

传感器网络中基于过滤的概率 Skyline 查询算法

信俊昌¹, 石凌旭², 王 培¹, 王之琼¹

(1. 东北大学 信息科学与工程学院, 辽宁 沈阳 110819; 2. 中国人民解放军后勤工程学院, 重庆 400311)

摘 要: 针对感知数据固有的不确定性问题,研究了无线传感器网络中概率 Skyline 查询的处理与优化技术. 首先分析了概率 Skyline 查询的性质,证明了概率 Skyline 查询的不可分解性,因而无法直接利用网内计算方法求解;进而提出了无线传感器网络中基于过滤的概率 Skyline 查询处理算法(filter-based probabilistic Skyline query processing algorithm in WSN, FPSP). FPSP 算法将感知数据划分为候选数据、相关数据和无关数据;只需要候选数据和相关数据即可求得概率 Skyline 查询结果,可以在传感器节点过滤无关数据以避免大量的数据网内传输. 仿真实验结果表明,FPSP 算法可以有效降低传感器节点的数据传输量,极大地延长了无线传感器网络的使用寿命.

关 键 词: 不确定性数据;无线传感器网络;概率 Skyline;查询处理;数据过滤

中图分类号: TP 311.13

文献标志码: A

文章编号: 1005-3026(2014)07-0944-05

Filter-Based Probabilistic Skyline Query Processing Algorithm in Wireless Sensor Network

XIN Jun-chang¹, SHI Ling-xu², WANG Pei¹, WANG Zhi-qiong¹

(1. School of Information Science & Engineering, Northeastern University, Shenyang 110819, China; 2. Logistic Engineering University of PLA, Chongqing 400311, China. Corresponding author: XIN Jun-chang, E-mail: xinjunchang@ise.neu.edu.cn)

Abstract: Due to the inherent uncertainty of sensing data, the processing and optimization techniques for probabilistic Skyline (PS) in wireless sensor networks (WSNs) were investigated. It has been proved that PS was not decomposable after analyzing its properties, so in-network aggregation techniques could not be used directly to improve the performance. Therefore, a filter-based probabilistic Skyline query processing algorithm in WSNs (FPSP) was proposed to evaluate the PS query in WSNs. The sensing data were divided into candidate data (CD), relevant data (RD), and irrelevant data (ID) by the proposed FPSP. The ID in each sensor node could be filtered directly so as to reduce data transmission cost, since PS result could be correctly obtained only according to CD and RD on the base station. The experimental results showed that most of the unnecessary data can be effectively filtered and the lifetime of WSNs can be greatly prolonged by the proposed FPSP algorithm.

Key words: uncertain data; wireless sensor network (WSN); probabilistic Skyline; query processing; data filtering

随着微电子技术、通信技术和嵌入式技术的飞速发展,无线传感器网络(wireless sensor network, WSN)以其巨大的商业应用前景和学术研究价值,得到了工业界和学术界的共同关注^[1]. 由于受到硬件设备、传感技术、通信质量和周围环境等多种因素的共同影响,传感器节点采集到的

感知数据往往是不精确或可信度很低的. 可以说,感知数据本质上是不确定数据,不确定性是感知数据的固有属性.

无论是传感器网络中的 Skyline 查询处理^[2-6],还是面向不确定数据的概率 Skyline 查询处理^[7-10]都取得了大量的研究成果. 其中,Chen

等^[2]研究了传感器网络中的连续 Skyline 查询. 文献[3]研究了传感器网络中的滑动窗口 Skyline. 文献[4]研究了传感器网络中的多 Skyline 查询优化问题,提出了基站与节点两阶段优化算法. Su 等^[5]提出了基于聚簇型路由结构的、以数据为中心的 Skyline 查询算法,通过不同节点发出的多 Skyline 查询共享同一感知数据搜集过程来降低查询的能量消耗. 文献[6]提出了基于过滤的 Skyline 节点连续查询算法,通过在传感器节点设置过滤器来避免不必要的数据传输. Pei 等^[7]提出了自下而上和自上而下两种方法,通过不断地定界、剪枝、精化的迭代过程来完成概率 Skyline 查询. 文献[8-9]研究了不确定数据流的概率 Skyline 查询. 文献[10]研究了主从结构分布式环境中的概率 Skyline 查询.

