

Information Retrieval

Sascha Kriewel
 sascha.kriewel@uni-duisburg.de

Übungsblatt 8

Bearbeitung bis **20. Juni 2005**
keine Abgabe

Aufgabe 1: Probability Ranking Principle

Es seien zehn Dokumente d_i gegeben und die folgenden Schätzwerte für die Wahrscheinlichkeit der Relevanz $P(R|d_i) = \pi(d_i)$:

1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Der Erwartungswert für die Relevanz (ρ) läßt sich dann einfach als die Summe der $P(d_i|R)$ über die gegebenen Dokumente bestimmen, analog dazu der Erwartungswert der Nicht-Relevanz (π) mit $P(d_i|\bar{R}) = 1 - P(d_i|R_i)$. Komplexer sieht der Fall für den Wert der Precision (ϕ) aus. Überlege, warum das so ist.

Bestimme Expected Recall und Expected Fallout nach jedem Dokument, und übertrage die Werte in ein Diagramm. Wie würdest Du die Werte für Expected Precision berechnen?

Aufgabe 2: Least Square Polynomials

In dieser Aufgabe soll der LSP-Ansatz noch einmal an einem anderen Beispiel schrittweise nachvollzogen werden. Zu einer Menge von Frage-Dokument-Paaren mit Relevanzurteilen auf einer binären Skala werden folgende Beschreibungsvektoren \vec{x} erstellt:

\vec{x}^T	$r(q_k, d_m)$	\vec{y}^T
(1,1)	R_1	(1,0)
(1,1)	R_2	(0,1)
(1,1)	R_2	(0,1)
(1,0)	R_1	(1,0)
(1,0)	R_1	(1,0)
(1,0)	R_1	(1,0)
(1,0)	R_2	(0,1)
(0,1)	R_1	(1,0)
(0,1)	R_2	(0,1)

Wir betrachten im Folgenden die lineare Polynomstruktur $\vec{v} = (1, x_1, x_2)$.

- (a) Wie sieht die Retrievalfunktion $\vec{e}(\vec{x})$ aus?
- (b) Welche Werte liefert die Funktion \vec{e}_{opt} für die verschiedenen Vektoren \vec{x} ?

- (c) Erstelle wie in der Vorlesung gezeigt die Momentenmatrix $M = \begin{pmatrix} \vec{v} \cdot \vec{v}^T & \vec{v} \cdot \vec{y}^T \\ \vec{v} \cdot \vec{v}^T & \vec{v} \cdot \vec{y}^T \end{pmatrix}$, die rechte und linke Seiten des Gleichungssystems $E(\vec{v} \cdot \vec{v}^T) \cdot A = E(\vec{v} \cdot \vec{y}^T)$ enthält.
- (d) Berechne iterativ die Retrievalfunktionen $\vec{e}^{(i)}(\vec{x})$ für $i = 1, 2, 3$. Die Auswahl des jeweils nächsten zu berechnenden Koeffizienten soll nach dem Kriterium der maximalen Reduktion der Reststreuung erfolgt. Bezeichne $M^{(i)} = (m_{kj}^{(i)})$ mit $k=1, \dots, n; j=1, \dots, L+n$ die Matrix M vor dem i ten Lösungsschritt (L — Anzahl der Komponenten von \vec{v} , n — Anzahl der Relevanzstufen). Die durch die Auswahl der j ten Komponente von \vec{v} erreichbare Minderung der Reststreuung berechnet sich dann nach der Formel

$$d_j^{(i)} = \frac{1}{m_{jj}^{(i)^2}} \cdot \sum_{k=L+1}^{L+n} m_{jk}^{(i)^2}$$

(In der Retrievalfunktion werden die noch nicht berechneten Koeffizienten jeweils = 0 gesetzt).

- (e) Welche Eigenschaft besitzen jeweils die Vektorkomponenten von $\vec{e}(\vec{x})$?
- (f) Vergleiche bei jedem Schritt die Werte von $e_1^{(i)}(\vec{x})$ mit $P(R_1|\vec{x})$. Was läßt sich über die Folge der Funktionen $e_1^{(i)}(\vec{x})$ aussagen?
- f) Wir fügen nun als zusätzliches Element der Lernstichprobe den Vektor $\vec{x} = (0, 0)^T$ mit dem Relevanzurteil R_2 hinzu. Berechne hierzu die Retrievalfunktion $e_1^{(3)'}(\vec{x})$ und vergleiche die Funktionswerte mit den Werten von $P(R_1|\vec{x})$. Welcher wesentliche Unterschied ergibt sich gegenüber den Ergebnissen von Teilaufgabe (e)?