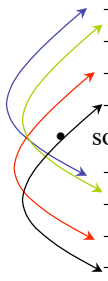
Proseminar WiSe 04/05

Benutzungsintuition und Transparenz im Semantic Web

Cristina Vertan
cri@nats.informatik.uni-hamburg.de

Was lernen wir im Proseminar ?

- Semantic Web ist eine wichtige Technologie, die die Funktionalität des aktuellen Webs ergänzt.
- Im Proseminar werden wichtige Aspekte und Methoden des Semantic Webs erklärt:
  - Daten-Annotationen
  - Metadaten
  - Ontologien
  - Inferenzregeln
- sowie Standard-Modelle und -Repräsentationssprachen.
  - XML
  - RDF,
  - RDFS, OWL, Topic Maps
  - RuleML



Struktur des Proseminars

- Erster Teil : (18.10. - 13.12.)
  - die wichtigsten Methoden und Sprachen des Semantic Webs werden dargestellt.
  - Übungen für ausgewählte Themen werden verteilt
  - Ein Vorgespräch über die vorgesehenen Lösungen wird stattfinden
- Zweiter Teil (03.01. - 31.01.)
  - Präsentation der Lösungen von Studenten
  - Ablieferung der Referate

Scheinkriterien

- Anwesenheit (maximal 2 begründete Abwesenheiten)
- Teilnahme am Vorgespräch
- Präsentation des Referats
- Schriftliche Form des Referats (die schriftliche Form basiert auf den Präsentationsfolien wird aber wie ein Artikel strukturiert: vollständige Sätze, Inhaltsverzeichnis, Überschriften usw.)

Warum ist die Entwicklung des Semantic Webs nötig?

Aktuelle Funktionalität des Webs

- Das WWW wurde für die menschliche Benutzung entworfen.
- Das Webinhalt enthält keine strukturelle Information. (z.B. die Informationen die von Datenbanken extrahiert werden, enthalten nach der Extraktion keine Information über ihre ursprüngliche Zuordnung in der Datenbank)
- Typische Operationen in WWW:
  - Suche
  - Kommunikation
- Die Suchwerkzeuge (AltaVista, Yahoo, Google) sind meistens auf Stichwort-Suche eingestellt.

Probleme der Suchwerkzeuge im aktuellen WWW

- **„High recall, low precision“:** Die nötigen Webseiten werden gefunden , aber zusammen mit anderen 28 758 nicht so relevanten oder total unnötigen Dokumenten.
  - Z.B. suche nach „*Java programming language*“ wird auch alle Dokumenten über die Insel Java , eventuell die dortige Sprache herausfinden
- **„Low or no recall“:** manchmal wichtige Seiten werden nicht gefunden. Das kann beobachtet werden wenn man dieselbe Anfrage an zwei oder drei Suchmaschinen gibt. Oft sind die Ergebnisse unterschiedlich.
- **Große Terminologie- und Sprachabhängigkeit**. Wichtige Informationen, die Synonyme oder Übersetzungen der Anfragewörter enthalten werden nicht gefunden
- **Ergebnisse sind einzelne Webseiten**. Wenn das Ergebnis in mehrere Dokumenten verstreut ist, muss man mehre Anfragen starten und dann die partiellen Informationen aggregieren.

Warum entstehen Probleme währen der WWW-Suche?

- Hauptproblem: die Bedeutung des Webinhaltes ist nicht für automatische Prozesse verfügbar.
- Für eine Maschine ist es z.B schwer zu unterscheiden zwischen:
  - *Das Buch ist eine gute Quelle für .....*

  Und
  - *Das Buch wäre eine gute Quelle für ...wenn nicht...*
- Textverstehen-Methoden (Sprachverarbeitung, KI) können zur Zeit keine Lösungen, die domäne- und sprachunabhängig sind, liefern.
- Lösung: Eine neue Methode für Datenannotation und Inferenzmechanismen im WWW

Definitionen des Semantic Webs

- Semantic = Bedeutung
- Semantic Web = Datenbedeutung die auch maschinell gefunden und bearbeitet sein kann
- Es gibt mehrere Sichtweisen,was „SemanticWeb“ ist:
  - „*Die Sicht von maschinell lesbaren Daten*“: Die Daten sind so dargestellt dass sie von den Computern interpretierbar sind, und das nicht nur für Darstellungszweck. (Berners-Lee)
  - „*Die Sicht von Intelligenten Agenten*“ . Das aktuelle Web soll maschinell lesbar, so dass intelligente Agenten Daten finden und manipulieren können.
  - „*Dies verteilte Datenbanksicht*“. Das Semantic Web wird für die Daten sein was das Web für Menschen ist. D.h. Semantic Web wird einen einhaltlichten Mechanismus für Speicherung und Durchsuchung der Daten bereitstellen. (W3C)

