max planck institut informatik

Ideen und Konzepte der Informatik

Websuche

Kurt Mehlhorn

max planck institut informatik

Suchmaschinen

- Google seit 1998, Konkurrenten Bing, DuckDuckGo
- Altavista etwas früher.
- Google: 4 Mio. Anfragen / Minute.
- 90% Marktanteil in Deutschland, Quasimonopol (Anzeigenpreise, Markteinstieg für Konkurrenten, Standard bei Android, vertikale Suchmaschinen)

Ich erkläre die Grundzüge der Google-Suchmaschine: keine Personalisierung, keine Tagesnachrichten, …

Websuche

- **Eingabe:** einige Worte, z. B. Kurt Mehlhorn.
- **Ausgabe:** die **wichtigsten** Webseiten, die die Schlüsselwörter enthalten (oder die profitabelsten)
- **Qualitätsmaß:** Nutzerzufriedenheit.

- Webseiten bestehen aus Inhalt und Verweisen; Content und Links.

Beispiel: Google-Suche nach Kurt Mehlhorn in 2020

About 284.000 results (0,77 seconds), Ausgabe beginnt mit:

Kurt Mehlhorn (MPI-INF) - Homepage
www.mpi-inf.mpg.de › ~mehlhorn
Kurt Mehlhorn. Algorithms and Complexity Group · Max-Planck-Institut für Informatik · Saarland Informatics Campus Campus E1 4 66123 Saarbruecken

Mehlhorn, Kurt | Max-Planck-Gesellschaft
www.mpg.de › informatik_wissM2
Translate this page
Kurt Mehlhorn ist Direktor und Wissenschaftliches Mitglied am Max-Planck-Institut für Informatik (seit 1990)

Kurt Mehlhorn - Google Scholar
scholar.google.com › citations
Professor of Computer Science, Max Planck Institute for Informatics, Saarland Informatics Campus - Cited by 22519 - Algorithms - Computational Geometry

Kurt Mehlhorn - Wikipedia
en.wikipedia.org › wiki › Kurt_Mehlhorn

Websuche nach Autoversicherung

About 1.580.000 results (0,54 seconds)

Ads

CHECK24: Autoversicherung 2020 - Wer vergleicht, spart mehr
Ad·www.check24.de/
CHECK24: Bester **Autoversicherung** Vergleich. TÜV "sehr gut". eVB bis zu 360 Tage gültig! Beim 15-fachen Testsieger bis zu 850 € sparenl.

DA Direkt Autoversicherung - eVB-Nummer sofort online
Ad·www.da-direkt.de/
069 71158180
Jetzt zum Testsieger wechseln und 10% Online-Rabatt erhalten. Nur noch 6 Tage. eVB-Nummer sofort online bei Abschluss.

Kfz-Versicherung ab 4,10€/Mon. - Allianz Direct
Ad·www.allianzdirect.de/
Berechne jetzt deinen günstigen Beitrag zur Kfz-Versicherung. Einfach wechseln und...

Search Results

Web results

Autoversicherung - HUK Coburg
www.huk.de › kfz-versicherung › au...
Translate this page
Ihre günstige **Autoversicherung**: Sie sparen 20% mit Kasko SELECT! ✓ Jetzt online Kfz-Versicherung berechnen & abschließen!

Wichtige Anmerkung

- Existierende Suchmaschinen (Google, Bing, …) haben noch wenig Textverständnis:
  - Suche nach Kurt Mehlhorn Ehefrau — kein Ergebnis
  - Suche nach Kurt Mehlhorn married to — Ena Mehlhorn
- Sie finden Webseiten, die gegebene Suchworte (search keys) enthalten und ordnen diese geschickt an (das ist die Leistung).
- Aktuelle Forschung: Textverständnis.

Drei Fragen

1. Woher kennen Suchmaschinen so viele Webseiten?

2. Wie finden Suchmaschinen die Webseiten, die Kurt und Mehlhorn enthalten?
   - Wie Seiten, die Mehlhorn enthalten?
   - Wie Seiten, die Kurt und Mehlhorn enthalten?

