Bedeutungsorientierte Suche mit Sprachverstehen statt Stichwortsuche

Eine Kurzpräsentation in zwölf Folien, 2013-11-29

Problem der Kunden und Lösung von SEMPRIA

1. Problem

Archivbesitzer (Verlage, Rundfunkhäuser, Web-Sites, Organisationen . . .): haben **Probleme**, ihre Archive voll und effizient zu **erschließen**

2. Lösung

Suchmaschine SEMPRIA-Suchmaschine hilft durch

- **höhere Vollständigkeit** von Suchergebnissen
- **höhere Genauigkeit** von Suchergebnissen
- **Fragefunktionalität** (Möglichkeit gezielter Fragen)

3. Wie?

eigene Technologie (60 Mannjahre) für deutsche (englische, chinesische) Texte: **automatisches Sprachverstehen durch tiefe semantische Analyse**

4. Alleinstellungsmerkmal der Lösung

durch Einsatz von Sprachverstehen deutliche Leistungssteigerungen

→ Nutzer (Leser, Redakteure, Analysten, . . .) **sparen Recherchezeit**

→ Nutzer **finden öfter das eine relevante Dokument**

SEMPRIA Entstehung

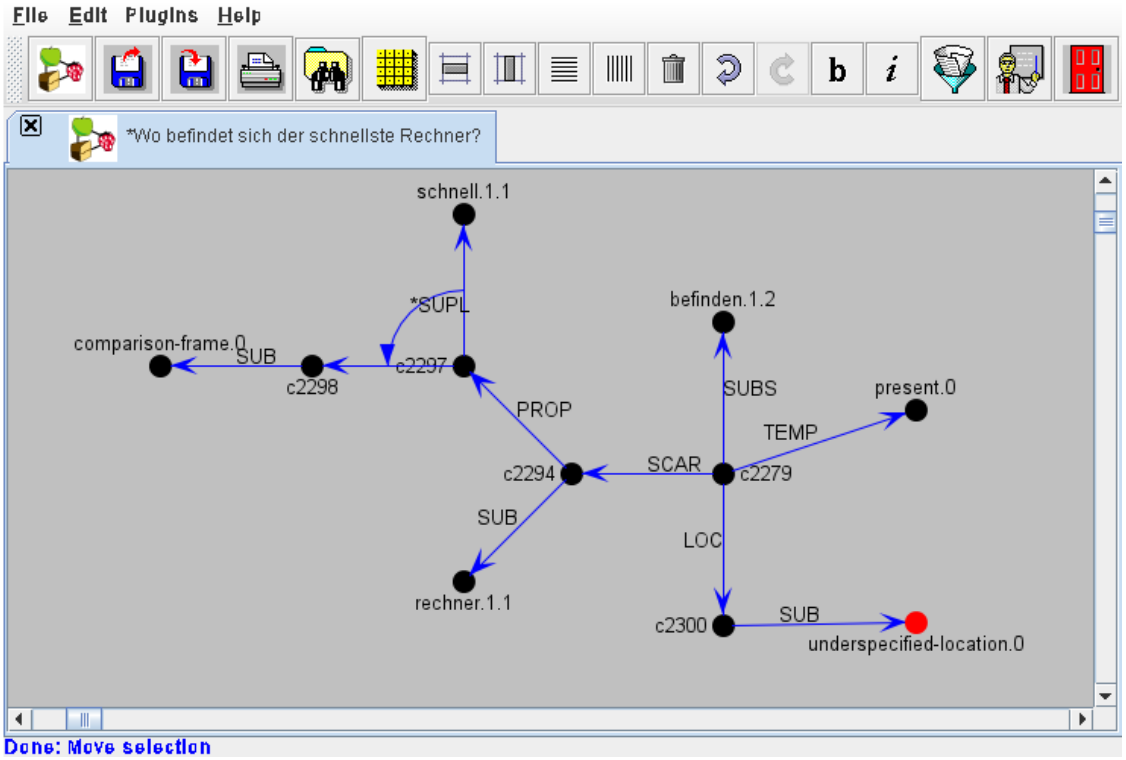
1992 bis heute:	Forschung der AG Intelligente Informations- und Kommunikationssysteme von Prof. Helbig (FernUniversität in Hagen) auf dem Gebiet der wissensbasierten Systeme http://pi7.fernuni-hagen.de/forschung/
2003 bis heute:	erfolgreiche Teilnahmen an internationalen Wettbewerben im Bereich Suche und Information Retrieval (CLEF)
2009:	Prof. Helbig und drei langjährige Mitarbeiter gründen die SEMPRIA GmbH

SEMPRIA Forschung und Unternehmen

- SEMPRIA GmbH mit Sitz in Düsseldorf
- Kooperation mit der AG Intelligente Informations- und Kommunikationssysteme (FernUniversität in Hagen)
- Aktuelle Arbeitsschwerpunkte der SEMPRIA GmbH:
  Anpassung der bedeutungsorientierten Suchmaschine SEMPRIA-Search an individuelle Kundenbedürfnisse, Vertrieb

SEMPRIA Kompetenz: Bedeutungsanalyse

Bedeutungsanalyse von Texten:
Natürliche Sprache $\rightsquigarrow$ Bedeutungsdarstellung durch semantische Netze



SEMPRIA Kompetenz: Wissensbasen

- **Aufbau lexikalisch-semantischer Ressourcen**
  - Synonyme, Unter-/Oberbegriffe, Nominalisierungen, Schreibvarianten: semi-automatisch aus Textarchiven
  - 150.000 Beziehungen in Begriffs-Ontologien
- **Aufbau logischer Regelsysteme**
  - Erfassung komplexer Beziehungen zwischen Wortbedeutungen
  - Interpretation von Funktionsverbgefügen und bildhaften Ausdrücken
  - mehrere tausend Axiome
- **Automatische Erzeugung von Wissensbasen aus Texten**
  - z.B. Wikipedia (60 Millionen Sätze)
- **Entwicklung von Werkzeugen zur Wissensakquisition**
  - Unterstützung des Wissensingenieurs beim semi-automatischen Wissenserwerb

