Carnegie Mellon

15-826: Multimedia Databases and Data Mining

Lecture #23: Multimedia indexing

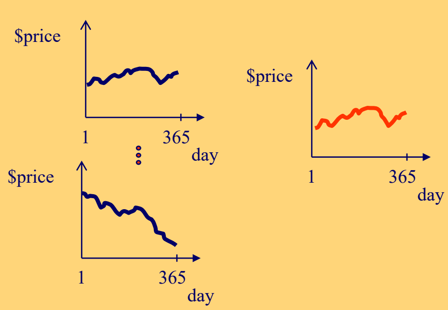
C. Faloutsos

1

Carnegie Mellon

Problem

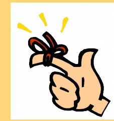
- Q: Find stocks similar to <MSFT>



15-826 Copyright (c) 2019 C. Faloutsos 2

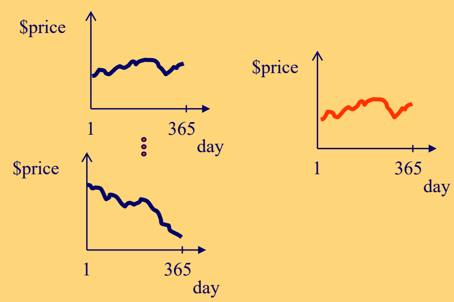
2

Carnegie Mellon



Solution

- Q: Find stocks similar to <MSFT>
- A: GEMINI: Extract features + SAM



3

Carnegie Mellon

Must-read Material

- MM Textbook, chapters 7, 8, 9 and 10.
- Myron Flickner, et al: *Query by Image and Video Content: the QBIC System* IEEE Computer 28, 9, Sep. 1995, pp. 23-32.
- 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.

4

Carnegie Mellon

Outline

Goal: 'Find similar / interesting things'

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

5

Carnegie Mellon

Indexing - Detailed outline

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

6

Carnegie Mellon

Multimedia - Detailed outline

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

15-826 #7

7

Carnegie Mellon

Problem

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

Allow fast, similarity queries

15-826 #8

8

Carnegie Mellon

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)

9

Carnegie Mellon

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'

10

Carnegie Mellon

Detailed problem defn.:

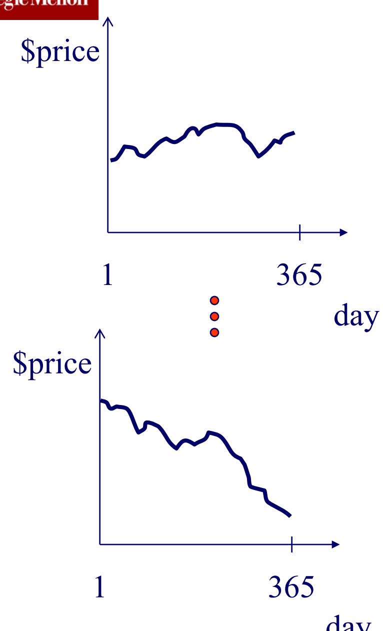
Problem:

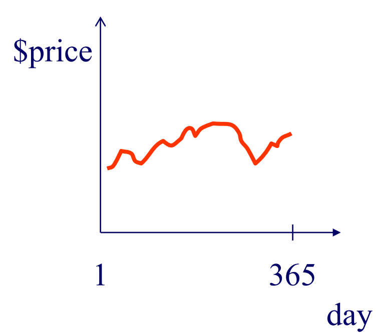
- given a set of multimedia objects,
- find the ones similar to a desirable query object

- for example:

11

Carnegie Mellon





distance function: by expert

(eg, Euclidean distance)

12

Carnegie Mellon

Types of queries

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

15-826 Copyright (c) 2019 C. Faloutsos #13

13

Carnegie Mellon

Design goals

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

15-826 Copyright (c) 2019 C. Faloutsos #14

14

Carnegie Mellon

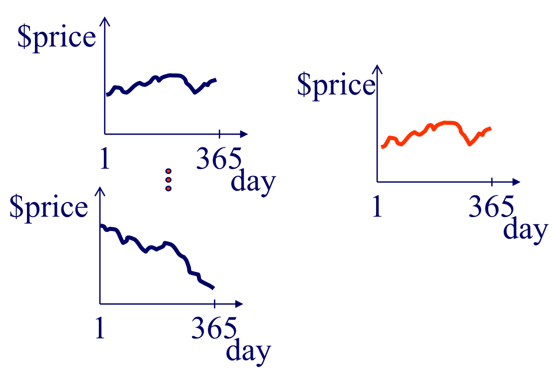
Multimedia - Detailed outline

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

Carnegie Mellon

Main idea

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



Carnegie Mellon

Main idea

- Seq. scanning works - how to do faster?