3. Wie finden sie die wichtigen Webseiten? (Fachbegriff für wichtig = relevant)

4. Basisalgorithmen: Suchen und Sortieren.

Wiederholung: Suchen und Sortieren

- Binärsuche findet ein Wort in einer geordneten Menge der Größe $2^k - 1$ mit k Vergleichen. 30 Vergleiche bei einer Menge der Größe eine Milliarde.

- Sortieren durch Mischen sortiert n Elemente mit 2n log n Vergleichen. Eine Million (2^{20}) Elemente mit $2 \cdot 2^{20} \cdot 20$ = 40 Millionen Vergleichen (weit weniger als 1 Sekunde auf meinem Notebook).

Drei Fragen

1. Woher kennen Suchmaschinen so viele Webseiten?

2. Wie finden Suchmaschinen die Webseiten, die Kurt und Mehlhorn enthalten?
   - Wie Seiten, die Mehlhorn enthalten?
   - Wie Seiten, die Kurt und Mehlhorn enthalten?

3. Wie finden sie die wichtigsten Webseiten? (Fachbegriff für wichtig = relevant)

4. Basisalgorithmen: Suchen und Sortieren.

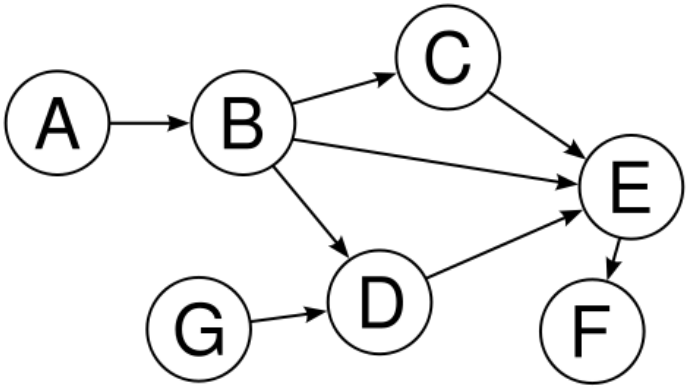
Web Crawler

- Kriechen übers Netz, indem sie von ein paar Startseiten (Seed Pages) ausgehend systematisch Verweisen (Links) folgen.
- Schicken eine Kopie jeder besuchten Seite zum Organisator des Webcrawls.
- **Ergebnis:** Google hat eine Kopie des ganzen erreichbaren Webs (mehrere Milliarden Seiten).

Graphen

Ein *Graph* besteht aus Knoten und Kanten.

Eine Kante verbindet zwei Knoten. Sie ist entweder gerichtet (Einbahnstraße) oder ungerichtet.



Straßennetzwerke, Firmengeflechte, Webgraph, Freundschaftsbeziehungen, Abhängigkeit von Aufgaben,… kann man als Graphen darstellen.

Systematische Durchmusterung

A ← Menge der Saatknoten

Solange es eine Kante (u,v) gibt mit u in A und v nicht in A

    Füge v zu A hinzu

Findet alle Knoten, die von den Saatknoten aus erreichbar sind. Saatknoten könnte Startseite von Wikipedia sein.

Statt Kante sagt man auch Verweis oder Link.

Anordnung nach Relevanz

- Suchmaschinen haben eine Kopie des erreichbaren Webs.
- Sie nummerieren die Webseiten nach ihrer Wichtigkeit durch. Wie das geht, lernen wir später.
- Analogie: Die wichtigsten Bücher der Weltliteratur.

Die zweite Frage

- Wie kann man Seiten finden, die Kurt und Mehlhorn enthalten?
  - Wie Seiten, die Mehlhorn enthalten?
  - Wie Seiten, die Kurt und Mehlhorn enthalten?

- Dazu: Vorkommen von Worten in Texten und Vorkommenslisten.

