


15-826: Multimedia Databases and Data Mining


Lecture #10: Fractals - case studies - I
C. Faloutsos



Reading Material

- Christos Faloutsos and Ibrahim Kamel, *Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension*, Proc. ACM SIGACT-SIGMOD-SIGART PODS, May 1994, pp. 4-13, Minneapolis, MN.
- (optional, but very useful: Manfred Schroeder *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* W.H. Freeman and Company, 1991)

15-826 Copyright: C. Faloutsos (2008) 2



Outline

Goal: 'Find similar / interesting things'

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

15-826 Copyright: C. Faloutsos (2008) 3

CMU SCS

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 (2008) 4

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 (2008) 5

CMU SCS

(Fractals mentioned before:)

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

15-826 Copyright: C. Faloutsos (2008) 6

CMU SCS

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 (2008) 7

CMU SCS

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 $(q_1 \times \dots \times q_E)$

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

15-826 Copyright: C. Faloutsos (2008) 8

CMU SCS

Examples: World's countries

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

pop

area

log(pop)

log(area)

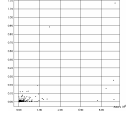
15-826 Copyright: C. Faloutsos (2008) 9

CMU SCS

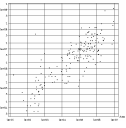
Examples: World's countries

- neither uniform, nor independent!

pop



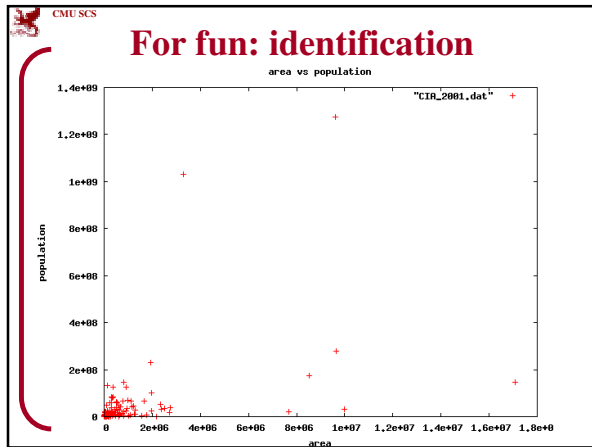
log(pop)

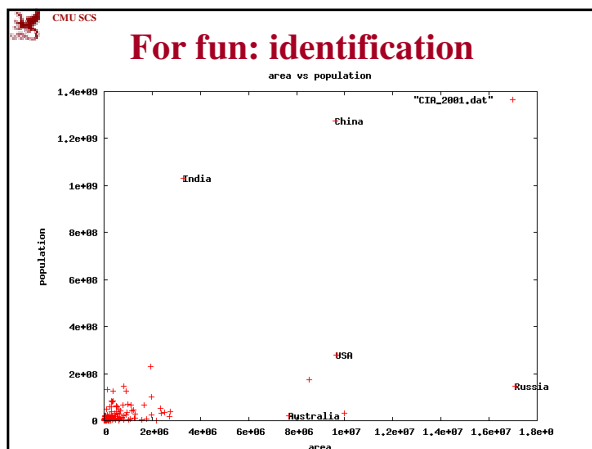


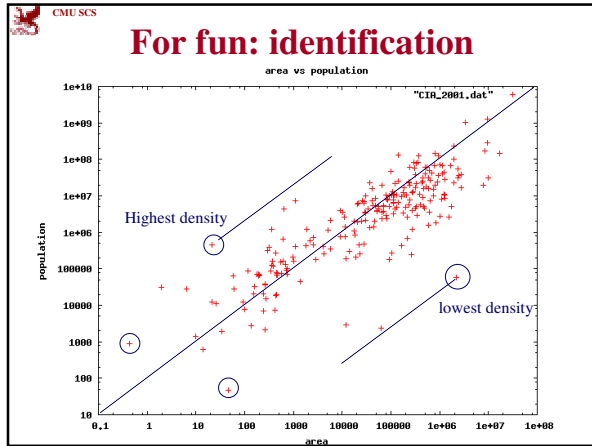
area

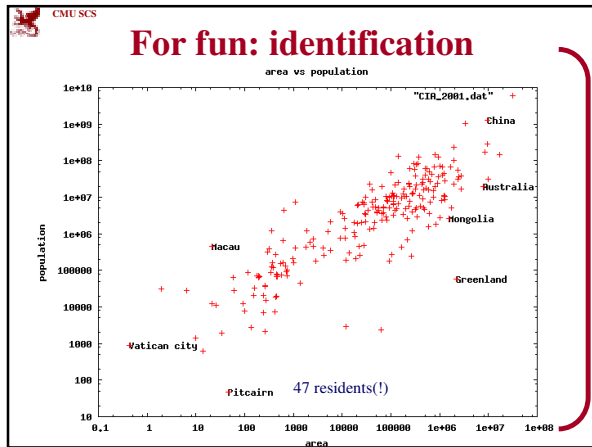
log(area)

