

UNIVERSIDAD SIMÓN BOLÍVAR

MINERÍA DE DATOS

CONSULTAS SKYLINE Y LA MINERÍA DE DATOS

FABIOLA DI BARTOLO

MARZO 2011

Agenda

- Introducción a las Consultas Skyline
- Consultas Skyline: definición y ejemplos.
- Aplicación de las Consultas Skyline en la Minería de Datos.
 - Agrupamiento de puntos Skyline y descubrimiento de sus vecinos.
 - Similitud entre Skylines.
- Conclusiones

Introducción a las Consultas Skyline



- Origen del término: se denomina Skyline al panorama creado por la silueta de las estructuras y edificios más altos de una ciudad.
- El Skyline de Manhattan podría ser calculado como el conjunto de edificios más altos y que a la vez, están cerca del río Hudson.
- El cálculo del Skyline es conocido como el Maximal Vector Problem, 1975 (identificar los maximales sobre una colección de vectores).
- Primer trabajo de investigación en bases de datos: *The skyline operator, 2001.*

Consultas Skyline: Motivación

Problemas típicos de decisión:

- **Agencia de viajes:** encontrar los hoteles con la mayor cantidad de estrellas, más cercanos a la playa y menos costosos.
- **Recursos humanos:** identificar aquellos aspirantes con mayor grado de instrucción académica, mayor experiencia laboral (en años) y menor sueldo esperado.
- **Sistema de recomendaciones:** conseguir los vehículos en venta con menor tiempo de uso, menor kilometraje y más económicos.

Consultas Skyline: Definición

Consulta Skyline en SQL:

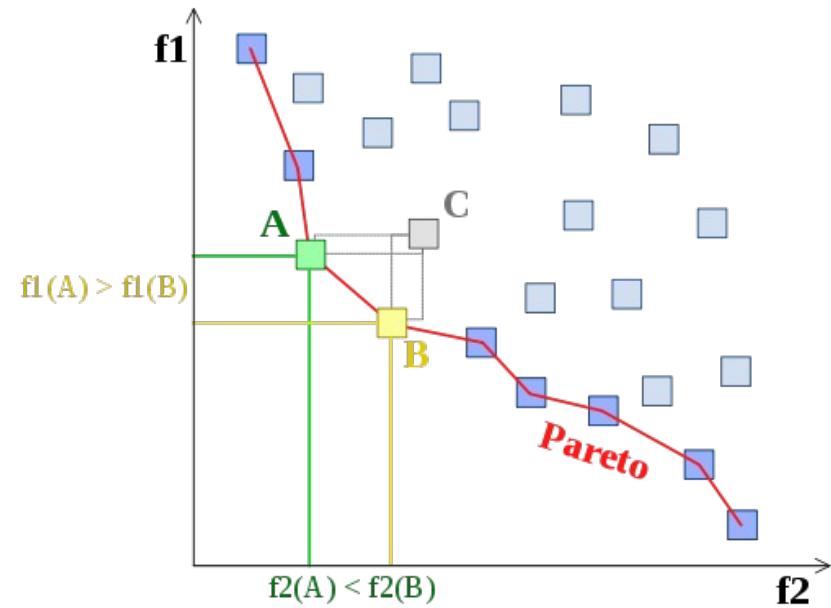
SELECT ... FROM ... WHERE ...

GROUP BY ... HAVING ...

SKYLINE OF [DISTINCT] d1 [MIN | MAX | DIFF], ..., dm [MIN | MAX | DIFF]

ORDER BY ...

- d_1, \dots, d_m son atributos numéricos.
- No se definen pesos para las preferencias.
- Filtra un conjunto de puntos interesantes sobre un gran conjunto de datos.
- Un punto es interesante si **no es dominado** por ningún otro punto.
Conforman el Conjunto de Pareto

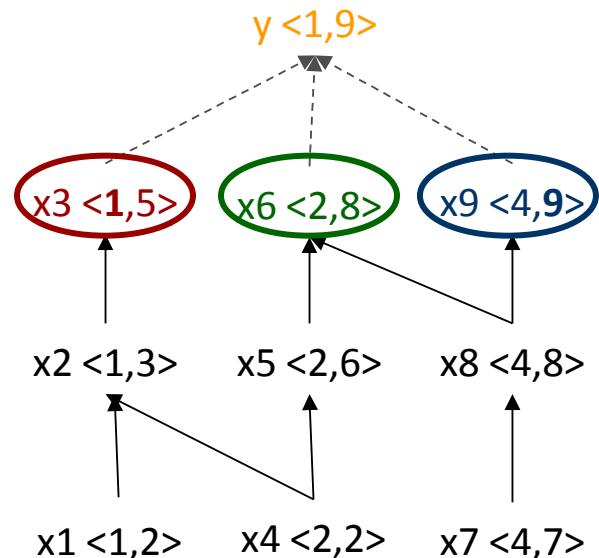


Consultas Skyline: Dominancia

- El Skyline define un **orden parcial** en el conjunto de datos según las preferencias dadas.
- Precedencia entre los elementos dada por la dominancia.
- x1 domina a x2**, si x1 es **mejor o igual** a x2 en todos las dimensiones y **mejor** a x2 en al menos una.
- Skyline=Conjunto de maximales o elementos no dominados.** Son los que mejor se adaptan a las preferencias.

