TU München, Fakultät für Informatik
Lehrstuhl III: Datenbanksysteme
Prof. Alfons Kemper, Ph.D.

Übung zur Vorlesung *Einsatz und Realisierung von Datenbanksystemen* im SoSe17

Maximilian E. Schüle (schuele@in.tum.de)
http://db.in.tum.de/teaching/ss17/impldb/

Blatt Nr. 08

Hausaufgabe 1

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 {Drucker}⇒ {Papier, Toner}.

VerkaufsTransaktionen

TransID	Produkt
111	Drucker
111	Papier
111	PC
111	Toner
222	PC
222	Scanner
333	Drucker
333	Papier
333	Toner
444	Drucker
444	PC
555	Drucker
555	Papier
555	PC
555	Scanner
555	Toner

Zwischenergebnisse

FI-Kandidat	Anzahl
{Drucker}	4
{Papier}	3
{PC}	4
{Scanner}	2
{Toner}	3
{Drucker, Papier}	3
{Drucker, PC}	3
{*Drucker, Scanner*}	
{Drucker, Toner}	3
{Papier, PC}	2
{*Papier, Scanner*}	
{Papier, Toner}	3
{*PC, Scanner*}	
{PC, Toner}	2
{*Scanner, Toner*}	

Abbildung 1: Ausgangssituation für den Apriori-Algorithmus

Vgl. Übungsbuch 17.6. Frequentitemsets sind alle nicht gestrichenen (wegen zu geringem Supports) bzw. nicht kursiv gesetzten (wegen nicht häufig auftretender Teilmenge).

Iteration	Item-Menge X	$\sigma(X)$	s(X)
1	{Drucker}	4	4/5
1	{Papier}	3	3/5
1	{PC}	4	4/5
1	~~{Scanner}~~	2	2/5
1	{Toner}	3	3/5
2	{Drucker, Papier}	3	3/5
2	{Drucker, PC}	3	3/5
2	{*Drucker, Scanner*}		
2	{Drucker, Toner}	3	3/5
2	~~{Papier, PC}~~	2	2/5
2	{*Papier, Scanner*}		
2	{Papier, Toner}	3	3/5
2	{*PC, Scanner*}		
2	~~{PC, Toner}~~	2	2/5
2	{*Scanner, Toner*}		
3	{*Drucker, Papier, PC*}		
3	{Drucker, Papier, Toner}	3	3/5
3	{*Drucker, PC, Toner*}		
3	{*Papier, PC, Toner*}		

Der Vollständigkeit halber im Nachfolgenden alle möglichen Assoziationsregeln.

Item-Menge X	$\sigma(X)$	$s(X)$	$c(X)$
$\emptyset \Rightarrow$ {Drucker}	4	4/5	4/5
$\emptyset \Rightarrow$ {Papier}	3	3/5	3/5
$\emptyset \Rightarrow$ {PC}	4	4/5	4/5
$\emptyset \Rightarrow$ {Toner}	3	3/5	3/5
$\emptyset \Rightarrow$ {Drucker, Papier}	3	3/5	3/5
{Drucker}$\Rightarrow$ {Papier}	3	3/5	3/4
{Papier}$\Rightarrow$ {Drucker}	3	3/5	3/3
$\emptyset \Rightarrow$ {Drucker, PC}	3	3/5	3/5
{Drucker}$\Rightarrow$ {PC}	3	3/5	3/4
{PC}$\Rightarrow$ {Drucker}	3	3/5	3/4
$\emptyset \Rightarrow$ {Drucker, Toner}	3	3/5	3/5
{Drucker}$\Rightarrow$ {Toner}	3	3/5	3/4
{Toner}$\Rightarrow$ {Drucker}	3	3/5	3/3
$\emptyset \Rightarrow$ {Papier, Toner}	3	3/5	3/5
{Papier}$\Rightarrow$ {Toner}	3	3/5	3/3
{Toner}$\Rightarrow$ {Papier}	3	3/5	3/3
$\emptyset \Rightarrow$ {Drucker, Papier, Toner}	3	3/5	3/5
{Drucker}$\Rightarrow$ {Papier, Toner}	3	3/5	3/4
{Drucker, Papier}$\Rightarrow$ {Toner}	3	3/5	3/3
{Drucker, Toner}$\Rightarrow$ {Papier}	3	3/5	3/3
{Papier}$\Rightarrow$ {Drucker, Toner}	3	3/5	3/3
{Papier, Toner}$\Rightarrow$ {Drucker}	3	3/5	3/3
{Toner}$\Rightarrow$ {Drucker, Papier}	3	3/5	3/3

Hausaufgabe 2

Die in Abbildung 2 dargestellten Relationen Mietspiegel und Kindergarten dienen der Bewertung von Wohngegenden im Großraum München. Für eine junge Familie ist ausschlaggebend, wie hoch die Lebenshaltungskosten gemessen an zu zahlender Miete und zu entrichtender Gebühr für den Kindergarten im jeweiligen Wohnort ausfallen. Illustrieren Sie die Ausführung einer Top-1-Berechnung (zur Bestimmung des günstigsten Wohnorts) für eine junge Familie mit zwei Kindern. Zeigen Sie die phasenweise Berechnung des Ergebnisses jeweils mit dem Threshold- und dem NRA-Algorithmus.