→ **Erfahrung im Aufbau und Nutzung großer Wissensbestände**

SEMPRIA Kompetenz: Computerlexika

Kennzahlen

- 35.000 semantisch umfassende Einträge (alle linguistischen Beschreibungsebenen inkl. semantische Rollen und Komplemente)
- 50.000 unterstützende Einträge (nur Morphologie und Syntax)
- 800.000 Eigennamen in circa 50 Klassen
- 1.500.000 Komposita mit Analysen, z.B. *((Erbschafts (steuer)) reform)*

→ **SEMPRIA-Search verwendet eines der größten semantischen Computerlexika**

SEMPRIA-Search Beispiel

Nutzeranfrage:

Import von Öl

Zusätzliche richtige Treffer gegenüber Stichwortsuche (→ Vollständigkeit):

importierten Öls

Öl importen aus . . .

Import von Erdgas und . . . Erdöl

führte 2011 . . . Öl ein

Vermiedene falsche Treffer der Stichwortsuche
(→ Genauigkeit durch semantische Filter):

um den Import mit Öl zu bezahlen

SEMPRIA-Search Einsatzbereiche

- **Suchfunktion (im Web oder in-house) für Textarchive von:**
  - Zeitungen, Zeitschriften
  - Radio, Fernsehen
  - Enzyklopädien, Akten, Dokumentation . . .
- **Alternatives Suchmodul für Content-Management-Systeme (CMS)**
  Typo3, WordPress, . . .
- **Unternehmensweite Suche** (*enterprise search*)
- **Archivierungsunterstützung**
  - Verschlagwortung, Verstichwortung, Themenseiten
  - Verlinkung, ähnliche Artikel
  - Duplikatserkennung . . .
- **Sentimentanalyse** (*opinion mining*, *issue management*, Öffentlichkeitsarbeit)
  in Vorbereitung

SEMPRIA-Search Merkmale 1

Traditionelle Suchmaschinen **beherrschen nur**:
Flexion, Derivation, Komposita (teils unvollständig)
Nutzung von Metadaten
facettierte Suche für einige Namensklassen

SEMPRIA-Search **beherrscht zusätzlich**:

Mehrwortausdrücke *Rio de Janeiro* stets als Einheit

Zahlen, Maße *10 km* ⟷ *10000 Meter* ⟷ *10.000 m* . . .

Idiome, Metonymie * *Handtuch werfen* ⟷ *aufgeben*

Funktionsverbgefüge * *Antrag stellen* ⟷ *beantragen*

Mehrdeutigkeiten (lexikalische, strukturelle) * *Lincoln*: Mensch, Auto, Stadt . . .

absolute Zeitangaben *31. Juli 13* ⟷ *31.07.2013*

relative Zeitangaben *vorgestern* (*am Mittwoch*) + Metadaten ⟷ *31.07.2013*

facettierte Suche mittels voller Semantik

* abhängig vom verfügbaren Hintergrundwissen

SEMPRIA-Search Merkmale 2

Bezüge von Pronomen * *Die X*[1] *kritisierte den Y. Sie*[1] *antwortete . . .*

Beziehungen zwischen Objekten *Was hat Firma X mit Kriegswaffen zu tun?*

Rolle von Beteiligten in Ereignissen
Kauf der Dresdner Bank vs. *Kauf durch Dresdner Bank*

Bedeutung aus mehreren Sätzen/Dokumenten *Wer exportiert seltene Erden?*
Dokument 1: *seltene Erde Neodym* + Dokument 2: *X führt Neodym aus*

semantische Suchvorschläge aus den Dokumenten

Fehlerrobustheit (Rechtschreibung, Zusammenschreibung)
Fussball WM ⟷ *Fußball-WM*

* abhängig vom verfügbaren Hintergrundwissen

Produkt: SEMPRIA-Search ASP

Gehostete Suchlösung:

- flexibel
- keine Hardwarekosten, kein Installations-Aufwand, kein Update-Aufwand
- hochverfügbare Lösung in deutschem Rechenzentrum
- preiswerte Monatsgebühren (Einstiegsversion: SEMPRIA-Search Webstarter)
- automatische Software-Updates nach Vereinbarung

Randbedingungen:

- Vereinbarung der Formate für die zu durchsuchenden Archive
- Festlegung der Internet-Schnittstelle zur Einrichtung und Aktualisierung des Suchindexes

Produkt: SEMPRIA-Search Server

Eine individuelle Installation beim Archiv-Anbieter.

Randbedingungen:

- Vereinbarung der Formate für die zu durchsuchenden Archive (wie bei SEMPRIA-Search ASP)
- Betriebssystem: Linux oder verwandtes Betriebssystem
- Software: keine weiteren Anforderungen
- Plattenplatz: Größe des Archivs (als reiner Text) mal Faktor 30 bis 60
- Server: 4–8 CPU-Kerne, 3 GHz (empfohlen), 4–8 GB RAM

Impressum

SEMPRIA GmbH
Grafenberger Allee 277–287
40237 Düsseldorf

Telefon: 0211/566693-57
Fax: 0211/566693-58
Web: http://www.sempria.de/
E-Mail: info@sempria.de

Geschäftsführer: Dr. Sven Hartrumpf
Handelsregister: Amtsgericht Düsseldorf, HRB 62168
UStID-Nr: DE268248179