Vom Suchen und Finden - Google und andere Ansätze

Norbert Fuhr

Universität Duisburg Essen
FB Ingenieurwissenschaften
Abteilung Informatik

3. Juli 2006

Gliederung

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:

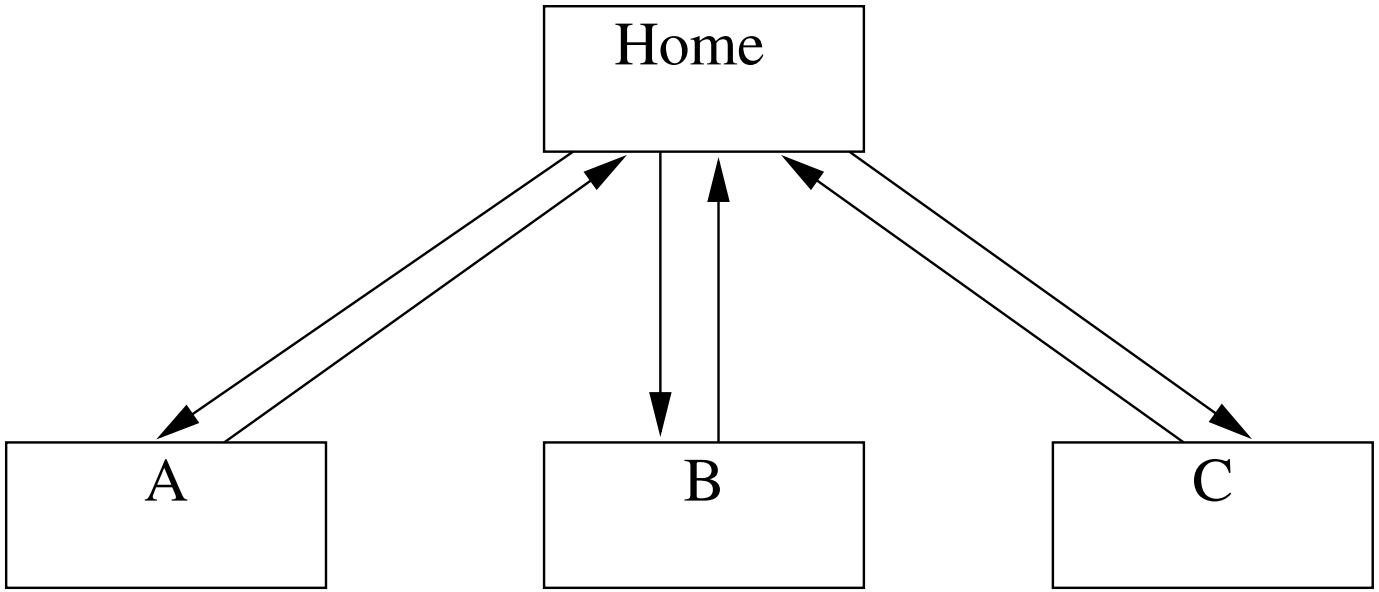
Nur die Dokumente werden ausgegeben, die alle Suchbedingungen erfüllen:

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
- $\longrightarrow$ Page Rank