15-826 #17

17

Carnegie Mellon

Idea: 'GEMINI'

(GEneric Multimedia INdexIng)

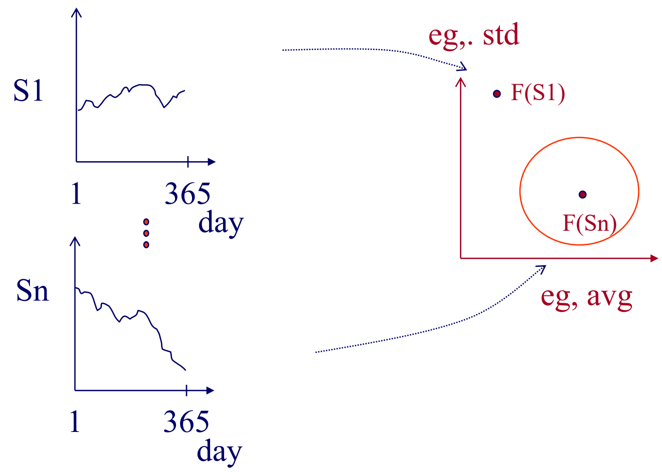
Extract a few numerical features, for a 'quick and dirty' test

15-826 #18

18

Carnegie Mellon

'GEMINI' - Pictorially



15-826 Copyright (c) 2019 C. Faloutsos #19

19

Carnegie Mellon

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) 2019 C. Faloutsos #20

20

Carnegie Mellon

GEMINI

Important: Q: how to guarantee no false dismissals?

A1: preserve distances (but: difficult/impossible)

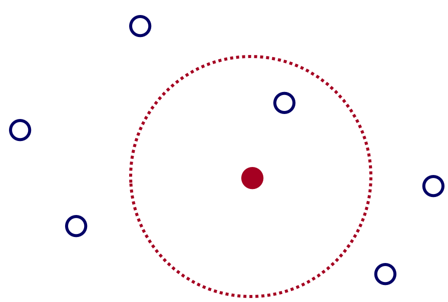
A2: **Lower-bounding lemma**: if the mapping 'makes things look closer', then there are **no** false dismissals

21

Carnegie Mellon

GEMINI

- 'proof' of lower-bounding lemma

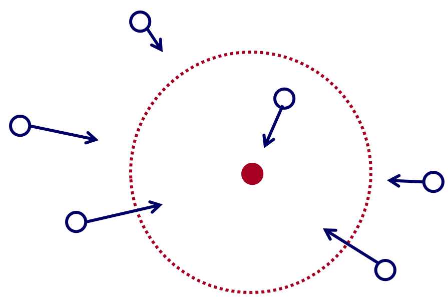


22

Carnegie Mellon

GEMINI

- 'proof' of lower-bounding lemma

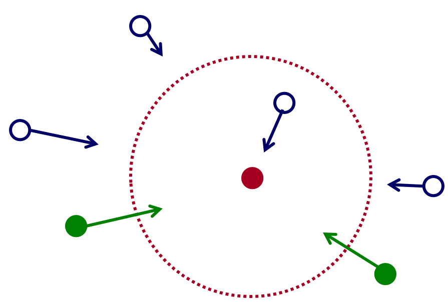


Lower-bounding: Makes objects look closer to each other (& to query object)

Carnegie Mellon

GEMINI

- 'proof' of lower-bounding lemma



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

Carnegie Mellon

GEMINI

Important:

Q: how to extract features?

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

15-826 #25

25

Carnegie Mellon

Time sequences

Q: what features?

15-826 #26

26

Carnegie Mellon

Time sequences

Q: what features?

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

15-826 Copyright (c) 2019 C. Faloutsos #27

27

Carnegie Mellon

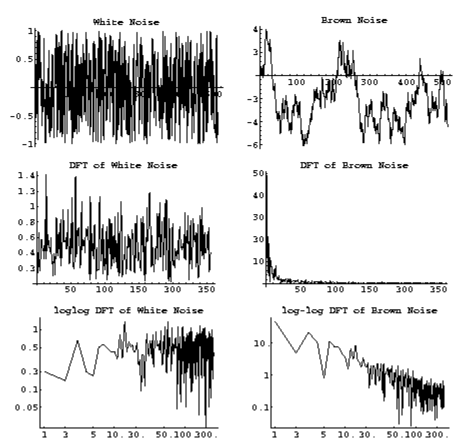
details

Time sequences

white noise brown noise

Fourier spectrum

... in log-log



15-826 #28

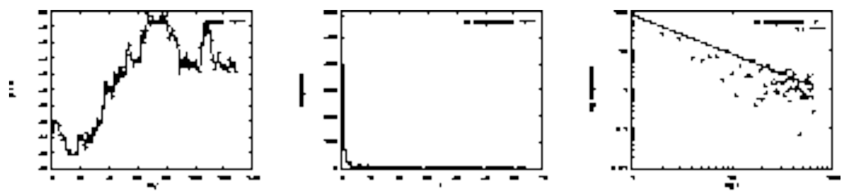
28

Carnegie Mellon

details

Time sequences

- Eg.:



