一种 ρ-支配轮廓查询的高效处理算法

信俊昌[1)] 白 梅[1)] 东 韩[2)] 王国仁[1)]
[1)](东北大学信息科学与工程学院 沈阳 110819)
[2)](国家海洋信息中心 天津 300171)

摘 要 近年来,作为重要的多目标决策手段的轮廓查询逐渐得到学术界的重视,相继提出了基于不同支配关系的多种轮廓变体查询.首先,通过对实际应用需求进行分析,提出了基于元组对应数值间比例值大小的 ρ-支配关系的定义,进而提出了 ρ-支配轮廓查询的概念.其次,对 ρ-支配轮廓的基本性质进行了细致而深入的分析,在此基础上,提出了基于分支定界的 ρ-支配轮廓查询算法(Branch and Bound ρ-Dominant Skyline Algorithm,BBDS),避免了对 R-树索引的多次访问,从而提高了 ρ-支配轮廓查询的执行效率.最后,通过大量的仿真实验对 ρ-支配轮廓查询的语义进行分析,并对 BBDS 算法的性能进行验证.实验结果表明,ρ-支配轮廓查询是轮廓查询语义的扩展和补充,而提出的 BBDS 算法则是求解 ρ-支配轮廓查询的高效算法.

关键词 轮廓查询;轮廓变体;ρ-支配关系;ρ-支配轮廓;分支定界
中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2011.01876

An Efficient Processing Algorithm for ρ-Dominant Skyline Query

XIN Jun-Chang[1)] BAI Mei[1)] DONG Han[2)] WANG Guo-Ren[1)]
[1)](*College of Information Science and Engineering, Northeastern University, Shenyang* 110819)
[2)](*National Marine Data and Information Service, Tianjin* 300171)

Abstract In recent years, as an important operator for multi-decision making, skyline query has attracted much attention from the academia gradually, and a variety of skyline variants based on different dominance relationships have been proposed successively by the database researchers. In this paper, firstly, through making the analysis of practical applications' requirements, the ρ-dominance relationship based on the ratio of corresponding values between the tuples is defined, and then the concept of ρ-dominant skyline query based on the ρ-dominance relationship is proposed. Next, by making a detailed and in-depth analysis of ρ-dominance's basic properties, a novel algorithm, named Branch and Bound ρ-Dominant Skyline Algorithm (BBDS), is developed. The BBDS algorithm avoids visiting R-tree index too many times, which can improve the ρ-dominant skyline query implementation efficiency greatly. Finally, through a large number of simulation experiments, the semantic of the ρ-dominant skyline query is analyzed, and meanwhile the performance of BBDS algorithm is verified by the simulation experiments. The simulation experimental results show that ρ-dominant skyline query based on ρ-dominance relationship is a new extension and complement of the traditional skyline query semantic and the BBDS algorithm proposed in this paper is proved to be a highly effective algorithm for solving ρ-dominant skyline queries.

Keywords skyline query; skyline variants; ρ-dominance; ρ-dominant skyline; branch and bound

收稿日期:2011-07-18;最终修改稿收到日期:2011-08-17.本课题得到国家自然科学基金重点项目(60933001)、国家杰出青年科学基金项目(61025007)、国家自然科学基金面上项目(61073063)和中央高校基本科研业务费专项资金(N090304007)资助.**信俊昌**,男,1977 年生,博士,讲师,主要研究方向为感知数据和不确定数据. E-mail: xinjunchang@ise.neu.edu.cn; xinjunchang@gmail.com. **白 梅**,女,1986 年生,博士研究生,主要研究方向为感知数据和不确定数据. **东 韩**,男,1981 年生,工程师,主要研究方向为高性能计算、并行计算、互联网服务和云计算技术. **王国仁**,男,1966 年生,教授,博士生导师,主要研究领域为 XML 数据管理、生物信息学、分布与并行数据库、多媒体索引技术图论、并行计算等.

1 引　言

作为多目标决策与数据挖掘的重要手段之一，轮廓查询[1-2]在许多实际应用中都发挥着非常重要的作用.在不同的应用场合中，“好坏”的评价标准往往也有所不同，在传统轮廓查询的基础上派生出了大量的轮廓变体查询，如多目标轮廓[3]、k-支配轮廓[4]、Top-k 频率轮廓[5]和 Top-k 轮廓[6]等.这些轮廓变体查询极大地丰富了轮廓查询的语义，扩大了轮廓查询的应用范围.

已有的轮廓变体查询中的支配关系都依赖于元组对应数值之间的大小关系，均没有考虑数值间的比例关系.然而，在一些实际应用中，数值间的大小关系不能完全说明问题，这时就需要用数值间的比例作为“好坏”关系的评价.例如，想选取一个温度和湿度都较低的地方做仓库，A 地点温度 20、湿度 30，B 地点温度 18、湿度 32，传统的支配不能区分地点 A 和 B 哪个更适合做仓库，而考虑数值大小关系时，发现它们的差值都为 2，同样不能决定 A 和 B 哪个地方更好，这时可以考虑数值的比例关系，发现 B 在两维与 A 的比值均小于 1.1，但是 A 在两维与 B 的比值不均小于 1.1，所以可以判定地点 B 更适合作为仓库.尤其当各维度上数值的数量级不同时，这种用比例关系衡量元组间的“好坏”关系将更能说明问题.

因为 ρ-支配关系与已有的轮廓变体查询中的支配关系有着明显的不同，所以已有的轮廓变体查询算法都无法直接用来计算 ρ-支配轮廓查询.由于 ρ-支配轮廓查询在日常生活中有着非常广泛的应用，因此，亟需设计高效的算法计算 ρ-支配轮廓查询以满足日益增长的实际应用需求.

