

Vom Suchen und Finden - Google und andere Ansätze

Norbert Fuhr

Universität Duisburg Essen
FB Ingenieurwissenschaften
Abteilung Informatik

3. Juli 2006

Gliederung

- 1 Einführung
- 2 Grundlagen
- 3 Erweiterte Suchfunktionen
- 4 Ausblick: Personalisierung

Einführung

- Suche in über 25 Milliarden Webseiten (Mai 2006)
- Hohe Treffergenauigkeit - „beste“ Antworten zuerst
- Probleme mit mehrdeutigen Begriffen
- Probleme, wenn das Suchziel nicht präzise formuliert werden kann

Retrievalverfahren

„Suche Wintersportort in den Alpen, in dem Skilanglauf und Rodeln angeboten wird“

Suchwort	Dokumente			
	d_1	d_2	d_3	d_4
Rodeln	1	1	1	1
Skilanglauf	1	1		1
Wintersportort	1	1	1	
Alpen	1	1	1	
Dokumentgewicht	4	4	3	2

Lineares Retrieval:

Dokumente werden nach fallender Anzahl der Treffer geordnet:

d_1, d_2, d_3, d_4

Retrievalverfahren

„Suche Wintersportort in den Alpen, in dem Skilanglauf und Rodeln angeboten wird“

Suchwort	Dokumente			
	d_1	d_2	d_3	d_4
Rodeln	1	1	1	1
Skilanglauf	1	1		1
Wintersportort	1	1	1	
Alpen	1	1	1	
Dokumentgewicht	1	1	0	0

Boolesches Retrieval:

Nur die Dokumente werden ausgegeben, die alle Suchbedingungen erfüllen (implizite UND-Verknüpfung)

d_1, d_2

Retrievalverfahren

„Suche Wintersportort in den Alpen, in dem Skilanglauf und Rodeln angeboten wird— *aber nicht in Frankreich*“

Suchwort	Dokumente			
	d_1	d_2	d_3	d_4
Rodeln	1	1	1	1
Skilanglauf	1	1		1
Wintersportort	1	1	1	
Alpen	1	1	1	
Frankreich		1	1	
Dokumentgewicht	1	0	0	0

Boolesches Retrieval mit Negation:

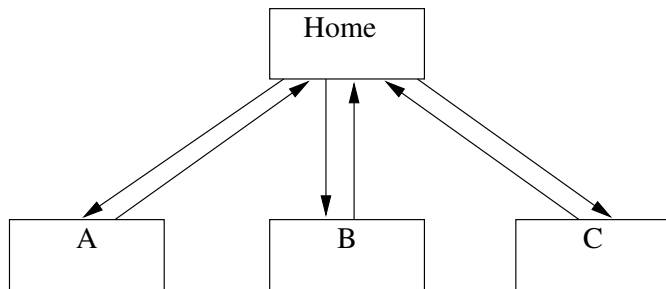
Nur die Dokumente werden ausgegeben, die alle Suchbedingungen erfüllen:

d_1

Page-Rank

- Zu den meisten Anfragen gibt es sehr viele Web-Seiten, die alle Bedingungen erfüllen
- Es wird ein Verfahren zur Rangordnung der gefundenen Seiten benötigt
- → Page Rank

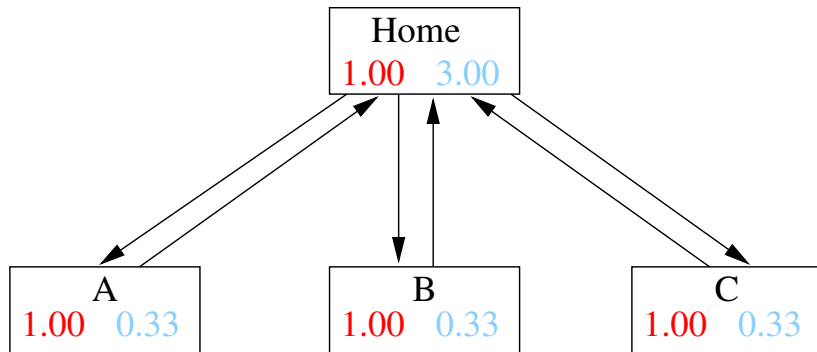
Page Rank: Idee



- Berücksichtige die Link-Struktur im Web
- Seiten, auf die häufig verwiesen wird, sind besser also solche, auf die kaum verwiesen wird