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

29

Carnegie Mellon

details

Time sequences

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

30

Carnegie Mellon

details

Time sequences

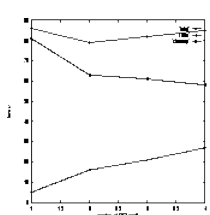
- brown noise: stock prices ($1/f^2$ energy spectrum)
- pink noise: works of art ($1/f$ spectrum)
- black noises: water reservoirs ($1/f^b$, $b>2$)
- (slope: related to 'Hurst exponent', for self-similar traffic, like, eg. Ethernet/web [Schroeder], [Leland+]

Carnegie Mellon

Time sequences - results

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

time



total

cleanup-time

r-tree time

coeff. kept

Carnegie Mellon

Time sequences - improvements:

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

15-826 Copyright (c) 2019 C. Faloutsos #33

33

Carnegie Mellon

Multimedia - Detailed outline

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

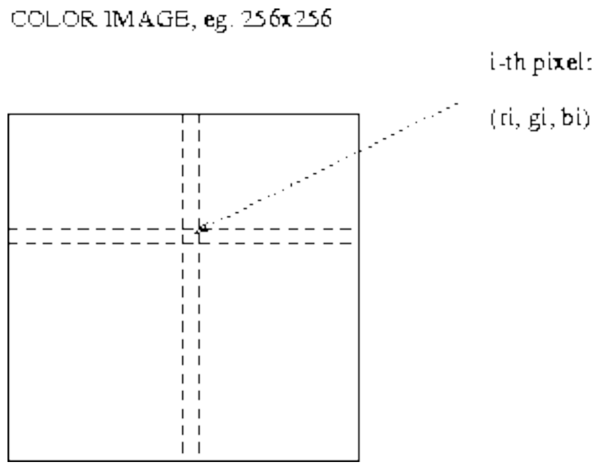
15-826 Copyright (c) 2019 C. Faloutsos #34

34

Carnegie Mellon

Images - color

what is an image?
A: 2-d array

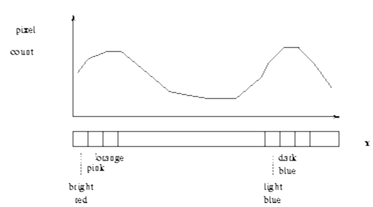


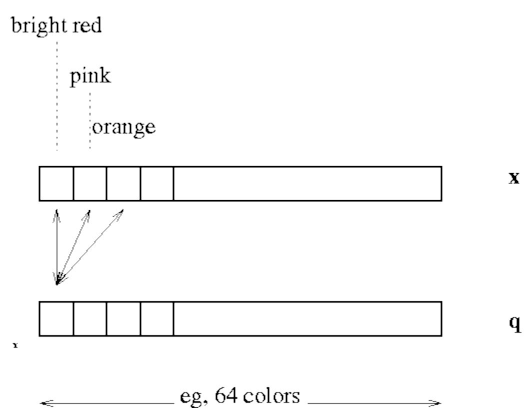
35

Carnegie Mellon

Images - color

Color histograms,
and distance function





36

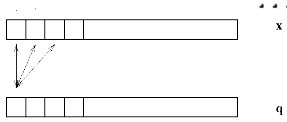
Carnegie Mellon

Images - color

Mathematically, the distance function is:

$$distance_{histogram}(\vec{x}, \vec{q}) = (\vec{x} - \vec{q}) \begin{bmatrix} a_{RR} & a_{RP} & \dots \\ a_{PR} & a_{PP} & \dots \\ \dots & \dots & \dots \end{bmatrix} (\vec{x} - \vec{q})^t$$

$$\dots = (\vec{x} - \vec{q}) \mathcal{A} (\vec{x} - \vec{q})^t$$



15-826 Copyright (c) 2019 C. Faloutsos #37

37

Carnegie Mellon

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) 2019 C. Faloutsos #38

38

Images - color

possible answers:

- avg red, avg green, avg blue

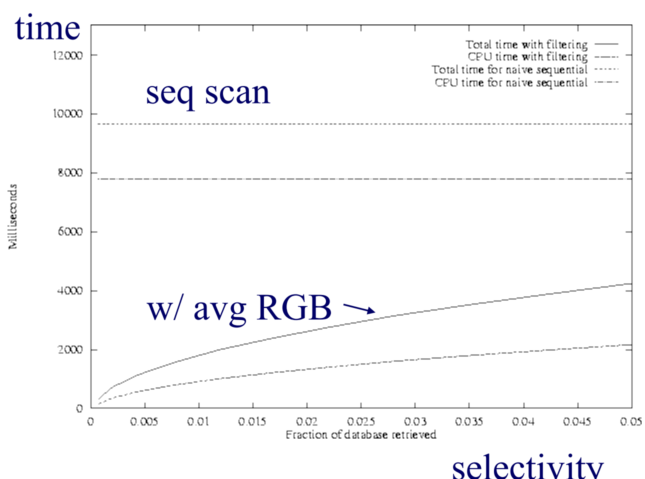
it turns out that this lower-bounds the histogram distance ->