Vorkommen von Worten in Texten

- **Text:** Adrian und Kurt unterrichten gemeinsam und …

Sortieren der vorkommenden Worte ergibt:

- Adrian gemeinsam Kurt und und unterrichten

Nun kann man leicht für jedes Wort die Anzahl der Vorkommen bestimmen.

Vorkommenslisten

- **Text1:** Adrian und Kurt unterrichten und …
- **Text2:** Adrian forscht

- Erzeuge Paare (Adrian 1), (und 1), …, (Adrian 2), … und sortiere

- (Adrian 1), (Adrian 2), (forscht 2), (Kurt 1), …

- Extrahiere Vorkommenslisten, etwa Adrian: 1 2

  Kurt: 1

Geordnete Vorkommenslisten

- Für **jedes mögliche Suchwort** (jedes Wort im Duden, Eigennamen, …) schreibt man auf, in welchen Dokumenten es vorkommt (> 1 Mio. Listen).

- Kurt: 94, 113, 217, 405, ….
- Mehlhorn: 20, 113, 405, 602, ….
- Kosta: 27, 405, ….

- Kleine Zahlen = wichtige Dokumente

Suche nach Mehlhorn

- Finde V-liste von Mehlhorn
  (Binärsuche in der Menge aller V-Listen)

Mehlhorn: 20, 113, 405, 602, ….

- und gib sie aus (genauer: gib eine Kurzfassung der Dokumente mit diesen Nummern aus und Verweise auf das vollständige Dokument).

Suche nach Kurt Mehlhorn

- Finde V-listen von Kurt und von Mehlhorn

  (Binärsuche)

Kurt: 94, 113, 217, 405, ….
Mehlhorn: 20, 113, 405, 602, ….

- Bestimme die gemeinsamen Einträge und gib sie aus:
  113, 405, …. Mischen der beiden Listen.

Geht das wirklich so schnell?

- *Oxford English Dictionary:* 616,500 words
  - Binärsuche braucht log 616,500 ≤ 20 Schritte

- Kurt: 240 000 000 Dokumente, 0.14 sec
- Mehlhorn: 1 560 000 Dokumente, 0.14 sec
- Kurt Mehlhorn: 592 000 Dokumente, 0.33 sec

- V-Listen sind lang, aber man braucht nur die ersten 10 gemeinsamen Einträge; man findet sie durch Mischen der beiden Listen.

Wie viel Platz braucht man?

- Zeit geht, wie steht es mit Speicherplatz?

- 10^7 Schlagworte, je mit einer V-liste der Länge ca. 10^6
- Gesamtlänge < 10^{13} Zahlen.
- Mein Notebook kann 4.0 10^9 Zahlen speichern (150 Gbyte Platte).
- 2500 kleine Rechner reichen.

