

openHPI: **Wie funktioniert eine Suchmaschine?**
Ranking: Boolesches Modell

Dr. Ralf Krestel

Hasso-Plattner-Institut

Universität Potsdam

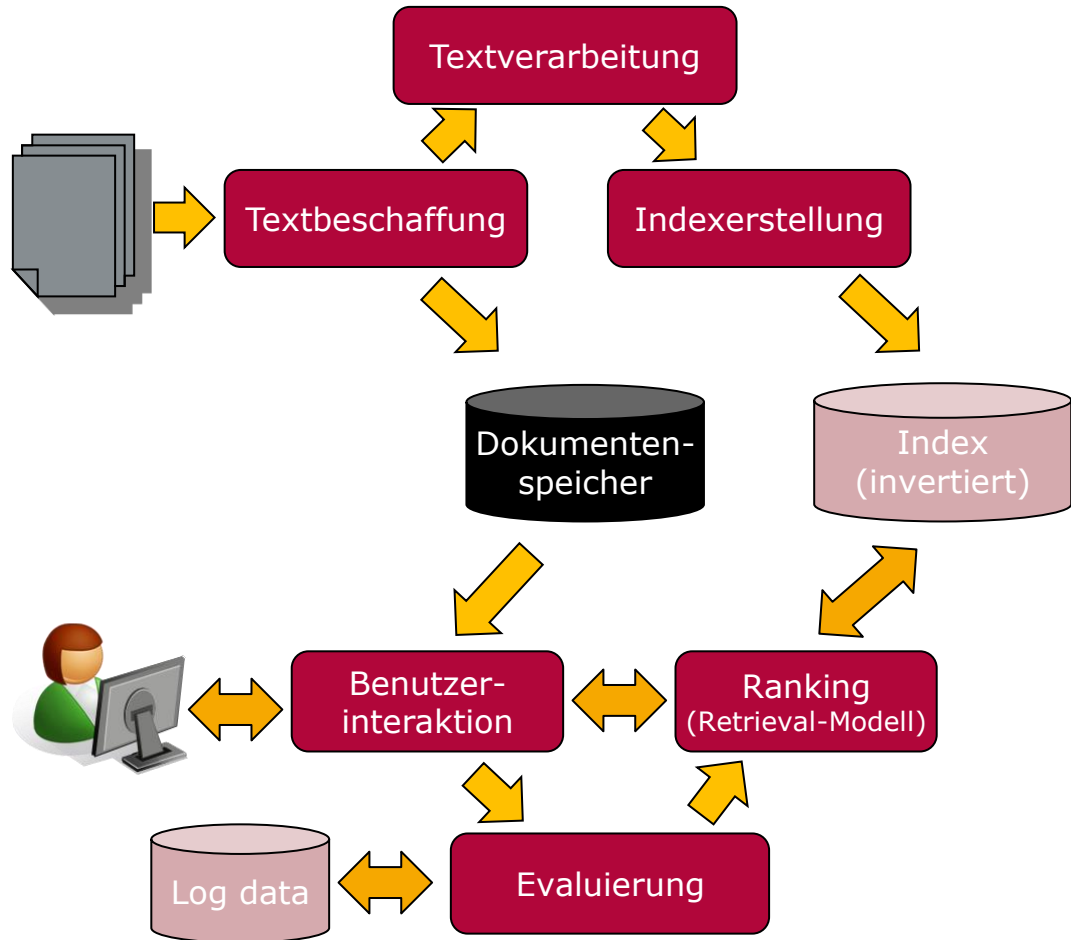
■ Grundlegende Bausteine

□ Indexierung

- Textbeschaffung
- Textverarbeitung
- Indexerstellung

□ Anfrage

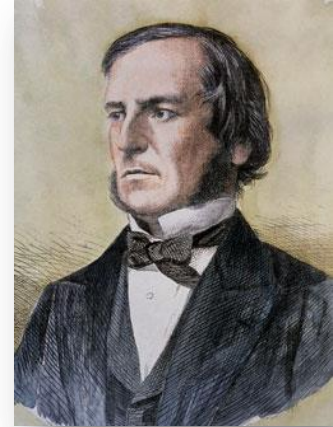
- Benutzerinteraktion
- Ranking
- Evaluierung



