

Internet Grundlagen

SEMANTIC WEB, SUCHMASCHINEN

Semantic Web

- Codierung von Bedeutung im Web
 - In Rechnerverständlicher Form
- Ermöglicht Automatische Auswertung von Bedeutungen
- Vorteile:
 - Daten können in Beziehung zueinander gesetzt werden
 - Neue Erkenntnisse können gewonnen werden



- I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A ‘Semantic Web’, which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The ‘intelligent agents’ people have touted for ages will finally materialize.
- – Tim Berners-Lee, 1999



- „The Semantic Web is not a separate Web, but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.”

Tim Berners-Lee, James Hendler, Ora Lassila

→ From a Web of Documents to a Web of Data



Das Web – Probleme

oder warum Semantic Web?

- Inhalte des Webs auf Menschen zugeschnitten
 - Layout, Struktur → einfache Nutzung
- Problem: Finden von (gesuchten) Informationen
- Suchmaschinen können Fundstellen suchen
 - Stichwortbasiert, ohne Kontext
 - Bsp.: Suche nach Apple bringt Frucht und Rechner
 - Wahl des richtigen Stichwortes ist entscheidend
 - Ergebnisse sind immer einzelne Webseiten
 - müssen von Menschen interpretiert und kombiniert werden
 - oft ist gesuchte Info auf mehrere Webseiten verteilt
→ Informationsintegration
 - Relevanz kann nur schwer durch Maschine geprüft werden



Web – Probleme

oder warum Semantic Web?

- Web ist heterogen:
 - unterschiedliche Informationsdarstellung: Bilder, Text, Audio
 - unterschiedliche Codierung: ASCII, Unicode, ISO...
 - unterschiedliche Sprachen
- → Informationen zu einem Thema sind nur schwer aufzufinden



Beispiele

- Gegenüberstellung von Informationen zum Wahlprogramm einzelner Parteien
- Verknüpfung von verteilt im Netz liegenden Informationen:
 - Vortrag, Termin in Hawthorne/NY
 - Reisebuchung von Berlin nach Hawthorne:
 - Berlin liegt in Deutschland/Europa
 - Hawthorne liegt in den USA/Amerika
 - → Flugbuchung notwendig, Anschlusszug notwendig oder Mietwagen



Beispiel: Abendliche Planung

- Essen gehen, Kino gehen, Cocktailbar
- Problem:
 - Finden eines guten Restaurants (je nach persönlicher Vorliebe, Preisklasse, Einschätzung durch andere)
 - Ermittlung des Kinoprogramms in unterschiedlichen Kinos mit Genrevorgabe, Reservierung von Karten
 - Cocktailbar sollte in der Nähe des Kinos liegen
- Mögliche Anfrage: Finde ein *Restaurant mit italienischer Küche in mittlerer Preislage*, und zeige mir die Kritiken zu den neuen Filmen der letzten zwei Wochen



Mögliche Anwendungsbereiche

- allgemein: wissensintensive Prozesse
- Beispiel:
 - kontextbezogene Informationsvernetzung
 - intelligentes Information Retrieval
 - personalisierte Wissensportale
 - Helpdesk-Systeme
- → Anwendungssoftware muss „logisch denken“
- → neues Wissen aus vorhandenem erschließen



Semantic Web

- Beschreibung von Daten und deren Semantik in rechnerverständlicher/-verarbeitbarer Form
- Daten brauchen Informationen darüber, wie sie zu strukturieren und zu interpretieren sind
- → Wissensrepräsentation im Web

Semantic Web Prinzipien (Auswahl)