Datos				
ID	a	b	c	
x1	1	2	...	
x2	1	3	...	
x3	1	5	...	
x4	2	2	...	
x5	2	6	...	
x6	2	8	...	
x7	4	7	...	
x8	4	8	...	
x9	4	9	...	

```
SELECT * FROM Datos d  
SKYLINE OF d.a MIN, d.b MAX
```



Dimensiones Skyline = a y b

Datos = {x1,x2,x3,x4,x5,x6,x7,x8,x9}

Skyline = {x3,x6,x9}

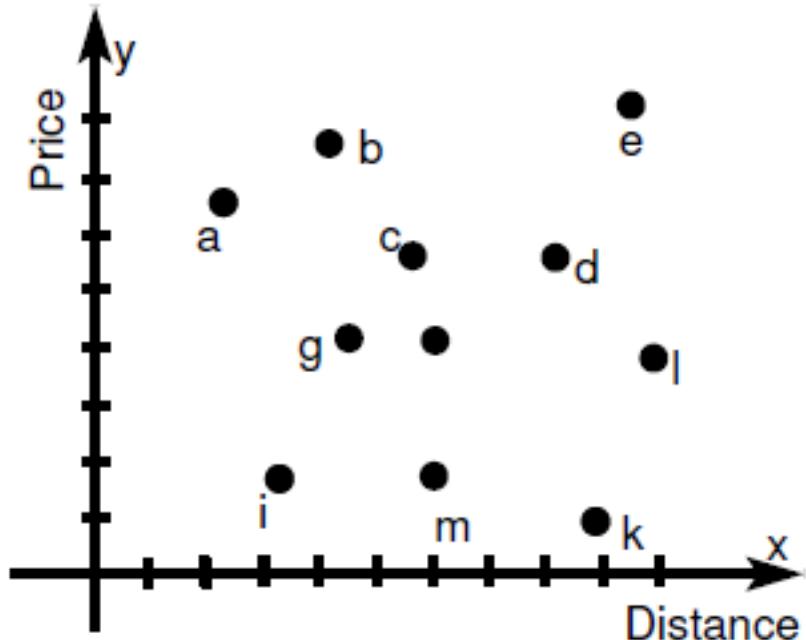
Consultas Skyline: Ejemplo (dataset de hoteles)

Objetivo:

Encontrar aquellos hoteles de Margarita más económicos y cercanos a la playa.

```
SELECT * FROM hoteles h  
WHERE h.location = 'Margarita'  
SKYLINE OF h.price MIN, h.distance MIN
```

- Los puntos interesantes son los que satisfacen total o parcialmente estas preferencias.



Consultas Skyline: Ejemplo (dataset de hoteles)

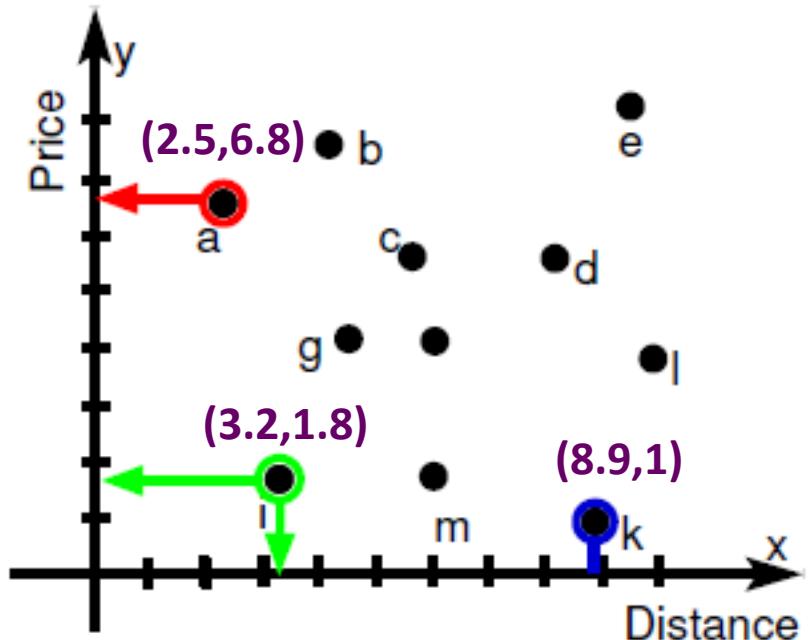
- En este dataset no existe un hotel que satisfaga ambas propiedades.

