Resource selection and data fusion for multimedia international digital libraries: An overview of the MIND project.

Forbes Gibb, Department of Computer and Information Sciences, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, United Kingdom

1. Introduction

MIND is a 30 month project funded by IST (IST-2000-26061) which commenced its activities in January 2001. It brings together partners from Europe and the USA with extensive experience in information retrieval and digital libraries. The partners are:

- University of Strathclyde, United Kingdom (Project co-ordinator)
- University of Dortmund, Germany
- University of Florence, Italy
- University of Sheffield, United Kingdom
- Carnegie Mellon University, USA

The inspiration for MIND grew out of the problems which users face when they have remote access to thousands of heterogeneous and distributed multimedia digital libraries. A user must know *where* to search, *how* to query different media, and how to *combine* information from diverse resources. As digital libraries continue to proliferate, in a variety of media and from a variety of sources, the problems of *resource selection*, *query formulation* and *data fusion* become major obstacles to effective search and retrieval.

The key goal of MIND is to develop a common system for identifying, searching and combining results from multiple digital libraries. MIND, therefore, is investigating methods for *resource description and selection* (i.e., gathering and updating information about digital libraries to assist in selecting those which are most likely to contain the information sought), *query processing* (i.e. modifying the terms contained in a query and transforming the query into the local command language), *data fusion* (i.e., the merging of different data retrieved from different digital libraries) and *information visualisation* (in particular, the automatic generation of surrogates and presentation of fused retrieved data).

This paper provides an overview of the following:

- The MIND architecture;
- Key concepts and definitions;
- Query processing;
- Resource description;
- Resource selection; and
- Data fusion.

2. The MIND Architecture

MIND is a distributed system and the architecture consists of four main layers [1] (see figure 1). The bottom layer is composed of digital library resources which may contain uni- or multimedia documents (i.e. combinations of text, speech and images). Each library will typically have their own schema, query language and communication interface. These may be open (i.e. co-operative) libraries, i.e. the indices and/or the documents are visible to external systems; or closed (i.e. non co-operative) libraries, i.e. the indices and/or the documents are visible to external systems. The latter, predominant case presents interesting research challenges in terms of estimating

collection size, term frequency and database similarity. The next layer is composed of the proxies which are representatives for one or more homogeneous digital libraries. These proxies may reside on different servers. The layer above the proxies contains the dispatcher. The dispatcher calls the proxies and therefore needs to know the location of the proxies; they must register themselves with the dispatcher by specifying their URL and port number. The dispatcher resides on a single server. The final layer is the user interface which can be a standalone application or a browser-based interface.

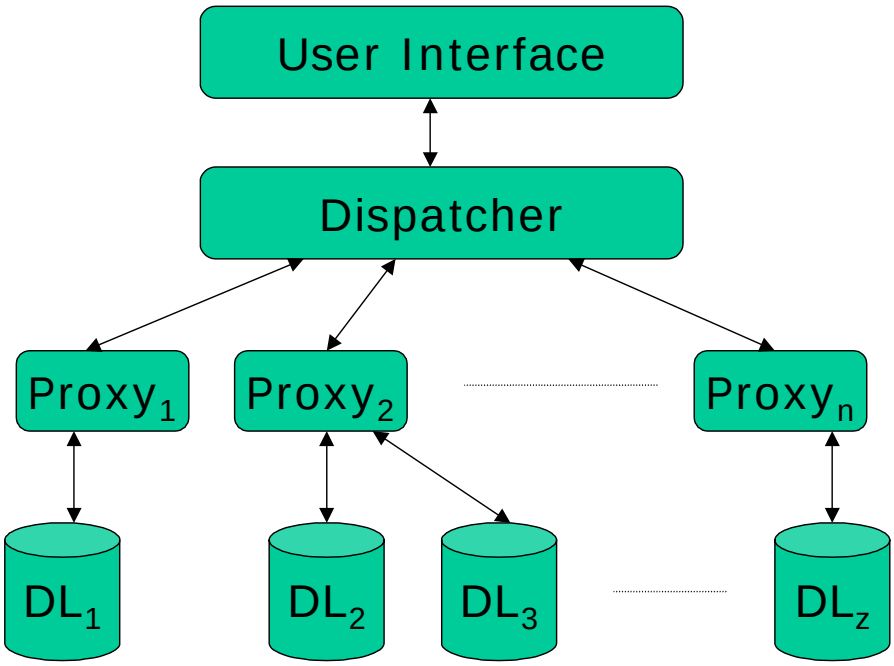


Figure 1. The MIND Architecture

3. Key Concepts and Definitions

3.1 Proxy

As noted above, a proxy is a representative of one or more homogeneous digital libraries which share the same schema, query language and communication interface. The functions which a proxy may carry out in connection with a specific digital library include:

- Generating global information features;
- Generating relevance feedback features;
- Creating document IDs where they are absent;
- Caching results;
- Estimating retrieval costs;
- mapping user queries to proprietary queries;
- Calculating and modifying RSV values;
- Creating summaries;
- Initiating resource description creation and updates.