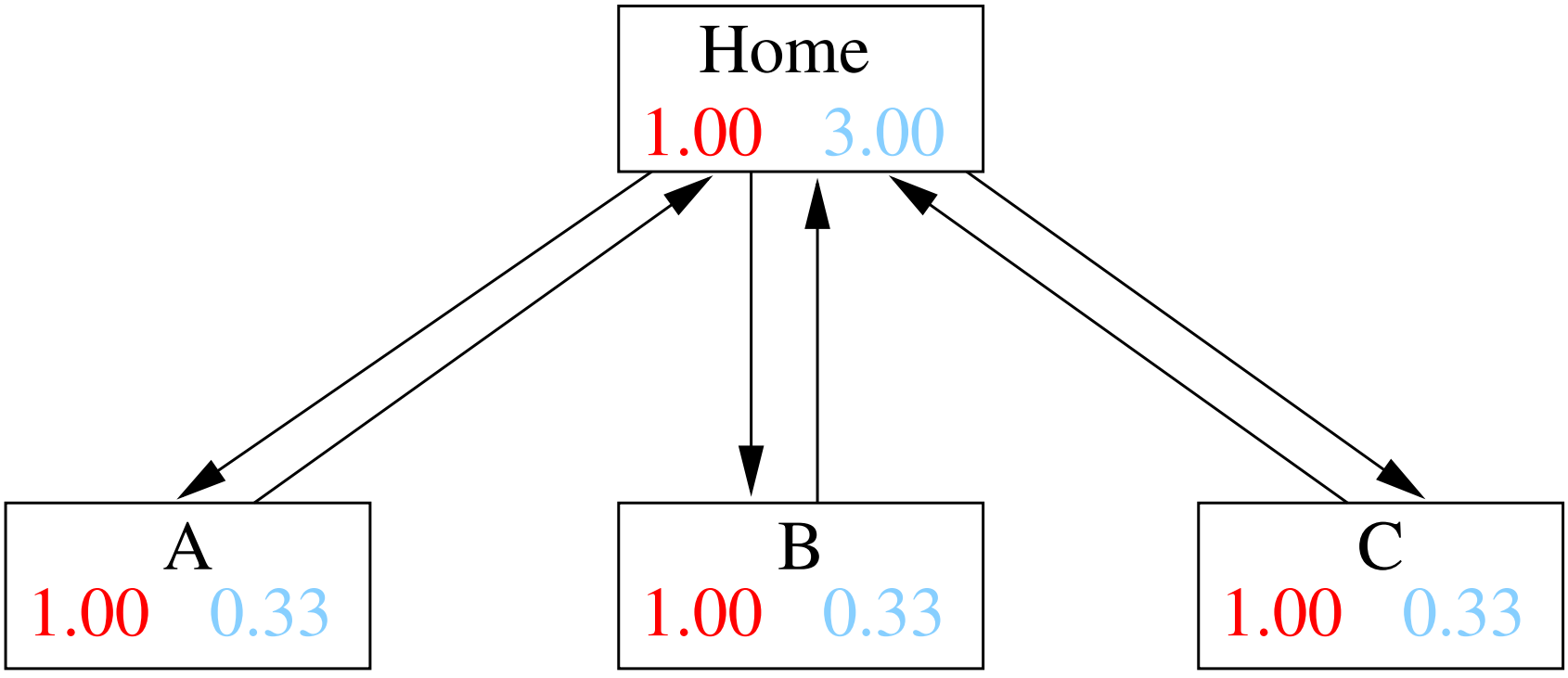
Page Rank: Idee



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

Page Rank-Algorithmus

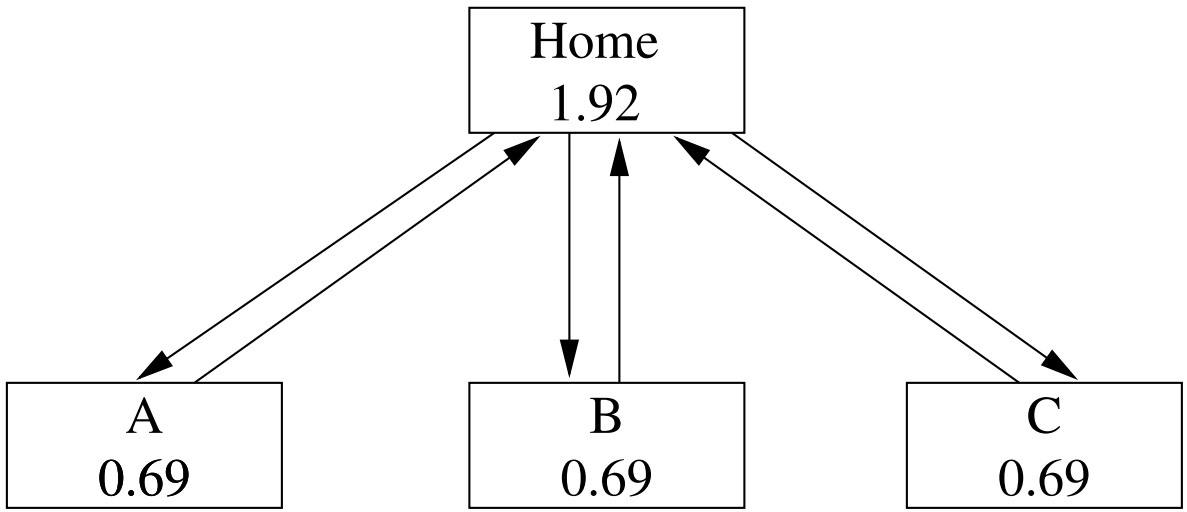
1. Gebe jeder Seite das initiale Gewicht 1
2. Gewicht eines ausgehenden Links = Seitengewicht/Gesamtzahl ausgehender Links
3. Gewicht einer Seite = Summe der Gewichte der eingehenden Links
4. Wiederhole Schritte 2 und 3, bis die Gewichte sich nicht mehr ändern



