


15-826: Multimedia Databases and Data Mining

Fractals - case studies - II
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



Metric trees - analysis

- Problem: How many disk accesses, for an M-tree?
- Given:
 - N (# of objects)
 - C (fanout of disk pages)
 - r (radius of range query - BIASED model)

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



Metric trees - analysis

- Problem: How many disk accesses, for an M-tree?
- Given:
 - N (# of objects)
 - C (fanout of disk pages)
 - r (radius of range query - BIASED model)
- NOT ENOUGH - what else do we need?

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

CMU SCS

Metric trees - analysis

- A: something about the distribution


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

CMU SCS

Metric trees - analysis

- A: something about the distribution

[Ciaccia, Patella, Zezula, PODS98]: assumed that the distance distribution is the same, for every object:



Paolo Ciaccia Marco Patella

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

CMU SCS

Metric trees - analysis

- A: something about the distribution

[Ciaccia+, PODS98]: assumed that the distance distribution is the same, for every object:

$F1(d) = \text{Prob}(\text{an object is within } d \text{ from object \#1})$
 $= F2(d) = \dots = F(d)$

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

CMU SCS

Metric trees - analysis

- A: something about the distribution
- Given our 'fractal' tools, we could try them - which one?

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

CMU SCS

Metric trees - analysis

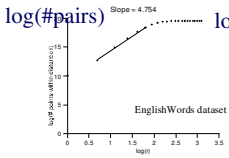
- A: something about the distribution
- Given our 'fractal' tools, we could try them - which one?
- A: Correlation integral [Traina+, ICDE2000]

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

CMU SCS

Metric trees - analysis

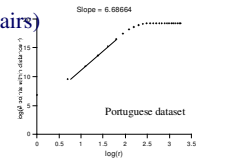
English dictionary



EnglishWords dataset

$\log(d)$

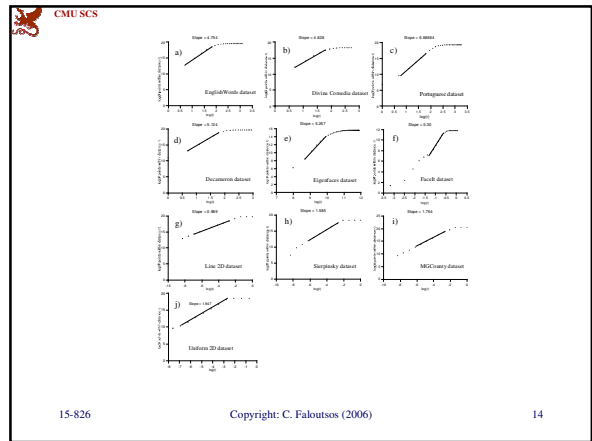
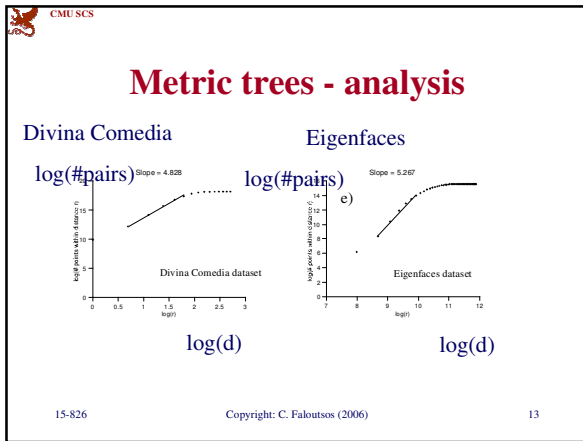
Portuguese dictionary



