ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.5, May 2010, pp.1020−1030
doi: 10.3724/SP.J.1001.2010.03703

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

无线传感器网络中一种近似 Skyline 查询处理算法*

潘立强+, 李建中, 骆吉洲

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Approximate Skyline Query Processing Algorithm in Wireless Sensor Networks

PAN Li-Qiang+, LI Jian-Zhong, LUO Ji-Zhou

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

+ Corresponding author: E-mail: panlq@hit.edu.cn

Pan LQ, Li JZ, Luo JZ. Approximate Skyline query processing algorithm in wireless sensor networks. *Journal of Software*, 2010,21(5):1020−1030. http://www.jos.org.cn/1000-9825/3703.htm

Abstract: Due to the limitation of wireless sensor networks in energy resources and the fact that part of Skyline query results can satisfy the users in many applications, this paper proposes an energy efficient approximate Skyline query processing algorithm to save the energy maximally according to the different requirements of applications. The proposed algorithm can compute an approximate Skyline result set only by making partial sensor nodes transmitting their sensing data back. And it is energy efficient because each sensor node transmits its sensing data back or not only depending on the information of itself, without the comparison with other data. Accordingly, communication cost is greatly reduced and network energy is greatly saved. Extensive experiments in simulation environment indicate that the proposed algorithm can process the approximate Skyline queries in wireless sensor networks energy efficiently according to different requirements of applications.

Key words: sensor network; algorithm; Skyline; query processing; approximate query

摘 要: 由于无线传感器网络的能源有限,且在许多应用中 Skyline 查询的部分结果即可满足用户需求,提出了一种近似Skyline查询处理算法,在满足用户查询需求的前提下最大化地节省能量.该算法仅需无线传感器网络中的部分传感器节点回传其感知数据即可计算出Skyline查询的一个近似结果集.由于该算法在处理查询时,每个传感器节点只需考察自身数据信息即可决定是否回传其感知数据,而无须与其他传感器节点的感知数据进行比较,因此可以避免大量的网内通信开销,从而节省网络能源.模拟环境下的大量实验结果表明,该算法可以根据用户的应用需求,节能地处理传感器网络中的近似 skyline 查询.

关键词: 传感器网络;算法;Skyline;查询处理;近似查询

中图法分类号: TP311 **文献标识码**: A

无线通信技术、微电子技术及嵌入式计算技术的快速发展使无线传感器网络得到了广泛应用,无线传感器

* Supported by the National Natural Science Foundation of China under Grant Nos.60533110, 60703012, 60773063 (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 NSFC/RGC Joint Research Scheme under Grant No.60831160525 (NSFC/RGC 联合资助项目)

Received 2008-09-27; Revised 2009-02-16; Accepted 2009-07-07

网络技术也迅速成为研究的热点[1,2].无线传感器网络由分布于特定区域的很多传感器节点构成,每个节点均具有一定的计算能力和存储能力.无线传感器网络可以被看作是一个新型的分布式数据库系统[3−5].人们通过对无线传感器网络发出查询来获取被监测区域的信息,以满足各种应用.由于传感器节点的能量、计算能力、存储能力、通信能力和网络传输带宽均有限,已有的数据库查询处理技术不适于无线传感器网络.目前,无线传感器网络数据查询处理算法的研究是传感器网络研究中的重要内容之一.

Skyline 查询作为一类多目标优化问题,近年来受到广泛关注,但是在无线传感器网络领域却少有研究.由于 Skyline 查询能够在没有目标满足查询条件时返回给用户一个近似的查询结果集,因此在无线传感器网络中具有广泛应用.例如,在如图 1 所示的一个智能大厦中,用户想找一个温度接近于 25°C、并且湿度接近于 35%的房间召开会议.由于 r_1,r_2,r_3 是最接近于查询目标的 3 个房间,因此 Skyline 查询会将房间 r_1,r_2,r_3 作为查询结果返回给用户.基于 Skyline 查询结果,用户可以选择自己满意的房间召开会议.又如,在野生鱼类的研究中,当科学家需要寻找生活在一定水温、一定水流速度,并具有一定种群数量的鱼群作为研究对象时,可以通过 Skyline 查询找到最接近于这一查询条件的多个目标样本,然后从中选择一个作为研究对象.

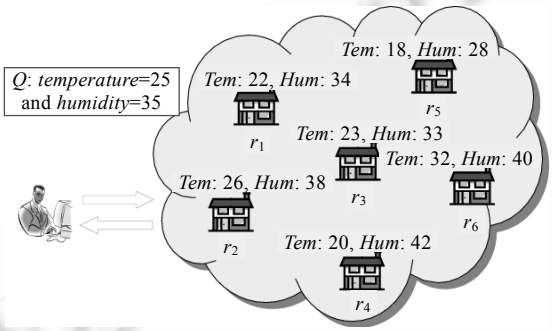


Fig.1 Meeting rooms with wireless sensor networks

图 1 部署有传感器网络的会议室

从上述两个例子中我们注意到,在许多情况下,用户并不需要全部的 Skyline 查询结果,其部分结果集就可以完全满足用户的应用需求.例如,在上述的两个例子中,用户只需要 3 个房间中的 1 个房间开会,科学家只需要多个目标样本中的 1 个样本作为研究对象.由于 Skyline 查询结果集中的多个查询结果从某种意义上讲对查询条件的近似程度是等价的,因此在上述例子中,Skyline 查询结果集中的任意一个结果均可近似满足用户的需求.尽管返回 Skyline 查询的全部结果可以提供给用户更多的选择,但是很多时候这是不必要的.例如,房间 r_1,r_2 和 r_3 的环境本身相差不大,其与查询条件在温度和湿度上的偏差分别为(3,1),(1,3)和(2,2),因此我们很难区分哪个房间与查询条件更近似,因此用户并不在意选择哪一个房间开会.

