


**15-826: Multimedia Databases  
and Data Mining**

*Multimedia indexing*  
C. Faloutsos




**Outline**

Goal: 'Find **similar** / **interesting** things'

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


15-826 Copyright: C. Faloutsos (2006) #2



**Indexing - Detailed outline**

- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
- fractals
- text
- Singular Value Decomposition (SVD)
- ➔ • multimedia
- ...


15-826 Copyright: C. Faloutsos (2006) #3



**Multimedia - Detailed outline**

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

15-826 Copyright: C. Faloutsos (2006) #4




**Problem**

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

Allow fast, similarity queries

15-826 Copyright: C. Faloutsos (2006) #5



**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)

15-826 Copyright: C. Faloutsos (2006) #6

CMU SCS

## 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'

15-826 Copyright: C. Faloutsos (2006) #7

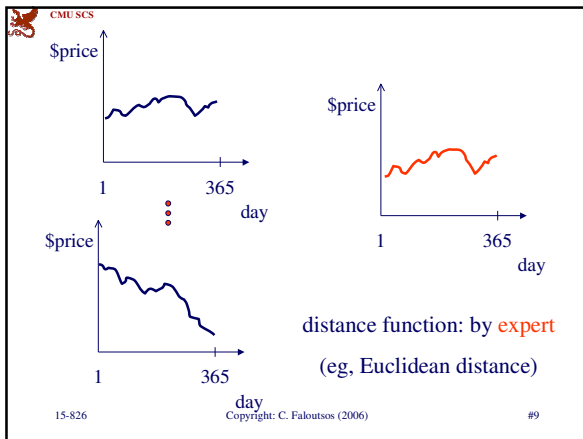
CMU SCS

## Detailed problem defn.:

Problem:

- given a set of multimedia objects,
- find the ones similar to a desirable query object
- for example:

15-826 Copyright: C. Faloutsos (2006) #8



CMU SCS

## Types of queries

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

15-826 Copyright: C. Faloutsos (2006) #10

CMU SCS

## Design goals

- Fast (faster than seq. scan)
- 'correct' (ie., no false alarms; no false dismissals)

15-826 Copyright: C. Faloutsos (2006) #11

CMU SCS

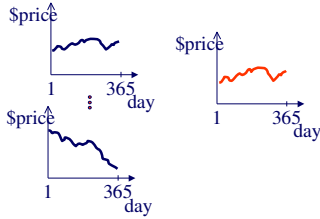
## Multimedia - Detailed outline

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

15-826 Copyright: C. Faloutsos (2006) #12

**Main idea**

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



15-826 Copyright: C. Faloutsos (2006) #13

**Main idea**

- Seq. scanning works - how to do faster?

15-826 Copyright: C. Faloutsos (2006) #14

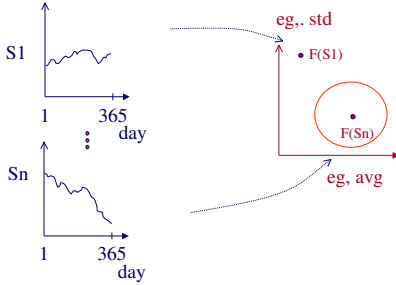
**Idea: ‘GEMINI’**

(GENeric Multimedia INdexIng)

Extract a few numerical features, for a ‘quick and dirty’ test

15-826 Copyright: C. Faloutsos (2006) #15

**‘GEMINI’ - Pictorially**



15-826 Copyright: C. Faloutsos (2006) #16

**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

15-826 Copyright: C. Faloutsos (2006) #17

**GEMINI**

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

15-826 Copyright: C. Faloutsos (2006) #18

CMU SCS

## GEMINI

Important:

Q: how to extract features?

A: *“if I have only one number to describe my object, what should this be?”*

15-826 Copyright: C. Faloutsos (2006) #19

CMU SCS

## Time sequences

Q: what features?

15-826 Copyright: C. Faloutsos (2006) #20

CMU SCS

## Time sequences

Q: what features?