15-826 Copyright: C. Faloutsos (2008) 10









Examples: TIGER files

- neither uniform, nor independent!

MG county

LB county

15-826 Copyright: C. Faloutsos (2008) 15

CMU SCS

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 (2008) 16

CMU SCS

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 (2008) 17

CMU SCS

R-trees - performance analysis

Q: OK - so we are told that the Hausdorff fractal dim. = $D0$ - Next step?
(also know that there are at most C points per page)

$D0=1$ $D0=2$




15-826 Copyright: C. Faloutsos (2008) 18

CMU SCS

R-trees - performance analysis

Hint: dfn of Hausdorff f.d.:



Felix Hausdorff (1868-1942)

15-826 Copyright: C. Faloutsos (2008) 19

CMU SCS

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 :

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

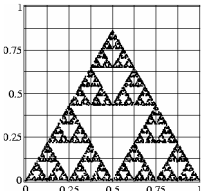
15-826 Copyright: C. Faloutsos (2008) 20

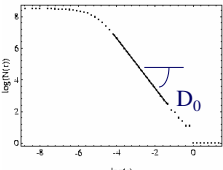
CMU SCS

Reminder

- Hausdorff fd:

r — $\log(\#\text{non-empty cells})$





15-826 Copyright: C. Faloutsos (2008) 21

CMU SCS

Reminder

- dfn of Hausdorff fd implies that

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

↙

non-empty cells of side r


15-826 Copyright: C. Faloutsos (2008) 22

CMU SCS

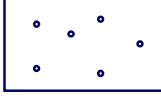
R-trees - performance analysis

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

$D_0=1$



$D_0=2$



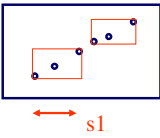
15-826 Copyright: C. Faloutsos (2008) 23

CMU SCS

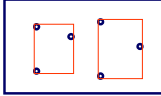
R-trees - performance analysis

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 (2008) 24

CMU SCS

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. = $D0$

D0=1

$s_1=s_2=s$

D0=2

15-826 Copyright: C. Faloutsos (2008) 25

CMU SCS

R-trees - performance analysis

A: (educated guess)

- $s=s_1=s_2 (= \dots)$ - square-like MBRs
- N/C non-empty cells = $K * s^{(-D0)}$

D0=1

s_1 s_2

D0=2

$\log(\#cells)$

$\log(s)$

15-826 Copyright: C. Faloutsos (2008) 26

CMU SCS

R-trees - performance analysis

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


$s = (C/N)^{1/D0}$

Q: sanity check: how does s change with $D0$?
 A:

15-826 Copyright: C. Faloutsos (2008) 27

CMU SCS

R-trees - performance analysis

Details of derivations: in [Kamel+, PODS 94]. 

Finally, expected side s of parent MBRs:

$$s = (C/N)^{1/D0}$$

Q: sanity check: how does s change with $D0$?
 A: s grows with $D0$
 Q: does it make sense?

Q: does it suffer from (intrinsic) dim. curse?

15-826 Copyright: C. Faloutsos (2008) 28

CMU SCS

R-trees - performance analysis

Q: Final-final formula (# disk accesses for range queries $q1$ x $q2$ x ...):
 A:

15-826 Copyright: C. Faloutsos (2008) 29

CMU SCS

R-trees - performance analysis

Q: Final-final formula (# disk accesses for range queries $q1$ x $q2$ x ...):
 A: # of parent-node accesses:

$$N/C * (s + q1) * (s + q2) * ... * (s + qE)$$

 A: # of grand-parent node accesses

15-826 Copyright: C. Faloutsos (2008) 30

CMU SCS

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 (2008) 31

CMU SCS

R-trees - performance analysis

Results: IUE (x-y star coordinates)

leaf accesses

15-826 Copyright: C. Faloutsos (2008) 32

CMU SCS

R-trees - performance analysis

Results: LB County

leaf accesses

15-826 Copyright: C. Faloutsos (2008) 33

CMU SCS

R-trees - performance analysis

Results: MG-county

leaf accesses

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

query side

15-826 Copyright: C. Faloutsos (2008) 34

CMU SCS

R-trees - performance analysis

Results: 2D- uniform

leaf accesses

(d) 2D-Uniform - Leaf accesses vs. query side

query side

15-826 Copyright: C. Faloutsos (2008) 35

CMU SCS

R-trees - performance analysis

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

15-826 Copyright: C. Faloutsos (2008) 36

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 (2008) 37

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]