由于 Skyline 查询结果集是通过数据之间的比较计算得出的.因此,计算 Skyline 查询的完整结果集与只计算 Skyline 查询的一个部分结果集,其代价是不同的.显然,计算并返回 Skyline 查询的全部结果需要更多的网络通信开销和工作负载,从而会消耗更多的网络能量.考虑到在能源极其宝贵的无线传感器网络中,在满足用户需求的前提下最大限度地节省能源开销是我们设计各种查询处理算法的首要目标.因此,当用户不严格需要全部的 Skyline 查询结果时,我们可以根据用户的需求计算并返回 Skyline 查询的部分结果集,从而有效降低传感器网络中的能量消耗,延长网络生命期.

本文首先形式化地定义了近似 Skyline 查询,然后给出了一种近似 Skyline 查询处理算法.该算法可以根据用户指定的近似度δ,使传感器网络只回传部分感知数据就可以计算得到 Skyline 查询的一个部分结果集.在模拟数据和真实数据上的大量实验表明,本文提出的算法能够在满足用户需求的前提下,高效节能地处理近似 Skyline 查询.本文第 1 节介绍相关工作.第 2 节给出近似 Skyline 查询的形式化定义.第 3 节介绍近似 Skyline 查询处理算法.第 4 节是实验结果及分析.第 5 节给出结论.

1 相关工作

目前,传感器网络中数据查询处理技术的研究主要集中在连续查询和近似查询两方面.在连续查询方面,人们主要研究如何结合具体的网络拓扑结构或其他系统特征生成优化的查询计划,高效节能地地将满足查询的感知数据回传到Sink节点[4,6−10].在近似查询处理方面,人们主要研究如何利用传感器网络中感知数据的时空相关性建立恰当的数学模型来近似地回答查询,避免大量感知数据的回传[5,11−14].

文献[15]研究了关系数据库系统中的Skyline查询问题,提出了Skyline查询操作符并给出了BNL和D&C两种算法.文献[16]给出了一种基于预排序的Skyline计算方法.文献[17]提出了两种改进的Skyline查询方法.文献[18]提出了NN算法,文献[19]对NN算法进行了改进,提出了BBS算法.这些方法只适用于集中式数据库系统,不适于分布式环境,更不适用于无线传感器网络.

文献[20]研究了分布式环境下的Skyline查询问题,给出了Web应用的BDS算法和IDS算法.BDS算法通过对存储在各服务器上的数据进行排序,可以找到包含Skyline数据的一个子集,从而将Skyline的计算空间由全部数据减少为部分数据.IDS算法在BDS算法的基础上采用了启发式方法,可以更快地找到包含Skyline数据的子集.文献[21]提出了PDS算法,该算法通过progressiveness和rank estimation策略进一步地提高了文献[20]中的算法性能.文献[22,23]研究了P2P环境下的Skyline查询问题.文献[22]给出了DSL算法,分布式地递增寻找Skyline点.文献[23]通过对查询子空间的peer节点进行估计,降低了peer节点的访问数及查询数据的传输量.此外,文献[23]基于平衡树结构在一定程度上解决了查询热点问题,平衡了Peers节点的查询负载.文献[24,25]还考虑了移动环境下的Skyline查询及数据存储问题.但是,这些方法均没有考虑传感器节点能量稀缺、计算和存储能力有限、通信带宽低的特点,因此也不适用于无线传感器网络.

文献[26,27]研究了传感器网络环境下的Skyline连续查询方法,其优化目标是如何降低查询结果的更新代价,而不是降低Skyline查询的计算代价,从而不适用于ad hoc查询.文献[28]介绍了一种网内Skyline查询处理算法,但是该算法讨论的是如何计算Skyline查询的全部结果,而不是讨论如何根据用户的需求来高效节能地求解Skyline查询的一个部分结果集.由于在许多情况下,Skyline查询的一个近似结果集即可满足用户需求,并且求解Skyline查询的完全结果集和求解其近似结果集的代价是不同的,因此文献[28]不适用于无线传感器网络中的近似Skyline查询.

2 问题定义

在无线传感器网络中,通常每个节点可以监测k个环境属性$A_1,\ldots,A_k$,传感器节点在某一时刻的监测值可以看作是一个k元组$r=\langle a_1,\ldots,a_k\rangle$,其中,$a_k$表示传感器节点在$A_k$上的监测值.在某一时刻,由所有传感器节点产生的感知数据构成数据空间R.假设对于任意的查询Q,该k元组与Q的查询条件在属性A_k上的偏差为d_k,则我们可以定义元组r相对于查询Q的距离向量为$d^r=\langle d_1,\ldots,d_k\rangle$.例如,图1中的房间$r_1$,$r_2$和$r_3$相对于查询$Q$的距离向量可以表示为$d^{r_1}=\langle 3,1\rangle$, $d^{r_2}=\langle 1,3\rangle$和$d^{r_3}=\langle 2,2\rangle$.