- 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

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

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

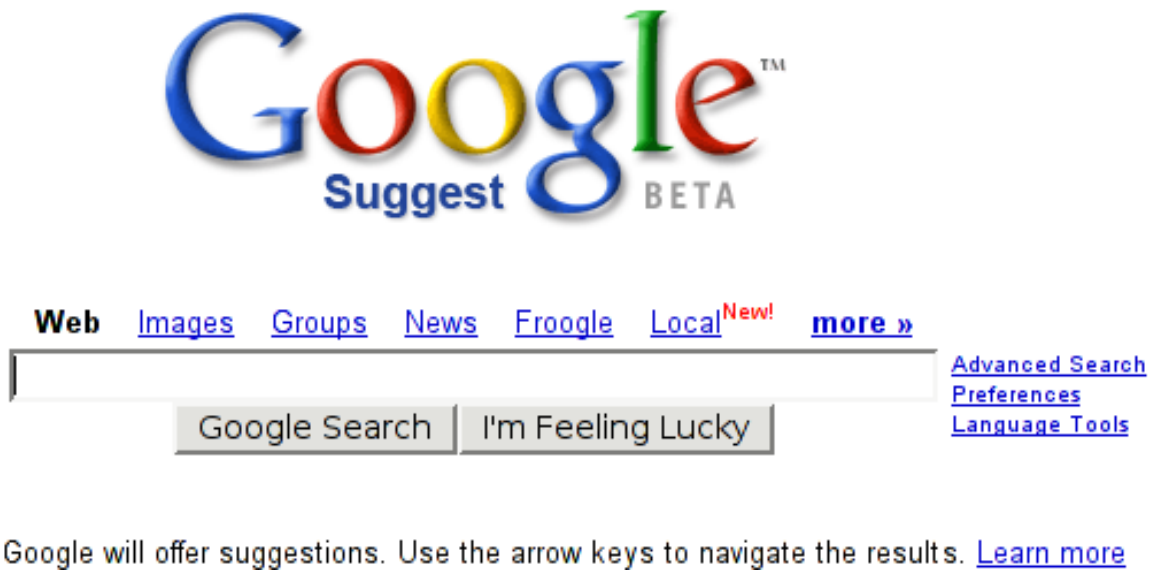
Stemming wird angeblich durch Google und MSN Search unterstützt, lässt sich aber nicht nachvollziehen

