

openHPI: **Wie funktioniert eine Suchmaschine?**
Ranking: Vektorraummodell

Dr. Ralf Krestel

Hasso-Plattner-Institut

Universität Potsdam

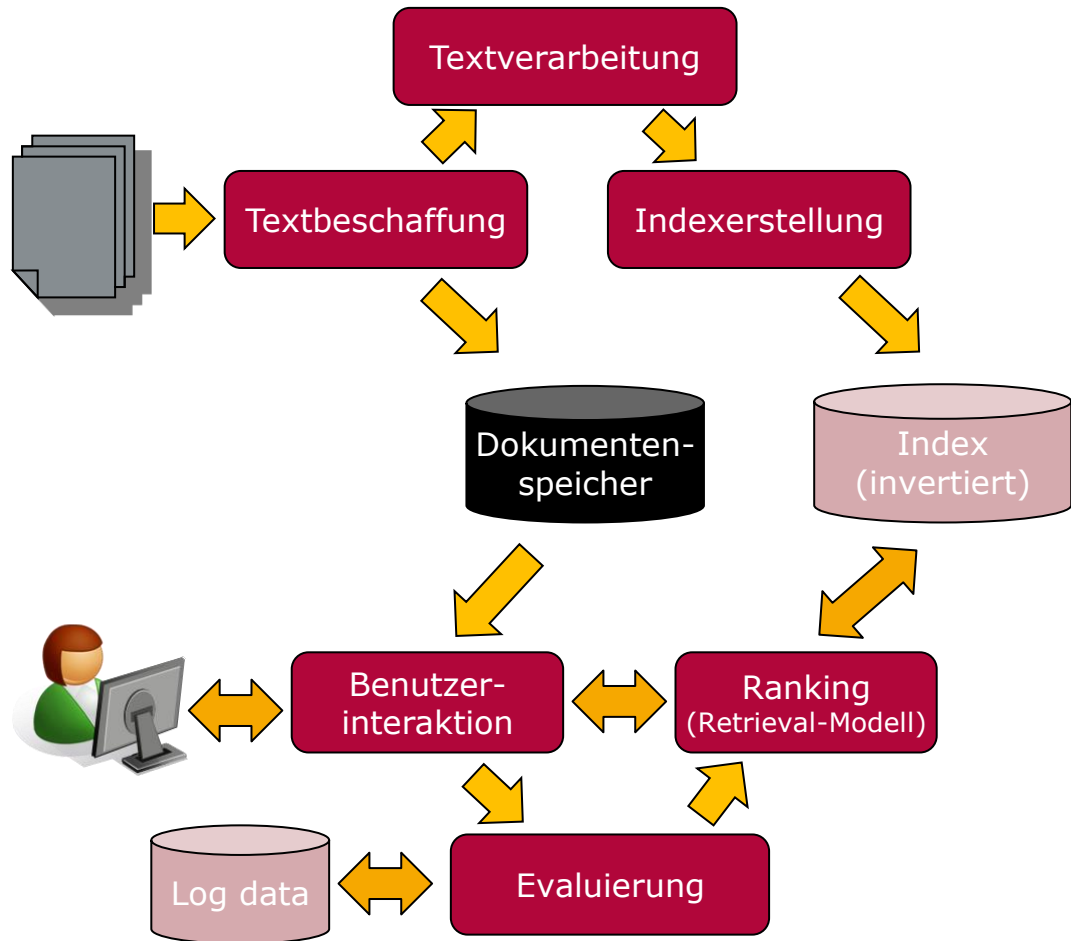
■ Grundlegende Bausteine

□ Indexierung

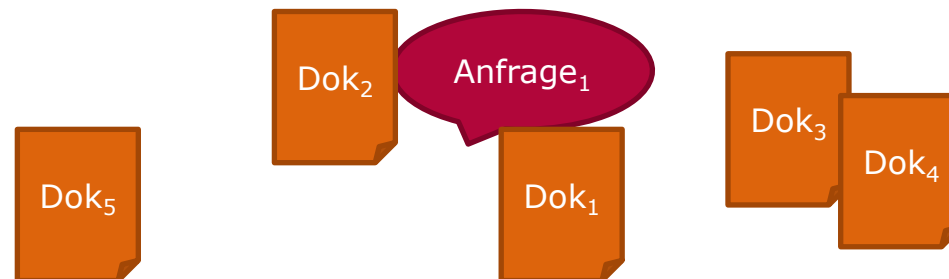
- Textbeschaffung
- Textverarbeitung
- Indexerstellung