1. Alles kann durch eine URI identifiziert werden

<http://www.magdeburg.de>

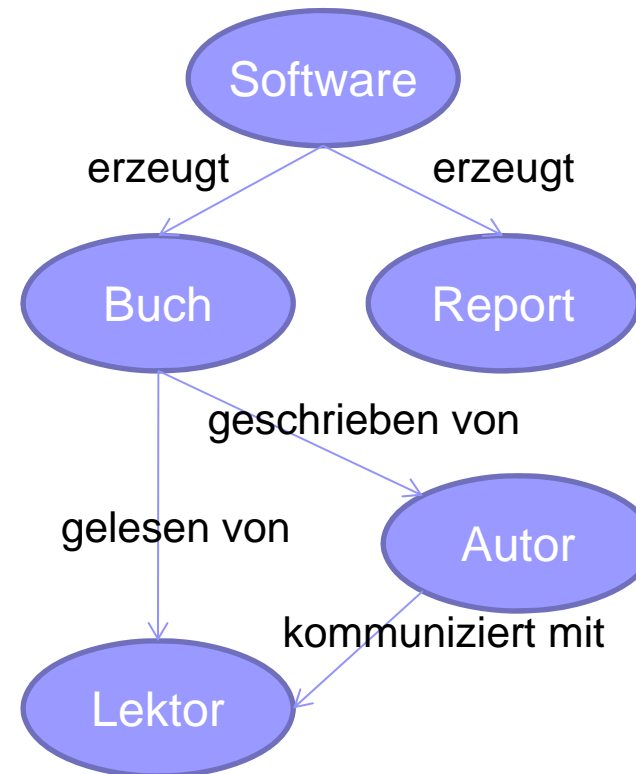
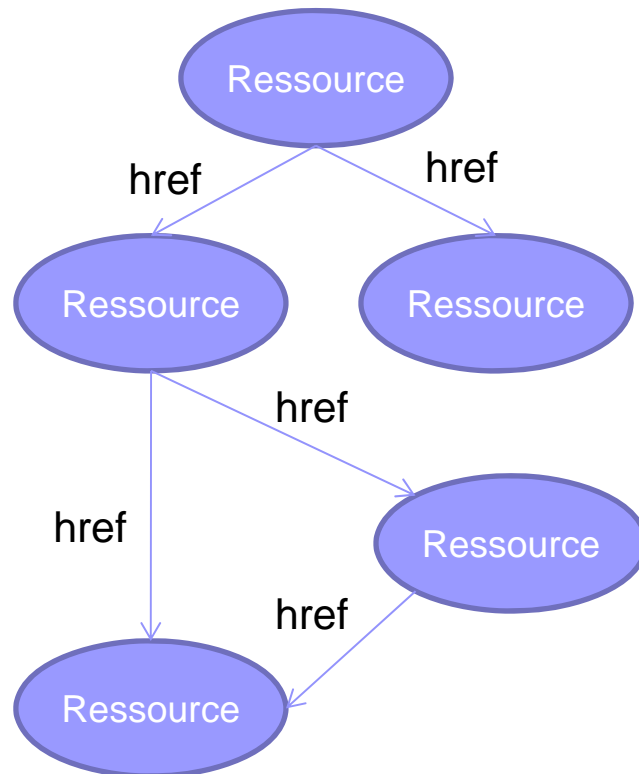