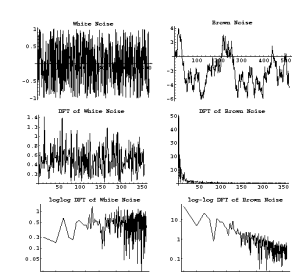
A: Fourier coefficients (we'll see them in detail soon)

15-826 Copyright: C. Faloutsos (2006) #21

CMU SCS

## Time sequences

white noise      brown noise



Fourier spectrum

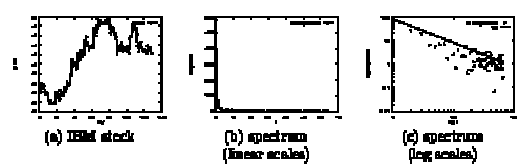
... in log-log

15-826 #22

CMU SCS

## Time sequences

- Eg.:



(a) IBM stock      (b) spectrum (linear scales)      (c) spectrum (log scales)

15-826 Copyright: C. Faloutsos (2006) #23

CMU SCS

## Time sequences

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

15-826 Copyright: C. Faloutsos (2006) #24

**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+])

15-826 Copyright: C. Faloutsos (2006) #25

**Time sequences - results**

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

time

# coeff. kept

← total

← cleanup-time

← r-tree time

15-826 Copyright: C. Faloutsos (2006) #26

**Time sequences - improvements:**

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

15-826 Copyright: C. Faloutsos (2006) #27

**Multimedia - Detailed outline**

- multimedia
  - Motivation / problem definition
  - Main idea / time sequences
  - ➔ – images (color, shapes)
  - sub-pattern matching
  - automatic feature extraction / FastMap

15-826 Copyright: C. Faloutsos (2006) #28

**Images - color**

what is an image?  
A: 2-d array

COLOR IMAGE, eg. 256x256

1-th pixel:  
(r, g, b)

15-826 Copyright: C. Faloutsos (2006) #29

**Images - color**

Color histograms, and distance function

15-826 Copyright: C. Faloutsos (2006) #30

CMU SCS

## Images - color

Mathematically, the distance function is:

$$\text{distance}_{\text{Histogram}}(\vec{x}, \vec{y}) = (\vec{x} - \vec{y})^T \begin{bmatrix} 0.25 & 0.25 & \dots \\ 0.25 & 0.25 & \dots \\ \dots & \dots & \dots \end{bmatrix} (\vec{x} - \vec{y})^0$$

$$\dots = (\vec{x} - \vec{y})^T A (\vec{x} - \vec{y})^0$$

15-826 Copyright: C. Faloutsos (2006) #31

CMU SCS

## Images - color

Problem: 'cross-talk':

- Features are not orthogonal ->
- SAMs will not work properly
- Q: what to do?
- A: feature-extraction question

15-826 Copyright: C. Faloutsos (2006) #32

CMU SCS

## 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

15-826 Copyright: C. Faloutsos (2006) #33

CMU SCS

## Images - color

performance: time

selectivity

15-826 Copyright: C. Faloutsos (2006) #34

CMU SCS

## Multimedia - Detailed outline

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

15-826 Copyright: C. Faloutsos (2006) #35

CMU SCS

## Images - shapes

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

15-826 Copyright: C. Faloutsos (2006) #36

CMU SCS

## Images - shapes

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

15-826 Copyright: C. Faloutsos (2006) #37

CMU SCS

## Images - shapes

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

15-826 Copyright: C. Faloutsos (2006) #38

CMU SCS

## Images - shapes

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

15-826 Copyright: C. Faloutsos (2006) #39

CMU SCS

## Images - shapes

- distance function: Euclidean, on the area, perimeter, and 20 'moments'
- Q: how to do dim. reduction?

