Bedeutungsorientierte (kognitive) Suche mit Sprachverstehen statt Stichwortsuche

Eine Kurzpräsentation, 2020-08-01

SEMPRIA GmbH, Grafenberger Allee 277–287, 40237 Düsseldorf, https://www.sempria.de/

Gliederung

* Motivation und wissenschaftlicher Hintergrund
* Problem von Such-Anwendern und Lösung von SEMPRIA
* Technologie und Kompetenzen bei SEMPRIA
* SEMPRIA-Suchmaschine: Systemmerkmale, Beispiele, Vorteile
* Fazit

Teil 1

Motivation und wissenschaftlicher Hintergrund

SEMPRIA Entstehung und Entwicklung

1992 bis 2015: Lehrgebiet Intelligente Informations- und Kommunikationssysteme von Prof. Helbig (FernUni Hagen); Forschung auf dem Gebiet der wissensbasierten Systeme http://pi7.fernuni-hagen.de/forschung/

2003 bis 2010: erfolgreiche Teilnahmen an internationalen Wettbewerben im Bereich Suche und Information Retrieval (CLEF)

2009: Prof. Helbig und langjährige Mitarbeiter gründen die SEMPRIA GmbH in Düsseldorf

Motivation und Einstieg

Mängel bisheriger Suchmaschinen und Sprachtechnologie

- **Mehrdeutigkeit weitgehend ignoriert**
  Münster, *Kohl*, *Maus*
  → **geringe Korrektheit** der Suchergebnisse
- **semantische Beziehungen zwischen Wörtern nicht berücksichtigt**
  Waffenimport vs. *führt . . . Haubitzen . . . ein*
  Südfruchteinfuhr vs. *importiert Datteln*
  → **geringe Vollständigkeit** der Suchergebnisse

SEMPRIA-Lösung

- Semantik-orientierter Ansatz – Bedeutung von Texten im System bestimmt und repräsentiert
- Repräsentation ist homogen und interoperabel: Wörter, Phrasen, Sätze, Texte
- Repräsentation ist universell: über Anwendungen und Sprachen hinweg (Deutsch, Englisch, Mandarin)

Teil 2

Problem von Suchenden und Lösung von SEMPRIA

Problem von Suchenden und Lösung von SEMPRIA

1. Problem

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

2. Lösung

SEMPRIA-Suchmaschine hilft durch

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

3. Wie?

eigene Technologie (70 Mannjahre) für deutsche (englische, chinesische) Texte:
automatisches Sprachverstehen durch tiefe semantische Analyse

4. Alleinstellungsmerkmal der Lösung

durch Einsatz von Sprachverstehen deutliche Leistungssteigerungen

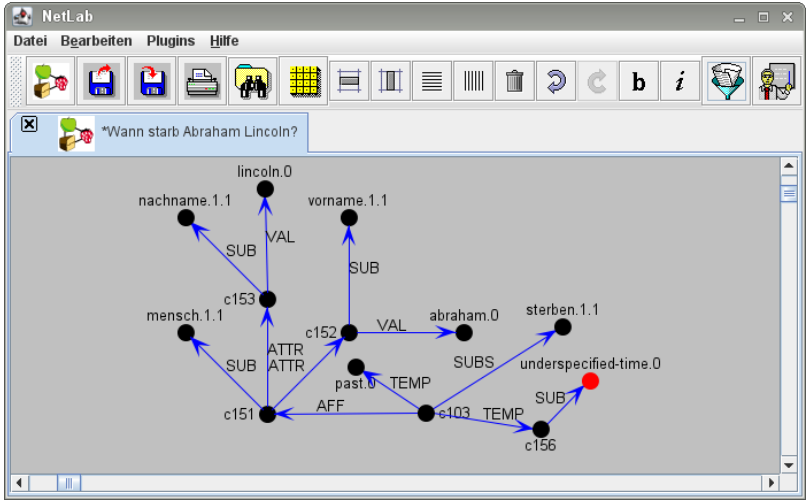
→ Nutzer (Leser, Redakteure, Analysten . . .) **sparen Recherchezeit**

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

Technologie und Kompetenzen bei SEMPRIA

Bedeutungsanalyse von Texten:

Natürliche Sprache $\rightsquigarrow$ Bedeutungsdarstellung mit semantischen Netzen



