MIND: An architecture for multimedia information retrieval in federated digital libraries

Henrik Nottelmann
University of Dortmund, Germany
Synopsis

1. Retrieval in Digital Libraries
2. Architecture
3. Terminology
4. Query process in detail
   • query transformation
   • resource selection
   • data fusion
5. Resource gathering in detail
6. Project organisation
MIND: An architecture for multimedia information retrieval in federated digital libraries

Retrieval in Digital Libraries

Henrik Nottelmann, University of Dortmund, Germany
Federated Digital Libraries

• Database-oriented approaches:
  – heterogeneity

• Information retrieval approaches:
  – vagueness and imprecision

• MIND bases on information retrieval approaches, extensions:
  – heterogeneity (e.g. query language, schema)
  – multimedia (text, facts, images, speech)
  – non-co-operative libraries (query interface only)
MIND Architecture

• Dispatcher:
  – library-independent work

• Co-operating proxies:
  – extend functionality of non-co-operating library
  – provide all information required by the dispatcher
  – standard implementation with textual resource descriptions (XML)
Terminology

- Schema
  - Attribute
    - Name
      - Media type
    - Data type
      - Domain
      - Predicate
    - Predicate
Terminology
Terminology
Query Transformation

- Heterogenous schemas
- Required: uncertain mapping between schemas, used to transform user query to proprietary query

Diagram:

- Dublin Core
  - title
  - creator
- RFC 1807
  - title
  - author
Query Transformation

- Heterogenous schemas
- Required: uncertain mapping between schemas, used to transform user query to proprietary query

Diagram:
- Dublin Core:
  - title
  - creator
- MARC 21:
  - 245
  - 100
  - 700
  - 710
Query Transformation

• Task:
  – transform user query to proprietary query

• Proxy:
  – transforms query condition by condition
Query Transformation

• Attribute/Predicate:
  – mapping modeled in probabilistic Datalog
    • probabilistic extension to Horn predicate logic
    • weights for facts and rules
  – certain mapping rules
    \[
    \text{dc_creator_equals}(D, V) \leftarrow \text{marc} \_ 100 \_ equals (D, V)
    \]
    \[
    \text{dc_creator_equals}(D, V) \leftarrow \text{marc} \_ 700 \_ equals (D, V)
    \]
    \[
    \text{dc_creator_equals}(D, V) \leftarrow \text{marc} \_ 710 \_ equals (D, V)
    \]
  – uncertain mapping rules
    \[
    0.4 \text{marc} \_ 100 \_ equals (D, V) \leftarrow \text{dc_creator_equals}(D, V)
    \]
  – rules and probabilities will be learned
Query Transformation

• Comparison value:
  – necessary, when domains do not match
    • dates: “2001-09-09” versus “September 9, 2001”
    • authors: “Fuhr, N.” versus “Norbert Fuhr”
    • classification schemas: DDC versus ACM
    • languages: German versus English
    • image colour histogram: different dimensions
  – transformation:
    • goal: automatic transformation
    • several methods possible, unclear which will be used
    • possibly: simple hardcoding in proxy
Resource Selection

• Task:
  – find relevant libraries w.r.t. the query

• Method:
  – decision-theoretic model
  – cost factors
    • computation and communication time
    • charges for delivery
    • retrieval quality
  – goal: retrieve many relevant documents at low expected costs
Resource Selection

• Task:
  – calculate optimum selection
    • vector $s = (s_1, ..., s_l)^T$
    • expected retrieval costs $EC_i(s_j)$
    • minimal overall (summed up) expected costs

• Proxies:
  – calculate $EC_i(j), 1 \leq j \leq n$

• Dispatcher:
  – calculates optimum selection $s = (s_1, ..., s_l)^T$

$s = (3, 0, 1, 2)^T$

$DL_1$ $s_1 = 3$

$DL_2$ $s_2 = 0$

$DL_3$ $s_3 = 1$

$DL_4$ $s_4 = 2$
Resource Selection
Resource Selection

• expected number of relevant documents in library

\[ E(\text{rel} \mid q, DL) = \sum_{d \in DL} P(\text{rel} \mid q, d) \]

\[ P(\text{rel} \mid q, d) = P(d \rightarrow q) \cdot P(\text{rel} \mid d \rightarrow q), P(\text{rel} \mid \neg d \rightarrow q) \approx 0 \]

\[ P(d \rightarrow q) = \sum_{c_i \in q} P(d \rightarrow c_i) \cdot P(c_i \rightarrow q) \]

\[ E(\text{rel} \mid q, DL) = P(\text{rel} \mid d \rightarrow q) \sum_{c_i \in q} P(c_i \rightarrow q) \sum_{d \in DL} P(d \rightarrow c_i) \]

– required: last sum of indexing weights
Resource Selection

- sum of indexing weights:
  - text, speech:
    - e.g. normalised tf idf values as indexing weight
  - facts, images:
    - feature vectors over continuous domain $V$
    - clusters $V_j \subseteq V$, centroid $v_i$
    - $f: V \times V \rightarrow [0,1]$ retrieval metric
    - approximation for indexing weight sum:
      $$\sum_{j} |V_j| \cdot f(v_j, value(c_i))$$
Data Fusion

• Task:
  – optimise overall retrieval quality

• Proxies:
  – modify weights of their documents (normalisation) based on global idf values
  – provide local df values
  – create summaries

• Dispatcher:
  – merges results w.r.t. normalised document weights
  – computes global idf values
Resource Description

• Schema
• Uncertain schema mapping
• Statistical description of collection:
  – text, speech: terms
    • document frequencies (df)
    • sum of indexing weights
  – facts, images: clusters of feature vectors
    • centroid vector, cluster radius (number of clusters determines granularity of metadata)
    • number of vectors in cluster
Resource Gathering

• Task:
  – create and update resource description

• Proxy:
  – uses query-based sampling for statistical descriptions
    • iterative retrieval of documents
    • assumption: union of results is representative for whole collection
    • extract resource description w.r.t. document sample
  – learns uncertain schema mapping rules
  – goal: learns library schema
Project Organisation

• Funded by the EU commission (FP 5)
• Duration:
  – January 2001 - June 2003
• Project participants:
  – University of Strathclyde (UK) (Coordinator)
  – University of Dortmund (Germany)
  – University of Florence (Italy)
  – University of Sheffield (UK)
  – Carnegie Mellon University (USA)
Conclusion

MIND deals with

• vagueness and imprecision
• heterogeneity
• multimedia
• resource selection
• data fusion
• non-co-operation (resource descriptions)
in federated digital libraries