<mailto:marcel.goetze@ovgu.de>



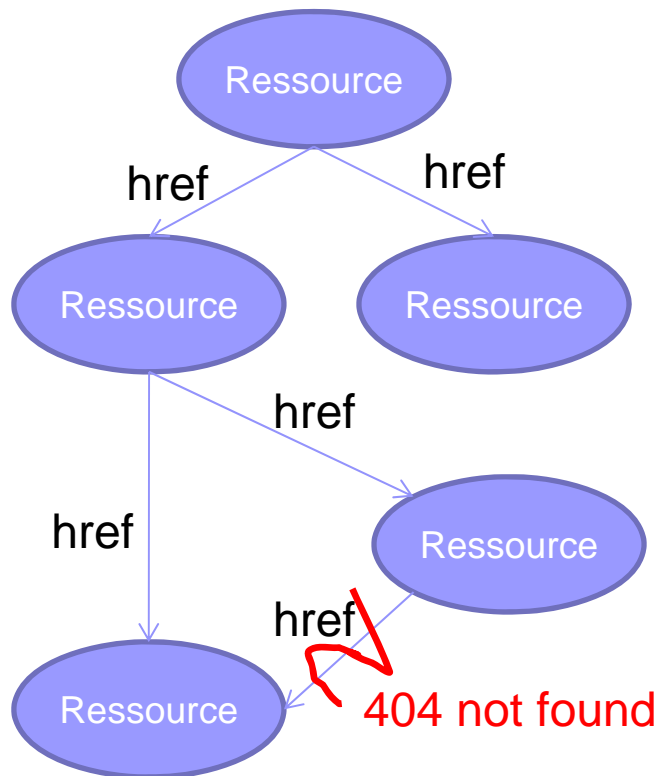
Semantic Web Prinzipien (Auswahl)

2. Ressourcen und Links können typisiert sein



Semantic Web Prinzipien (Auswahl)

3. Unvollständige Informationen sind unproblematisch



- im Semantic Web können fehlende Informationen rekonstruiert werden



Semantic Web Prinzipien (Auswahl)

4. Eine absolute Wahrheit ist nicht notwendig

- neues Wissen kann aus vorhandenen Informationen geschlussfolgert werden
 - Marcel arbeitet in der Otto-von-Guericke-Universität
 - Stefan ist Kollege von Marcel
- Stefan arbeitet ebenfalls an der OvG-Uni



Semantic Web Prinzipien (Auswahl)

5. Evolution wird unterstützt

- Informationen können auf einfache Weise in einen neuen Kontext gesetzt werden
- Beispiel: Stefan bekommt einen Ruf an eine andere Universität → neue Verknüpfung

Web 3.0?

- Weiterführende semantische Vernetzung?
- Internet 2.0: jedes Gerät hat eine eigene IP-Adresse
 - IPv6: 128 Bit lange Adresse = 2^{128} Adressen
 - reicht, für $6,65 \cdot 10^{23}$ Adressen/m² Erdoberfläche
 - Jeder kann jederzeit online sein
- Vielleicht aber auch nicht nur jedes Gerät sondern jeder Blog, jeder Artikel, Autor,...
- Verarbeiten von Informationen erfolgt dezentral, getrennt vom Medium, unabhängig vom Autor
 - Man liest nicht mehr ein Blog sondern von einem Autor in mehreren Blogs
 - Orte wären irrelevant. Bild wird durch IP bestimmt, nicht Position
 - Neue Zugangswege wären notwendig
 - Artikel werden nicht mehr in einem Blog veröffentlicht sondern „zur Verfügung gestellt“
 - Informationen werden anhand ihrer IP identifiziert und entsprechend zusammengestellt

Zusammenfassung

- WWW: Dienst im Internet
- Weltweites Gewebe durch Verlinkung
- Protokoll: Hypertext Transfer Protokoll
- Web 2.0:
 - Webseite als Plattform für Anwendungen
 - Daten wichtiger als Aussehen
 - Nutzung verteilter Daten und Anwendungsteile
 - Ausnutzen der Kompetenz vieler Web-Nutzer

Vom URL-Raten zur Suchmaschine

- Web 1.0: URL-Raten, Web 2.0: Suchmaschinen
- Letztere analysieren Webseiten
 - Robots durchforsten das Web, in der Regel durch das Folgen von Links
 - Analyse des Anfangs oder des gesamten Textes (Altavista, Fireball)
- Auswertung von Metatags
- Auswertung von Verlinkung: PageRank (Google)

Dienste und deren Nutzung

Suchmaschinen

- Arten von Suchmaschinen:
 - Manuell erstellte Kataloge
 - Automatisch erstellte Indizes
 - Suchmaschinen für spezielle Zwecke
- Datenbank von Suchmaschinen:
<http://www.suchlexikon.de/>

