

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

Fractals - case studies - I

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
 - z-ordering
 - R-trees
 - misc
- ➔ • fractals
 - intro
 - applications
- text


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



Indexing - Detailed outline

- fractals
 - intro
 - applications
 - disk accesses for R-trees (range queries)
 - dimensionality reduction
 - selectivity in M-trees
 - dim. curse revisited
 - "fat fractals"
 - quad-tree analysis [Gaede+]


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



(Fractals mentioned before:)

- for performance analysis of R-trees
- fractals for dim. reduction

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



Case study#1: R-tree performance

Problem

- Given
 - N points in E-dim space
- Estimate # disk accesses for a range query ($q_1 \times \dots \times q_E$)

(assume: 'good' R-tree, with tight, cube-like MBRs)

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

Case study#1: R-tree performance

Problem

- Given
 - N points in E-dim space
 - with fractal dimension D
- Estimate # disk accesses for a range query ($q1 \times \dots \times q_E$)

(assume: 'good' R-tree, with tight, cube-like MBRs)
Typically, in DB Q-opt: uniformity + independence

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

Examples: World's countries

- BUT: area vs population for ~200 countries (1991 CIA fact-book).

pop log(pop)

area log(area)

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

Examples: World's countries

- neither uniform, nor independent!

pop log(pop)

area log(area)

15-826 Copyright: C. Faloutsos (2006) 9

Examples: TIGER files

- neither uniform, nor independent!

MG county LB county

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

How to proceed?

- recall the [Pagel+] formula, for range queries of size $q1 \times q2$

$$\#DiskAccesses(q1, q2) = \sum (x_{i,1} + q1) * (x_{i,2} + q2)$$

But:

formula needs to know the $x_{i,j}$ sizes of MBRs!

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

R-trees - performance analysis

I.e: for range queries - how many disk accesses, if we just now that we have

- N points in E-d space?

A: can not tell! need to know distribution

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

R-trees - performance analysis

Q: OK - so we are told that the **Hausdorff** fractal dim. = D_0 - Next step?
(also know that there are at most C points per page)

$D_0=1$ $D_0=2$

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

R-trees - performance analysis

Hint: dfn of Hausdorff f.d.:

Felix Hausdorff (1868-1942)

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

Reminder:
Hausdorff or box-counting fd:

- Box counting plot: $\text{Log}(N(r))$ vs $\text{Log}(r)$
- r : grid side
- $N(r)$: count of non-empty cells
- (Hausdorff) fractal dimension D_0 :

$$D_0 = -\frac{\partial \log(N(r))}{\partial \log(r)}$$

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

Reminder

- Hausdorff fd:

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

Reminder

- dfn of Hausdorff fd implies that

$N(r) \sim r^{(-D_0)}$

non-empty cells of side r

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

R-trees - performance analysis

Q (rephrased): what is the side s_1, s_2, \dots of parent nodes, given N data points, packed by C , with f.d. = D_0

$D_0=1$ $D_0=2$

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

R-trees - performance analysis

Q (rephrased): what is the side s_1, s_2, \dots of parent nodes, given N data points, packed by C , with f.d. = D_0

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

R-trees - performance analysis

Q (rephrased): what is the side s_1, s_2, \dots of parent nodes, given N data points, packed by C , with f.d. = D_0

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

R-trees - performance analysis

A: (educated guess)

- $s=s_1=s_2 (= \dots)$ - square-like MBRs
- N/C non-empty cells = $K * s^{(-D_0)}$

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

R-trees - performance analysis

Details of derivations: in [PODS 94].
Finally, expected side s of parent MBRs:

$$s = (C/N)^{1/D_0}$$

Q: sanity check: how does s change with D_0 ?
A:

15-826 Copyright: C. Faloutsos (2006) 22

R-trees - performance analysis

Details of derivations: in [Kamel+, PODS 94].
Finally, expected side s of parent MBRs:

$$s = (C/N)^{1/D_0}$$

Q: sanity check: how does s change with D_0 ?
A: s grows with D_0
Q: does it make sense?
Q: does it suffer from (intrinsic) dim. curse?

