Task 2.6
Advanced Access Structures for Complex Similarity Measures

Participants:

UMIT	Sören Balko, Hans-J. Schek
ISTI-CNR	Giuseppe Amato
MUNI	Pavel Zezula

Giuseppe Amato
giuseppe.amato@isti.cnr.it

Task 2.6: Objectives (JPA2)

- General Objective
  - Develop and enhance state of the art techniques for efficient similarity search
- Specific Objectives (JPA2)
  - develop index structures that
    - efficiently support nearest neighbour, range and ranking queries
    - operate on any kind of multimedia data
    - are generic in the metric distance measure to be employed
- Operational Objectives (JPA2)
  - Integrate existing metric indexing with VA-file like quantisation approaches
  - Design of distributed access methods for similarity search in metric spaces
  - Adoption of similarity search techniques in digital libraries by way of XML encoded metadata
  - Survey of the state of the art on similarity search in metric spaces

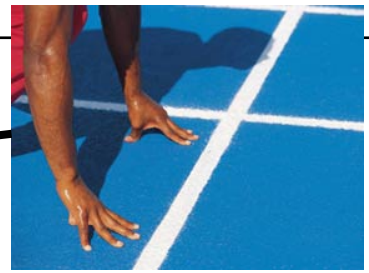
Overview of the presentation

- Similarity search and digital libraries: basics
- Work carried out in JPA2
  - Integrate existing metric indexing with VA-file like quantisation approaches
  - Design of distributed access methods for similarity search in metric spaces
  - Adoption of similarity search techniques in digital libraries by way of XML encoded metadata
- Future work

Overview of the presentation

- Similarity search and digital libraries: basics
- Work carried out in JPA2
  - Integrate existing metric indexing with VA-file like quantisation approaches
  - Design of distributed access methods for similarity search in metric spaces
  - Adoption of similarity search techniques in digital libraries by way of XML encoded metadata
- Future work

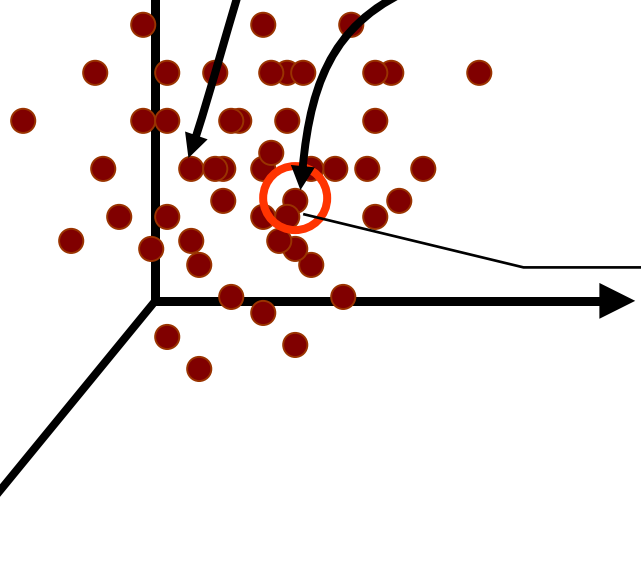
Similarity search: intuition



Feature extraction

Feature extraction

Query image



Query neighbourhood

Similarity search and Digital Libraries

- Similarity search offers a new search paradigms to multimedia intensive Digital Libraries
  - Sometimes the content of documents is not sufficiently described in manually generated metadata
  - Query by examples, using the similarity search (or content based search) paradigm is a good solution
- Examples:
  - Was this shot (a shot very similar to this) used in other movies?
  - Is there another picture similar to this?
  - Is there a picture containing something similar to this?
  - ...
- Similarity search is intended to be used with (no to substitute) other search paradigms
- Particularly useful to solve unplanned or unforeseen searches
  - Example: was a censored shot used in some documentary without permission?
    - Clearly if someone used a censored shot, no record was maintained of it
      - similarity search can help to easily discover such occurrences

