

Relevance-oriented search with HyREX

Norbert Gövert, Norbert Fuhr,
Mohammad Abolhassani, Kai Großjohann

Universities of Dortmund and Duisburg

Outline

- HyREX
- Weighting and ranking
- Implementation
- (Preliminary) Evaluation

HyREX: Hypermedia Retrieval Engine for XML

XIRQL:

- XML Information Retrieval Query Language
- XQL (XPath) + Information Retrieval

Added value:

- Weighting and ranking
- Relevance-oriented search
- Datatypes and vague predicates
- Semantic relativism

Weighting and Ranking

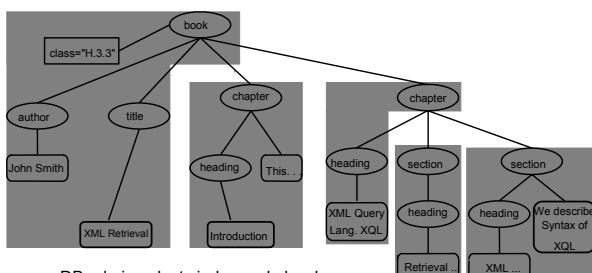
Aim: Retrieve the most specific document component
which answers the query

Either: Invent new weighting mechanisms for XML
retrieval

Or: Generalize well-performing Weighting schemes for
traditional IR applications

- 1 Define "atomic" units: **index nodes**
 - Application of weighting formulas
 - Relevance-oriented search: retrievable answers

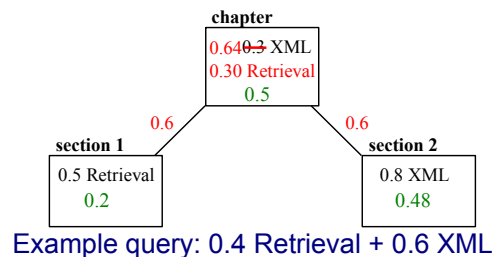
Weighting and Ranking II: Index nodes



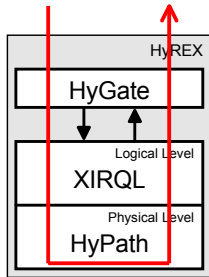
- DB admin selects index node borders
- Apply (BM 25) weighting scheme

Weighting and Ranking III: Retrieval

- Propagate index node weights to their upper level index nodes via *augmentation*
- RSV is the scalar product of query term weights and index node weights

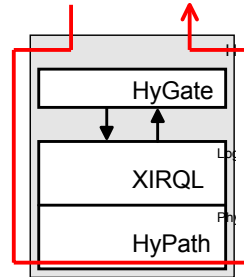


Implementation: Plain HyREX



- CAS+CO topics:
- Automatic conversion to XIRQL

Implementation II: HyPath



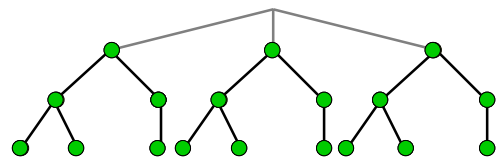
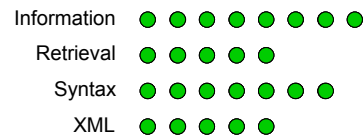
Optimization:

- Provide efficient operator for relevance-oriented search at the physical level
- Bypass XIRQL level

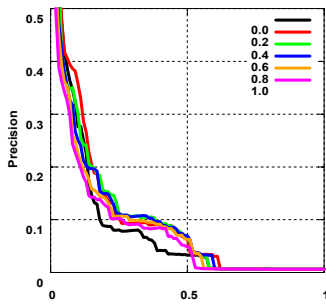
Implementation III: Strategies

- At the level of inverted lists:
 - Sequential read
 - Parallel read
- Parallel read: minimize memory consumption
 - Inverted list entries (posting) contain information about index nodes, augmentation weights, nesting, ...
 - Sorting order for postings: preorder regarding index nodes

Implementation IV: Parallel read



(Preliminary) Evaluation: Effectiveness



- Impact of augmentation factors:
- Small values for lower recall
 - Higher values else

Conclusions

- HyREX: native XML IR system
- Expressiveness of XIRQL covers INEX topics
- Index nodes and augmentation of weights as generalization of classical IR weighting

Outlook

- **Improving efficiency**
 - Early determination of top-ranking documents:
 - Sorting query terms according to importance
 - Sorting inverted lists according to weights
 - Approximate ranking by skipping documents with low indexing weights
- **Improving effectiveness**
 - Adjustment of augmentation factors: level, number of children, type of element, relevance feedback...
 - Other methods for computing trade-off between specificity of answers and indexing weights
- **Consider trade-off between efficiency and effectiveness**