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

R-trees - performance analysis

Q: Final-final formula (# disk accesses for range queries $q_1 \times q_2 \times \dots$):
A:

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

R-trees - performance analysis

Q: Final-final formula (# disk accesses for range queries $q_1 \times q_2 \times \dots$):

A: # of parent-node accesses:

$$N/C * (s + q_1) * (s + q_2) * \dots * (s + q_E)$$

A: # of grand-parent node accesses

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

R-trees - performance analysis

Q: Final-final formula (# disk accesses for range queries $q_1 \times q_2 \times \dots$):

A: # of parent-node accesses:

$$N/C * (s + q_1) * (s + q_2) * \dots * (s + q_E)$$

A: # of grand-parent node accesses

$$N/(C^2) * (s' + q_1) * (s' + q_2) * \dots * (s' + q_E)$$

$$s' = (C^2/N)^{1/D_0}$$

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

R-trees - performance analysis

Results: IUE (x-y star coordinates)

leaf accesses

(a) IUE - Leaf accesses vs. query side

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

R-trees - performance analysis

Results: LB County

leaf accesses

(b) LB County - Leaf accesses vs. query side

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

R-trees - performance analysis

Results: MG-county

leaf accesses

(a) MG County - Leaf accesses vs. query side

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

R-trees - performance analysis

Results: 2D- uniform

leaf accesses

(a) 2D-UNIFORM - Leaf accesses vs. query side

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

CMU SCS

R-trees - performance analysis

Conclusions: usually, <5% relative error, for range queries

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

CMU SCS

Indexing - Detailed outline

- fractals
 - intro
 - applications
 - disk accesses for R-trees (range queries)
 - dimensionality reduction
 - selectivity in M-trees
 - dim. curse revisited
 - "fat fractals"
 - quad-tree analysis [Gaede+]
 -




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

CMU SCS

Case study #2: Dim. reduction

Problem definition: 'Feature selection'

- given N points, with E dimensions
- keep the k most 'informative' dimensions [Traina+, SBBD'00]

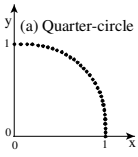




Caetano Traina Agma Traina Leejay Wu

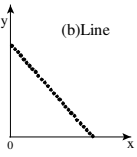
15-826 Copyright: C. Faloutsos (2006) 33

CMU SCS

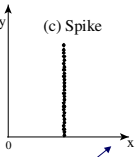
Dim. reduction - w/ fractals



(a) Quarter-circle



(b) Line



(c) Spike

not informative

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

CMU SCS

Dim. reduction

Problem definition: 'Feature selection'

- given N points, with E dimensions
- keep the k most 'informative' dimensions

Re-phrased: spot and drop attributes with strong (non-)linear correlations

Q: how do we do that?

