Studienprojekt TaxoSearch Spezifikation

Semantisch gestützte Suche im Internet
Lehrstuhl für Computerlinguistik
Ruprecht-Karls-Universität Heidelberg
WS 2002-2003

vorgestellt von

Thorsten Beinhorn, Vesna Cvoro, Khaled Dhaoui und Christian Pretzsch

Was ist TaxoSearch?

- Tool zur automatischen semantischen Erweiterung einer Suchanfrage
- Schnittstelle für Internet-Suche mittels mehrerer Web-Suchmaschinen
- Information-Retrieval System zur Aufbereitung der Suchergebnisse

Ziele des Projekts TaxoSearch

- Erhöhung der Trefferquote durch automatische Erweiterung der Suchanfrage
- Begrenzung der Suchergebnisse auf relevante Dokumente
- Sinnvolles Ranking der Ergebnisse durch Information Retrieval

Namensfindung

- Ursprünglich: OntoSearch
  - Ontologie + Search
  - Aber OntoSearch™ 1.0 von Ontos AG (ontosearch.com)
- Jetzt : TaxoSearch
  - Taxonomie + Search

Was versteht man unter „Ontologie“?

- „Ontologie ist eine explizite begriffliche Formalisierung eines Anwendungsbereiches. Sie dient der Wissensrepräsentation zum Zwecke der zwischenmenschlichen Kommunikation, aber vor allem auch der computergestützten Wissensverarbeitung.“
- Eine Ontologie ist eine explizite, formale Spezifizierung einer gemeinsamen Konzeptualisierung
  - eine explizite Spezifikation von Begriffen (concepts) und deren Beziehungen in einem Bereich (domain)
- meist Taxonomie, Klassifikationsordnung oder eine Katalogerstellung von Objekten und deren Zusammenhang

Taxonomie

- Lehre von der Bildung der verschiedenen Kategorien (Taxa), in die die Naturgegenstände eingeteilt werden.

- Teilgebiet der Linguistik, auf dem man durch Segmentierung und Klassifikation sprachlicher Einheiten den Aufbau eines Sprachsystems beschreiben will.
  - WordNet

WordNet

- WordNet ist eine semantische (allgemeinorientierte) Wortdatenbank, die an der University of Princeton erstellt wurde
  - http://www.cogsci.princeton.edu/~wn/
- WordNet ist eine hierarchische Datenbank, die erfassten Konzepte werden in Baumform konzeptualisiert.
- WordNet ist ein erweitertes Lexikon, dessen Design an aktuelle psycholinguistische Theorien des menschlichen Wortgedächtnisses angelehnt ist.
- Englische Nomen, Verben und Adjektive werden in sogenannten *Synonym Sets* organisiert, von denen jeder ein zugrundeliegendes lexikalisches Konzept repräsentiert. Verschiedene Beziehungen verbinden diese Synonym Sets.