15-826 Copyright: C. Faloutsos (2006) #40

CMU SCS

## Images - shapes

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

15-826 Copyright: C. Faloutsos (2006) #41

CMU SCS

## Images - shapes

- Performance: ~10x faster

log(# of I/Os)

# of features kept

← all kept

15-826 Copyright: C. Faloutsos (2006) #42

CMU SCS

## Other shape features


15-826 Copyright: C. Faloutsos (2006) #43


CMU SCS

## Other shape features

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

shape “structuring element”



R=1 


15-826 Copyright: C. Faloutsos (2006) #44




CMU SCS

## Other shape features

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

shape “structuring element”



R=0.5   
 R=1   
 R=2 


15-826 Copyright: C. Faloutsos (2006) #45




CMU SCS

## Other shape features

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

shape “structuring element”



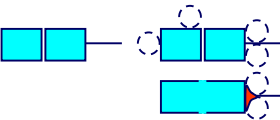
R=0.5   
 R=1   
 R=2 

15-826 Copyright: C. Faloutsos (2006) #46

CMU SCS

## Morphology: closing

- fill in small gaps
- very similar** to ‘alpha contours’




15-826 Copyright: C. Faloutsos (2006) #47

CMU SCS

## Morphology: closing

- fill in small gaps



‘closing’, with R=1

15-826 Copyright: C. Faloutsos (2006) #48



**Morphology: opening**

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

15-826 Copyright: C. Faloutsos (2006) #49

**Morphology: opening**

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

‘opening’ with  $R=1$

15-826 Copyright: C. Faloutsos (2006) #50

**Morphology**

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

15-826 Copyright: C. Faloutsos (2006) #51

**Morphology**

- Features: areas of openings ( $R=1, 2, \dots$ ) and closings

15-826 Copyright: C. Faloutsos (2006) #52

**Multimedia - Detailed outline**

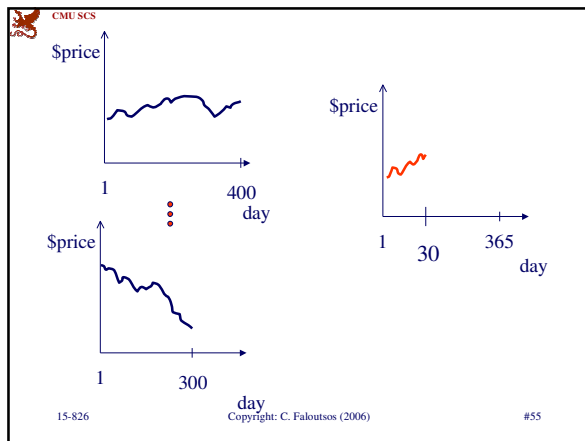
- multimedia
  - Motivation / problem definition
  - Main idea / time sequences
  - images (color; shape)
  - ➡ sub-pattern matching
  - automatic feature extraction / FastMap

15-826 Copyright: C. Faloutsos (2006) #53

**Sub-pattern matching**

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

15-826 Copyright: C. Faloutsos (2006) #54



## Sub-pattern matching

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

15-826 Copyright: C. Faloutsos (2006) #56

## Sub-pattern matching

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

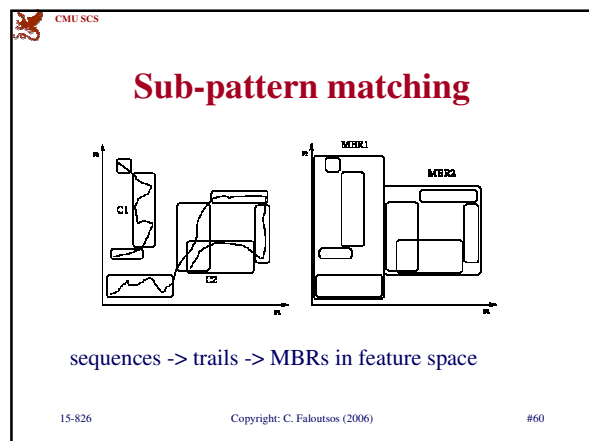
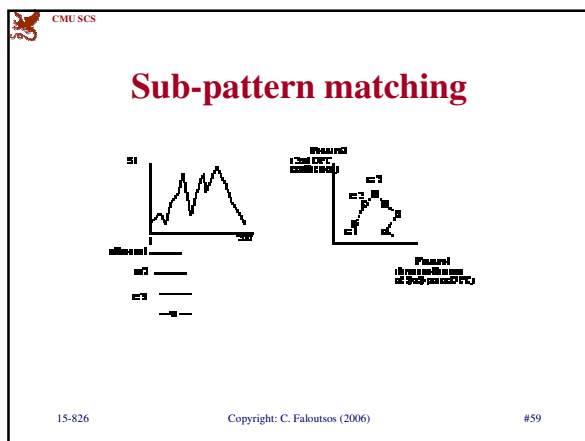
15-826 Copyright: C. Faloutsos (2006) #57

## Sub-pattern matching

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

Pictorially:

15-826 Copyright: C. Faloutsos (2006) #58



CMU SCS

## Sub-pattern matching

Q: do we store all points? why not?