Tippfehlerkorrektur

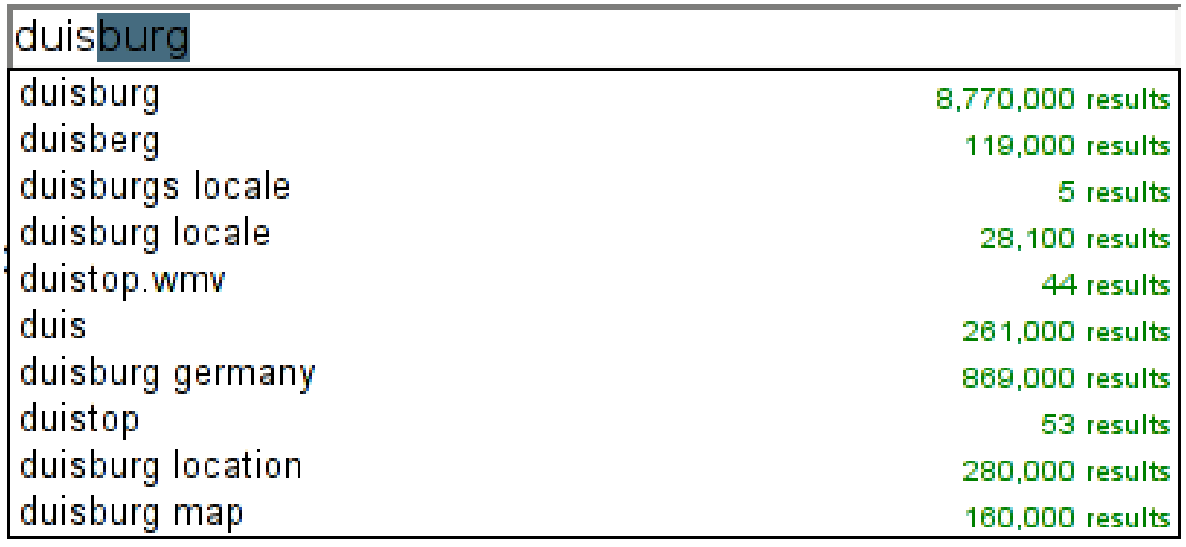
(Seltene) Tippfehler werden von Google erkannt und Vorschläge zur Korrektur gemacht

Google Suggest

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



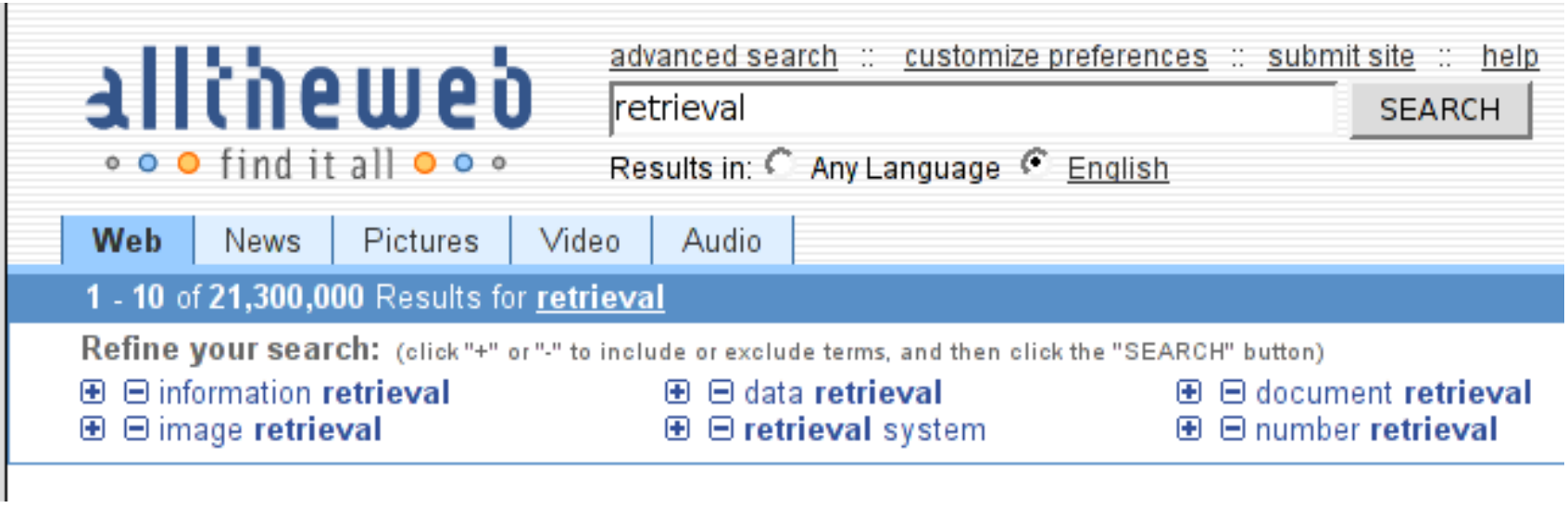
Google Suggest (2)



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

Verfeinerung der Suche

`http://alltheweb.com/`



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?