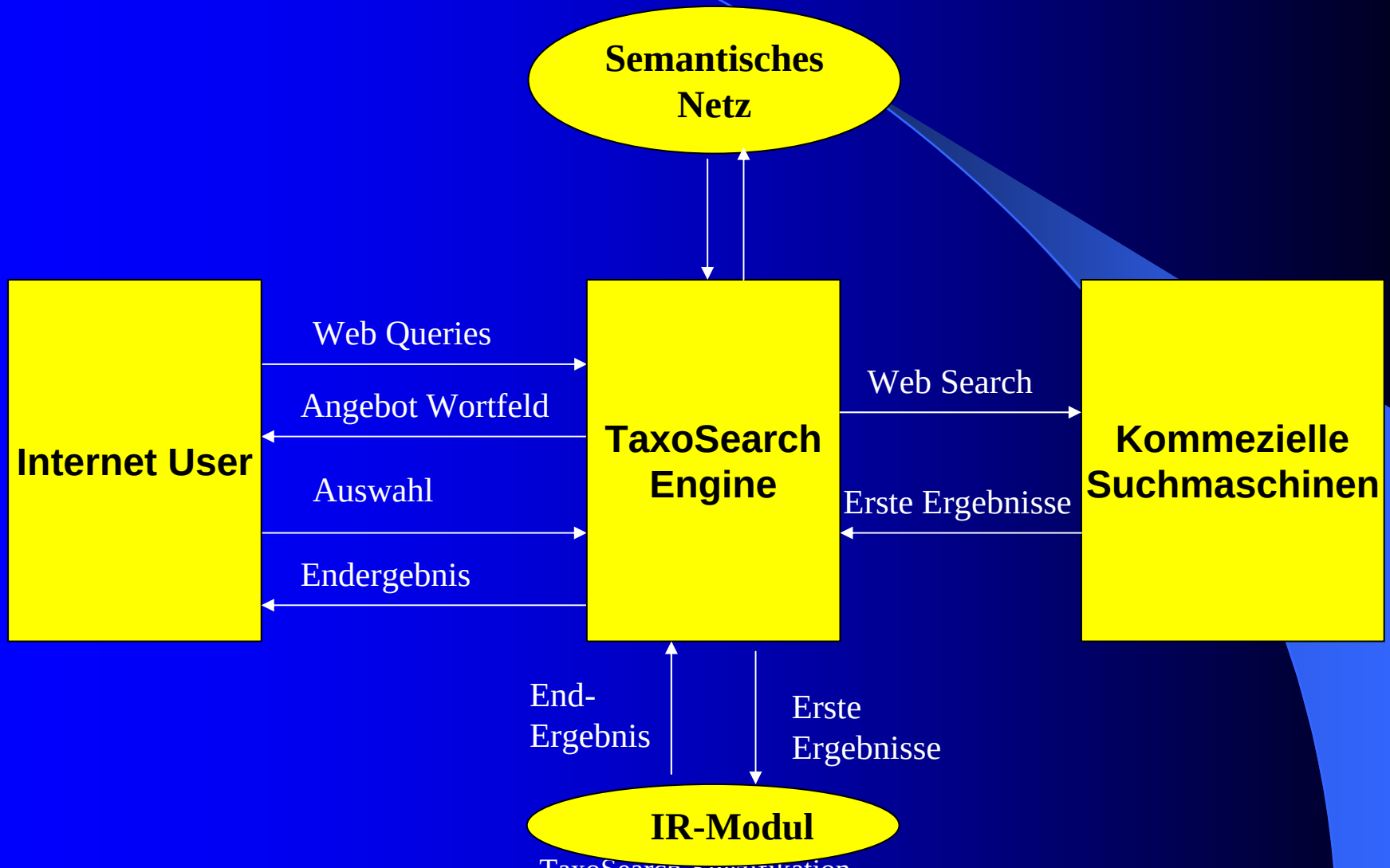
WordNet II

- WordNet speichert seine Informationen nicht nach alphabetischen, sondern nach konzeptuellen Gesichtspunkten.
  - Begriffe gleicher oder ähnlicher Bedeutung werden zusammen abgelegt.
  - die Datenbank von WordNet umfaßt im Gegensatz zu früheren Projekten ungefähr 95600 verschiedene Wortformen, welche in ca. 70100 Bedeutungsklassen (Synonym Sets) eingeordnet sind.
  - WordNet unterteilt das Lexikon in 5 Kategorien : Nomen, Verben, Adjektive, Adverben und Funktionswörter (nicht implementiert). Diese Einteilung beruht auf Untersuchungen über Wortassoziationen von Fillenbaum und Jones (1965)
    - Der Nachteil dieser Vorgehensweise besteht darin, daß mache Wörter in mehr als eine syntaktische Kategorie hineinfallen können und somit mehrmals gespeichert werden müssen.