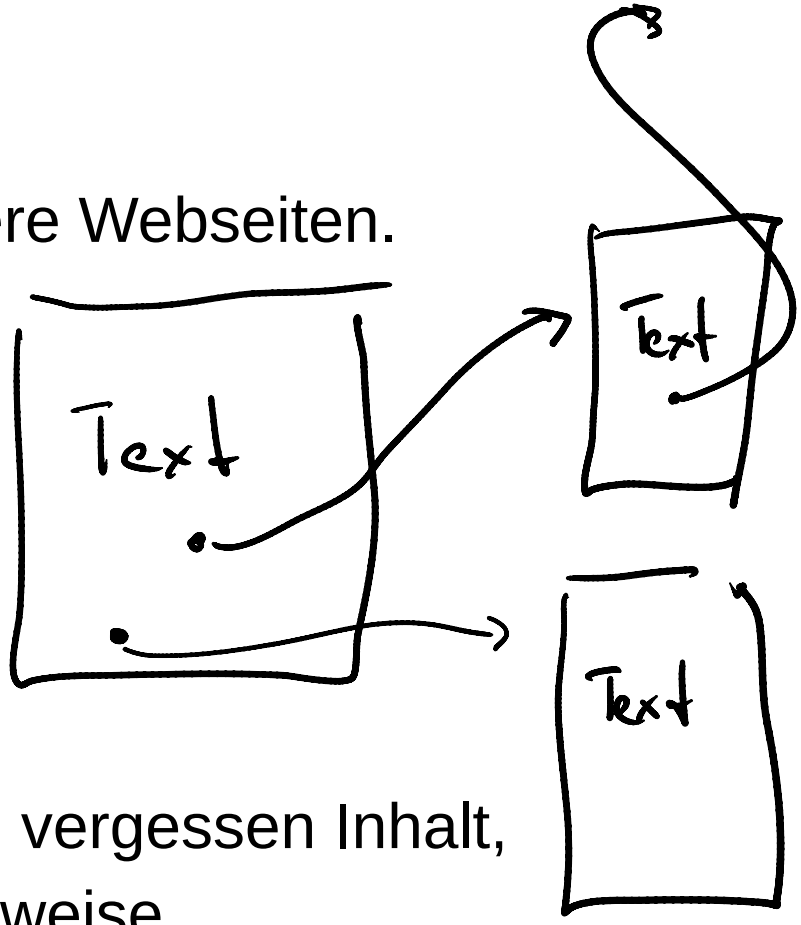
Anordnung nach Relevanz

- Wie ordnet man eine Milliarde Webseiten nach ihrer Relevanz? Was ist das wichtigste Buch?

- **Zentrale Idee:** Ignoriere den Inhalt und konzentriere dich auf die Links.

Gestalt einer Webseite

- Text und Verweise (Links).
- Die Links verweisen auf andere Webseiten.



- **Bestimmung von Relevanz:** vergessen Inhalt, konzentrieren uns auf die Verweise.

Das Prinzip von Pagerank

Eine Seite ist wichtig, wenn wichtige Seiten auf sie zeigen.

Ein Mensch ist wichtig, wenn wichtige Leute ihn für wichtig halten.

Jon Kleinberg (98),

Sergey Brin / Larry Page (98)

Vom Ergebnis her denken

- b_w = Relevanz der Seite w.

- Wir tun so, als ob wir schon wüssten, dass es diese Größe gibt, und fragen uns nach ihren Eigenschaften, etwa:
  - Wenn ich Relevanz b habe und auf 5 andere Seiten zeige, dann gebe ich an jede Relevanz weiter.

