Information Retrieval in Digital Libraries: Dealing with Structure

Norbert Fuhr
University of Dortmund, Germany

December 2000
IR in Digital Libraries

- ad-hoc retrieval
- filtering (alerting)
- networked retrieval
- cross-lingual retrieval
- multimedia retrieval (speech, images, video)

→ fulltext retrieval in structured documents

→ searching and browsing in complex information structures
Query languages for XML retrieval

W3C requirements for XML query language:
“Queries MUST be able to express simple conditions on text”

XQL:
• conditions wrt. to document structure
• Result: subtrees of XML documents
• Boolean retrieval only
XIRQL: XML IR query language

- weighting for uncertain document representations
- datatypes with vague predicates
- relevance-oriented retrieval
XIRQL: Weighting

- probabilistic document indexing
- term independence
- operators: Boolean connectors, weighted sum
  \[ a \lor b, a \land b, \neg a, 0.4a + 0.6b \]
- result: ranking of document nodes
XIRQL: datatypes with vague predicates

- Extensible type hierarchy
  
  text – western-language – English/German

- Set of vague predicates for each datatype

- Datatypes of XML elements defined in extended DTD

Examples:

- chapters with the word "extensible" in the heading:
  
  //chapter[heading contains "extensible"]

- documents with class similar to "H.3.3" and author sounds like "Maier":
  
  //document[@class elsim "H.3.3" and author sounds "Maier"]
XIRQL: Relevance-Oriented Search

- “classical IR” queries: content-oriented, without structural conditions
- traditional IR concept of “document” not clear in XML context
- retrieval strategy: retrieve most specific subtree answering the query
- possible answers: some subtrees too fine-grained/inappropriate
XIRQL: Relevance-Oriented Search (cont.)

index nodes: Types of possible answers
Searching and Browsing in Federated Digital Libraries

- different types of digital libraries
  - Fulltext, Table of Contents, Bibliographic, Reference
- links in and between DL’s not supported
  - explicit - Citations
  - implicit - Coauthoring
- non-uniform user interfaces
- multiple access points
- insufficient functionality
Solution: Daffodil

- **virtual** Digital Library with one access point
- **additional** functionality by exploitation of synergies
- **strategic** support of information seeking process

Implement **strategic** support:

Adopt Bates characterization of information search activities [Bates:90]
Moves

- basic units in the information search process
- moves correspond to the functions of today’s digital libraries

Examples: process a set of query conditions; follow a link

Moves layer
- deal with heterogeneity on syntactical level
- synergies by combining metadata from different DLs
- based on wrapper technology
Strategies

Stratagems

Tactics

Moves

User Interface

• one or a handful of moves with the purpose of improving or speeding a search

*Examples:* generalizing or specialize a search term *(improve recall)*

determine journals or conferences in answer set

find coauthor of an author

**Tactics layer**

– deal with heterogeneity on semantical level
– implementation of powerful browsing facilities
  by using explicit and implicit linking information from various sources
Strategies

Stratagems

Tactics

Moves

- exploit information structures of a given search domain, like author names, classifications, thesauri, journals or conferences

Examples:

area scan (browse a classification)

subject search (browse a thesaurus)

journal/conference browse (browse the aggregation)

Stratagem layer

- implement user guidance through information structures by a given search scheme (strategic plan)
Strategies

- plans which comprise moves, tactics, and stratagems, in order to process a complete information search, i.e. cover an information need

*Example:* simple strategy for paper search in the computer science: search for author met at a conference, search his/her publications, followed by an coauthor search with authors of relevant documents found in the first step
Why an Agent-Based Architecture?

Need for a *flexible* and *extensible* system:

- to add new services (Wrapper) to the system,
- to distribute the agents to different places,
- for different realisations for certain tasks e.g. query expansion:
  - thesaurus
  - word net
  - Excite refine query
- for adaptivity: competitive agents give estimates of costs/benefits,
- for proactivity: agent monitor the search process, suggest tactics,stratagems
The Daffodil Prototype

<table>
<thead>
<tr>
<th>Strategem View</th>
<th>Query View</th>
<th>Attribute View</th>
</tr>
</thead>
<tbody>
<tr>
<td>Message View</td>
<td>Result View</td>
<td>Document View</td>
</tr>
</tbody>
</table>
Strategem selection view