- Ein **Retrieval-Modell** ist ein mathematisches Gerüst, um den Ranking-Vorgang formal zu beschreiben
 - Wie würde ein Benutzer die Relevanz eines Dokuments abschätzen?
 - Modell sollte mit Benutzereinschätzung korrelieren
 - Grundlage vieler Ranking-Algorithmen
 - Kann implizit gegeben sein

- Theorie über **Relevanz**
 - Komplexes Konzept mit vielen zu berücksichtigenden Faktoren
 - Benutzer oft uneinig über Relevanzabschätzungen (inter-annotator disagreement)
 - Retrieval-Modelle treffen Annahmen zur Vereinfachung
 - Thematische vs. Benutzer-Relevanz

- Einfachstes und ältestes Retrieval-Modell
- Dokumente = Menge (nicht Multimenge!) von Wörtern
- Arbeitet auf dem einfachsten Index
- Anfragesprache = **Boolesche Ausdrücke** auf indexierten Termen
- Nur **binäres Ranking** (relevant/nicht relevant)
 - Alle Dokumente, die auf die Anfrage passen, sind **gleich relevant**
- Das Finden basiert auf der Zugehörigkeit der Dokumente zu einer Menge
 - „Finde alle Dokumente, die durch das Wort „Fisch“ indexiert wurden!“
 - „Finde alle Dokumente, die durch das Wort „Fisch“ oder „Fleisch“ indexiert wurden!“
 - WAHR oder FALSCH
 - „Exact-Match“-Semantik



George Boole
(1815–1864)

https://upload.wikimedia.org/wikipedia/commons/c/ce/George_Boole_color.jpg

■ Verknüpfungen

- Konjunktion AND $\wedge$
- Disjunktion OR $\vee$
- Negation NOT $\neg$

■ Wahrheitstabeln

$\wedge$	0	1
0	0	0
1	0	1

$\vee$	0	1
0	0	1
1	1	1

$\neg$	
0	1
1	0

- Können durch Klammerung beliebig kombiniert werden

Ranking: Boolesches Modell

Beispiel

d_1 Aquarium Berlin	Aquarium Berlin Das Aquarium Berlin besteht seit 1913 . Beispielsweise stehen in der Krokodilhalle Pflanzen des tropischen Regenwaldes .	aquarium berlin aquarium berlin bestehen beispielsweise stehen krokodilhalle pflanze tropisch regenwald
d_2 Aquarium	Aquarium Das Aquarium (v . lat . Aqua „ Wasser “) ist die am weitesten verbreitete Art des Vivariums . Bei den meisten Aquarien handelt es sich um tropische Süßwasseraquarien oder Meerwasseraquarien . Links : Berliner Aquarium	aquarium aquarium lat aqua wasser weit verbreiten art vivariums meist aquarium handeln tropisch süßwasseraquarium meerwasseraquarium links berlin aquarium

aqua	2		meerwasseraquarium	2	
aquarium	1	2	meist	2	
art	2		pflanze	1	
beispielsweise	1		regenwald	1	
berlin	1	2	stehen	1	
bestehen	1		süßwasseraquarium	2	
handeln	2		tropisch	1	2
krokodilhalle	1		verbreiten	2	
lat	2		vivarium	2	
links	2		wasser	2	
			weit	2	

- Anfrage $q_1 = \text{“aquarium AND pflanzen“}$
 - Ergebnismenge = $\{d_1\}$
- Anfrage $q_2 = \text{“vivarium OR pflanzen“}$
 - Ergebnismenge = $\{d_1, d_2\}$

