www.trebleclef.eu

TrebleCLEF

Multimatch System Overview
Franca Debole, ISTI-CNR,Pisa, Italy

TrebleCLEF Summer School on
Multilingual Information Access
Hotel Santa Croce in Fossabanda, Pisa, Italy
15-19 June 2009

MultiMatch
System Overview

Franca
Debole

ISTI – CNR

Outline

- ## MM System Idea

- ## MM System Architecture

- ## MM System Modules

- ## MM System Details

Web from:

- Museums
- Libraries
- Archives
- Newspapers
- News agencies
- Personal Pages
- Blogs

Museums Databases

Van Gogh Museum

Musée d'Orsay

National Gallery

Image Similarity

Full Text

Media Repository
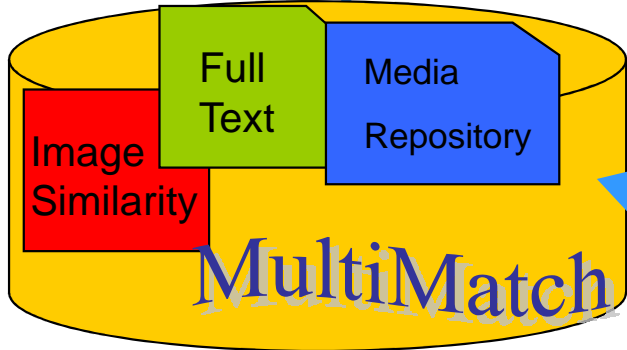
MultiMatch

Interface language: En Sp It Du

Overview | Web | Archives | Video | Image | Audio

van gogh — Search — Show advanced search / Translate my query

I write in: English Spanish Italian Dutch

English | Spanish | Italian | Dutch

Overview results for the query "van gogh"

1 - 4 Web results of about 100 results for van gogh in English (0.32 seconds)

1 - 4 Image results of about 9 results for van gogh in English (0.54 seconds)

WebMuseum: Gogh, Vincent van: Vineyards
www.ibiblio.org/wm/paint/au...
Add to workspace
Show properties

Vincent Van Gogh (1853 - 1890) Artwork Images, Exhibitions, Reviews
wwar.com/masters/v/van_gogh...
Add to workspace
Show properties

Browse by Artist - Norton Simon Museum
www.nortonsimon.org/collect...
Add to workspace
Show properties

Browse by Artist - Norton Simon Museum
www.nortonsimon.org/collect...
Add to workspace
Show properties

Painting by Vincent Van Gogh entitled 'Women of Arles' in the Hermitage in Saint Petersburg. In the foreground, right, two half-length figures of women. On the left some flowers and a woman taking care of a garden
multimatch01.isti.cnr.it:80...
Add to workspace
Show properties

Painting by Vincent Van Gogh of the 'Prison Rounds', in the Pushkin Museum in Moscow. A group of prisoners circle is dragging itself along in the courtyard of a prison, under the vigil eyes of the guards
multimatch01.isti.cnr.it:80...
Add to workspace
Show properties

