Carnegie Mellon Univ.
Dept. of Computer Science
15-415 - Database Applications

C. Faloutsos
Extra Lecture: Multimedia Databases
(R&G 26.6.1, 29.5)
(not in the exam)

Outline

• Motivation / problem definition
  • Main idea / time sequences
    • images
    • sub-pattern matching
    • automatic feature extraction / FastMap

Problem

Given a large collection of (multimedia) records (e.g. stocks)
Allow fast, similarity queries
Applications

• time series: financial, marketing (click-streams!), ECGs, sound;
• images: medicine, digital libraries, education, art
• higher-d signals: scientific db (eg., astrophysics), medicine (MRI scans), entertainment (video)

Sample queries

• find medical cases similar to Smith's
• Find pairs of stocks that move in sync
• Find pairs of documents that are similar (plagiarism?)
• find faces similar to ‘Tiger Woods’

Detailed problem defn.:

Problem:
• given a set of multimedia objects,
• find the ones similar to a desirable query object

• for example:
distance function: by expert
(eg. Euclidean distance)

Types of queries

• whole match vs sub-pattern match
• range query vs nearest neighbors
• all-pairs query

Design goals

• Fast (faster than seq. scan)
• ‘correct’ (ie., no false alarms; no false dismissals)
Outline

• Motivation / problem definition

Main idea / time sequences
• images
• sub-pattern matching
• automatic feature extraction / FastMap

Main idea

• Eg., time sequences, ‘whole matching’, range queries, Euclidean distance

Main idea

• Seq. scanning works - how to do faster?
Idea: ‘GEMINI’

(GEneric Multimedia INdexIng)
Extract a few numerical features, for a ‘quick and dirty’ test

‘GEMINI’ - Pictorially

GEMINI

Solution: Quick-and-dirty filter:
• extract \( n \) features (numbers, eg., avg., etc.)
• map into a point in \( n \)-d feature space
• organize points with off-the-shelf spatial access method (‘SAM’)\)
• discard false alarms
Important: Q: how to guarantee no false dismissals?
A1: preserve distances (but: difficult/impossible)
A2: Lower-bounding lemma: if the mapping ‘makes things look closer’, then there are no false dismissals

• ‘proof’ of lower-bounding lemma

Lower-bounding: Makes objects look closer to each other (& to query object)
GEMINI

• ‘proof’ of lower-bounding lemma

Lower-bounding:
Makes objects
look closer to each
other (& to query
object)
⇒ ONLY false
alarms

Important:
Q: how to extract features?
A: “if I have only one number to describe my
object, what should this be?”

Time sequences
Q: what features?
Time sequences

Q: what features?
A: Fourier coefficients (etc)

Time sequences - results

- keep the first 2-3 Fourier coefficients
- faster than seq. scan
- NO false dismissals (see book)

Time sequences - improvements:

- improvements/variations: [Kanellakis +Goldin], [Mendelzon+Rafiei]
- could use Wavelets, or DCT
- could use segment averages [Yi+2000]
Outline

- Motivation / problem definition
- Main idea / time sequences
- Images
  - sub-pattern matching
  - automatic feature extraction / FastMap

Images - color

what is an image?
A: 2-d array

Color histograms, and distance function
Mathematically, the distance function is:

$$distance_{histogram}(\tilde{x}, \tilde{y}) = (\tilde{x} - \tilde{y})^T \begin{bmatrix} g_{RR} & g_{RP} & \ldots \\ g_{PR} & g_{PP} & \ldots \\ \vdots & \vdots & \ddots \end{bmatrix} (\tilde{x} - \tilde{y})$$

Problem: ‘cross-talk’:
- Features are not orthogonal ->
- SAMs will not work properly

- Q: what to do?
- A: feature-extraction question

Possible answers:
- avg red, avg green, avg blue

It turns out that this lower-bounds the histogram distance ->
- no cross-talk
- SAMs are applicable
Images - color

- Performance: time

- seq scan

- w/ avg RGB

Multimedia - Detailed outline

- multimedia
  - Motivation / problem definition
  - Main idea / time sequences
  - images (color; shape)
  - sub-pattern matching
  - automatic feature extraction / FastMap

Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 'moments'
- (Q: how to normalize them?)
Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 'momsents'
- (Q: how to normalize them?
- A: divide by standard deviation)

Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 'moments'
- (Q: other 'features' / distance functions?

Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 'moments'
- (Q: other 'features' / distance functions?
- A1: turning angle
- A2: dilations/erosions
- A3: ... )
Images - shapes

- Performance: ~10x faster

log(# of I/Os)

# of features kept

all kept

Conclusions

- GEMINI works for any setting (time sequences, images, etc)
- uses a ‘quick and dirty’ filter
- faster than seq. scan
- (but: how to extract features automatically?)

Outline

- Motivation / problem definition
- Main idea / time sequences
- images
- sub-pattern matching
  automatic feature extraction / FastMap
FastMap

Automatic feature extraction:

• Given a dissimilarity function of objects
• Quickly map the objects to a (k-d) 'feature' space.
• (goals: indexing and/or visualization)

FastMap

<table>
<thead>
<tr>
<th></th>
<th>O1</th>
<th>O2</th>
<th>O3</th>
<th>O4</th>
<th>O5</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>O2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>O3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>O4</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>O5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

FastMap

• Multi-dimensional scaling (MDS) can do that, but in $O(N^{**2})$ time
Main idea: projections

We want a linear algorithm: FastMap [SIGMOD95]

FastMap - next iteration
Results

Documents / cosine similarity -> Euclidean distance (how?)

Results

bb reports

recipes

Applications: time sequences

• given $n$ co-evolving time sequences
• visualize them + find rules [ICDE00]
Applications - financial

- currency exchange rates [ICDE00]

Variations

- Isomap [Tenenbaum, de Silva, Langford, 2000]
- LLE (Local Linear Embedding) [Roweis, Saul, 2000]
- MVE (Minimum Volume Embedding) [Shaw & Jebara, 2007]
Conclusions

• GEMINI works for multiple settings
• FastMap can extract ‘features’ automatically (→ indexing, visual d.m.)

Main references

• Myron Flickner, et al: Query by Image and Video Content: the QBIC System
• Journal of Intelligent Inf. Systems, 3, 3/4, pp. 231-262, 1994 (An earlier, more technical version of the IEEE
  Computer ’95 paper.)
• FastMap: C. Faloutsos and K.I. Lin FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of
  Traditional and Multimedia Datasets ACM SIGMOD 95, pp. 163-174.

References

  Proc. ACM SIGMOD, Minneapolis, MN.
References


References


References

- Lawrence Saul & Sam Roweis. An Introduction to Locally Linear Embedding (draft)
References
