Information Retrieval - Sommer 2006
Dipl.-Inform. Sascha Kriewel, LF 137
kriewel@is.informatik.uni-duisburg.de

Übungsblatt 7

Aufgabe 18: Probabilistisches Retrieval

Im Folgenden seien die Ergebnisse zweier Suchanfragen an ein probabilistisch arbeitendes IR-System zu sehen. In der ersten Spalte habe eine Sucherin die aus ihrer Sicht relevanten Dokumente zur Suchanfrage markiert. Die zweite Spalte enthält die Dokumentengewichtung als Grundlage des Ranking.

+	0,8	Nonclausal Temporal Deduction
+	0,8	A deduction model of belief and its logics
	0,8	Relevance Logic
+	0,8	Description Logics
	0,75	Automated Deduction for Deontic Logics
+	0,75	Automated Deduction in Nonclassical Logics
+	0,75	Nonclassical logics as generalized Galois logics
+	0,6	Intelligent Automated Deduction in Nonstandard logics
	0,6	Weak deduction mechanisms
+	0,5	Substructural Logics
	0,5	Automated Deduction for Categorial Grammar Logics
	0,3	A history of natural deduction
+	0,3	Mechanising Deduction in the Logics of Practical Reasoning
	0,1	Proof Theory of Finite-Value Logics
	0,1	Assignments: Logics for Computer Science
	0,1	Natural deduction

+	0,8	Probability Ranking Principle in IR
+	0,75	Optimum Polynomial Retrieval Functions Based on the PRR
+	0,6	Probability Spaces of Information Retrieval
	0,6	Probability theory and statistical inference
+	0,6	Lemur Retrieval Applications
+	0,5	Minimum Probability of error image retrieval
	0,5	Probability kinematics in information retrieval
	0,3	Introduction to Information Retrieval
	0,1	Cloud parameter retrieval
	0,1	Information Retrieval Interaction
	0,1	Tutorial: Web Information Retrieval

(a) Was bedeutet es konkret, dass ein Retrievalsystem „probabilistisch" arbeitet?

(b) Natürlich kann man anhand zweier Beispiele nicht ernsthaft statistisch argumentieren, aber wie kann man prinzipiell nachprüfen, ob die Retrievalgewichtungen tatsächlich probabilistisch sind? Überprüfe es

anhand der Beispiele? Was ist das Ergebnis?

(c) Was bewirkt Relevance Feedback?

Aufgabe 19: BIR-Modell

(a) Wiederhole das BIR-Modell. Wie wird es hergeleitet? Beschreibe die zugrunde liegenden Annahmen.

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

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

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	$\bar{R}$

(*i*) Berechne die Termgewichte c_{ik}.

(*ii*) In welcher Reihenfolge werden die Dokumente auf Grundlage dieser Werte gerankt?

(*iii*) 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?

Aufgabe 20: Hierarchisches Clustering
Abgabe bis 1. Juni 2006, 12 Uhr:
`iruebg-abgaben@is.informatik.uni-duisburg.de`

Für eine umfangreiche Dokumentenkollektion seien die Dokumente bereits in sieben Gruppen klassifiziert worden. Nun sollen diese Dokumente in zwei grössere Cluster eingeteilt werden. Dazu hat man für die Repräsentanten der einzelnen Klassifikationen die folgenden Abstände d ermittelt:

$dist(x,y)$	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	0	2	2	17	16	6	9
2	2	0	4	9	10	9	5
3	2	4	0	13	10	7	8
4	17	9	13	0	1	10	11
5	16	10	10	1	0	11	15
6	6	10	7	10	11	0	3
7	9	5	8	11	15	3	0

Es sei die Verschiedenheitsfunktion

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

gegeben.

(a) Schreibe eine kleine Java-Anwendung, welche die Ähnlichkeitsmatrix z.B. aus einer komma-separierten Textdatei einliest, die entstehenden Cluster durch hierarchisches Bottom-Up-Clustering berechnet und dann die Cluster mit ihren Repräsentanten ausgibt.

(b) Wenn noch keine Ähnlichkeits- oder Abstandsmatrix 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. Was ändert sich, wenn man statt Abständen Ähnlichkeiten betrachtet? Führe ein Clustering für die angegebenen Repräsentanten durch.