Suchmaschinen, der manuell erstellte Katalog

- Generell: von Menschen gemacht
 - Von einer zusammen arbeitenden Gruppe → Redaktion
 - Von vielen Beteiligten unabhängig voneinander → Folksonomy
- Meist hierarchische Präsentation des Katalogs
 - Beispiel: Yahoo, Web.de
- Eignung: Suche nach einem Thema, Sachgebiet, Stichwort
- Vorteil: Redaktion kann Inhalt eines Dokuments berücksichtigen
- Nachteil: Zusammenhänge können verloren gehen:
 - Beispiel: Name einer Person + Sachverhalt

Suchmaschinen, der automatisch erstellte Katalog

- Software (Robot, Crawler, Spider) browsen vollautomatisch durchs Netz → Von Link zu Link
- Indexierungssoftware analysiert und strukturiert Daten
- Suchmaschinen arbeiten auf Begriffen, ohne die Relevanz eines Wortes für den Inhalt des Dokumentes zu berücksichtigen
→ für die Suchanfrage wichtig zu wissen
- Eingrenzen des Suchraumes: Welche Begriffe könnten im Zusammenhang mit dem Suchwort stehen
 - explizites Ausschließen von Begriffen

Suchmaschinen, wichtigste Befehle

- Plus (+): Verknüpfung zweier Wörter, beide müssen im Ergebnisdokument vorkommen
 - +Fahrrad
- Minus (-): schließt ein Wort aus, das nachfolgende darf nicht im Ergebnisdokument vorkommen
 - -preis
- Anführungsstriche: Verbinden von Worten zu einer Phrase. Wird wie ein Wort behandelt
 - „Der oide Depp“
- Trotzdem: nicht jede Suchmaschine erfasst das gesamte Web

Barrieren für Suchmaschinen

■ Die Internetseite ...


- ... ist nicht verlinkt
- ... ist zu versteckt
- ... ist zu aktuell
- ... darf nicht indexiert werden
- ... hat zuviel Text
- ... hat keinen Text
- ... ist nicht frei zugänglich
- ... hat ein unbekanntes Dateiformat

→ Unsichtbares Netz

Suchmaschinen, Google

- Automatisches Browsen durch Links → Robot
- Relevanzsortierung durch PageRank
 - Larry Page und Sergey Brin
 - Grundprinzip: Je mehr Links auf eine Seite verweisen, desto höher ist das „Gewicht“ der Seite
 - Je höher das „Gewicht“ der verweisenden Seiten, desto höher der Effekt → wichtiger die Seite
 - Ziel: Liste der zu einem Suchbegriff wichtigsten Seiten

Suchmaschinen, Google → Benutzung

- Plus (+), Minus (-), Anführungsstriche
- Oder (|): Verknüpfung zweier Wörter, beide können im Ergebnisdokument vorkommen
 - Strand | Beach
- Berechnungen: $((3*8)/6)^2$ 
- Einheiten umrechnen: foot in cm
- Zug, Kino und Wetterauskunft
- Erweiterte Syntax



Google [Erweiterte Suche](#)

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web [Optionen anzeigen...](#) Ergebnisse 1 - 10 von ...

Wetter für Magdeburg - [Zu iGoogle hinzufügen](#)

4 °C
Wind: S mit 6 km/h
Feuchtigkeit: 88 %

Mo.	Di.	Mi.	Do.
			
9 °C 3 °C	8 °C 5 °C	9 °C 3 °C	10 °C 4 °C

[Wetter Magdeburg, Sachsen-Anhalt, Deutschland Wetter ...](#)
Das **Wetter** in **Magdeburg**, Sachsen-Anhalt, Deutschland. **Wettervorhersage** für die nächsten 3 Tage inkl. Temperatur, Wetterzustand und Regenwahrscheinlichkeit.

