15-826: Multimedia Databases and Data Mining

Lecture #9: Fractals - introduction

C. Faloutsos

Must-read 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.

Recommended Material

optional, but very useful:

  (on reserve in the library)
  – Chapter 10: boxcounting method
  – Chapter 1: Sierpinski triangle
Outline

Goal: ‘Find similar / interesting things’
- Intro to DB
- Indexing - similarity search
- Data Mining

Indexing - Detailed outline
- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
  - z-ordering
  - R-trees
  - misc
- fractals
  - intro
  - applications
- text

Intro to fractals - outline
- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots
Problem #1: GIS - points

Road end-points of Montgomery county:

- Q1: how many d.a. for an R-tree?
- Q2: distribution?
  - not uniform
  - not Gaussian
  - no rules??

Problem #2 - spatial d.m.

Galaxies (Sloan Digital Sky Survey w/ B. Nichol)

- ‘spiral’ and ‘elliptical’ galaxies
  (stores and households ...)
- patterns?
- attraction/repulsion?
- how many ‘spi’ within r from an ‘ell’?

Problem #3: traffic

- disk trace (from HP - J. Wilkes); Web traffic - fit a model
- how many explosions to expect?
- queue length distr.?
Problem #3: traffic

Q: Then, how to generate such bursty traffic?
Common answer:

- Fractals / self-similarities / power laws
- Seminal works from Hilbert, Minkowski, Cantor, Mandelbrot, (Hausdorff, Lyapunov, Ken Wilson, …)

Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots

What is a fractal?

= self-similar point set, e.g., Sierpinski triangle:

- zero area;
- infinite length!
Definitions (cont’d)

• Paradox: Infinite perimeter; Zero area!
• ‘dimensionality’: between 1 and 2
• actually: \( \log(3)/\log(2) = 1.58 \ldots \)

Dfn of fd:

ONLY for a perfectly self-similar point set:

\[
\frac{\log(n)}{\log(f)} = \frac{\log(3)}{\log(2)} = 1.58
\]

Intrinsic (‘fractal’) dimension

• Q: fractal dimension of a line?
• A: 1 (= \( \log(2)/\log(2) \))
Intrinsic ('fractal') dimension

- Q: fractal dimension of a line?
- A: 1 (= log(2)/log(2))

- Q: definition for a given set of points?
- A: \( n(n <= r)^{-\alpha} \) (power law: \( y = x^\alpha \))
- \( \text{fd} = \text{slope of } (\log(n(n)) vs \log(r)) \)

- Q: fractal dimension of a plane?
- A: \( n(n <= r)^{-\beta} \)
Intrinsic (‘fractal’) dimension

- Algorithm, to estimate it?
  Notice
- avg nn( <=r ) is exactly
  \[ \frac{\text{tot#pairs}(<=r)}{N} \]
  including ‘mirror’ pairs

Sierpinsky triangle

\[ \log(\#\text{pairs within } <=r) \]
\[ = \log(\text{r}) \]
\[ = \text{‘correlation integral’} \]

Observations:

- Euclidean objects have integer fractal dimensions
  – point: 0
  – lines and smooth curves: 1
  – smooth surfaces: 2
- fractal dimension -> roughness of the periphery
Important properties

- $fd = \text{embedding dimension} \rightarrow \text{uniform pointset}$
- a point set may have several $fd$, depending on scale

1-d

2-d
Important properties

Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
  - More examples and tools
  - Discussion - putting fractals to work!
  - Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots

Problem #1: GIS points

Cross-roads of Montgomery county:
  • any rules?
Solution #1

$$\log(\text{pairs(\text{within } \leq r)})$$

A: self-similarity ->

• $\leftrightarrow$ fractals
• $\leftrightarrow$ scale-free
• $\leftrightarrow$ power-laws

(y = $x^\alpha$, $F=C\cdot r^{\alpha(-2)}$)

• avg#neighbors(<= r) 
  $\approx r^D$

$\log(\text{pairs(\text{within } \leq r)}$$

1.51

Examples: MG county

• Montgomery County of MD (road endpoints)
Examples: LB county

- Long Beach county of CA (road end-points)

