

# Wie funktioniert eine Suchmaschine?

---

Sommercamp Juli 2024

Friedrich-Schiller-Universität Jena

[maik.fröbe@uni-jena.de](mailto:maik.fröbe@uni-jena.de)

[heinrich.reimer@uni-jena.de](mailto:heinrich.reimer@uni-jena.de)

[ines.zelch@uni-jena.de](mailto:ines.zelch@uni-jena.de)

Suchmaschine? Welche war denn die erste?

Eine von diesen hier?



bing

Google



DuckDuckGo

amazon

Яндекс

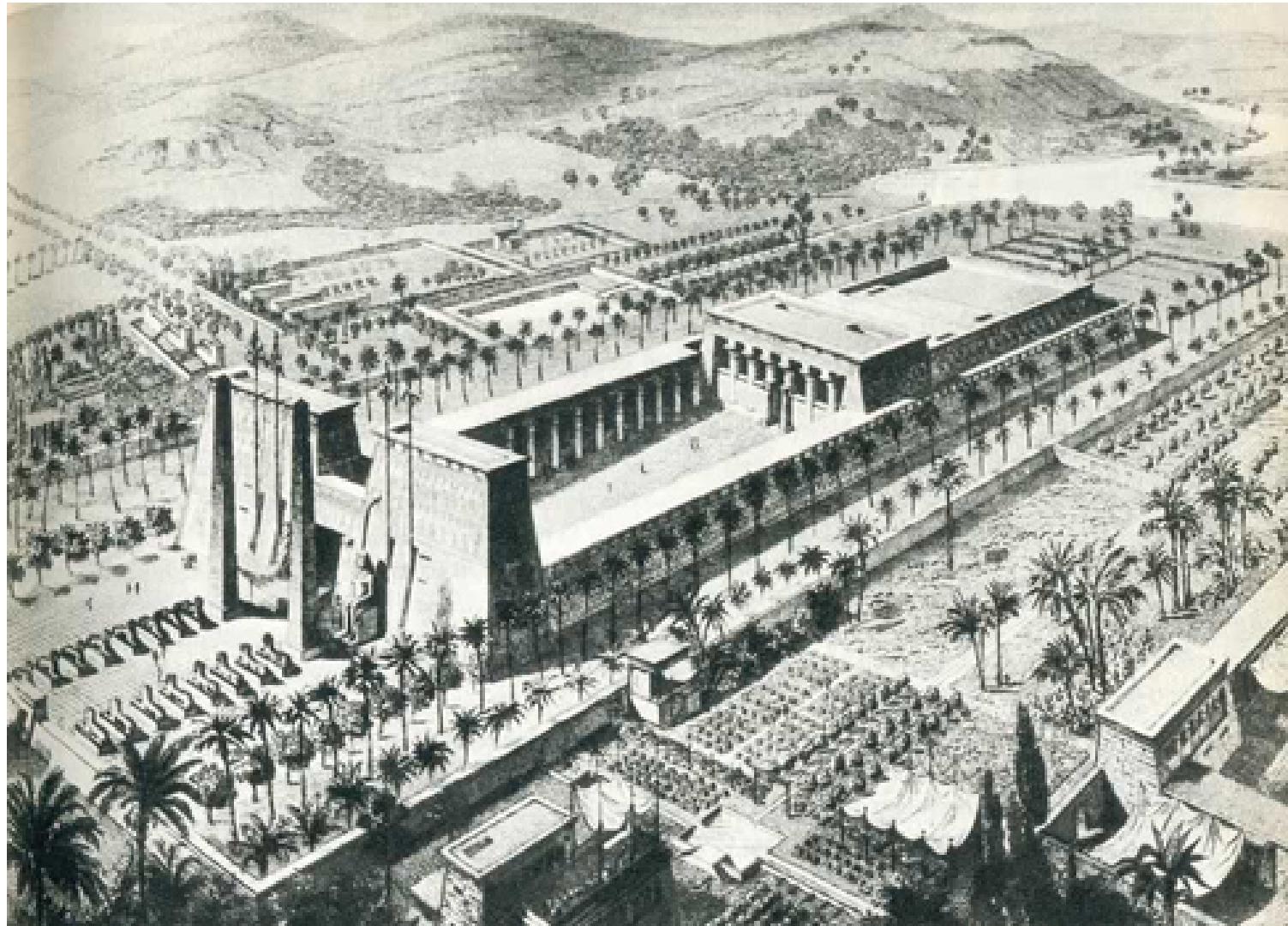
Найдётся всё

YAHOO!