本文详尽地阐明了 ρ-支配轮廓查询的语义，并深入地分析了 ρ-支配轮廓的性质，进而提出了基于分支定界的 ρ-支配轮廓查询算法 BBDS.归结起来，本文的主要贡献如下：

(1) 提出了 ρ-支配关系以及基于该支配关系的 ρ-支配轮廓查询的概念；

(2) 深入分析了 ρ-支配轮廓查询的性质，提出了基于分支定界的 ρ-支配轮廓查询算法(Branch and Bound ρ-Dominant Skyline，BBDS)；

(3) 设计了详细的性能评价实验，实验结果表明 BBDS 算法可以有效地处理 ρ-支配轮廓查询.

本文第 2 节回顾相关研究工作；第 3 节介绍 ρ-支配轮廓查询的定义；第 4 节分析 ρ-支配轮廓的性质，进而提出基于分支定界的 ρ-支配轮廓查询处理算法 BBDS；第 5 节介绍仿真实验结果，同时对实验结果进行分析；第 6 节对全文进行总结.

2 相关工作

在传统轮廓查询的基础上，多种轮廓变体查询被相继提出.Balke 和 Guntzer[3]研究了基于多目标函数的多目标检索问题，提出了支持 Top-k 查询和轮廓查询的一般形式的多目标检索算法.Lee 等人[6]研究了依据用户特定偏好来进行动态查找的个性化轮廓排序查询，提出了访问压缩存储结构的新颖算法降低存储负载.Sharifzadeh 等人[7]提出了基于多个查询点距离的空间轮廓查询，利用该查询的几何性质提高空间轮廓查询的计算效率.

Chan 等人[4]提出了依赖于元组子维度空间上的支配关系的 k-支配轮廓查询，并提出了单遍扫描、两遍扫描和顺序访问三种不同的算法计算 k-支配轮廓.除 k-支配轮廓之外，Chan 等人[5]考虑了每个元组成为轮廓元组的子维度空间数量，提出了 Top-k 频率轮廓查询，同时提出了近似算法快速计算 Top-k 频率轮廓.

Lin 等人[8]提出了总体支配面积最大的 k-最具代表性轮廓查询.针对 k-最具代表性轮廓查询存在的问题，Tao 等人[9]提出了使得非代表性轮廓点与最近代表性轮廓点之间距离最小化的基于距离的代表性轮廓查询.Papadias 等人[10]提出了返回支配元组最多的 k 个元组的 Top-k 支配能力排序轮廓，并提出了基本的求解算法.接着，Yiu 等人[11]提出了多种高效算法计算 Top-k 支配能力排序轮廓.吴俊杰等人[12]扩充 Top-k 支配能力排序轮廓的概念，提出了考虑权重的 Top-k 支配能力排序轮廓.

Xia 等人[13]提出了允许数值差异的 ε-支配关系，进而提出了 ε-轮廓的概念，并设计了多种算法计算 ε-轮廓.

在上述的轮廓变体查询中，文献[3]、文献[6]和文献[7]的研究内容与本文关系不大.k-支配轮廓和 Top-k 频率轮廓，都是通过维度特征来控制结果数目；k-最具代表性轮廓查询和 Top-k 支配能力排序轮廓，都是依赖数据支配面积来控制结果数目；ε-轮廓是通过增减数值来控制结果数目；而本文提出的 ρ-支配轮廓是通过调整数值间的比例关系来控制结果数目.

此外，Levandoski 等人[14]研究了扩展现有数据库管理软件支持轮廓变体查询的框架.Zhang 等

人[15]提出了轮廓变体查询框架将多种轮廓变体查询进行统一.然而,本文提出的 ρ-支配轮廓不具备该框架要求的轮廓变体查询的有理性和传递性等性质,因而无法利用该框架计算 ρ-支配轮廓.

3 问题描述

本文中假设所有的元组在各维度上的取值都不为负,如果出现负值,需要经过预处理变成正值后再进行处理.首先,回顾一下传统支配关系和传统轮廓的相关概念[1],如定义 1 和 2 所示.

定义 1. 元组 t_x 支配元组 t_y(记为 $t_x \succ t_y$)当且仅当在所有维度 t_x 都不比 t_y 差,且在至少一个维度 t_x 比 t_y 好,即 $\forall k, t_x[k] \leq t_y[k]$ 且 $\exists l, t_x[l] < t_y[l]$.

定义 2. 元组集合 D 中所有不被其它元组支配的元组构成了 D 的轮廓,简记为 $Sky(D)$.

如图 1 所示,元组 t_a、t_b 和 t_c 支配元组 t_e,元组 t_c 和 t_d 支配元组 t_f,而折线上的 t_a、t_b、t_c 和 t_d 均不被其它元组所支配,共同构成了图 1 的传统轮廓.

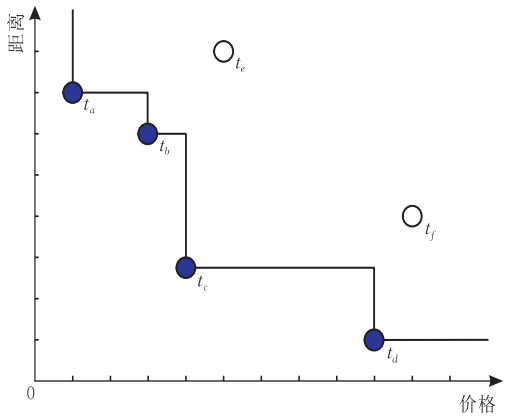


图 1 轮廓举例

接着,介绍 ρ-支配关系和 ρ-支配轮廓的形式化定义,如定义 3 和 4 所示.

定义 3. 元组 t_x ρ-支配元组 t_y(记为 $t_x \succ^{\rho} t_y$)当且仅当在所有维度 t_x 与 t_y 之间的比值都不比 ρ 差,且在至少一维 t_x 与 t_y 之间的比值比 ρ 好.即 $\forall k$, $t_x[k]/t_y[k] \leq \rho$ 且 $\exists l, t_x[l]/t_y[l] < \rho$.

定义 4. 元组集合 D 中所有不被其它元组 ρ-支配的元组构成了 D 的 ρ-支配轮廓,简记为 ρ-$Sky(D)$.

