15-826: Multimedia Databases and Data Mining

Lecture #25: Multimedia indexing

C. Faloutsos

Outline

Goal: ‘Find similar / interesting things’

• Intro to DB
• Indexing - similarity search
• Data Mining

Must-read Material

• MM Textbook, chapters 7, 8, 9 and 10.
• Journal of Intelligent Inf. Systems, 3, 3/4, pp. 231-262, 1994 (An earlier, more technical version of the IEEE Computer '95 paper.)
• FastMap: Textbook chapter 11; Also in: 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.

Indexing - Detailed outline

• primary key indexing
• secondary key / multi-key indexing
• spatial access methods
• fractals
• text
• Singular Value Decomposition (SVD)
• Multimedia
  – DSP
  – indexing
Multimedia - Detailed outline

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

Problem

Given a large collection of (multimedia) records (eg. 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:

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)
Multimedia - Detailed outline

- multimedia
  - 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

Idea: ‘GEMINI’

(GEneric Multimedia INdexIng)
Extract a few numerical features, for a ‘quick and dirty’ test
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
Lower-bounding: Makes objects look closer to each other (& to query object)

- ‘proof’ of lower-bounding lemma

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 (we’ll see them in detail soon)

Time sequences

- white noise
- brown noise

Fourier spectrum

... in log-log

Time sequences

• Eg.:

- IBM stock
- spectrum (linear scale)
- spectrum (log scale)

Time sequences

• conclusion: colored noises are well approximated by their first few Fourier coefficients
• colored noises appear in nature:
Time sequences

- brown noise: stock prices ($1/f^2$ energy spectrum)
- pink noise: works of art ($1/f$ spectrum)
- black noises: water reservoirs ($1/f^b$, $b>2$)
- (slope: related to ‘Hurst exponent’, for self-similar traffic, like, eg. Ethernet/web [Schroeder], [Leland+])

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]

Multimedia - Detailed outline

- multimedia
  - Motivation / problem definition
  - Main idea / time sequences
  - images (color, shapes)
  - sub-pattern matching
  - automatic feature extraction / FastMap
what is an image?
A: 2-d array

Images - color

Mathematically, the distance function is:

$$\text{distance}_{\text{histogram}}(\vec{\xi}, \vec{\eta}) = \begin{bmatrix} a_{nn} & a_{nr} & \cdots \\ a_{rn} & a_{rr} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} (\vec{\xi} - \vec{\eta})^t$$

Problem: ‘cross-talk’:
- Features are not orthogonal $\Rightarrow$
- SAMs will not work properly

• Q: what to do?
• A: feature-extraction question
Images - color

possible answers:
• avg red, avg green, avg blue

it turns out that this lower-bounds the histogram distance ->
• no cross-talk
• SAMs are applicable

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 ‘moments’
• (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?
• A1: turning angle
• A2: dilations/erosions
• A3: ... )
Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 ‘moments’
- Q: how to do dim. reduction?
- A: Karhunen-Loève (= centered PCA/SVD)

Performance: ~10x faster

![Graph showing performance improvement]

Other shape features?

- Morphology (dilations, erosions, openings, closings) [Korn+, VLDB96]
  - “structuring element”
  - $R=1$

Other shape features
Other shape features

• Morphology (dilations, erosions, openings, closings) [Korn+, VLDB96]

Morphology: closing

• fill in small gaps
• very similar to ‘alpha contours’
Morphology: opening

- ‘closing’, for the complement =
- trim small extremities

Morphology: opening

- ‘closing’, for the complement =
- trim small extremities

Morphology

- Closing: fills in gaps
- Opening: trims extremities
- All wrt a structuring element:

Morphology

- Features: areas of openings (R=1, 2, …) and closings
Morphology

- Powerful method:
- ‘pattern spectrum’ [Maragos+]
- ‘skeletonization’ of images
- ‘Alpha-shapes’ [Edelsbrunner]
- Book: *An introduction to morphological image processing*, by Edward R. Dougherty

Multimedia - Detailed outline

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

Sub-pattern matching

- Problem: find sub-sequences that match the given query pattern

\[ \text{\$} \uparrow \quad \text{\$} \quad \text{\$} \quad \text{\$} \quad \text{\$} \quad \text{\$} \]

Sub-pattern matching

- Q: how to proceed?
- Hint: try to turn it into a ‘whole-matching’ problem (how?)

\[ \text{\$} \uparrow \]
Sub-pattern matching

- Assume that queries have minimum duration $w$; (e.g., $w=7$ days)
- Divide data sequences into windows of width $w$ (overlapping, or not?)

Sub-pattern matching

- Assume that queries have minimum duration $w$; (e.g., $w=7$ days)
- Divide data sequences into windows of width $w$ (overlapping, or not?)
- A: Sliding, overlapping windows. Thus: trails

Pictorially:

sequences -> trails -> MBRs in feature space
Sub-pattern matching

Q: do we store all points? why not?

Sub-pattern matching

Q: how to do range queries of duration \(w\)?

A: R-tree; find qualifying stocks and intervals

Sub-pattern matching

Q: how to do range queries of duration \(w\)?

A: R-tree; find qualifying stocks and intervals
Q: how to do range queries of duration >\(w\) (say, \(2*w\))?

A: Two range queries of radius epsilon and intersect (or two queries of smaller radius and union – see paper)

(improvement [Moon+2001])
- use non-overlapping windows, for data
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?)

Multimedia - Detailed outline

- multimedia
  - Motivation / problem definition
  - Main idea / time sequences
  - images (color; shape)
  - 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

- 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?)

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.)

References

References


References

