Bases de données avancées

Bases de données multimédia

Université de Cergy-Pontoise
Master Informatique M1
Cours BDA
Multimedia DB

- Multimedia data include images, text, video, sound, spatial data
- Data of large volume; their use is widespread
  - Photos, personal videos
  - Digital television, online music
  - Web pages (text, images, video, sound)
  - Maps, directions, GPS
- "non-traditional" data requiring special management incompatible with conventional DB
  - Very weak structure (image, sound, video, text) or very specific (spatial)
  - Specific methods of access / modification
- The special case of image DB
Images

• Unstructured content
  – Matrix of dots (pixels)
  – Each point color information

• Further comments
  – Textual, structured, semantic ...

• Search images in a collection
  – needs
    • Detect objects in images
    • Summarizing a collection of images
    • Search images on various criteria (similarity, objects)
  – methods
    • Based on the annotations
    • Based on the content
Methods of image search

• By annotation
  – First method developed, using the text
  – Problem: how to annotate an image?
  – disadvantages
    • Ambiguity inherent subjectivity
    • Depends on the language and application
    • Expensive manual annotation

• The visual content
  – Basic idea: extract visual features (descriptors) and use the similarity of these characteristics?
  – Advantages: extraction program, language-independent, less choice problems and dependence on the application
  – Problem: the semantic gap = gap between the intention of the user and the search criteria used for annotation
Applications of content search

• Medical images: detection of pathologies
• Satellite Images: detection of object, meteorological phenomena
• Audiovisual: find a film plane or a character, copy detection
• Investigation: Fingerprint research, identification of a face
• Culture: seeking information about a work of art, e.g. monument from a photo
• …
Types of content-based search

• Search by Example: The mostly used
  – It gives a picture and seeks similar images in descending order of similarity
  – Variant: given an area of an image and find images containing similar areas
  – Generally it is satisfied with the k best answers (top-k)

• Search with relevance feedback
  – Extension that tries to reduce the semantic gap
  – Among the best answers, the user marks those that are not relevant
  – The system changes the similarity function to "stick" to the indication
  – several iterations
Image Descriptors

• Feature descriptor extracted from the image
  – Extraction method
  – Similarity measure
  Ex. color histogram

• Signature = vector representing a descriptor
  – Ex. components of the color histogram

• Images’ space description
  – Multidimensional space where each signature is a point
  – Number of dimensions = number of components comprising the signature vector
  – Image similarity = signatures in the feature space are near
Types of descriptors

• Global descriptor
  – Approximate description of the entire image
  – Describes an atmosphere, an overall impression
    Ex. color, texture, shape descriptor
  – Produces a signature (point in the feature space) per picture
  – Large Feature space

• Local descriptor
  – Accurate description of parts of the image
  – Suitable for objects (parts of images)
    Ex descriptor areas, points of interest
  – Produced several signatures picture
  – Feature space of smaller size
Example: color histogram

- Space representation of color (eg RGB)
  - Each color represented by the intensity values of the components
  - Objective: two colors close to the eye have close representations

- Sampling colors in the image
  - For eg RGB 24 bits (8 bits = 256 levels per component)
  - Sampling of 6 values component $\rightarrow 6^3=216$ colors
  - Choice values: eg levels 0, $256/6$, $2 \times 256/6$, ..., $5 \times 256/6$
  - Result: Histogram of 216 colors = 216 vector elements

- Example of calculating the color histogram
  - $C$ for each color among the 216 colors of the histogram as the percentage of pixels of the image with a color "near" $c$
  - Histogram vector = 216 values between 0 and 100
Indexing of Images

• Index = data structure that allows to accelerate search

• General Idea: organize points in space description to quickly find the nearest neighbors of a given point

• Same type of problem in conventional DB
  – Given a value, find the objects in the database that match this value
  – The index allows not to browse the entire database

• ... But it needs special index structures
  – multidimensional space
  – Similarity search, no exact match
Architecture of an image storing/querying system

- Image Database
- Image
- Signature
- Description space
- Index
- Indexing
- Search
- Image Query
- Signature
The curse of dimensionality

• For a point in the feature space, the index should allow to quickly find neighboring images… but in a multidimensional space there are problems!