根据定义 1 和 3 可知,当 $\rho=1$ 时,ρ-支配关系等价于支配关系.根据定义 2 和 4 可知,当 $\rho=1$ 时,ρ-$Sky(D)=Sky(D)$.因而可以说,传统轮廓是 ρ-支配轮廓在 $\rho=1$ 时的特例.

图 2 展示了 $\rho=2$ 时的 ρ-支配轮廓示例.图中的方形元组代表了对应的圆形元组与原点之间连线的中点,根据定义 1 和 3 可知,当 $\rho=2$ 时,被方形元组支配的元组一定被原始元组 ρ-支配.图中的阴影部分表明了元组 t_e 的 2-支配区域.因为元组 t_b 在阴影部分内,所以元组 t_e 2-支配 t_b.同理可得,元组 t_a、t_c 和 t_e 2-支配元组 t_b,元组 t_a、t_b、t_c 和 t_d 2-支配元组 t_e,元组 t_a、t_b、t_c 和 t_d 2-支配元组 t_f.因为元组 t_a、t_c 和 t_d 均不被集合中任何其它元组 2-支配,所以它们共同构成了图 2 的 ρ-支配轮廓($\rho=2$).

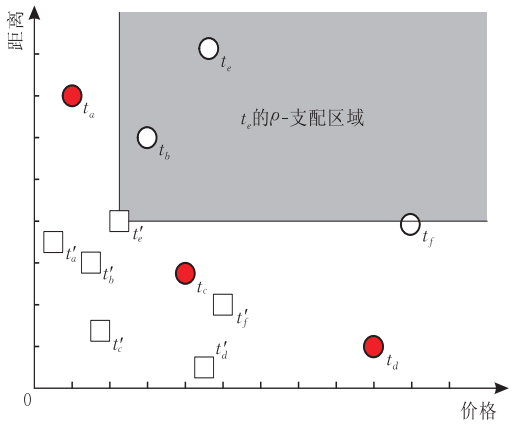


图 2 ρ-支配轮廓举例($\rho=2$)

4 ρ-支配轮廓查询处理

本节中,首先分析 ρ-支配轮廓的基本性质,然后介绍 BBDS 算法的具体处理过程.

4.1 基本性质

当元组间的支配关系满足传递性条件时,相应的轮廓变体查询可以使用基于 R-树的分支定界轮廓算法(Branch and Bound Skyline,BBS)求解[15].幸运地,当 $\rho \leq 1$ 时,ρ-支配关系满足传递性,具体如定理 1 所示.

定理 1. 设 $\rho \leq 1$,如果 3 个元组 t_x、t_y 和 t_z 满足 $t_x \succ^{\rho} t_y$ 且 $t_y \succ^{\rho} t_z$,那么 $t_x \succ^{\rho} t_z$.

证明. 根据定义 3 可知,

$t_x \succ^{\rho} t_y \Rightarrow \forall k, t_x[k]/t_y[k] \leq \rho \wedge \exists l, t_x[l]/t_y[l] < \rho$,

$t_y \succ^{\rho} t_z \Rightarrow \forall k, t_y[k]/t_z[k] \leq \rho \wedge \exists j, t_y[j]/t_z[j] < \rho$.

由于 $t_x[k]/t_y[k] \leq \rho$ 且 $t_y[k]/t_z[k] \leq \rho$,两端分别相乘得 $t_x[k]/t_z[k] \leq \rho^2$.又因为 $\rho \leq 1$,所以 $\rho^2 \leq \rho$.因此,$\forall k, t_x[k]/t_z[k] \leq \rho$.同理可证,$\exists l, t_x[l]/t_z[l] < \rho$.

根据定义 3,可得 $t_x \succ^{\rho} t_z$. 证毕.

根据定理 1,当 $\rho \leq 1$ 时,ρ-支配轮廓可以用 BBS

算法求解.然而,当 $\rho>1$ 时,ρ-支配关系不再具备传递性,需要研究新算法计算 $\rho>1$ 时的 ρ-支配轮廓.

在介绍 $\rho>1$ 时,ρ-支配轮廓查询求解算法之前,先分析 ρ-支配关系和 ρ-支配轮廓的一些重要性质.

引理 1.　设 $\rho>1$,如果两个元组 t_x 和 t_y 满足 $t_x \succ t_y$,那么 $t_x \succ^{\rho} t_y$.

证明.　根据定义 1 可知,

$$t_x \succ t_y \Rightarrow \forall k, t_x[k] \leq t_y[k] \wedge \exists l, t_x[l] < t_y[l].$$

由于 $t_x[k] \leq t_y[k]$,所以 $t_x[k]/t_y[k] \leq 1$.因为 $\rho>1$,所以 $\forall k$, $t_x[k]/t_y[k] \leq \rho$.同理可证,$\exists l, t_x[l]/t_y[l] < \rho$.

根据定义 3,可得 $t_x \succ^{\rho} t_y$.　　证毕.

根据引理 1 可知,当 $\rho>1$ 时,一个元组支配另一个元组意味着它同时也 ρ-支配该元组,所以,ρ-支配轮廓是传统轮廓的子集,具体如定理 2 所示.

定理 2.　设 $\rho>1$,如果集合 D 中的元组 t_x 满足 $t_x \in \rho\text{-}Sky(D)$,那么 $t_x \in Sky(D)$,即 $\rho\text{-}Sky(D) \subseteq Sky(D)$.

证明.　用反证法证明.

假设 $t_x \notin Sky(D)$,根据定义 2 可知,集合 D 中存在元组 t 满足 $t \succ t_x$.又根据引理 1,$t \succ t_x \Rightarrow t \succ^{\rho} t_x$.根据定义 4,$t_x \notin \rho\text{-}Sky(D)$,与已知条件相矛盾.

因此,$\rho\text{-}Sky(D) \subseteq Sky(D)$.　　证毕.

根据定理 2 可知,只有属于传统轮廓的元组才有可能成为 ρ-支配轮廓元组,因而可以先计算元组集合的轮廓 $Sky(D)$,再对 $Sky(D)$ 中的每个元组逐一判断进而得到 ρ-支配轮廓,称为 ρ-支配轮廓的基准算法.

