# Gesucht und Gefunden:
# Die Funktionsweise einer Suchmaschine

Prof. Dr. Peter Becker
FH Bonn-Rhein-Sieg
Fachbereich Informatik
peter.becker@fh-bonn-rhein-sieg.de

Vortrag im Rahmen des “Studieninformationstags”
16. Oktober 2007

# Eine Suchmaschine und ihr Einfluß auf die deutsche Sprache


Gefunden im Usenet:

> *Ist vielleicht eine dumme Frage, aber was heißt "google", bzw. wie geht es?*

Die (verkürzte) Antwort:

> *Google (`www.google.com`) ist eine Internet-Suchmaschine.*

# Eigenschaften einer Internet-Suchmaschine

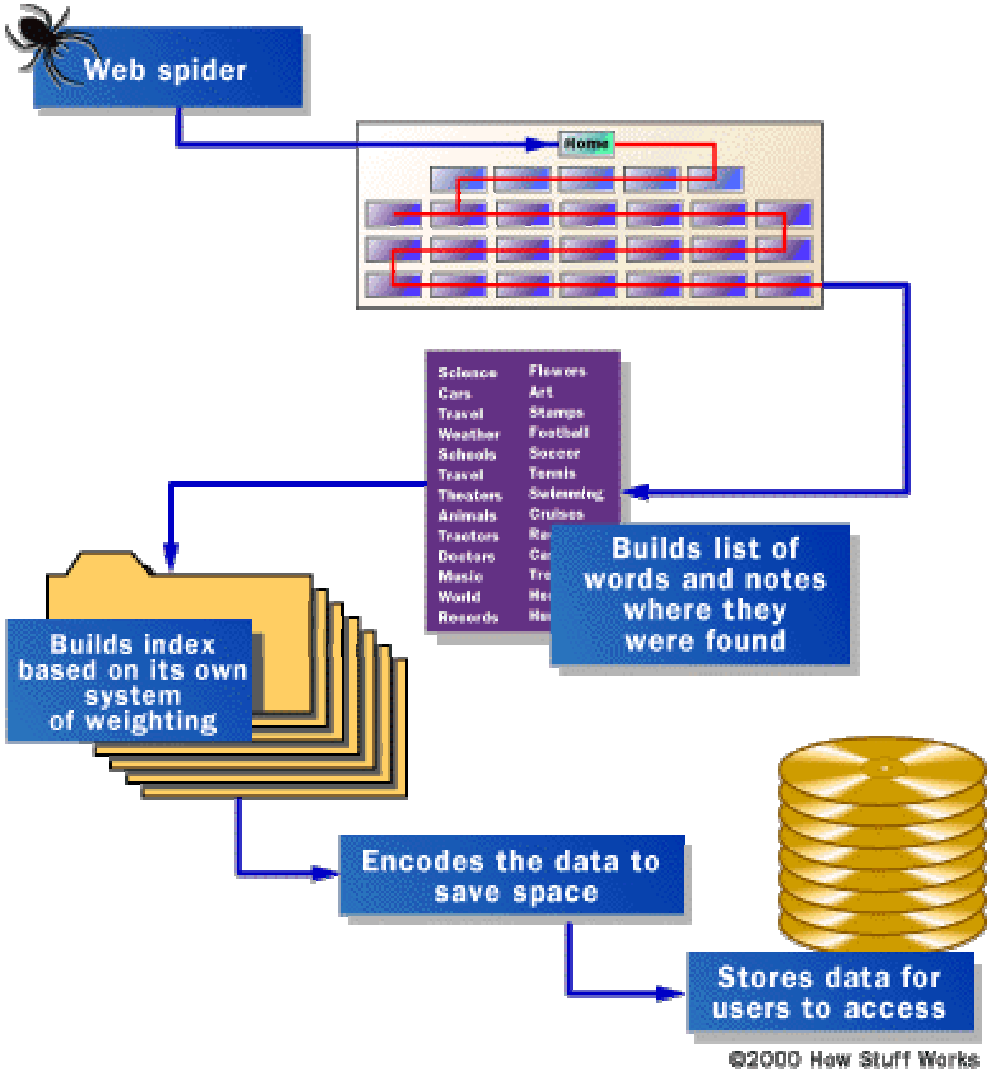
Internet-Suchmaschinen dienen dazu,

- Benutzer bei der Suche nach Informationen zu unterstützen,
- die auf anderen Web-Seiten angeboten werden.