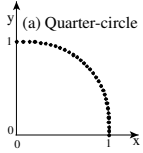




Caetano Traina Agma Traina Leejay Wu

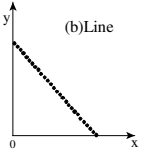
15-826 Copyright: C. Faloutsos (2008) 38

CMU SCS

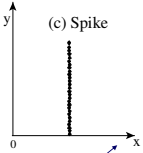
Dim. reduction - w/ fractals



(a) Quarter-circle



(b) Line



(c) Spike

not informative

15-826 Copyright: C. Faloutsos (2008) 39

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 (2008) 40

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 (2008) 41

CMU SCS

Dim. reduction - w/ fractals

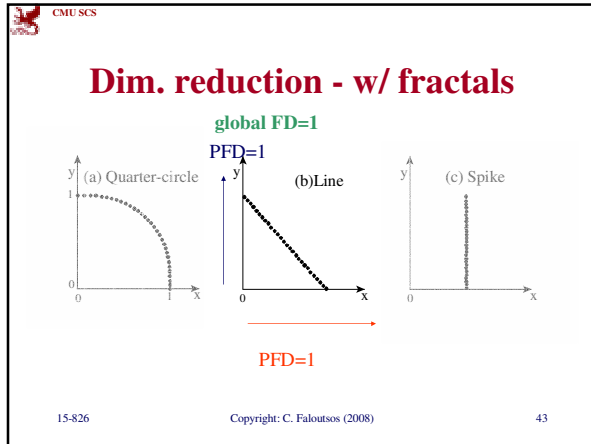
global FD=1

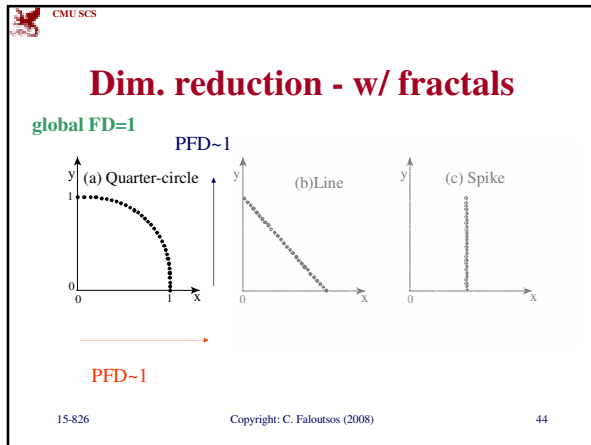
(a) Quarter-circle (b) Line (c) Spike

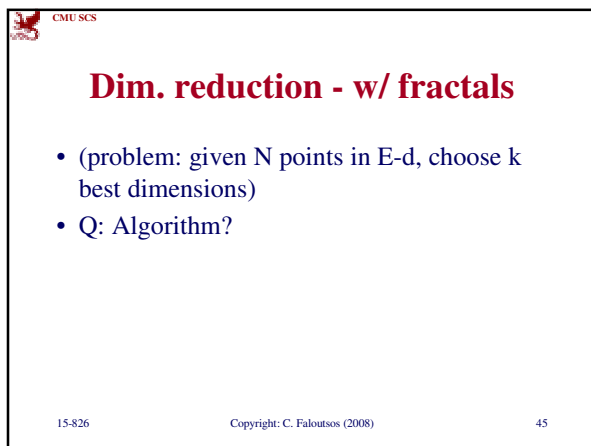
PFD=1


PFD=0

15-826 Copyright: C. Faloutsos (2008) 42








 CMU SCS

Dim. reduction - w/ fractals

- Q: Algorithm?
- A: e.g., greedy - forward selection:
 - keep the attribute with highest partial fd
 - add the one that causes the highest increase in pfd
 - etc., until we are within *epsilon* from the full f.d.