15-826 Copyright: C. Faloutsos (2006) #61

CMU SCS

## Sub-pattern matching

Q: how to do range queries of duration  $w$ ?

15-826 Copyright: C. Faloutsos (2006) #62

CMU SCS

## Sub-pattern matching

(improvement [Moon+2001])

- use non-overlapping windows, for data

15-826 Copyright: C. Faloutsos (2006) #63

CMU SCS

## 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?)

15-826 Copyright: C. Faloutsos (2006) #64

CMU SCS

## Multimedia - Detailed outline

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

15-826 Copyright: C. Faloutsos (2006) #65

CMU SCS

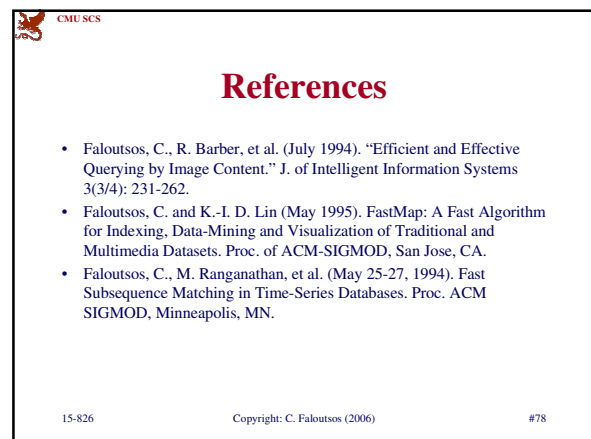
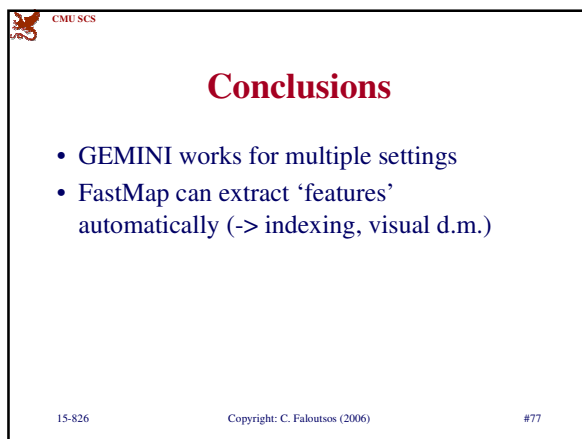
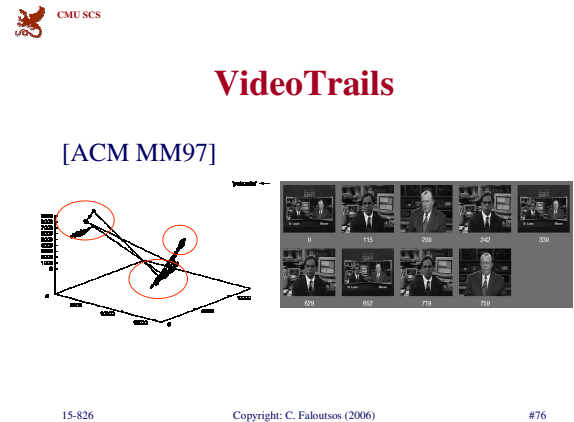
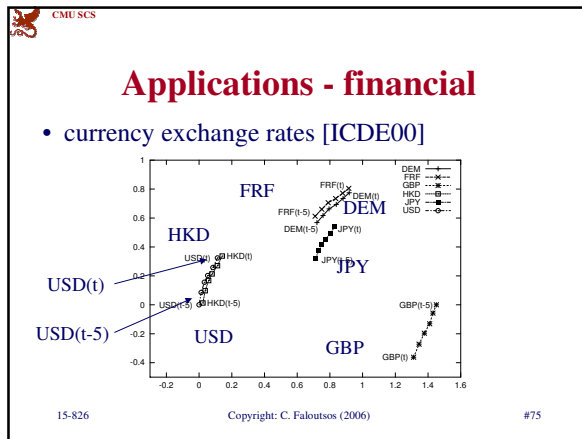
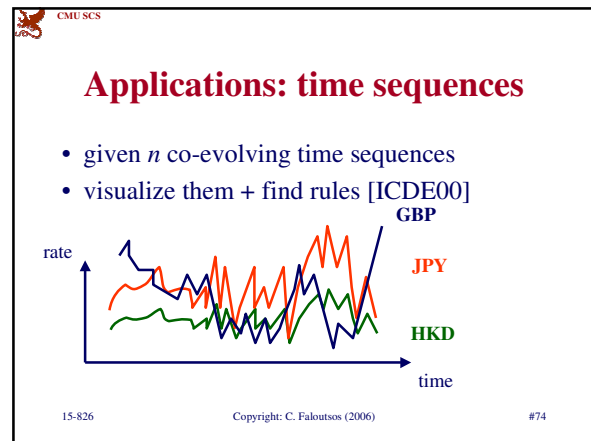
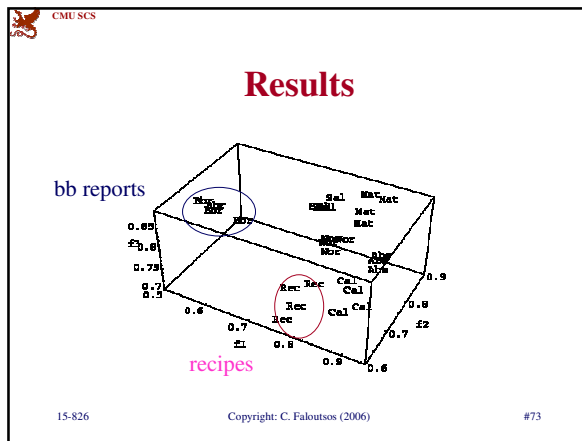
## FastMap