定义1. 假设元组r和s与查询Q的距离向量分别$d^r=\langle d_1^r,\ldots,d_k^r\rangle$和$d^s=\langle d_1^s,\ldots,d_k^s\rangle$,如果$d_i^r\le d_i^s$对$\forall 1\le i\le k$成立且$d_i^r<d_i^s$对某个$1\le i\le k$成立,则称元组$r$支配元组$s$,记为$r\to s$;否则,称元组$r$不支配元组$s$,记为$r\not\to s$.

定义2. 对于任意的查询Q,如果该查询返回的结果集B满足$B=\{r\mid r\in R且\nexists s\in R,s\to r\}$,则我们称该查询为Skyline查询,集合$B$称为Skyline查询结果集,记为$SK$.

引理1. 假设SK为Skyline查询结果集,则对$\forall s\in R-SK$,$\exists r\in SK$使得$r\to s$.

引理2. 对于$\forall r,s,t\in R$,如果$r\to s$且$s\to t$,则$r\to t$.即支配关系满足传递性.

定义3. 假设$SK\subseteq R$是Skyline查询结果集,如果查询Q返回的结果集D满足$D\subseteq SK$且$|D|\ge\delta\cdot|SK|$,$0<\delta\le 1$,则该查询称为δ-近似Skyline查询,集合D称为δ-近似Skyline查询结果集,记为ASK_δ.

3 算法描述

由定义 3 可知,处理近似 Skyline 查询的最简单方法是在无线传感器网络中进行 Skyline 查询,然后将查询结果 *SK* 的一个子集返回给用户.显然,该方法由于回传了大量无用数据,是浪费能量的.由于用户只需要 *SK* 的一个子集,因此,我们只需回传属于 *SK* 的部分数据就可以满足用户的查询需求.下面我们介绍近似 Skyline 查询处理算法——AS(approximate Skyline)算法.

AS 算法的基本思想是:每个传感器节点都计算自己的感知数据属于 Skyline 查询结果集的概率,当且仅当该概率值大于某一阈值(由 Sink 节点计算并下发)时才将自己的感知数据回传.最后,Sink 节点在回传的数据集合上计算查询结果并返回给用户.该算法的关键是如何在每个传感器节点上计算感知数据的概率值,以及 Sink 节点如何计算阈值.下面,我们先给出相关的概念,然后讨论概率值和过滤阈值的计算.

定义 4. 如果某传感器节点上的监测数据 r 不能被其他节点的监测数据所支配,则称监测数据 r 有效.监测数据 r 有效的概率(即 $P\{r\in SK\}$)称为数据 r 的有效率,记为 p_r.

由于每个传感器节点在任意时刻均最多产生 1 个感知数据,数据 r 的有效率也可以被看成是产生该数据的传感器节点的数据有效率.

定义 5. 给定 $1\geq p^*>0$,如果要求传感器网络中的任意感知数据 r 当且仅当满足 $p_r\geq p^*$时被回传,则称 p^*为传感器网络的数据发送阈值.

定义 6. 假设 R 是查询时刻由所有传感器节点产生的感知数据构成的数据空间,对于$\forall r\in R$,r 相对于查询 Q 的距离向量为 $d^r=\langle d_1,\ldots,d_k\rangle$,则我们称由所有 d^r 构成的集合 $R^Q=\{d^r|r\in R\}$为感知数据在查询 Q 上的距离空间.

3.1 数据有效率的计算

为方便说明问题,我们假定每个节点的感知数据是一个具有两个属性的二维数据.因此,每个感知数据的距离向量 d^r 都可以表示成二维距离空间上的一个点,如图 2 所示.直观上,如果感知数据 r 的距离向量 d^r 能够支配距离空间中的点越多,则说明 r 的有效率越大;反之,p_r 越小.为了刻画 d^r 能够支配的距离向量的个数,下面引入最大边界和支配区域的概念.

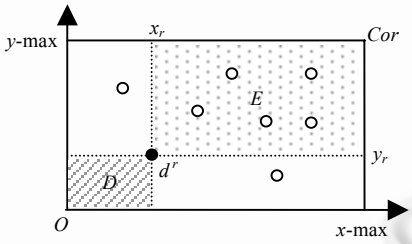


Fig.2 Two dimensional distance space

图 2 二维距离空间

定义 7. 设 R^Q 是 R 到查询 Q 的距离空间,定义 R^Q 在第 i 维上的最大边界为 $\max\{d_i^r,r\in R\}$,记为 i_max.向量 $\langle 0,\ldots,0\rangle$和$\langle 1_\max,\ldots,k_\max\rangle$分别称为 R^Q 的最小边界点和最大边界点.

定义 8. 对于$\forall r\in R$,称以 d^r 为左下(右上)角顶点、以 R^Q 的最大(最小)边界点为右上(左下)角顶点的 k-维长方体为 r 的(被)支配区域.

在图 2 中,最大边界点为 *Cor*,最小边界点为 *O*,数据 r 的支配区域为 *E*,它的被支配区域为 *D*.如果感知数据 r 的被支配区域表示为 D_r,则 $r\in SK$ 当且仅当 D_r 中不包含任何其他感知数据对应的距离向量,即

$$p_r=P\{r\in SK\}=P\{\forall s\in R(s\neq r),d^s\notin D_r\}.$$

这说明,只需要计算传感器网络中其他节点的感知数据所对应的距离向量均不出现在 D_r 中的概率.