Similarity search: metric space approach

- Metric spaces:
  - No assumption on object representation
  - Distance d between objects should satisfy the following:

$$(1)\ d(O_x, O_y) = d(O_y, O_x) \quad \text{(symmetry)}$$

$$(2)\ 0 < d(O_x, O_y) < \infty, O_x \neq O_y \quad \text{(positiviness)}$$

$$(3)\ d(O_x, O_x) = 0 \quad \text{(reflexivity)}$$

$$(4)\ d(O_x, O_y) \leq d(O_y, O_z) \leq d(O_z, O_y) \quad \text{(triangle ineq.)}$$

  - Vector spaces + Minkowsky distances are metric spaces

Overview of the presentation

- Similarity search and digital libraries: basics
- Work carried out in JPA2
  - Integrate existing metric indexing with VA-file like quantisation approaches (UMIT)
  - Design of distributed access methods for similarity search in metric spaces
  - Adoption of similarity search techniques in digital libraries by way of XML encoded metadata
- Future work

Integrate existing metric indexing with VA-file like quantisation approaches

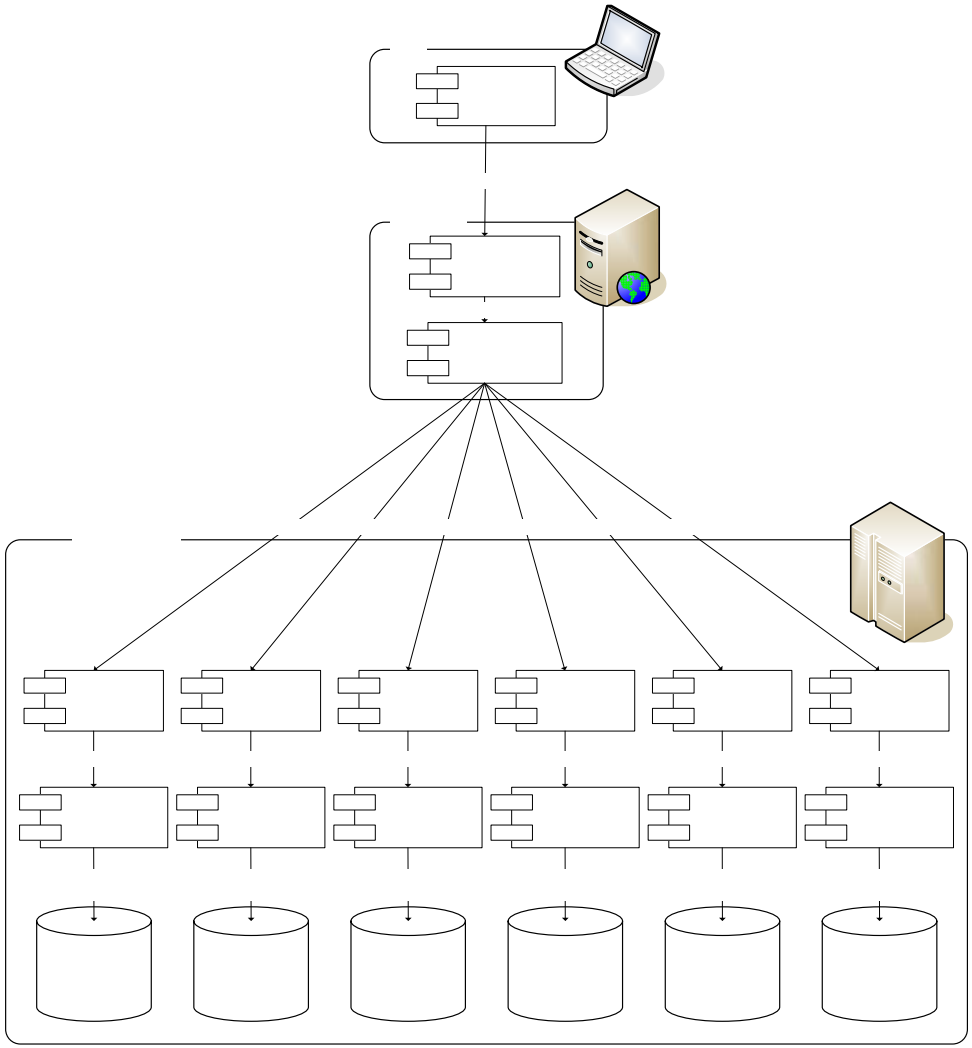
- Generic metric access method (MAM) for arbitrary distance measures and data domains
  - Efficient kNN query processing including cost-balanced ranking support
  - Straightforward, scalable intra-query parallelism
  - Constrained main memory consumption and modest overall resource consumption
- Prototype implementation (CBIR, sequence matching, others)
- Formal foundations (correctness proof, analytic cost models etc.)