3.2 Dispatcher

The dispatcher is responsible for managing the query process and calls the proxies when necessary to undertake tasks which are specific to a digital library. Each of the libraries has an associated ID (proxy ID, internal ID). It is also

responsible for undertaking other activities which are independent of individual digital libraries. These include:

- Re-ranking of documents;
- Elimination of duplicate documents; and
- Generating global values.

3.3 Schema

A schema is a set of attributes (name, data type) which is used to structure documents.

3.4 Data Types

A data type consists of a set of valid values, a set of valid predicates, and the media type (i.e. text, fact, image or speech). The attributes are used to model metadata (e.g. author, title, year of publication), in which case the media type is **fact**; and to contain the content of a document, in which case the media type is either **text**, **image** or **speech**.

3.5 Documents

A document is an instance of schema and consists of a set of values corresponding to the set of attributes associated with a data type. A document has an ID of the form (digital library ID, digital library internal ID). If the library does not use an internal ID one is generate by the proxy. A document can contain more than one media type and can also be part of larger document, in which case the attribute **is-part** is used.

3.6 User Query

The user query is the only query which is visible to either the user or the dispatcher. This query contains:

- The optimal number of (relevant) documents that should be retrieved;
- A set of conditions expressed as the tuple (type, weight, attribute, predicate, value, global condition information);
- Relevance feedback data;
- The digital libraries which should be searched;
- User-defined cost parameters; and
- Global condition-independent information.

3.7 Proprietary Query

The proprietary query is transparent to the user and can be interpreted as the translation (where necessary) of the user query using the query language specific to an individual digital library. Clearly a user query may have to be translated into many proprietary queries. The syntax of a proprietary query differs from a user query in two important regards: firstly, the library is implicit rather than explicit; and secondly, an additional attribute (a transformation precision value) is included to reflect the accuracy of the mapping of the user query using the query language of the digital library.

3.8 Result Set

The result set is the set of weighted documents retrieved by the translated user query and consists of a list of tuples of the form (document, RSV obtained from

digital library, new RSV calculated using global information). The result type (which is determined by the dispatcher) may be one of the three following forms:

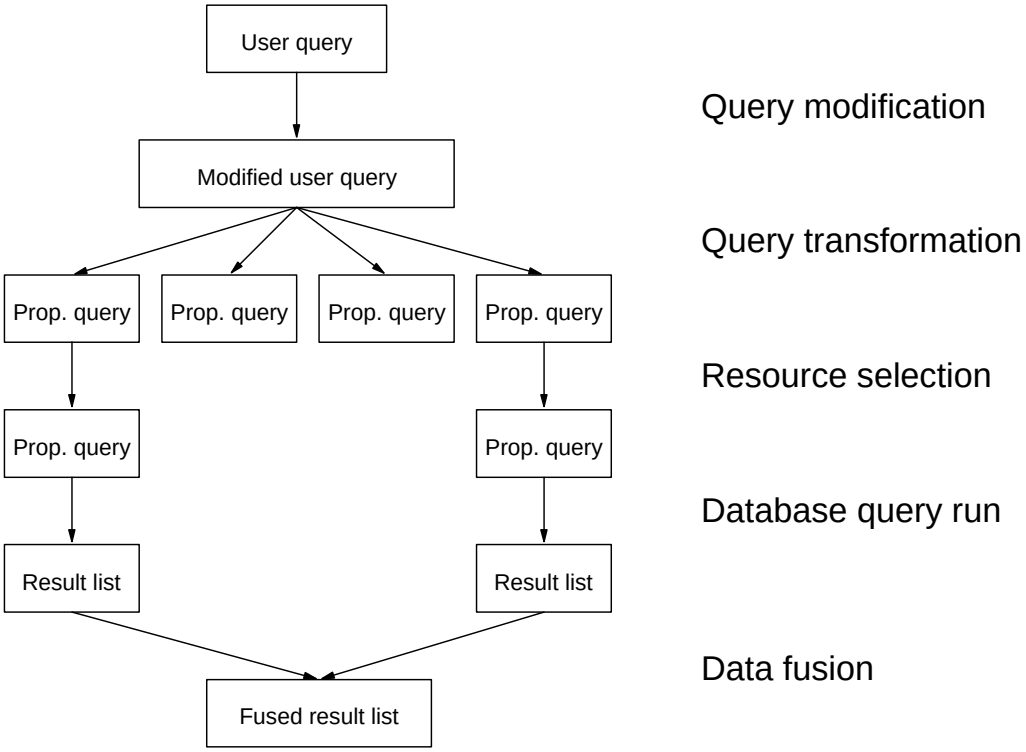
- Empty (i.e. the document is restricted to metadata);
- Summary (these are created by the proxy); or
- Complete.

Empty or summary documents may be presented to the user as complete documents may carry a high intellectual or economic overhead and the data may need to be compressed to facilitate online evaluation.