设 $f(X_1,\ldots,X_k)$是传感器网络中的感知数据在其数据空间上分布的概率密度函数,对于给定的查询 Q,由定义

1 可知,感知数据 r 到 Q 的距离向量 d^r 是 r 的函数,即 $d^r=g_Q(r)$.由函数 g_Q 和 f,传感器节点可以计算得到 d^r 在其值域[0,1_max]×…×[0,k_max]分布的概率密度函数 $G_Q(X_1,…,X_k)$.于是,对$\forall s(\neq r)$,$d^s\in D_r$ 的概率等于

$$\int_{D_r} G_Q(X_1,...,X_k)\mathrm{d}X_1...\mathrm{d}X_k,$$

进而有

$$P\{d^s\notin D_r\}=1-\int_{D_r} G_Q(X_1,...,X_k)\mathrm{d}X_1...\mathrm{d}X_k.$$

由于传感器网络中的 n 个节点相互独立地监测周围环境,故 $p_r=[1-\int_{D_r} G_Q(X_1,...,X_k)\mathrm{d}X_1...\mathrm{d}X_k]^{n-1}$.在实际应用中,由于传感器网络中的节点数量巨大,因此计算上式的开销非常大.注意到,$1-\int_{D_r} G_Q(X_1,...,X_k)\mathrm{d}X_1...\mathrm{d}X_k$ 即反映了 p_r 的大小.因此,在实际应用时可以直接认为

$$p_r=1-\int_{D_r} G_Q(X_1,...,X_k)\mathrm{d}X_1...\mathrm{d}X_k \tag{1}$$

概率密度函数 $f(X_1,…,X_k)$可由 Sink 节点通过对近期传感器网络中的感知数据进行抽样统计来获得,并将它周期性地广播到整个网络中.概率密度函数 $f(X_1,…,X_k)$还可以支持无线传感器网络中的其他应用[13,14].

3.2 数据发送阈值的选择

数据发送阈值 p^*由 Sink 节点计算并随查询 Q 一同下发.传感器节点根据数据有效率 p_r 是否大于 p^*来决定是否将其感知数据 r 回传.p^*的选取与网络能量消耗和查询结果对 SK 的近似程度直接相关.本节将讨论如何根据用户指定的$\delta(0<\delta\leq 1)$计算得到一个数据发送阈值 p^*,使得最终得到的δ-近似 Skyline 查询结果集 ASK_δ满足 $E(|ASK_\delta|)\geq\delta\cdot E(|SK|)$.

先暂时假设 Sink 节点可以获得传感器网络中所有 n 个节点的数据有效率 p_i.将 p_i 从大到小排成一个序列 S,不妨设 $S=\langle p_1,…,p_n\rangle$.由定义 4 可知,任一节点 N_i 上的感知数据 r_i 属于 SK 的概率等于 p_i,因此,在由 n 个节点构成的传感器网络中,SK 所包含的感知数据个数$|SK|$的数学期望为 $E(|SK|)=\sum_{1\leq i\leq n} p_i$.如果 p^*满足 $p_m\geq p^*>p_{m+1}$,则用 p^*作为数据发送阈值得到的近似 Skyline 查询结果集的大小的数学期望为 $\sum_{1\leq i\leq m} p_i$.对用户给定的δ,如果 $E(|ASK_\delta|)\geq\delta\cdot E(|SK|)$成立,亦即 $\sum_{1\leq i\leq m} p_i\geq\delta\cdot\sum_{1\leq i\leq n} p_i$ 成立.

基于以上分析我们可以看出,选择 p^*使得 $E(|ASK_\delta|)\geq\delta\cdot E(|SK|)$成立且使网络中传输的数据量最小,等价于选择最小的 $m\in\{1,…,n\}$使得 $\sum_{1\leq i\leq m} p_i\geq\delta\cdot\sum_{1\leq i\leq n} p_i$ 成立.

在实际应用中,由于 Sink 节点不能有效获得 $p_1,…,p_n$,因此不能按照上述过程来计算 p^*.AS 算法采用线性回归模型,通过历史数据来估计查询 Q 的数据发送阈值 p^*.设 Sink 节点的近期历史数据是对网络中的所有 n 个节点进行 t 次抽样得到的.对于第 i 次抽样得到的感知数据集合 R_i,Sink 节点可以在 R_i 上计算出节点 $N_j(1\leq j\leq n)$对查询 Q 的数据有效率 $p_j^{(i)}$.进而,按照上一段介绍的 p^*计算过程,Sink 节点可以计算得到在第 i 次样本数据上的数据发送阈值 $p_{(i)}^*$.这样,对于 t 个样本数据,Sink 节点总共可以得到 t 个数据发送阈值 $p_{(1)}^*,...,p_{(t)}^*$,令 $Y=\langle p_{(1)}^*,...,p_{(t)}^*\rangle^T$.

用户提交查询 Q 后,Sink 节点收集其一跳邻居节点上感知数据对 Q 的数据有效率,不妨设这些节点的标号依次是 $N_1,…,N_w$,计算 $x=\sum_{1\leq j\leq w} p_j$ (即 x 是 Sink 节点的一跳邻居节点上感知数据对 Q 的数据有效率之和).同时,对 $1\leq i\leq t$ 计算在第 i 次样本数据上,Sink 节点的一跳邻居节点上感知数据对 Q 的数据有效率之和 $x_{(i)}=\sum_{1\leq j\leq w} p_j^{(i)}$.令 $X=\langle x_{(1)},…,x_{(t)}\rangle^T$.接下来,Sink 节点利用上述 Y 和 X 求解线性回归模型 $Y=\alpha\cdot X+\beta$中的系数α和β,然后令 $p^*=\alpha\cdot x+\beta$ 作为本次查询的数据发送阈值.