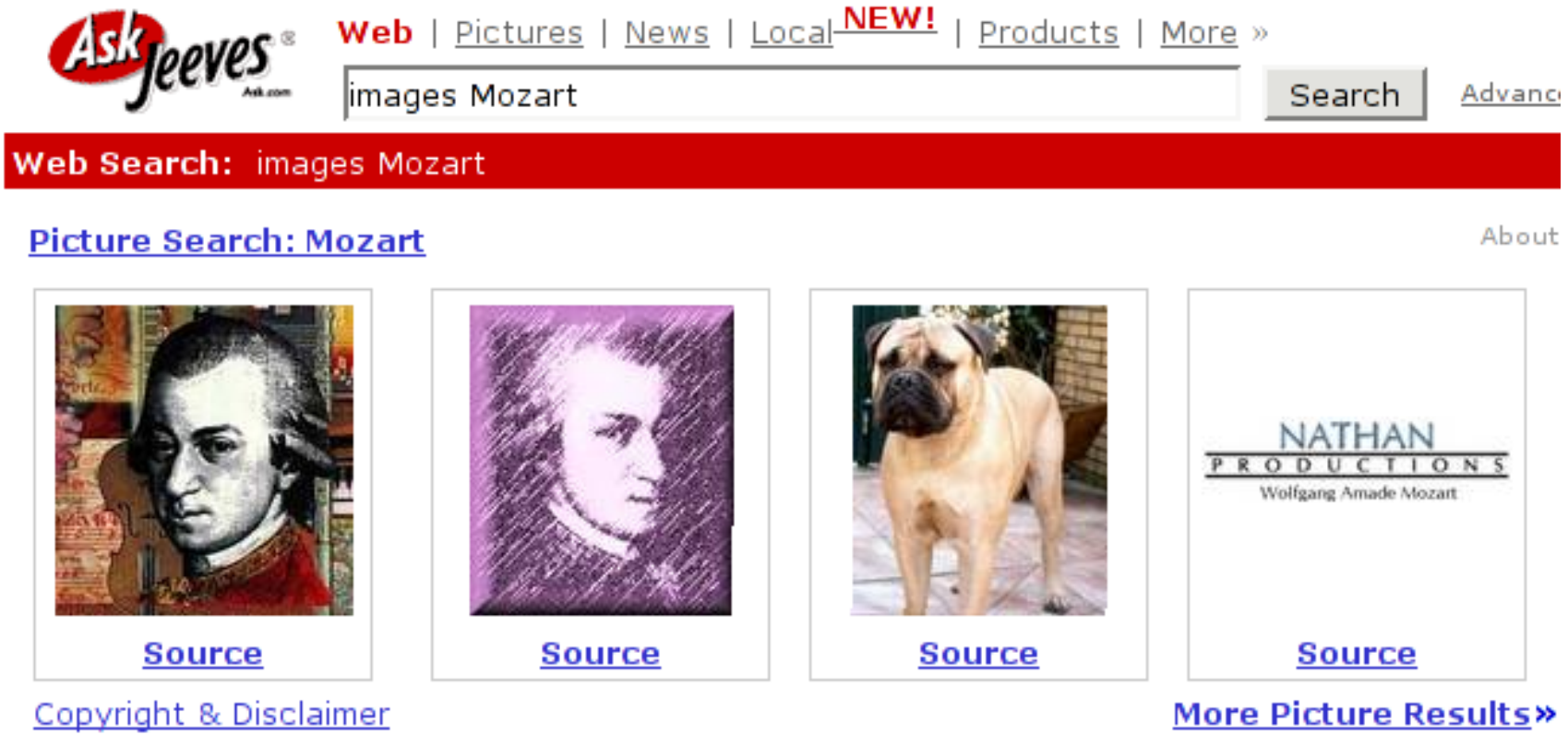
Bessere Suchergebnisse durch Suche nach bestimmten Objekttypen

Objekttypen bei Google

- Objekttypen müssen explizit bei der Anfrage ausgewählt werden

Objekttypen bei AskJeeves

http://www.ask.com



Objekttypen bei AskJeeves (2)

→ 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/`

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 Klavinius (Illustrator)

