\}\right)$。又因样本容量 $|S|=O\left(\frac{\log_{f(\varepsilon')}\delta}{\hat{P}_X(i_1)}\right)$，所以(ε,δ)-近似 *Top*(*k*)算法通信代价和计算代价分别为 $O\left(\frac{\log_{f(\varepsilon')}\delta}{\hat{P}_X(i_1)}+h\sum_{j=1}^{L}\{(1+\omega_j)m_j+1\}\right)$ 和 $O\left(\left(\frac{\log_{f(\varepsilon')}\delta}{\hat{P}_X(i_1)}\right)\ln\left(\frac{\log_{f(\varepsilon')}\delta}{\hat{P}_X(i_1)}\right)\right)$。

6 实验结果与分析

本模拟实验采用 MATLAB 编写的程序运行，计算机配置为：处理器 Pentium(R) 4 CPU 3.06GHz，内存为 504MB，Dell Optiplex 210L 台式电脑。模拟实验中的温度感知数据来源于来自 Berkeley Intel 实验室[16]的传感器网络实测，考察感知温度的真实传感器网络。

第 1 组实验主要考察网络规模不同时，(ε, δ)与优化样本大小的关系；其中，ε 与 δ 的变化范围分别为 0.07~0.28，0.03~0.24。图 1 考察了网络节点数量为 1 600 时，根据推论计算的优化样本容量与(ε, δ)的关系。从图 1 中观察到仅需少量的样本就可以满足(ε, δ)的误差要求。例如当 ε=0.12、δ=0.09 时，样本容量为 204。即当均匀抽样的样本大小仅占全局网络的 13%时，就能满足算法输出的近似 Top-*k* 结果的平均相对误差小于 ε/(1+ε)=0.12/(1+0.12)的概率大于 0.91。

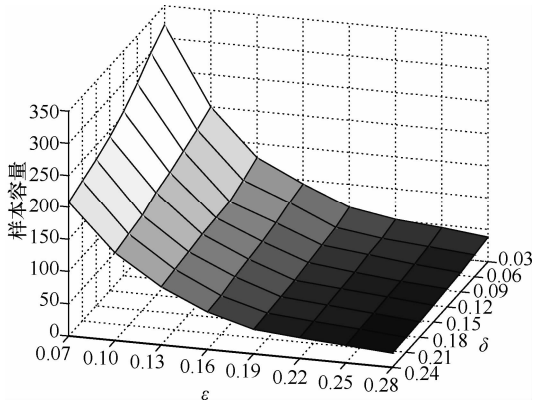

图 1　网络规模为 1 600 时，均匀随机抽样的样本容量

第 1 组实验说明：均匀抽样算法仅需少量的样本就可以满足(ε, δ)的误差要求，可以为传感器网络节约大量的能量，延长网络寿命。

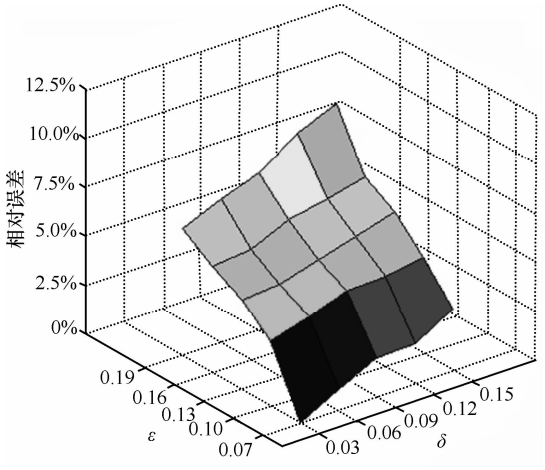

图 2　网络规模为 400 时，Top-8 的平均相对误差

第 2 组实验考察本文提出的优化抽样算法是否能够满足(ε, δ)的近似要求。当网络规模为 400 时，通过考察图 2 看到算法输出的结果满足(ε, δ)的近似要求，并使得相对误差可以任意小。当 ε=0.16， δ=0.12 时，Top-8 的平均相对误差为 7.56%；当 ε=0.10，δ=0.09 时，Top-8 的平均相对误差为 5.66%。图 3 考察了网络大小为 1600 时，算法输出的近似 Top-64。当 ε=0.13，δ=0.09 时，Top-64 平均相对误差为 9.01%。当 ε=0.10，δ=0.06 时，Top-64 的相对误差为 5.94%。