AS 算法主要包括以下 6 步:

(1) Sink 节点收集其一跳邻居节点当前的感知数据.

(2) Sink 节点根据其收集到的感知数据和历史数据按照用户指定的近似度δ计算数据传送阈值 p^*.

(3) Sink 节点向无线传感器网络中发布查询 Q 和 p^*.

(4) 传感器节点计算其感知数据 r 的有效率 p_r.如果 $p_r \geq p^*$,则将 r 向 Sink 节点回传.

(5) 路由节点对回传的数据进行过滤,丢弃被其他数据支配的感知数据.

(6) Sink 节点在返回的数据集合上计算 ASK_δ,并输出.

定理 1. 设 ASK_δ和 SK 分别表示 AS 算法返回的结果集和 Skyline 查询结果集,则 $ASK_\delta \subseteq SK$.

证明:用反证法.假设$\exists r \in ASK_\delta - SK$,则由引理 1 可知,必存在 $s \in SK$ 使得 $s \rightarrow r$.由 AS 算法的第(5)步和第(6)步可知,ASK_δ中不存在满足支配关系的感知数据,故 $s \notin ASK_\delta$.由于 $s \rightarrow r$,根据定义 8 可知 $D_s \subseteq D_r$,于是,由公式(1)有 $p_s \geq p_r$.由 $r \in ASK_\delta$可知,$p_r \geq p^*$,于是,$p_s \geq p^*$也成立.因此 AS 算法将回传 s.注意到,算法的第(5)步和第(6)步不可能过滤掉 s,故 $s \in ASK_\delta$,矛盾. □

由上面的定理可知,AS 算法计算得到的结果集 ASK_δ是 Skyline 查询结果集 SK 的子集.因此,ASK_δ可以作为 Skyline 查询的近似结果返回给用户.

4 实验结果及分析

我们用 Java 开发了一个无线传感器网络模拟环境.基于该模拟环境,我们测试了本文所提出算法的性能,并与 MFTAC 算法进行了比较.MFTAC 算法是文献[28]提出的一种 Skyline 查询处理算法.该算法通过网内计算来减少数据通信量,从而节省能量.用本文算法与 MFTAC 算法进行比较来考察近似 Skyline 查询相对于确切的 Skyline 查询的节能效率.在本文的实验中,我们主要考察了当节点数量、节点通信半径和查询属性的数量变化时,各种算法执行查询时的能量消耗;此外,我们还考察了在这些参数下 AS 算法返回的近似结果集满足用户查询精度要求δ的概率 ACC_δ.

为了直观地考察算法的节能性,我们用“能量消耗比”作为衡量算法节能效率的标准.它反映了所给算法相对于 Naïve 方法节省能量的效率.在本文中,Naïve 方法是在处理查询时,将所有传感器节点的感知数据全部回传到 Sink 节点,然后在 Sink 节点计算 SK 或 ASK_δ.对于算法 $method \in \{AS,MFTAC\}$,其能量消耗比定义为 $f_{method}=E_{method}/E_{Naïve}$,其中,$E_{method}$表示算法 $method$ 处理一次查询所需消耗的能量,$E_{Naïve}$表示 Naïve 方法处理一次查询所需消耗的能量.在本文中,我们假定任意两个传感器节点间每传输一个数据需要消耗 1 个单位的能量.

4.1 实验设置

实验中采用的生成数据为多维独立数据,其中,各维属性的取值独立均匀分布于整数区间[1,1000].在我们的模拟环境中,传感器节点随机分布在 300m×300m 的平面区域内,Sink 节点位于监测区域的左上角,其坐标用(0,0)表示.传感器节点的数量变化范围是 200 个~800 个,默认设置为 300 个.由于查询属性个数会影响到数据元组之间的支配关系,进而影响 Skyline 查询结果集的大小,因此,实验设定查询属性个数的变化范围是 2~5,默认设置是 3.由于传感器节点通信半径会影响到传感器网络的拓扑结构,进而影响查询处理的能量开销;由于通信半径小于 30m 时网络可能不连通;通信半径大于 110m 时几乎所有节点都可以在 3 跳之内将感知数据送达 Sink 节点,从而失去了传感器网络多跳转发感知数据的特性,因此,实验设定节点的通信半径的变化范围是 30m~110m,默认设置为 50m.实验中,当考察某特定参数变化时,其他参数均设置为默认值.

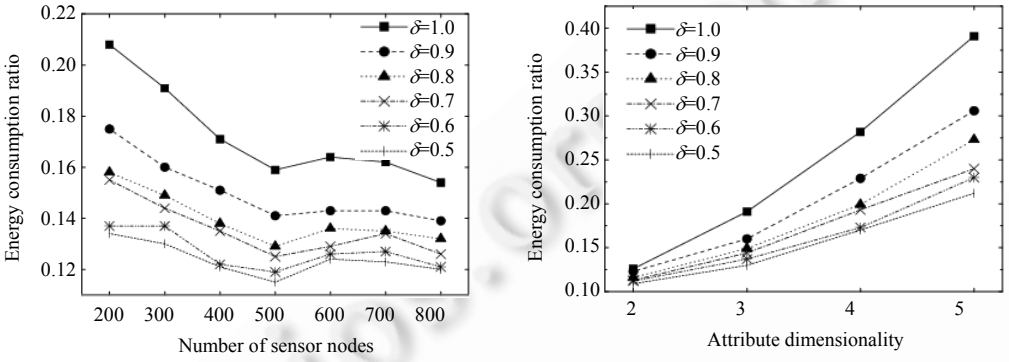
4.2 近似Skyline查询处理算法性能

本节主要研究 AS 算法在不同查询精度下的性能.实验考察了 AS 算法的能耗比和 AS 算法满足查询精度δ的概率 ACC_δ.其中,δ依次取 1.0,0.9,0.8,0.7,0.6,0.5.对不同的δ,AS 算法在多维独立数据集上相对于 Naïve 方法的能耗比如图 3 所示.其中,图 3(a)~图 3(c)分别给出了能耗比随传感器节点数量、查询属性个数和节点通信半径的变化情况.