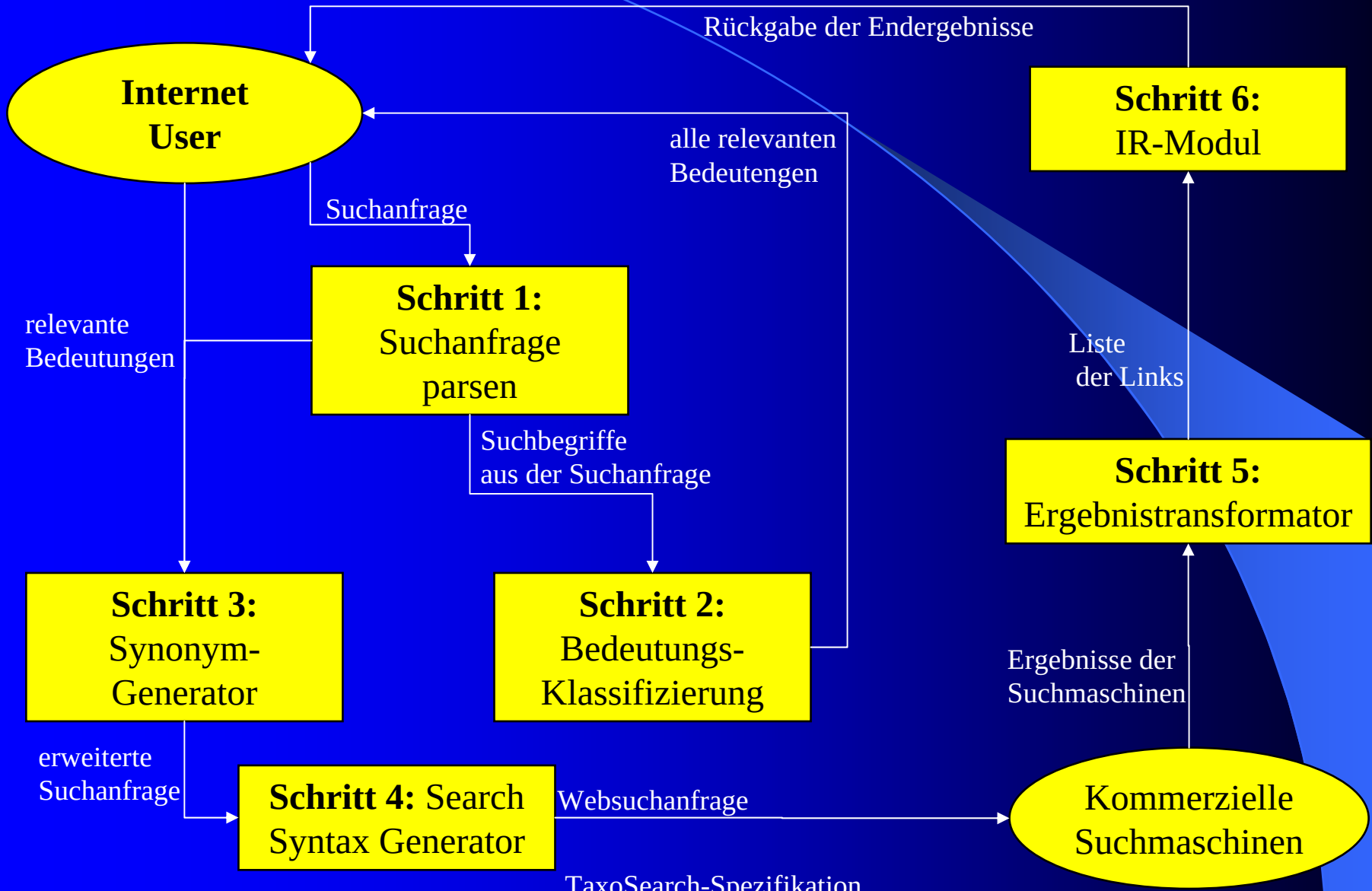
Wordnet III

- **Synonyme :** Dies ist die wichtigste Beziehung in WordNet. Dabei wird eine erweiterte Definition verwendet, nach der zwei Ausdrücke synonym sind, wenn der Austausch des einen Wortes durch das andere den Wahrheitswert einer Aussage nicht verändert. Ein Beispiel für Synonyme sind z.B. Fantasie und Vorstellungskraft.
  Antonyme : Antonyme sind Wörter gegensätzlicher Bedeutung. Ein Beispiel dafür ist aufgehen - untergehen.
  Hyponyme : Ein Hyponym ist eine Verfeinerung, besitzt also eine speziellere Bedeutung als das Wort, auf das es sich bezieht. Beispiel : Haus ist Hyponym von Gebäude.
  Hypernym : Hypernym ist das Gegenteil von Hyponym, also eine Verallgemeinerung des Ursprungswortes. Beispiel : Tier ist Hypernym von Katze.
  Meronym : Meronym ist ein Teil einer Gesamtheit. Beispiel : Rad ist Meronym von Auto.
  Holonym : Holonym ist das Gegenteil von Meronym. Beispiel : Baum ist Holonym von Ast.
  Morphologische Relationen : Morphologische Relationen beschreiben die syntaktische Abhängigkeit eines Wortes vom Kontext. Beispiel: Pluralformen.

Schematischer Aufbau



Funktionsweise TaxoSearch



Schritt 1: Suchanfrage

- Eingabe der Suchanfrage über GUI

Schritt 2: Bedeutungsklassifizierung

- Bei Wörtern mit mehr als einer Bedeutung werden diese dem Benutzer zur Auswahl zurückgegeben
- Benutzer wählt gewünschtes Bedeutungsfeld aus

Schritt 3: Synonym Generator

- Benutzer hat Bedeutungsfeld des / der Suchbegriffe festgelegt
- Synonym Generator erweitert die Suchanfrage durch Synonyme

Schritt 4: Search Syntax Generator

- Aufbereitung der Suchanfrage-URL für die unterschiedlichen Suchmaschinen
- Verfügbare Suchmaschinen zunächst nur google.com und altavista.com

Schritt 5: Ergebnistransformator

- Auslesen der Ergebnisseiten aller Suchmaschinen
- Ausfiltern doppelter Suchtreffer
- Aufbereiten der Ergebnisse für IR-Modul

Schritt 6: IR-Modul

- Analyse der gefundenen Dokumente
- Ranking der Dokumente nach Relevanz
- Methode: Vektormodell

(Schritt 7: Ergebnisse anzeigen)

- Ausgabe der Ergebnisliste
- Eventuell auf dynamisch generierter HTML-Seite

Erwartete Probleme

- Hohes Datentransfervolumen
- Noch keine empirischen Untersuchungen, ob dieses Verfahren die Trefferquote wirklich steigert
- Begrenzt durch Datenbestand in WordNet