- Los puntos que conforman el Skyline son {a,i,k}:

a= menor distancia a la playa.

k= menor precio.

i= no tiene la menor distancia ni el menor precio, pero tiene menor precio que a y menor distancia que k.



El resto de los puntos son dominados por el conjunto {a,i,k}

Consultas Skyline: Ejemplo (dataset de hoteles)

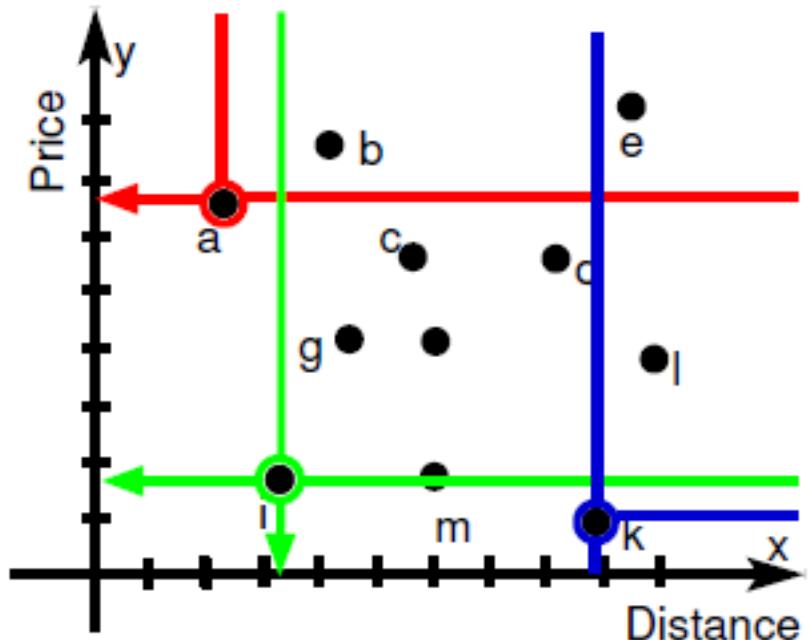
- En este dataset no existe un hotel que satisfaga ambas propiedades.

- Los puntos que conforman el Skyline son {a,i,k}:

a= menor distancia a la playa.

k= menor precio.

i= no tiene la menor distancia ni el menor precio, pero tiene menor precio que a y menor distancia que k.



El resto de los puntos son dominados por el conjunto {a,i,k}

Consultas Skyline en Minería de Datos: Trabajos Relacionados

Ofrecen técnicas que pueden ser utilizadas en tareas de agrupamiento, clasificación o detección de desviaciones y anomalías.

- Descubrimiento de patrones para la recomendación de objetos skyline y sus vecinos:
 - Thick Skylines (2004).
 - Subspace Skyline Clusters (2007).

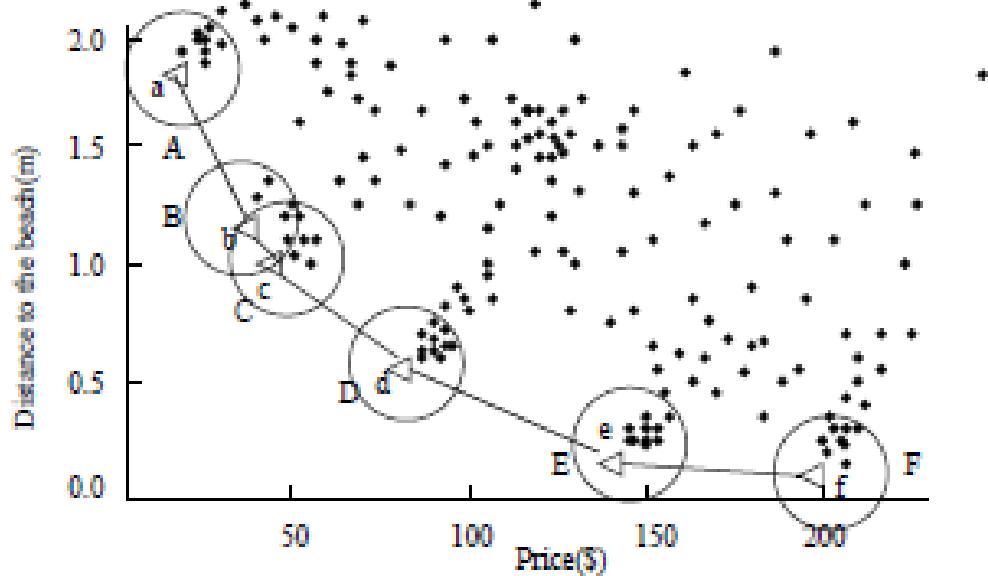
- Medidas de similitud para comparar diferentes skylines:
 - SkyDist (2010)

Consultas Skyline en Minería de Datos:

Recomendación de puntos skyline y sus vecinos (Thick Skyline)

Ejemplo: Si el skyline devuelve pocos puntos y los hoteles en el skyline están llenos... ¿existen otros hoteles cercanos que aún sean buenos candidatos?