Solution#2: spatial d.m.
Galaxies (‘BOPS’ plot - [sigmod2000])

- 1.8 slope
- plateau!
- repulsion!

log(#pairs within ≤ r)
log(r)

log(#pairs)
Spatial d.m.

log(#pairs within <= r )

- 1.8 slope
- plateau!
- repulsion!!
- duplicates

Solution #3: traffic

- disk traces: self-similar:

Solution #3: traffic

- disk traces (80-20 ‘law’ = ‘multifractal’)
80-20 / multifractals

- p : (1-p) in general
- yes, there are dependencies

More on 80/20: PQRS

- Part of ‘self-* storage’ project [Wang+’02]
More on 80/20: PQRS

- Part of ‘self-* storage’ project [Wang+’02]

Solution#3: traffic

Clarification:
- fractal: a set of points that is self-similar
- multifractal: a probability density function that is self-similar

Many other time-sequences are bursty/clustered: (such as?)

Example:
- network traffic

http://repository.cs.vt.edu/lbl-conn-7.tar.Z
Web traffic

- [Crovella Bestavros, SIGMETRICS'96]

1000 sec; 100sec
10sec; 1sec

Tape accesses

# tapes needed, to retrieve n records?
(# days down, due to failures / hurricanes / communication noise...)

The figure shows plots for the three graphical methods...
Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems

More tools and examples
- Discussion - putting fractals to work!
- Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots

A counter-intuitive example

- avg degree is, say 3.3
- pick a node at random
  - guess its degree, exactly (-> “mode”)

A: 1!!
A counter-intuitive example

- avg degree is, say 3.3
- pick a node at random - what is the degree you expect it to have?
  - A: 1!!
  - A': very skewed distr.
- Corollary: the mean is meaningless!
- (and std -> infinity (!))

Rank exponent $R$

- Power law in the degree distribution [SIGCOMM99]
- ![Graph](image)

More tools

- Zipf's law
- Korcak’s law / “fat fractals”
A famous power law: Zipf’s law

- Q: vocabulary word frequency in a document - any pattern?

freq.

aaron  

zoo

log(freq)

log(rank)

Bible - rank vs frequency (log-log)

similarly, in many other languages; for customers and sales volume; city populations etc etc
A famous power law: Zipf’s law

- Zipf distr:
  \[ \text{freq} = \frac{1}{\text{rank}} \]

- Generalized Zipf:
  \[ \text{freq} = \frac{1}{(\text{rank})^a} \]

Olympic medals (Sidney):

\[ y = -0.9676x + 2.3054 \]
\[ R^2 = 0.9458 \]

Olympic medals (Sidney’00, Athens’04):
**TELCO data**

Count-frequency plot of real and synthetic data

- Count of customers
- 'best customer'

**SALES data – store#96**

Count-frequency plot for store no. 96.

- Count of products
- "aspirin" and # units sold

**More power laws: areas – Korcak’s law**

Scandinavian lakes

Any pattern?
More power laws: areas – Korcak’s law

Scandinavian lakes area vs complementary cumulative count (log-log axes)

More power laws: Korcak

Japan islands: area vs cumulative count (log-log axes)

(Korck’s law: Aegean islands)
Korcak’s law & “fat fractals”

How to generate such regions?

Q: How to generate such regions?
A: recursively, from a single region

so far we’ve seen:

- concepts:
  - fractals, multifractals and fat fractals
- tools:
  - correlation integral (= pair-count plot)
  - rank/frequency plot (Zipf’s law)
  - CCDF (Korcak’s law)
so far we’ve seen:

• concepts:
  – fractals, multifractals and fat fractals
• tools:
  – correlation integral (pair-count plot)
  – rank/frequency plot (Zipf’s law)
  – CCDF (Korcak’s law)

Road map

• Motivation – 3 problems / case studies
• Definition of fractals and power laws
• Solutions to posed problems
• More tools and examples
  • Discussion - putting fractals to work!
  • Conclusions – practitioner’s guide
  • Appendix: gory details - boxcounting plots

Other applications: Internet

• How does the internet look like?
Other applications: Internet

• How does the internet look like?
• Internet routers: how many neighbors within $h$ hops?