• Main problem: the points tend to be equidistant
  – When we seek the neighborhood of a point is recovered either nothing or almost all BD!

  \(\rightarrow\) index no longer useless, as do a sequential scan
  – The problem increases with increasing number of dimensions

• Solutions
  – Reduce the number of dimensions of local descriptors, project spaces of smaller size (eg find separately on groups of dimensions of the color histogram and merge results)
  – Approximate search by hash (LSH)
  – Classification (interactive) points for a given query
Index Types

• General Idea: group of close points in bundles, packages super-packets, etc.
  – A hierarchy of packages we obtain = tree
  – neighboring points $\rightarrow$ in the same package, super-package, etc..

• Most of the time:
  – Space Division
    • Division of the feature space into zones, areas into sub-areas, etc.
    • Product division of the data $\rightarrow$ paquet = packet zone points
  – Data Division
    • Dividing the data into packets, packets into sub-packets, etc.
    • A subdivision of space $\rightarrow$ zone = region (bounding sphere) package containing the points
Example Index: KD-tree

- Division of space according to each dimension
  - At each step a size and pivot value is chosen
  - The two sub-areas continue recursively for each
  - It stops when the number of points of an area is under a given threshold

Example: points in a 2D space
Search with KD-tree

- Simple case points in a single zone
  - Depth-first traversal from the root
  - The coordinates of the motion guide the course
Search with KD-tree (continued)

• More complex cases: points in several areas
  – The depth-first traversal can follow several paths
  – Can possibly give priority to the most interesting areas
• Multimedia in the relational database
  – Consideration from SQL3 and object-relational model
  – General Idea: introduction of types of multimedia objects with properties / methods specific

• Standard SQL / MM (2001), several types multimedia
  – FullText: text (words, phrases, stemming, ...)
  – Spatial: spatial and geometric data
    • Types: ST_Geometry, ST_Point, ST_Curve, ST_MultiPolygon ...
  – Still Image: Image
    • Types: SI_StillImage, SI_ColorHistogram, SI_Texture, ...
Multimédia in Oracle

• Extension *interMedia*, named *Oracle Multimedia* found in version 11g

• Data Types
  – Adaptation of SQL type / MM in Oracle
    • Four types: ORDAudio, ORDDoc, ORDImage, ORDVideo
    • Type ORDDicom dedicated to medical imaging (DICOM)
    • Type ORDSource for storing multimedia content
  – ORDImage = content (ORDSource) + image properties
    • Properties: size, content size, format ...
    • Metadata stored in the content, standard formats (XMP, EXIF) based on XML
  – Oracle also supports type SQL / MM Still Image

• Key Features
  – Storage: in a BLOB or outside of DB
  – Modification of content and properties
  – Signature extraction and metadata from the content
  – Research on properties, metadata and content
Image search in Oracle

• approaches considered
  – By the attributes that accompany the image in the database tables
  – By the properties of the image (size, format, ...)
  – By XML metadata
  – The content from the signature image

• Search by content
  – Signature images: **ORDImageSignature**
    • Contains descriptions of color, texture and shape as BLOBs
    • Methods to extract the signature of an image, measure the similarity of two signatures
  – Special index for signatures **ORDImageIndex**

• Problem: image descriptors imposed
Sources

• *Cours BD multimédia*, Valérie Gouet, CNAM Paris

• *Local-Descriptor Matching for Image Identification Systems*, Eduardo Valle, thèse de doctorat, 2008

• *SQL Multimedia and Application Packages (SQL/MM)*, Jim Melton, Andrew Eisenberg, 2001

• Oracle Multimedia User's Guide / Reference 11g

• Oracle interMedia Reference 10g