Ofrece técnicas para recomendar no sólo los objetos en el skyline, sino también los objetos cercanos dentro de una ϵ -distancia de cada uno de estos.



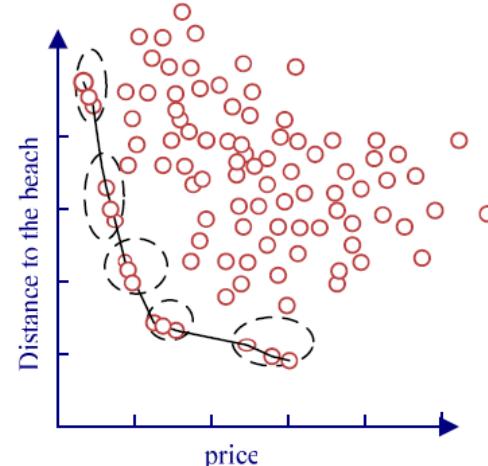
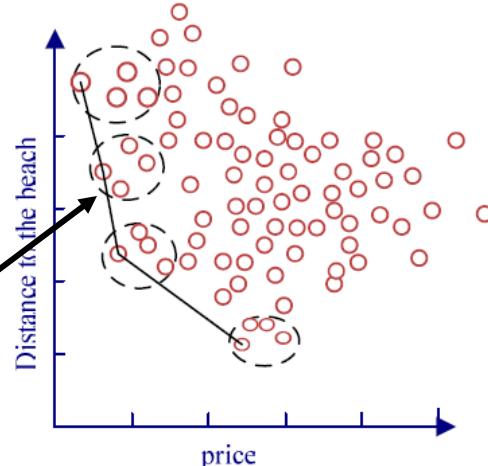
Thick Skyline of N.Y. hotels.

Consultas Skyline en Minería de Datos: Recomendación de puntos skyline y sus vecinos (Subspace Skyline Clusters)

- Permite hallar los vecinos en un subespacio de dimensiones del skyline.
- No todas las dimensiones son necesarias para hallar similitudes interesantes entre los objetos.
- + dimensiones => mayor cardinalidad del skyline.
- Genera patrones que reflejan las correlaciones entre los puntos, lo cual le proporciona al usuario, los objetos deseados agrupados en **skyline clusters**. (Evitando un posterior procesamiento manual de los datos).

Ej: Skyline de
pocos puntos +
vecinos

Subspace Skyline
Clusters



Ej: Agrupación
de los puntos
skyline por
clusters

Consultas Skyline en Minería de Datos: Recomendación de puntos skyline y sus vecinos (Subspace Skyline Clusters)

Idea:

- Encontrar los **top-k** vecinos del skyline en los subespacios skyline.
- Los elementos son ordenados por su distancia Manhattan al skyline.
- Cada algoritmo tiene como entrada una lista ordenada en cada dimensión y el subespacio a explorar. Devuelven los skyline clusters en ese subespacio.

Búsqueda de los vecinos para la creación de clusters:

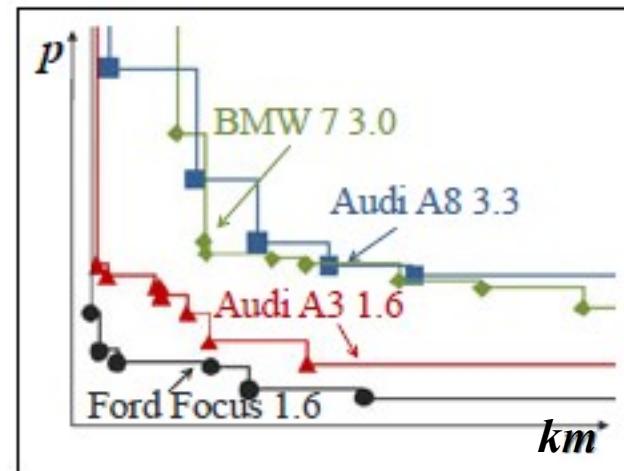
- **Enfoque 1:** accede en paralelo a las listas ordenadas en cada dimensión y encuentra los top-k objetos que hayan sido visitados en todas las listas.
- **Enfoque 2:** crea un umbral tomando los menores valores encontrados en todas las listas. Va tomando los objetos cercanos mientras que el score de los top-k objetos no sea mayor a ese umbral.

Consultas Skyline en Minería de Datos: Similitud entre Skylines

Mercado de carros usados:

- El skyline puede ser utilizado para identificar el comportamiento de cada modelo.
- Pueden existir ofertas muy malas, pero sólo las que están en el skyline o cerca son las que potencialmente encontraran un cliente.
- El skyline de las ofertas marca en cierto grado el precio justo de los vehículos para cada kilometraje.