UMIT
Sören Balko, Hans-J. Schek

Prototype (1)

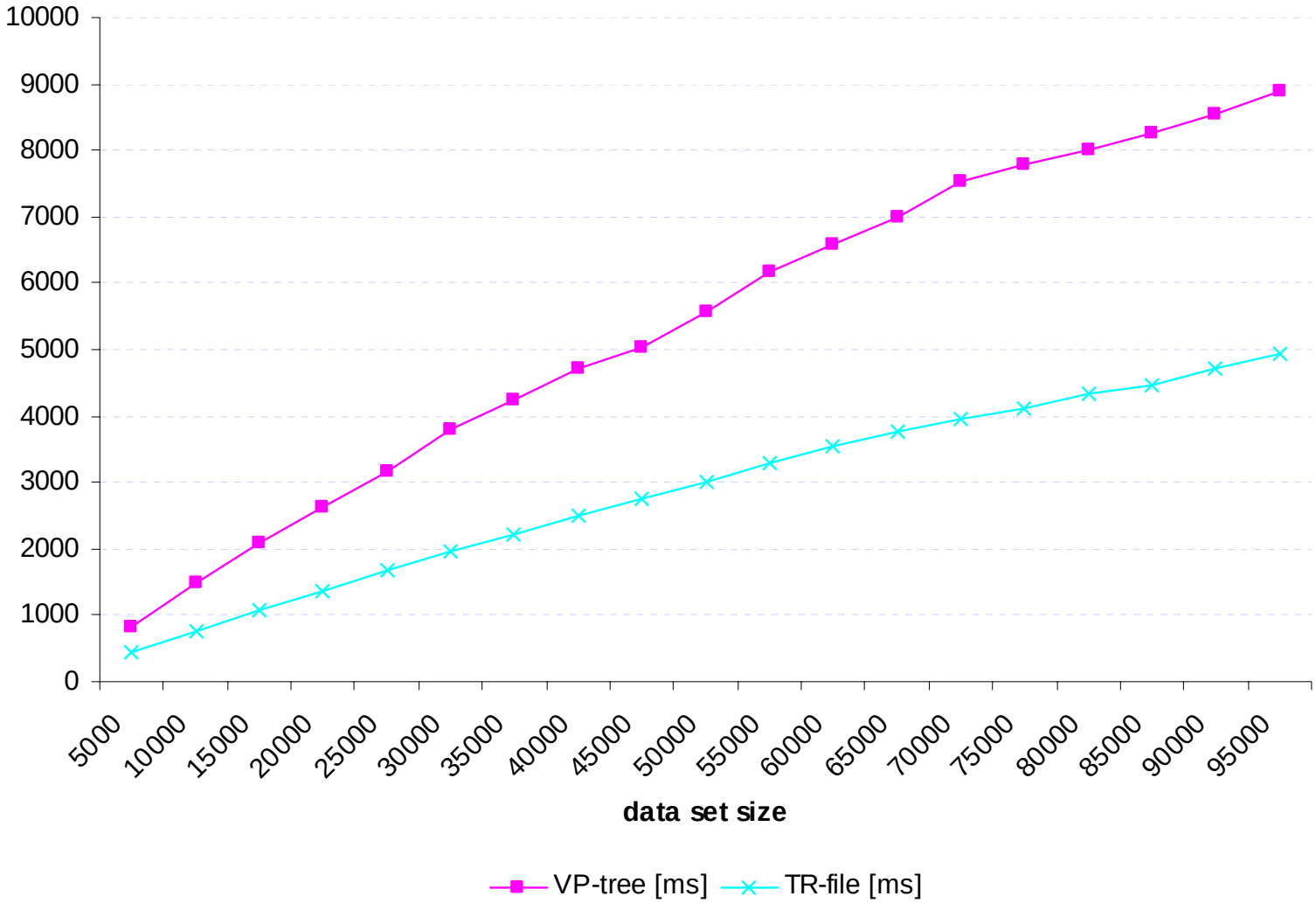
- Content-based image retrieval prototype
  - Index-supported similarity search in roughly 100.000 images
  - Earth Mover's Distance atop 5-region color signatures (of 9 or 12 most frequent colors)
  - Intra-query-parallelism with 6 worker nodes processing disjoint index chunks