基准算法的具体处理过程如算法 1 所示.首先,初始化元组集合 S 和堆 H(第 1 行);其次,将 R-树中根的所有项加入 H(第 2 行);接着,当 H 不为空时,取出 H 中的首元素 e(第 3～4 行);如果 e 是中间结点,那么判断 e 是否被 S 中的元组支配.若 e 不被 S 中的任何元组支配,则将 e 中所有不被 S 中的任何元组支配的结点加入 H(第 5～9 行);如果 e 是叶结点,那么同样判断 e 是否被 S 中的元组支配.若 e 不被 S 中的任何元组支配,则将 e 加入到集合 S 中(第 10～12 行).经过上述计算过程(第 1～12 行),集合 S 中得到了元组集合 D 中的轮廓.逐一取出集合 S 中的轮廓元组 p,如果 p 被集合 D 中的任何元组 ρ-支配,则将该元组从集合 S 中移除(第 13～15 行);最后,基准算法返回元组集合 D 的 ρ-支配轮廓 S.

算法 1.　基准算法.

输入:参数 ρ 和元组集合 D 的 R-树索引 R

输出:ρ-支配轮廓 $\rho\text{-}Sky(D)$;

```
1. List S=∅; Heap H=∅;
2. H.insertAllEntry(R.root);
                          //将 R-树根中所有项加入 H
3. while(H is not empty)          //遍历 R-树
4.   e=H.removeTopEntry();   //取堆中首个结点
5.   if (e 是中间结点)            //中间结点的处理
6.     if (e.min 不被 S 中的任何元组支配)
7.       for (each child e_i of e)
8.         if (e_i.min 不被 S 中的任何元组支配)
9.           H.insertEntry(e_i); //加入堆排序
10.  else                         //叶结点的处理
11.    if (e 不被 S 中的任何元组支配)
12.      S=S+{e};
13. for (each point p in S)
14.  if (p 被 D 中的元组 ρ-支配) //基于 R-树判断
15.    S=S-{p};
16. Return S.
```

基准算法需要多次访问 R-树以便逐一确认每个轮廓元组是否为 ρ-支配轮廓元组,无疑将消耗大量的计算时间,需要进一步优化.

引理 2.　设 $\rho>1$,如果 3 个元组 t_x、t_y 和 t_z 满足 $t_x \succ t_y$ 且 $t_y \succ^{\rho} t_z$,那么 $t_x \succ^{\rho} t_z$.

证明.　根据定义 1 可知,

$$t_x \succ t_y \Rightarrow \forall k, t_x[k] \leq t_y[k] \wedge \exists l, t_x[l] < t_y[l].$$

根据定义 3 可知,

$$t_y \succ^{\rho} t_z \Rightarrow \forall k, t_y[k]/t_z[k] \leq \rho \wedge \exists j, t_y[j]/t_z[j] < \rho.$$

因为 $t_y[k]/t_z[k] \leq \rho$,所以 $t_y[k] \leq \rho \times t_z[k]$.又因为 $t_x[k] \leq t_y[k]$,所以 $t_x[k] \leq \rho \times t_z[k]$,两端同除以 $t_z[k]$ 得 $t_x[k]/t_z[k] \leq \rho$.同理可证,$\exists l, t_x[l]/t_z[l] < \rho$.

根据定义 3,可得 $t_x \succ^{\rho} t_z$.　　证毕.

根据引理 2 可知,除了能支配元组 t_y 的元组 t_x 之外,所有能被元组 t_y ρ-支配的元组都能被元组 t_x ρ-支配.因此,只需保留最多被一个元组支配的元组,即 2-轮廓带(Skyband)[10] 就可以保证 ρ-支配轮廓结果的正确性,如定理 3 所示.

定义 5.　元组集合 D 中,最多被 $(k-1)$ 个元组支配的元组构成了 D 的 k 层轮廓带,简记为 $Sky^k(D)$.

定理 3.　设 $\rho>1$,$\rho\text{-}Sky(D) = \rho\text{-}Sky(Sky^2(D))$.

证明.

(1)证明 $\rho\text{-}Sky(D) \subseteq \rho\text{-}Sky(Sky^2(D))$.

假设 $t \notin \rho\text{-}Sky(Sky^2(D))$,那么 $t \notin Sky^2(D)$ 或者 $\exists t' \in Sky^2(D)$,$t' \succ^{\rho} t$.如果 $t \notin Sky^2(D)$,那么 $\exists t' \in D$,满足 $t' \succ t$,根据引理 1,$t' \succ^{\rho} t$,所以 $t \notin$

ρ-$Sky(D)$；如果 $\exists t' \in Sky^2(D)$，$t' \succ^{\rho} t$，那么 $t \notin \rho$-$Sky(D)$. 因此，ρ-$Sky(D) \subseteq \rho$-$Sky(Sky^2(D))$.

(2) 证明 ρ-$Sky(Sky^2(D)) \subseteq \rho$-$Sky(D)$.

假设 $t \notin \rho$-$Sky(D)$，那么根据定义 4，$\exists t' \in D$，满足 $t' \succ^{\rho} t$. 如果 $t' \in Sky^2(D)$，那么 $t \notin \rho$-$Sky(Sky^2(D))$；如果 $t' \notin Sky^2(D)$，那么 $\exists t'' \in D$，满足 $t'' \neq t'$ 且 $t'' \succ t'$，根据引理 2，$t'' \succ^{\rho} t$，所以 $t \notin \rho$-$Sky(Sky^2(D))$. 因此，ρ-$Sky(Sky^2(D)) \subseteq \rho$-$Sky(D)$.

综上所述，ρ-$Sky(D) = \rho$-$Sky(Sky^2(D))$. 证毕.

根据定理 3 可知，所有被两个或两个以上元组支配的元组都不影响 ρ-支配轮廓的计算，因此，可以在计算过程中将这样的元组(或者 R-树结点)直接丢弃，以便避免不必要的比较操作，从而提高查询效率.