Page Rank-Algorithmus

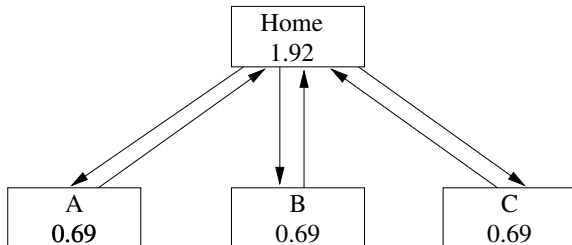
- 1 Gebe jeder Seite das initiale Gewicht 1
- 2 Gewicht eines ausgehenden Links =
Seitengewicht/Gesamtzahl ausgehender Links
- 3 Gewicht einer Seite = Summe der Gewichte der
eingehenden Links
- 4 Wiederhole Schritte 2 und 3, bis die Gewichte sich nicht
mehr ändern



- Das Verfahren konvergiert leider nicht
- **Dämpfungsfaktor** notwendig

Verbesserter Page Rank-Algorithmus

- 1 Gebe jeder Seite das initiale Gewicht 1
- 2 Gewicht eines ausgehenden Links = Seitengewicht $\cdot d$ / Gesamtzahl ausgehender Links
- 3 **Gewicht einer Seite = $(1 - d) + d \cdot$ Summe der Gewichte der eingehenden Links**
- 4 Wiederhole Schritte 2 und 3, bis die Gewichte sich nicht mehr ändern



Page-Rank-Formel

$PR(X)$ Page-Rank-Gewicht der Seite X

$C(X)$ Anzahl der von der Seite X ausgehenden Links

d Dämpfungsfaktor

$$PR(A) = (1 - d) + d \left(\frac{PR(T_1)}{C(T_1)} + \dots + \frac{PR(T_n)}{C(T_n)} \right)$$

Random Surfer-Modell:

- d : Wahrscheinlichkeit, dass Benutzer auf einen der ausgehenden links klickt
- $1 - d$: Wahrscheinlichkeit, dass Benutzer zu einer zufälligen anderen Seite springt

Vor- und Nachteile von Page Rank

- + Page Rank bevorzugt populäre Seiten
- + Gute Ergebnisse für die Suche nach Home Pages
- + - Page Rank bevorzugt Einstiegsseiten von Web Sites
- Zu engeren thematischen Anfragen liefern andere Verfahren bessere Ergebnisse

Linguistische Flexions- und Derivationsformen

Stemming/Grund- und Stammformreduktion

- Suche nach allen Flexionsformen eines Wortes:
Haus – Hauses – Häuser
schreiben – schreibt – schrieb – geschrieben
- Suche nach allen Wörtern mit dem gleichen Wortstamm (Derivationsformen):
Formatierung – Format – formatieren

Stemming wird angeblich durch Google und MSN Search unterstützt, lässt sich aber nicht nachvollziehen

Tippfehlerkorrektur

The screenshot shows the Google search interface. The search bar contains the text "Duisbrg". Below the search bar, the text "Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland" is visible. The search results section is titled "Web" and shows "Ergebnisse 1 - 10 von ungefähr 320 für Duisbrg. (0,25 Sekunden)". A red text "Meinten Sie: [Duisburg](#)" is displayed above the first search result. The first result is titled "Die Wohnheim-Tutoren des Studentenwerks Essen-Duisburg" and includes the URL "tudu.no-ip.org/tudupublic/home/" and other links like "Im Cache" and "Ähnliche Seiten".