Mietspiegel	
Ort	Miete
Garching	800
Ismaning	900
Unterföhring	1000
Nymphenburg	1500
Bogenhausen	1600
Grünwald	1700

Kindergarten	
Ort	Beitrag
Grünwald	-100
Unterföhring	0
Bogenhausen	100
Ismaning	200
Garching	250
Nymphenburg	300

WohnLage	
Ort	Lage
Grünwald	München-Süd
Unterföhring	München-Nord
Ismaning	München-Nord
Garching	München-Nord
Bogenhausen	München-City
Nymphenburg	München-City

Abbildung 2: Münchner Wohnlagen zur Berechnung der monatlichen Kosten für eine Familie.

Siehe Lösungsbuch

Hausaufgabe 3

Geben die Relation Klausur:

MatrNr	Vorbereitungszeit	Note
1	150	1.7
2	70	2.7
3	450	2.0
4	180	1.7
5	2500	1.3

a) Formulieren Sie die Anfrage, die die MatrNr in der Skyline für die Attribute Vorbereitungszeit und Note erzeugt (kleiner ist jeweils besser) in SQL mit Hilfe des Skyline Operators.

b) Formulieren Sie die Anfrage in SQL ohne Skyline Operator.

c) Bestimmen Sie das Ergebnis der Anfrage.

```
with Klausur (MatrNr, Vorbereitungszeit, Note) as(
  values (1,150,1.7),(2,70,2.7),(3,450,2.0),(4,180,1.7),(5,2500,1.3)
)
```

SQL mit Skyline:

```
select MatrNr from Klausur k skyline of k.Vorbereitungszeit min, k.Note min
```

SQL ohne Skyline:

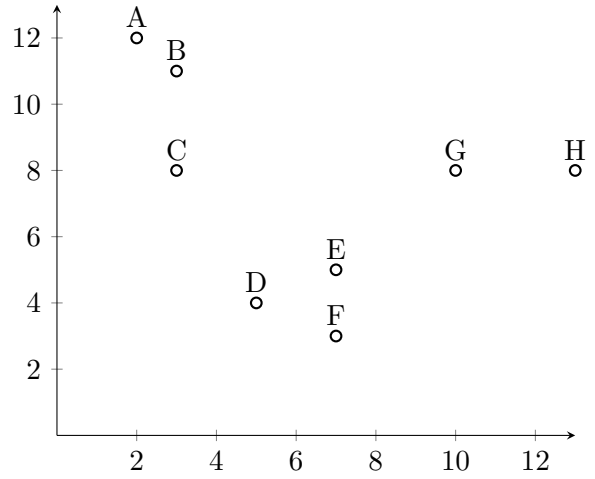
```
select MatrNr from Klausur k
where not exists (
   select * from klausur dom
   where
      dom.Vorbereitungszeit <= k.Vorbereitungszeit and
      dom.Note <= k.Note and (
         dom.Vorbereitungszeit < k.Vorbereitungszeit or
         dom.Note < k.Note)
)
```

Ergebnis:

1) Ist in Skyline (Kann in Vorbereitungszeit nur von MatrNr 2 dominiert werden, dort ist aber Note schlechter)
2) Ist in Skyline (Minimum für Vorbereitungszeit)
3) Ist nicht in Skyline, dominiert von MatrNr 1
4) Ist nicht in Skyline, dominiert von MatrNr 1
5) Ist in Skyline (Minimum für Note)

Hausaufgabe 4

Folgende Datenpunkte im euklidischen Raum seien gegeben:



Punkt	X	Y
A	2	12
B	3	11
C	3	8
D	5	4
E	7	5
F	7	3
G	10	8
H	13	8

Clustern Sie die Punkte mithilfe des *k-means*-Verfahren in 3 Cluster. Nutzen Sie als initiale Clusterzentren die Werte A, B und C. Wenn ein Punkt zu mehreren Clustern die gleiche Distanz hat, wird er dem Cluster der näher am Nullpunkt liegt zugeordnet. Geben Sie für jede Iteration jeweils die Zuordnung und die Mittelpunkte der Cluster an.
Eine Iteration des K-Means-Algorithmus kann wie folgt ausgewertet werden:

```
with points(id,x,y) as (
        VALUES ('A', 2, 12), ('B', 3, 11), ('C', 3,8), ('D', 5,4),
        ('E',7,5),('F',7,3),('G',10,8),('H',13,8)
),
clusters_0(cid,x,y) as (
  VALUES ('1', 2, 12), ('2', 3, 11), ('3', 3,8)
),
clusters_1(cid, x,y, count) as (
  select cid, avg(px), avg(py), count(*) from (
    select cid, p.x as px, p.y as py, rank() OVER (
      partition by p.id
      order by (p.x-c.x)*(p.x-c.x)+(p.y-c.y)*(p.y-c.y) asc,
        (c.x*c.x+c.y*c.y) asc)
    from points p, clusters_0 c
  ) x
  where x.rank=1
  group by cid
)
```