Skylines de las ofertas de diferentes modelos de carros.
(Precio vs kilometraje)

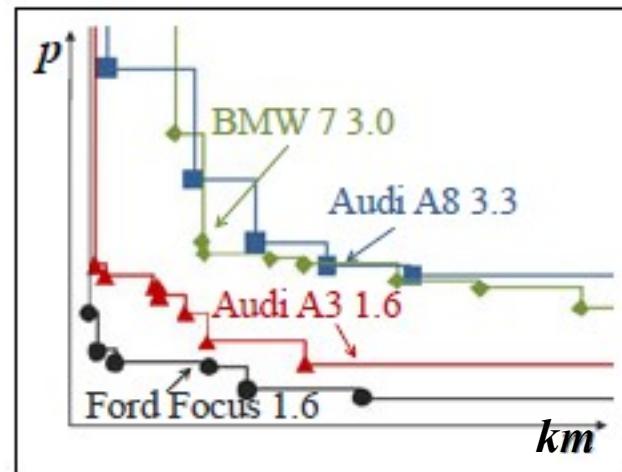


Consultas Skyline en Minería de Datos: Similitud entre Skylines

Mercado de carros usados:

- Cada skyline caracteriza un modelo de carro distinto. Similitudes entre skylines identifican modelos de carros similares.
- Un sistema de recomendación puede encontrar que el Focus es una excelente opción al Audi A3.
- Empleando medidas de similitud entre skylines, pueden realizarse tareas de agrupamiento, clasificación, detección de valores atípicos, etc.

Skylines de las ofertas de diferentes modelos de carros.
(Precio vs kilometraje)



Consultas Skyline en Minería de Datos: Similitud entre Skylines

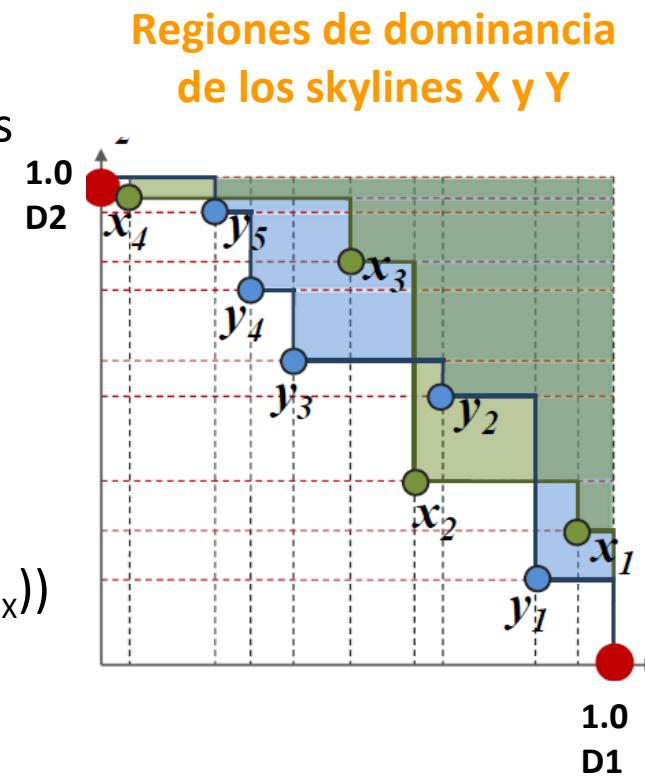
Para medir la similitud entre dos skylines proponen:

- **SkyDist:** función de distancia entre skylines.
- Si dos skylines son similares, entonces el SkyDist debe ser un valor pequeño.
- Dos skylines son similares si están formados por puntos similares.
- Dos skylines pueden ser considerados similares si sus puntos dominan aproximadamente a los mismos puntos en el espacio de datos.

Consultas Skyline en Minería de Datos: Similitud entre Skylines

Regiones de dominancia y el SkyDist:

- Sea D el espacio numérico de los datos.
- **Región de dominancia** $\text{DOM}_x =$ unión de todas las regiones dominadas por los puntos en el skyline. ($\text{DOM}_x = \text{región verde}$, $\text{DOM}_y = \text{región azul}$)
- **Región de no-dominancia** $\text{DOM}_x = D - \text{DOM}_x$
- $\text{SkyDist}(X,Y) = \text{Vol}((\text{DOM}_x - \text{DOM}_y) \cup (\text{DOM}_y - \text{DOM}_x))$
(Volumen del area entre los skylines X y Y)



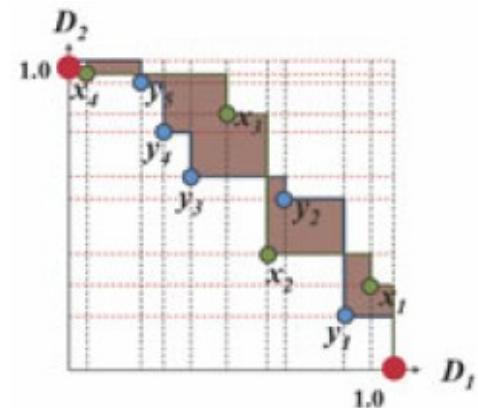
Consultas Skyline en Minería de Datos: Similitud entre Skylines