System Architecture

Information Extraction

User Interface

Searchable Metadata Repository

Enhanced Information Retrieval
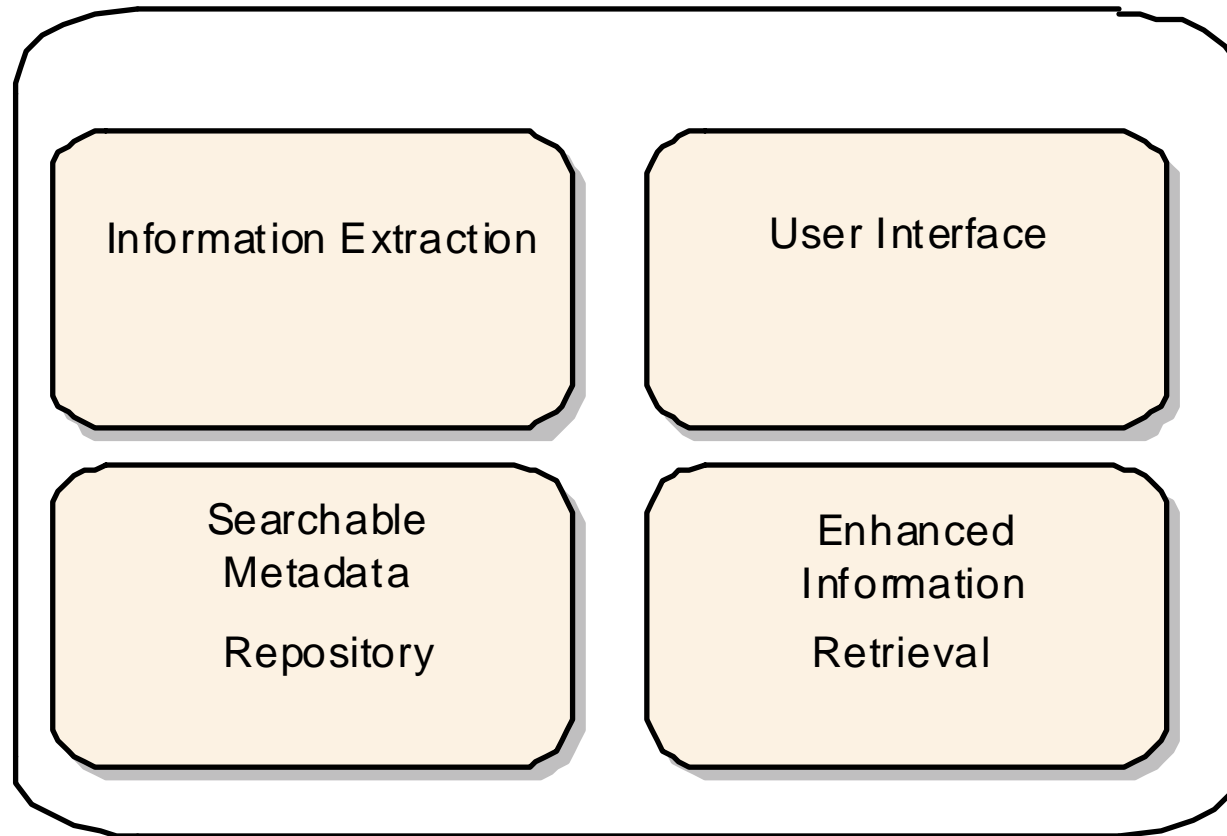
Information Extraction

- Locates MultiMatch Objects (via Web crawl + CH Selection)

- Translates and augments MultiMatch objects to produce MultiMatch Metadata records for them

- Generates Term Clouds

- Classificates automatically the metadata

Searchable Metadata Repository

- Stores MultiMatch metadata documents
- Allows these documents to be optimally searched
- Supported by MILOS
  - Lucene plug-in for full-text searches
  - Fusion plug-in for multi-modal searches
  - Entity disambiguator
  - MPEG-7 index plug-in

Enhanced Information Retrieval

- Public-facing Web Services used by the UI:
  - Image Browse
  - Query Expansion
  - Translation to 6 Languages (+GER & +POL)
  - Video playback
  - Creator/Creation (suggest; retrieve; relationship networks)

User Interface

- Searching
  - Standard search
  - Specialist audio, video and image interfaces
  - Search suggestions
  - Query Translation
- Result Display
  - *Creator/Creation profile*
  - *Creator/Creation network*
  - *Term clouds*
  - *Timelines*
  - *Geographical maps*
- Log-in / UI Personalisation