Prototype (2)



UMIT
Sören Balko, Hans-J. Schek

Experimentation



Overview of the presentation

- Similarity search and digital libraries: basics
- Work carried out in JPA2
  - Integrate existing metric indexing with VA-file like quantisation approaches
  - Design of distributed access methods for similarity search in metric spaces (MUNI+CNR)
  - Adoption of similarity search techniques in digital libraries by way of XML encoded metadata
- Future work

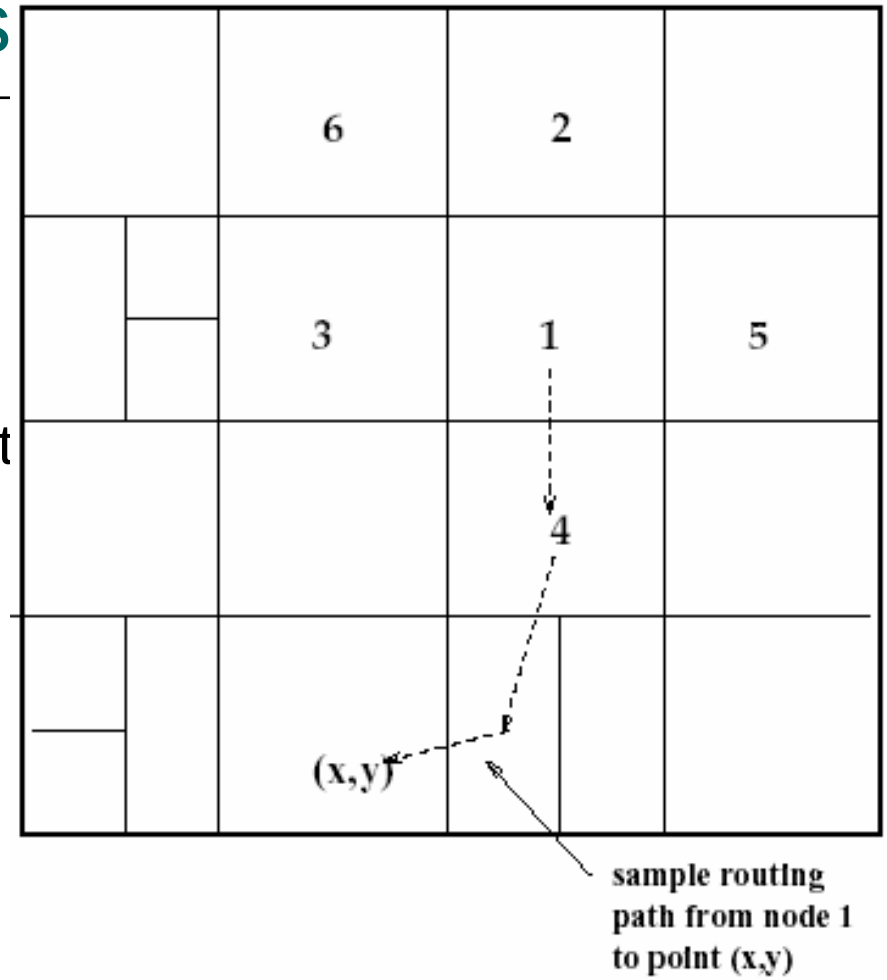
Distributed index for similarity search in metric spaces: MCAN