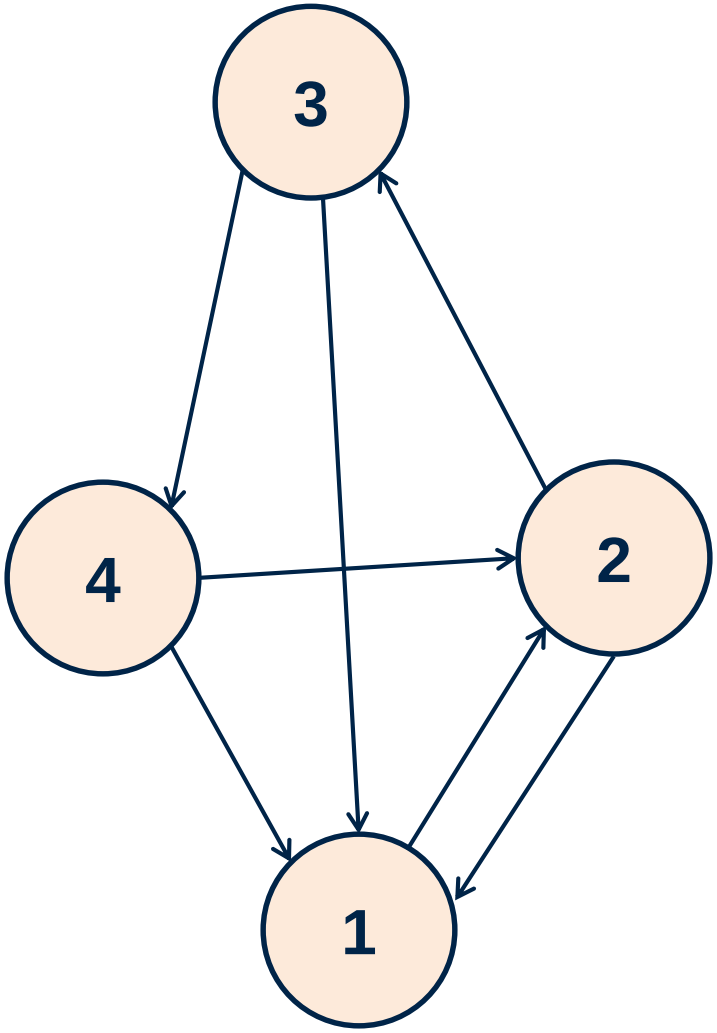
Etwas genauer

b_w = Wichtigkeit der Seite w.

- Jede Seite w gibt an jeden Nachfolger den gleichen Bruchteil seiner Wichtigkeit weiter.
  - (also bei 3 Nachfolgern, jedem)
- Jeder Knoten sammelt die ihm mitgeteilte Wichtigkeit auf; w sammelt s_w auf.

Forderung: $b_w = s_w$

Beispiel



$$b_1 = s_1 =$$

$$b_2 = s_2 = b_1 + \frac{b_4}{2}$$

$$b_3 = s_3 = \frac{b_2}{2}$$

$$b_4 = s_4 = \frac{b_3}{2}$$

$$b_1 = \frac{7}{21}, \quad b_2 = \frac{8}{21}, \quad b_3 = \frac{4}{21}, \quad b_4 = \frac{2}{21}$$

Wie berechnen?

1. Man stellt das Gleichungssystem auf und löst es: sehr aufwendig.

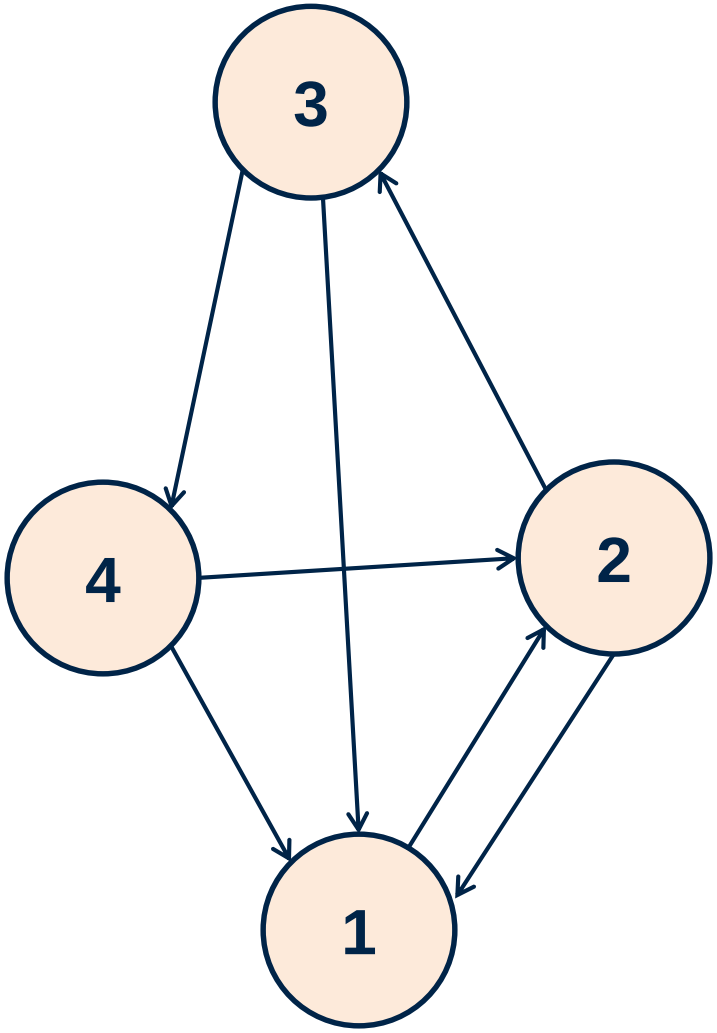
2. Man simuliert das System.

