MIND: Resource Selection and Data Fusion in Multimedia Distributed Digital Libraries

Stefano Berretti[1], Jamie Callan[2], Henrik Nottelmann[3], Xiao Mang Shou[4], Shengli Wu[5]
[1] Dip. Sistemi e Informatica, Università degli Studi di Firenze, Firenze, Italy
[2]School of Computer Studies, Carnegie Mellon University, Pittsburgh, PA, USA
[3]Institute of Informatics and Interactive Systems, University of Duisburg-Essen, Duisburg, Germany
[4]Dept. of Information Studies, University of Sheffield, Sheffield, UK
[5]Dept. Computer and Information Sciences, University of Strathclyde, Glasgow, UK
berretti@dsi.unifi.it, callan@cs.cmu.edu, nottelmann@uni-duisburg.de,
x.m.shou@shef.ac.uk, shengli.wu@cis.strath.ac.uk

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software

General Terms

Design

Keywords

Networked Retrieval, Data Fusion, Resource Selection

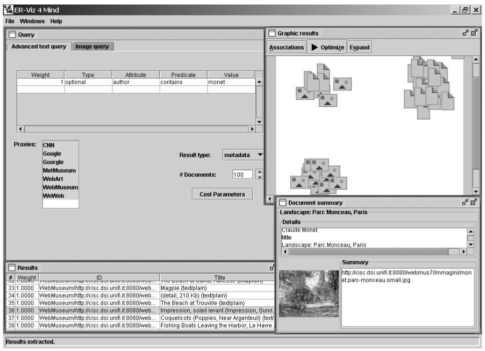
The MIND Project

With thousands of DLs in the web, a user has to select relevant DLs (*resource selection*), reformulate the query w.r.t. each DL (*schema mapping*) and reorganise the different results (*data fusion*). The EU project MIND attempts to overcome these ineffective manual tasks, giving the user the impression of dealing with a homogeneous multimedia DL, handling text, facts, images, and speech. All required steps are done automatically and invisible to a user.

System Architecture

Because of MIND's modular architecture extending the system as well as distributing the components on different machines is easy. Communication is done via SOAP.

Proxies extend the functionality of the corresponding cooperating or non-cooperating DL (e.g. cost estimation for resource selection) and hide the inherent heterogeneity in federated DLs (e.g. by transforming queries and documents). With a standard proxy implementation and textual *resource descriptions* for library-specific information, new DLs can be integrated easily, since only a *wrapper* communicating with the corresponding DL has to be implemented. Media-specific work is carried out in media-specific proxy components. Following the standard structure of federated systems, MIND also has a mediating instance, the *dispatcher*, for proxy-independent work (e.g. selecting DLs). We kept the *data fuser* separate to allow for the use of different implementations.



Weight	Type	Attribute	Predicate	Value
1	optional	author	contains	monet

Proxies: CNN, Google, Goorgle, MetMuseum, WebArt, WebMuseum, WnWeb

Result type: metadata

Documents: 100

Cost Parameters

Landscape: Parc Monceau, Paris

Details

Claude Monet
title
Landscape: Parc Monceau, Paris

Summary

http://cisc.dsi.unifi.it:8080/webmus7/Immagini/monet.parc-monceau.small.jpg

#	Weight	ID	Title
33	1.0000	WebMuseum/http://cisc.dsi.unifi.it:8080/web...	Magpie (text/plain)
34	1.0000	WebMuseum/http://cisc.dsi.unifi.it:8080/web...	(detail, 210 Kb) (text/plain)
35	1.0000	WebMuseum/http://cisc.dsi.unifi.it:8080/web...	The Beach at Trouville (text/plain)
36	1.0000	WebMuseum/http://cisc.dsi.unifi.it:8080/web...	Impression, soleil levant (Impression, Sunri...
37	1.0000	WebMuseum/http://cisc.dsi.unifi.it:8080/web...	Coquelicots (Poppies, Near Argenteuil) (text/
38	1.0000	WebMuseum/http://cisc.dsi.unifi.it:8080/web...	Fishing Boats Leaving the Harbor, Le Havre ...

Results extracted.

Figure 1: The MIND system interface

User interfaces

Different *user interfaces* (UIs) can be connected to the MIND backend. Within the project we developed a HTML-based and two Java-based UIs.

One of the Java UIs presents documents in a graphic 2D Information Display Space (IDS) (figure 1). The left side contains the query panel, supporting query specification by text and visual features, and a single ranked list of result items ordered by matching score. On the upper right, the graphic result panel retrieval shows the IDS where results are associated with icons. In this example, close icons represent documents returned by the same digital library. Every time a document is selected in one of the two result views, information (text and/or images) about the corresponding document are shown in the document summary panel (shown on the lower right part of the interface).

Alternatively, documents can be presented in the UI organised in a topical hierarchical structure, built using complete link hierarchical clustering, where documents can be browsed in a "scatter-gather" way. This way of presenting documents is also very useful to identify document duplicates, which could not be picked up by the data fusion due to the use of different document identifications. Despite the retrieved documents being very differently represented (from full text to title only), the hierarchical clustering appears to be quite robust.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.