Von Web zu Semantic Web -Anwendungen-

- Es gibt bereits mehrere Anwendungsgebiete die direkt vom Semantic Web profitieren können:
  - Wissensmanagement
  - eBusiness
  - eLearning
  - Personal intelligent agents
  - Sprachverarbeitung

Wissensmanagement -heutiger Zustand

- Wissensmanagement beschäftigt sich mit dem Erwerb, dem Zugang und der Pfelge existierenden Wissens in einer Organisation.
- Wissensmanagement ist sehr wichtig für internationale Organisationen, die geographisch verteilte Bereiche haben.
- Die existierende Information(Wissen) ist zur Zeit sehr schwach (oder überhaupt nicht) strukturiert (z.B.: Text, Audio und Video).
- Folgende Begrenzungen aktuellen Technologien worden festgestellt:
  - Informationssuche: Die Firmen sind abhängig von Stichwort-basierter Suche
  - Informationsextraktion: Die maschinell gefundenen Dokumente müssen von Menschen durchsucht werden um die relevante Information zu extrahieren. (existente Werkzeuge sind qualitativ noch schwach)
  - Informationspflege: Inkonsistenzen in der Terminologie und veraltete Informationen sind schwer zu identifizieren.
  - Informationszugang: Restriktionen zu sämtlichen Daten, an denen man keine proprietären Rechte hat, lassen sich schwer definieren.

Wissensmanagement -Semantic Web Perspektiven

- Das Wissen wird in konzeptuellen Räumen, entsprechend seiner Bedeutung, organisiert.
- Die Datenpflege, Inkonsistenztests und die Extraktion neuen Wissens werden automatisch durchgeführt.
- Stichwortsuche wird durch „Query answering“ ersetzt: das gesuchten Wissen wird gefunden, extrahiert und intuitiv und transparent für den Benutzer dargestellt.
- Mehrdokument-„Query answering“ wird unterstützt.
- Unterschiedliche „Sichten“ derselben Daten wird möglich.

Business-to-Consumer eCommerce (B2C)- heutiger Zustand

- Typische kommerzielle Erfahrung der Web-Benutzer.
- Typisches Szenario: Besuch von ein (mehreren) On-line Shops, Durchsuchung von deren Angeboten, Selektion und Bestellung von Produkten.
- Ideal-Szenario: Man sammelt Informationen über existierende Produkte, deren Merkmalen und Preise, und danach entscheidet der Benutzer wo er kauft.
- Zur Zeit existieren sog. „shopbots“ die mit den Wrappers jedes online-shops interagieren
- Begrenzungen der existierenden shopbots:
  - Basiert auf Stichwortsuche
  - Wrapperprogrammierung ist zeitaufwändig und reagiert sehr sensibel auf Änderungen in der Organisation des online-shops

B2C - Semantic Web Perspektiven

- Entwicklung von Agenten die die Produktinformationen interpretieren können, d.h.:
  - Preise und Produktinformation werden korrekt extrahiert und geliefert, Lieferung und Privacy-Regeln werden interpretiert und mit den Benutzeranforderungen verglichen.
  - Zusätzliche Information über die Zuverlässigkeit des Ladens wird von anderen Quellen gefunden
  - Die Wrappers werden keine bedeutende Rolle mehr spielen

Technologien des Semantic Webs

Wie entsteht das Semantic Web?

- Semantic Web wird „on top“ des aktuellen Webs gebaut.
- Kein revolutionärer wissenschaftlicher Fortschritt ist nötig.
- Partielle Lösungen zu allen Teilen sind bereits vorhanden.
- Heutige Herausforderungen:
  - Integration
  - Standardisierung
  - Werkzeugentwicklung
  - Benutzerakzeptanz.
- Folgende Technologien spielen eine Hauptrolle in der Implementierung des Semantic Webs:
  - Explizite Metadaten
  - Ontologien
  - Logik (Inferenzregeln)
  - Agenten

Explizite Metadaten -1-

- Der Webinhalt ist für Menschen und nicht für die Maschinen formatiert.
- Standard Repräsentationssprache ist HTML:
- Beispiel: Beschreibung eines Beratungsbüros:

```
<h1> Willkomen in unserem Zentrum .. </h1>
...Für allgemeine Informationen rufen Sie bitte unsere
   Sekretärin ...
Für fachliche Fragen rufen sie bitte Dr. ....
<h2> Sprechstunden </h2>
Mo - Do 11 - 17 Uhr
Fr 8 - 11.30 Uhr
Während der Ferienzeit  in <a href="...">Hamburg </a> wird
   unser Zentrum geschlossen.
```

Wie unterscheidet die Maschine zwischen Sekretärin und Dr.?