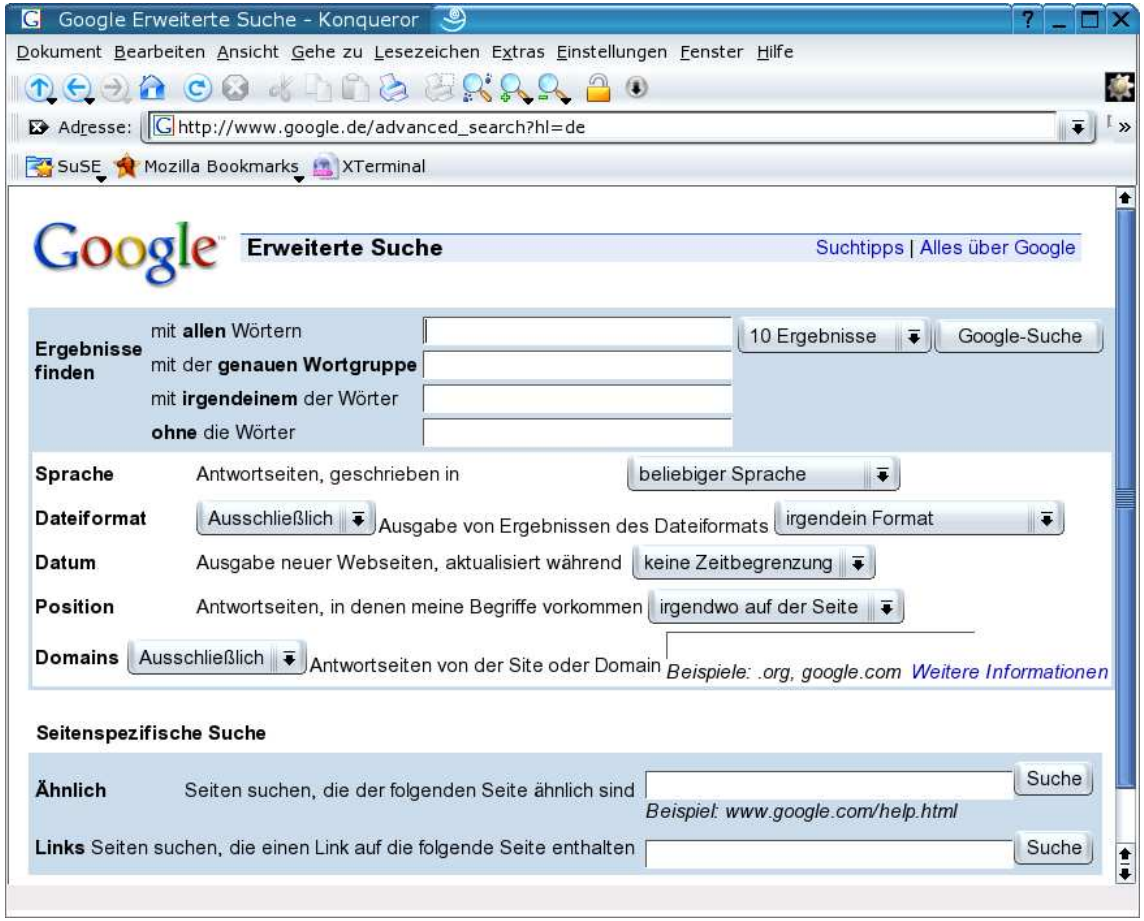
Jede Suchmaschine führt drei wesentliche Dinge durch. ↪