从图 3(a)可以看出,AS 算法的能耗比随传感器节点数量增多而逐渐减小,这说明传感器网络规模越大,节点数目越多,AS 方法的节能效果越好.这是因为 Naïve 方法需要回传所有感知数据,且节点越多需要回传的数据也越多,进而消耗的能量也越多;而 AS 算法仅根据节点的数据有效率,回传部分有效率高的数据,因此当传感器节点数量增多时,该算法需要回传的数据增加较少,从而能耗比降低.

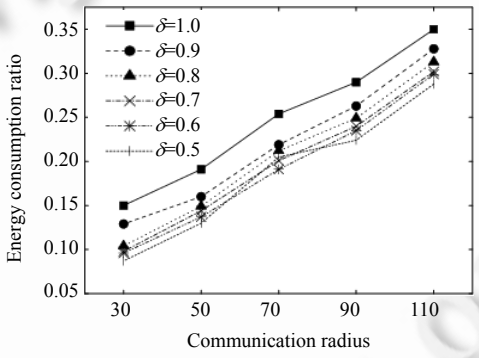
从图 3(b)可以看出,AS 算法的能耗比随查询属性个数的增加而增大.这是因为距离空间维数增加导致满足支配关系的数据相对减少,进而导致 *SK* 中的数据增多,需要回传的数据也因此增多,最终导致 AS 算法能耗增加;但 Naïve 方法不考虑感知数据之间的支配关系,回传所有数据,其数据传输量与查询属性的个数无关,因此当查询属性的个数增加时,AS 算法的能耗比也增加.

从图 3(c)可以看出,通信半径越小,AS 算法的能耗比也越小.这表明,在通信半径有限、感知数据需要多跳转发的大规模传感器网络中,AS 算法能够取得更好的节能效果.由于感知数据在回传 Sink 节点过程中需要被转发的次数随传感器节点通信半径的增大而减小,因此当传感器节点通信半径增大时,网络中的数据通信量减小,而且需要回传的感知数据越多,其数据通信量减少的越显著.由于 Naïve 方法回传所有感知数据,因此当传感器节点通信半径增大时,其数据通信量减少的更多,从而导致 AS 算法的能耗比增大.



(a) Energy consumption ratio vs. number of nodes
(a) 能量消耗比对节点个数

(b) Energy consumption ratio vs. number of attributes
(b) 能量消耗比对属性个数

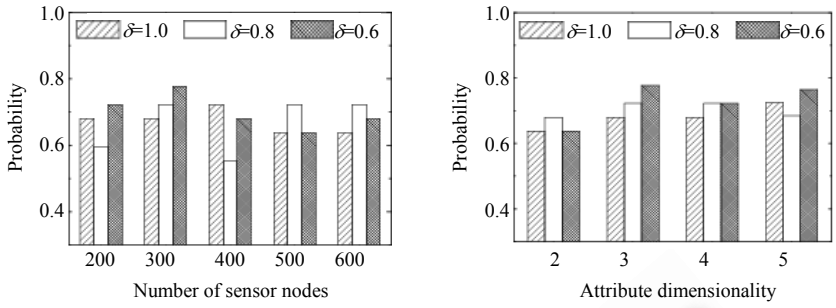


(c) Energy consumption ratio vs. transmission radius
(c) 能量消耗比对传输半径

Fig.3 Energy consumption ratios of AS algorithm with different δ

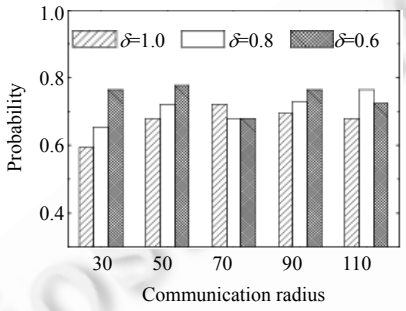
图 3 AS 算法在不同δ下的能量消耗比

AS 算法在多维独立数据集上满足不同查询精度要求的概率如图 4 所示.其中,图 4(a)~图 4(c)分别给出了 ACC_δ随节点数量、查询属性个数和节点通信半径的变化情况.从图中可以看出,ACC_δ随节点数量、查询属性个数、通信半径的变化趋势不明显.这是因为在 AS 算法中,各传感器节点依据数据发送阈值 p^*决定是否回传其感知数据,最终影响 ACC_δ;而 p^*是由历史数据和小规模抽样计算得到的,它受节点数量、查询属性个数、通信半径的影响较小.从图 4 中我们还可以看出,当放松对查询的精度要求时,近似查询的准确率往往会有所提高.这是因为,由数据发送阈值的选择过程可知,δ越小,p^*越大,从而使得回传的数据具有更高的有效率,即回传的数据属于 *SK* 的可能性更大,因此在这些数据上计算出的 ASK_δ满足精度要求的概率也随之增大.