- no cross-talk
- SAMs are applicable

39

Images - color

performance:



40

Carnegie Mellon

Multimedia - Detailed outline

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

Carnegie Mellon

Images - shapes

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

Carnegie Mellon

Images - shapes

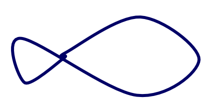
- distance function: Euclidean, on the area, perimeter, and 20 'moments'
- (Q: how to normalize them?
- A: divide by standard deviation)

43

Carnegie Mellon

Images - shapes

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

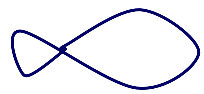


44

Carnegie Mellon

Images - shapes

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



Carnegie Mellon

Images - shapes

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

Carnegie Mellon

details

Images - shapes

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

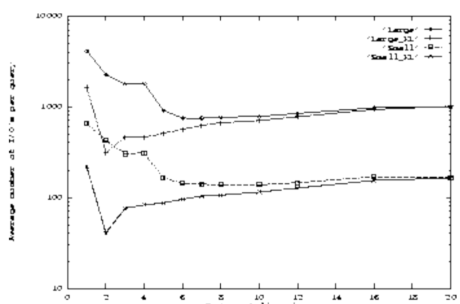
Carnegie Mellon

details

Images - shapes

- Performance: ~10x faster

log(# of I/Os)

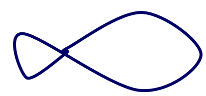


← all kept

# of features kept

Carnegie Mellon

Other shape features?



49

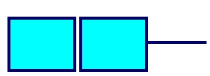
Carnegie Mellon

details

Other shape features

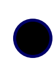
- Morphology (dilations, erosions, openings, closings) [Korn+, VLDB96]

shape



"structuring element"

R=1



50

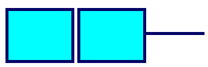
Carnegie Mellon

details

Other shape features

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

shape

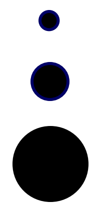


"structuring element"

R=0.5

R=1

R=2



15-826 Copyright (c) 2019 C. Faloutsos #51

51

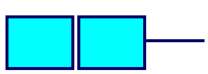
Carnegie Mellon

details

Other shape features

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

shape



"structuring element"

R=0.5

R=1

R=2

15-826 Copyright (c) 2019 C. Faloutsos #52

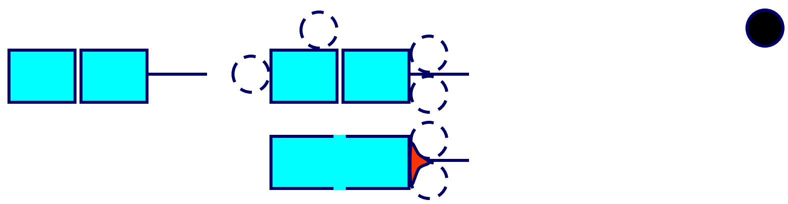
52

Carnegie Mellon

details

Morphology: closing

- fill in small gaps
- **very similar** to 'alpha contours'



