
Aufbau einer Suchmaschine

Prof. Dr. Clemens Cap

Universität Rostock

<http://www.internet-prof.de>

Komponenten einer Suchmaschine

Eine Suchmaschine besteht aus 4 Komponenten

- Crawling:** Herunterladen von Seiten aus dem Internet
- Indexing:** Herstellen von Datenstrukturen, die Suchwörter mit Seiten assoziieren
- Ranking:** Anordnung von Seiten auf der Liste von "Treffern"
Zugleich: Mechanismus zur Bewertung von Seiten
- Searching:** Umwandeln der Benutzereingabe in eine Abfrage

Crawler

Aufbau und Vorgehensweise

Der Crawler lädt die Seiten aus dem Internet

Der Crawler fragt sich dabei

Wie finde ich ein initiales Pool von Seiten?

- Anmeldungen von Domänen bei Domain-Registry und whois-Registry
- Spezielle Anmeldeseite der Suchmaschinen
- Sammlung bisher gefundener Links

Welche Seiten soll ich überhaupt laden?

- Tiefe der Seite in der Hierarchie
- Anzahl der Links auf eine Seite, die gefunden wurden

Wann soll ich eine Seite erneutladen?

- Steuerung durch meta-Tags möglich
- Einfluß der Änderungshäufigkeit in der Vergangenheit

Crawler

Zentrale Fragen für Webmaster

Kann der Crawler eine Seite überhaupt **finden**?

- **Zugriffsschutz:** Seite darf nicht durch Login geschützt sein
- **Suchformular:** Seite muß verlinkt sein
Crawler kann nicht sinnvolle Eingaben in Suchformulare raten
- **Aussperrung:** Robot-Datei und meta tags können Crawler aussperren

Kann der Crawler eine Seite überhaupt **auswerten**?

- **Frames:** Nicht von jedem Crawler richtig behandelt
- **Dynamische Seiten:** Inhalte, die erst durch Javascript aufgebaut werden
Inhalte, die erst durch User Interaktion entstehen
- **Graphische Seiten:** Texte, die eigentlich Graphiken oder Spezialformate sind
Aber: Konnektoren für diverse Formate PDF, PPT, DOC

Deep Web: Jener Teil des Web, der nicht durch Suchmaschinen zu finden ist

Google Webmaster Tools: Informieren über Probleme des Google Crawlers

Crawler Sitemap Protokoll

Mechanismus, um den Crawler über vorhandene Seiten zu informieren
Speziell für Sites mit Deep Web Inhalten

Angabe von Zusatzinfo möglich

- Letztes Update der Seite
- Änderungshäufigkeit der Seite
- Bedeutung der Seite relativ zu anderen auf der Website

Anbieten von Sitemaps

- **Datei robots.txt** direkt im Root-verzeichnis der Domäne
Darin ein Eintrag, der die URL angibt:
Bsp: `sitemap: http://www.bsp.de/sitemap.xml`
- **Submission URL** der Suchmaschine selber

**Empfehlung: Sitemaps
anbieten, falls erforderlich !**

Crawler

Sitemap Protokoll Beispiel

```
<urlset
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
                      http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">
  <url>
    <loc>http://www.wikipedia.org</loc>
    <lastmod>2006-11-18</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.8</priority>
  </url>
</urlset>
```

Crawler

Doorpages zur Manipulation ??

Doorpages sind Seiten zum "Anfüttern" der Suchmaschine

- Suchmaschine sieht Doorpage
- User sieht andere Inhalte (unter derselben URL)
- Implementierung: Inhalt steuern aufgrund http UserAgent oder IP Adresse

Effekte

- Seite erhält anderes (besseres?) Ranking
- Seite wird unzutreffenden Suchwörtern zugeordnet
- Letztlich: "Betrug" der Suchmaschine

Daher

- Von guten Suchmaschinen erkannt und penalisiert
- Manche Suchmaschinen drohen mit schwarzer Liste
- Juristisch problematisch (unlauterer Wettbewerb)

Empfehlung: Keine Doorpages einsetzen !