(a) ACC_δ vs. number of nodes (b) ACC_δ vs. number of attributes

(a) ACC_δ 对节点个数 (b) ACC_δ 对属性个数



(c) ACC_δ vs. transmission radius

(c) ACC_δ对传输半径

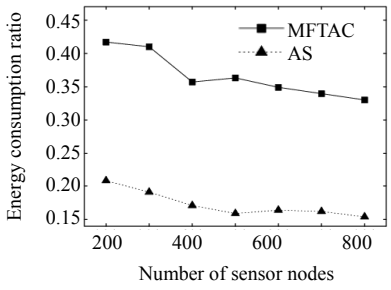
Fig.4 ACC_δ of AS algorithm with different δ

图 4 AS 算法在不同δ下的 ACC_δ

4.3 与MFTAC算法的比较

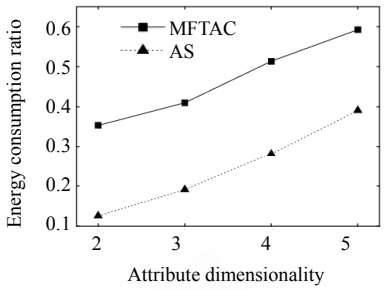
图 5 给出了 AS 算法和 MFTAC 算法在多维独立数据集上处理 Skyline 查询时的实验结果.从中我们可以看出,AS 算法的能耗要远远小于 MFTAC 算法的能耗.这说明在处理 Skyline 查询时,AS 算法会节省更多的网络能量.其原因是,MFTAC 算法在处理 Skyline 查询时需要每个节点的感知数据都与其他节点的感知数据进行比较,从而增加了网络中的数据通信量,消耗了更多能量;而 AS 算法不需要感知数据之间的网内比较,从而节省了能源.尽管 AS 算法并不能保证总是得到全部的 Skyline 查询结果,但是在能源极其宝贵的无线传感器网络中,当用户不严格要求得到全部的 Skyline 查询结果时,AS 算法可以有效满足用户需求并节省大量能源.

从图 5(a)中我们可以看出,随着节点数量的增多,相对于 MFTAC 算法,AS 算法的能耗比下降得更快.这说明传感器网络的规模越大,AS 算法相对于 MFTAC 算法的节省效果越好.从图 5(b)可以看出,随着查询属性个数的增加,AS 算法与 MFTAC 算法的能耗比差距逐渐减小.这是因为查询属性个数的增加会使满足支配关系的感知数据减少,从而导致 *SK* 中的数据急剧增加,因此需要回传的数据也迅速增多,AS 算法的节能效果下降.



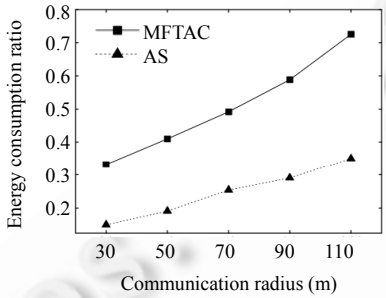
(a) Energy consumption ratio vs. number of nodes

(a) 能量消耗比对节点个数



(b) Energy consumption ratio vs. number of attributes

(b) 能量消耗比对属性个数



(c) Energy consumption ratio vs. transmission radius

(c) 能量消耗比对传输半径

Fig.5 Energy consumption ratio of AS algorithm against that of MFTAC algorithm

图 5 AS 算法与 MFTAC 算法关于能量消耗比的比较

5 结 论

Skyline 查询作为一类复杂查询,在无线传感器网络中具有广泛应用.由于在多数情况下,部分 Skyline 查询结果即可满足用户需求,为了最小化传感器网络的能量消耗,本文提出了一种近似 Skyline 查询处理算法.该算法根据用户需求,仅使部分传感器节点回传其感知数据即可计算得到 Skyline 查询的一个近似结果集,从而极大地降低无线传感器网络中的通信开销,节省网络能量.我们在模拟环境下测试了本文算法,并与精确的 Skyline 查询处理算法进行了比较.实验结果表明,本文的算法可以根据用户的应用需求,高效节能地处理传感器网络中的近似 Skyline 查询.

References:

[1] Akyildiz IF, Su WL, Sankarasubramaniam Y, Cayirci E. A survey on sensor networks. IEEE Communications Magazine, 2002, 40(8):102−114. [doi: 10.1109/MCOM.2002.1024422]

[2] Cullar DE, Estrin D, Strvastava M. Overview of sensor networks. IEEE Computer, 2004,37(8):41−49.

[3] Fung WF, Sun D, Gehrke J. Cougar: The network is the database. In: Franklin MJ, Moon B, Ailamaki A, eds. Proc. of the 2002 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2002. 621.

[4] Madden S, Franklin M, Hellerstein J, Hong W. The design of an acquisitional query processor for sensor networks. In: Halevy AY, Ives ZG, Doan AH, eds. Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2003. 491−502.

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

[6] Manjhi A, Nath S, Gibbons PB. Tributaries and deltas: Efficient and robust aggregation in sensor network streams. In: Özcan F, ed. Proc. of the 2005 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2005. 287−298.

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

[8] Silberstein A, Munagala K, Yang J. Energy-Efficient monitoring of extreme values in sensor networks. In: Chaudhuri S, Hristidis V, Polyzotis N, eds. Proc. of the 2006 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2006. 169−180.

[9] Abadi DJ, Madde S, Lindner W. Reed: Robust, efficient filtering and event detection in sensor networks. In: Böhm K, Jensen CS, Haas LM, Kersten ML, Larson PÅ, Ooi BC, eds. Proc. of the 31st Int'l Conf. on Very Large Data Bases. New York: ACM Press, 2005. 769−780.

