

HyREX – Hypermedia Retrieval Engine for XML

XML retrieval: search with respect to content *and* structure

Problem: XML query languages like XPath, XQL, XQuery do not consider the needs of Information Retrieval

Solution: query language which takes into account intrinsic imprecision and vagueness of IR

XIRQL: XML Information Retrieval Query Language

Weighting and ranking: probabilistic weighting of document content and query conditions → ranked retrieval results

Relevance-oriented search: retrieve relevant *parts* of a document by choosing the most specific element(s) that satisfy a given information need.

Data types: assign data types to XML elements (e. g. person names, dates, names); specific **vague predicates** for searching (e. g. phonetic similarity of names, approximate matching of dates).

Structural relativism: support uncertainty and vagueness for structural query conditions.

Architecture

HyGate

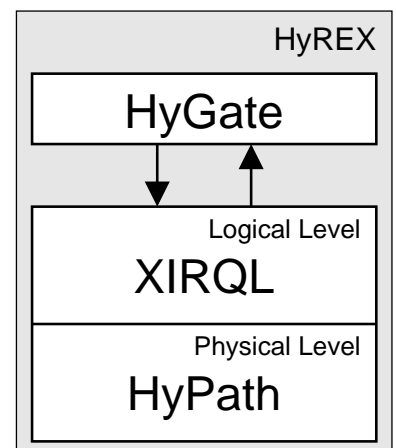
- user interface for searching and browsing
- query formulation assistant: query by example
- presentation of retrieval results: tile bars, tree maps

XIRQL

- transformation of XIRQL syntax into path algebra
- optimization of path algebra expressions
- mapping onto physical operators

HyPath

- efficient access paths for content and structure
- application specific selection of access paths



Implementation

- object-oriented design → extensible IR system architecture
- implemented in Perl (time-critical parts in C)
- open source software