IE pipeline

act Metadata + Digital Artifact

Native
Metadata
+ Digital
Object

Crawled data &
other

**Metadata
Transformation**

Metadata
+ Digital
Object

Translation

Metadata
+ Digital
Object

Metadata
+ Digital
Object

Classification

Summarisation

**Term Cloud
Generation**

Metadata
+ Digital
Object

Metadata
+ Digital
Object

Metadata
+ Digital
Object

MultiMatch Object for
Storage
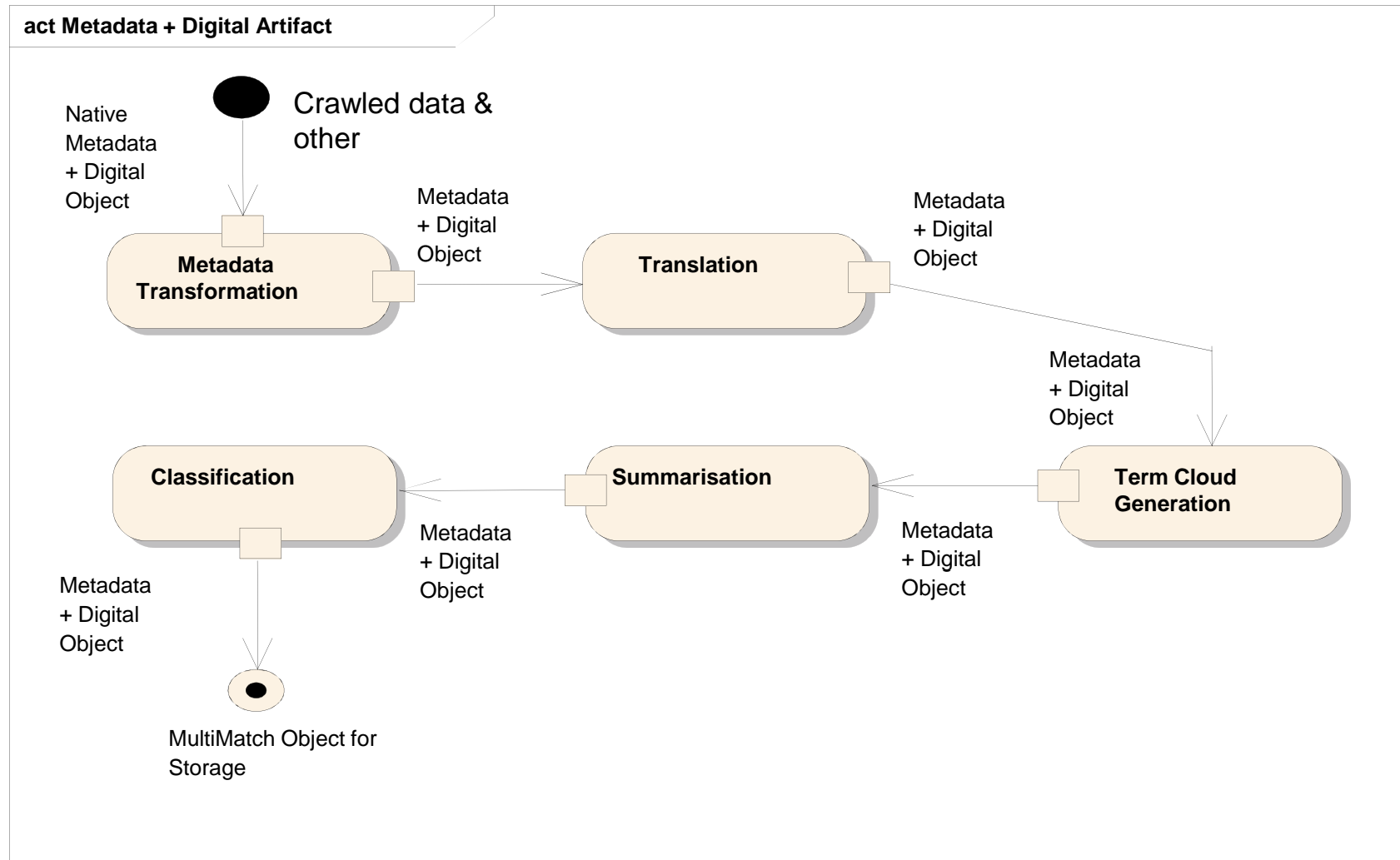
Crawling

- Web page from whitelist of CH sites
  - Fine arts and artists (movements and styles)
  - History and Events (historical events)
  - Architecture and views

- Wikipedia contents
  - Idem
  - Literature (biographies,reviews of works)
  - Architecture and views (places)
  - Films theatre and Music (film directors)

- Dynamic data (RSS, Feeds)

Term Clouds

- Identify most important/ relevant/ informative terms in documents
- Score for each term can be based on
  - tf.idf
  - Term position
  - Document syntax
- Process Lucene index
  - Generate corpus statistics
  - Generate and insert static data

