GESIS Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V.

IZ InformationsZentrum Sozialwissenschaften

Retrieval on Heterogeneous Document Collections
The Project CARMEN

Metadata
Documents Retrieval
Heterogeneity

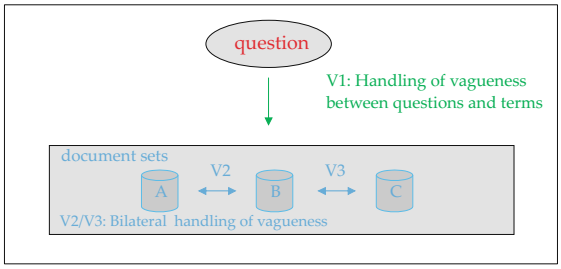
Strategies in CARMEN

- Metadata (Dublin Core in RDF, Meta-Maker, digital signatures)
- Retrieval on structured documents and heterogeneous data types (search engine and gatherer for XML documents)
- Methods for treatment of resisting semantic heterogeneity

Semantic Heterogeneity

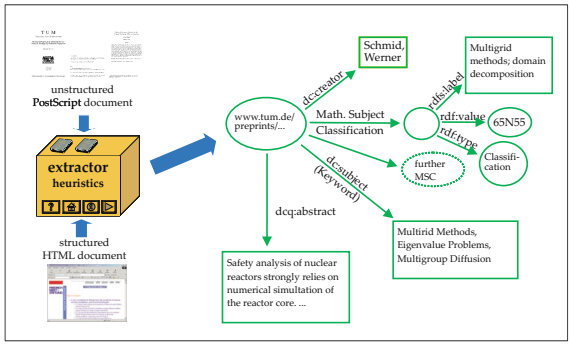
Different data collections using different thesauri or classifications for content description or containing varying or no meta data at all, or intellectually indexed documents meeting even completely un-indexed Internet pages

Query Translation



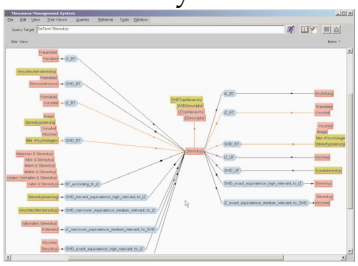
- Handling of vagueness between different document sets
- Query translation using semantic relations
- Document set specific queries

Metadata Extraction



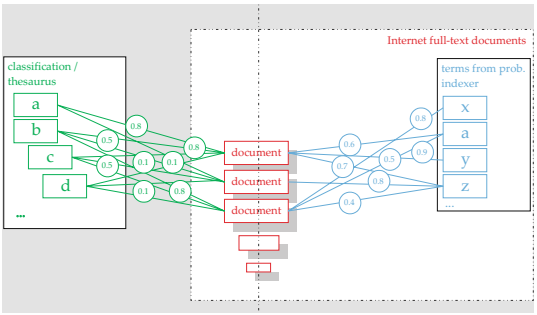
- Enriching poorly indexed documents with meta data during the gathering process
- This meta data is to be available for retrieval
- PostScript documents for the mathematics domain from PostScript documents (thesis papers); extracting meta data using keywords and style information
- Internet documents from the social sciences (html files)

Intellectually created relations



- Cross-concordances between different thesauri and classifications
- Semantic relations created by documentary and professional experts
- Software tool: SIS-TMS by ICS-FORTH

Semantic relations based on quantitative-statistical methods



- Based on the analyses of double corpora
- "Simulated double corpora" with assigned controlled vocabulary and full text terms
- Conditional probability for the co-occurrence of terms
- Creation of semantic relations between classes or terms and Internet full text terms
- Automatic indexing for Internet documents without controlled vocabulary

The project CARMEN ("Content Analysis, Retrieval and MetaData: Effective Networking") handles the semantic differences between heterogeneous data sources of mathematics, physics and social sciences. The treatment of this heterogeneity uses among other ways the extraction of meta data and the statistical correlation of terms in documents from different collections. Query transfer modules will support the user searching documents by translating the query into semantically collection specific queries.

Sponsored by bmb+f Federal Ministry of Education and Research
GMD Projektträger Fachinformation

Robert Strötgen * German Social Science Information Centre (IZ) * Lennéstr. 30 * D-53113 Bonn / Germany * http://www.gesis.org/iz/ * soe@bonn.iz-soz.de