分析 2-轮廓带中的元组的特征，可以进一步得到下面的定理 4.

定理 4. 设 $\rho > 1$，如果轮廓元组 t 不是 ρ-支配轮廓元组，那么 t 要么被另一个轮廓元组 ρ-支配，要么被它所支配的 2-轮廓带元组 ρ-支配.

证明. 因为元组 t 不是 ρ-支配轮廓元组，所以 $\exists t' \in D$，满足 $t' \succ^{\rho} t$. 那么 t' 可能是轮廓元组，也可能是非轮廓的 2-轮廓带元组，还可能是非 2-轮廓带元组. 下面分情况加以讨论.

(1) 如果 t' 是轮廓元组，直接满足结论；

(2) 如果 t' 是非轮廓的 2-轮廓带元组，根据定义 5 可知，元组 t' 要么被轮廓元组 t 支配，要么被非 t 的其它轮廓元组支配. 如果元组 t' 被轮廓元组 t 支配，直接满足结论. 如果元组 t' 被非 t 的其它轮廓元组支配，那么根据引理 2 可知，支配元组 t' 的轮廓元组 ρ-支配元组 t，满足结论.

(3) 如果 t' 是非 2-轮廓带元组，根据定义 5 可知，至少存在一个非 t 的轮廓元组支配 t'，根据引理 2 可知，该轮廓元组 ρ-支配元组 t，满足结论.

综上所述，如果轮廓元组 t 不是 ρ-支配轮廓元组，那么 t 要么被另一个轮廓元组 ρ-支配，要么被它所支配的 2-轮廓带元组 ρ-支配. 证毕.

根据定理 4 可知，如果一个轮廓元组不是 ρ-支配轮廓元组，只要它不被自己所支配的 2-轮廓带元组 ρ-支配，那么它一定被另一个轮廓元组所 ρ-支配. 因此，非轮廓的 2-轮廓带元组只对支配它的轮廓元组的判断有直接影响，与其它的轮廓元组的判断无关，可以在判定后直接丢弃，只需要保留所有的轮廓元组即可保证 ρ-支配轮廓结果的正确性.

4.2 BBDS 算法

定理 2 确定了哪些元组(轮廓元组)是候选的 ρ-支配轮廓元组，而定理 3 和定理 4，确定了哪些元组对 ρ-支配轮廓的计算没有影响，哪些元组影响较小. 根据前面的分析，可得出基于分支定界的查询算法.

基于分支定界的 ρ-支配轮廓查询算法的具体处理过程如算法 2 所示. 首先，初始化 ρ-支配轮廓元组集合 S、非 ρ-支配轮廓元组的轮廓元组集合 SS 和堆 H(第 1 行)；其次，将 R-树中根的所有项加入 H(第 2 行)；接着，当 H 不为空时，取出 H 中的首元素 e(第 3～4 行)；如果 e 是中间结点，那么判断 e 是否被 $S+SS$ 中的两个元组支配. 若 e 不被 $S+SS$ 中的两个元组支配，则将 e 中所有不被 $S+SS$ 中的两个元组支配的结点加入 H(第 5～9 行)；如果 e 是叶结点(第 10 行)，那么首先判断 e 是否被 $S+SS$ 中的元组支配(第 11 行). 若 e 不被 $S+SS$ 中的任何元组支配，则继续判断它是否被 $S+SS$ 中的任何元组 ρ-支配(第 12 行). 若 e 不被 $S+SS$ 中的任何元组 ρ-支配，则将 e 加入到集合 S 中(第 13 行)，否则，将 e 加入到集合 SS 中(第 14～15 行). 接着，将集合 S 中被 e ρ-支配的元组移动到集合 SS 中(第 16～17 行). 如果 e 只被 $S+SS$ 中的一个元组支配，那么需要用 e 来判定支配它的元组是否为 ρ-支配轮廓元组(第 18～21 行). 最后，BBDS 算法返回元组集合 D 的 ρ-支配轮廓 S.

算法 2. 基于分支定界的 ρ-支配轮廓查询算法.

输入：参数 ρ 和元组集合 D 的 R-树索引 R；

输出：ρ-支配轮廓 ρ-$Sky(D)$；

```
1.  List S=∅; List SS=∅; Heap H=∅;
2.  H.insertAllEntry(R.root);
                                  //将 R-树根中所有项加入 H
3.  while (H is not empty)        //遍历 R-树
4.   e=H.removeTopEntry();       //取堆中首个结点
5.   if (e 是中间结点)            //中间结点的处理
6.     if (e.min 不被(S+SS)中的两个元组支配)
7.       for (each child e_i of e)
8.         if (e_i.min 不被(S+SS)中的两个元组支配)
9.           H.insertEntry(e_i); //加入堆排序
10.  else                         //叶结点的处理
11.    if (e 不被(S+SS)中的任何元组支配)
12.      if (e 不被(S+SS)中的任何元组 ρ-支配)
13.        S=S+{e};              //ρ-支配轮廓点
14.      else
15.        SS=SS+{e};            //非 ρ-支配轮廓点
16.      D'=getRhoDominated(S,e)
17.      S=S-D'; SS=SS+D';
18.    elseif (e 只被(S+SS)中的一个元组支配)
19.      p'=getDominating(S,e)
20.      if (p' 被 e 所ρ-支配)
21.        S=S-{p'}; SS=SS+{p'};
22. Return S.
```

5　实验结果与分析

本节的实验中，使用 Visual C++语言实现了基准算法(BA)和基于分支定界的 ρ-支配轮廓查询算法(BBDS). 实验环境为 Pentium IV 2.4 GHz CPU、512 MB DDR 内存、80 GB 硬盘和 Windows XP 操作系统.

使用轮廓查询的标准测试数据生成器[10]分别生成独立分布、正相关分布和反相关分布的实验数据. 由于独立分布与正相关的性质比较接近，因此本节的实验仅列举了各算法在独立分布和反相关分布下的性能变化规律. 实验中所考察的主要参数的变化范围及默认值如表 1 所示.

