Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Peer Kröger
Johannes Aßfalg, Karsten Borgwardt

Knowledge Discovery in Databases

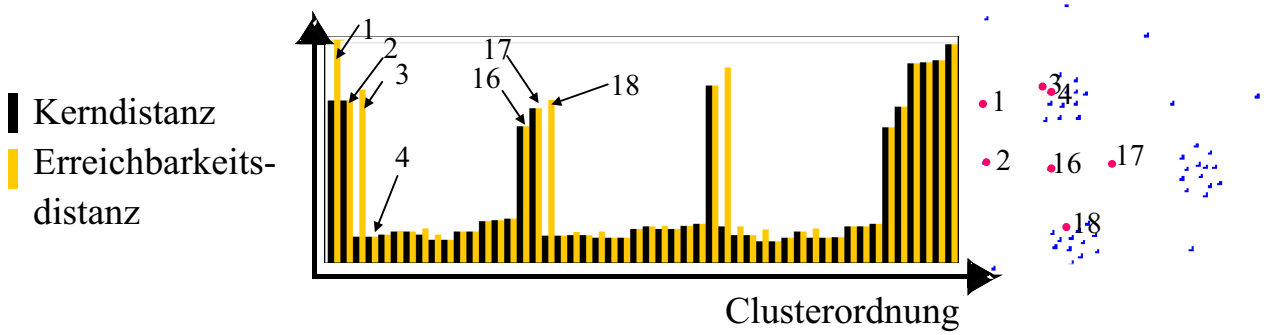
WS 2005/06

Übungsblatt 5

Abgabe aller mit Hausaufgabe *markierten Aufgaben bis Donnerstag, 8.12.2005, 8:30 Uhr, vor der Vorlesung beim Dozenten oder im Übungsbriefkasten*

In der Woche vom 28.11.05 bis 2.12.05 entfallen Übungen und Vorlesung. Die Bearbeitungszeit für dieses Blatt beträgt daher 2 Wochen.

Aufgabe 5-1 Zusammenhang zwischen DBSCAN und OPTICS
Hausaufgabe



Sei OPTICS auf eine Datenbank mit den Parametern ε und *MinPts* angewandt worden.

Geben Sie ein Verfahren an, wie man aus dem Resultat des OPTICS Laufes (Clusterordnung, Erreichbarkeitsdiagram und Kerndistanzdiagram) das DBSCAN-Clustering für ein gegebenes $\varepsilon' \leq \varepsilon$ extrahieren kann! Benutzen Sie möglichst intuitiven Pseudocode.

Kann aus dem OPTICS-Ergebnis eine eindeutige Clusterzugehörigkeit abgeleitet werden, die DBSCAN bzgl. des gegebenen $\varepsilon' \leq \varepsilon$ erzeugen würde? Mit anderen Worten: stimmt das Ergebnis ihres Verfahrens exakt mit dem Ergebis eines DBSCAN-Laufes bzgl. ε' überein? Begründen Sie Ihre Antwort!

Aufgabe 5-2 Join-Schritt des Apriori-Algorithmus

Gegeben ist die Menge L_k der k-Frequent Itemsets. Im Join-Schritt des Apriori-Algorithmus wird aus dieser Menge die Menge C_{k+1} der $(k+1)$-Kandidaten berechnet.

Beweisen Sie, daß der Join-Schritt korrekt ist, d.h. daß $C_{k+1} \supseteq L_{k+1}$.

Aufgabe 5-3 Apriori-Algorithmus
Hausaufgabe

Gegeben ist die Menge der Items $I = \{$A, B, C, D, E, F, G, H, I, K, L, M$\}$.

Weiterhin ist eine Menge von Transaktionen T laut folgender Tabelle gegeben:

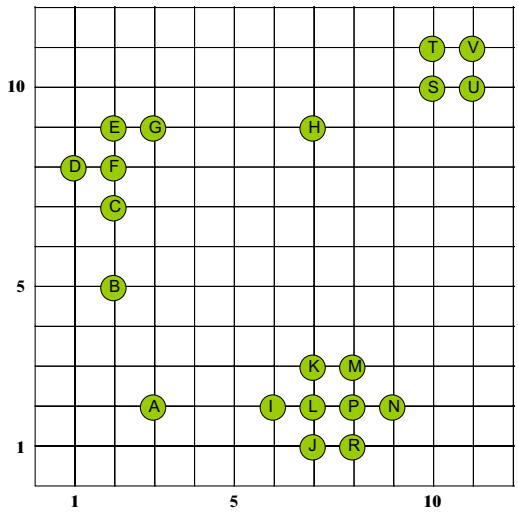
Menge der Transaktionen T

Transaktions ID	gekaufte Items
1	B E G H
2	A B C E G H
3	A B C E F H
4	B C D E F G H L
5	A B E K H
6	B E F G H I K
7	A B D G H
8	A B D G
9	B D F G
10	C E F
11	A C E F H
12	A B E G

Bestimmen Sie zum minimalen Support von 30% die häufig auftretenden Itemsets. Verwenden Sie dazu den Apriori-Algorithmus. Geben Sie insbesondere die Kandidatenmengen nach den Join-Schritten und nach den Prune-Schritten an, sowie die häufig auftretenden Itemsets mit ihrem jeweiligen Support.

Aufgabe 5-4 Outlier Detection

Gegeben ist - wieder einmal - der folgende 2-dimensionale Datensatz:



Verwenden Sie als Distanzfunktion auf den Punkten wieder die Manhattan-Distanz (L_1-Norm):

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

Bestimmen Sie die Outlier-Punkte der Tiefe 1 und 2 mit Hilfe der Depth-Based Outlier Detection!