

Searching non-text information objects

1

Non-text digital objects

- Music
- Speech
- Images
- 3D models
- Video
- ?

2

Ways to query for something

1. Query by category/ theme
 - easiest - work done ahead of time
 2. Query by describing content
 - text-based query
 - text-based retrieval?
 3. Query by example
 - "similar to"
 - imprecise example - sketch
- query text docs and non-text objects with 2
 - don't often do doc search by 3
 - big move to do music, images by 3

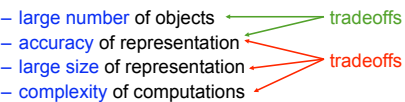
3

Query by describing content

- text-based queries
- where get text-based content?
 - author labels
 - metadata
 - URLs
 - text near imbedded objects
 - html pages
 - group tagging
 - folksonomy
 - Flickr

4

Query by example

- How represent objects?
 - features of a class of objects (e.g. image)
 - how compare features?
 - what data structures?
 - what computational methods?
 - Issues
 - large number of objects
 - accuracy of representation
 - large size of representation
 - complexity of computations
- 
- The diagram shows four items in a list: 'large number of objects', 'accuracy of representation', 'large size of representation', and 'complexity of computations'. Two green arrows point from 'large number of objects' and 'accuracy of representation' to the word 'tradeoffs'. Two red arrows point from 'large size of representation' and 'complexity of computations' to the word 'tradeoffs'.

5

Features

- typically vector of numbers characterizing object representation
- "similar to" = close in vector space
 - threshold
 - Euclidean distance?
 - other choices for distance metric

6

Example: content- based image search

7

First example method: color histogram

- k colors
- histogram: % pixels each color
- k×k matrix A of color similarity weights
- histogram defines feature vectors
- $\text{dist}_{\text{histo}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}-\mathbf{y})^t A(\mathbf{x}-\mathbf{y})$

$$= \sum_{i=1}^k \sum_{j=1}^k a_{ij} (x_i - y_i)(x_j - y_j)$$

- cross-talk: quadratic terms needed
- not Euclidean distance

8

color histograms: reducing complexity

- compute RED_{avg} , $\text{GREEN}_{\text{avg}}$, BLUE_{avg}
 - over all pixels
- use to construct 3D-vector
- use Euclidean distance
- get close candidates
- examine close candidates with full histogram metric

9

color histograms: observations

- works for certain types of images
 - sunset canonical example
- color histogram global property
- this only small part of work:
 - QBIC system, IBM, 1995

10

Second example method: a region-based representation

- region-based features of images
- query processed in same way as collection
- space-conscious: use bit vectors
- levels of representation:
 - store bit vector for each region
 - store bit vector for each image
- get close candidates: compare image bit vectors
- compare top k candidates using region bit vectors

11

Processing images of collection & query

- segment into homogeneous regions
 - 14 dimensional feature vectors
- threshold and transform
 - high-dimensional bit vectors - store
 - XOR for distance between regions
- build image feature vector
 - n region bit-vectors + weights \Rightarrow 1 image feature vector
 - L_1 distance between feature vectors
- transform image vector
 - one high-dimensional bit vector for image - store

12

Components region feature vector

- color moments - 9 dim
 - role similar to histogram
- bounding box region - 5 dim
 - ln(aspect ratio)
 - ln(bounding box size)
 - density = # pixels / bounding box size
 - centroid x
 - centroid y

weight regions proportional to sq. root of area

13

Observations: region based

- **Example** of one regional method
 - lots of research, lots of places!
- This method uses **sampling** heavily
 - produce bit vectors
- Part of larger project - multiple media
 - CASS, Princeton, 2004

14

Example: Image ranking

- given similarity measures
- use PageRank style
- define

$$\mathbf{v} = \alpha(1/n) + (1-\alpha)\mathbf{S}\mathbf{v}$$

- where
 - n is the number of images to be ranked
 - S is a matrix of image-image similarities
 - column normalized, symmetric
 - \mathbf{v} is the vector of VisualRanks
 - α is the usual parameter

15

Observations: Image rank

- intention to use on images returned by other means
 - e.g. text based
- graph undirected
- tested on Google image search
 - VisualRank, Google, 2008
- Deployed?

16

Image search: Summary of techniques

- Techniques seen
 - aggregate/average features
 - sample
 - coarse screening followed by more accurate
- Goals
 - reduce dimension
 - reduce complexity of distance metric
 - reduce space

17

Image search: Commercial search engines

- Use everything you can afford to use
- Text still king!?

18