SAPIR
Search in Audio-Visual

Diligent

SAPIR – Search in Audio Visual Content

A DIgital Library Infrastructure on Grid ENabled Technology

The SAPIR Project 1/2

- The searchable space created by the massive amounts of existing video and multimedia files greatly exceeds the area searched by today's major engines.
- Traditional search engines are limited to searching in the associated text and meta-data of the multimedia content. If content providers don't clearly or accurately describe their multimedia files, or use inaccurate tags, the current method falls short.
- SAPIR goal is to provide searches over *huge* quantities of multimedia objects, using both text and multimedia features and exploiting the similatiry and search-by-examples queries
- Centralized search engines prove not to be efficient nor scalable for this task

The SAPIR Project 2/2

- SAPIR is based on a scalable, completely decentralized, largely self-organizing P2P system where peers act both as client and servers and the users produce audio-visual content using multiple devices

- The project aims at proving in practice the theoretical advantages shown by multimedia P2P systems like MCAN and MChord

- This technology can provide a significant advantage to the European community over existing, centralized, text-only search engines and can be applied to various fields such as government services, tourism, healthcare, and more.

Challenge

- In order make significative tests with the proposed infrastructure, the project needs a huge amount of data
- Starting from June, we started extracting metadata from the Flickr archive (www.flickr.com)
- The data collected so far represents the **world's largest multimedia metadata collection available**
  - Target collection: **100 million** of images
  - Already processed: **40 million** of images
  - For each image we keep text data and 3 MPEG-7 features → **160 millions of processed features**.

Data Challenge: Procedures

The Data Challege is being conducted as follows

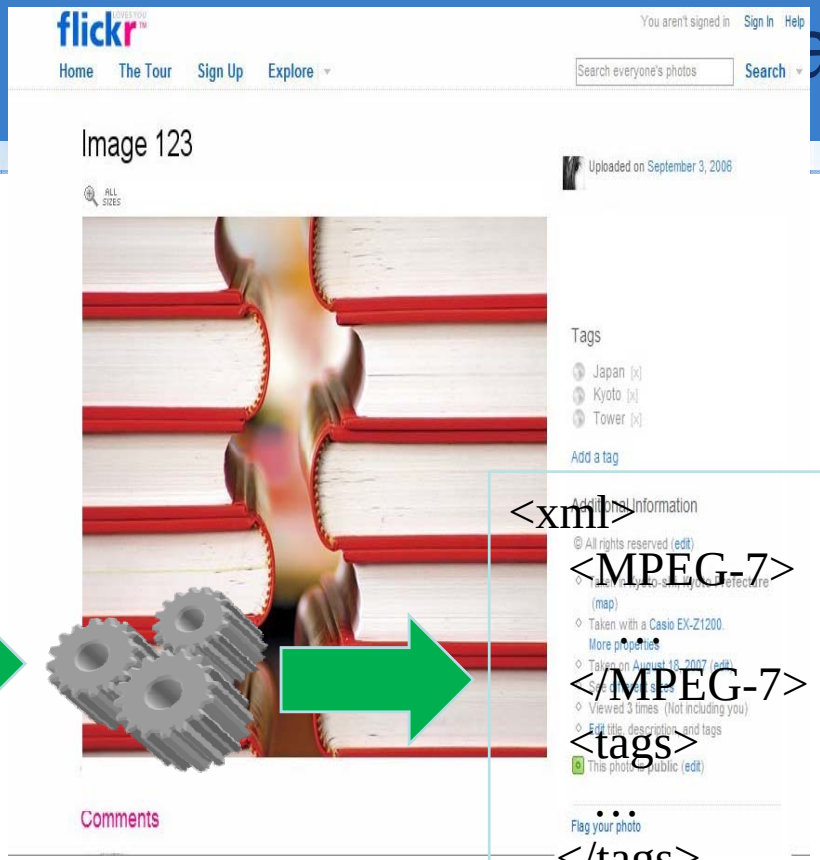
- Each job consists in a pilot job. When it starts, it
  - Download from a SAPIR server and install a package that contains
    - The configuration files
    - The sw needed to run the application
    - The SAPIR application
  - Starts the SAPIR application
- The SAPIR application
  - download from the SAPIR server the ID range of the photos to download
  - Starts a series of threads
  - Each thread
    - Download an image form the ID range
    - Extract the text and MPEG-7 features
    - Upload the features on the SAPIR database

Pilot Jo

SAPIR

IDs

CE



<xml>
<MPEG-7>
...
</MPEG-7>
<tags>
...
</tags>
</xml>

SAPIR

Data Challenge : images

- Preparation and porting of the Sapir application to the grid
  - Starting date 16/06
  - Computing resources: non-grid nodes and 3 PPS-Sites managed by DILIGENT
    - PPS-SNS, PPS-CNR, PPS-ESRIN
  - Total number of images processed 5 M

- Data Challenge
  - Starting date 16/07
  - Total number of images processed till the 26/09 > 36 M

Data Challenge : jobs

- Number of jobs submitted :
  - Preparation phase 250 jobs a day
  - Data Challenge
    - From 16/07 to 29/07 500 jobs a day
    - From 20/07 till now 1000 jobs a day
- Variable number of images processed per jobs
  - Preparation phase 250 images per job
  - Data Challenge 1000 images per job

Data Challenge: performance

	Jobs		Images	
Phase	Submitted	Processed	Expected	Processed
Preparation	7500	5200	1.875 M	1.3 M
DC 1st part	7500	5000	3.75 M	2.5 M
DC 2nd part	37000	28000	37 M	28 M

- Total number of jobs: 52 k
- Total number of jobs processed successfully: 38.2 k
- Total number of jobs failed: 13.8 k
- Failure rate: 26%
- Total number of images expected: 42.62 M
- Total number of images processed: 31.8 M (4.2 TB)