

openHPI: **Wie funktioniert eine Suchmaschine?**
Interaktion: Anfragenpräzisierung

Dr. Ralf Krestel
Hasso-Plattner-Institut
Universität Potsdam

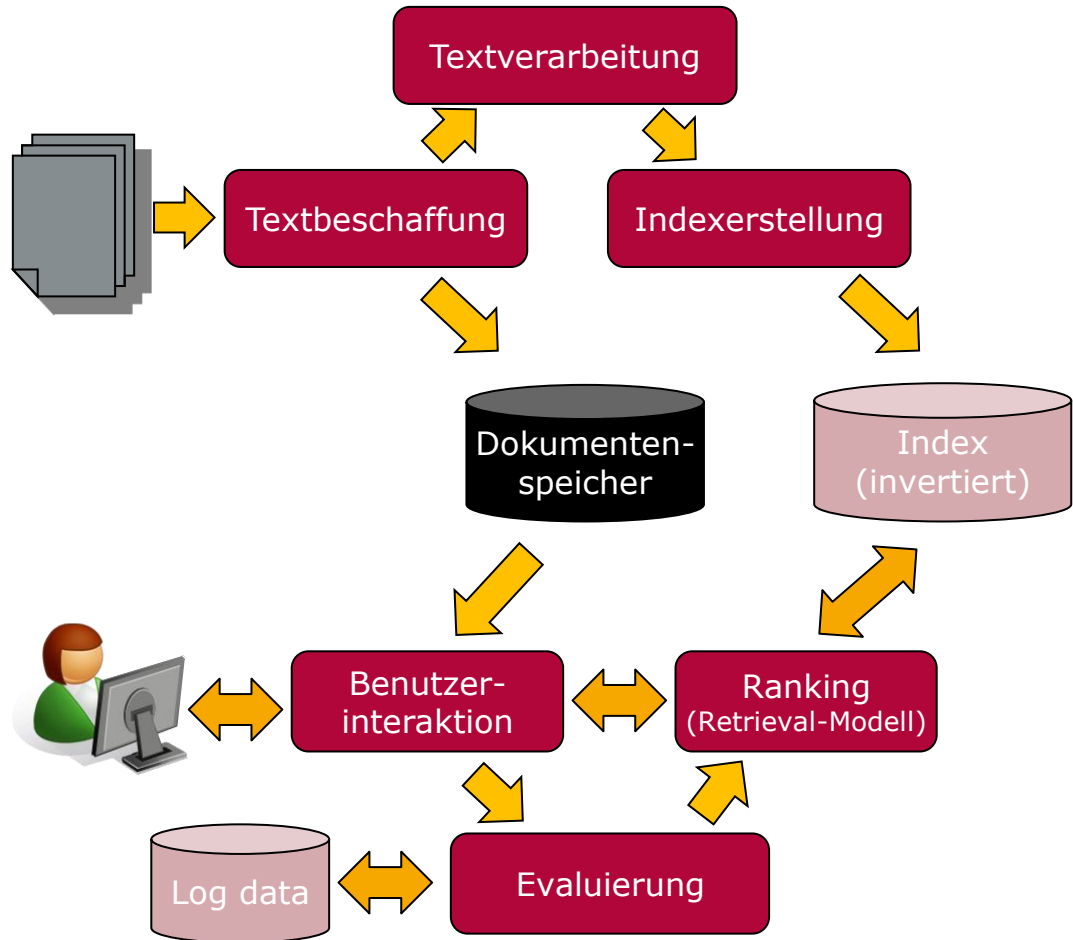
■ Grundlegende Bausteine

□ Indexierung

- Textbeschaffung
- Textverarbeitung
- Indexerstellung

□ Anfrage

- Benutzerinteraktion
- Ranking
- Evaluierung



- Anfragen oft mehrdeutig und ungenau
- Hilfe durch Suchmaschine, um Ergebnisse zu verbessern
 - **Ähnliche** Suchanfragen/Vorschlägen von Suchanfragen
 - Mit Hilfe von Log-dateien
 - **Personalisierung** und **Lokalität**
 - Ausnutzen zusätzlicher Informationen über Nutzer
 - Anpassen der Anfrage durch den Nutzer (**query refinement**)
 - Erfordert Erfahrung und Können des Nutzers
 - Anfrageerweiterung (**query expansion**)
 - Hinzufügen von Stichwörtern zu einer Anfrage
 - **Relevanz-Feedback**
 - Durch Nutzer: Eye-Tracking, Mausbewegung, Klicks
 - **Pseudo-Relevanz-Feedback**
 - Relevanzlernen ohne Nutzerinteraktion

- Einsatz in frühen Suchmaschinen
 - Von Hand erstellte Begriffshierarchie
 - Kontrolliertes Vokabular

- Einsatz zur Anfrageerweiterung
 - Ergänzen von Synonymen
 - Ober-/Unterbegriffe
 - Verbessert die Sucheffektivität