□ Anfrage

- Benutzerinteraktion
- Ranking
- Evaluierung

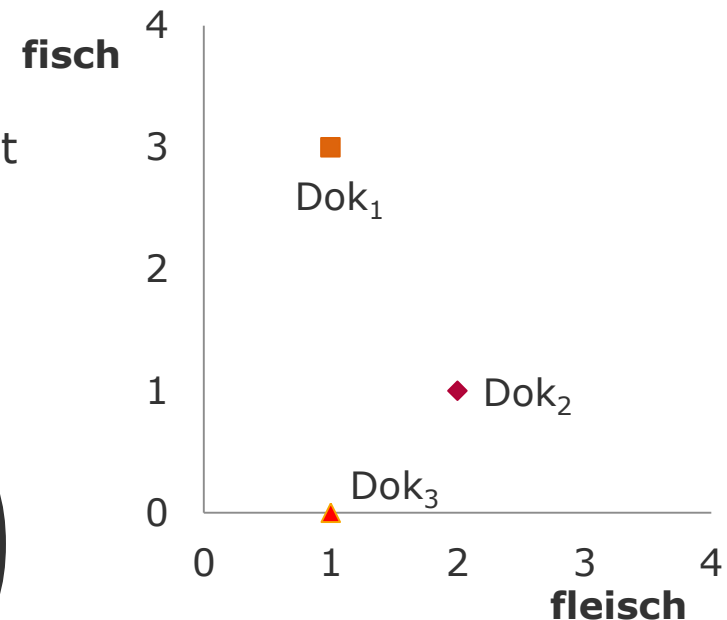


- Beispiel: Bibliothek
 - Thematisch ähnliche Bücher stehen räumlich nah beieinander
- Können wir das Konzept für Information Retrieval nutzen?
- Idee: Repräsentation von Dokumenten und Anfragen als Punkte in einem abstrakten Raum
 - Ähnlichkeit, bzw. Relevanz lässt sich dann durch die Entfernung im Raum messen



- Vorgesprochen von Gerard Salton 1975; Auch heute noch **sehr populär**
- Dokumente und Anfragen werden als Punkte im n -dimensionalen, reellen Vektorraum $\mathbb{R}^n$ dargestellt, wobei n der Größe des indexierten Vokabulars ist
 - Typischerweise ist n sehr groß ($>100,000$)
- Jeder Indexterm hat seine **eigene Dimension**
 - Terme können **gewichtet** werden
 - Einfachste Gewichtung: Anzahl der Vorkommen eines Terms in einem Dokument
- Dokumentensammlung kann als Matrix von Gewichten dargestellt werden
- Beispiel:
 - Dok_1 = "fleisch fisch fisch fisch"
 - Dok_2 = "fleisch fleisch fisch"
 - Dok_3 = "fleisch"

$$\begin{pmatrix} 1 & 3 \\ 2 & 1 \\ 1 & 0 \end{pmatrix}$$



■ Beispiel:

- $\text{Dok}_1 = \text{"fleisch fisch fisch fisch"}$
- $\text{Dok}_2 = \text{"fleisch fleisch fisch"}$
- $\text{Dok}_3 = \text{"fleisch"}$
- $\text{Anfrage}_1 = \text{"fisch"}$

■ Distanzmaß

- Euklidische Distanz → Dok_3 am nächsten
- Besser geeignet: **Kosinusähnlichkeit** → Dok_1 am nächsten

