Towards Web-scale Content Search: the SAPIR Approach

Fausto Rabitti
Pavel Zezula
SAPIR Goals

Develop cutting-edge technology to index and search large scale audio-visual information by content

Support Web2.0 MM content production: personal producer VS professional producers
• The datasets:
  – **CoPhIR:**
    *Content-based Photo Image Retrieval Test-Collection*
  – **BBC Videos**

• State of the art in searching:
  – **MUFIN** for similarity search in P2P
  – **MINERVA** for text search in P2P
  – Optimized Threshold *Algorithms* for merging results
  – **Metric-Cache** for improving efficiency and efficacy of similarity search results
• Today **scalability** issues already **put brake** on growth of multimedia search engines
• The amount of row data is still **growing exponentially**
• Content enrichment techniques produce more and more **heavy features**
• The quality of multimedia search would greatly benefit from solving **scalability issues**
Chorus Gap Analysis: Scalability Challenges

• Breakthroughs are urgent

• Scalability considerations must be taken into account at all stages of:
  – **indexing** – content enrichment, and
  – **retrieval** – query evaluation
Effectiveness Improvement via Scalability

Query image

Search in 1M Flickr images

Search in 10M Flickr images

Search in 50M Flickr images
Content-based Photo Image Retrieval

100M images + metadata + MPEG-7 VDs

http://cophir.isti.cnr.it/

- largest publicly available collection of high-quality images metadata: **106 Million images**.
- Each contains five MPEG-7 VDs:
  - Scalable Color, Color Structure, Color Layout, Edge Histogram, Homogeneous Texture.
- and other textual information:
  - title, tags, comments, etc.
- Photos have been crawled from the Flickr photo-sharing site.
Content Searching Architecture

Extensibility
- metric space
  - Edit distance
  - Jaccard’s coef.
  - Hausdforff distance
  - Minkowski distance
  - Mahalanobis distance
  - etc.

Scalability
- P2P structure

Search infrastructure
- Minkowski distance
- Mahalanobis distance
- etc.

Cloud computing infrastructure as a service

http://mufin.fi.muni.cz/
Image Search Demo

Extensibility
- COPHIR
  - color structure
  - scalable color
  - color layout
  - edge histogram
  - homogeneous texture

Scalability
- M-Chord + M-Tree

MUFIN
SEARCH
ENGINE

6 x IBM server x3400

http://mufin.fi.muni.cz/imgsearch/
<table>
<thead>
<tr>
<th>Technique</th>
<th>CPUs</th>
<th>100k</th>
<th>1M</th>
<th>10M</th>
<th>50M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential scan</td>
<td>1</td>
<td>4.3s</td>
<td>43.4s</td>
<td>7.2min</td>
<td>36min</td>
</tr>
<tr>
<td>M-Tree</td>
<td>1</td>
<td>1.4s</td>
<td>12s</td>
<td>1.8min</td>
<td>-</td>
</tr>
<tr>
<td>Parallel sequential scan</td>
<td>16</td>
<td>0.4s</td>
<td>2.7s</td>
<td>27s</td>
<td>2.3min</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.3s</td>
<td>0.5s</td>
<td>5.4s</td>
<td>27s</td>
</tr>
<tr>
<td>M-Chord</td>
<td>16</td>
<td>0.29s</td>
<td>0.45s</td>
<td>1.7s</td>
<td>5.9s</td>
</tr>
<tr>
<td>M-Chord with approximation</td>
<td>16</td>
<td>0.31s</td>
<td>0.38s</td>
<td>0.44s</td>
<td>2.6s</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.3s</td>
<td>0.36s</td>
<td>0.43s</td>
<td>0.45s</td>
</tr>
<tr>
<td>M-Chord with approximation and disk</td>
<td>32</td>
<td>0.75s</td>
<td>0.82s</td>
<td>0.87s</td>
<td>1.2s</td>
</tr>
</tbody>
</table>
How SAPIR achieve scalability

- 10M network, 500 peers, memory-based
- Batch of 250 queries started from 10 peers

![Graph showing execution time and throughput vs number of CPUs]
Demo Invitation:
SAPIR Image and Video similarity search
Demo Invitation:
SAPIR Image GPS search (e.g. Matera)
Demo Invitation:
SAPIR combined search (Videos)
Demo Invitation:
SAPIR Video
Demo Invitation:
MUFIN image similarity search