4. Search and Retrieval

There are five key processes involved in search and retrieval (see Figure 2):

- Query modification
- Query transformation
- Resource selection
- Query execution
- Data fusion



4.1 Query Modification

Query modification is concerned with altering the components of a query and can be viewed as having a primarily semantic focus. Each of the proxies is responsible for generating global information features which are specific to the digital libraries which it represents. For instance, descriptors may include:

- The number of documents in a digital library which contain a text condition term (i.e. the df values);
- The colour histogram for an image; or
- The word error rate for a transcribed piece of speech.

The dispatcher will then collect these library specific information features and use them to generate global values which can be used to modify the query conditions and condition weights. Relevance feedback data can also be used to modify the query. Relevance feedback data can be supplied directly by the user or calculated by the interface on the basis, for instance, of binary relevance judgements made by the user when evaluating results.

4.2 Query Transformation

Query transformation is concerned with converting from the schema used to represent the user query to the schema specific to each of the digital libraries and can be viewed as having a primarily syntactic focus. Each condition of the user query is transformed into a set of proprietary conditions. The output of this transformation is termed a proprietary query within the MIND framework. This transformation is the responsibility of the proxies and is transparent to the user. A key issue with query transformation is the confidence that can be placed in the mapping between the MIND schema and that used by each of the digital libraries. This is a common problem when considering co-operation between digital libraries as the mappings may be incomplete. Incompleteness may be the result of differing data structures, differing command languages, different weighting algorithms, etc. MIND therefore incorporates a precision value for the transformation which is used, inter alia, to rank the documents which are retrieved from individual digital libraries.

4.3 Query Execution

Once the query has been transformed into the schema of the digital library(ies) local to the proxy it is then executed. The result is set of pairs (document, RSV obtained from digital library).

4.4 Resource Selection

The key goal of resource selection is to identify the digital libraries which are most likely to produce relevant documents in response to the user query. In addition the dispatcher is responsible for establishing the optimal number of documents that should be retrieved in order to minimise the overall costs. MIND incorporates a decision-theoretic model (proposed by Fuhr [2]) which considers a number of cost factors including: time, money and quality [3]. The proxy is responsible for calculating retrieval costs while the dispatcher is responsible for calculating the optimal number of documents to be retrieved.

4.5 Data Fusion

The documents which are retrieved from each of the selected digital libraries will be weighted using the information which is specific to that digital library. These weights (RSVs) will therefore reflect the local term frequencies and occurrences. The proxies are responsible for calculating new RSVs (i.e. normalising RSVs) in order to improve retrieval quality. The dispatcher then reranks the documents based on the new RSVs. In addition, duplicate documents may be retrieved from individual digital libraries. The dispatcher is also responsible for detecting and eliminating duplicate documents. The detection of duplicates and quasi-duplicates presents a number of research as well as operational issues. For instance, the same document may be stored in slightly different ways by individual digital libraries. The elimination of duplicates may also produce a set of documents which is fewer than that specified in the user query. In this case the query is re-submitted with a higher value for required number of documents.

4.6 Resource Gathering

Resource gathering is concerned with creating a new, or updating an existing, resource descriptor. Resource descriptors are used to store information about the content of a digital library and are created by query based sampling of each digital library [4]. The descriptors are held by the proxy using a standard format and are used inter alia to calculate the likely number of relevant documents that are contained in a digital library and the cost of retrieving a pre-determined number of documents form that library. The proxy also routinely updates its associated resource descriptors using iterative query-based sampling. Resource descriptors consist of two parts: a specification of the database schema and a specification of content features. The database specification consists of a schema name, a set of attributes (name, data type) where the data type defines the domain (i.e. the set of valid values), a metadata flag, the media type and a set of valid predicates.

5. Summary

MIND addresses a number of problems raised by distributed digital libraries:

- Heterogeneity of database schema, media and command languages;
- Resource description creation and updating;
- Optimal resource selection;
- Data fusion of duplicate and quasi-duplicate documents;
- Multimedia documents and associated content representation.

Some of the open issues which are currently being explored include:

- Estimating digital library size;
- Sampling algorithms;
- Extending the decision-theoretic framework;
- Fusion of different media;
- Resource description elicitation;
- Relevance estimation;
- Digital library similarity; and
- Visualisation of fused results.

---

[1] Nottelmann, H. Test-bed architecture specification. Dortmund: The MIND Consortium, 2001. (Deliverable D1.1)
[2] Fuhr, N. A decision-theoretic approach to database selection in networked IR. ACM Transactions on Information Systems, 1999, 17(3), 229-249.
[3] Nottelmann, H. and Fuhr, N. Resource selection framework and methods. Dortmund: The MIND Consortium, 2002. (Deliverable D3.1)
[4] Pala, P. and Berretti, S. Definition of content metadata structure for text, audio and images. Florence: The MIND Consortium, 2001. (Deliverable D2.1)