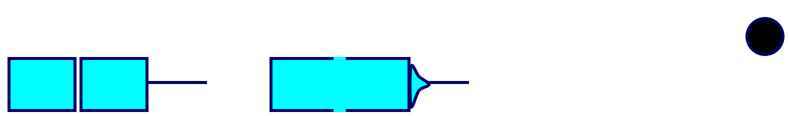
53

Carnegie Mellon

details

Morphology: closing

- fill in small gaps



'closing',
with R=1

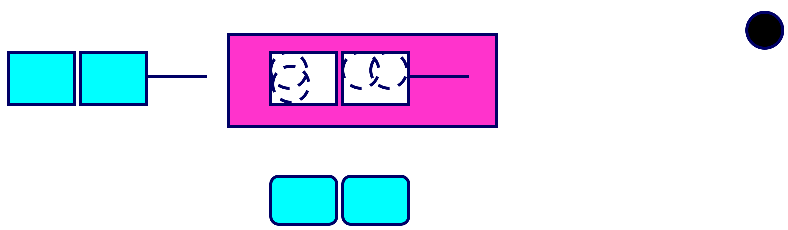
54

Carnegie Mellon

details

Morphology: opening

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



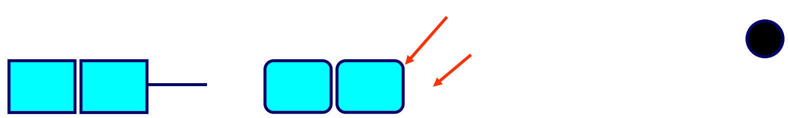
55

Carnegie Mellon

details

Morphology: opening

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



‘opening’
with R=1

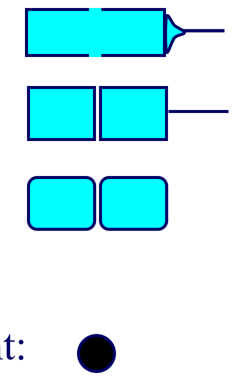
56

Carnegie Mellon

details

Morphology

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



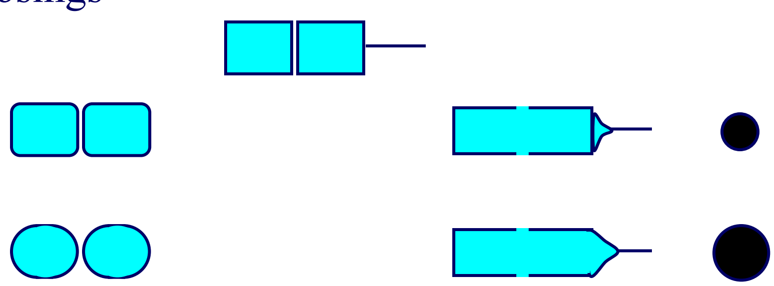
57

Carnegie Mellon

details

Morphology

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



58

Carnegie Mellon

Morphology

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

Carnegie Mellon

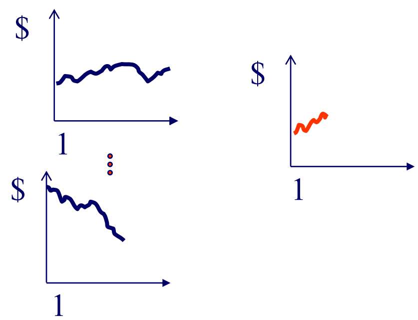
Multimedia - Detailed outline

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

Carnegie Mellon

Sub-pattern matching

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

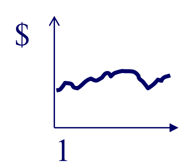


61

Carnegie Mellon

Sub-pattern matching

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



62

Carnegie Mellon

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?)

63

Carnegie Mellon

Sub-pattern matching

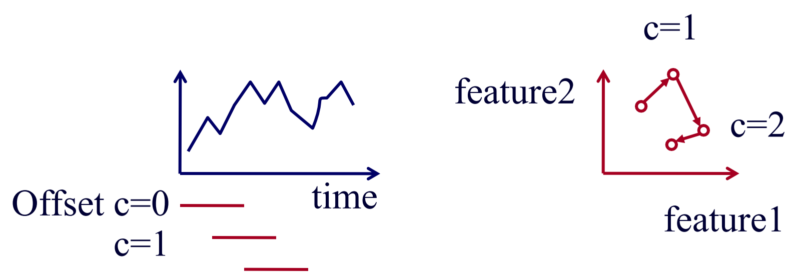
- Assume that queries have minimum duration *w;* (eg., *w*=7 days)
- divide data sequences into windows of width *w* (overlapping, or not?)
- A: sliding, overlapping windows. Thus: trails

Pictorially:

64

Carnegie Mellon

Sub-pattern matching

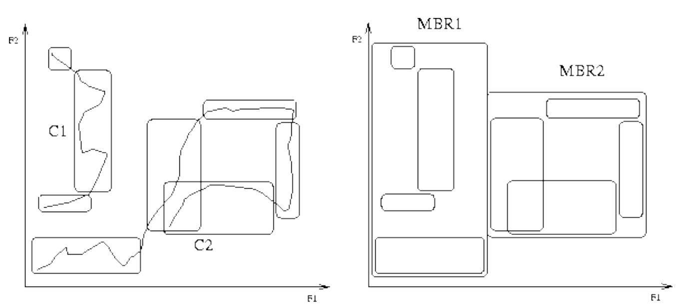


15-826 #65

65

Carnegie Mellon

Sub-pattern matching



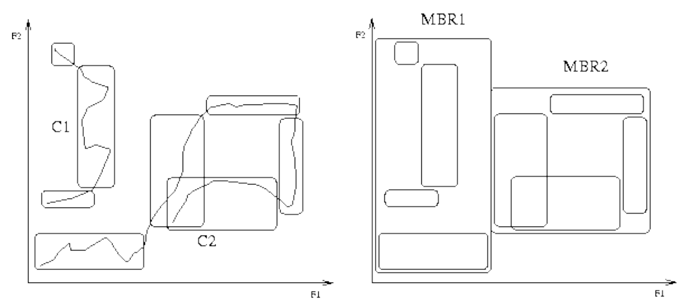
sequences -> trails -> MBRs in feature space

15-826 #66

66

Carnegie Mellon

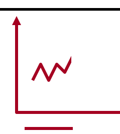
Sub-pattern matching



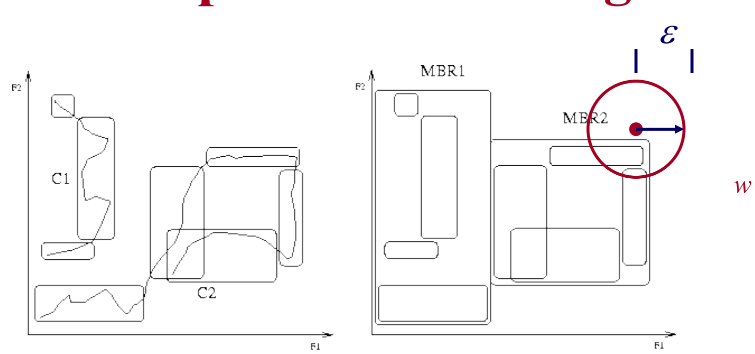
Q: do we store all points? why not?