Meinten Sie: [Duisburg](#)

[Die Wohnheim-Tutoren des Studentenwerks Essen-Duisburg](#)

Die Wohnheim-Tutoren des Studentenwerks Essen-Duisburg.

tudu.no-ip.org/tudupublic/home/ - 8k - [Im Cache](#) - [Ähnliche Seiten](#)

[findus der Landesfachstellen](#)

... Schlagwort war: **Duisbrg** Weitere Schlagworte: Modernisierung; Stadtbibliothek,

Az BUB, Aufsatz, Aufsatzdatenbank Ja, ist verfügbar. ...

www.datronic.de/cgi-bin/findus.pl?customer=ifs&suchfeld1=schlagwort&wildcard1=istgleich&suchf... - 8k -

[Im Cache](#) - [Ähnliche Seiten](#)

(Seltene) Tippfehler werden von Google erkannt und Vorschläge zur Korrektur gemacht

Google Suggest

<http://www.google.com/webhp?complete=1&hl=en>



Web [Images](#) [Groups](#) [News](#) [Froogle](#) [Local](#)^{New!} [more »](#)

Google Search

I'm Feeling Lucky

[Advanced Search](#)

[Preferences](#)

[Language Tools](#)

As you type, Google will offer suggestions. Use the arrow keys to navigate the results. [Learn more](#)

Google Suggest (2)

duisburg	
duisburg	8,770,000 results
duisberg	119,000 results
duisburgs locale	5 results
duisburg locale	28,100 results
duistop.wmv	44 results
duis	261,000 results
duisburg germany	869,000 results
duistop	53 results
duisburg location	280,000 results
duisburg map	160,000 results

- Zeigt mögliche Vervollständigungen des Suchwortanfangs, zusammen mit Häufigkeiten
- Listet auch häufige Tippfehler (aber nur in der Vervollständigung)

Verfeinerung der Suche

http://alltheweb.com/

The screenshot shows the alltheweb search engine interface. At the top left is the logo "alltheweb" with the tagline "find it all" below it. To the right of the logo are navigation links: "advanced search", "customize preferences", "submit site", and "help". Below these links is a search input field containing the word "retrieval" and a "SEARCH" button. Under the search field, it says "Results in: Any Language" and "English". Below the search field is a horizontal menu with tabs for "Web", "News", "Pictures", "Video", and "Audio". A blue banner below the menu displays "1 - 10 of 21,300,000 Results for retrieval". Below the banner is a section titled "Refine your search:" with a subtext "(click '+' or '-' to include or exclude terms, and then click the 'SEARCH' button)". This section contains six suggestions for 2-word groups, each with a plus and minus icon: "information retrieval", "data retrieval", "document retrieval", "image retrieval", "retrieval system", and "number retrieval".

Vorschlag von 2-Wortgruppen bei der Anfrage mit einem einzelnen Wort

Verwandte Suchbegriffe

[content-based image](#)

[retrieval](#)

[data mining](#)

[digital library](#)

[image database](#)

[information retrieval](#)

[low level features](#)

[machine translation](#)

[memory](#)

[phase retrieval](#)

[prefrontal](#)

[query expansion](#)

[retrieval performance](#)

[retrieval system](#)

[search engine](#)

www.scirus.com

„Refine your search“

(zu *retrieval*)

Statistisch verwandte Wörter:

Häufig zusammen mit dem

Suchbegriff auftretende Wörter

Objekttypen



Die nächste Google-Version?

Bessere Suchergebnisse durch Suche nach bestimmten Objekttypen

Objekttypen bei Google



Web [Bilder](#) [Groups](#) [Verzeichnis](#) [News](#) [Froogle](#)^{Neu!}

uni-colleg duisburg-essen 2005

Google-Suche Auf gut Glück!

