



Übung zur Vorlesung *Einsatz und Realisierung von Datenbanken im SoSe20*

Maximilian {Bandle, Schüle}, Josef Schmeißer (i3erdb@in.tum.de)

<http://db.in.tum.de/teaching/ss20/impldb/>

Blatt Nr. 08

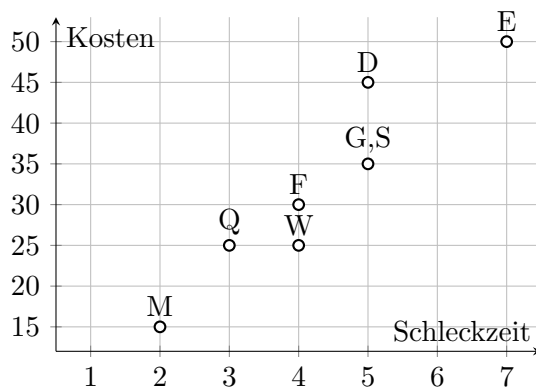
Hausaufgabe 1

Der Datenbanken-Lehrstuhl möchte wissen, mit welchem Eis der Gefrierschrank bestückt werden soll. Die Kosten sollen möglichst gering sein, aber die Schleckzeit möglichst groß. Hierfür wurde ein Test mit handelsüblichen Eissorten durchgeführt.

Eis			
<u>id</u>	Name	Schleckzeit (<i>min</i>)	Kosten (<i>ct</i>)
D	Double-Stieleis	5	45
E	Eiskonfekt	7	50
F	Frucht-Stieleis	4	30
G	Großes Stieleis	5	35
M	Mini-Stieleis	2	15
Q	Quetschtüte	3	25
S	Sandwich-Eis	5	35
W	Waffeltüte	4	25

Wir betrachten die Skyline über das **Maximum** des Attributs *Schleckzeit* sowie das **Minimum** des Attributs *Kosten* der Tabelle *Eis*.

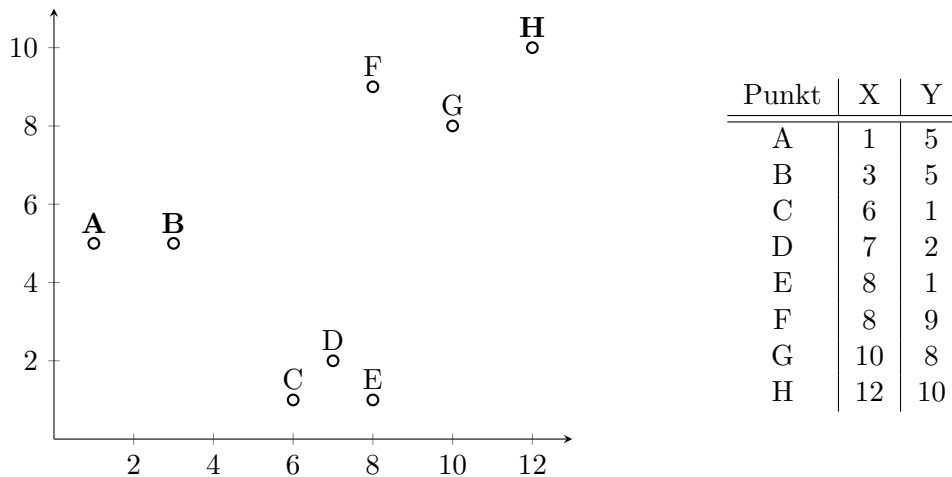
- Geben Sie die Anfrage, die die oben genannte Skyline mithilfe des Skyline-Operators berechnet.
- Geben Sie die Anfrage, die die oben genannte Skyline in SQL-92 berechnet, an (d.h. ohne Skyline-Operator).
- Vervollständigen Sie das unten gezeigte Diagramm. Zeichnen Sie alle Dominanzachsen ein.



- Geben Sie die Kürzel aller in der Skyline enthaltenen Tupel an.

Hausaufgabe 2

Gegeben seien folgende Datenpunkte, die im Plot und der Tabelle dargestellt sind. Die Punkte sollen mithilfe des k -Means-Algorithmus in drei Cluster aufgeteilt werden.



Als initiale Clusterzentren werden dabei folgende Punkte gewählt:
 Cluster (C_1) \rightarrow A; Cluster (C_2) \rightarrow B; Cluster (C_3) \rightarrow H.

- a) Führen Sie die Zuordnung für die erste Iteration qualitativ durch, indem sie das zugehörige Feld ankreuzen. Eine Rechnung oder Begründung ist nicht erforderlich.

	A	B	C	D	E	F	G	H
C_1								
C_2								
C_3								

- b) Berechnen Sie den Mittelpunkt M_3 von Cluster C_3 für die erste Iteration (Rechenweg angeben).
 c) Nennen Sie die Bedingung, nach der k -Means das Clustering optimiert.
 d) Geben Sie die Terminierungsbedingung von k -Means an.