Simulation

- Gib jedem Knoten 1000 Wichtigkeitspunkte.
- Tue wiederholt
  - Jeder Knoten verteilt seine Wichtigkeitspunkte gleichmäßig auf seine Nachfolger.

b_w = Anzahl der Wichtigkeitspunkte nach vielen Simulationsschritten (normalisiert).

Beispiel für Simulation



$$b_1 = \frac{7}{21}, \quad b_2 = \frac{8}{21}, \quad b_3 = \frac{4}{21}, \quad b_4 = \frac{2}{21}$$

Werbung

- Neben den Antworten der Suchmaschine gibt es auch noch „bezahlte Antworten (= Anzeigen)“.
- Hier bezahlen Firmen die Suchmaschine dafür, dass bei bestimmten Suchwörtern bestimmte Anzeigen gezeigt werden, etwa
  - Anfrage Auto führt zu Anzeige von autoscout24.de.
  - Wenn Nutzer auf die Anzeige klickt, wird die Suchmaschine bezahlt. Im Dollarbereich.
- Anzeigenplätze werden in einer Auktion versteigert, siehe Kurs Auktionen und verteiltes Entscheiden.

Prinzipien der Websuche – Zusammenfassung

- Dokumente werden nach Wichtigkeit geordnet.
- Wichtigkeit wird in einem selbstreferentiellen Prozess bestimmt.
- geordnete V-Liste für jedes Schlagwort.
- Suche: Finde V-Liste für jedes Schlagwort in der Frage und bilde Durchschnitt. Gib Dokumente in Reihenfolge aus.