- The basic idea is:
  - To extend CAN (Content Addressable Networks) to support storage and retrieval of metric space objects enabling CAN to perform similarity searching in the metric spaces -> MCAN
- CAN:
  - CAN is a distributed hash-based data structure
  - using an hash function, every object is mapped in a N-dimensional space
  - the space is divided in zones (chunks of the entire hash table)
  - each physical machine corresponds to one zone
  - each CAN node holds information about adjacent zones

MUNI and ISTI-CNR
Pavel Zezula, Giuseppe Amato

CAN: Data acces

- CAN uses greedy routing
  (i.e. jumps to the neighbor zone nearest to the target point)

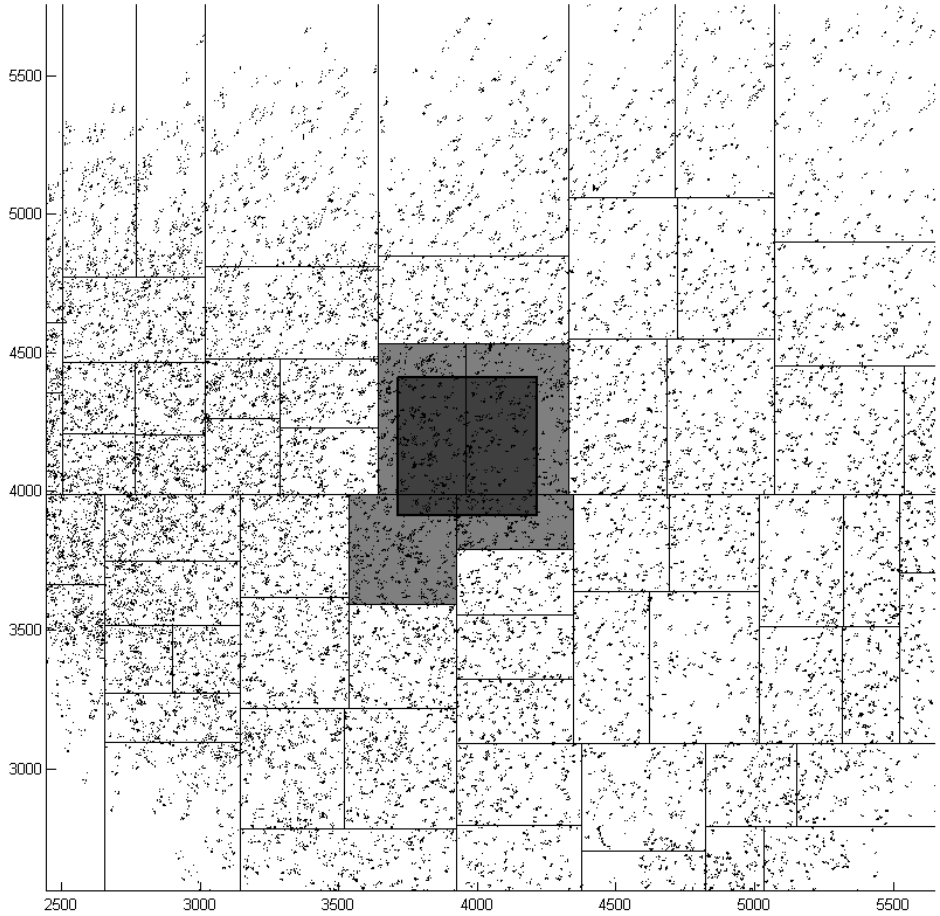


MCAN (ISTI-CNR and MUNI)