67

Carnegie Mellon



Sub-pattern matching

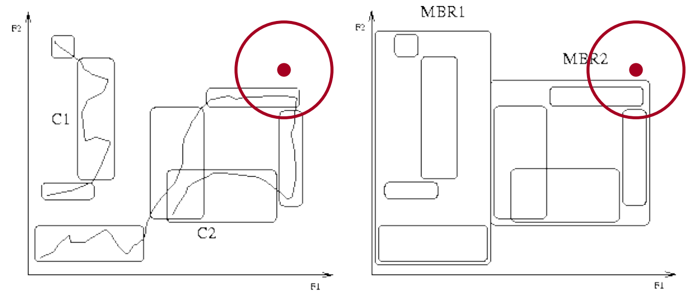


Q: how to do range queries of duration *w*?

68

Carnegie Mellon

Sub-pattern matching



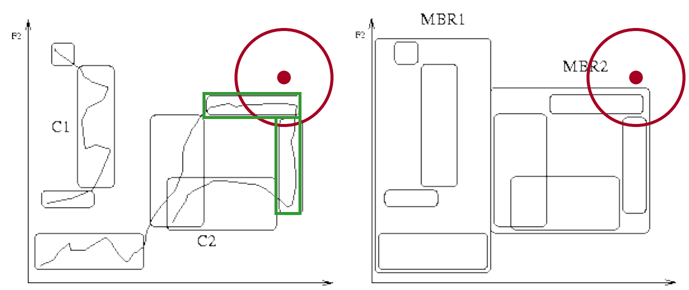
Q: how to do range queries of duration *w*?

A: R-tree; find qualifying stocks and intervals

69

Carnegie Mellon

Sub-pattern matching



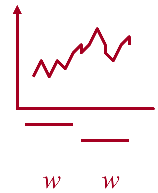
Q: how to do range queries of duration *w*?

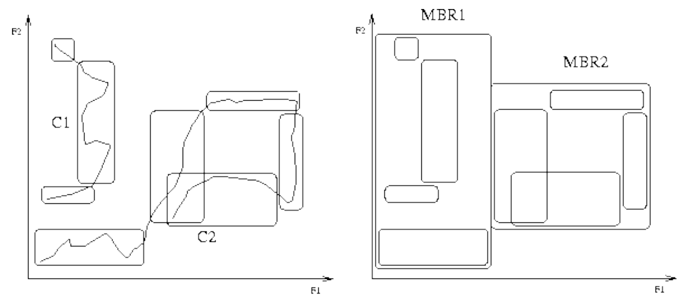
A: R-tree; find qualifying stocks and intervals

70

Carnegie Mellon

Sub-pattern matching





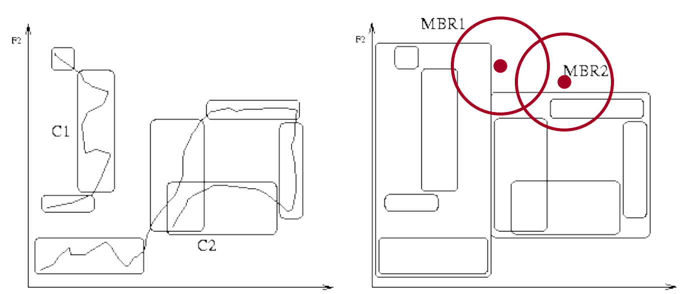
Q: how to do range queries of duration >*w* (say, *2*w*)?

15-826 Copyright (c) 2019 C. Faloutsos #71

71

Carnegie Mellon

Sub-pattern matching



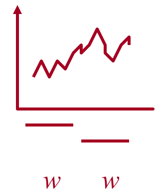
Q: how to do range queries of duration >*w* (say, *2*w*)?

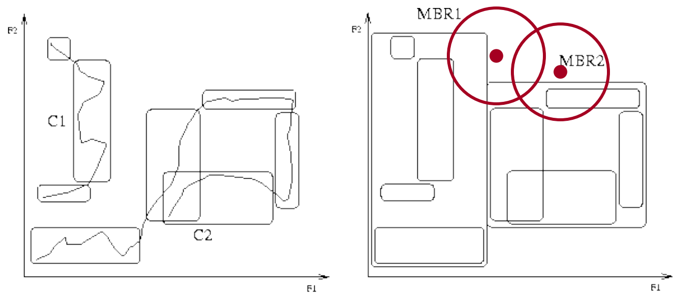
15-826 Copyright (c) 2019 C. Faloutsos #72

72

Carnegie Mellon

Sub-pattern matching





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)

73

Carnegie Mellon

Sub-pattern matching

(improvement [Moon+2001])

- use non-overlapping windows, for data

15-826 Copyright (c) 2019 C. Faloutsos #74

74

Carnegie Mellon

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?)