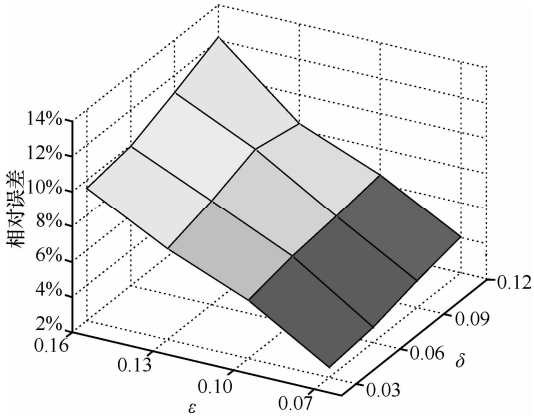

图 3　网络规模为 1 600 时，Top-64 的平均相对误差

从实验结果看到算法输出的结果不仅能保证(ε, δ)的近似要求使得相对误差可以任意小，而且接近真实的 Top-k 结果。进一步地，实验结果验证了理论的正确性。

第 3 组实验考察样本大小固定时针对不同的 Top-k 查询，抽样算法运行次数与相对误差之间的关系。从实验结果可以观察到网络规模固定时，随着抽样次数的增加，较大的抽样比例使得误差下降速度更快。如图 4 与图 5 所示，网络节点数量为 800，样本容量为 96 时，抽样算法运行 5 次使得 Top-16、Top-32、Top-48、Top-64 及 Top-80 的平均相对误差分别从 4.12%、5.77%、7.83%、9.75%、11.43%下降到 3.54%、5.04%、6.25%、7.68%、8.92%；当样本容量为 110 时，运行抽样算法 5 次使得 Top-16、Top-32、Top-48、Top-64 及 Top-80 的平均相对误差分别从 3.87%、4.93%、6.27%、8.04%、9.44%下降到 2.62%、3.6%、4.04%、4.93%、6.13%。实验结果说明，本文提出的优优化抽样算法能高效地处理连续的 Top-k 查询。

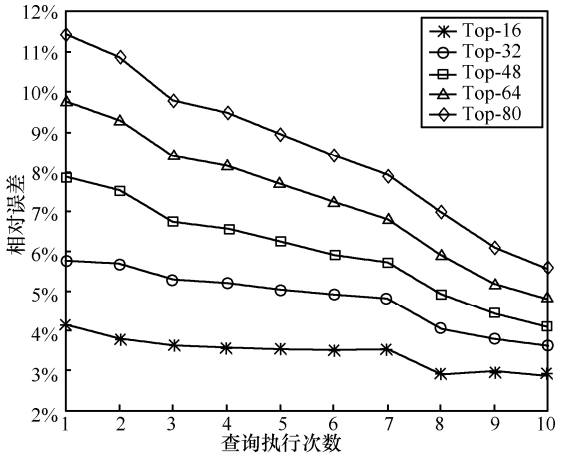

图 4　网络规模为 800，样本容量为 96 时的平均相对误差

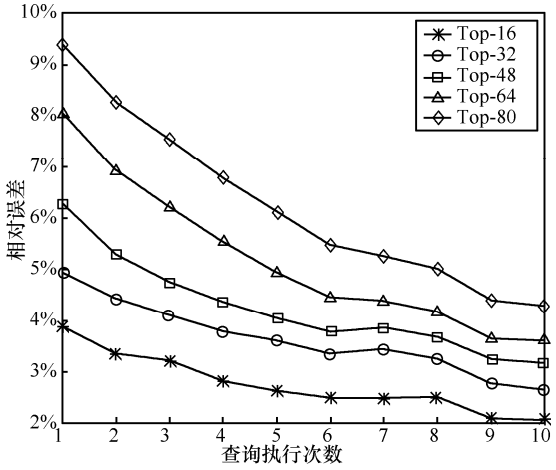

图 5　网络规模为 800，样本容量为 110 时的平均相对误差

第 2 组实验与第 3 组实验说明在传感器网络环境下，优化的抽样算法适和处理 Top-k 查询。由于采用独立均匀地抽样方式，即使网络数据变化比较剧烈，方差较大时，算法也能显示出高效性与健壮性。

第 4 组实验考察当 k 值变化时，算法输出的近似 Top-k 结果的误差界与样本大小的关系。当网络规模固定时，随着抽样比例的增加，误差明显地下降；当样本大小固定时，k 值越小误差越小。如图 6 所示当网络节点数量为 400 时，抽样比例从 9%增加到 18.75%时，Top-24 的相对误差从 10.03%下降到 2.87%；当抽样比例为 13%时，Top-8、Top-16、Top-24、Top-32、Top-40 的平均相对误差分别为 4.57%、5.46%、6.7%、8.08%和 9.36%。