- Defining special hash function such that similar objects are placed close one to each other
  - Close means on the same node on in close nodes
  - Distances on MCAN representation are shorter than in the original metric space
- Similarity search is performed in the MCAN representation
  - Filtering-out remaining non qualifying objects

MCAN: Range Query

- The query has been mapped into the vector CAN space
- Because the map is contractive the candidate objects have distance from the query $\leq r$
- We access only the zones which own the candidate objects

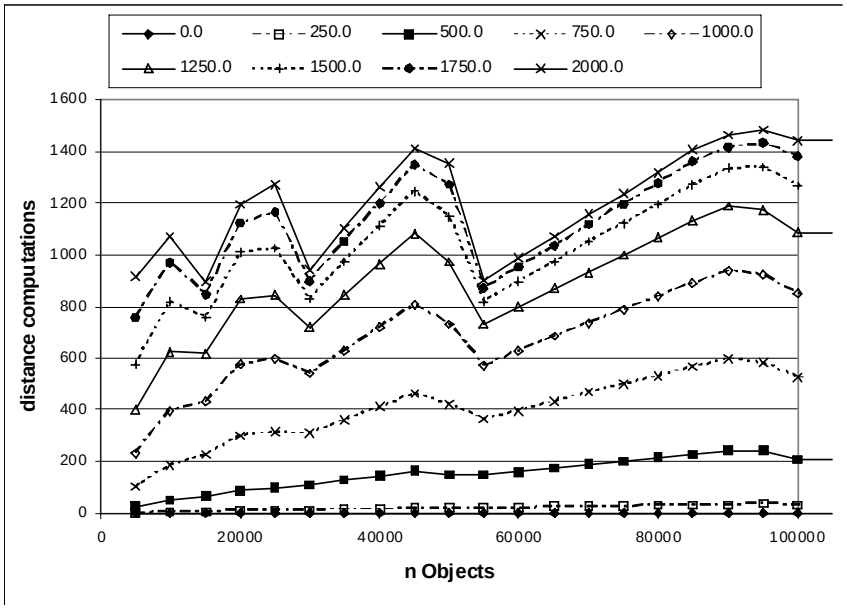


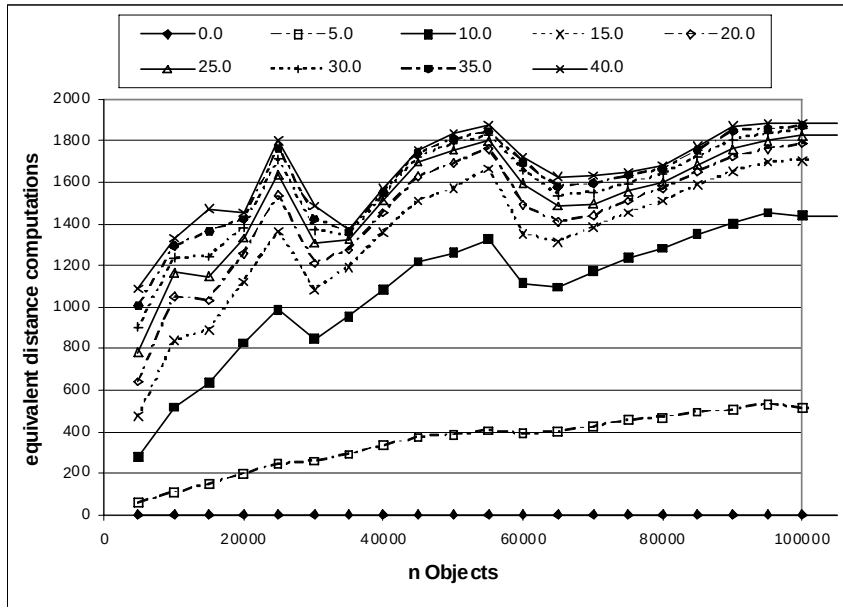
MUNI and ISTI-CNR
Pavel Zezula, Giuseppe Amato