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

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

[Werbung](#) - [Unternehmensangebote](#) - [Alles über Google](#) - [Google.com in English](#)

©2005 Google - Suche auf 8.058.044.651 Web-Seiten

- Objekttypen müssen explizit bei der Anfrage ausgewählt werden

Objektypen bei AskJeeves

http://www.ask.com



[Web](#) | [Pictures](#) | [News](#) | [Local](#) **NEW!** | [Products](#) | [More](#) »

images Mozart

Search

[Advanced](#)

Web Search: images Mozart

Picture Search: [Mozart](#)

[About](#)



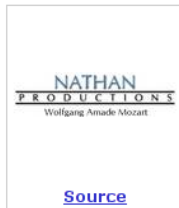
[Source](#)



[Source](#)



[Source](#)



[Source](#)

[Copyright & Disclaimer](#)

[More Picture Results»](#)

Sponsored Web Result

Objektypen bei AskJeeves (2)

The screenshot shows the AskJeeves search interface. At the top left is the AskJeeves logo. To its right are navigation links: [Web](#), [Pictures](#), [News](#), [Local](#), [NEW!](#), [Products](#), and [More >](#). Below these is a search input field containing the text "pocket pc" and a "Search" button. To the right of the search button is a link for [Advanced Options](#). A red banner below the search bar displays "Web Search: pocket pc".

The main content area is titled "Products for 'pocket pc'" and includes links for "Email This" and "About". It features three product listings, each with a small image of the device and its name:

- [iPAQ hx4705 PDA \(HP \(Hewlett-Packard\)\)](#)
- [iPAQ RX3115 Mobile Media Companion PDA \(...\)](#)
- [iPAQ rx3715 PDA \(HP \(Hewlett-Packard\)\)](#)

At the bottom of the product area, there is a "Find:" section with links for [PDAs](#) and [All Categories](#).

→ Gesuchte Objekttypen werden automatisch vom System erkannt

Objekttypen bei Google (2)

- Call-by-Call-Vorwahlen: (089) 12345678
- Sendungen verfolgen: 1Z9999W99999999999
- Stadtpläne: Duisburg
- Wertpapier-Informationen: DE0006231004
- Wörterbuch deutsch-englisch: conclusion en-de
- Zugverbindungen: Duisburg München 13:30
- Definitionen: definiere Suchmaschine

Personalisierung bei Amazon

http://www.amazon.de/

The screenshot shows the Amazon.de homepage with a personalized recommendation section. At the top, there is a navigation bar with the Amazon logo, a shopping cart icon, and links for 'WUNSCHZETTEL', 'MEIN KONTO', 'HILFE', and 'IMPRESSUM'. Below this is a category bar with buttons for 'HOME', 'NORBERTS SHOP', 'BÜCHER', 'ENGLISH BOOKS', 'ELEKTRONIK & FOTO', 'MUSIK', 'DVD', 'VHS', 'SOFTWARE', 'PC- & VIDEO-SPIELE', 'KÜCHE, HAUS & GÄRTEN', and 'SPIELWAREN & KINDERWELT'. A search bar is present with the text 'Schnellsuche: Alle Produkte' and a 'LOS' button. The main content area features a section titled 'Persönliche Empfehlungen' with a sub-header 'Ihre persönlichen Empfehlungen'. The text reads: 'Hallo, Norbert Fuhr. Entdecken Sie die heute vorgestellten Empfehlungen. (Wenn Sie nicht Norbert Fuhr sind, [klicken Sie hier.](#))'. Below this, there is a sub-section 'Buch-Empfehlungen' featuring 'The Stone Monkey (Lincoln Rhyme Novels (Paperback))' by Jeffery Deaver. The book cover is shown, and the text describes the plot: 'When a vicious smuggler known as the Ghost scuttles a ship filled with undocumented Chinese immigrants less than a mile from New York harbor, only a handful of survivors--and the Ghost himself--manage to escape the burning vessel. Lincoln Rhyme, the quadriplegic NYPD forensic detective first... [Mehr dazu](#)'.

