

openHPI: **Wie funktioniert eine Suchmaschine?**
Wie Gut sind die Ergebnisse?

Dr. Ralf Krestel
Hasso-Plattner-Institut
Universität Potsdam

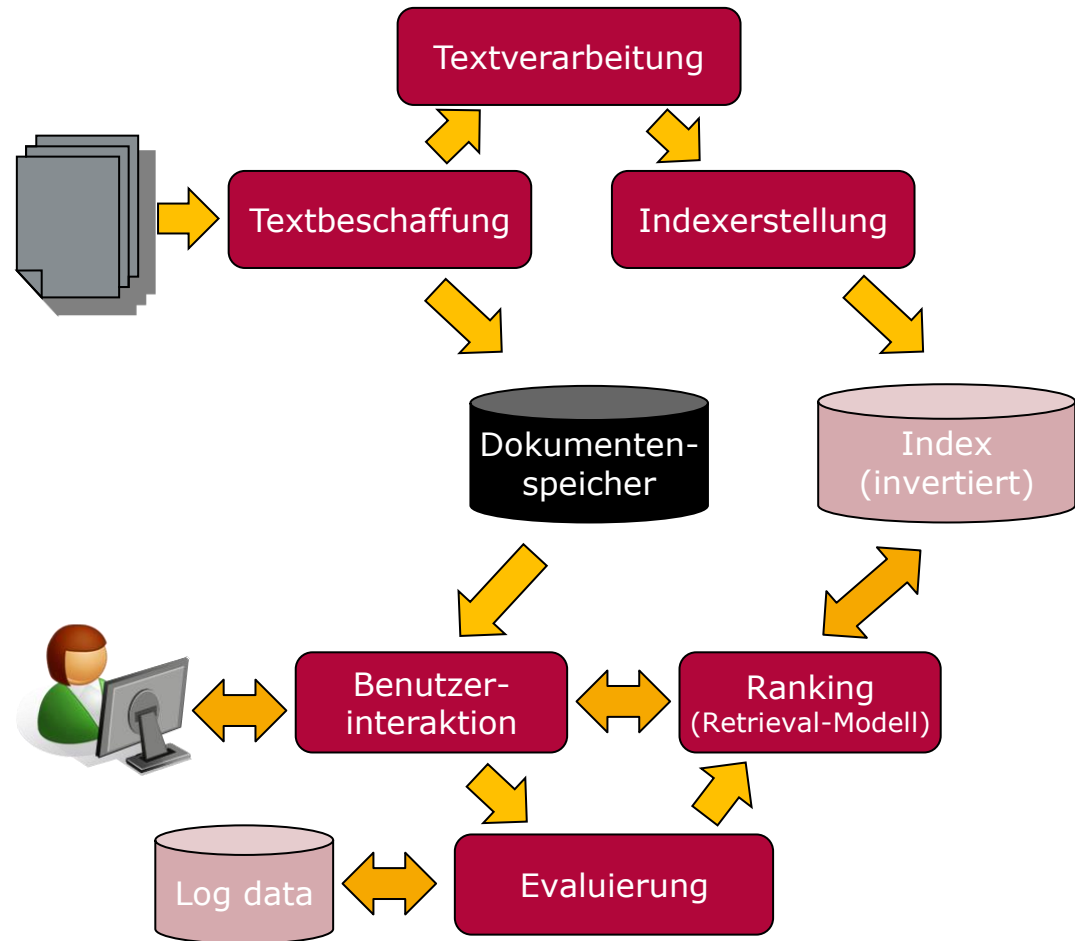
■ Grundlegende Bausteine

□ Indexierung

- Textbeschaffung
- Textverarbeitung
- Indexerstellung

□ Anfrage

- Benutzerinteraktion
- Ranking
- Evaluierung



Wie Gut sind die Ergebnisse?

Was ist Wichtig bei einer Suchmaschine?

- Qualität der Ergebnisse (Effektivität)
- Geschwindigkeit (Effizienz)
- Bedienbarkeit (Design)



www.jolyon.co.uk

<https://www.webpagefx.com/blog/wp-content/uploads/2009/04/havenworks.jpg>
<http://www.greenbookblog.org/wp-content/uploads/2013/01/needle.jpg>

Wie Gut sind die Ergebnisse?

Wie kann ich das Messen?

- Bedienbarkeit (Design)
 - Benutzerbefragung
 - A/B-Tests
 - Eye-Tracking
 - Klickdaten (wo, wann, wie lange, wie oft, ...)
 - ...
- Geschwindigkeit (Effizienz)
 - Indexierungsdauer
 - Indexgröße
 - Durchschnittliche **Antwortzeit** für eine Anfrage
 - **Anfragendurchsatz**
 - Wie viele Anfragen können pro Sekunde bearbeitet werden

- Populärste Metrik
- Verdeutlicht mehrere mögliche Problemfelder
 - **Kapazitätsplanung**: Ist mehr Hardware notwendig?
 - Erfüllt das System die gesetzten **Anforderungen**?
- **Aber**: Antwortzeiten werden nicht Berücksichtigt
 - Weniger als 150ms (=keine Verzögerung)
- **Antwortzeit** und **Durchsatz** sind zwei sich **entgegenstehende** Ziele
 - Privater Koch vs. Restaurant
 - Wenn einzelne Anfragen verzögert werden können, kann das System auf Durchsatz optimiert werden
 - Anfragen umsortieren, für höheren Durchsatz
- Für Suchmaschinen ist der Durchsatz **keine** einstellbare Variable
 - Alle Anfragen **müssen** beantwortet werden
 - Optimieren der **Antwortzeit** und der **Hardwarekosten**

Wie Gut sind die Ergebnisse?