1. Sie durchsucht systematisch das Internet nach Web-Seiten.

2. Sie erstellt einen Index für die in den Web-Seiten enthaltenen Wörter.



3. Sie erlaubt es Benutzern, nach Kombinationen der Wörter aus dem Index zu suchen.

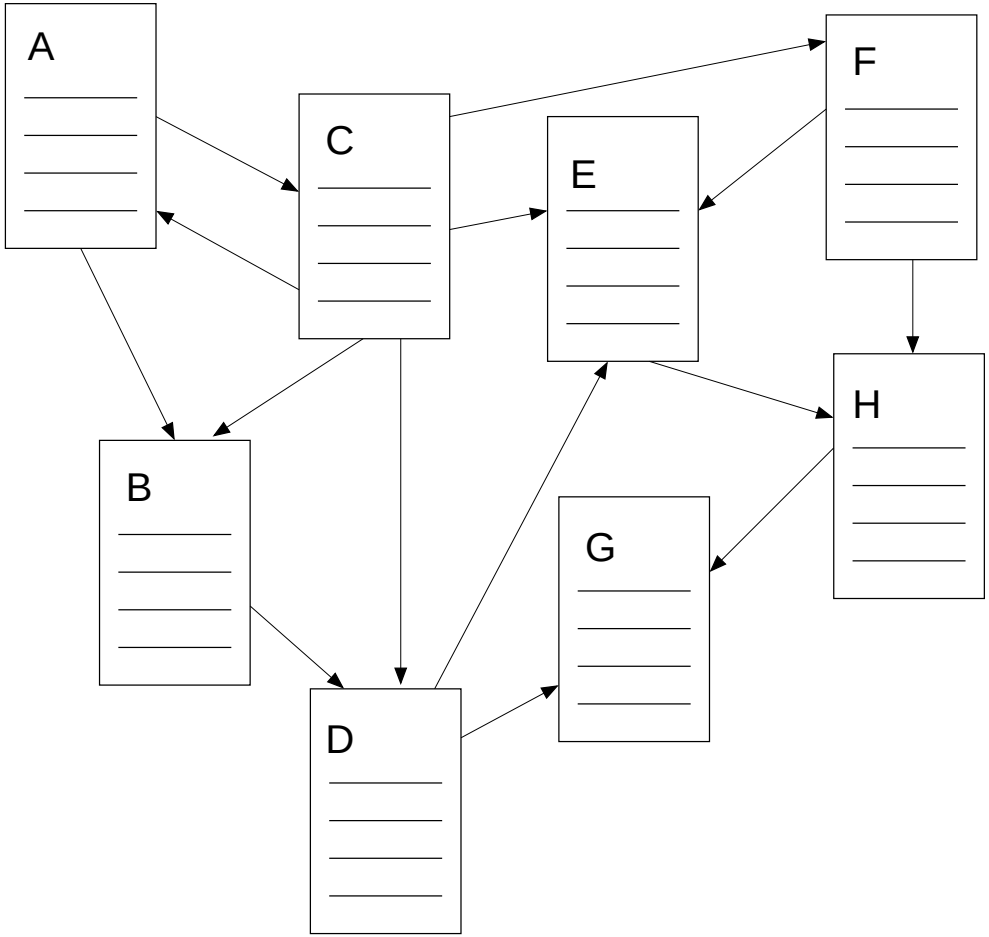


# Web spider: Die Suche nach Web-Seiten ...

... oder: Wie verlaufe ich mich nicht im World Wide Web?

Ausgehend von einem Startknoten sollen alle erreichbaren Web-Seiten gefunden werden.

☞ Graphensuche



# Algorithmus: Breitensuche

1. Stelle die Adresse der Startseite in eine Warteschlange
2. Ist die Warteschlange leer? Wenn ja, dann STOP.
3. Nehme den erste Adresse der Warteschlange, lösche diese aus der Warteschlange und hole die zugehörige Web-Seite aus dem Internet.
4. Durchsuche die Web-Seite nach Adressen von Web-Seiten. Füge Adressen, die noch nicht in der Warteschlange sind oder waren, an diese an.
5. Gehe zu Schritt 2.

# Von Web-Seiten zur Mathematik (1)

Die Sprache der Marsianer besteht nur aus 8 Worten:

ah, fipp, krack, lurks, ru, schnurp, znarg, zong

# Von Web-Seiten zur Mathematik (2)

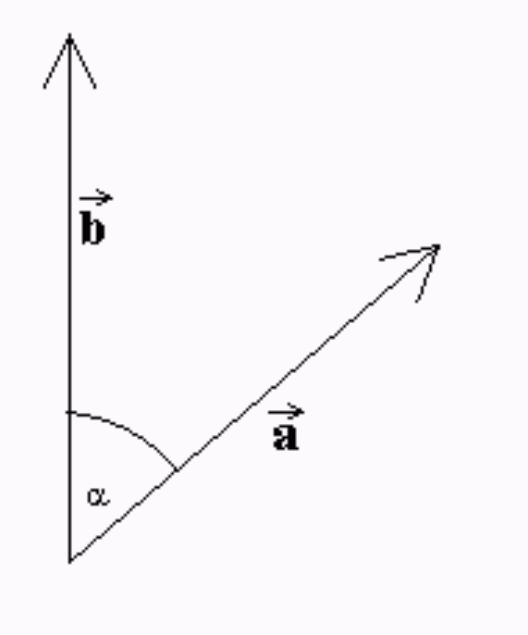
Texte werden als Vektoren repräsentiert.

fipp ah fipp krack ah lurks ah
schnurp ah fipp ah krack krack
krack fipp ah zong ah lurks

$$\begin{pmatrix} 7 \\ 4 \\ 4 \\ 2 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

**Einfacher Ansatz:** Eine Vektorkomponente drückt die Häufigkeit des zugehörigen Wortes in der Web-Seite aus.

# Einschub: Skalarprodukt




$$\vec{a} \cdot \vec{b} = |a| \cdot |b| \cdot \cos \alpha$$

$|a|$ ist der Betrag von $\vec{a}$,
$|b|$ ist der Betrag von $\vec{b}$ und
$\alpha$ ist der Winkel zwischen $\vec{a}$ und $\vec{b}$.

Für $\vec{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ und $\vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$

ist $\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i \cdot b_i$

und $|a| = \sqrt{a_1^2 + \cdots + a_n^2}$.

