School of Engineering

Institut für angewandte Informations-
technologie InIT

Information Engineering 1: Information Retrieval

Praktikum 4b: Rangierung

Relevanz-Rangierung

(Beispiel von Elke Mittendorf)

Es seien die folgenden Dokumente und Anfragen gegeben:

- q = „Terrorismus, bekämpfen“ = Φ1 Φ2
- D1 = „Gegenmassnahmen gegen Terrorismus“ = Φ3 Φ1
- D2 = „Kampf gegen den Terror“ = Φ2 Φ1
- D3 = „Sicherheit bei asymmetrischer Bedrohung und asymmetrische Sicherheit“ = Φ5 Φ6 Φ7 Φ6 Φ5
- D4 = „Terror bekämpfen“ = Φ1 Φ2
- D5 = „Extremismus und Gewalt“ = Φ8 Φ9
- D6 = „Terrorismus und innere Sicherheit“ = Φ1 Φ10 Φ5

Relevant erwiesen sich R(q)={D1, D2, D3, D6}

- Rangieren Sie die Informationsobjekte aus unserem Beispiel mit dem BIR (RSJ-Gewichtung). Diese Rangierung ist a posteriori (weiss zum Zeitpunkt des Vergleichs, welche Dokumente relevant sind).

$$RSV(q, d_j) := \sum_{\phi_i \in \Phi(q) \cap \Phi(d_j)} \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

  (zur Berechnung der p_i und q_i siehe BIR.pdf)

- Vergleichen Sie mit probabilistischer idf-Gewichtung

$$RSV(q, d_j) := \sum_{\varphi_i \in \Phi(q) \cap \Phi(d_j)} idf(\varphi_i)$$, mit

$$idf(\varphi_i) := \log\left(\frac{1+n}{1+df(\varphi_i)}\right)$$

- Vergleichen Sie mit tf.idf-Cosinus (siehe Folie 72)

Ein paar Anregungen zur Diskussion:

- Bleibt die Reihenfolge gleich?
- Ist immer das gleiche Dokument der beste Match?
- Was ist der fundamentale Unterschied zwischen BIR/RSJ und der probabilistischen idf-Gewichtung/tf.idf-Cosinus?
- Was ist mit D2 gegenüber D4?
- Was ist mit D1 gegenüber D6?
- Bonus-Verständnisfrage: was wäre die optimale Rangierung für das Beispiel?