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

CMU SCS

Dim. reduction

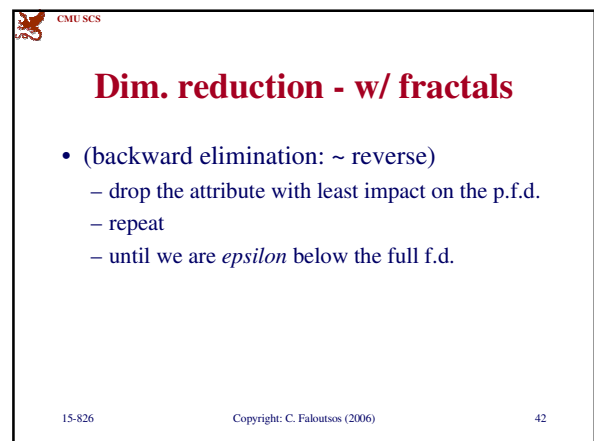
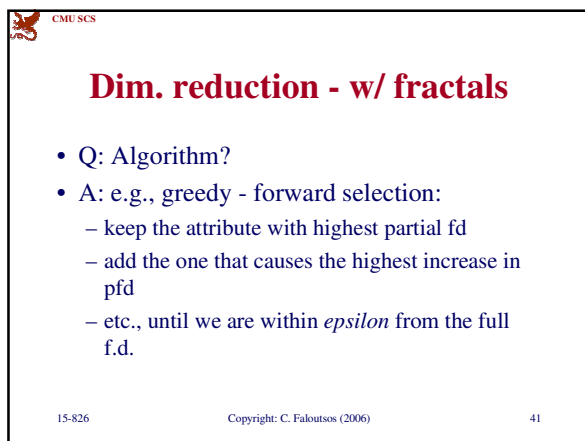
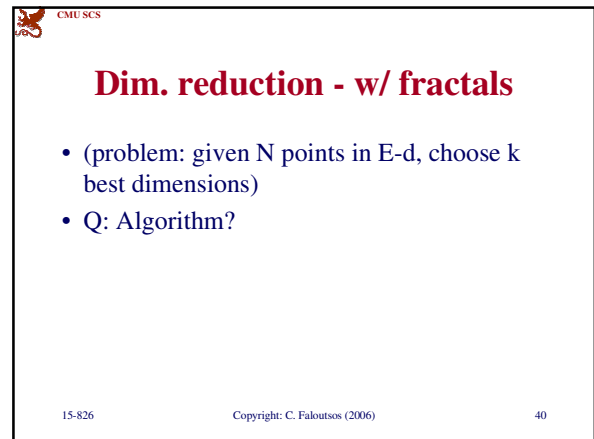
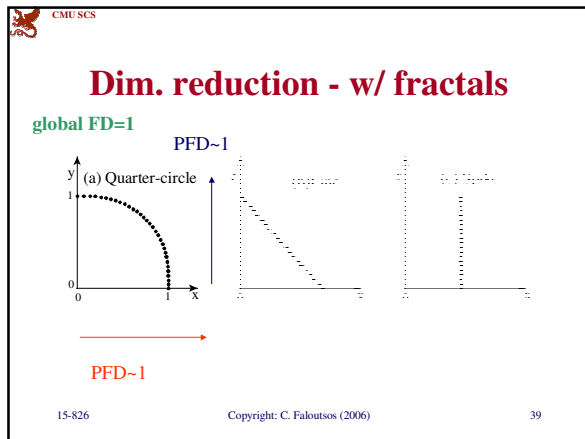
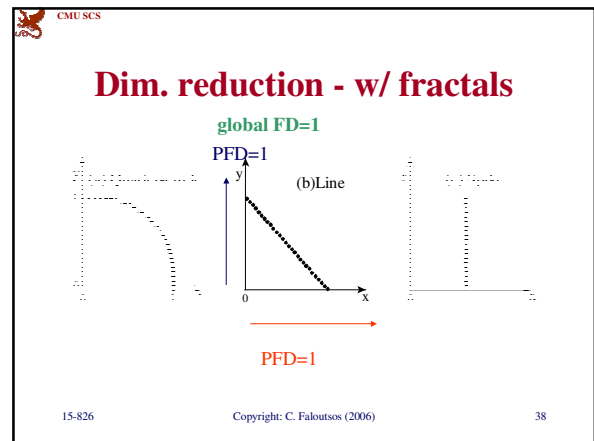
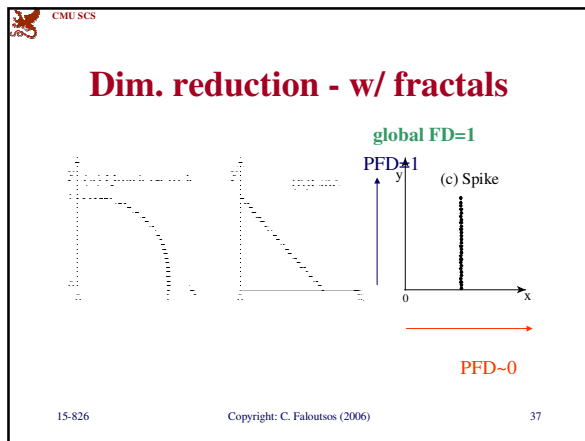
A: Hint: correlated attributes do not affect the intrinsic/fractal dimension, e.g., if


$$y = f(x, z, w)$$

we can drop y

(hence: 'partial fd' (PFD) of a set of attributes = the fd of the dataset, when projected on those attributes)

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



CMU SCS


Dim. reduction - w/ fractals

- Q: what is the smallest # of attributes we should keep?