amazon.de | WUNSCHZETTEL | MEIN KONTO | HILFE | IMPRESSUM

Auf Englisch: *
Harry Potter 6
Jetzt vorbestellen

HOME | NORBERTS SHOP | BÜCHER | ENGLISH BOOKS | ELEKTRONIK & FOTO | MUSIK | DVD | VHS | SOFTWARE | PC- & VIDEO-SPIELE | KÜCHE, HAUS & GÄRTEN | SPIELWAREN & KINDERWELT

Ihre persönliche Seite | Ihre persönlichen Empfehlungen | Ihre Lieblingsshops | Neu für Sie

Schnellsuche: Alle Produkte | | | Stöbern: Bücher

Ihre Empfehlungen

[Alle Produkte](#)
[Alles gebraucht](#)

Ihre Favoriten

[Bücher](#)
[English Books](#)

Mehr Shops
[Zeitschriften](#)
[Musik](#)
[Klassik](#)

Persönliche Empfehlungen

Hallo, **Norbert Fuhr**. Entdecken Sie die heute vorgestellten Empfehlungen. (Wenn Sie nicht Norbert Fuhr sind, [klicken Sie hier.](#))

Buch-Empfehlungen

The Stone Monkey (Lincoln Rhyme Novels (Paperback))

Amazon.com

When a vicious smuggler known as the Ghost scuttles a ship filled with undocumented Chinese immigrants less than a mile from New York harbor, only a handful of survivors--and the Ghost himself--manage to escape the burning vessel. Lincoln Rhyme, the quadriplegic NYPD forensic detective first... [Mehr dazu](#)

Personalisierung und Kontext

- 

1. [Kinesiologie-Kinder finden ihr Gleichgewicht. Wissenswertes, Spiele, Lieder und Geschichten.](#)
von Barbara Innecken

- 

2. [Atlas der Globalisierung](#)
von Hermann Scheer (Vorwort), u. a.

- 

3. [Brain-Gym mit Maxi \(Kartenspiel\)](#)
von Beate Walter, Haralds Klavinus (Illustrator)

- 

4. [Evolution und Lebenskunst](#)
von Dietmar Hansch

- 

5. [Das Jazzbuch](#)
von Joachim-Ernst Berendt, Günther Huesmann

Google Search History



Search History (Beta) for fuhr@uni-duisburg.de - [Pause](#)

May 6, 2005

[Remove items](#)

[uni-colleg duisburg 2005](#)

[Veranstaltungen des Uni-Colleg - Sommersemester 2005](#) - 11:38am
[www.uni-duisburg-essen.de/presse/events/uni_colleg_somme r2005.shtml](#)

[google suggest](#)

[Google](#) - 11:38am
[www.google.com/webhp?complete=1&h=en](#)

[Norbert Fuhr](#)

[http://is6-www.informatik.uni-dortmund.de/ir/](#) - 11:37am
[Prof. Dr.-Ing. Norbert Fuhr](#) - 11:37am
[www.is.informatik.uni-duisburg.de/staff/fuhr.html](#)

Searches with no clicked results:

[uni-colleg duisburg](#), [personal google](#)

Search Activity May 2005						
S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	1	2	3	4
Today, May 6						

of searches

Total searches: 8

<http://www.google.com/searchhistory/>

Zusammenfassung

- Die besten Suchmaschinen (Google, MSN Search) basieren auf **Page Rank**, bei dem die Link-Struktur des Web berücksichtigt wird.
- Hilfen zur **Verfeinerung der Suche** werden nur wenige angeboten.
- Für spezielle Suchen liefert die Einschränkung auf **Objekttypen** bessere Ergebnisse.

- Ausblick
 - Personalisierte Suche
 - Kontext-bezogene Suche