Cálculo de la SkyDist

□ Monte-Carlo Sampling (distancia aproximada)

- Se toma una cantidad de puntos al azar del conjunto de datos.
- Determina la proporción aproximada de los puntos que se encuentran en la región de dominancia de un skyline y que a la vez **no** están en la región de dominancia del otro skyline.
- Se calcula el radio entre los puntos que caen en la región del SkyDist y los que no.
- Este radio da una aproximación de la distancia entre los dos skylines.

SkyDist entre los
skylines X y Y
(en espacio 2D)

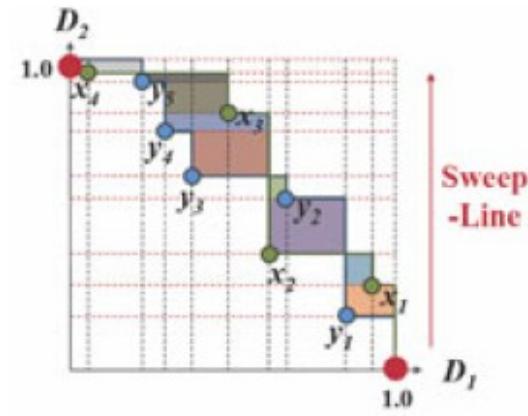


Consultas Skyline en Minería de Datos: Similitud entre Skylines

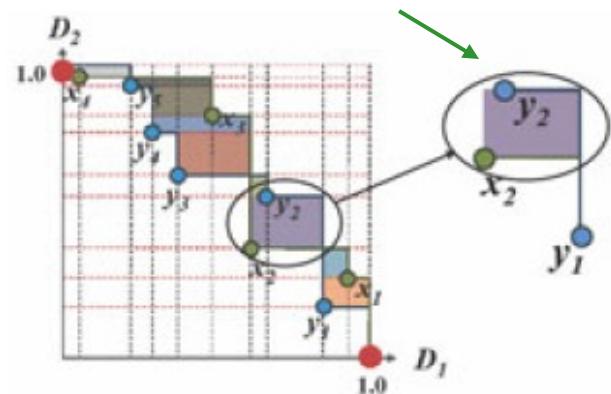
Cálculo de la SkyDist

- Método para calcular la distancia exacta entre los dos skylines:
 - Ordenar los puntos entre los skylines por una dimensión (Ej. D2) y almacenarlos en una lista.
 - Tomar las regiones de dominancias exclusivas de cada punto skyline (descartar las regiones de dominancia de los puntos que son dominados por estos).
 - Recorrer la lista ordenada, para calcular el área de cada rectángulo en cada parada.
 - La suma de los rectángulos determina la región del SkyDist.

SkyDist 2D de los skylines X y Y
(suma de los rectangulos)



$$\text{Area} = (x_2 \cdot D_1 - y_1 \cdot D_1) (y_2 \cdot D_2 - x_2 \cdot D_2)$$



Consultas Skyline en Minería de Datos: Similitud entre Skylines

Experimento 1: agrupamiento de los modelos de carros utilizando SkyDist.

Datos reales:

Sitio online de venta de vehículos:
<http://www.autoscout24.de>

Muestra:

1519 carros usados del 2000

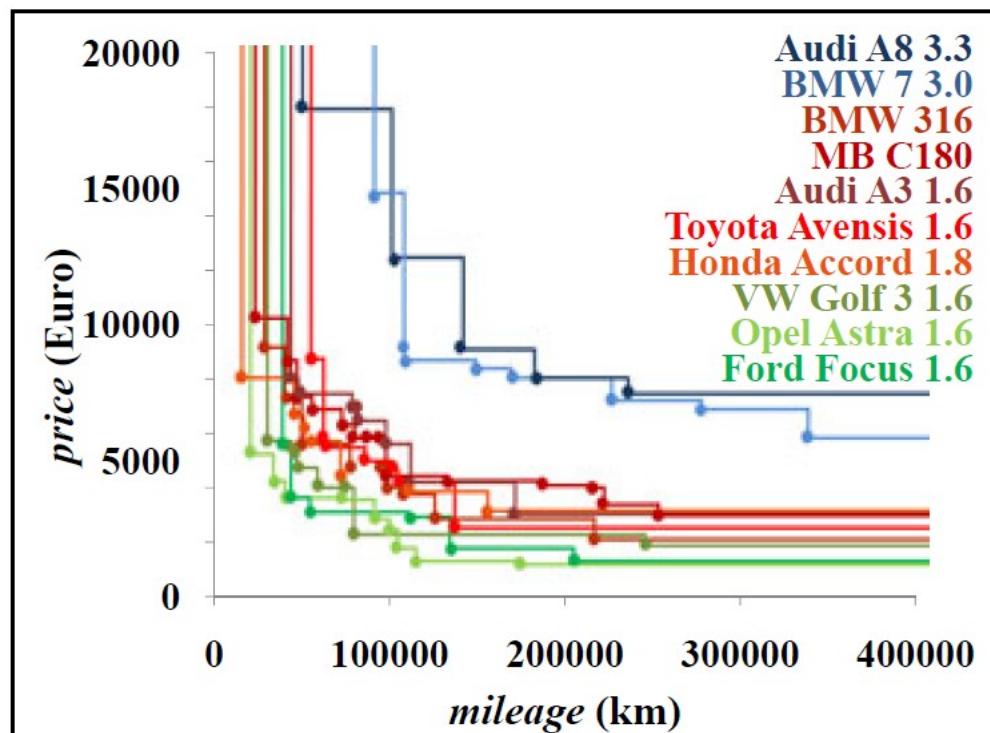
Grupos:

Compactos, medianos, lujosos.

Algoritmos utilizados con el SkyDist:

PAM, DBSCAN y SingleLink.

Todos los algoritmos combinados con la SkyDist produjeron grupos 100% puros según la clasificación real.



Consultas Skyline en Minería de Datos: Similitud entre Skylines

Experimento 1: agrupamiento de los modelos de carros utilizando SkyDist.

Algoritmos agrupamiento empleados:

- **PAM (método k-medias grupos: k=3).** 1990

Encuentra el conjunto de los k objetos que mejor caracterizan el dataset. A partir de estos puntos crea los clusters mediante la **similitudes** entre esos puntos y el resto de los objetos.

- **DBSCAN (alcance cluster: $\epsilon=10$, #puntos min por cluster: MinPts=2).** 1996

Agrupamiento basado en densidad, los clusters son áreas de alta densidad de objetos separados por áreas de baja densidad.

- **Dendograma Single Link (maxDistance=90 entre clusters a agrupar).** 1988

Agrupamiento jerárquico. Inicialmente parte de un cluster por objeto y va agrupando los clusters en cada iteración según la similitud entre ellos hasta quedar uno solo. El dendograma es la jerarquía dada por el orden en la unión de esos clusters.

Consultas Skyline en Minería de Datos: Similitud entre Skylines

Experimento 1: agrupamiento de los modelos de carros utilizando SkyDist.

Dendograma de Single Link usando SkyDist vs métricas convencionales

Distancia Euclíadiana entre los vectores p y q

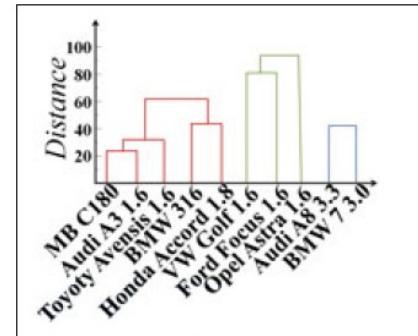
$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}.$$

Distancia Manhattan entre los vectores p y q

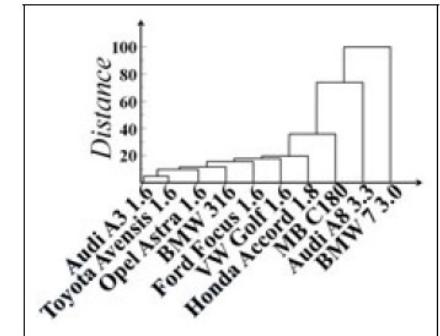
$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^n |p_i - q_i|,$$

Coseno entre los vectores A y B

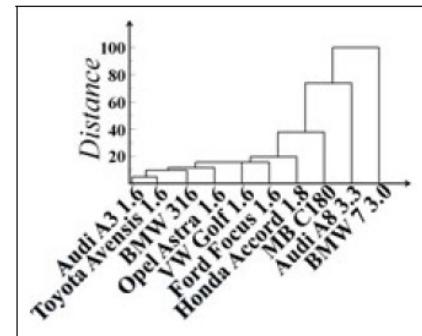
$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$



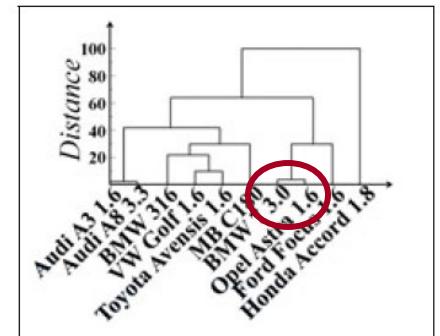
(a) SkyDist.



(b) Euclidean.



(c) Manhattan.



(d) Cosine.

