

**Information Retrieval**  
**Prof. Dr. N. Fuhr**

Sascha Kriewel  
sascha.kriewel@uni-duisburg.de

Abgabe bis 2. Juni 2004

**Übungsblatt Nr. 5**

**Aufgabe 1: Probabilistisches Clustering**

Es seien die folgenden Merkmalsvektoren zu Dokumenten  $d_1, d_2, d_3$  und  $d_4$  mit Termen  $t_1, t_2$  und  $t_3$  gegeben:

$$\begin{array}{ll} d_1 : (1,1,0) & d_2 : (1,0,1) \\ d_3 : (0,1,1) & d_4 : (1,0,0) \end{array}$$

Es sollen zwei Cluster gebildet werden, als *Seed*-Dokumente wähle  $d_1$  und  $d_3$ . Initialisiere die Parameter  $n^j = 1$  und  $p^j = 1/k$ . Berechne die  $q_i^j$  nach der verbesserten Schätzformel:

$$q_i^j = \frac{p^j + \sum_{d_m \in D} x_{m_i} \cdot P(C^j | x_m)}{n_j + 1}$$

Führe zwei Iterationen (Schritte 4 und 5) des vorgestellten EM-Algorithmus durch. Wieso ist es in diesem Beispiel wichtig gewesen, die verbesserte Schätzfunktion zu benutzen, statt der Schätzformel

$$q_i^j = \frac{\sum_{d_m \in D} x_{m_i} \cdot P(C^j | x_m)}{n_j}$$

?

10 Punkte

**Aufgabe 2: Suchaktivitäten und Systembeteiligung**

Suche Dir ein Recherchethema Deiner Wahl. Spezifiziere Deinen Informationswunsch, ohne dabei zu sehr in Details zu gehen: was ist das Suchthema, was möchtest Du mit der Suche erreichen, welche Art von Information oder Materialien erwartest Du als Ergebnis Deiner Suche? Entwickle und beschreibe dann eine Suchstrategie, nach der Du vorgehen möchtest. Identifiziere dabei verwendete Taktiken und Strategeme. Erkläre, warum Du Deine Suchstrategie derart strukturiert hast.

Führe anschliessend Deine Recherche mit Dir bekannten Quellen durch, z.B. Websuchmaschinen (URLs aus der Präsenzübung), CiteSeer (<http://citeseer.ist.psu.edu/cs>), Scirus (<http://www.scirus.com>), Suche der UB (letztes Übungsblatt), Daffodil (<http://www.daffodil.de>). Welche Art von Unterstützung (Ebenen nach Bates) bieten die gewählten Systeme dabei?

20 Punkte

### **Allgemeine Bemerkungen zum Übungsbetrieb**

Aufgrund des Feiertags fällt die Übung am Montag den 31. 5. 2004 aus. Die korrigierten Übungsaufgaben aus Woche 4 und 5 gibt es am nächsten Übungstermin, dem 7. 6. 2004 zurück. In der nächsten Woche gibt es wie gewohnt ein neues Übungsblatt.