Suchmaschinen, Google → Erweiterte Syntax

- cache: Sucht in von Google gespeicherten Seiten
- define: Suche nach Definitionen
- filetype: Suche nach bestimmten Dateiendungen
- inanchor: Suche nur in Links
- intitle: Suche nur im Titel einer Seite
- inurl: Suche nur in der Adresse
- intext: Suche nach Begriffen, die nur im Text vorkommen
- link: Ausgabe aller Seiten, die auf eine bestimmte verlinken
- site: Suche auf eine bestimmte Domain eingrenzen.
- related: Sucht nach ähnlichen Seiten

Spezielle Suchmaschinen, Wolfram Alpha

- Weniger Suchmaschine, eher Antwortmaschine
- Entwickelt vom Mathematica-Erfinder Stephen Wolfram
- Daten wurden von 100 Mitarbeitern manuell aufbereitet
- Fragen können in Suchbegriffen oder direkt gestellt werden
- Sehr gut bei Fakten zu Mathematik, Technik, Naturwissenschaften, Linguistik, Wirtschaft
- Beispiel: How old is Barack Obama?



How old is barack obama?



Input interpretation:

age of Barack Obama (politician) today

Result:

48 years 4 months 3 days

[Show details](#)

Computed by: [Wolfram Mathematica](#)

[Source information »](#)

Download as: [PDF](#) | [Live Mathematica](#)

Give us your feedback:

[send](#)



What is Pi? =

Assuming "Pi" is a mathematical constant | Use as a [movie](#) instead

Input:

Mathematica form

π

Decimal approximation:

More digits

3.1415926535897932384626433832795028841971693993751058209749...

Property:

π is a transcendental number

Continued fraction:

Fraction form | More terms

[3; 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 1, 84, 2, 1, 1, 15, ...]

Alternative representations:

More

$$\pi = 180^\circ$$

$$\pi = -i \log(-1)$$

$$\pi = 2E(0)$$

$E(m)$ is the complete elliptic integral of the second kind [»](#)

i is the imaginary unit [»](#)

$\log(x)$ is the natural logarithm [»](#)

[More information »](#)

Series representations:

More

$$\pi = 4 \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1}$$

$$\pi = -2 + 2 \sum_{k=1}^{\infty} \frac{2^k}{\binom{2k}{k}}$$

Microsoft Bing

- Suchmaschine von Microsoft
- Seit 3. Juni 2009 online
- Entscheidungsmaschine
 - Hilfe bei Kaufentscheidungen, Reservierungen, Reisevorbereitungen, etc.
- Gute Suche nach Bildern und Videos
 - Incl. Interaktiver Vorschaufunktion

Spezielle Suchmaschinen, Bildindex

- Spezielle Suchmaschine für Bilder
- Bildindex der Kunst und Architektur
- 2 Millionen Bilder aus 13 europäischen Ländern
- Nach Künstler, Ort, Porträt und Themen katalogisiert
- Suche in unterschiedlichen Bereichen möglich
 - Jahr, Genre,

<http://www.bildindex.de>

Spezielle Suchmaschinen,

■ Metasuchmaschinen

- Weiterleiten einer Suchanfrage an viele Suchmaschinen
- Oft langsamer
- Erste Metasuchmaschine: MetaCrawler
- Deutsche Metasuchmaschine: MetaGer

■ Weitere Spezialsuchmaschinen:

- Medienarchive, Bildarchive, Mensuchsuchmaschinen
- Nachrichtendienste, Bibliotheken und Buchkataloge
- Beispiel: <http://www.bundesarchiv.de/index.html.de>

Zusammenfassung

- Suchmaschinen: manuell erstellt oder automatisch
- Automatisch: Robots browsen durchs Netz → Indexierung von Webseiten nach Suchbegriffen und Schlüsselwörtern
- Suchanfragen haben spezielle Syntax
 - +, -, „“, |,
- Größter Teil des Netzes ist nicht sichtbar
- Zugang zu Informationen teilweise über spezielle Suchmaschinen