Searchable Metadata Repository
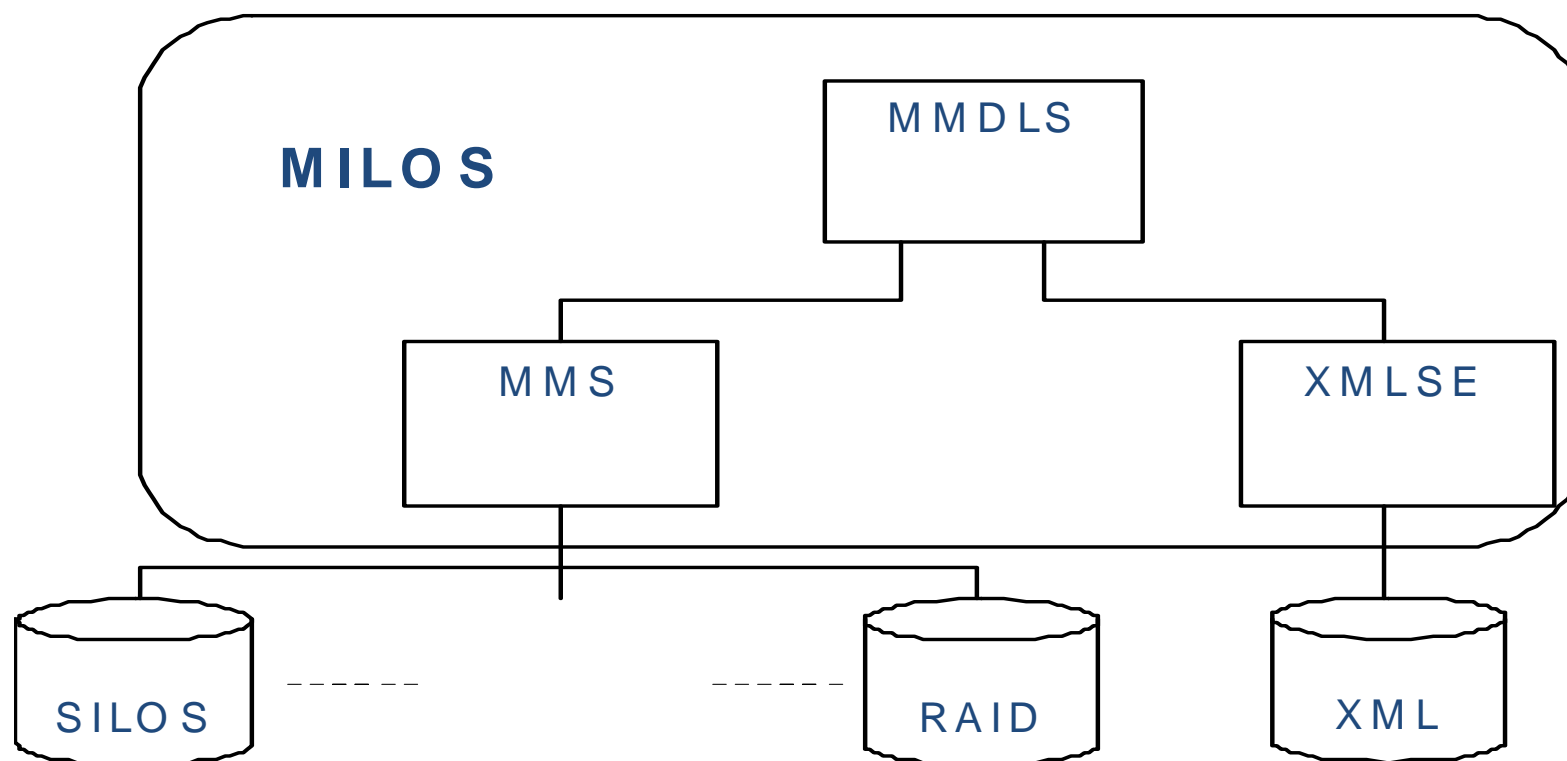
Searchable Repository MILOS