但无线传感器网络中的 Skyline 查询处理算法未考虑数据的不确定性,而概率 Skyline 查询处理算法未考虑传感器网络的分布式环境,因而都无法直接应用于无线传感器网络中的概率 Skyline 查询. 本文在分析概率 Skyline 查询性质的基础上,提出了无线传感器网络中基于过滤的概率 Skyline 查询处理算法 (filter-based probabilistic Skyline query processing algorithm in WSN, FPSP), 并通过实验验证了算法的有效性.

1 基于过滤的概率 Skyline 查询算法

1.1 问题描述

定义 1 Skyline 给定一个元组集合 D , 其中不被其他元组所支配的元组构成了 D 的 Skyline. 元组 t_i 支配元组 t_j 当且仅当元组 t_i 在任何维度都不比 t_j 差, 并且在至少一维 l 比 t_j 好.

定义 2 Skyline 概率 给定一个不确定元组集合 U , 由 U 中不确定元组所构成的可能世界的集合为 $X = \{W_1, W_2, \dots, W_n\}$. 设 U 中的不确定元组 t 及可能世界子集 $X' \subseteq X$ 满足条件:

- 1) 对任一可能世界 $W \in X'$, 不确定元组 t 属于 W 的 Skyline, 即 $t \in \text{Skyline}(W)$;
- 2) 对任一可能世界 $W \in X - X'$, 不确定元组 t 不属于 W 的 Skyline, 即 $t \notin \text{Skyline}(W)$.

那么, 不确定元组 t 的 Skyline 概率为可能世界子集 X' 中的所有可能世界的存在概率之和, 即

$$P_{\text{Sky}}(t) = \sum_{W \in X'} P(W).$$

显然, 如果不确定元组集合 U 中的不确定元组之间相互独立, 那么不确定元组 t 的 Skyline 概率等于不确定元组 t 的存在概率 $P(t)$ 和支配 t 的

每个不确定元组 t' 均不存在的概率 $\prod_{t' > t} (1 - P(t'))$ 的乘积, 即 $P_{\text{Sky}} = P(t) \times \prod_{t' > t} (1 - P(t'))$.

定义 3 概率 Skyline 给定一个不确定元组集合 U 和一个阈值 p , 则 U 中 Skyline 概率大于 p 的所有不确定元组共同构成了不确定元组集合 U 的概率 Skyline, 即 $\text{PS}(U) = \{t | P_{\text{Sky}}(t) > p\}$.

1.2 性质分析

定理 1 概率 Skyline 查询不可分解.

证明: 假设不确定元组集合 $U = \{t_1, t_2, t_3, t_4\}$, 如图 1 所示. 根据定义 2 可得, $P_{\text{Sky}}(t_1) = 0.4$, $P_{\text{Sky}}(t_2) = 0.56$, $P_{\text{Sky}}(t_3) = 0.42$, $P_{\text{Sky}}(t_4) = 0.3$. 根据定义 3 可得 $\text{PS}(U) = \{t_2\}$.

令 $U = U_1 \cup U_2$, 且 $U_1 = \{t_1, t_4\}$, $U_2 = \{t_2, t_3\}$. 同理可得, $\text{PS}(U_1) = \emptyset$, $\text{PS}(U_2) = \{t_2, t_3\}$.

仅根据 $\text{PS}(U_1) \cup \text{PS}(U_2) = \{t_2, t_3\}$, 无论用何种方式, 都无法知道 U 的概率 Skyline 为 $\{t_2\}$. 也就是, $\text{PS}(U) \neq g(\text{PS}(U_1) \cup \text{PS}(U_2))$. 因此, 概率 Skyline 查询不可分解.

证毕.

根据定理 1 可知, 概率 Skyline 不可分解, 也就不能直接利用网内计算技术^[3-4]来提高无线传感器网络中概率 Skyline 查询的处理效率.