Wie wirklich alles begann . . .

# Suchmaschinengeschichte

Die ersten Suchmaschinen ...



# Suchmaschinengeschichte

Die ersten „Suchmaschinen“ ... waren Bibliotheksangestellte!



# Suchmaschinengeschichte

Die ersten „Suchmaschinen“ ... waren Bibliotheksangestellte mit Zettelkatalogen

## FROM CARD CATALOG TO THE BOOK ON THE SHELF

**THE CARD CATALOG**  
is an alphabetical list of books found in the Library

**THE THREE WAYS OF FINDING A BOOK IN THE CATALOG**

- UNDER AUTHORS SURNAME
- UNDER TITLE OF BOOK
- UNDER SUBJECT WITH WHICH BOOK DEALS

**THE CALL NUMBER**  
Directs you to the book's location on the shelf and is found in the upper left hand corner of the catalog card also on the back of the book which is on the shelf.

**ARRANGEMENT OF BOOKS**  
A numerical system is followed in correct order

**CLASSIFICATION**

000-099	general works
100-199	Philosophy
200-299	Religion
300-399	Sociology
400-499	Linguistics
500-599	Natural Sciences
600-699	Useful Arts
700-799	Fine Arts
800-899	Literature
900-999	Geography

Fiction is not classified but is arranged on the shelves alphabetically by author

**PEABODY VISUAL AIDS**  
PUBLISHED BY FOLLETT BOOK COMPANY CHICAGO

Prepared under the direction of Miss Ruby Ethel Gurdiff for the Peabody Library School Course in Teaching the Use of the Library. Planned by Martha Edmondson, lettered by Mr. McComb.

# Aktuelle „Krone“ der Suchevolution



bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

# Aktuelle „Krone“ der Suchevolution



bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

# Aktuelle „Krone“ der Suchevolution



bing

Google

amazon

Яндекс

Найдётся всё

YAHOO!

Wie funktioniert eine Suchmaschine?

# Suchmaschinenarchitektur

# Suchmaschinenarchitektur

## Allgemein

### Ziele

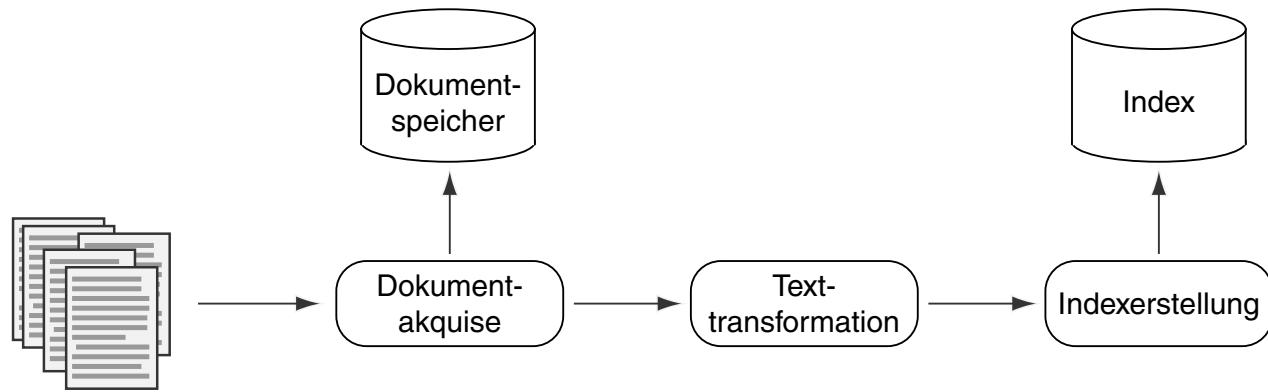
- Effektivität: Qualität der Suchergebnisse
- Effizienz: Geschwindigkeit des Zurücklieferns der Suchergebnisse

### Aufbau

- Indexierungsprozess:  
Datenstrukturen für die Suche anlegen
- Anfrageprozess:  
Suchanfragen verstehen, Ergebnisse liefern, Nutzerverhalten beobachten