Indexer

Arten invertierter Indizes

Gegeben: Menge von Dokumenten + Suchwort oder Suchphrase
Gesucht: Dokumente, die Wort oder Phrase enthalten

File / Filesystem

- Datenstruktur, welche Dokumente auf die enthaltenen Wörter abbildet

Invertierter Index

- Datenstruktur, welche Wörter auf die sie enthaltenden Dokumente abbildet
- Findet durch Durchschnittsbildung alle Texte, die bestimmte Wörter enthalten

Voller invertierter Index

- Datenstruktur, die Wörter auf Dokumente und Ort des Auftretens abbildet
- Findet alle Texte, in denen bestimmte Phrasen auftreten

Indexer

Invertierter Index

Text0: it is what it is

Text1: what is it

Text2: it is a banana

Beispiel Text-Sammlung

"a": {2}
"banana": {2}
"is": {0, 1, 2}
"it": {0, 1, 2}
"what": {0, 1}

Doku Nummer

Invertierter Index

"a": {(2, 2)}
"banana": {(2, 3)}
"is": {(0, 1), (0, 4), (1, 1), (2, 1)}
"it": {(0, 0), (0, 3), (1, 2), (2, 0)}
"what": {(0, 2), (1, 0)}

(Doku Nummer, Position in Text)

Voller invertierter Index

Indexer

Invertierter Index im Einsatz

Text0: it is what it is

Text1: what is it

Text2: it is a banana

Suche nach Worten:

what, is, it

"a": {2}

"banana": {2}

"is": {0, 1, 2}

"it": {0, 1, 2}

"what": {0, 1}

$$\{0, 1, 2\} \cap \{0, 1, 2\} \cap \{0, 1\} = \{0, 1\}$$

Indexer

Voll invertierter Index im Einsatz

Text0: it is what it is

Text1: what is it

Text2: it is a banana

Suche nach Phrase:

"what is it"

"a": {(2, 2)}

"banana": {(2, 3)}

"is": {(0, 1), (0, 4), (1, 1), (2, 1)}

"it": {(0, 0), (0, 3), (1, 2), (2, 0)}

"what": {(0, 2), (1, 0)}


Ranking

Aufgabe und Bedeutung

User sehen sich typischerweise die ersten 10 – 40 Treffer an

Daher

- Möglichst gute Platzierung wichtig – oder –
- Bezahlter Link erforderlich



The image shows a Google search interface for the term "Ajax". The search bar contains "Ajax" and the search button is labeled "Suche". Below the search bar, there are radio buttons for "Suche: Das Web" (selected), "Seiten auf Deutsch", and "Seiten aus Deutschland". The search results are displayed under the heading "Web" and show "Ergebnisse 1 - 10 von ungefähr 81.300.000 für Ajax. (0,07 Sekunden)".

The first search result is "Ajax (Programmierung) - Wikipedia". The snippet reads: "Ajax ['eidzæks] ist ein Acronym für die Wortfolge **Asynchronous JavaScript and XML**. Es bezeichnet ein Konzept der asynchronen Datenübertragung zwischen einem ...". The URL is "de.wikipedia.org/wiki/Ajax_(Programmierung) - 84k - Im Cache - Ähnliche Seiten".

The second search result is "Ajax - Wikipedia". The snippet reads: "Ajax bezeichnet:.. zwei griechische Heerführer im Trojanischen Krieg aus Homers Ilias:.. den Sohn des Königs Telamon von Salamis, genannt „der Telamonier“, ...". The URL is "de.wikipedia.org/wiki/Ajax - 18k - Im Cache - Ähnliche Seiten".

The third search result is "AJAX Quellensammlung | Dr. Web Weblog". The snippet reads: "Das Weblog für Webdesigner, Webworker und Seitenbetreiber." The URL is "www.drweb.de/weblog/weblog/?p=454 - 53k - 17. Juni 2007 - Im Cache - Ähnliche Seiten".