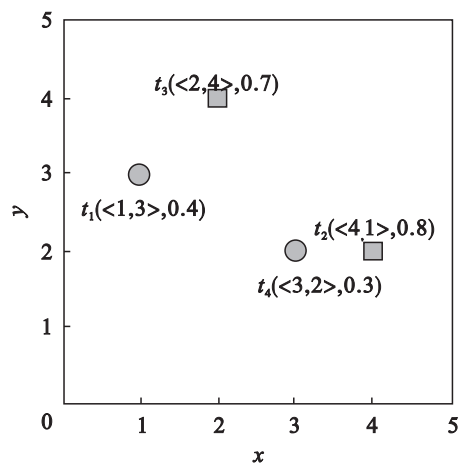


图1 概率 Skyline 不可分解示例

Fig. 1 Example of PS query is not decomposable

定理 2 给定一个不确定元组集合 U . 如果元组 $t_j \in U_i \subseteq U$ 满足: $P(t_j) \times \prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) < p$, 则元组 t_j 不属于 U 的概率 Skyline.

证明: 根据定义 2 和 3 直接得证.

证毕.

定理 3 给定一个不确定元组集合 U . 如果元组 $t_j \in U_i \subseteq U$ 满足: $\prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) < p$, 则删除元组 t_j 不会影响 U 的概率 Skyline 的计算.

证明:因为 $\prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) < p$ 且 $P(t_j) \leq 1$, 可得 $P_{\text{Sky}}(t_j) = P(t_j) \times \prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) < p$. 因此 $t_j \notin \text{PS}(U_i)$, 且 $t_j \notin \text{PS}(U)$.

删除元组 t_j , 只影响被 t_j 支配的元组的 Skyline 概率. 假设 $t_j > t_i$, 由于所有支配 t_j 的元组一定也支配 t_i , 故 $\prod_{t_k \in U_i - \{t_j\}, t_k > t_i} (1 - P(t_k)) < p$, 不会将 t_i 误判为概率 Skyline 元组.

证毕.

定理 2 明确指出了元组子集 U_i 中的哪些元组一定不属于 U 的概率 Skyline, 也即指出了哪些元组可能成为 U 的概率 Skyline 元组; 定理 3 指出了元组子集 U_i 中的哪些元组不会影响 U 的概率 Skyline 的计算, 可以直接删除. 并非所有不属于 $\text{PS}(U)$ 的元组都可以删除, 满足 $P(t_j) \times \prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) < p$ 且 $\prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) \geq p$ 的元组会影响其他元组的 Skyline 概率计算, 因而需要保留.

1.3 算法描述

根据定理 2 和 3, 无线传感器网络中的感知数据可以划分为候选数据、相关数据和无关数据三类.

定义 4 候选数据 在传感器节点的感知数据子集 $U_i \subseteq U$ 中, 满足 $P(t_j) \times \prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) \geq p$ 的元组 t_j 称为概率 Skyline 查询的候选数据 (candidate data, CD).

定义 5 无关数据 在传感器节点的感知数据子集 $U_i \subseteq U$ 中, 满足 $\prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) < p$ 的元组 t_j 称为概率 Skyline 查询的无关数据 (irrelevant data, ID).

定义 6 相关数据 在传感器节点的感知数据子集 $U_i \subseteq U$ 中, 满足 $P(t_j) \times \prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) < p$ 且 $\prod_{t_k \in U_i, t_k > t_j} (1 - P(t_k)) \geq p$ 的元组 t_j 称为概率 Skyline 查询的相关数据 (relevant data, RD).

传感器节点汇总数据、分类数据和删除无关数据的具体过程如算法 1 所示. 首先, 传感器节点合并所有子节点发送的数据, 将候选数据合并到候选数据集, 相关数据合并到相关数据集 (第 1 ~ 4 行); 其次, 将本地感知数据加入候选数据集 (第 5 行); 再次, 计算相关数据集内每个元组的 Skyline 概率, 并删除其中的无关数据 (第 6 ~ 16 行); 接着, 计算候选数据集内每个元组的 Skyline 概率, 并删除无关数据, 同时将新的相关数据从候

选数据集移动到相关数据集 (第 17 ~ 31 行). 最后, 将包含候选数据和相关数据的局部结果提交父节点 (第 32 行).

