Information Retrieval

Vektorraummodell

Stefan Birkner

7. September 2010

Vektorraummodell

- Keine exakten Treffer sondern zur Anfrage ähnliche Dokumente
- Grundannahme: Dokumente werden durch ihre Worte repräsentiert
- Keine Semantik („Bag of Words“)

Vektorraummodell

Jedes Dokument und die Anfrage werden als Vektor kodiert.

Häuser in Italien $D_1 = (1, 1, 0, 0)$

Häuser in Italien und um Italien $D_2 = (1, 2, 0, 0)$

Gärten und Häuser in Italien $D_3 = (1, 1, 1, 0)$

Gärten in Italien $D_4 = (0, 1, 1, 0)$

Gärten und Häuser in Frankreich $D_5 = (1, 0, 1, 1)$

Häuser in Italien $Q = (1, 1, 0, 0)$

Vektorraummodell

- Es wird die Ähnlichkeit zwischen jedem Dokument und der Anfrage berechnet.
- Dafür werden Ähnlichkeitsmaße verwendet.
- Beispiel: Skalarprodukt, Kosinus-Ähnlichkeitsmaß

Vektorraummodell

Beispiel: Skalarprodukt als Ähnlichkeitsmaß

$D \cdot Q = (d_1, d_2, d_3, d_4) \cdot (q_1, q_2, q_3, q_4) = d_1 \cdot q_1 + d_2 \cdot q_2 + d_3 \cdot q_3 + d_4 \cdot q_4$

- $D_1 \cdot Q = (1, 1, 0, 0) \cdot (1, 1, 0, 0) = 1 \cdot 1 + 1 \cdot 1 + 0 \cdot 0 + 0 \cdot 0 = 2$
- $D_2 \cdot Q = (1, 2, 0, 0) \cdot (1, 1, 0, 0) = 1 \cdot 1 + 2 \cdot 1 + 0 \cdot 0 + 0 \cdot 0 = 3$
- $D_3 \cdot Q = (1, 1, 1, 0) \cdot (1, 1, 0, 0) = 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 0 = 2$
- $D_4 \cdot Q = (1, 0, 1, 1) \cdot (1, 1, 0, 0) = 1 \cdot 1 + 0 \cdot 1 + 1 \cdot 0 + 1 \cdot 0 = 1$
- $D_5 \cdot Q = (1, 1, 0, 0) \cdot (1, 1, 0, 0) = 1 \cdot 1 + 1 \cdot 1 + 0 \cdot 0 + 0 \cdot 0 = 2$

Reihenfolge der Dokumente: D_2, D_1, D_3, D_5, D_4

Literatur

- Andreas Henrich: Information Retrieval 1
- `http://www.uni-bamberg.de/minf/ir1_buch/`
- Kapitel 5