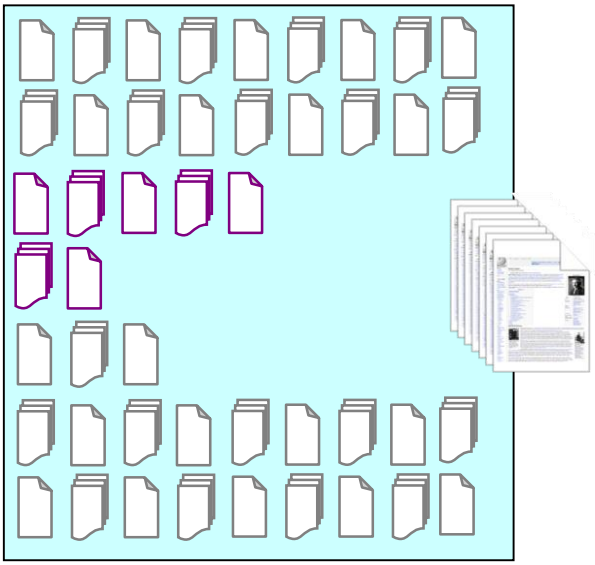
Aktuelle Forschung

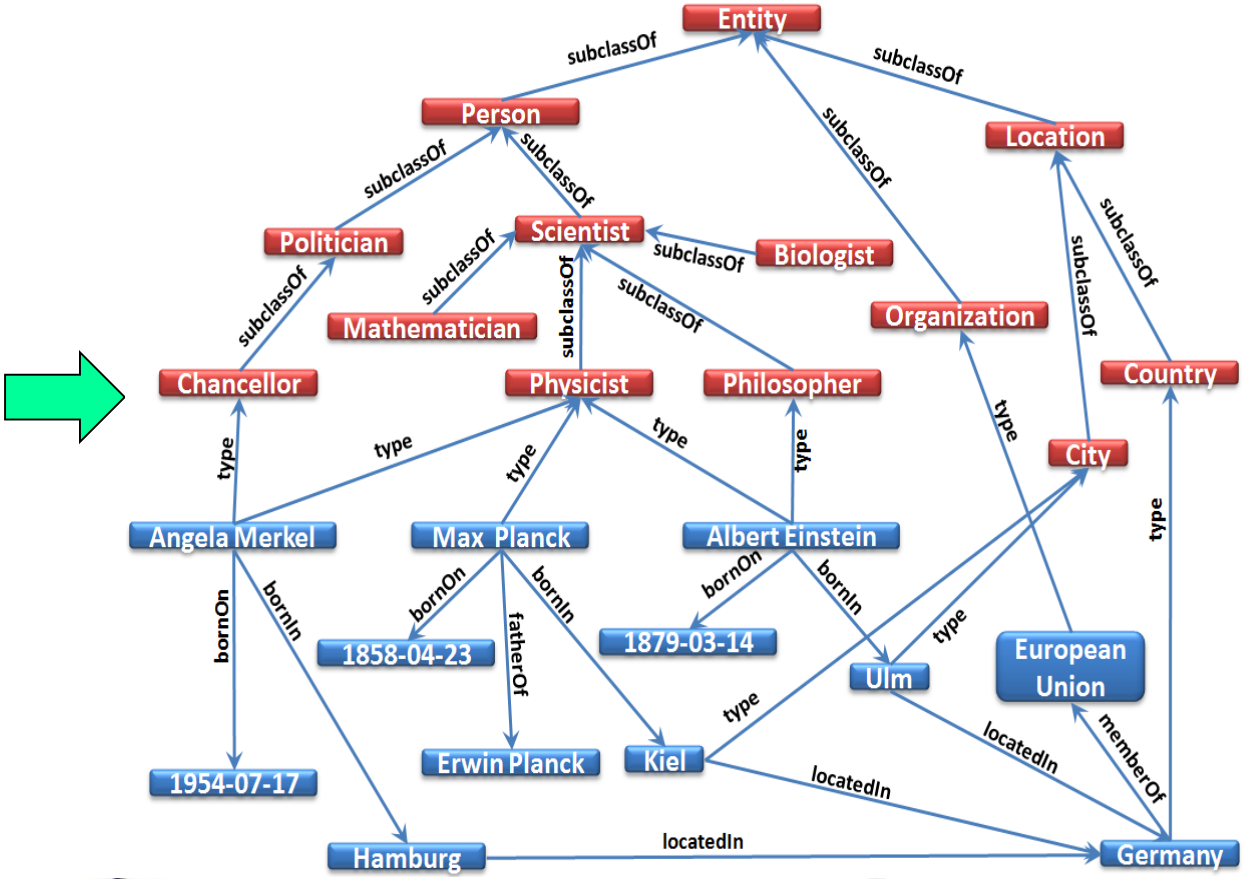
- Gerhard Weikum, MPI für Informatik
- Von Information zu Wissen

Schritt 1

- Benutze WordNet Kategorien:
  - Mann ≤ Mensch ≤ Säugetier ≤ Tier
- Sammle Fakten:
  - KM ist Informatiker, KM geboren in Ingolstadt, KM verheiratet mit Ena, KM geboren 1949, KM Direktor MPI-INF, …
  - Beginne mit Wikipedia Infoboxen.
  - Dann einfache Aussagesätze in Texten.
- Großes Problem: Konsistenz.

Approach: Harvesting Facts from Web





Cyc SUMO

TextRunner

YAGO-NAGA

DBpedia

SIGMA SEMANTIC INFORMATION MASHUP

IWP

umbel

True Knowledge

Carnegie Mellon

DBLife

WikiTax2WordNet

WolframAlpha

ReadTheWeb

freebase

Beantwortung komplexer Fragen

- Wer war deutscher Nationaltrainer als Schweinsteiger geboren wurde?
  - Finde Geburtsjahr von Schweinsteiger
  - Finde Deutschen Nationaltrainer in diesem Jahr
- Was haben Manfred Pinkal, Michael Dell und Renee Zellwenger gemeinsam?
  - Finde ein X, mit dem Pinkal, Dell und Zellwenger in Relation stehen (born-in, lebt, arbeitet, studiert, verheiratet-mit)
- Politiker, die auch Wissenschaftler sind
  - Finde ein X, das sowohl Politiker als auch Wissenschaftler ist

…

Jeopardy! (dt. Gefahr)

- US Quizshow
- 3 Spieler
- Quizmaster stellt Fragen, Spieler drücken Buzzer
- Richtige (falsche) Antworten werden belohnt (bestraft)
- In 2011, IBMs Watson gewinnt.