[10] Yang X, Lim HB, Ozsu MT, Tan KL. In-Network execution of monitoring queries in sensor networks. 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. 521−532.

[11] Deshpande A, Guestrin C, Madden S, Hellerstein JM, Hong W. Model-Driven data acquisition in sensor networks. In: Nascimento MA, Özsu MT, Kossmann D, Miller RJ, Blakeley JA, Schiefer KB, eds. Proc. of the 30th Int'l Conf. on Very Large Data Bases. New York: ACM Press, 2004. 588−599.

[12] Deshpande A, Guestrin C, Hong W, Madden S. Exploiting correlated attributes in acquisitional query processing. In: Proc. of the 21st Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2005. 143−154.

[13] Chu D, Deshpand A, Hellerstein JM, Hong W. Approximate data collection in sensor networks using probabilistic models. In: Liu L, Reuter A, Whang KY, Zhang JJ, eds. Proc. of the 22nd Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2006.

[14] Silberstein A, Braynard R, Ellis CS, Munagala K, Yang J. A sampling-based approach to optimizing top-*k* queries in sensor networks. In: Liu L, Reuter A, Whang KY, Zhang JJ, eds. Proc. of the 22nd Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2006.

[15] Borzsonyi S, Kossmann D, Stocker K. The Skyline operator. In: Proc. of the 22nd Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2001. 421−430.

[16] Chomicki J, Godfrey P, Gryz J, Liang D. Skyline with presorting. In: Dayal U, Ramamritham K, Vijayaraman TM, eds. Proc. of the 19th Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2003. 717−816.

[17] Tan KL, Eng PK, Ooi BC. Efficient progressive Skyline computation. In: Apers PMG, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass RT, eds. Proc. of the 27th Int'l Conf. on Very Large Data Bases. New York: ACM Press, 2001. 301−310.

[18] Kossmann D, Ramsak F, Rost S. Shooting stars in the sky: A online algorithm for Skyline queries. In: Kaufmann M, ed. Proc. of the 28th Int'l Conf. on Very Large Data Bases. New York: ACM Press, 2002. 275−286.

[19] Papadias D, Tao Y, Fu G, Seeger B. An optimal and progressive algorithm for Skyline queries. In: Halevy AY, Ives ZG, Doan AH, eds. Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2003. 467−478.

[20] Balke WT, Guntzer U, Zheng JX. Efficient distributed skylining for Web information systems. In: Bertino E, Christodoulakis S, Plexousakis D, Christophides V, Koubarakis M, Böhm K, Ferrari E, eds. Proc. of the 9th Int'l Conf. on Extending Database Technology. Berlin: Springer-Verlag, 2004. 256−273.

[21] Lo E, Yip K, Lin KI, Cheung DW. Progressive skylining over Web-accessible database. Data and Knowledge Engineering, 2006, 57(2):122−147. [doi: 10.1016/j.datak.2005.04.003]

[22] Wu P, Zhang CJ, Feng Y, Zhao BY, Agrawal D, Abbadi AE. Parallelizing Skyline queries for scalable distribution. In: Ioannidis YE, Scholl MH, Schmidt JW, Matthes F, Hatzopoulos M, Böhm K, Kemper A, Grust T, Böhm C, eds. Proc. of the 10th Int'l Conf. on Extending Database Technology. Berlin: Springer-Verlag, 2006. 112−130.

[23] Wang SY, Ooi BC, Tung AKH, Xu LZ. Efficient Skyline query processing on peer-to-peer networks. In: Proc. of the 23rd Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2007. 1126−1135.

[24] Huang ZY, Jensen CS, Lu H, Ooi BC. Skyline queries against mobile lightweight devices in MANETs. In: Liu L, Reuter A, Whang KY, Zhang JJ, eds. Proc. of the 22nd Int'l Conf. on Data Engineering. Washington: IEEE Computer Society Press, 2006. 66.

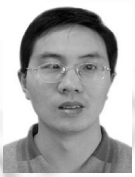
[25] Huang ZY, Lu H, Ooi BC, Tung AKH. Continuous Skyline queries for moving objects. IEEE Trans. on Knowledge and Data Engineering, 2006,18(12):1645−1658. [doi: 10.1109/TKDE.2006.185]

[26] Chen HK, Zhou SG, Guan JH. Towards energy-efficient Skyline monitoring in wireless sensor networks. In: Langendoen K, Voigt T, eds. Proc. of the 4th European Wireless Sensor Networks Conf. Berlin: Springer-Verlag, 2007. 101−116.

[27] Xin JC, Wang GR, Zhang XY. Energy-Efficient Skyline queries over sensor network using mapped Skyline filters. In: Dong GZ, Lin XM, Wang W, Yang Y, Yu JX, eds. Proc. of the Joint 9th Asia-Pacific Web Conf. and 8th Int'l Conf. on Web-Age Information Management. Berlin: Springer-Verlag, 2007. 144−156.

[28] Kwon Y, Choi J, Chung YD, Lee SK. In-Network processing for Skyline queries in sensor networks. IEICE Trans. of Communications, 2007,E90-B(12):3452−3459. [doi: 10.1093/ietcom/e90-b.12.3452]

潘立强(1978－),男,山东掖县人,博士生,主要研究领域为无线传感器网络,数据流.



骆吉洲(1975－),男,博士,副教授,主要研究领域为海量数据库压缩,无线传感器网络.

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