75

Carnegie Mellon

Multimedia - Detailed outline

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

76

Carnegie Mellon

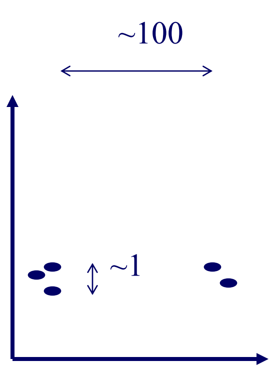
FastMap

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

Carnegie Mellon

FastMap

	O1	O2	O3	O4	O5
O1	0	1	1	100	100
O2	1	0	1	100	100
O3	1	1	0	100	100
O4	100	100	100	0	1
O5	100	100	100	1	0



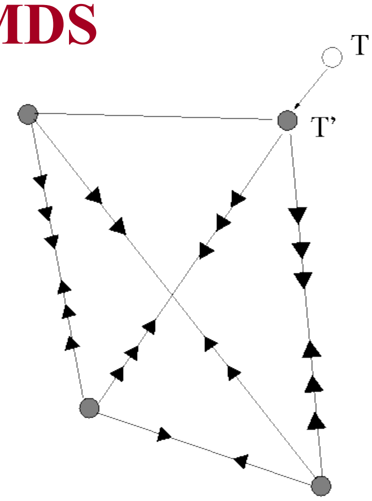
FastMap

- Multi-dimensional scaling (MDS) can do that, but in O(N**2) time

79

MDS

Multi Dimensional Scaling

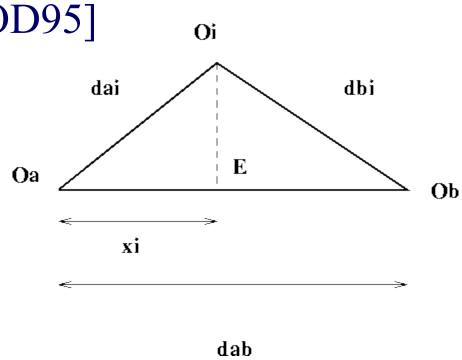


80

Carnegie Mellon

Main idea: projections

We want a linear algorithm: FastMap [SIGMOD95]

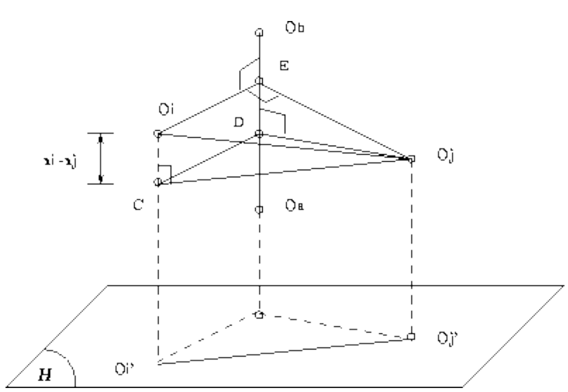


15-826 Copyright (c) 2019 C. Faloutsos #81

81

Carnegie Mellon

FastMap - next iteration



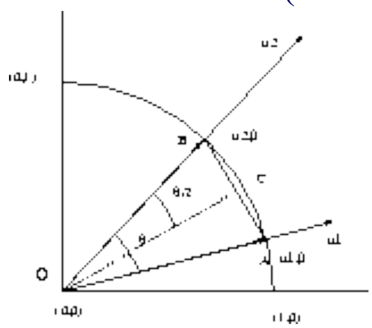
15-826 Copyright (c) 2019 C. Faloutsos #82

82

Carnegie Mellon

Results

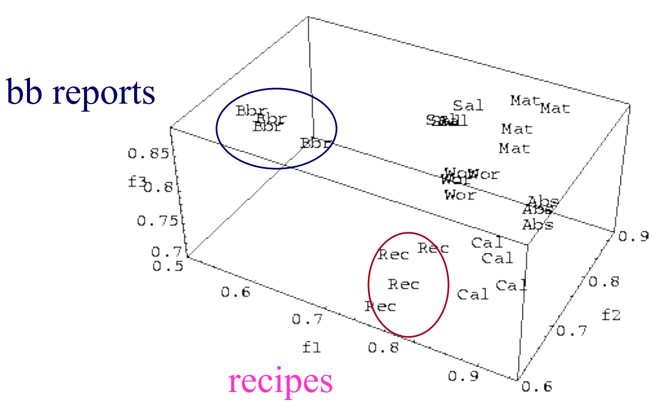
Documents /cosine similarity -> Euclidean distance (how?)