The sidebar on the right contains three advertisements, each with a red circle around the word "Anzeigen":

- AJAX framework for .NET**: Codelessly **AJAX**-enable any new or existing ASP.NET application. URL: www.telerik.com/radajax
- Enterprise AJAX Framework**: Top-grade Backbone **Ajax** Solution Tutorial, Samples, Tools, Learn now. URL: www.backbase.com
- Beyond AJAX**: URL: [www.backbase.com](#)

Ranking Vorgehen

Wort-Zählung

Seite gut, wenn Suchwort oft aufscheint

- Naiver Ansatz, leicht zu betrügen
- Heute nicht mehr im Einsatz

Linktopologie

Seite ist gut, wenn sie gut im Internet verlinkt ist

- Kernfrage: Was ist "gut verlinkt"?
- 2 Kernideen: HITS Algorithmus und Pagerank; später Modifikationen

Heuristiken

Naheliegende Verbesserungen

Forschung

Hier wird noch nachgedacht

- Textanalyse
- Topic Nets
- Ontologische Analyse
- usw.

Ranking

Wortzählung

Wortzählende Suchmaschinen werten positiv, wenn

- Suchbegriff ist im **Linktext**
- Suchbegriff ist im **Titel** der Seite `<title>`
- Suchbegriff ist in den **<meta> tags**

Keywords

```
<meta name="keywords" content="Meta-Tag, Metadaten, Suchmaschinen">
```

Description

```
<meta name="description" content="Seite des Projekts ...">
```

- Suchbegriff steht am **Anfang im Text**
- Suchbegriff steht **oft im Text**

Heute:

- Entsprechende Maßnahmen sind wichtig
- Dürfen nicht übertrieben werden, da sonst Penalisierung
- Reichen alleine nicht aus

Ranking Vorgehen

Wort-Zählung

Seite gut, wenn Suchwort oft aufscheint

- Naiver Ansatz, leicht zu betrügen
- Heute nicht mehr im Einsatz

Linktopologie

Seite ist gut, wenn sie gut im Internet verlinkt ist

- Kernfrage: Was ist "gut verlinkt"?
- 2 Kernideen: HITS Algorithmus und Pagerank; später Modifikationen

Heuristiken

Naheliegende Verbesserungen

Forschung

Hier wird noch nachgedacht

- Textanalyse
- Topic Nets
- Ontologische Analyse
- usw.

Eigene Lehreinheit

Searching

Die Möglichkeiten

Freitextsuche

Eingabe von Suchwörtern oder Suchphrasen

Detailliertes Formular

Genauere Erfassung in einem Formular

- Query by Example
- Query by Form

Abfragesprache

Spezielle Programmiersprache zur Abfrage

Searching Freitextsuche

Ad
Fläche

Snippets

Freitext
Eingabe

Ad
Fläche

Google

Web Bilder Groups News Froogle Mehr

Microsoft Exchange Server

Suche

[Erweiterte Suche](#)
[Einstellungen](#)

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web

Ergebnisse 1 - 10 von ungefähr 127.000.000 für Microsoft Exchange Server . (0,35 Sekunden)

[Hosted Exchange](#) ab 3,49€

Anzeigen

www.vivio.de Bankentochter bietet professionelle Kommunikation für Selbständige

[Exchange Server](#)

manageengine.adventnet.com Monitor **Microsoft Exchange Server** Availability & Performance stats

Anzeigen

[Outlook im Netzwerk](#)

Gemeinsame Kalender, Kontakte
Emails, Aufgaben ohne **Exchange**
www.publicshareware.de

[Cortado](#)

Eigenes **Exchange** Email Konto
incl. Direct Push Mail Service.
www.cortado.de

[Microsoft Exchange Server](#)

Top Angebote jetzt
sofort zum Kracherpreis kaufen!
www.ebay.de

[Microsoft Exchange Server](#) - offizielle Homepage

Microsoft Exchange Server 2003 bietet Unternehmen eine Kommunikationsplattform, die alle Mitarbeiter und deren Wissen miteinander verbindet.

www.microsoft.com/germany/exchange/default.aspx - 16k - [Im Cache](#) - [Ähnliche Seiten](#)

[Microsoft Exchange Server](#) - [[Diese Seite übersetzen](#)]

Get product and licensing information, downloads, updates, and technical articles about **Microsoft Exchange Server**, the messaging and collaboration server.

