

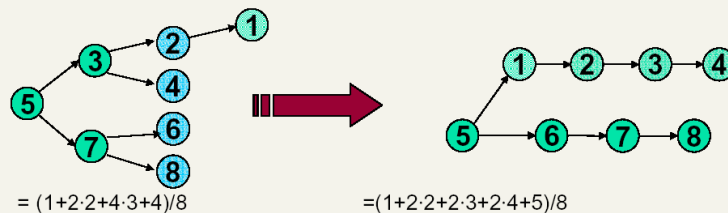
Musterlösung Übung 2

Alexander Hinneburg

6. November 2006

Blocking

- Binary search down to 4-term block;
- Then linear search through terms in block.
- 8 documents: binary tree ave. = 2.6 compares
- Blocks of 4 (binary tree), ave. = 3 compares



1 Hierarchisches Clustering

Gesucht sei eine Laufzeit und Speicherplatz-effiziente Implementierung des Algorithmus für hierarchisches Clustering Complete-Linkage. Die Eingabe sei die Distanzmatrix mit allen paarweisen Distanzen zwischen den Instanzen $D = \{d_{x,y}\}$ mit $x, y = 1 \dots, n$.

Der generische Algorithmus für bottom-up hierarchisches Clustering initialisiert die Liste der aktuellen Cluster mit den einzelnen Instanzen (Einer-Mengen) als initialen Clustern. Die Distanzen zwischen den initialen Clustern ist durch D gegeben. In $n - 1$ Schritten wird eine Hierarchie (binärer Baum) aufgebaut, der die n initialen Cluster als

Blätter hat. In jedem Schritt werden die zwei Cluster zusammengefaßt, welche die kleinste Distanz zu einander haben, s.d. die Liste der aktuellen Cluster um eins schrumpft. Nach $n - 1$ Schritten bleibt ein Cluster übrig, der alle Instanzen enthält.

Bei Complete Linkage ist die Distanz $d(C_i, C_j)$ zwischen zwei Clustern C_i und C_j die maximale paarweise Distanz zwischen zwei Instanzen aus C_i und C_j :

$$d(C_i, C_j) = \max\{d(x, y) : x \in C_i, y \in C_j\} \quad (1)$$

Seien im Schritt l ($1 \leq l \leq n - 1$) C_i und C_j die zwei Cluster mit der kleinsten Distanz zueinander, dann werden sie vereinigt und bilden einen neuen Cluster $C_l = C_i \cup C_j$. Für den neuen Cluster C_l muß die Distanz zu allen noch aktiven Clustern C_k berechnet werden. Dies wird effizient mit dem Lance-Williams Aktualisierungsschema für Complete Linkage gemacht

$$d(C_l, C_k) = \max\{d(C_i, C_k), d(C_j, C_k)\} \quad (2)$$

Dies ist viel effizienter, als wenn man die Distanzen über die Original-Definition (1) Neuberechnet. Überlegen Sie sich ein Beispiel dafür, um die Idee zu verstehen. Nach der Aktualisierung der Distanzen zu dem neuen Cluster C_l werden die zusammengefaßten Cluster C_i und C_j aus der Liste der aktuellen Cluster gestrichen (und auch die Distanzen zu diesen Clustern werden nicht mehr gebraucht).

Die Ausgabe soll eine Tabelle mit drei Spalten sein `0(idx1,idx2,linkdist)`. Die beiden Indizes `x` verweisen auf die zusammengefaßten Cluster und `linkdist` speichert die Distanz zwischen diesen Clustern. Die initialen Cluster sollen nicht in der Ausgabe auftauchen, werden aber implizit mitgedacht indem die Indizes auf sie verweisen.

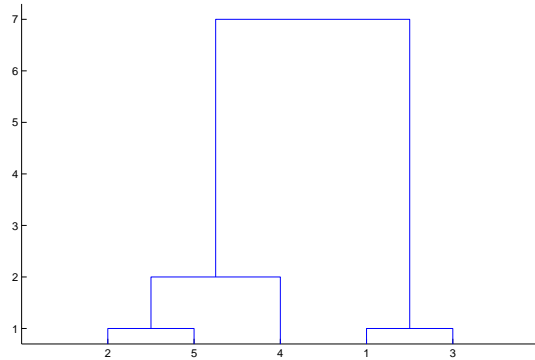
Ein Beispiel, die Instanzen sind durch ein-dimensionale Vektoren (also Zahlen) beschrieben $X = \{4, 9, 3, 10, 8\}$. Die Distanzmatrix D ist:

$$d = \begin{pmatrix} 0 & 5 & 1 & 6 & 4 \\ 5 & 0 & 6 & 1 & 1 \\ 1 & 6 & 0 & 7 & 5 \\ 6 & 1 & 7 & 0 & 2 \\ 4 & 1 & 5 & 2 & 0 \end{pmatrix}$$

Speicherplatz-effizienter ist es nur die obere Dreiecksmatrix als Vektor zu speichern: $D = (5, 1, 6, 4, 6, 1, 1, 7, 5, 2)$. Die Ausgabe von Complete-Linkage ist dann:

idx1	idx2	LinkDist
2	5	1
1	3	1
4	6	2
7	8	7

Als Dendrogramm sieht das Ganze so aus, die X-Achse gibt den Index in der Datenmenge X an.



Hinweise und Links zur weiterer Literatur finden Sie auf der Vorlesungswebseite.

Complete Linkage kann schneller berechnet werden, wenn zuerst alle Zeilen der Distanzmatrix sortiert werden, was $O(n^2 \log n)$ kostet. So kann das Minimum immer in $O(n)$ gefunden werden. Das Aktualisieren der Distanzmatrix mittel Lance-Williams Schema kostet $O(n)$, ebenso das Finden der minimalen Distanz des neuen Clusters zu den restlichen Clustern.

Bei Single Linkage geht die Berechnung sogar noch schneller. Zuerst wird ein Feld mit der kleinsten Distanz eines Clusters zum Rest für jedn Cluster berechnet (kostet $O(n^2)$). Dann werden die Cluster gesucht, die zusammengefasst werden sollen. Das Zusammenfassen kostet $O(n)$, ebenso das Aktualisieren des Feldes mit den kleinsten Distanzen. Für Single Linkage gilt, dass wenn für einen Cluster C_k einer der zusammengefaßten Cluster C_i oder C_j die Minimumdistanz hatte, dann hat der zusammengefaßte Cluster $C_i \cap C_j$ die minimale Distanz zu C_k . Für Complete Linkage gilt dies nicht.