Automatic feature extraction:

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

15-826 Copyright: C. Faloutsos (2006) #66






 CMU SCS
 

## References

- Flickner, M., H. Sawhney, et al. (Sept. 1995). "Query by Image and Video Content: The QBIC System." IEEE Computer 28(9): 23-32.
- Goldin, D. Q. and P. C. Kanellakis (Sept. 19-22, 1995). On Similarity Queries for Time-Series Data: Constraint Specification and Implementation. Int. Conf. on Principles and Practice of Constraint Programming (CP95), Cassis, France.
- Flip Korn, Nikolaos Sidiropoulos, Christos Faloutsos, Eliot Siegel, Zenon Protopapas: *Fast Nearest Neighbor Search in Medical Image Databases*. VLDB 1996: 215-226


15-826 Copyright: C. Faloutsos (2006) #79

 CMU SCS
 

## References

- Leland, W. E., M. S. Taqqu, et al. (Feb. 1994). "On the Self-Similar Nature of Ethernet Traffic." IEEE Transactions on Networking 2(1): 1-15.
- Moon, Y.-S., K.-Y. Whang, et al. (2001). Duality-Based Subsequence Matching in Time-Series Databases. ICDE, Heidelberg, Germany.
- Rafiei, D. and A. O. Mendelzon (1997). Similarity-Based Queries for Time Series Data. SIGMOD Conference, Tucson, AZ.

15-826 Copyright: C. Faloutsos (2006) #80

 CMU SCS
 

## References

- Schroeder, M. (1991). Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise. New York, W.H. Freeman and Company.
- Yi, B.-K. and C. Faloutsos (2000). Fast Time Sequence Indexing for Arbitrary Lp Norms. VLDB, Cairo, Egypt.

15-826 Copyright: C. Faloutsos (2006) #81