www.microsoft.com/exchange/ - 26k - [Im Cache](#) - [Ähnliche Seiten](#)

Searching

Query by Example / by Form



Erweiterte Suche

[Suchtipps](#) | [Über Google](#)

Ergebnisse finden	mit allen Wörtern	<input type="text"/>	10 Ergebnisse ▾ <input type="button" value="Google-Suche"/>
	mit der genauen Wortgruppe	<input type="text"/>	
	mit irgendeinem der Wörter	<input type="text"/>	
	ohne die Wörter	<input type="text"/>	
Sprache	Antwortseiten, geschrieben in	<input type="text" value="beliebiger Sprache"/>	
Dateiformat	<input type="text" value="Ausschließlich"/> Ausgabe von Ergebnissen des Dateiformats	<input type="text" value="irgendein Format"/>	
Datum	Ausgabe neuer Webseiten, aktualisiert während	<input type="text" value="keine Zeitbegrenzung"/>	
Position	Antwortseiten, in denen meine Begriffe vorkommen	<input type="text" value="irgendwo auf der Seite"/>	
Domains	<input type="text" value="Ausschließlich"/> Antwortseiten von der Site oder Domain	<input type="text"/> Beispiele: .org, google.com Weitere Informationen	
Nutzungsrechte	Ergebnisse zurückgeben, die	<input type="text" value="nicht nach Lizenz gefiltert sind"/>	
SafeSearch	<input checked="" type="radio"/> Kein Filter <input type="radio"/> Filtern mit SafeSearch		

Seitenspezifische Suche

Ähnlich	Seiten suchen, die der folgenden Seite ähnlich sind	<input type="text"/>	<input type="button" value="Google-Suche"/>
		Beispiel: www.google.com/help.html	
Links	Seiten suchen, die einen Link auf die folgende Seite enthalten	<input type="text"/>	<input type="button" value="Google-Suche"/>

Searching Abfragesprache

Google **Erweiterte Suche** [Suchtipps](#) | [Über Google](#)

Ergebnisse finden mit **allen** Wörtern
mit der **genauen Wortgruppe**
mit **irgendeinem** der Wörter
ohne die Wörter

10 Ergebnisse

Sprache Antwortseiten, geschrieben in Deutsch

Dateiformat Ausgabe von Ergebnissen des Dateiformats Microsoft Powerpoint (.ppt)

Datum Ausgabe neuer Webseiten, aktualisiert während die letzten 3 Monate

Position Antwortseiten, in denen meine Begriffe vorkommen im Hauptteil der Seite

Domains Antwortseiten von der Site oder Domain msn.com
Beispiele: .org, google.com [weitere Informationen](#)

Nutzungsrechte Ergebnisse zurückgeben, die nicht nach Lizenz gefiltert sind

SafeSearch Kein Filter Filtern mit [SafeSearch](#)

allintext: "Microsoft Exchange Server" -Lotus site:msn.com filetype:ppt

Searching

Mögliche Features

Automatische Query-Modifikation

- Entfernen von sog. **Stopwörtern** (häufige Wörter, die in Suche oft sinnlos)
Bsp: wer, wo, was, der, die, das
- Erweiterung der Suche (sog. **Query Extension**) um häufige Begleitwörter

Phrasensuche vs. Wortsuche

- **Wortsuche:** Wenn Wörter nur hintereinandergesetzt, dann wird Reihenfolge nicht beachtet oder nur zum Ranking benutzt
Bsp: to be or not to be = to be or not
- **Phrasensuche:** Wenn Wörter unter Anführungszeichen gesetzt, dann wird genau nach dieser Wortgruppe gesucht
Bsp: "to be or not to be"

Searching

Google Abfragesprache

filetype: type	Suche nach Dokumenten mit entsprechendem Typ Bsp: PDF, DOC etc.
allintitle:	Suche nachfolgende Worte in Titel
allintext:	Suche nachfolgende Worte in Text der Seite
allinurl:	Suche nachfolgende Worte in URL
allinanchor:	Suche nachfolgende Worte in Ankertext
link: url	Welche Seiten verlinken auf diese URL ?
related: url	Welche Seiten sind ähnlich wie diese Seite ?