15-826

Copyright: C. Faloutsos (2006)

43

CMU SCS


Dim. reduction - w/ fractals

- Q: what is the smallest # of attributes we should keep?
- A: we should keep at least as many as the f.d. (and probably, a few more)

15-826

Copyright: C. Faloutsos (2006)

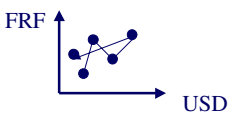
44

CMU SCS

Dim. reduction - w/ fractals

- Results: E.g., on the ‘currency’ dataset
- (daily exchange rates for USD, HKD, BP, FRF, DEM, JPY - i.e., 6-d vectors, one per day - base currency: CAD)

e.g.: FRF




USD

15-826

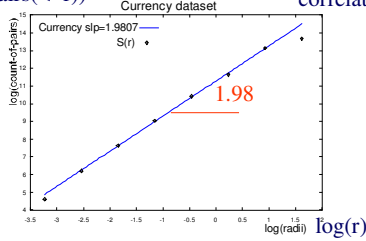
Copyright: C. Faloutsos (2006)

45

CMU SCS

E.g., on the ‘currency’ dataset

$\log(\#pairs(\leq r))$ Currency dataset correlation integral




Currency slope=1.9807
 $S(r)$
1.98
 $\log(radius)$ $\log(r)$

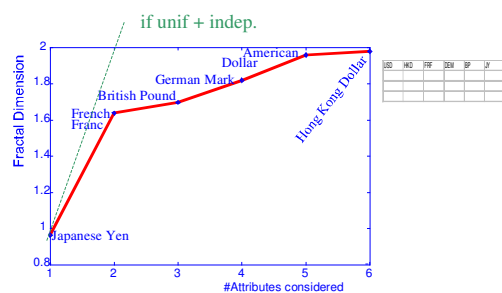
15-826

Copyright: C. Faloutsos (2006)

46

CMU SCS

E.g., on the ‘currency’ dataset



if unif + indep.

Fractal Dimension

#Attributes considered


Japanese Yen, French Franc, British Pound, German Mark, American Dollar, Hong Kong Dollar

USD	HKD	FRF	DEM	JPY

15-826

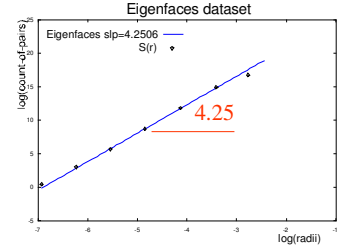
Copyright: C. Faloutsos (2006)

47

CMU SCS

E.g., on the eigenface dataset

16-d vectors, one for each of ~1K faces



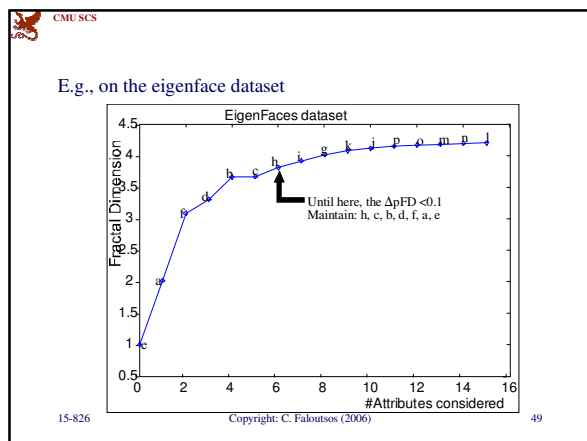
Eigenfaces dataset

Eigenfaces slope=4.2506
 $S(r)$
4.25
 $\log(radius)$

15-826

Copyright: C. Faloutsos (2006)

48



CMU SCS

Dim. reduction - w/ fractals

Conclusion:

- can do non-linear dim. reduction

global FD=1

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

CMU SCS

References

- [PODS94] Faloutsos, C. and I. Kamel (May 24-26, 1994). *Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension*. Proc. ACM SIGACT-SIGMOD-SIGART PODS, Minneapolis, MN.
- [Traina+, SBBD'00] Traina, C., A. Traina, et al. (2000). *Fast feature selection using the fractal dimension*. XV Brazilian Symposium on Databases (SBBD), Paraiba, Brazil.

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