Messung der Qualität I

- Klassischer Ansatz: **Benchmarks**
- Testsuite bestehen aus
 - **Dokumenten**
 - **Anfragen**
 - **Relevanzurteile** („gold standard“, „ground truth“)
 - (Einheitliches Evaluationsmaß zur Vergleichbarkeit)
- Erste Benchmark für Information Retrieval: Cranfield Experimente 1960er
- Seit den 1990: TREC (**T**ext **RE**trieval **C**onference)
 - Jährliche Wettbewerbe
 - Unterschiedliche IR-Aufgaben (Blogs, Spam, Patente, ...)
 - Motor des Fortschritts im IR-Bereich

Wie Gut sind die Ergebnisse?

Messung der Qualität II

- Falls Zugriff auf eine populäre Suchmaschine besteht:
 - **Online-Messung** während Live-Betrieb
 - Z.B. A/B-Tests
 - Auswertung der **Log-Dateien**
- Kann zum **Tuning** und zur **Evaluation** genutzt werden
- Viel mehr Anfragen als in einer statischen Testsuite
 - Aber **ungenauer** (Klicks sind **keine Relevanzurteile**)
- Privatsphäre der Nutzer
 - Daten können nicht weitergegeben werden (**Keine Vergleichbarkeit**)
- Typischer Inhalt
 - Benutzer ID oder Session ID (login, toolbar, cookies,...)
 - Anfragerterme eines Benutzers
 - Gelieferte Ergebnisliste + Klicks
 - Zeitstempel für alle Nutzeraktionen

Wie Gut sind die Ergebnisse?

Populäre Qualitätsmaße I

- A sei die Menge aller relevanten Dokumente
 - Von denen wir möglicherweise nicht alle finden...
- B sei die Menge der gefundenen Dokumente
 - Von denen aber wahrscheinlich nicht alle relevant sind...

	Relevant	Nicht Relevant
Gefunden	$A \cap B$	$\bar{A} \cap B$
Nicht Gefunden	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$

- **Recall** $R = \frac{|A \cap B|}{|A|}$
- **Precision** $P = \frac{|A \cap B|}{|B|}$
- Funktioniert für Boolesche Suche
- Annahme: Wir sind an **allen** relevanten Dokumenten interessiert

Wie Gut sind die Ergebnisse?

Populäre Qualitätsmaße II

- Suchalgorithmus der Precision maximiert
 - Kein Ergebnis zurückliefern!
- Suchalgorithmus der Recall maximiert
 - Alles zurückliefern!
- Beides wichtig
 - **F-Maß** maximieren

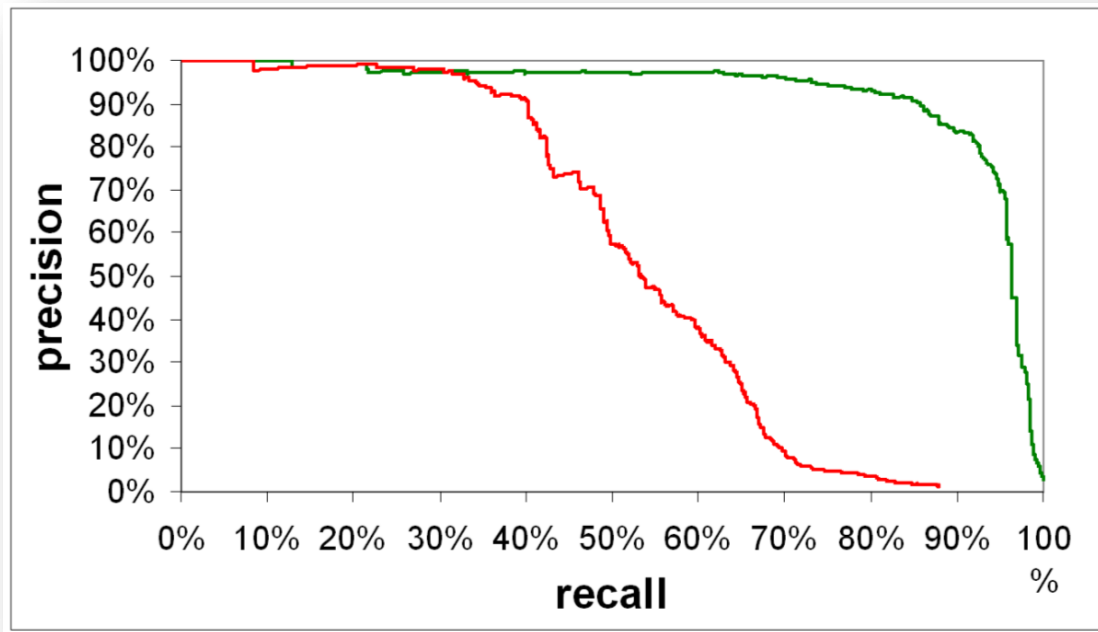
$$F = \frac{1}{\frac{1}{2} \left(\frac{1}{R} + \frac{1}{P} \right)} = \frac{2RP}{R + P}$$

Wie Gut sind die Ergebnisse?

Populäre Qualitätsmaße III

- Für Rankings

- **Durchschnittlicher Precision-Recall-Graph**



- **MRR** (**M**ean **R**eciprocal **R**ank)
- **MAP** (**M**ean **A**verage **P**recision)
- **NDCG** (**N**ormalized **D**iscounted **C**umulative **G**ain)