表 1　仿真实验参数

参　数	数据维度	数据集大小/K	ρ
默认值	3	100	1.5
变化范围	2,3,4,5	100, 300, 500, 700, 900	0.9, 1, 1.5, 2, 2.5

首先，考查 ρ 值大小对结果集大小的影响. 图 3 中给出了数据集大小为 100 K、维度为 3 维时，ρ 值大小由 0.9 变化到 2.5 时，ρ-支配轮廓大小的变化规律. 从图中可以看出，无论在独立分布还是在反相关分布的数据集中，随着 ρ 值的不断增大，ρ-支配轮廓大小逐渐变小. 这是因为 ρ 值的增加，意味着元组更加容易被 ρ-支配，使得元组成为 ρ-支配轮廓的可能性减少，ρ-支配轮廓的结果数量也随之减少. 在反相关分布下，随着 ρ 值的增加，ρ-支配轮廓大小急剧减少，这是因为反相关分布特征导致的，当 ρ 值增加时，元组间被 ρ-支配的概率大大提高，从而导致了ρ-支配轮廓的结果数量骤减.

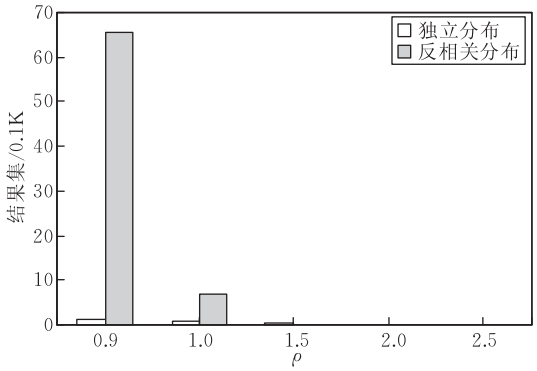


图 3　ρ 值大小对结果集大小的影响

其次，考查 ρ 值大小对各个算法性能的影响. 图 4 给出了数据集大小为 100 K、维度为 3 维时，ρ 值大小由 1.5 变化到 2.5 时各算法性能的变化规律. 从图中可以看出，无论在独立分布还是在反相关分布的数据集中，各算法的响应时间随着 ρ 值增加而影响不大. 在反相关分布下，各算法的响应时间都大于独立分布下算法的响应时间. 无论 ρ 值如何变化，在同样的数据分布下，BBDS 算法的响应时间都要小于 BA 算法的响应时间.

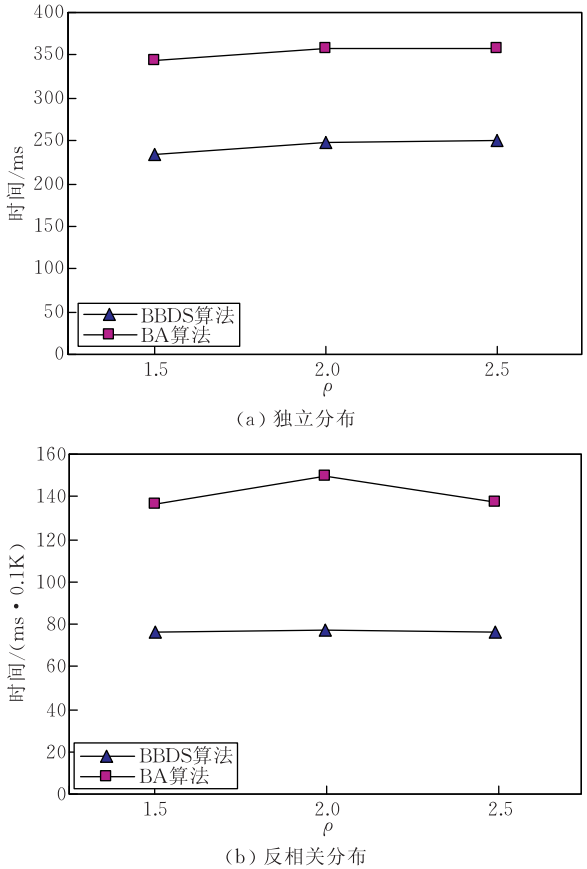
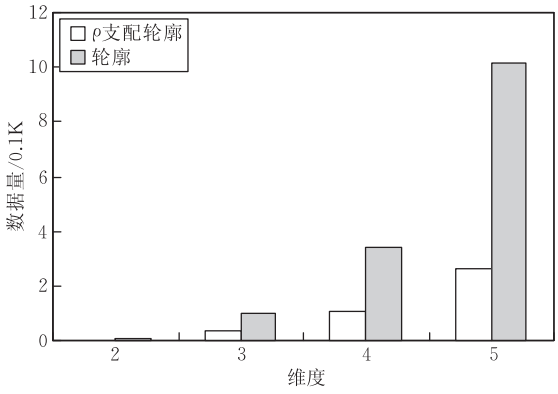


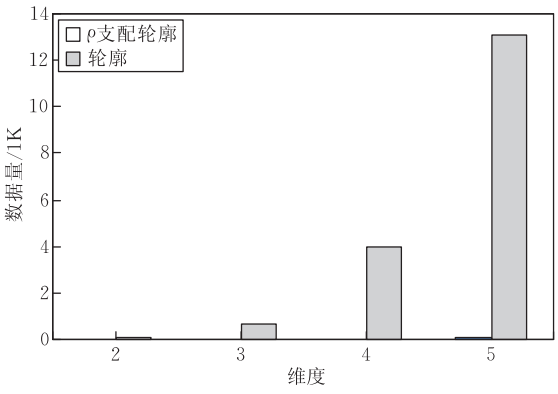
图 4　ρ 值大小对性能的影响

接着，考查维度大小对各个算法性能的影响. 图 5、图 6 中分别给出了数据集大小为 100 K、ρ 值为 1.5 时，维度由 2 维变化到 5 维时，结果集的大小变化以及各算法性能的变化规律. 无论在独立分布还是在反相关分布的数据集中，从图 5 中可以看出，返回的 ρ-支配轮廓集合数量和轮廓集合数量都随着维度增加而增加，同时由图 6 中可见，各算法的响应时间也随着数据维度的增加而相应的增加. 这是因为数据维度的增加，意味着元组间关系判别时需要计算的次数增加，同时，在数据集大小相同时，数据维度的增加还意味着元组被支配和 ρ-支配的概率降低，使得元组成为轮廓和 ρ-支配轮廓的可能性提高，轮廓和 ρ-支配轮廓的结果数量往往也同时增加，这些都导致了 ρ-支配轮廓计算时需要进行的计算次数增加. 在反相关分布下，轮廓集合的数量要远

远大于独立分布下轮廓集的数量，从而导致了各算法在反相关分布下的响应时间远远高于对应独立分布下的响应时间. 无论维度如何变化，在同样的数据分布下，BBDS 算法的响应时间始终低于 BA 算法的响应时间.