# Suchmaschinenarchitektur

## Indexierungsprozess



# Suchmaschinenarchitektur

## Indexierungsprozess: Dokumentakquise

### 1. Crawler

- Werkzeug zum Entdecken und Herunterladen von Dokumenten
- Folgt bspw. den Verlinkungen in Dokumenten (Web Crawler)

### 2. Konverter

- Vereinheitlichung von Dokumentformaten (HTML)
- Vereinheitlichung der Kodierung (z.B. ASCII oder UTF)

### 3. Dokumentspeicher

- Originale und konvertierte Dokumente, Metadaten, Versionshistorie

# Suchmaschinenarchitektur

## Indexierungsprozess: Texttransformation

- Extraktion von (Index-) Termen (repräsentative Wörter eines Dokuments)
- Vokabular: Menge aller Wörter in einem Dokument
- Index: Terme verweisen auf Liste aller den Term enthaltenden Dokumente

## Komponenten

- Segmentierer
- Stoppwortfilter
- Stemmer
- Lemmatisierer

# Suchmaschinenarchitektur

## Indexierungsprozess: Indexerstellung

- Extraktion von (Index-) Termen (repräsentative Wörter eines Dokuments)
- Vokabular: Menge aller Wörter in einem Dokument
- Index: Terme verweisen auf Liste aller den Term enthaltenden Dokumente

## Komponenten

- Segmentierer: Trennung in Struktureinheiten (Wörter, Sätze, Absätze)
- Stoppwortfilter: Entfernung häufiger und domänen-spezifischer Wörter  
z.B. *und, oder, der, die* oder *Wikipedia* in Wikipedia-Artikeln
- Stemmer: Zurückführung auf eine Grundform durch Abschneiden  
z.B. *Lauf* → *lauf/* und *laufen* → *lauf/en*, aber *lief* → *lief/*
- Lemmatisierer: Überführung in Stammform (*saw* → *see*)  
z.B. *laufen* → *laufen*, und *lief* → *laufen*

# Suchmaschinenarchitektur

Index: Datenstruktur ähnlich Schlagwortverzeichnis

---

$T \rightarrow \text{Postings}$  (Posting Lists, Postlists)

---

$t_1$	$\rightarrow$	$d_1, w_{1,1}$	$d_2, w_{1,2}$		
$t_2$	$\rightarrow$	$d_1, w_{2,1}$	$d_2, w_{2,2}$	$d_4, w_{2,4}$	
$t_3$	$\rightarrow$	$d_1, w_{3,1}$	$d_2, w_{3,2}$	$d_4, w_{3,4}$	$d_5, w_{3,5}$
$t_4$	$\rightarrow$	$d_2, w_{4,2}$			
$t_5$	$\rightarrow$	$d_1, w_{5,1}$			
	$\vdots$				