Hausaufgabe 3

Gegeben seien Datenpunkte, welche im nachfolgenden Listing aufgeführt sind. Die Punkte sollen mithilfe des k -Means-Algorithmus in drei Cluster aufgeteilt werden. Als initiale Clusterzentren werden hierbei die jeweiligen Datenpunkte aus der `clusters_0`-Hilfsrelation gewählt.

```
with points (pid, x, y) as (values('A',1,5), ('B',3,5), ('C',6,1),
  ('D',7,2), ('E',8,1), ('F',8,9), ('G',10,8), ('H',12,10)
), clusters_0 (cid,x,y) as (values ('1',1e0,5e0), ('2',3e0,5e0),
  ('3',12e0,10e0))
```

- a) Formulieren Sie eine Iteration des k -Means-Algorithmus in SQL, die Ihnen die Clusterzentren zurückgibt. Nutzen Sie dazu eine Unterabfrage, die das Kreuzprodukt aus Clustern und Punkten berechnet und mit einer Window-Funktion pro Punkt ein Ranking der Cluster anhand der euklidischen Distanz erstellt.
 c) Berechnen Sie nun auf Grundlage Ihrer vorgehenden Anfrage die Zuordnung der Datenpunkte zu den jeweiligen Clusterzentren.
 d) Formulieren Sie nun Clusterberechnung als rekursive SQL-Anfrage mit folgendem Schema: `clusters_n (cid,step,x,y,delta)`. Nehmen Sie initial die gegebenen Clusterzentren. Verwenden Sie als Vorlage im Rekursionsschritt Ihre Anfrage aus Teilauf-

gabe a , welche die Clusterzentren pro Iteration Neuberechnet (**assign**). Hinweis: Nutzen Sie für die Fixpunkteriteration δ als die Summe aller Änderungen in Schritt **step**, um die Terminierungsbedingung des k -Means-Algorithmus zu formulieren. Ihre Anfrage soll terminieren, wenn die neu zugewiesenen Zentren gleich den vorherigen sind: $\delta = 0$.

Hausaufgabe 4

Alex und Max möchten sich für ihre neue Firma ein Fortbewegungsmittel zulegen. Hilf ihnen, die drei günstigsten bei 40.000 km Fahrleistung pro Jahr zu finden, wenn sie das Auto 5 Jahre lang nutzen wollen. Wende den NRA- und Threshold-Algorithmus an und bilde eine Skyline.

Einheit	Treibstoff	Preis
1l	Diesel	1,00€
1l	Benzin	1,50€
1l	Kerosin	1,00€
1kWh	Strom	0,10€

Kosten		Verbrauch	
Gefährt	Kosten	Gefährt	Verbrauch
Privatjet	2.500.000€	Privatjet	0,2l/km (Kerosin)
Elektroauto	80.000€	Elektroauto	20kWh/100km (Strom)
Cabrio	40.000€	Cabrio	4l/100km (Diesel)
Limousine	35.000€	Limousine	5l/100km (Diesel)
Transporter	20.000€	Transporter	6l/100km (Benzin)
Combi	25.000€	Combi	5l/100km (Benzin)
Sport-Coupé	25.000€	Sport-Coupé	4l/100km (Benzin)

Hausaufgabe 5

Zeigen Sie die weiteren Phasen des Apriori-Algorithmus für unser Beispiel in Abbildung 1 (hier ist lediglich bis inkl. 2. Phase dargestellt). Damit eine Menge von Produkten ein Frequentitemset ist, muss sie in mindestens $3/5$ aller Verkäufe enthalten sein, d.h. $minsupp = s_0 = 3/5$. Gehen Sie für die Assoziationsregeln von einer minimalen Konfidenz von $k_0 = 0$ aus und berechnen Sie die Konfidenz der Assoziationsregel $\{\text{Drucker}\} \Rightarrow \{\text{Papier, Toner}\}$.

VerkaufsTransaktionen		Zwischenergebnisse	
TransID	Produkt	FI-Kandidat	Anzahl
111	Drucker	{Drucker}	4
111	Papier	{Papier}	3
111	PC	{PC}	4
111	Toner	{Scanner}	2
222	PC	{Toner}	3
222	Scanner	{Drucker, Papier}	3
333	Drucker	{Drucker, PC}	3
333	Papier	{Drucker, Scanner}	
333	Toner	{Drucker, Toner}	3
444	Drucker	{Papier, PC}	2
444	PC	{Papier, Scanner}	
555	Drucker	{Papier, Toner}	3
555	Papier	{PC, Scanner}	
555	PC	{PC, Toner}	2
555	Scanner	{Scanner, Toner}	
555	Toner		

Abbildung 1: Ausgangssituation für den Apriori-Algorithmus