15-826 Copyright: C. Faloutsos (2008) 46

 CMU SCS

Dim. reduction - w/ fractals

- (backward elimination: ~ reverse)
 - drop the attribute with least impact on the p.f.d.
 - repeat
 - until we are *epsilon* below the full f.d.

15-826 Copyright: C. Faloutsos (2008) 47

 CMU SCS

Dim. reduction - w/ fractals

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

15-826 Copyright: C. Faloutsos (2008) 48

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)

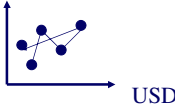
15-826 Copyright: C. Faloutsos (2008) 49

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



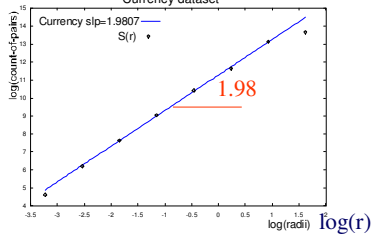
USD

15-826 Copyright: C. Faloutsos (2008) 50

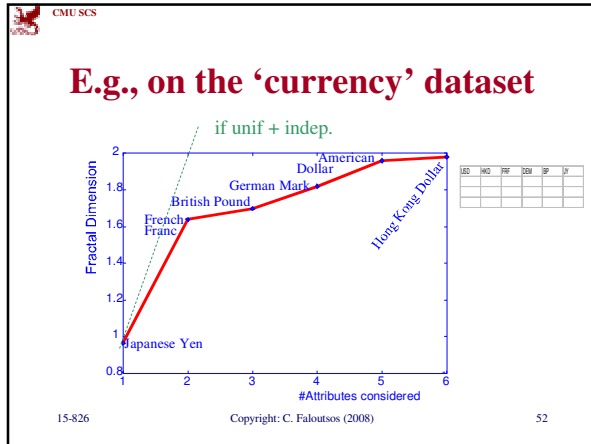
CMU SCS

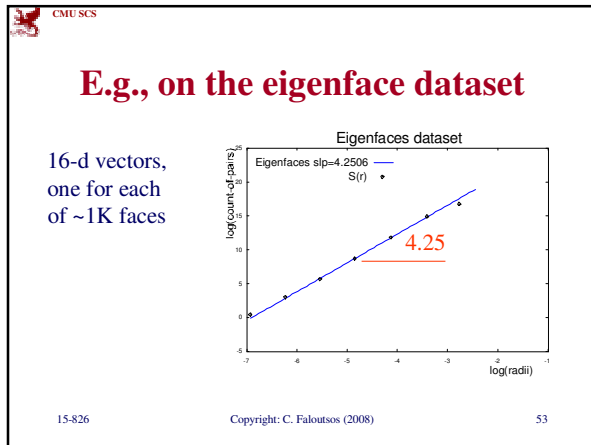
E.g., on the ‘currency’ dataset

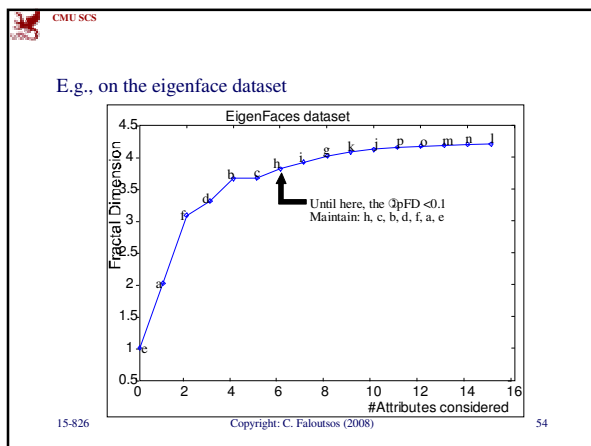
log(#pairs($\leq r$)) correlation integral



15-826 Copyright: C. Faloutsos (2008) 51







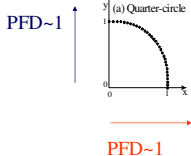
CMU SCS

Dim. reduction - w/ fractals

Conclusion:

- can do non-linear dim. reduction

global FD=1



15-826 Copyright: C. Faloutsos (2008) 55

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+, SBBB'00] Traina, C., A. Traina, et al. (2000). *Fast feature selection using the fractal dimension*. XV Brazilian Symposium on Databases (SBBB), Paraiba, Brazil.

15-826 Copyright: C. Faloutsos (2008) 56