- Beispiel: MeSH-Thesaurus
 - Medical Subject Headings

Tree Number(s): C10.597.617.576, C23.888.592.612.553

MeSH Unique ID: D019547

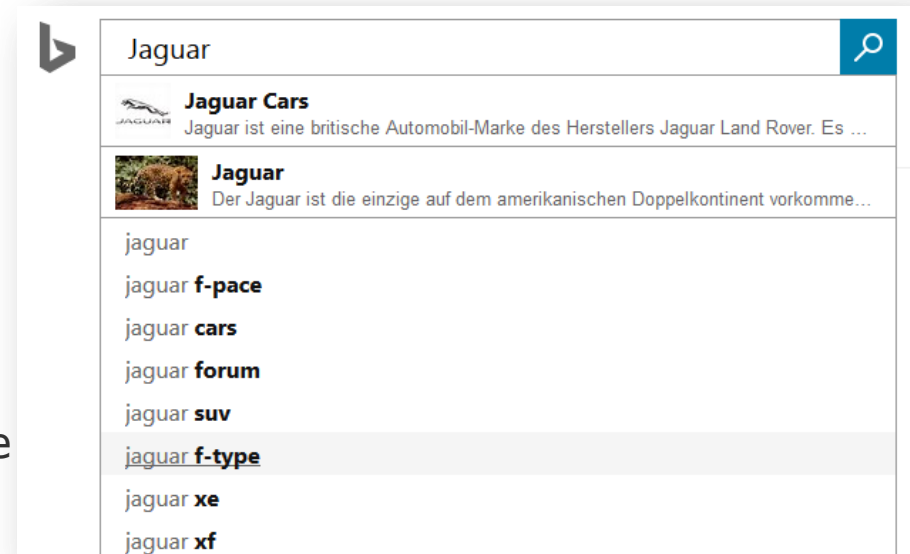
Entry Terms:

- Neck Pains
- Pain, Neck
- Pains, Neck
- Neck Ache
- Ache, Neck
- Aches, Neck
- Neck Aches
- Cervicalgia
- Cervicalgias
- Cervicodynia
- Cervicodynias
- Neckache
- Neckaches
- Cervical Pain
- Cervical Pains
- Pain, Cervical
- Pains, Cervical
- Posterior Cervical Pain

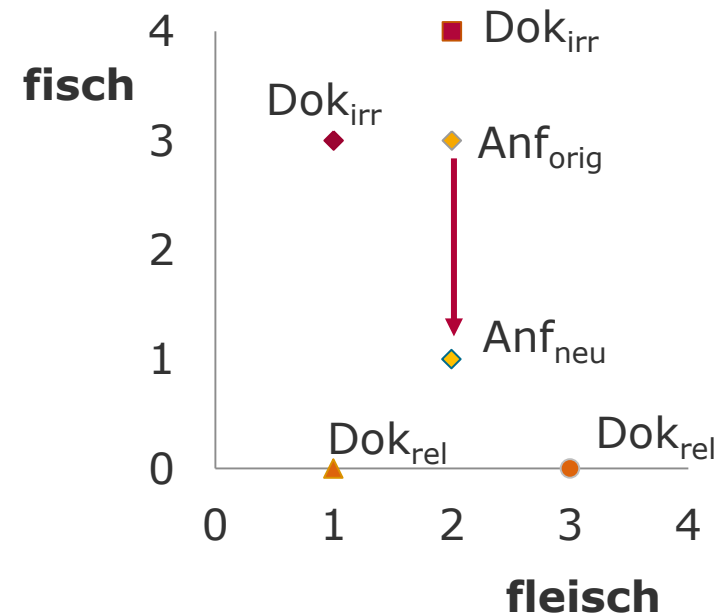
Interaktion: Anfragenpräzisierung

Anfrageerweiterung: Halbautomatisch

- Heutzutage selten Benutzung eines Thesaurus
 - Für allgemeine Suche nicht sehr hilfreich
- Automatische und halbautomatische Verfahren
 - Ziel: Verbessern der Trefferquote durch finden **verwandter Begriffe**
 - Halbautomatische Techniken verlangen **Auswahl vom Nutzer**
- 30% der Yahoo-Nutzer klicken auf Erweiterungsvorschlag
- Ansätze basieren auf der Analyse von **gemeinsam auftretenden Termen**
 - im gesamten Korpus,
 - in Sammlungen von Anfragen oder
 - in den höchst gelisteten Dokumente eines Suchanfrageergebnisses



- **Benutzer kennzeichnet** relevante (und/oder irrelevante) Dokumente in der erhaltenen Ergebnisliste
- **Suchmaschine passt** die Anfrage mit Hilfe von Termen aus den gekennzeichneten Dokumenten **an**
- Allgemeine Idee:
 - Anfragen werden durch Begriffe erweitert, welche häufig in relevant-markierten Dokumenten vorkommen
 - Oder bekommen ein höheres Gewicht
- Relevanz-Feedback im Vektorraum-Modell:
Rocchio's Algorithmus (Rocchio 1971)
 - Idee: Die Anfrage im Vektorraum ist ein Punkt, welcher verschoben wird
 - In Richtung der relevanten Dokumente
 - Weg von den Irrelevanten



- Pseudo-Relevanz-Feedback geht davon aus, dass die **obersten Treffer** der Ergebnisliste relevante Dokumente sind
 - **Kein Feedback** des Nutzers notwendig
 - Erweiterungsterme hängen von der gesamten originalen Anfrage ab, für welche die Ergebnisliste erstellt wurde
 - Qualität hängt von der Anzahl der **relevanten Treffer** innerhalb der obersten Ergebnisse ab
- Problem falls wenig relevante Dokumente in Top-10
 - Z.B. „Jaguar“ → 9/10 Auto, 1/10 Katze
 - Verschiebung der Bedeutung (**topic drift**)



- In vielen Fällen effektiv, aber nur sehr selten direkt eingesetzt
 - Suchmaschinennutzer haben sich auf irrelevante Ergebnisse eingestellt
 - Diversifizierung von Ergebnislisten, um jedem was zu bieten

- Indirekt wird es aber von einigen Anwendungen verwendet
 - „more like this“, „similar pages“
 - Zum Aufbau von Nutzerprofilen (Personalisierung)

- Vorschlagen von Anfragen ist populärer
 - Weniger genau, funktioniert aber auch, falls originale Anfrage scheitert
 - Annahme: Nutzer sucht viele relevante Dokumente
 - Sonst wären die ersten Ergebnisse schon ausreichend