❖ Basic functionalities:
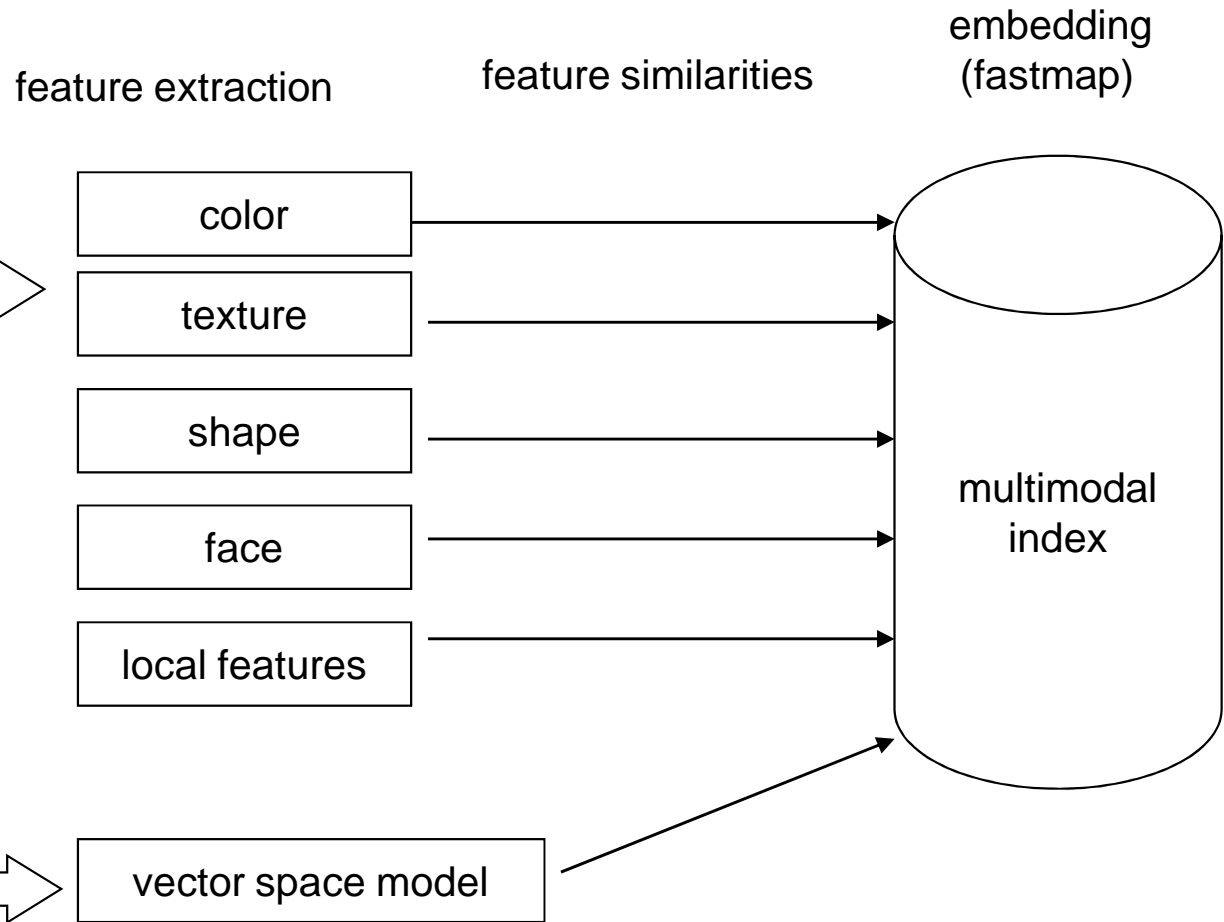 - Document storage
  - Efficient and effective retrieval
  - Efficient and effective metadata management
  - Searching and managing multimedia documents
  - Scalability

❖ Additional needed functionalities
  - Managing heterogeneous documents
  - Managing heterogeneous metadata
  - Applications provided with metadata schema independece
  - Plugins :: LuceneMLingual, Fusion (CMSE)

MultiLingual Plugin

- Milos plugin for handling full-text search in multiple languages.

- Ingest original language documents
  - Translate every document in every language-index

- Insert translation into ML index with a reference to the source document.

- Use classic Lucene functionalities (stemming, search language, ...)

Cross Modal Search Engine

- Data indexing

feature extraction   feature similarities   embedding (fastmap)