Consultas Skyline en Minería de Datos: Similitud entre Skylines

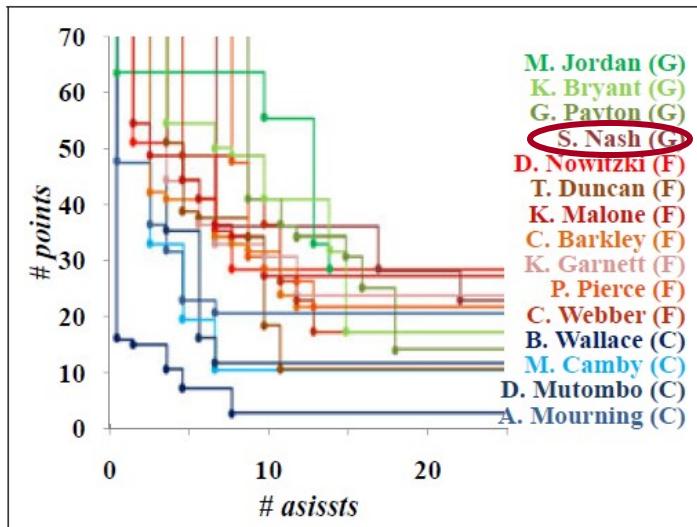
Experimento 2: agrupamiento de jugadores de la NBA utilizando SkyDist.

Datos reales: estadísticas de la NBA: <http://www.NBA.com>

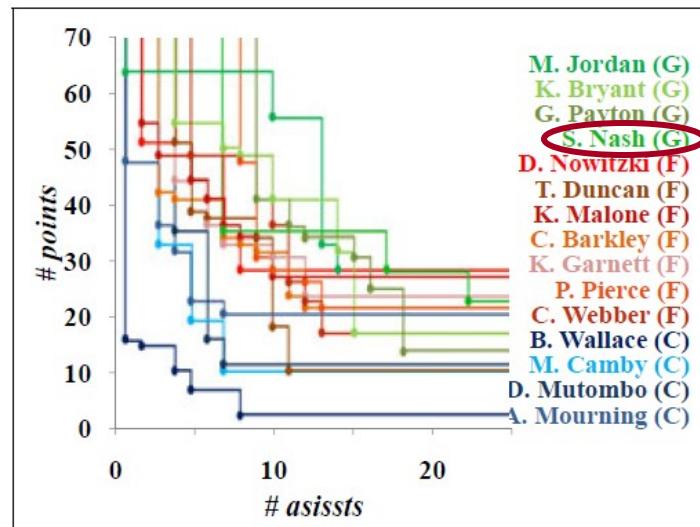
Muestra: estadísticas de los jugadores talentosos (1991 y 2005) con más de 500 juegos.

Grupos: **centrales (G), defensas (F), delanteros (C).**

Algoritmos utilizados: PAM ($k=3$), DBSCAN ($\epsilon=4$, MinPts=2) y SingleLink(maxDistance=96).



(a) PAM



(b) Single Link, DBSCAN.

Conclusiones

- El skyline de un conjunto de datos es muy útil para capturar las características más interesantes de estos datos.
- Los experimentos en tareas de aprendizaje no supervisado ofrecen buenos resultados utilizando skyline y la métrica de similitud SkyDist.
- Las técnicas de descubrimiento de patrones en los objetos skyline y puntos cercanos permiten el desarrollo de sistemas eficientes de recomendación de productos que le faciliten al usuario las tareas de búsqueda y selección.
- El cálculo del skyline, del SkyDist y el descubrimiento de clusters en el skyline, deberían ser integrados a las técnicas actuales de minería de datos.

Referencias Bibliográficas

- **SkyDist: Data Mining on Skyline Objects**; Christian Böhm, Annahita Oswald, Claudia Plant, Michael Plavinski and Bianca Wackersreuther ; ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING; Lecture Notes in Computer Science, 2010, Volume 6118/2010, 461-470, DOI: 10.1007/978-3-642-13657-3_49
- **Progressive Subspace Skyline Clusters Mining on High Dimensional Data**; Rong Hu, Yansheng Lu, Lei Zou and Chong Zhou; EMERGING TECHNOLOGIES IN KNOWLEDGE DISCOVERY AND DATA MINING. Lecture Notes in Computer Science, 2007, Volume 4819/2007, 268-279, DOI: 10.1007/978-3-540-77018-3_28
- **Mining Thick Skylines over Large Databases**; Wen Jin, Jiawei Han and Martin Ester; KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2004; Lecture Notes in Computer Science, 2004, Volume 3202/2004, 255-266, DOI: 10.1007/978-3-540-30116-5_25
- **The Skyline operator**; Borzsony, S.; Kossmann, D.; Stocker, K.; Passau Univ. Data Engineering, 2001. Proceedings. 17th International Conference on Data Engineering. IEEE Computer Society Washington, DC, USA 2001. ISBN:0-7695-1001-9.