Die Clusterzentren können mit folgender Abfrage ausgegeben werden

```
select * from clusters_1
```

Die Zuordnung kann mit folgender Abfrage ausgewertet werden

```
select cid,pid from (
    select cid, p.id as pid, rank() OVER (
        partition by p.id
        order by (p.x-c.x)*(p.x-c.x)+(p.y-c.y)*(p.y-c.y) asc,
            (c.x*c.x+c.y*c.y) asc)
    from points p, clusters_1 c
) x
where x.rank=1
```

Hausaufgabe 5

Entwerfen Sie einen Algorithmus, um den Klassifikationsbaum, wie er in Abbildung 3 dargestellt ist automatisch zu ermitteln.

Schadenshöhe			
wiealt	Geschlecht	Autotyp	Schäden
45	w	Van	gering
18	w	Coupé	gering
22	w	Van	gering
38	w	Coupé	gering
19	m	Coupé	hoch
24	m	Van	hoch
40	m	Coupé	hoch
40	m	Van	gering
⋮	⋮	⋮	⋮

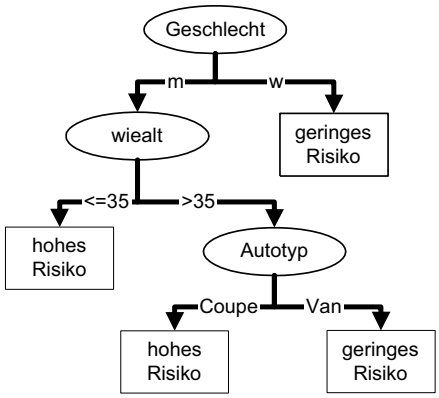


Abbildung 3: Klassifikationsschema für Haftpflicht-Risikoabschätzung.

Hausaufgabe (wird nicht in der Übung besprochen)

Die in Abbildung 4 dargestellten Relationen Autos und Unterhalt dienen der Bewertung von Autos. Eine junge Studierende sucht ein Auto mit guter Balance zwischen Sportlichkeit und Kosten. Sie überlegt sich wie die drei Werte Preis, PS und monatlicher Unterhalt in einen Score umberechnet werden können und nutzt schließlich folgende Formel:

$Preis - (100 * PS) + 24 * Unterhalt$

Zeigen Sie die phasenweise Berechnung der Top-3 Ergebnisse jeweils mit dem Threshold- und dem NRA-Algorithmus. Prüfen sie vor der Berechnung ob Teile der Scoringformel schon innerhalb jeder Relation vorberechnet werden können.

Auto	Preis	PS
Seat Leon	25000€	200
Audi A1	17000€	96
Citroen DS 4	20679€	100
Mini One	16500€	75
Mercedes C-Klasse	35000€	160
Porsche Cayenne	80100€	420

Auto	Unterhalt p. Monat
Seat Leon	215€
Audi A1	220€
Citroen DS 4	225€
Mini One	262€
Mercedes C-Klasse	290€
Porsche Cayenne	430€

Abbildung 4: Autokauf und -Unterhaltskosten.

Der erste Teil der Formel $Preis - (100 * PS)$ kann schon für jede Reihe in der Relation Autokauf vorberechnet werden.

Auto	PreisPS
Seat Leon	5000
Audi A1	7400
Mini One	9000
Citroen DS 4	10679
Mercedes C-Klasse	19000
Porsche Cayenne	38100

Bei der Top-3 Berechnung wird dann mit dieser sortierten Relation und der sortierten Unterhaltskosten Relation gearbeitet.

Threshold Algorithmus

Zw. Ergebnis: Phase 1

Auto	Score
Threshold	10160
Seat Leon	10160

Zw. Ergebnis: Phase 2

Auto	Score
Seat Leon	10160
Threshold	12680
Audi A1	12680

Zw. Ergebnis: Phase 3

Auto	Score
Seat Leon	10160
Audi A1	12680
Threshold	14400
Mini One	15288
Citroen DS 4	16079

Zw. Ergebnis: Phase 4

Auto	Score
Seat Leon	10160
Audi A1	12680
Mini One	15288
Citroen DS 4	16079
Threshold	16967

NRA Algorithmus

NRA: Phase 1

Auto	Score
Seat Leon	10160

NRA: Phase 2

Auto	Score
Seat Leon	10160
Audi A1	12680

NRA: Phase 3

Auto	Score
Seat Leon	10160
Audi A1	12680
Citroen DS 4	14400[p]
Mini One	14400[p]

NRA: Phase 4

Auto	Score
Seat Leon	10160
Audi A1	12680
Mini One	15288
Citroen DS 4	16079