Einführung in IR - Winter 2006/07
Dipl.-Inform. Sascha Kriewel, LF 137
sascha.kriewel@uni-due.de

Übungsblatt 8 — Clustering
keine Abgabe

Aufgabe 15: k-means

Es seien die folgenden Dokumente mit zweidimensionaler Repräsentation vorgegeben: $d_1 = (2, 10), d_2 = (2, 5), d_3 = (8, 4), d_4 = (5, 8), d_5 = (7, 5), d_6 = (6, 4), d_7 = (1, 2), d_8 = (4, 9)$.

Daraus ergibt sich für die Euklidische Distanz der Punkte

$$\sqrt{\sum_{i=1}^{n} |x_{i,1} - x_{i,2}|^2}$$

die folgende gegebene Abstandsmatrix:

$dist(d_i, d_j)$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0	$\sqrt{25}$	$\sqrt{72}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{4}$	$\sqrt{53}$	$\sqrt{41}$
4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
6						0	$\sqrt{29}$	$\sqrt{29}$
7							0	$\sqrt{58}$
8								0

Um die Distanzen in Ähnlichkeitswerte umzurechnen, kann man sich der Formel $sim(d_i, d_j) = e^{-dist(d_i,d_j)^2}$ bedienen, die Werte zwischen 0 und 1 für die Ähnlichkeiten liefert. Alternativ können wir auch die Distanzen für das Clustering benutzen.

Benutze nun den k-means-Algorithmus, um die Dokumente in 3 Cluster einzuordnen. Als initiale Seeds der Cluster sollen d_1, d_4 und d_7 herhalten. Bestimme am Ende jedes Durchlaufs die neuen Cluster, die Zentroiden und die neue Distanz- oder Ähnlichkeitsmatrix. Wie viele Iterationen des Algorithmus sind nötig, bis die Cluster stabil sind?

Aufgabe 16: Hierarchisches Clustering

Benutze agglomeratives Single-Link-, Complete-Link- und Average-Link-Clustering, um die Dokumente aus Aufgabe 15 zu clustern. Das Resultat eines hierarchischen Clusterings wird oft als sogenanntes Dendogramm dargestellt. Recherchiere, was man darunter versteht, und verwende es zur Darstellung Deiner Lösungen.

Zur Erinnerung:

- Beim Single-Link-Clustering werden in jedem Schritt die beiden Cluster mit der geringsten minimalen, paarweisen Distanz bzw. der größten maximalen, paarweisen Ähnlichkeit zusammengeführt (es werden die beiden einander nächsten bzw. ähnlichsten Punkte in den beiden Clustern betrachtet).
- Beim Complete-Link-Clustering werden in jedem Schritt die beiden Cluster mit der geringsten maximalen, paarweisen Distanz bzw. der größten minimalen, paarweisen Ähnlichkeit zusammengeführt (es werden die beiden einander entferntesten bzw. verschiedensten Punkte in den beiden Clustern betrachtet).
- Beim Average-Link-Clustering betrachtet man entsprechend den Durchschnitt der paarweisen Distanzen bzw. Ähnlichkeiten.