83

Carnegie Mellon

Results

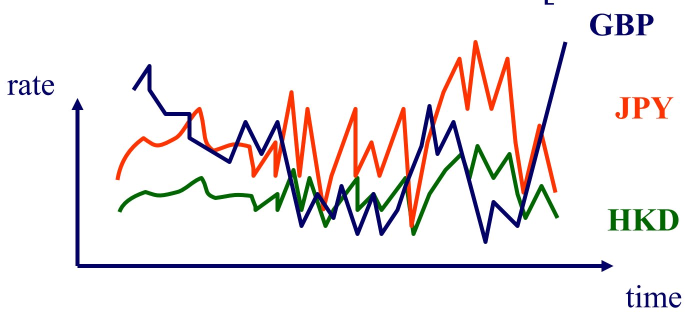


84

Carnegie Mellon

Applications: time sequences

- given *n* co-evolving time sequences
- visualize them + find rules [ICDE00]



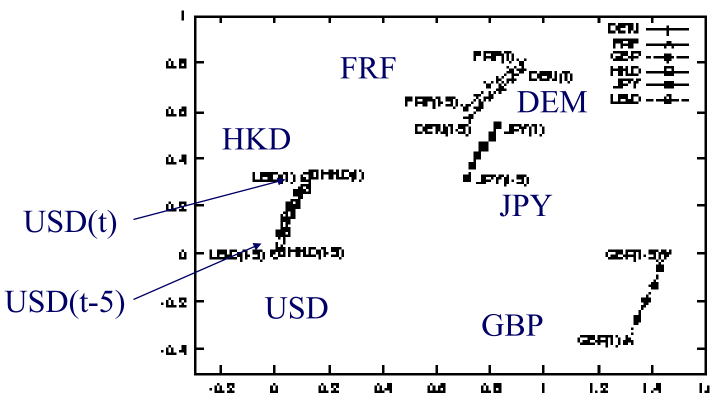
15-826 Copyright (c) 2019 C. Faloutsos #85

85

Carnegie Mellon

Applications - financial

- currency exchange rates [ICDE00]



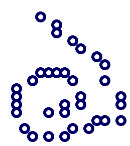
15-826 Copyright (c) 2019 C. Faloutsos #86

86

Carnegie Mellon

Variations

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

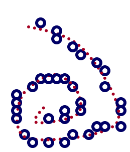


87

Carnegie Mellon

Variations

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



88

Carnegie Mellon

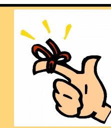
Conclusions

- GEMINI works for multiple settings
- FastMap can extract 'features' automatically (-> indexing, visual d.m.)

15-826 Copyright (c) 2019 C. Faloutsos #89

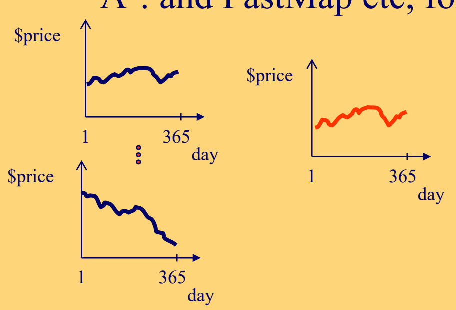
89

Carnegie Mellon



Solution

- Q: Find stocks similar to <MSFT>
- A: GEMINI: Extract features + SAM
  - A': and FastMap etc, for feature extraction



15-826 Copyright (c) 2019 C. Faloutsos 90

90

Carnegie Mellon

References

- Faloutsos, C., R. Barber, et al. (July 1994). "*Efficient and Effective Querying by Image Content*." J. of Intelligent Information Systems 3(3/4): 231-262.
- Faloutsos, C. and K.-I. D. Lin (May 1995). *FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets*. Proc. of ACM-SIGMOD, San Jose, CA.
- Faloutsos, C., M. Ranganathan, et al. (May 25-27, 1994). *Fast Subsequence Matching in Time-Series Databases*. Proc. ACM SIGMOD, Minneapolis, MN.

91

Carnegie Mellon

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

92

Carnegie Mellon

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.

Carnegie Mellon

References

- Lawrence Saul & Sam Roweis. *An Introduction to Locally Linear Embedding* (draft)
- Sam Roweis & Lawrence Saul. *Nonlinear dimensionality reduction by locally linear embedding*. Science, v.290 no.5500 , Dec.22, 2000. pp.2323--2326.
- Schroeder, M. (1991). *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. New York, W.H. Freeman and Company.
- B. Shaw and T. Jebara. "*Minimum Volume Embedding*" . Artificial Intelligence and Statistics, AISTATS, March 2007.

Carnegie Mellon

References

- Josh Tenenbaum, Vin de Silva and John Langford. *A Global Geometric Framework for Nonlinear dimensionality Reduction*. Science 290, pp. 2319-2323, 2000.
- Yi, B.-K. and C. Faloutsos (2000). *Fast Time Sequence Indexing for Arbitrary Lp Norms*. VLDB, Cairo, Egypt.