# Von Web-Seiten zur Mathematik (3)

Je stärker zwei Vektoren zu Web-Seiten in die gleiche Richtung zeigen, desto ähnlicher sind die zugehörigen Web-Seiten.

Mit dem Cosinusmaß können wir die Ähnlichkeit von zwei Webseiten $a$ und $b$ sogar messen:

$$\cos(a, b) = \frac{\vec{a} \cdot \vec{b}}{|a| \cdot |b|}$$

a | fipp ah fipp krack ah lurks ah schnurp ah fipp ah krack krack krack fipp ah zong ah lurks

b | znarg ah fipp krack ah ru krack krack krack lurks ah schnurp ah fipp fipp ah znarg ah lurks

c

znarg ru znarg krack ru lurks ru schnurp ru znarg ru krack krack krack znarg ru zong ru lurks

$$a = \begin{pmatrix} 7 \\ 4 \\ 4 \\ 2 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, b = \begin{pmatrix} 6 \\ 3 \\ 4 \\ 2 \\ 1 \\ 1 \\ 2 \\ 0 \end{pmatrix}, c = \begin{pmatrix} 0 \\ 0 \\ 4 \\ 2 \\ 7 \\ 1 \\ 4 \\ 1 \end{pmatrix}$$

$$\cos(a, b) = \frac{75}{\sqrt{87} \cdot \sqrt{71}} = 0.954$$

$$\cos(a, c) = \frac{36}{\sqrt{87} \cdot \sqrt{87}} = 0.413$$

# Von Web-Seiten zur Mathematik (4) ...

Aber wir wollen doch gar nicht Web-Seiten untereinander vergleichen, sondern Web-Seiten mit einer Anfrage!

Kein Problem: Hierzu betrachten wir eine Anfrage auch als Web-Seite und überführen diese in einen Vektor.

q | ah ah ah ah fipp fipp lurks

$$q = \begin{pmatrix} 4 \\ 2 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

# ... um eine Rangordnung der Web-Seiten zu berechnen!

$$q = \begin{pmatrix} 4 \\ 2 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, a = \begin{pmatrix} 7 \\ 4 \\ 4 \\ 2 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, b = \begin{pmatrix} 6 \\ 3 \\ 4 \\ 2 \\ 1 \\ 1 \\ 2 \\ 0 \end{pmatrix}, c = \begin{pmatrix} 0 \\ 0 \\ 4 \\ 2 \\ 7 \\ 1 \\ 4 \\ 1 \end{pmatrix}$$

$$\cos(q, a) = \frac{38}{\sqrt{21} \cdot \sqrt{87}} = 0.89 \qquad \cos(q, b) = \frac{32}{\sqrt{21} \cdot \sqrt{71}} = 0.83$$

$$\cos(q, c) = \frac{2}{\sqrt{21} \cdot \sqrt{87}} = 0.05$$

Anfrage:

ah ah ah ah fipp fipp lurks

1. fipp ah fipp krack ah lurks ah schnurp ah fipp ah krack krack krack fipp ah zong ah lurks

Ähnlichkeit = 0.89

2. znarg ah fipp krack ah ru krack krack krack lurks ah schnurp ah fipp fipp ah znarg ah lurks

Ähnlichkeit = 0.83

3. znarg ru znarg krack ru lurks ru schnurp ru znarg ru krack krack krack znarg ru zong ru lurks

Ähnlichkeit = 0.05

# Nicht mehr als so ein bisschen Mathe?

Natürlich mehr! Bei der Aufstellung der Vektoren berücksichtigen Suchmaschinen z.B. auch ...

- ... die Selektivität von Worten:

  **nicht selektiv:** der, die, das, ein, wie, weil, wenn, ...
  **wenig selektiv:** Computer, Internet, Informatik, Rechner, ...
  **selektiv:** HTML, Browser, Router, Modem, Linux, ...
  **stark selektiv:** Quicksort, Vektorraum, Retrieval, ...

- ... wo das Wort auftritt:

  **ausgezeichnete Stellen:** Titel, Verweise, ...
  **hervorgehobene Stellen:** fett, kursiv, größer, ...
  **normaler Text**

# Was haben wir gelernt?

- Algorithmen sind so etwas wie exakt definierte Rechenverfahren.

  Sie sind ein wesentlicher Bestandteil der Informatik und und ihre Beherrschung ist ein wesentliches Lernziel des Informatikstudiums.

- Mathematik steckt manchmal in Dingen, in denen man sie gar nicht vermuten würde.

- Modellierung und Abstraktion ist wichtig!

  Wir modellieren Texte als Vektoren. So können wir mathematische Operationen nutzen, um mit Texten umzugehen. Die Ergebnisse (hier Rangliste) übertragen wir auf unser Problem und erhalten so eine Lösung.