Select stratagem: plain

Plain Search
Subject Search
Aggregation Browse
Author Search
Area Scan
Query view (Plain Search)

New condition

Active query conditions

Author
Keywords
Title
Year

Inactive query conditions
none

Submit  Reset  Apply
Result view

<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Model for Representing and Retrieving Heterogeneous Structured Documents Based on Evidential Reasoning</td>
<td>Mountia Lalmas (Achilles)</td>
</tr>
<tr>
<td>A logical model of information retrieval based on situation theory</td>
<td>M. Lalmas, C. J. Rijssbergen (Achilles)</td>
</tr>
<tr>
<td>Dempster-Shafer's Theory of Evidence Applied to Structured Documents: Modeling Uncertainty</td>
<td>Mountia Lalmas (HC/IB, Achilles)</td>
</tr>
<tr>
<td>From a qualitative towards a quantitative representation of uncertainty in a situation theory-based model of an information retrieval system</td>
<td>Mountia Lalmas (Achilles)</td>
</tr>
<tr>
<td>Information Retrieval and Situation Theory</td>
<td>Th.W.Ch. Hulbers M. Lalmas, C.J. van Rijssbergen (Coa)</td>
</tr>
<tr>
<td>Information retrieval and Dempster-Shafer’s theory of evidence</td>
<td>M. Lalmas (Achilles)</td>
</tr>
<tr>
<td>Logic-based Information Retrieval: Is it really worth it?</td>
<td>P. D. Bruza, M. Lalmas (Achilles)</td>
</tr>
<tr>
<td>Logical models in information retrieval: Introduction and overview</td>
<td>Mountia Lalmas (Achilles)</td>
</tr>
<tr>
<td>Short Queries, Natural Language and Spoken Document Retrieval: Experiments at Glasgow University</td>
<td>Fabio Crestani Mark Sanderson Marcus Theophyllactou, Mountia Lalmas (Coa)</td>
</tr>
<tr>
<td>The use of logic in information retrieval modelling</td>
<td>Mountia Lalmas (Achilles)</td>
</tr>
<tr>
<td>Theories of Information and Uncertainty for the modelling of Information Retrieval: an application of Situation Theory and Dempster-Shafer’s Theory of Evidence</td>
<td>Mountia Lalmas (Achilles)</td>
</tr>
</tbody>
</table>
A Decision-Theoretic Approach to Database Selection in Networked IR

Author(s):
N. Fuhr, Try Homepage

Year: 1996
Month: Jul, 3
Pages: 229 -- 229

Journal: ACM Transactions on Information Systems

Keywords:

Abstract:
In networked IR, a client submits a query to a broker, which is in contact with a large number of databases. In order to yield a maximum number of relevant documents at minimum cost, the broker has to make estimates about the retrieval cost of each database, and then decide for each database whether or not to use it for the current query, and if, how many documents to retrieve from it. For this purpose, we develop a decision-theoretic model using the following parameters for each database: expected recall–precision curve, expected number of relevant documents, and cost factors for query processing and document delivery. For computing the overall optimum, a divide-and-conquer algorithm is given. If there are several brokers knowing different databases, a preselection of brokers can only be ...

Possible actions on this document
- This document has the following external links:
  - fulltext at lb informatik.uni-dortmund.de
  - fulltext at www.acm.org
- Find other documents cited by this document.
- Find other documents citing this document.

Possible actions on picked values:
- Query other documents with picked values instantly.
Attribute view: Classification

- Information Systems
  - GENERAL
  - MODELS AND PRINCIPLES
  - DATABASE MANAGEMENT
  - INFORMATION STORAGE AND RETRIEVAL
  - INFORMATION SYSTEMS APPLICATIONS
  - INFORMATION INTERFACES AND PRESENTATION
  - MISCELLANEOUS
Attribute view: Journal

- journals
  - International Journal on Digital Libraries
    - Volume 1 (1997)
      - Number 1
      - Number 2
      - Number 3
      - Number 4

- conferences
Federated Digital Libraries
Conclusions

Fulltext retrieval in XML documents

- weighting
- datatypes
- relevance-oriented search

ls6-www.cs.uni-dortmund.de/ir/projects/carmen/

Searching and Browsing in federated digital libraries

- strategic support, high-level searching and browsing functions
- users guided through the information structure
- Agent-based architecture: flexibility, extensibility, adaptivity and proactivity.

www.daffodil.de