■ Exkurs: Berechnung des Kosinus zweier Vektoren

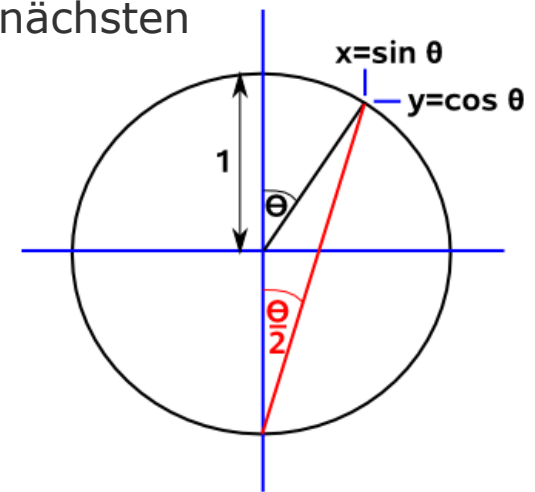
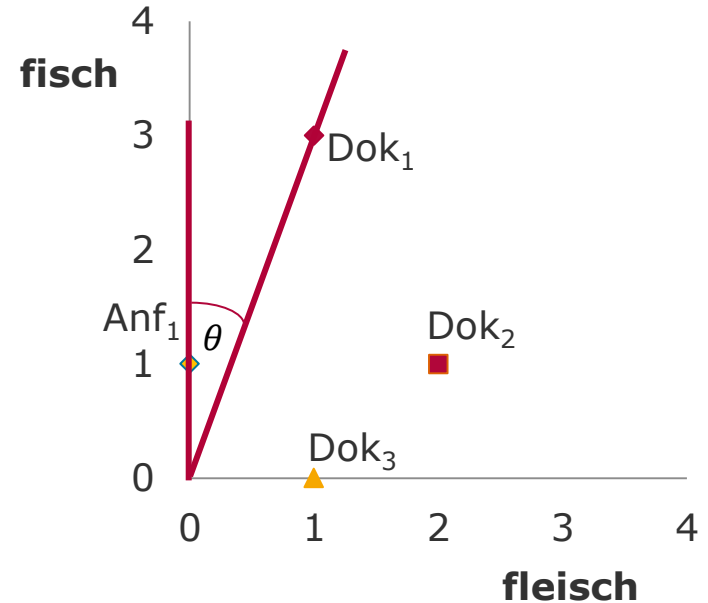
$$\square \cos(\theta) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

$$\square x \cdot y = \sum_{i=1}^n x_i y_i$$

Skalarprodukt

$$\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$$

Euklidische Norm



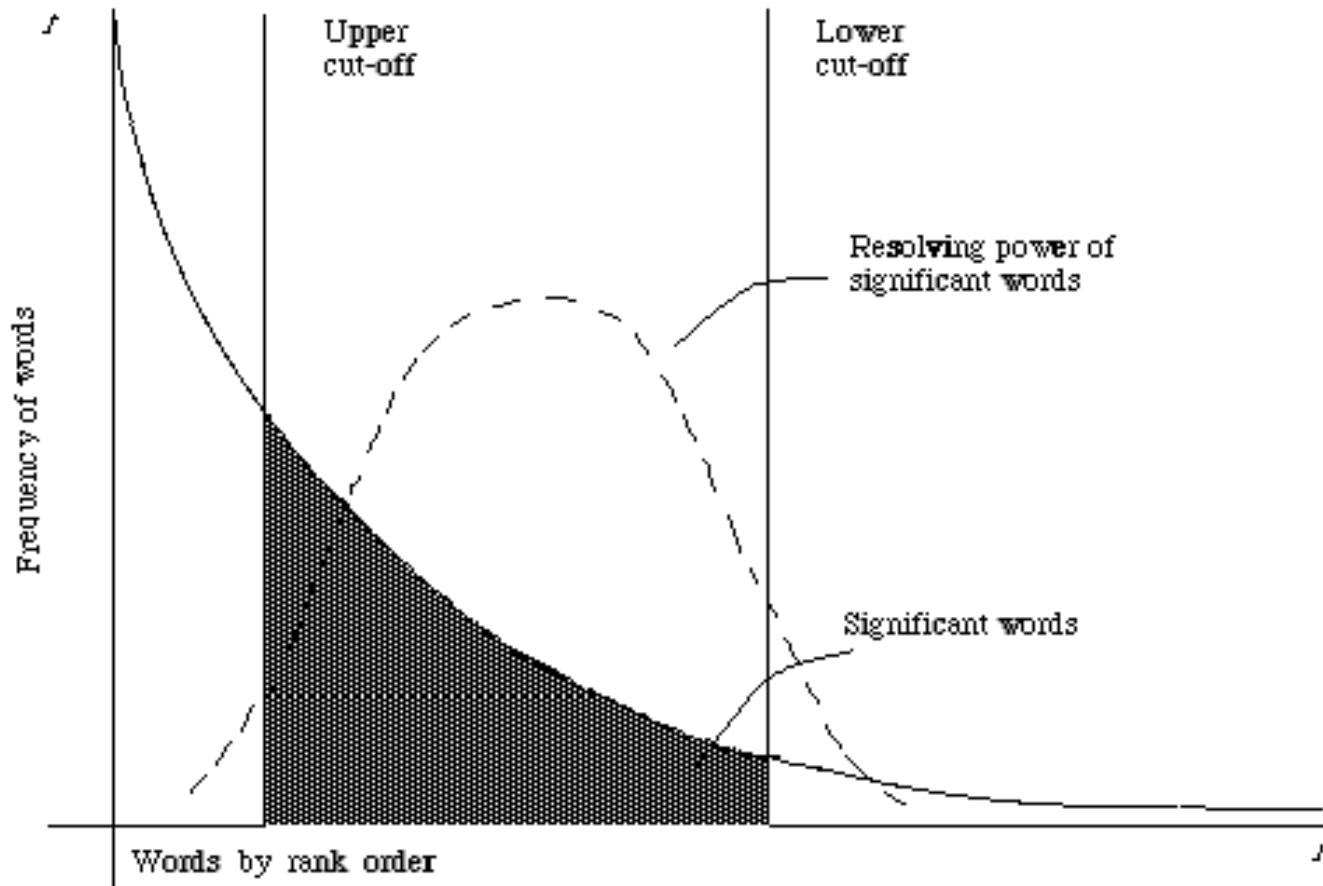
- „Wiederholung von Wörtern ist ein Zeichen der Betonung.“
(Hans Peter Luhn 1957)
 - Statt absoluter Anzahl der Wörter können wir die relative Anzahl nehmen
 - **Relative Vorkommenshäufigkeit** (term frequency)

