Aufgabenblatt 6: Ähnlichkeitsmaße

Aufgabe 1 Interpretation von Ähnlichkeitsmaßen

Queryvektoren $Q = (q_1, ..., q_n)$ bzw. Dokumentvektoren $D = (d_1, ..., d_n)$, die nur aus Werten aus $\{0, 1\}$ bestehen, können als Repräsentation von Dokumenten durch Mengen von Indextermen aufgefasst werden. Interpretieren Sie für diesen Fall die folgenden Ähnlichkeitsfunktionen im Sinne von Mengenoperationen:

a) Das Skalarprodukt $s_k(Q, D) = \sum_{i=1}^{n} q_i d_i$

b) Den Jaccard-Koeffizienten $s_2(Q, D) = \frac{\sum_{i=1}^{n} q_i d_i}{\sum_{i=1}^{n} q_i + \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} q_i d_i}$

Wie lassen sich die Booleschen Operatoren AND und OR mit den beiden Ähnlichkeitsfunktionen in Verbindung bringen?

Aufgabe 2 Ähnlichkeitsfunktionen im Vektorraum

Berechnen Sie für die Zeilenvektoren der folgenden Matrix das Skalarprodukt und die fünf anderen in der Vorlesung angegebenen Ähnlichkeitsfunktionen mit dem Queryvektor Q und interpretieren Sie die Ergebnisse.

```
0 0 1 0 0 0
0 0 1 0 1 0
0 0 1 0 1 1
0 1 1 1 1 1
1 0 1 0 0 0
1 0 1 0 1 0
1 0 1 0 1 1
1 1 1 1 1 1
```

Queryvektor:

```
Q = 1 0 1 0 0 0
```

Aufgabe 3 Kurven gleicher Ähnlichkeit

Die Ähnlichkeitsmaße geben die Ähnlichkeit zweier Dokumentvektoren, bzw. eines Queryvektors und eines Dokumentvektors an. Umgekehrt kann man fragen, welche Vektoren zu einem gegebenen Vektor die gleiche Ähnlichkeit haben.

Untersuchen Sie diese Frage im zweidimensionalen Fall. Wählen Sie dazu die Vektoren $(1, 1)$, $(1, 2)$ und $(2, 1)$ und berechnen Sie zunächst jeweils das Ähnlichkeitsmaß zu sich selbst. Anschließend geben Sie für jeden der Vektoren die Menge derjenigen Vektoren an, die den selben Ähnlichkeitswert haben. Stellen Sie diese Menge graphisch in einem Koordinatensystem dar. Berechnen Sie die Aufgabe für

a) das Skalarprodukt,
b) das Cosinusmaß,
c) das Dice-Maß
d) das Overlap-Maß.

Aufgabe 4 Relevanz

a) Beschreiben Sie die Unterschiede zwischen dem umgangsprachlichen Gebrauch des Begriffs Relevanz (im Sinne von “wichtiges Dokument”) und der Definition von Relevanz, wie sie in der Vorlesung gegeben wurde.

b) Welche Einflussfaktoren bei der Informationssuche werden durch die Definition von Relevanz, wie sie im IR verwendet wird, im Allgemeinen nicht erfasst?