

MultiMatch: Multilingual / Multimedia Access to Cultural Heritage

Franca Debole
ISTI-CNR

MM Achievements

- MultiMatch developed a **Web search engine** specialized in the **cultural heritage** domain.
- Queries can be expressed in **multiple languages**
- **Multilingual** and multimedia retrieval
- Access to **multiple sources of information**
 - Web sites containing authoritative cultural heritage data (e.g. museums, cultural institutions, CH educational sites, tourist information portals)
 - IPR protected CH material, provided by content providers (e.g. Alinari, Sound and Vision, Biblioteca Virtual de Cervantes)
 - Other OAI compliant resources

The Project

- VI FP Project
- Technology-enhanced Learning and Access to Cultural Heritage
- Started: May 2006
- First prototype: August 2007
- Second Prototype: July 2008
- Evaluation and field trials: October 2008

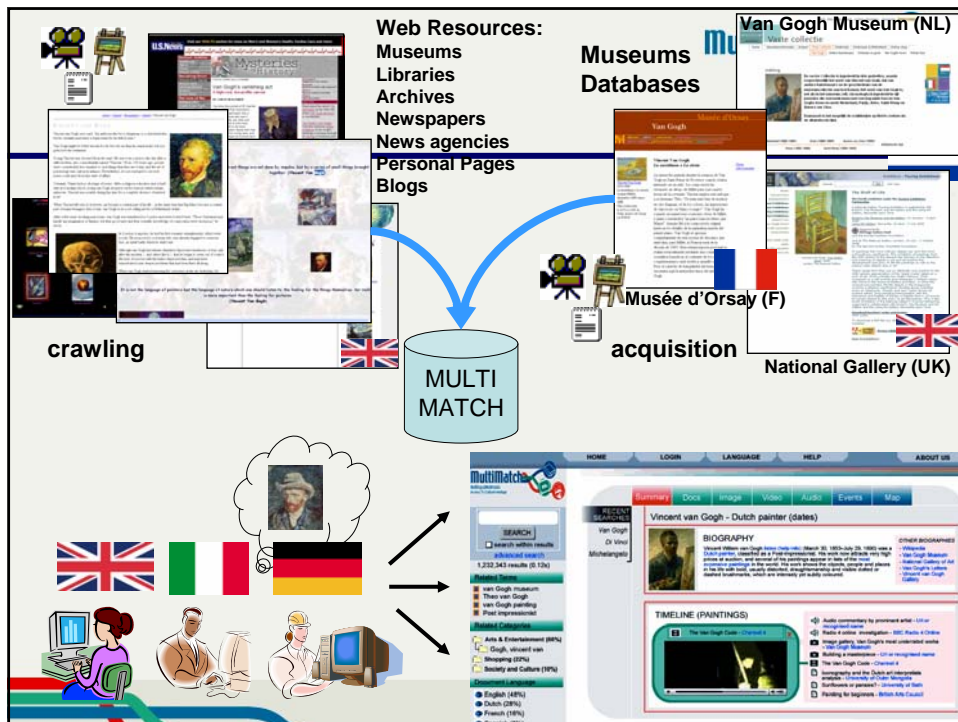
MultiMatch Partners

- **Cultural Heritage**
 - Fratelli Alinari Istituto Edizioni Artistiche SpA (Alinari)
 - Netherlands Institute for Sound and Vision(Sound and Vision)
 - University of Alicante – Biblioteca Miguel de Cervantes (UA-BVMC)
- **Industry**
 - OCLC PICA (FDI)
 - WIND Telecomunicazioni S.p.A. (WIND)
- **Academia**
 - Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR)
 - University of Sheffield (USFD)
 - Dublin City University (DCU)
 - University of Amsterdam (UvA)
 - University of Geneva (UniGE)
 - Universidad Nacional de Educación a Distancia (UNED)

MultiMatch Vision



- Information needed is likely to be available on the Internet but to access it **language, media, and source boundaries** need to be overcome
- Key ideas underlying the MultiMatch vision
 - **Multiplicity**
 - » Queries and retrieved items in multiple languages
 - » Queries and retrieved items in multiple media
 - » Items selected from multiple sources
 - **Aggregation**
 - » Presentation of aggregated results
 - » Relationships between retrieved items



Main Research Challenges

- **Focused crawling** for acquisition of CH-related information from heterogeneous MM resources
- CH **concept and relation extraction** using information extraction and text mining techniques
- Multimedia search
- Mixed media search
- **Multilingual management** with support for query formulation, cross-language retrieval