Wie wird automatisch die Öffnungszeit gerechnet?

Explizite Metadaten -2-

- Ersetzung von HTML durch Metadatensprachen, die den Inhalt repräsentieren können.
- Metadaten = Daten über Daten
- Für das o.g. Beispiel:

```
<zentrum>
   <dienst> Beratung</dienst>
   <zentrumName> ..... </zentrumName>
   <staff>
      <fachlicheBeratung> Dr. .... </fachlicheBeratung>
      <sekretariat> ..... </sekretariat>
   </staff>
</zentrum>
```

- Repräsentationssprache: **XML**:Teil der Datenbedeutung ist mit Metadaten darstellbar.

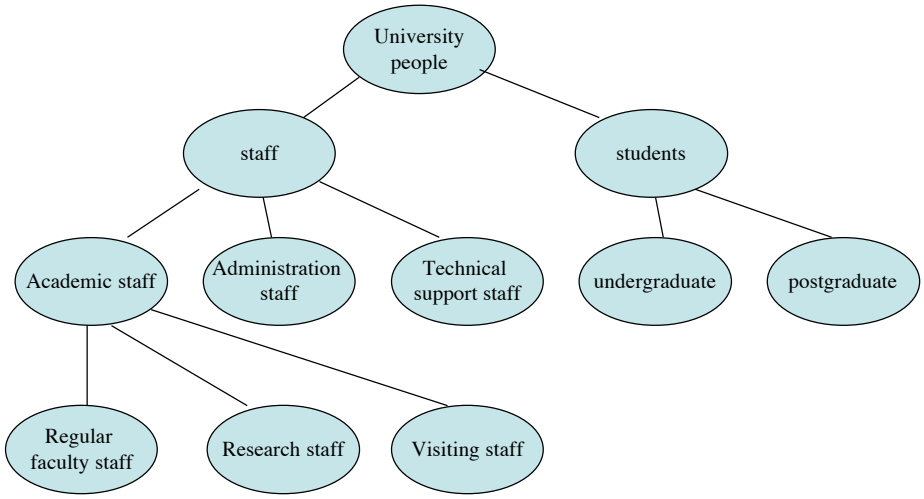
Ontologien -1-

- In der Informatik:„Eine Ontologie ist eine explizite und formale Repräsentation einer Konzeptualisierung.“ (R. Struder).
- Eine Ontologie beschreibt formal ein Diskursdomäne und enthält:
  - Begrenzte Liste von Konzepten (Objektklassen) und
  - Beziehungen zwischen diesen Konzepten
    - Vererbung (isSubclassOf)
    - Merkmale (isRelatedWith, isSimilarWith)
    - Wertbegrenzungen (z.B. Sprechstunden hat nur das wissenschaftliche Personal)
    - „disjoint statements“ (*fachliche Staff* und *administrative Staff* sind disjoint)
    - Logische Beziehungen zwischen Objekten (ein Zentrum muss mindestens 5 fachliche Berater haben)

Ontologien -2-

- Die Metadaten in XML dargestellt werden, werden in Ontologien organisiert. Insofern lassen sich mögliche Überlappungen vermeiden
- Z.B. in einer Institution könnte <staff> nur das fachliche Personal sein, in einer anderen das gesamte Personal.
- Das Problem wird durch eine Ableitung an der entsprechenden Ontologie gelöst.
- Die Organisation von Konzepten in einer Ontologie, sowie die Ableitung von Metadaten auf die Ontologie, ermöglicht auch die Definition von Inferenzregeln.

Ontologie -Beispiel



Logik

- Logik beschäftigt sich mit Denkprinzipien:
  - Formale Sprachen für Wissensrepräsentation
  - Bedeutung von Äußerungen (deklaratives Wissen), d.h. beschrieben wird: **was** entsteht und nicht **wie** entsteht(es).
  - Inferenzregeln die implizite Wissen in explizite Wissen umwandeln.

z.B.:wenn wir wissen :

```
Prof(X)  faculty(X)
faculty(X)  staff(X)
Prof(michael)
```

Können wir schliessen, daß,:

```
faculty(michael)
Staff(michael)
Prof(X)  staff(X)
```

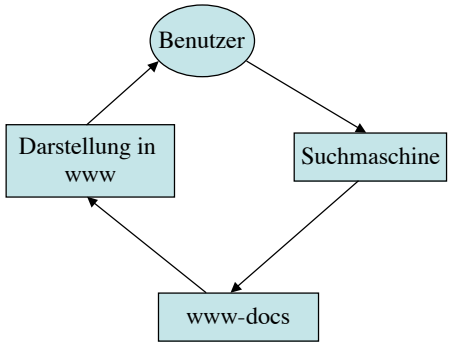
Das Wissen wird normalerweise aus Ontologien extrahiert.