算法 1 传感器节点的查询处理

输入: 子节点消息集合 M , 本地感知数据集 R , 查询阈值 p .

输出: 提交父节点的数据集合 S .

算法描述:

- 1) for each element m in M do
 //合并子节点数据
- 2) tempCD = tempCD + m . CD;
- 3) tempRD = tempRD + m . RD;
- 4) end for
- 5) tempCD = tempCD + R ;
 //加入本地感知数据
- 6) for each element t in tempRD do
- 7) $P = 1$; //初始化累积概率
- 8) $D = t$. getDominating (tempCD + tempRD);
 //发现 t 的支配元组
- 9) for each element d in D do
- 10) $P = P \times (1 - P(t))$; //计算累积概率
- 11) if $P < p$ then
- 12) tempRD. Delete (t); //丢弃无关元组
- 13) break;
- 14) end if
- 15) end for
- 16) end for
- 17) for each element t in tempCD do
- 18) $P = 1$; //初始化累积概率
- 19) $D = t$. getDominating (tempCD + tempRD);
 //发现 t 的支配元组
- 20) for each element d in D do
- 21) $P = P \times (1 - P(t))$; //计算累积概率
- 22) if $P < p$ then
- 23) tempCD. Delete (t); //丢弃无关元组
- 24) break;
- 25) end if
- 26) end for
- 27) if $P(t) \times P < p$ then //候选移动到相关
- 28) tempRD. Add (t);
- 29) tempCD. Delete (t);
- 30) end if
- 31) end for
- 32) return $S = \langle \text{tempCD}, \text{tempRD} \rangle$

基站汇总数据、删除无关数据获取最终概率 Skyline 结果的具体过程如算法 2 所示. 首先,

基站合并所有子节点发送的数据,将候选数据合并到候选数据集,相关数据合并到相关数据集(第 1~4 行);其次,计算候选数据集内每个元组的 Skyline 概率,并删除不属于最终概率 Skyline 结果的无关数据(第 6~15 行);最后,候选数据集中剩余的数据即为概率 Skyline 结果(第 16 行)。

算法 2 基站的查询处理

输入:子节点消息集合 M ,查询阈值 p

输出:提交父节点的数据集合 S

算法描述:

```
1)  for each element  $m$  in  $M$  do
                                //合并子节点数据
2)      tempCD = tempCD +  $m$ . CD;
3)      tempRD = tempRD +  $m$ . RD;
4)  end for
5)  for each element  $t$  in tempCD do
6)       $P = 1$ ;                //初始化累积概率
7)       $D = t$ . getDominating( tempCD + tempRD );
                                //发现  $t$  的支配元组
8)      for each element  $d$  in  $D$  do
9)           $P = P \times (1 - P(t))$ ;    //计算累积概率
10)         if  $P(t) \times P < p$  then
11)             tempCD. Delete( $t$ );    //非查询结果
12)             break;
13)         end if
14)     end for
15) end for
16) return  $S = \text{tempCD}$ 
```

2 实验结果及分析

实验中,随机地在面积为 n 个平方单位的区域内产生 n 个传感器节点,使得每个节点所占的平均面积为 1 个平方单位,将节点之间的通信半径设置为 $2\sqrt{2}$ 个单位长度,规定节点可以发送的最大数据包为 48 字节。所有实验均在一台具有 Inter® Core™ i7 - 2600 处理器、8.00 GB 内存、500 GB 硬盘的 PC 机上完成。

实验中使用的仿真数据为 Skyline 查询的标准测试数据集,并为每个元组按照均匀分布分配概率。因为独立分布与正相关分布的性质比较接近,实验中主要考察算法在独立分布和反相关分布下的算法性能。实验中主要考察 FPSP 算法和集中式算法(centralized algorithm, CA)的通信代价。

图 2 给出了传感器节点数量变化对算法性能的影响。通信代价随着节点数量的增加而增加,FPSP 算法通信代价的增加速度相对缓慢。节点数量的增加意味着感知元组数量的增加,所以 CA 算法的通信代价随着节点数量增加而呈线性增加。FPSP 算法过滤了无关数据,因而通信代价显著降低且增长速度低于 CA 算法。又因为反相关分布下查询结果数量略高于独立分布,所以 FPSP 算法在反相关分布下的通信代价高于独立分布下的通信代价。