Content sources overview

- Additional content from project partners
 - Sound and Vision – video
 - Alinari – images
 - BVMC – text/ web pages
 - UvA – web documents / internet audio
 - AISA photographic agency
 - Michael Plus MLA (UK)
 - Michael Plus ICIMSS(PL)
 - Michael Plus ONB (AT)
 - The European Library (TEL)
 - Teche RAI (video)(IT)
 - Selection of 35 European OAI sources

Indexing and information extraction



- Automatic extraction of indexing features for all media (text, speech, images, video) and crawled data
- Automatic generation of inter-document links
- Development of algorithms for classification and information extraction
 - Creators, type/genre, subject, place/time, art objects/works
- Semantic enrichment of documents




Multimedia search




- Similarity search based on visual features (low level and high level – faces, objects, etc.)
- Efficient retrieval
- Support of relevance feedback and interactive search
- Combined text and visual search


Query:



QueryRequest:



Result:



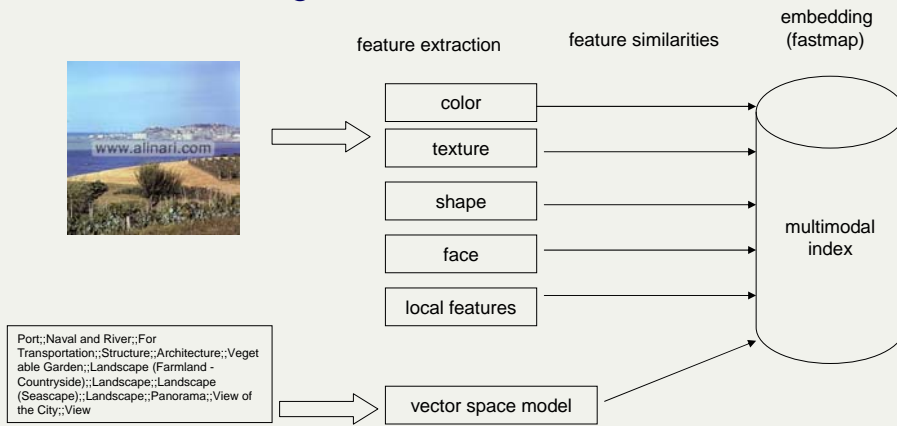
The result grid shows 15 small image thumbnails arranged in three rows of five. Each thumbnail includes a similarity score (e.g., 'Similarity: 0.12345'), a 'relevance' button, and a 'title' button. The first result is labeled 'Query Image'.



MultiMatch CMSE



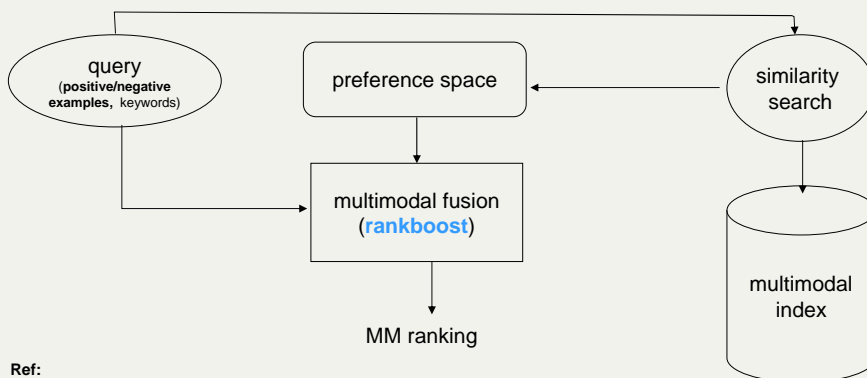
Data indexing



CMSE



Data retrieval



Ref:
 Combining multimodal preferences for multimedia information retrieval.
 Eric Bruno, Jana Kludas, and Stéphane Marchand-Maillet.
 MIR 2007, Augsburg, Germany.



Multilingual support

- Provide system with monolingual and multilingual search functionalities (initially four languages, extended to six)
- Provide effective translation strategies e.g. multilingual **dictionaries**, **machine translation**, thesaurus term expansion, combined hybrid translation methods
- Multilingual query expansion
 - Relevance and pseudo relevance feedback