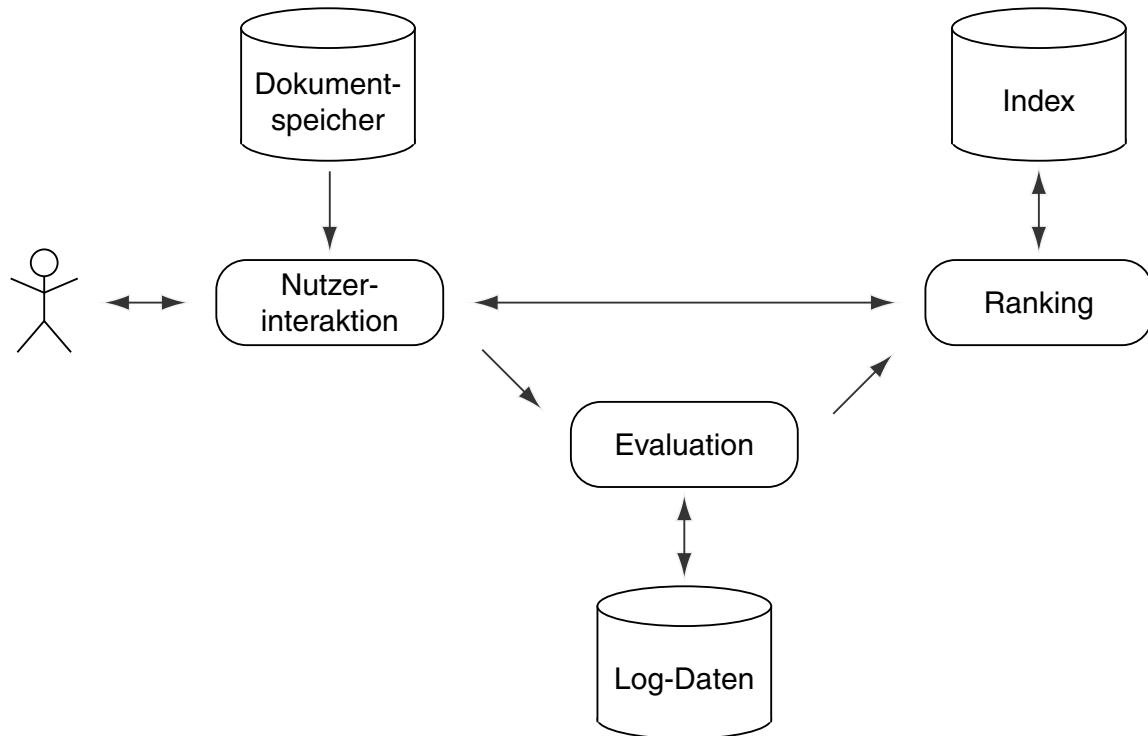
---

Für jeden Term  $t$

- Liste der Dokumente  $d$ , die  $t$  enthalten
- Optional: Gewicht  $w$  für die Bedeutung von  $t$  in  $d$  (z.B. Häufigkeit)

# Suchmaschinenarchitektur

## Anfrageprozess



# Anfrageprozess

## Anfragevorverarbeitung

- Im Prinzip analog Dokumentvorverarbeitung
  - Unwichtige Wörter ignorieren
  - Normalisieren
- Und ganz wichtig: Rechtschreibkorrektur!

bibliothek von alexanderer

# Anfrageprozess

## Ranking

Beispielanfrage: bibliothek von alexandria

Gegeben: Index mit Postings

Gesucht: Alle Dokumente, die alexandria und bibliothek enthalten

Anfrageergebnis: [2, ...] [8, ...] [41, ...] [77, ...] ...

# Anfrageprozess

## Ranking

Beispielanfrage: bibliothek von alexandria

Gegeben: Index mit Postings

Gesucht: Alle Dokumente, die alexandria und bibliothek enthalten

Anfrageergebnis: [2, ...] [8, ...] [41, ...] [77, ...] ...

- Anfrageergebnis sortieren! Ranking als „Geheimformel“ einer Suchmaschine
- Einsatz maschineller Lernverfahren
- Gewichtete Summe aus vielen Hundert Signalen
  - Häufigkeit der Anfrageterme im Dokument
  - Wichtigkeit des Dokuments im Web
  - Aktualisierungsdatum
  - Textqualität
  - ...

# Anfrageprozess

## Ranking

Beispielanfrage: bibliothek von alexandria

Gegeben: Index mit Postings

Gesucht: Alle Dokumente, die alexandria und bibliothek enthalten

Anfrageergebnis: [2, ...] [8, ...] [41, ...] [77, ...] ...

- Anfrageergebnis sortieren! Ranking als „Geheimformel“ einer Suchmaschine
- Einsatz maschineller Lernverfahren
- Gewichtete Summe aus vielen Hundert Signalen
  - Häufigkeit der Anfrageterme im Dokument
  - Wichtigkeit des Dokuments im Web
  - Aktualisierungsdatum
  - Textqualität
  - ...
- Schlüssel fürs Lernen: Nutzerverhalten!

# Nutzerverhalten

# Nutzerverhalten

## Suchinteraktionen bewerten

- Anfrage: bibliothek von alexandria
- Annahme: alle 42 Mio. Klicks für die Anfrage sehen wie folgt aus
  - Suchergebnis  $d_1$
  - $d_2$
  - $d_3$  (geklickt)
  - $d_4$
  - $d_5$
  - $d_6$
  - $d_7$
  - $d_8$
  - $d_9$
  - $d_{10}$
- Was sollte die Suchmaschine daraus lernen?