Portuguese dataset

$\log(d)$

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

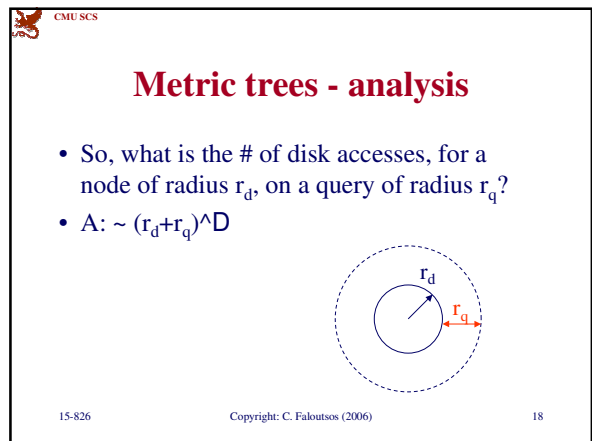
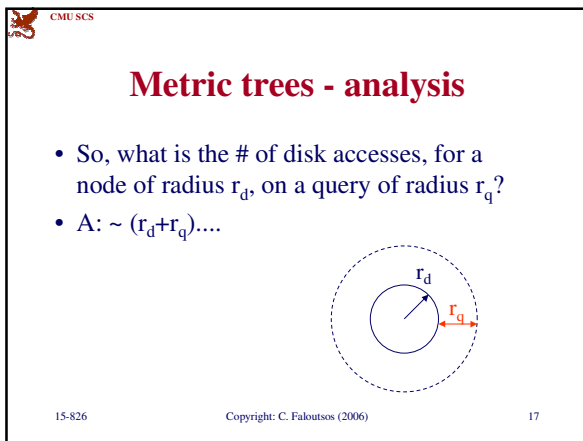
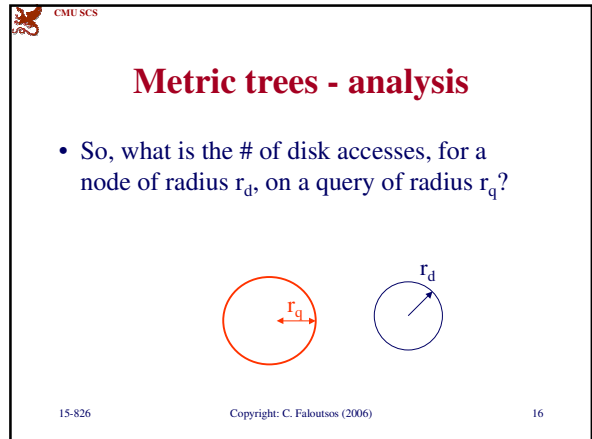


Metric trees - analysis

	Data Set	N (# Objects)	Dimension	Distance Function	Distance Exponent D
Real Metric datasets	English	25,143	NA	L_{edit}	4.753
	Divina Commedia	12,701	NA	L_{edit}	4.827
	Decamerone	18,719	NA	L_{edit}	5.124
	Portuguese	21,473	NA	L_{edit}	6.686
	Facell	1,056	NA	Not divulged	6.821
Real vector datasets	MGCounty	15,559	2	L_2	1.752
	Eigenfaces	11,900	16	L_2	5.267
	Sierpinsky	9,841	2	L_2	1.584
Synthetic datasets	2D Line	20,000	2	L_2	0.989
	Uniform 2D	10,000	2	L_2	1.947

Table 2 - Datasets used in the experiments.

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



CMU SCS

Accuracy of selectivity formulas

$\log(\#d.a.)$

$\log(rq)$

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

CMU SCS

Fast estimation of D

- Normally, D takes $O(N^2)$ time
- Anything faster? suppose we have already built an M-tree

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

CMU SCS

Fast estimation of D

- Hint:

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

CMU SCS

Fast estimation of D

- Hint:

ratio of radii:
 $r1^D * C = r2^D$

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

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 [Gaedde+]

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

CMU SCS

Dim. curse revisited

- (Q: how serious is the dim. curse, e.g.):
- Q: what is the search effort for k-nn?
 - given N points, in E dimensions, in an R-tree, with k-nn queries ('biased' model)

[Pagel, Korn + ICDE 2000]

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

Reminder: Hausdorff Dimension (D_0)

- r = side length (each dimension)
- $B(r)$ = # boxes containing points $\propto r^{D_0}$

$r = 1/2 \quad B = 2$	$r = 1/4 \quad B = 4$	$r = 1/8 \quad B = 8$
$\log r = -1$	$\log r = -2$	$\log r = -3$
$\log B = 1$	$\log B = 2$	$\log B = 3$

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

Reminder: Correlation Dimension (D_2)

- $S(r) = \sum p_i^2$ (squared % pts in box) $\propto r^{D_2}$
 \propto #pairs(within $\leq r$)

$r = 1/2 \quad S = 1/2$	$r = 1/4 \quad S = 1/4$	$r = 1/8 \quad S = 1/8$
$\log r = -1$	$\log r = -2$	$\log r = -3$
$\log S = -1$	$\log S = -2$	$\log S = -3$

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

Observation #1

- How to determine avg MBR side l ?
 N = #pts, C = MBR capacity

Hausdorff dimension: $B(r) \propto r^{D_0}$

$$B(l) = N/C = l^{-D_0} \Rightarrow l = (N/C)^{1/D_0}$$

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

Observation #2

- k -NN query $\rightarrow \epsilon$ -range query
 For k pts, what radius ϵ do we expect?