Agenten -1-

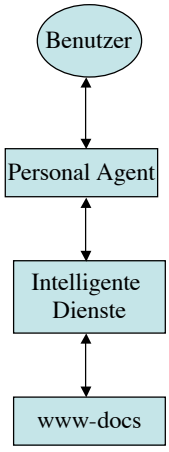
- Agenten sind Softwarekomponenten die autonom arbeiten.
- Z.B.Ein Agent im Semantic Web:
  - bekommt Aufgaben und Preferenzen vom Benutzer,
  - sucht Informationen in Web-Quellen,
  - kommuniziert mit anderen Agenten,
  - vergleicht Informationen über Benutzeranforderungen und -präferenzen,
  - wählt einige Optionen aus und
  - antwortet dem Benutzer.

Agenten -2-

Heute



Zukunft



Agenten und Semantic Web

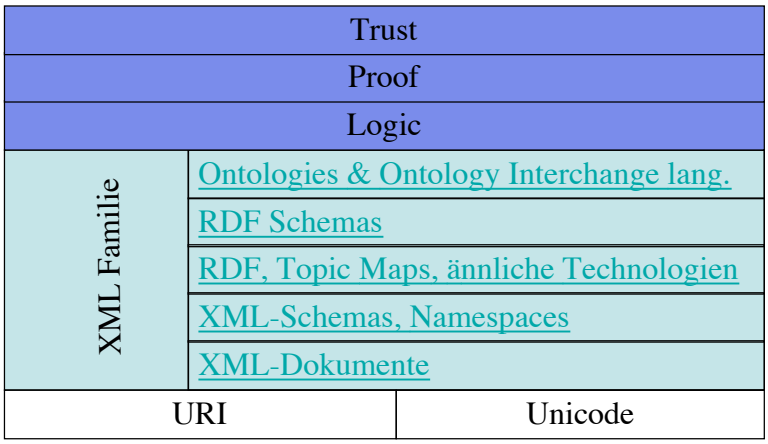
- Metadaten werden für die Identifizierung und die Informationextraktion von Webquellen benutzt.
- Ontologien werden die Websuche unterstützen bei der Interpretation der gefundenen Information und der Kommunikation mit anderen Agenten.
- Logik wird für die Bearbeitung der gefundenen Information und für Schlussfolgerungen benutzt.

Architektur des Semantic Webs

Sprachen für das Semantic Web

- XML - oberflächige Syntax für strukturierte Dokumente. Die Sprache kann aber keine Iinformation über die Bedeutung dieser Dokumente geben
- XML Schema - beschreibt syntaktische Regeln für XML tags.
- RDF - ist ein Datenmodell für Objekte und Beziehungen zwischen Objekten
- RDFS - beschreibt Merkmale und Beziehungen zwischen RDF-Objekte
- OWL ist eine vollständige Sprache für Ontologie-darstellung

„Layer-cake“ Architektur (nach Tim Berners-Lee)



Organisation des Proseminars

Themen (Teil 1)

- 18.10. Einführung „Semantic-Web“
- 25.10. Metadaten-Annotationen mit XML,; XML-Schema
- 01.11. Semantische Beschreibung der Daten mit RDF /RDFS
- 08.11. Ontologien
- 15.11. OWL - Web Ontology Language
- 22.11. Logik und Inferenzregeln
- 29.11. Topic Maps
- 06.12. Semantic Web Services
- 13.12. Agenten und Trust

Referatthemen (Teil 2)

- 03.01. Übungen zum Thema XML
- 10.01. Übungen zum Thema RDF/RDFS
- 17.01. Übungen zum Thema OWL
- 24.01. Übungen zum Thema Inferenz und Logik
- 31.01. Übungen zum Thema Topic Maps

Quellen

- Literaturquellen sowie Online-Ressourcen finden Sie auf der Proseminars-Webseite:

http://nats-www.informatik.uni-hamburg.de/view/SemanticWeb04/WebHome

- Die Referate der Teilnehmer werden an die Web-Seite angehängt. Bitte registrieren Sie sich unter
- http://nats-www.informatik.uni-hamburg.de/view/TWiki/TWikiRegistration

Und geben als Group “SemanticWebGroup” ein.

Kontakt

- Dr. Cristina Vertan
  - F- 211
  - E-mail: cri@nats.informatik.uni-hamburg.de
  - Tel: 42883 2519
  - Sprechstunden: Di. 10-12 und nach Vereinbarung
- Sekretariat Karin Jarck:
  - F-205
  - E-mail: jarck@nats.informatik.uni-hamburg.de
  - Tel: 42883 2433
  - Das Sekretariat ist täglich zwischen 9-13 Uhr besetzt.