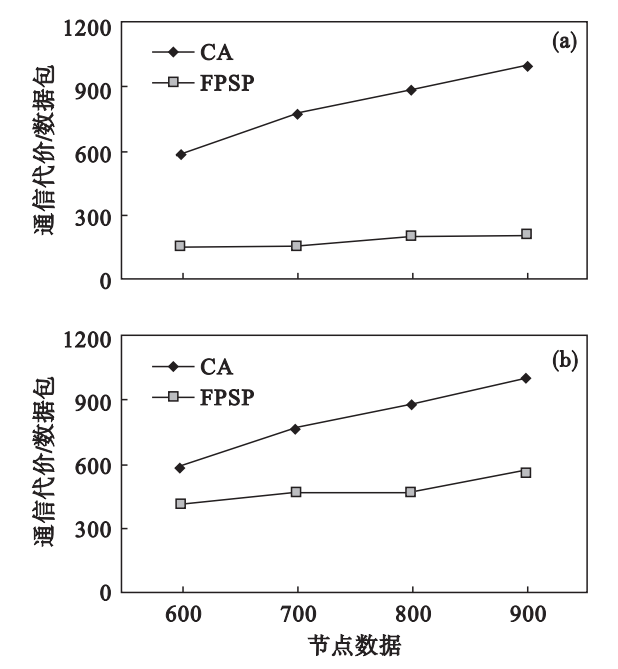


图 2 传感器节点数量和通信代价

Fig. 2 Number of nodes vs. communication cost

(a) — 独立分布; (b) — 反相关分布。

图 3 给出了数据维度变化对算法性能的影响。随着感知数据维度的增加,算法的通信代价也相应增加。原因在于数据维度的增加使得元组被支配的概率降低,导致了概率 Skyline 元组数量的增加,也就直接导致了通信代价的增加。FPSP 算法的通信代价始终低于 CA 算法的通信代价,进一步验证了 FPSP 算法的有效性。同样地,FPSP 算法在反相关分布下的通信代价高于独立分布下的通信代价。

图 4 给出了概率 Skyline 阈值变化对算法性能的影响。随着概率 Skyline 阈值的增加,算法的通信代价随之降低。原因在于概率 Skyline 阈值增加,使得概率 Skyline 查询结果的数量降低,也就直接导致了通信代价的降低。同样地,FPSP 算法的通信代价始终低于 CA 算法的通信代价,FPSP 算法在反相关分布下的通信代价高于独立分布下的通信代价。

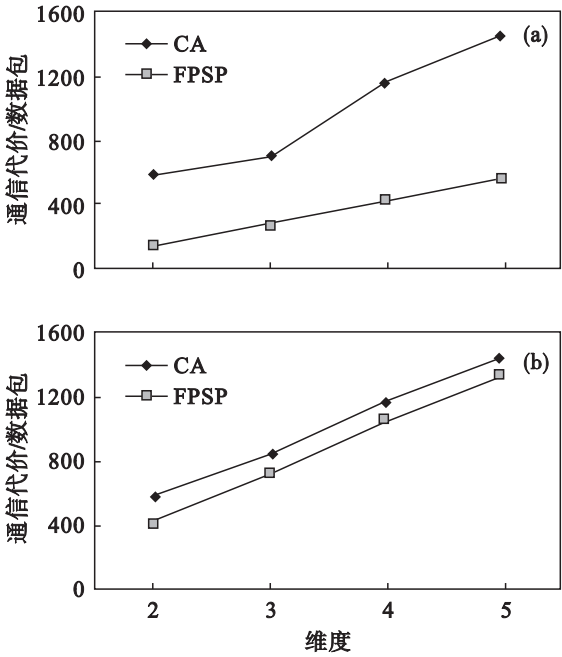


图 3 感知数据维度和通信代价
Fig. 3 Dimensionality vs. communication cost
(a) —独立分布; (b) —反相关分布.