- Its largest airport is named for a World Word II hero; its second largest, for a World War II battle.
- Almost exactly equal to the mass of 1000 cubic centimeters of water; it is a base unit in the metric system.
- Just add 273.15 to your Celsius readings to get this.

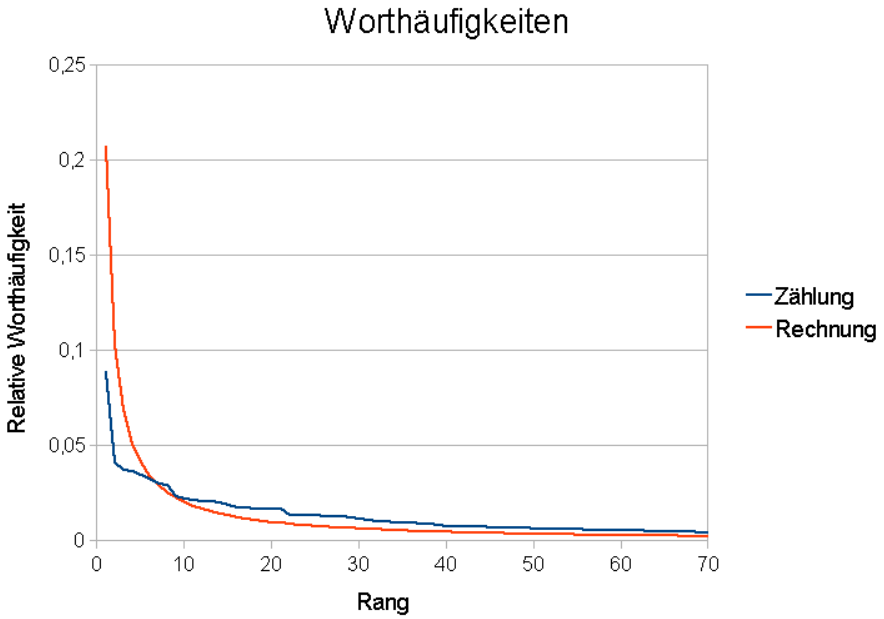
ENDE

Große Textkorpora

- 30 Formen stellen 31,8 % der Wörter:
  - die, der, und, in, zu, den, das, nicht, von, sie, ist, des, sich, mit, dem, dass, er, es, ein, ich, auf, so, eine, auch, als, an, nach, wie, im, für
- Weitere 70 Formen stellen weitere 15,3 % der Wörter:
  - man, aber, aus, durch, wenn, nur, war, noch, werden, bei, hat, wir, was, wird, sein, einen, welche, sind, oder, zur, um, haben, einer, mir, über, ihm, diese, einem, ihr, uns, da, zum, kann, doch, vor, dieser, mich, ihn, du, hatte, seine, mehr, am, denn, nun, unter, sehr, selbst, schon, hier, bis, habe, ihre, dann, ihnen, seiner, alle, wieder, meine, Zeit, gegen, vom, ganz, einzelnen, wo, muss, ohne, eines, können, sei

Zipfsches Gesetz, Power Laws, 20 – 80 Regel

- 20% der Worte bilden 80% eines Texts
  - 4% = 20% von 20% bilden 64% …
  - 0.8% bilden 51,2% …



Gilt ähnlich auch für

- Verteilung von Vermögen
- Größe von Städten
- Einkommensverteilung
- Gesundheitskosten

Durchschnittswerte sind stark irreführend bei Zipfscher Verteilung

- Durchschnittsvermögen eines Deutschen = 88.000 Euro
- 10% verfügen über 61 Prozent
- 5% verfügen über 46%
- 1% verfügen über 23%
- 27% haben kein Vermögen

Zahlen von 2007