# Nutzerverhalten

## Suchinteraktionen bewerten

- Anfrage: bibliothek von alexandria
- Annahme: alle 42 Mio. Klicks für die Anfrage sehen wie folgt aus
  - Suchergebnis  $d_1$
  - $d_2$
  - $d_3$  (geklickt)
  - $d_4$
  - $d_5$
  - $d_6$
  - $d_7$
  - $d_8$
  - $d_9$
  - $d_{10}$
- Präferenzen  $d_3$  ist besser als  $d_1, d_2$  und  $d_4$

# Nutzerverhalten

## Suchinteraktionen bewerten

- Anfrage: bibliothek von alexandria
  - Annahme: alle 42 Mio. Klicks für die Anfrage sehen wie folgt aus
    - Suchergebnis  $d_1$
    - $d_2$
    - $d_3$  (geklickt)
    - $d_4$
    - $d_5$
    - $d_6$
    - $d_7$
    - $d_8$
    - $d_9$
    - $d_{10}$
  - Präferenzen  $d_3$  ist besser als  $d_1, d_2$  und  $d_4$
- 
- Suchmaschinen zeichnen alles auf, was man mit ihnen macht
  - ... um die Ranking-Signale besser auszutarieren

# Nutzerverhalten

## Suchende als Versuchskaninchen

- Beim Suchen nimmt man immer an vielen „Experimenten“ teil
- Große Suchmaschinen optimieren ständig alles
  - Ranking-Mechanismen
  - Größe des Suchfensters
  - Einblenden von Bildern, Videos, usw.
  - Länge der angezeigten Textauszüge (Snippets)
  - Blautöne (ja, selbst das!)
  - ...
- Suchende mit den Varianten beobachten (Test- und Kontrollgruppe)
- Abschätzen, was besser funktioniert

# Nutzerverhalten

## Suchende als Versuchskaninchen

- Beim Suchen nimmt man immer an vielen „Experimenten“ teil
- Große Suchmaschinen optimieren ständig alles
  - Ranking-Mechanismen
  - Größe des Suchfensters
  - Einblenden von Bildern, Videos, usw.
  - Länge der angezeigten Textauszüge (Snippets)
  - Blautöne (ja, selbst das!)
  - ...
- Suchende mit den Varianten beobachten (Test- und Kontrollgruppe)
- Abschätzen, was besser funktioniert
- Schon das ist ggf. nicht unbedenklich
- **Aber es kommt noch „besser“ ...**

# Nutzerverhalten

## Persönliche Profile

- Einige Suchmaschinen speichern das Verhalten **personalisiert**
- Wenn man dann noch bedenkt, dass eine solche Suchmaschine
  - über 90% Marktanteil hat,
  - einen Großteil der im Web angezeigten Werbung ausspielt,
  - einen eigenen Web-Browser zur Verfügung stellt,
  - und und und,
- ... dann wird klar, was diese Suchmaschine alles über jeden weiß

# Nutzerverhalten

## Persönliche Profile

- Einige Suchmaschinen speichern das Verhalten **personalisiert**
- Wenn man dann noch bedenkt, dass eine solche Suchmaschine
  - über 90% Marktanteil hat,
  - einen Großteil der im Web angezeigten Werbung ausspielt,
  - einen eigenen Web-Browser zur Verfügung stellt,
  - und und und,
- ... dann wird klar, was diese Suchmaschine alles über jeden weiß



# Nutzerverhalten

## Weniger „übergriffige“ Wettbewerber (Beispiele)

- Brave Search [ <https://search.brave.com/> ]
  - Eigener Index; sammelt wohl keine personalisierten Daten
- DuckDuckGo [ <https://duckduckgo.com/> ]
  - Bing-Ergebnisse; wohl wenig Nutzerinformationen weitergereicht
- Mojeek [ <https://www.mojeek.com/> ]
  - Eigener Index; sammelt wohl keine personalisierten Daten
- Startpage [ <https://www.startpage.com/> ]
  - Google-Ergebnisse; wohl wenig Nutzerinformationen weitergereicht
- Swisscows [ <https://swisscows.com/> ]
  - Bing-Ergebnisse; eigener Index für Deutsch; sammelt wohl keine personalisierten Daten
- You [ <https://www.you.com/> ]
  - Bing-Ergebnisse; ‘private mode’ sammelt wohl keine personalisierten Daten

# Zusammenfassung

- Suchmaschinenindex

Offline erstellte Datenstruktur zum schnellen Zugriff.

- Anfrageprozess

Suchintention „erraten“ und passende Ergebnisse anzeigen.

- Wichtigste Datenquelle

Nicht unbedingt Webseiten ... sondern Nutzerverhalten!