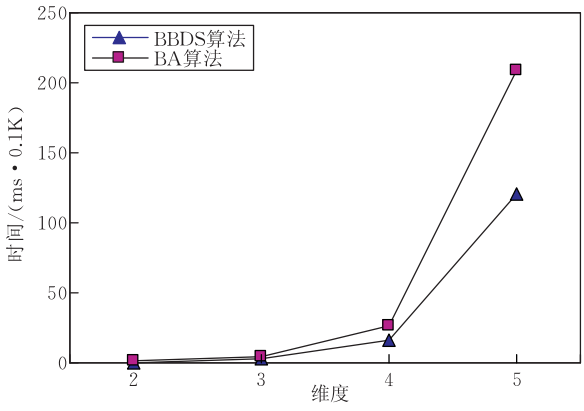
(a) 独立分布



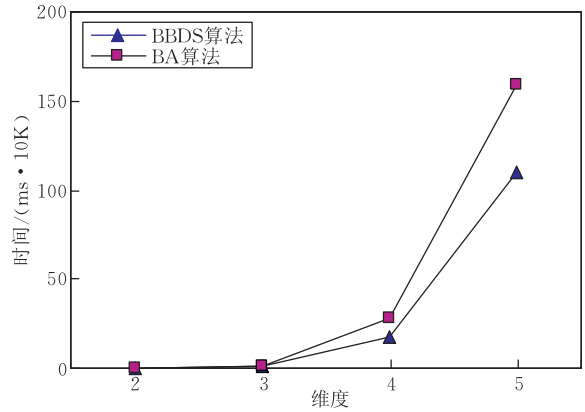
(b) 反相关分布

图 5 维度对结果集大小的影响

最后，考查数据集大小对各算法性能的影响. 图 7 给出了数据维度为 3、ρ 值为 1.5 时，数据集由 100 K 变化到 900 K 各算法性能的变化规律. 从图中可以看出，无论在独立分布还是在反相关分布的数据集中，各算法的响应时间都随着数据集的增加而相应地增加. 这是因为数据集大小的增加意味着元组成为轮廓的门槛降低，算法终止前需要处理的数据量增加，因此算法的响应时间相应增加. 同样，在反相关分布下，元组的被支配概率降低，使得元组成为轮廓的可能性提高，求 ρ-支配轮廓时需要进行的运算次数相应增加，因而算法在反相关分布下的响应时间高于独立分布下的响应时间. 同时，与数据维度的增加对算法性能的影响相比，数据集大小增加的影响要小很多，因此，在数据集增加时，算法响应时间的增速相对较慢，没有数据维度增加时变化的明显. 无论数据集大小如何变化，在同样的数据分布下，BBDS 算法的响应时间始终低于 BA 算法的响应时间.

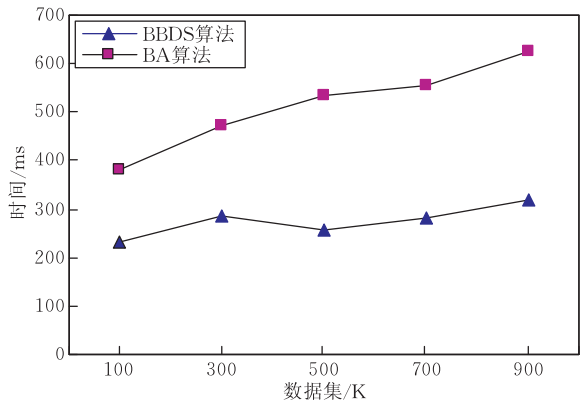


(a) 独立分布

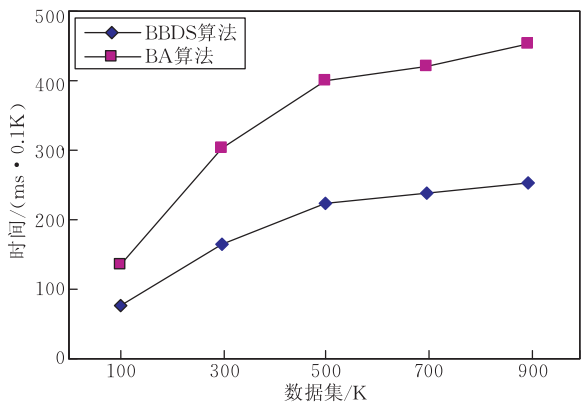


(b) 反相关分布

图 6 维度对算法性能的影响



(a) 独立分布



(b) 反相关分布

图 7 数据集大小对算法性能的影响

综上所述，无论在独立分布，还是在反相关分布下，BBDS 算法的性能始终优于 BA 算法，并且 BBDS 算法的响应时间在 10 s 到 1000 s 的量级，说明 BBDS 算法是计算 ρ-支配轮廓有效算法.

6 结　论

本文对 ρ-支配轮廓查询处理技术进行了深入的研究. 首先，通过对实际应用需求进行分析，提出了 ρ-支配关系以及基于该支配关系的 ρ-支配轮廓查询的概念；接着，深入分析了 ρ-支配轮廓查询的性质，提出了基于分支定界的 ρ-支配轮廓查询算法(Branch and Bound ρ-Dominant Skyline，BBDS)；最后，设计了详细的性能评价实验，实验结果表明 BBDS 算法可以有效地处理 ρ-支配轮廓查询. BBDS 算法的性能可以满足应用需求，是计算 ρ-支配轮廓的有效算法.

