Multimedia Information Retrieval in Networked Digital Libraries

Norbert Fuhr
University of Duisburg-Essen
Germany
MIND Project

- Task: Development of methods for accessing large numbers of multimedia digital libraries through a single system
- Funded by the EC under FP5 (2/01-7/03)
- Participants:
  - Carnegie-Mellon University
  - University of Duisburg-Essen
  - University of Florence
  - University of Sheffield
  - University of Strathclyde
MIND Architecture

User Interface

Data fuser

Dispatcher

Proxy1

Wrapper1

DL1

Proxy2

Wrapper2

DL2

Proxy3

Wrapper3

DL3

Proxy4

Wrapper4

DL4
Query Processing

- **Query Transformation**: Map between heterogeneous schemas (Dublin Core, MARC 21)
- **Resource Selection**: Determine relevant libraries (more cost effective than querying all libraries)
- **Database Query Run**: Query each selected library (task of “wrappers”)
- **Document Transformation**: Map between heterogeneous schemas (Dublin Core, MARC 21)
- **Data Fusion**: Fuse library results into one single ranked list or several clusters
Retrieval in Federated Digital Libraries

Tasks:

• Extract DL metadata ("resource description")

• Select relevant libraries ("resource selection")

• Communicate with libraries ("wrapper")

• Combine results ("data fusion")
Resource Description
Resource Description: Query-based Sampling

- For non-cooperative DLs
- Generate sequence of queries
- Collect answer documents (~ 300)
- Generate resource description from sampled documents
Resource description for images

- Feature vectors are clustered at different levels of granularity
- For each cluster we consider:
  - Cluster centre $c_i$
  - Cluster radius $r_i$
  - Cluster population $P_i$
- The smaller the cluster radii, the finer the approximation of feature vector density.
Demo: Resource descriptions for images

- Java application demonstrating resource description process for images:
  - Selection of resource to be described
  - Acquisition of individual document descriptors
  - Processing of document descriptors and extraction of resource descriptors at several granularity levels
Resource selection
Resource selection

- Querying all accessible DLs is too expensive
  - Thus route query only to „best“ libraries
- Two competing approaches:
  - Resource ranking:
    - Compute similarity of DL to query (heuristically)
    - Select fixed number of top-ranked DLs
  - Decision-theoretic framework:
    - Assign costs to retrieval (money, time, retrieval quality)
    - Compute selection which minimises costs
      - #selected DLs not fixed
    - Described in this talk
Multimedia retrieval model

- Query conditions $c$ (attribute, predicate, comparison value), weight $Pr(q \leftarrow c)$
  - E.g. term, year number, colour histogram
- Probabilistic indexing weights $Pr(c \leftarrow d)$
  - E.g. BM25 for text, histogram similarity for images
- Linear retrieval function:
  \[ Pr(q \leftarrow d) = \sum_{c_i \in q} Pr(q \leftarrow c_i) \cdot Pr(c_i \leftarrow d) \]
- Mapping onto $Pr(\text{rel}|q,d)$:
  - Linear/logistic function
Probability of Relevance

- Probability of relevance computed based on probability of implication (score)
  \[ \Pr(\text{rel} \mid q, d) = f(\Pr(q \leftarrow d)) \]
  - Linear function with constant \( \Pr(\text{rel} \mid q \leftarrow d) \)
  - New approach: logistic function

\[
f(x) := \frac{\exp(b_0 + b_1 \cdot x)}{1 + \exp(b_0 + b_1 \cdot x)}
\]

- Evaluation: better approximation quality than linear function
Costs Sources

- Computation and communication time
  - Affine linear function
- Charges for delivery
  - Linear function
- Retrieval quality (most interesting for IR)
  - $C^+ < C^-$ for relevant (non-relevant) documents
  - Estimating retrieval quality:
    for result set of $s_i$ documents, estimate the number $r_i$ of relevant documents contained
Estimating Retrieval Quality – Method 1

Relevant Documents in DL

• Resource description:
  – Expectation of indexing weights for terms (images, facts: vector clusters, retrieval function)

• Resource selection:
  – Estimate number of relevant documents in $DL_i$
  – Estimate number $r_i$ of relevant documents in the result set

  • Apply recall-precision-function
    – Approximated by linear function (defined by starting point $P(0)$)
Estimating Retrieval Quality – Method 2

Simulated Retrieval on Sample

- Resource description:
  - Store complete index of sample

- Resource selection:
  - Simulate retrieval on indexed sample (for all media types)
  - Derive distribution of probabilities of relevance
    - Assumed to be representative for whole collection
  - Estimate number \( r_i \) of relevant documents in result set
Estimating retrieval Quality – Method 3

Modelling Indexing Weights

- **Resource description:**
  - Expectation/variance of indexing weights

- **Resource selection:**
  - Approximate indexing weight distribution
    - New: normal distribution
  - Document score distribution
    - Also normal distribution
  - Proceed as for M2
  - Todo: other media types
Estimating retrieval quality

- **Methods:**
  1. Estimate #rel.docs. in DL, apply linear recall-precision function
  2. Simulate retrieval on sample, extrapolate to DL
  3. Normal distribution for indexing weights → normal distribution for retrieval scores

- **Results:**
  - Use logistic mapping for retrieval score → probability of relevance
  - $3 \sim \text{CORI} > 1 > 2$
Data fusion
Data fusion

- Rank data
  - Rank
  - Score
  - Surrogates
  - Duplicates across ranks

- Full document

- Broad information
  - Collection information
  - User preferences
Text data fusion

• Using a combination of evidence
  – Original rank position of documents
  – Re-ranking based on similarity of surrogate to query
    • Surrogate could be title or text summary
  – Promotion of documents found to be similar to others in the rankings
Speech data fusion

- SpeechBot has 50% Word Error Rate on average
  - 17.56% in 300 top ranked documents
  - No need to treat speech differently for fusion

“Speech is spoken and when recognised word errors occur”
Image data fusion – score normalisation

Data fusion

DL Search Engine

DF Search Engine

query

results

un-normalized scores

normalized scores

s=0.8

σ=0.9

s=0.6

σ=0.8

s=0.5

σ=0.8

s=0.3

σ=0.5

s=0.2

σ=0.4
Image data fusion
Presentation of retrieval results
2D Information Display Space (IDS)

• Presenting simultaneously multiple properties about documents
  – each document is represented through a visual object (VO)
  – Visual features of each VO encode relevant properties of documents
    • Position, size, shape, colour
  – Documents sharing the same value of one relevant property (selected by user) are displayed with the same value of the corresponding visual feature
Sample query session

Presentation of results
Sample query session

Presentation of results
MMIR in Networked DLs – Open Issues

- Vague schema mappings for heterogeneous environments
- Cross-media searches
- Resource selection for non-textual media
- Decentralized (P2P) architectures
JXTA Search Architecture

Two-level architecture:

1. Search hubs
2. Provider peers
Conclusions and Outlook

• MIND provides methods for
  – resource description,
  – resource selection and
  – result fusion
in federated multimedia DLs

• Open Issues:
  – Heterogeneous, cross-media environments
  – Time-dependent media
  – Decentralized (P2P) architectures