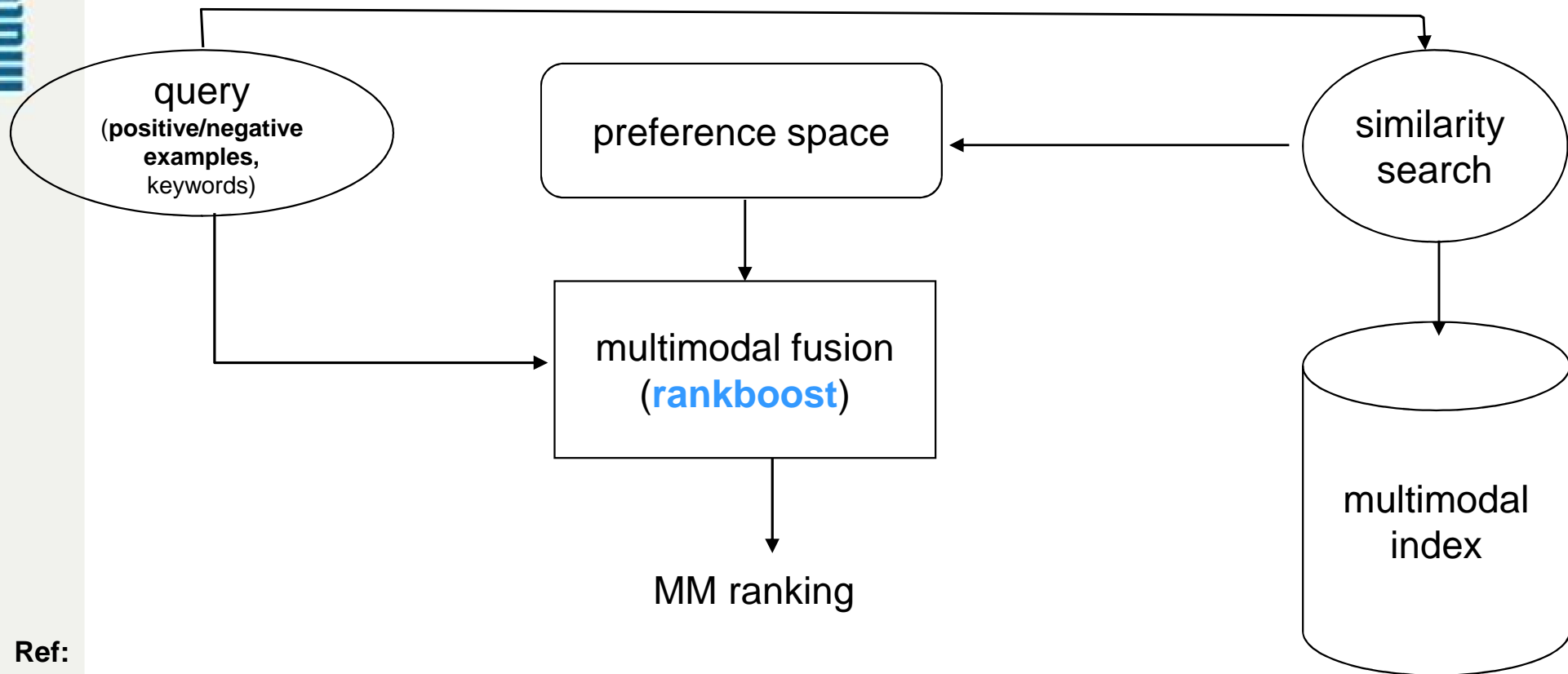| color |
| texture |
| shape |
| face |
| local features |

multimodal index

Port;;Naval and River;;For Transportation;;Structure;;Architecture;;Veget able Garden;;Landscape (Farmland - Countryside);;Landscape;;Landscape (Seascape);;Landscape;;Panorama;;View of the City;;View

vector space model

CMSE

- ## Data retrieval

query
(**positive/negative examples,** keywords)

preference space

similarity search

multimodal fusion
(**rankboost**)

MM ranking

multimodal index

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

Translation Query Service

- Input:
  - Text string to be translated

- Output:
  - Object containing the complete hybrid system translation, plus term-for-term translation where available.

Hybrid Query Translation

For example given an Italian query "Rievocazione storica in paese", we want to translate it into English :

1. "Rievocazione storica" is detected as a phrase using dictionary-based phrase identification with greedy-parsing algorithm;

2. We re-formulate query in the following manner:

    <html><!-- Historical reenactment --> in paese</html>

    where the term within the javascript tag <!-- --> will be skipped by Worldlingo.

3. The hybrid translation output is "Historical reenactment in country".

Support to MLingual

- Existing dictionaries have been refined (English, Dutch, Italian, Spanish) and extended to include German and Polish.
  - Total of 30 bilingual translation lexicons have been built.
- Multiple translations are now ranked based on their frequency in the MM text collections.
- Accuracy of translations needs to be manually checked!

Query Suggestion

- Web service to provide "auto-complete" functionality to users
  - User types in some characters to start
  - Service sends back matching entries

- Suggestion using the list of creators/creations instead
  - It ensure that suggestions would be found in MM Document Collection

UI

I will show you something….

ASAP

Comments ?
Questions?
Or
......

Let's try the Beta System?