- **Vorsicht!** Exklusive Nutzung des Negationsoperators führt meistens zu sehr großen Ergebnismengen
 - Anfrage $q_3 = \text{"NOT aquarium"}$
- „**BUT NOT**“ kann statt „AND NOT“ benutzt werden, was eher unserem Sprachgefühl entspricht
 - Anfrage $q_4 = \text{"aquarium BUT NOT vivarium"}$
- „**In der Nähe von**“-Suchen sind möglich
 - Anfrage $q_5 = \text{"aquarium /s berlin"}$
- **Platzhalter** (Wildcard)-Anfragen sind möglich
 - Anfrage $q_6 = \text{"aqua*"}$
- Durch „**OF**“ werden Suchen von Untermengen bestimmter Größe ermöglicht
 - Anfrage $q_7 = \text{"2 OF \{aquarium, vivarium, berlin\}"}$
 - Anfrage $q_8 = \text{"(aquarium AND vivarium) OR (aquarium AND berlin) OR (vivarium AND berlin)"}$

- Einfaches Modell, **einfach zu verstehen**, vorhersehbare Ergebnisse
- Viele Features können mit einbezogen werden
 - Datum, Dokumenttyp,...
- Komplexe Anfragen können effizient bearbeitet werden
- Falls alle Dokumentrepräsentationen sich unterscheiden, kann **jede beliebige Untermenge von Dokumenten** mit einer geeigneten Anfrage gefunden werden
 - Leider nur theoretischer Vorteil, da die „geeignete“ Anfrage meistens unbekannt ist



Ranking: Boolesches Modell

Suchen nach Zahlen

■ Informationsbedürfnis: Zeitungsartikel zu Präsident Lincoln

1. Anfrage q_9 = "lincoln"

- Ergebnis: Autos, Orte, Leute,...

2. Anfrage q_{10} = "president AND lincoln"

- Ergebnis: "Ford Motor Company today announced that Darryl Hazel will succeed Brian Kelley as president of Lincoln Mercury."

3. Anfrage q_{11} = "president AND lincoln AND NOT (automobile)"

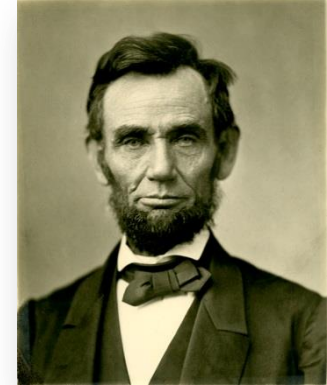
- Nicht im Ergebnis: "President Lincoln's body was in a nine-car funeral train."

4. Anfrage q_{12} = "president AND lincoln AND biography AND birthplace AND gettysburg AND NOT (automobile)"

- Ergebnis: $\emptyset$

5. Anfrage q_{13} = "president AND lincoln AND biography OR birthplace OR gettysburg AND NOT (automobile OR car)"

- Ergebnis könnte sein: "President Lincoln Read the entire book online! Abraham Lincoln Rese..."



Abraham Lincoln
(1809–1865)



- Schwer die **Größe der Ergebnisliste** zu kontrollieren
- **Kein Ranking!**
Nur ungeordnete Menge von Ergebnissen
- Effektivität hängt komplett vom Benutzer ab
 - Einfache Anfragen liefern meist keine guten Ergebnisse
 - Komplexe Anfragen sind schwer zu erstellen
- Meistens sind alle relevanten Dokumente auch relevant.
 - Aber: Viele **relevante Dokumente werden nicht gefunden**
- Keine Unterstützung für **Ähnlichkeitssuche**
- Wird von **Domainexperten** eingesetzt, z.B. Juristen, Westlaw
 - Are there any cases which discuss negligent maintenance or failure to maintain aids to navigation such as lights, buoys, or channel markers?
 - NOT NEGLECT! FAIL! NEGLIG! /5 MAINT! REPAIR! /P NAVIGAT! /5 AID EQUIP! LIGHT BUOY "CHANNEL MARKER"



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