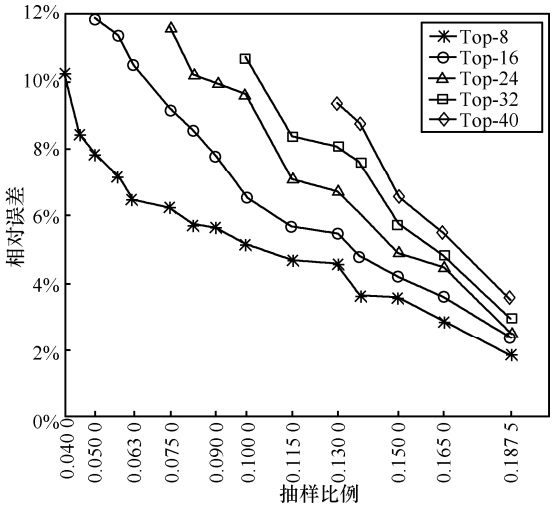

图 6　网络规模为 400 时，Top-k 查询结果的相对误差

第 5 组实验主要考察抽样算法在特定的网络规模下，执行一段时间内的 Top-k 查询的性能；网络

节点数量从 400 变化至 1 200。图 7~图 9 的实验结果显示出基于抽样的近似 Top-k 算法能够适应各种不同的网络规模和不同的 Top-k 查询。随着抽样次数增加，相对误差有明显下降，算法输出的结果非常接近真实的 Top-k 结果；这也使得阈值设置更加准确，准确的阈值可以过滤掉大量的非 ε-近似 Top-k 数据数据，降低了 $d_{\mathrm{Filter}(k)}$的更新频率，减少了网络通信能量，延长了网络寿命。

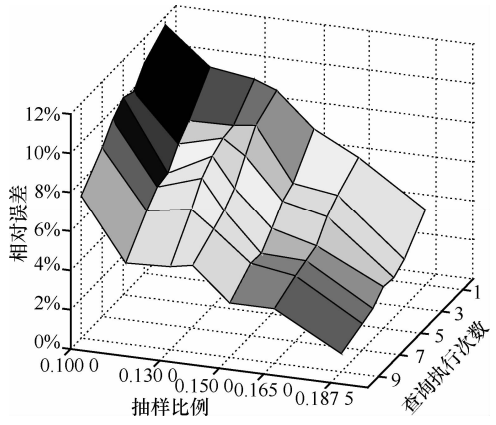

图 7　网络规模为 400 时，连续 Top-32 查询结果的相对误差

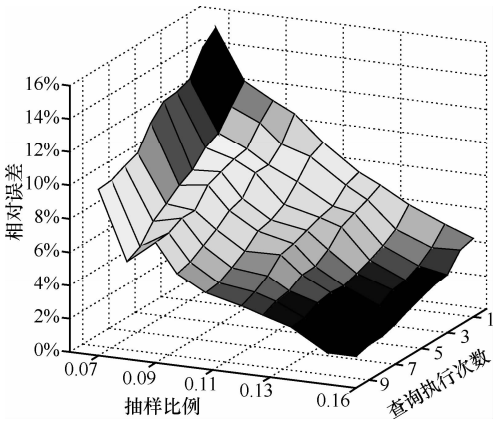

图 8　网络规模为 800 时，连续 Top-48 查询结果的相对误差

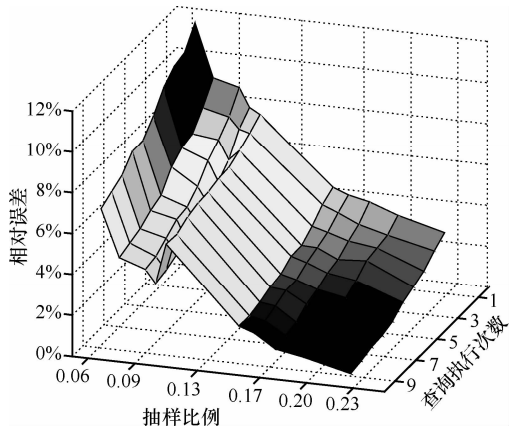

图 9　网络规模为 1 200 时，连续 Top-72 查询结果的相对误差

7　结束语

在传感器网络的实际应用中，用户仅需要近似结果而不是精确结果。已有的 Top-k 查询处理算法无法满足用户的任意误差要求。本文对无线传感器网络中近似 Top-k 查询处理技术进行了研究。首先给出了(ε, δ)-近似 Top-k 查询形式化定义和严格的语义定义，并证明了优化的样本容量。其次，设计了优化的抽样算法，通过数据过滤技术来减少通信能量的消耗。文中对算法正确性进行了证明，并给出算法的理论分析。通过模拟实验验证了优化的抽样算法不仅能高效地处理近似 Top-k 查询，而且有效地减小查询中的通信开销，对于延长网络寿命有很大的作用。

参考文献：

[1] ZEINALIPOUR-YAZTI D, VAGENA Z, GUNOPULOS D, *et al*. The threshold join algorithm for Top-k queries in distributed sensor networks[A]. ACM International Conference Proceeding Series[C]. 2005. 61-66.