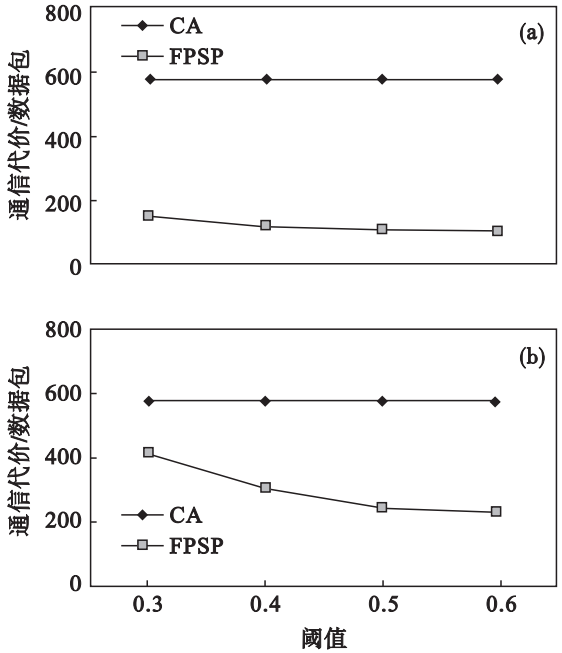


图 4 概率 Skyline 阈值和通信代价
Fig. 4 PS threshold vs. communication cost
(a) —独立分布; (b) —反相关分布.

3 结 论

本文深入分析了概率 Skyline 查询的基本性质,提出了无线传感器网络中基于过滤的概率 Skyline 查询处理算法 FPSP. FPSP 算法通过对传感器节点的感知数据进行分类,提前舍弃那些不影响概率 Skyline 查询结果的无关数据,从而

降低了无线传感器网络中的数据传输总量. 仿真实验结果表明,FPSP 算法在节约网络通信代价方面与集中式算法 CA 的性能相比有了非常显著的提高.

参考文献:

[1] Yick J, Mukherjee B, Ghosal D. Wireless sensor network survey[J]. *Computer Networks*,2008,52(12) :2292 - 2330.

[2] Chen H G,Zhou S G, Guan J H. Towards energy-efficient Skyline monitoring in wireless sensor networks [C]// *Proceedings of the 4th European Conference on Wireless Sensor Networks*. Heidelberg: Springer-Verlag, 2007: 101 - 116.

[3] Xin J C,Wang G R,Chen L, *et al*. Continuously maintaining sliding window skylines in a sensor network [C]// *Proceedings of the 12th International Conference on Database Systems for Advanced Applications*. Heidelberg: Springer-Verlag,2007:509 - 521.

[4] Xin J C, Wang G R, 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-Verlag,2009:247 - 262.

[5] Su I F,Chung Y C, Lee C, *et al*. Efficient Skyline query processing in wireless sensor networks [J]. *Journal of Parallel and Distributed Computing*, 2010, 70 (6) : 680 - 698.

[6] 信俊昌,王国仁. 无线传感器网络中 Skyline 节点连续查询算法[J]. *计算机学报*,2012,35(11) :2415 - 2430. (Xin Jun-chang, Wang Guo-ren. Continuous Skyline nodes query processing over wireless sensor networks[J]. *Chinese Journal of Computers*,2012,35(11) :2415 - 2430.)

[7] Pei J, Jiang B, Lin X M, *et al*. Probabilistic Skylines on uncertain data [C]//*Proceedings of the 33rd International Conference on Very Large Databases*. New York: ACM, 2007:15 - 26.

[8] Ding X F,Lian X,Chen L, *et al*. Continuous monitoring of Skylines over uncertain data streams [J]. *Information Sciences*,2012,184(1) :196 - 214.

[9] Zhang W J,Lin X M, Zhang Y, *et al*. Probabilistic Skyline operator over sliding windows[C]//*Proceedings of the IEEE 25th International Conference on Data Engineering*. Washington D C: IEEE Computer Society, 2009: 1060 - 1071.

[10] Ding X F, Jin H. Efficient and progressive algorithms for distributed Skyline queries over uncertain data [J]. *IEEE Transactions on Knowledge and Data Engineering*,2012,24 (8) :1448 - 1462.