MUNI and ISTI-CNR
Pavel Zezula, Giuseppe Amato

MCA: RQ Experimental Results

- Keeping the number of objects per nodes limited we can limit the cost of a Range Query
- The average parallel cost is strictly related to the range





Overview of the presentation

- Similarity search and digital libraries: basics
- Work carried out in JPA2
  - Integrate existing metric indexing with VA-file like quantisation approaches
  - Design of distributed access methods for similarity search in metric spaces
  - Adoption of similarity search techniques in digital libraries by way of XML encoded metadata (ISTI-CNR)
- Future work

Similarity search in digital libraries with XML encoded metadata

- XML is a good option for encoding metadata
  - Easily readable by humans
  - Easily manageable by computers
- Native XML databases can be used to deal with XML encoded metadata
  - Special techniques have to be used to mix traditional database searches and full text searches
- Emerging trends:
  - Using XML to also to encode low level features extracted from multimedia documents
    - E.g MPEG7/21

ISTI-CNR
Giuseppe Amato

Complex similarity search on XML data

- The objective here is to develop techniques to
  - Allow complex queries that mix
    - Traditional search capabilities
    - Full text search capabilities
    - Similarity based search capabilities

  on XML encoded metadata
- Example:
  - Search for videos related to Iraqi war, taken on July 2005, containing a shot showing a scene similar to a given picture

ISTI-CNR
Giuseppe Amato

The XMLSE (XML Search Engine)

- XMLSE is able to deal with any arbitrary XML file.
  - Special indexes are used to index
    - Textual elements
    - Visual descriptor elements
  - XQuery syntax was extended to deal with similarity operators: XQuery+
  - Special techniques to obtain an effiecient XQuery+ query processor able to deal with
    - semi structured data
    - path expressions containing wildcards
    - approximate/similarity match