[2] BABOCK B, OLSSTON C. Distributed Top-k monitoring[A]. SIGMOD[C]. 2003. 28-39.

[3] CHAUDHURI S, GRAVANO L, MARIAN A. Optimizing Top-k selection queries over multimedia repositories[J]. IEEE Transactions on Knowledge and Data Engineering, 2004,16(8): 992-1009.

[4] SEBASTIAN M, PETER T, GERHARD W. KLEE: a framework for distributed Top-k query algorithm[A]. VLDB[C]. 2005. 637-648.

[5] ADAM S, REBECCA B, CARLA S E, *et al*. A sampling-based approach to optimizing Top-k queries in sensor networks[A]. ICDE[C]. 2006. 68-78.

[6] RONALD F, AMNON L, MONI N. Optimal aggregation algorithm for middleware[A]. PODS[C]. 2001. 102-113.

[7] THEOBALD M, WEIKUM G, SCHENKEL R. Top-k query evaluation with probabilistic guarantees[A]. VLDB[C]. 2004. 648-659.

[8] ARAI B, DAS G, GUNOPULOS D, *et al*. Anytime measuers for Top-k algorithms[A]. VLDB[C]. 2007. 648-659.

[9] GUNTAER U, BALKE W T, KIEBLING W. Optimizing multi-feature queries for image databases[A]. VLDB[C]. 2000. 419-428.

[10] CAO P, WANG Z. Efficient Top-k query calculation in distributed netwrks[A]. PODC[C]. 2004. 206-215.

[11] LIU Y, LI J Z, GAO H, *et al*. Enabling-approximate querying in sensor networks[A]. VLDB[C]. Lyou, France, 2009.169-180.

[12] CHENG S Y, LI J Z. Sampling based(ε, δ)-approximate aggregation algorithm in sensor networks[A]. ICDCS[C]. Montreal, Canada, 2009.273-280.

[13] CHENG S Y, LI J Z. Bernoulli sampling based(ε, δ)- approximate aggregation algorithm in sensor networks[A]. INFOCOM[C]. San

Diego USA, 2010.1181-1189.

[14] CHU D, DESHPANDE A, HELLERSTEIN J M, *et al*. Approximate data collection in sensor networks using probabilistic models[A]. ICDE[C]. 2006. 48-59.

[15] MICHAEL M, ELI U. Probability and Computing: Randomized Algorithms and Probabilistic Analysis[M]. Cambridge Press, 2005.

[16] Intel lab dataset[EB/OL]. http://db.csail.mit.edu/labdata/labdata.html.

作者简介：

毕冉（1985-），女，黑龙江牡丹江人，哈尔滨工业大学博士生，主要研究方向为传感器网络。

李建中（1950-），男，黑龙江哈尔滨人，哈尔滨工业大学教授、博士生导师，主要研究方向为传感器网络、数据挖掘和数据仓库。

程思瑶（1982-），女，黑龙江哈尔滨人，哈尔滨工业大学博士生，主要研究方向为对等计算和传感器网络。

（上接第 44 页）

[6] FAN P Z, YUAN W N, TU Y F. Z-complementary binary sequences[J]. IEEE Signal Processing Letters, 2007, 14(8): 509-512.

[7] CHEN H H, YEH Y C, ZHANG X. Generalized pairwise complementary codes with set-wise uniform interference-free windows[J]. IEEE Journal on Selected Areas in Communications, 2006, 24(1): 65-74.

[8] FENG L F, FAN P Z, TANG X H. Generalized pairwise Z-complementary codes[J]. IEEE Signal Processing Letters, 2008, 15(1): 377-380.

[9] ZHANG Z Y, CHEN W, ZENG F X. Z-complementary sets based on sequences with periodic and aperiodic zero correlation zone[J]. EURASIP Journal on Wireless Communications and Networking, 2009, 2009(3):1-8.

[10] LEOPOLD B, MARKUS A. Periodic complementary binary sequences[J]. IEEE Transactions on Information Theory, 1990, 36(6):1487-1494.

[11] FENG K Q, PETER J S S, XIANG Q. On aperidic and periodic complementary binary sequences[J]. IEEE Transactions on Information Theory, January 1999, 45(1):296-303.

作者简介：

李玉博（1985-），男，河北衡水人，燕山大学博士生，主要研究方向为扩频序列设计。

许成谦（1961-），男，陕西城固人，博士，燕山大学教授、博士生导师，主要研究方向为编码理论、密码学、信号设计等。