SEMPRIA Kompetenz: Wissensbasen

- **Aufbau lexikalisch-semantischer Ressourcen**
  - Synonyme: *ausführen* und *exportieren*
  - Unterbegriffe und Oberbegriffe: *Einkommensteuer* und *Steuer*
  - Nominalisierungen: *Ausfuhr*
  - Schreibvarianten: *geographisch* und *geografisch*
  - Fremdsprachliches, Fachsprachliches: *soziale Medien* und *Social Media*
  - Geographisches Wissen: *Flingern* ist Teil von *Düsseldorf*
  - zusammen: über 200.000 Beziehungen
- **Aufbau logischer Regelsysteme**
  - Erfassung komplexer Beziehungen zwischen Wortbedeutungen: *ausführen* und *Exporteur*
- **Automatische Erzeugung von Wissensbasen aus Texten**
  - z.B. deutsche Wikipedia (125 Millionen Sätze, Stand 2020-08-01)
- **Entwicklung von Werkzeugen zur Wissensakquisition**
  - Unterstützung des Wissensingenieurs beim semi-automatischen Wissenserwerb

SEMPRIA-Suchmaschine: Systemmerkmale, Beispiele, Vorteile

SEMPRIA-Search Beispiel

Suchanfrage:

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.

Suchanfrage: ***Wer war der Erfinder der Blindenschrift?***
auch: **Wer erfand die Blindenschrift?**

Antwort:
Louis Braille.
Laut Kahlisch beherrschen etwa zehn Prozent . . . die Punktschrift, die von Louis Braille erfunden wurde.

Phänomene:

- Verbindung der Begriffe *Erfinder* und *erfinden* (*ein Erfinder erfindet etwas*)
- Synonymie zwischen *Blindenschrift* und *Punktschrift*

SEMPRIA-Search Produkfeatures

Traditionelle Suchmaschinen **beherrschen nur** (teils unvollständig):
Flexion (*Haus* ↔ *Hauses*), Derivation (*ändern* ↔ *Änderung*), Komposita (*Motor* ↔ *E-Motor*)

SEMPRIA-Search **beherrscht zusätzlich**:

Mehrwortausdrücke	*Sankt Augustin* (als Einheit)
Zahlen, Maße	*10 km* ↔ *10000 Meter* ↔ *10.000 m*
	90 EUR ↔ *neunzig Euro* ↔ *90,00 €*
relative Zeitangaben	*vorgestern* UND Publikationsdatum=13.03.2020
	↔ *11.03.2020*
Idiome, Metonymie	*Handtuch werfen* ↔ *aufgeben*
Funktionsverbgefüge	*Antrag stellen* ↔ *beantragen*
Mehrdeutigkeiten (lexikalische, strukturelle)	*Gerresheim* Ort
	Person
Rolle von Ereignis-Beteiligten	*Kauf der Dresdner Bank*
	Kauf durch Dresdner Bank

SEMPRIA-Search Produkfeatures 2

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

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

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
Eingabe: *för* ➡ Vorschläge: *öffentliche Fördermittel*, *förderfähige Kosten*, . . .

Integration von FAQs in die Suchvorschläge
Eingabe: *Wa* oder *Büro* ➡ Vorschläge: *Wann ist das Büro geöffnet?*, . . .

Fehlerrobustheit für Suchanfragen (Rechtschreibung, Zusammenschreibung)
Fusball WM ⬌ *Fußball-WM*

* abhängig vom verfügbaren Hintergrundwissen

SEMPRIA-Suchmaschine für FIRMA

Produkte: SEMPRIA-Search SaaS

Suchmaschine als Software-as-a-Servive (SaaS)

- kein Installations-Aufwand, kein Update-Aufwand, keine Hardwarekosten
- hochverfügbare Lösung in deutschem Rechenzentrum
- preiswerte Monatsgebühren (nach Zahl der Dokumente, Dokumenten-Updates und Suchanfragen)
- automatische Software-Updates

Randbedingungen:

- Integration eines kurzen Programmcode-Schnipsels im CMS (für Suche und Suchvorschläge)

Anwendungsszenario für FIRMA

- SEMPRIA-Search SaaS gehostet für FIRMA
- Verwendete Dokumente: von FIRMA
  optional: PDFs, Webseiten weiterer Portale, Veranstaltungsdatenbanken ...
- Anpassung an die Besonderheiten von Inhalten und Umgebung:
  - Vokabular der Dokumente und Anfragen
  - typische Anfragen (aus Such-Logdatei)
- Aktualisierung der Daten:
  - wöchentlich,
  - täglich,
  - zu definierten Uhrzeiten oder
  - in Realzeit (d.h. bei Veröffentlichung eines neuen Dokuments)
- Auswertungen der Suchanfragen: monatlich, wöchentlich oder ...
- Auswertungen zum Sprachgebrauch der Dokumente: regelmäßig (besonders: Tippfehler)

Fazit

- **Such-Ergebnisse: signifikant vollständiger und genauer.**
- **Frage-Antwort-Funktionalität: Präzise Antworten auf gezielte Fragen.**
- **Recherche für den Nutzer: schneller und effektiver.**
- **Nutzererlebnis wird komfortabler: Suchvorschläge, aufbereitete Treffer, FAQ.**

Impressum

SEMPRIA GmbH
Grafenberger Allee 277–287
40237 Düsseldorf

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

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