

Information Retrieval
Prof. Dr. N. Fuhr

Sascha Kriewel
sascha.kriewel@uni-duisburg.de

keine Abgabe,
Besprechung am 30. Mai 2005

Übungsblatt Nr. 5

Aufgabe 1: Clusterverfahren

Für eine umfangreiche Dokumentensammlung seien die Dokumente in fünf Gruppen klassifiziert worden. Nun sollen zur Präsentation für den Benutzer diese Dokumente in zwei größere Gruppen eingeteilt werden. Dazu hat man für die Repräsentanten der einzelnen Klassifikationen die folgenden Ähnlichkeiten d ermittelt:

$sim(x, y)$	1	2	3	4	5
1	0	2	2	17	16
2	2	0	4	9	10
3	2	4	0	13	10
4	17	9	13	0	1
5	16	10	10	1	0

- (a) Berechne die zwei Gruppen C_1 und C_2 mit Hilfe des hierarchischen single-linkage Verfahrens. Verwende zur Erstellung der Cluster die Verschiedenheitsfunktion

$$D(C_i, C_j) = \min_{x \in C_i, x \in C_j} sim(x, y)$$

- (b) Schreibe eine kleine Java-Anwendung, welche die Ähnlichkeitsmatrix z.B. aus einer komma-separierten Textdatei einliest, zu einem bestimmten α die entstehenden Cluster berechnet und dann die Cluster mit ihren Repräsentanten ausgibt. Ersetze nun die Verschiedenheitsfunktion durch andere Varianten aus der Vorlesung.
- (c) Wenn noch keine Ähnlichkeitsmatrix gegeben ist, muß diese zunächst berechnet werden. Seien die Dokumentvektoren für Repräsentanten von Dokumentgruppen wie folgt gegeben:

D_i	t_1	t_2	t_3	t_4	t_5	t_6	t_7
D_1	0	3	0	0	0	2	2
D_2	3	1	2	4	1	0	0
D_3	3	0	0	0	3	0	1
D_4	0	1	0	3	0	0	2
D_5	2	2	4	3	1	3	0

Wähle ein geeignetes Ähnlichkeitsmaß für Vektoren und erweitere Deine Anwendung um den Schritt der Berechnung der Ähnlichkeitsmatrix aus vorgegebenen Dokumentenvektoren. Führe dann ein Clustering für die angegebenen Repräsentanten durch.

Aufgabe 2: Probabilistisches Retrieval - BIR-Modell

Gegeben seien die folgenden simplifizierten Dokumente mit den Termen $a, b, c, d, e, f, g, h, i, j, k, l$:

$$\begin{aligned}
 d_1 &= a d i & d_6 &= a d h l \\
 d_2 &= a b i k & d_7 &= a c e f h j \\
 d_3 &= a d f i & d_8 &= a d e h \\
 d_4 &= a b c d e f g i l & d_9 &= a b c e f g j k l \\
 d_5 &= a b i j k l & d_{10} &= a c h k
 \end{aligned}$$

Zu den Anfragen $q_1 = (e, f, g)$, $q_2 = (d, e, f)$, $q_3 = (b, e, f, g)$, und $q_4 = (a, c, i)$ gibt der Benutzer folgende Relevanzbeurteilungen ab:

d_i	1	2	3	4	5	6	7	8	9	10
$r(q_1, d_i)$	\bar{R}	R	\bar{R}	R	\bar{R}	\bar{R}	R	\bar{R}	R	\bar{R}
$r(q_2, d_i)$	R	\bar{R}	R	R	\bar{R}	R	R	\bar{R}	\bar{R}	\bar{R}
$r(q_3, d_i)$	\bar{R}	R	\bar{R}	R	\bar{R}	\bar{R}	R	\bar{R}	R	R
$r(q_4, d_i)$	R	R	R	R	R	\bar{R}	\bar{R}	\bar{R}	R	\bar{R}

- (a) Berechne die Termgewichte c_{ik} .
- (b) In welcher Reihenfolge werden die Dokumente auf Grundlage dieser Werte gerankt?
- (c) Berechne die Wahrscheinlichkeiten $P(R|q, \vec{x})$ für die Dokumente auf den beiden möglichen Wegen (direkt/über das BIR-Modell) und vergleiche die Ergebnisse. Wodurch ist der Unterschied zu erklären?