Correlation dimension: $S(r) \propto r^{D_2}$

$$S(\epsilon) = \frac{k}{N-1} = (2\epsilon)^{D_2}$$

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

Observation #3

- Estimate avg # query-sensitive anchors:
 How many **expected** q will touch **avg** page?
 Page touch: q stabs ϵ -dilated MBR(p)

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

Asymptotic Formula

- k -NN page accesses as $N \rightarrow \infty$
 C = capacity
 D = fractal dimension ($=D_0 \sim D_2$)

$$P_{all}^{L\infty}(k) \approx \sum_{j=0}^h \left\{ \frac{1}{C^{h-j}} + \left[1 + \left(\frac{k}{C^{h-j}} \right)^{1/D} \right]^D \right\}$$

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

Asymptotic Formula

$$P_{all}^{L_{\infty}}(k) \approx \sum_{j=0}^h \left\{ \frac{1}{C^{h-j}} + \left[1 + \left(\frac{k}{C^{h-j}} \right)^{1/D} \right]^D \right\}$$

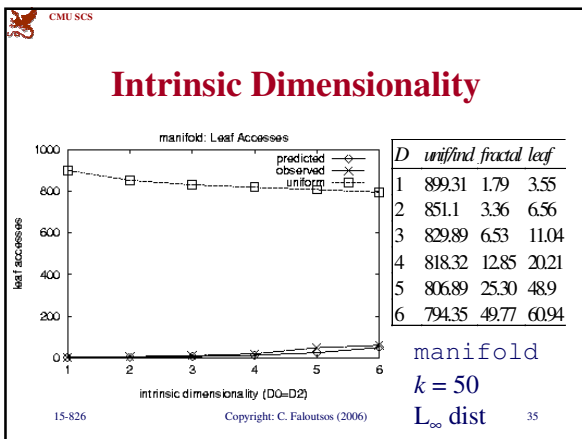
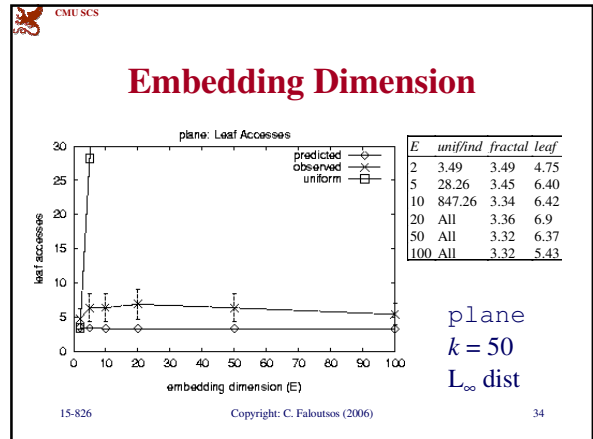
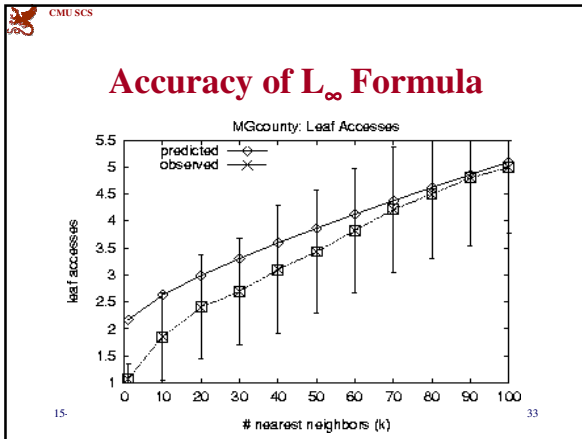
- NO mention of the embedding dimensionality!!
- Still have dim. curse, but on f.d. D

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

Synthetic Data

- plane
 - $D_0 = D_2 = 2$
 - embedded in E -space
 - $N = 100K$
- manifold
 - $E = 8$
 - $D_0 = D_2$ varies from 1-6
 - line, plane, etc. (in 8-d)

15-826 Copyright: C. Faloutsos




Non-Euclidean Data Set

E	unif	ind	fractal	leaf
2	3.49	2.53	4.72	±1.81
10	847.26	2.53	6.42	±2.11
20	all	2.53	7.76	±4.12
50	all	2.53	6.15	±2.82
100	all	2.53	5.64	±2.32

sierpinski $k = 50$, L_{∞} dist

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




CMU SCS

Conclusions

- Worst-case theory is **over-pessimistic**
- High dimensional data can exhibit good performance if **correlated, non-uniform**
- Many real data sets are **self-similar**
- Determinant is **intrinsic** dimensionality
 - multiple fractal dimensions (D_0 and D_2)
 - indication of how far one can go

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



CMU SCS

References

- Ciaccia, P., M. Patella, et al. (1998). *A Cost Model for Similarity Queries in Metric Spaces*. PODS.
- Pagel, B.-U., F. Korn, et al. (2000). *Deflating the Dimensionality Curse Using Multiple Fractal Dimensions*. ICDE, San Diego, CA.
- Traina, C., A. J. M. Traina, et al. (2000). *Distance Exponent: A New Concept for Selectivity Estimation in Metric Trees*. ICDE, San Diego, CA.

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