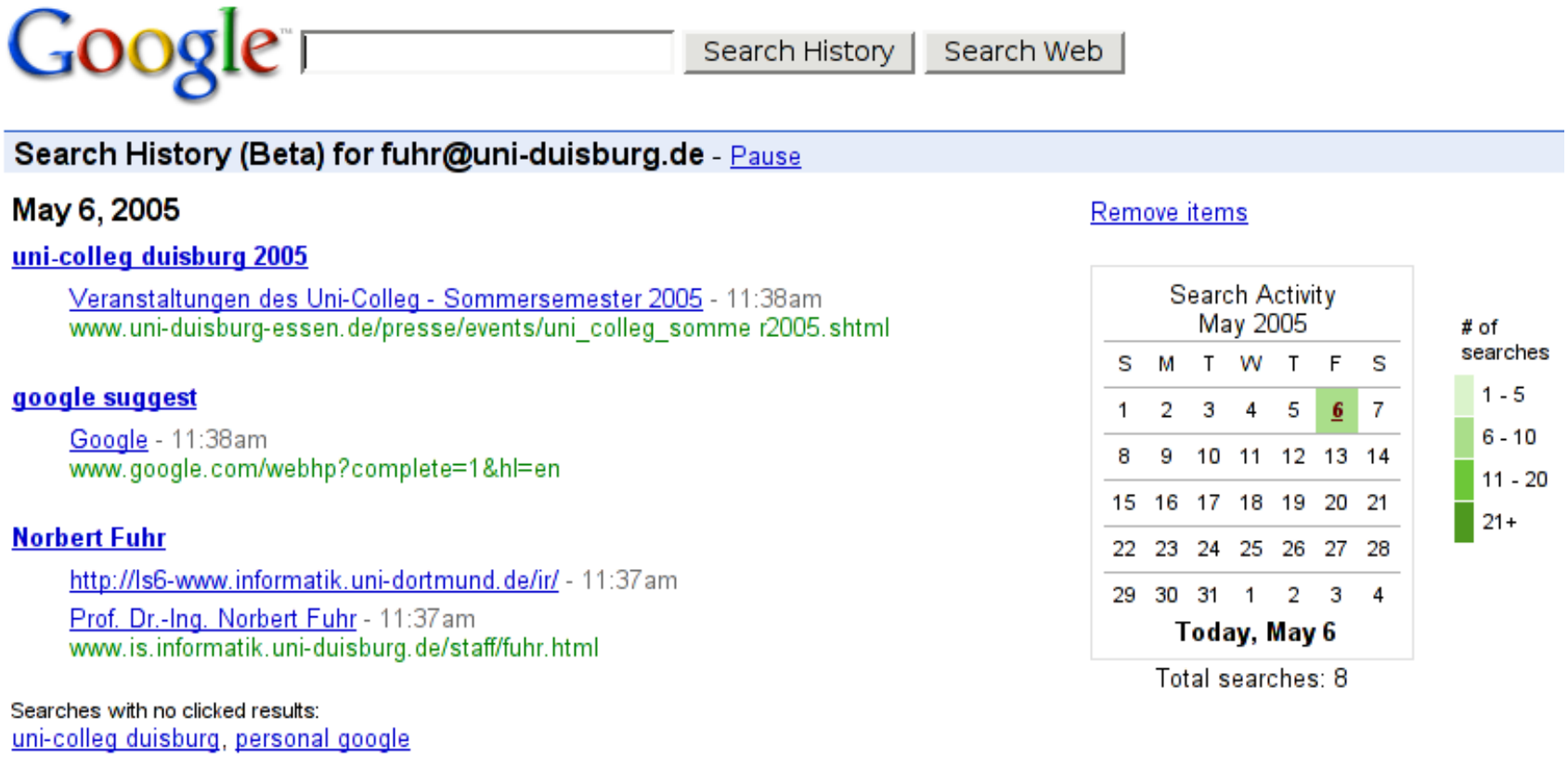
---

4. 
Evolution und Lebenskunst
von Dietmar Hansch

---

5. 
Das Jazzbuch
von Joachim-Ernst Berendt, Günther Huesmann

Google Search History



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