$$tf(t, d) = \frac{\text{Häufigkeit von } t \text{ in } d}{\text{Anzahl Terme in } d}$$

- Nicht alle Wörter sind in einer Sammlung gleich wichtig
 - Das Wort „Psychologie“ kann in einer Sammlung von Informatikartikeln sehr aufschlussreich sein, in einer Sammlung von Psychologieartikeln eher nicht
 - Daher: Informationswert (**discriminative power**) eines Terms $disc(t)$
- Gewichtung: **Kombination** $tf(t, d) \cdot disc(t)$

Ranking: Vektorraummodell

Discriminative Power



Luhn 1958

- „Unterscheidung $disc(t)$ am besten anhand der Spezifität eines Terms“ (Karen Spärck Jones 1972)

- In wie vielen Dokumenten kommt ein Term vor?
- Inverse davon ist die Spezifität eines Terms

- **Inverse Dokumenthäufigkeit**

N =Anzahl Dokumente in Sammlung

$$idf(t) = \log \frac{N}{n_t}$$

n_t =Anzahl Dokumente, welche Term t enthalten

- Je spezifischer ein Term, desto wichtiger für die Suche

- Diese Kombination von tf und idf zur Gewichtung nennt man **TF*IDF**

- Verschiedene Varianten möglich
- Normalisierung erfolgt meist schon bei den Gewichten

+1 um Null-Gewichte zu vermeiden

$$w_{td} = \frac{(\log(f_{td}) + 1) \log \frac{N}{n_t}}{\sqrt{\sum_{t=1}^{T_d} \left((\log(f_{td}) + 1) \log \frac{N}{n_t} \right)^2}}$$

T_d =Anzahl Terme in Dokument d

- Einfach und klar
- **Intuitive Anfrage** (Menge von Wörtern)
 - Einfache Benutzbarkeit
- Ergibt ein **Ranking** nicht nur eine Menge
- **Anpassbar** auf unterschiedliche Szenarien/Dokumentsammlungen
 - Unterschiedliche Abstands/Ähnlichkeitsfunktionen
 - Unterschiedliche Normalisierungen
 - Unterschiedliche Gewichtungsmethoden
- **Sehr gute Suchergebnisse**



- Hochdimensionale Vektorräume benötigen spezialisierte Algorithmen
- Beruht auf implizite Annahmen, welche im Allgemeinen nicht gelten:
 - **Cluster-Hypothese:**
 - „Dokumente, die im Vektorraum eng beieinander liegen, sind für Anfragen gleich relevant“
 - **Unabhängigkeitsannahme:**
 - „Das Vorkommen eines Terms ist unabhängig von dem Vorhandensein anderer Terme im Dokument“



<https://www.youtube.com/watch?v=gYNUOkqzcVA>