ISTI-CNR
Giuseppe Amato

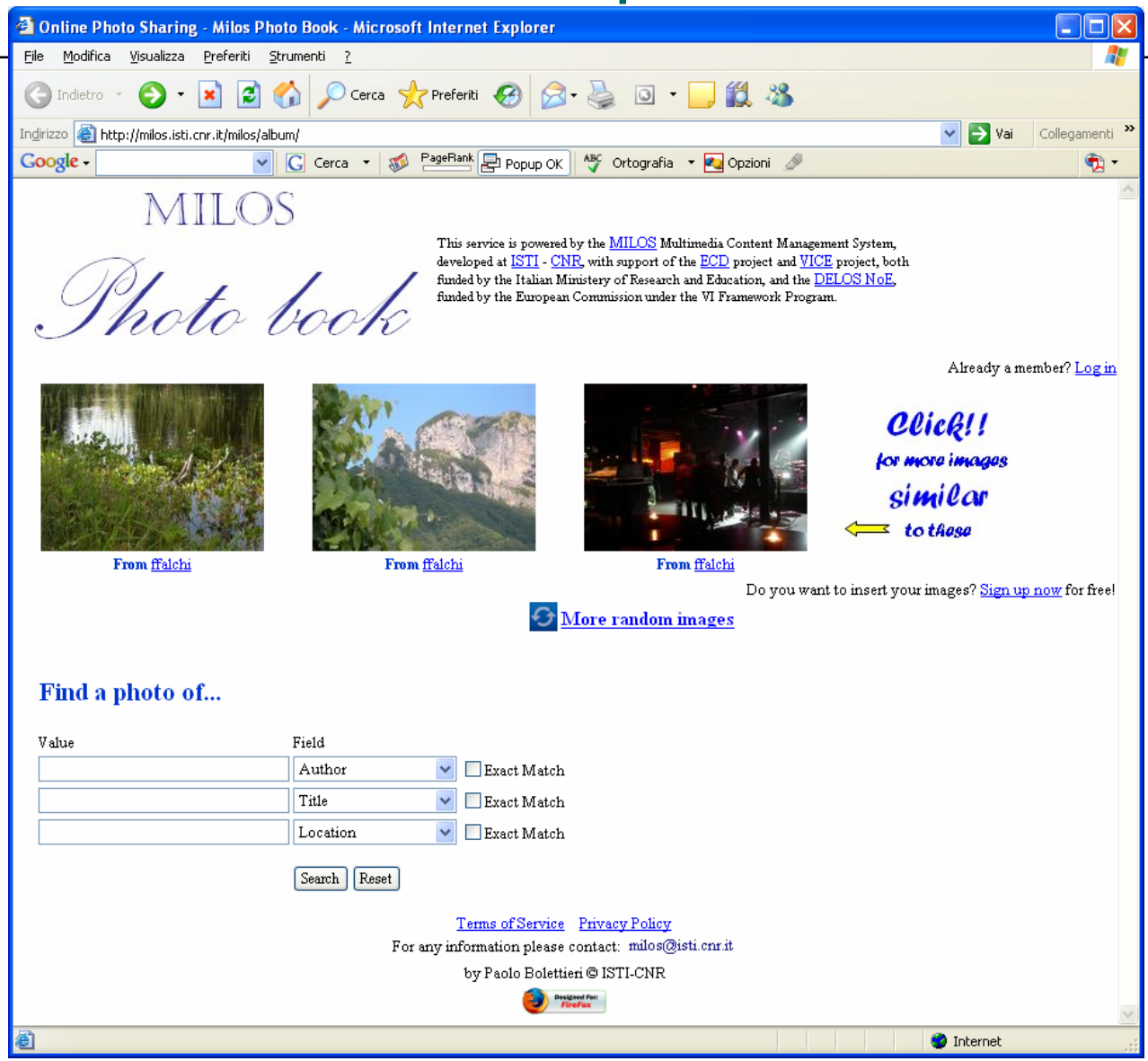
Query example

- Search for images taken by John Smith related to his holydays in Prague, similar to a given picture (e.g. showing Charles Bridge)

```
for  $a in /Mpeg7,  $b in /Mpeg7
where
    $a//MediaUri='D:\ANSAnumb\104.jpg' and
    $b//Author=“John Smith” and
    $b//FreeTextAnnotation ~ “holydays in Prague” and
    $ a//VisualDescriptor ~  $b//VisualDescriptor
return  $b
```

ISTI-CNR
Giuseppe Amato

An application example using XMLSE
On-line demos at http://milos.isti.cnr.it



ISTI-CNR
Giuseppe Amato

Overview of the presentation

- Similarity search and digital libraries: basics
- Work carried out in JPA2
  - Integrate existing metric indexing with VA-file like quantisation approaches
  - Design of distributed access methods for similarity search in metric spaces
  - Adoption of similarity search techniques in digital libraries by way of XML encoded metadata
- Future work

Task 2.6: Objectives (JPA3)

- General Objective
  - Develop and enhance state of the art techniques for efficient similarity search
- Specific Objectives (JPA3=JPA2)
  - develop index structures that
    - efficiently support nearest neighbour, range and ranking queries
    - operate on any kind of multimedia data
    - are generic in the metric distance measure to be employed
- Operational Objectives (JPA3)
  - Enhancement of techniques for distributed access methods
  - Investigation of new directions for building index structures specifically addressed to efficient/effective image searching
  - Investigate techniques that are also tuneable in addition to scalable
  - Dissemination of knowledge

Similarity search: The metric space approach

- For more information, see our book, written thanks to the DELOS support:

"Similarity search: the metric space approach",
by Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, Michal Batko

published by Springer