MachineTranslation

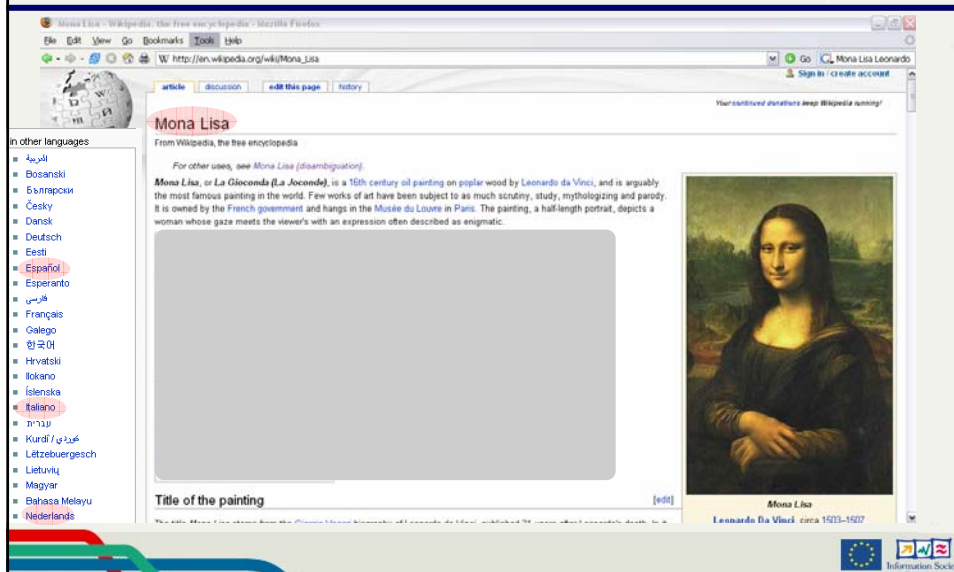


- WorldLingo commercial machine translation system used under licence
- Supports all 30 language pairs for the six selected languages
- Easy to use and integrate into prototype
- Good well documented API

Dictionary-based Query Translation

- Word-by-word (phrase-by-phrase) translation via a bilingual dictionary look-up
- Translation resources
 - General translation lexicon
 - FREELANG dictionary* www.freelang.net
 - Universal dictionary* www.dicts.info
 - Domain-specific translation lexicon
 - Wikipedia* www.wikipedia.org
 - Others
 - XDXF archives* xdxf.revdanica.com/down/

Domain-specific Translation Example



The screenshot shows a web browser displaying the Wikipedia article for "Mona Lisa". The browser's address bar shows the URL "http://en.wikipedia.org/wiki/Mona_Lisa". The article title "Mona Lisa" is highlighted in red. Below the title, there is a list of "in other languages" with "Español" selected. The main text of the article is visible, starting with "Mona Lisa, or La Gioconda (La Joconde), is a 16th century oil painting on poplar wood by Leonardo da Vinci, and is arguably the most famous painting in the world. Few works of art have been subject to as much scrutiny, study, mythologizing and parody. It is owned by the French government and hangs in the Musée du Louvre in Paris. The painting, a half-length portrait, depicts a woman whose gaze meets the viewer's with an expression often described as enigmatic." To the right of the text is a small image of the Mona Lisa painting. The browser's status bar at the bottom shows "Mona Lisa" and "Leonardo Da Vinci circa 1503-1507".

Hybrid Translation

- Full machine translation remove ambiguity in translation, but can have poor coverage of important concepts in CH domain
 - Often shown to be effective in CLIR evaluation campaigns
- Dictionaries offer alternative translations and flexibility, but are often highly ambiguous
 - Can be preferred by users since operation is most transparent, even if retrieval accuracy can be lower!

Hybrid Translation

- Proposed alternative for MultiMatch – hybrid translation combining full machine translation and domain-specific bilingual dictionaries:
 - Search text for phrases contained in bilingual dictionary and mark them
 - Pass text of machine translation system
 - Replace or append known phrases in translated output

G.Jones, F.Fantino, E.Newman and Y.Zhang (CLIA 2008)

E.Newman, Y.Zhang, M.Fuller, G.Jones, C.Peters (LaTech 2008)(submitted)

Hybrid Translation Examples

- Original (Italian): Ubicazione della Stele di Rosetta
- MT (to English): Location of the Stele of Rosette
- Hybrid (to English): Location of Rosetta stone

- Original (Italian): Arnaldo Pomodoro pittura
- MT (to English): arnaldo tomato painting
- Hybrid(to English): Arnaldo Pomodoro painting

Conclusions

- Further evolutions of MultiMatch
 - Use of MultiMatch technologies to build large scale Digital Libraries and a large scale search engine specialized for Cultural Heritage
 - » Enlarge the content base through the access to a complete set of CH sites and crawling of a significant part of the web
 - » Enlarge the number of languages managed (possibly to cover all EU languages)
 - » Invest on system efficiency, system scalability, and robustness

Further information

- MultiMatch Web site
 - <http://www.multimatch.org>
- Project coordination: Pasquale Savino
 - pasquale.savino@isti.cnr.it
- Technical coordination: Giuseppe Amato
 - giuseppe.amato@isti.cnr.it
- User group coordination: Sam Minelli
 - sam@alinari.it