参　考　文　献

[1] Borzsonyi S, Kossmann D, Stocker K. The skyline operator//Proceedings of the 17th International Conference on Data Engineering. Heidelberg, Germany, 2001: 421-430

[2] Wei Xiao-Juan, Yang Jing, Li Cui-Ping, Chen Hong. Skyline query processing. Journal of Software, 2008, 19(6): 1386-1400(in Chinese)
(魏小娟，杨婧，李翠平，陈红. Skyline 查询处理. 软件学报，2008，19(6)：1386-1400)

[3] Balke W-T, Guntzer U. Multi-objective query processing for database systems//Proceedings of the 30th International Conference on Very Large Data Bases. Toronto, Canada, 2004: 936-947

[4] Chan C-Y, Jagadish H, Tan K-L et al. Finding k-dominant skylines in high dimensional space//Proceedings of the ACM SIGMOD International Conference on Management of Data Chicago. Illinois, USA, 2006: 503-514

[5] Chan C-Y, Jagadish H, Tan K-L et al. On high dimensional skylines//Proceedings of the 10th International Conference on Extending Database Technology. Munich, Germany, 2006: 478-495

[6] Lee J, You G, Hwang S. Personalized top-k skyline queries in high-dimensional space. Information Systems, 2009, 34(1): 45-61

[7] Sharifzadeh M, Shahabi C. The spatial skyline queries//Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul, Korea, 2006: 751-762

[8] Lin X, Yuan Y, Zhang Q et al. Selecting stars: The k-most representative skyline operator//Proceedings of the 23rd International Conference on Data Engineering. Istanbul, Turkey, 2007: 86-95

[9] Tao Y, Ding L, Lin X et al. Distance-based representative skyline//Proceedings of the 25th International Conference on Data Engineering. Shanghai, China, 2009: 892-903

[10] Papadias D, Tao Y, Fu G et al. An optimal progressive algorithm for skyline queries//Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. San Diego, California, USA, 2003: 467-478

[11] Yiu M L, Mamoulis N. Efficient processing of top-k dominating queries on multi-dimensional data//Proceedings of the 33rd International Conference on Very Large Data Bases. University of Vienna, Austria, 2007: 483-494

[12] Wu Jun-Jie, Xin Jun-Chang, Wang Guo-Ren, Zhou Shi-Yong. The k-dominating ranking skyline algorithm. Journal of Computer Research and Development, 2009, 46(S): 133-139(in Chinese)
(吴俊杰，信俊昌，王国仁，周诗咏. k 支配能力排序轮廓查询算法. 计算机研究与发展，2009，46(S)：133-139)

[13] Xia T, Zhang D, Tao Y. On skylining with flexible dominance relation//Proceedings of the 24th International Conference on Data Engineering. Cancún, México, 2008: 1397-1399

[14] Levandoski J, Mokbel M F, Khalefa M. FlexPref: A framework for extensible preference evaluation in database systems//Proceedings of the 26th International Conference on Data Engineering. Long Beach, California, USA, 2010: 828-839

[15] Zhang Z, Lu H, Ooi B C et al. Understanding the meaning of a shifted sky: A general framework on extending skyline query. VLDB Journal, 2010, 19(2): 181-201

XIN Jun-Chang, born in 1977, Ph. D., lecturer. His research interests include sensor data management and uncertain data management.

BAI Mei, born in 1986, Ph. D. candidate. Her research interests include sensor data management and uncertain data management.

DONG Han, born in 1981, engineer. His research interests include high performance computing, parallel computing, Internet services and cloud computing.

WANG Guo-Ren, born in 1966, professor, Ph. D. supervisor. His research interests include XML management, bioinformatics, distributed database, and parallel computing.

Background

As an important operator of multi-criteria decision making and user preference applications, skyline queries play a very important role in our daily life. In some different applications, traditional dominance relationship cannot meet the people's requirements. So several skyline variants based on different dominance relationships have been proposed by the database researchers recently, such as multi-objective skyline, k-dominate skyline, top-k frequency skyline and top-k skyline. However, all the dominance relationships of the existing skyline query variants are only dependent on the values of the objects. But when the values of the tuples change greatly, the values of the objects cannot distinguish which one is better, so this paper propose ρ-dominance relationship to solve this problem.

The ρ-dominance relationship make use of the ratio of the corresponding values to measure the "better or bad" relationships between two tuples, which can be widely used in our daily life. And ρ-dominance relationship is much more stable than traditional dominance relationship when the value of the tuples changes. In this paper, first of all, the ρ-dominance relationship based on the ratio of corresponding values between tuples is defined by analyzing the actual application requirements, and then the concept of ρ-dominant skyline query is proposed. The ρ-dominant skyline cannot use the traditional means to solve because of it has no transitivity. Secondly, a novel algorithm Branch and Bound ρ-Dominant Skyline Algorithm (BBDS) is developed through the detailed and in-depth analysis of its basic properties. BBDS improve the ρ-dominant skyline query implementation efficiency greatly by avoiding visiting R-tree index too many times. The last but not the least, the semantic of ρ-dominant skyline query is analyzed and the performance of BBDS algorithm is evaluated through various simulations. The simulation results show that the proposed ρ-dominant skyline query based on ρ-dominance relationship is a new extension and complement of the traditional skyline query semantic and the proposed BBDS algorithm is a highly effective algorithm for solving ρ-dominant skyline queries.

Skyline query problem has been a hot topic in the database research, and ρ-dominant skyline proposed in this paper is a new extension and complement of skyline semantic, so it has good research value. The proposed algorithm BBDS will have a good prospect in some decision-making problems.

This research is supported by the State Key Program of National Natural Science of China (Grant No. 60933001), the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61025007), the National Natural Science Foundation of China (Grant No. 61073063), and the Fundamental Research Funds for the Central Universities (Grant No. N090304007).