(reminder: our tool-box:)

• concepts:
  – fractals, multifractals and fat fractals
• tools:
  – correlation integral (= pair-count plot)
  – rank/frequency plot (Zipf’s law)
  – CCDF (Korčak’s law)

Internet topology

• Internet routers: how many neighbors within $h$ hops?

Reachability function: number of neighbors within $r$ hops, vs $r$ (log-log).
Mbone routers, 1995
More power laws on the Internet

degree vs rank, for Internet domains (log-log) [sigcomm99]

log(rank) vs log(degree)

-0.82

log(rank)

log(degree)

-2.2

log(rank)

log(degree)

-0.47

Even more power laws on the Internet

Scree plot for Internet domains (log-log) [sigcomm99]
Fractals & power laws:

appear in numerous settings:

- medical
- geographical / geological
- social
- computer-system related

More apps: Brain scans

- Oct-trees; brain-scans

\[\log_2(\text{octree leaves}) = 2.63 - 2.91 \times \text{level} \]

More apps: Medical images

[Burden et al, SPIE '93]:

- benign tumors: \(fd \sim 2.37\)
- malignant: \(fd \sim 2.56\)
More fractals:

- cardiovascular system: 3 (!)
- lungs: 2.9

Fractals & power laws:

appear in numerous settings:

- medical
- geographical/geological
- social
- computer-system related

More fractals:

- Coastlines: 1.2-1.58

1.0 1.1 1.3
More fractals:

- the fractal dimension for the Amazon river is 1.85 (Nile: 1.4)
  [ems.gphys.unc.edu/nonlinear/fractals/examples.html]
More power laws

- Energy of earthquakes (Gutenberg-Richter law) [simscience.org]

Fractals & power laws:

appear in numerous settings:
- medical
- geographical / geological
- social
- computer-system related

More fractals:

stock prices (LYCOS) - random walks: 1.5
Even more power laws:

• Income distribution (Pareto’s law)
• size of firms
• publication counts (Lotka’s law)

Even more power laws:

library science (Lotka’s law of publication count); and citation counts:
(citeeर.nj.nec.com 6/2001)

Even more power laws:

• web hit counts [w/ A. Montgomery]
Fractals & power laws:

appear in numerous settings:
- medical
- geographical / geological
- social
- computer-system related

Power laws, cont’d

- In- and out-degree distribution of web sites
  [Barabasi], [IBM-CLEVER]

\[
\log \text{indegree} = \text{constant} - \log \text{freq}
\]

from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]
“Foiled by power law”

- [Broder+, WWW’00]

“The anomalous bump at 120 on the x-axis is due a large clique formed by a single spammer.”

Power laws, cont’d

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]
- length of file transfers [Crovella+Bestavros ‘96]
- duration of UNIX jobs [Harchol-Balter]

Even more power laws:

- Distribution of UNIX file sizes
- web hit counts [Huberman]
Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots

What else can they solve?

- separability [KDD’02]
- forecasting [CIKM’02]
- dimensionality reduction [SBBD’00]
- non-linear axis scaling [KDD’02]
- disk trace modeling [Wang+’02]
- selectivity of spatial/multimedia queries
  [PODS’94, VLDB’95, ICDE’00]
- ...

Settings for fractals:

Points; areas (→ fat fractals), eg:
Settings for fractals:

Points; areas, eg:
- cities/stores/hospitals, over earth’s surface
- time-stamps of events (customer arrivals, packet losses, criminal actions) over time
- regions (sales areas, islands, patches of habitats) over space

Settings for fractals:

- customer feature vectors (age, income, frequency of visits, amount of sales per visit)

Some uses of fractals:

- Detect non-existence of rules (if points are uniform)
- Detect non-homogeneous regions (eg., legal login time-stamps may have different fd than intruders’)
- Estimate number of neighbors / customers / competitors within a radius
Multi-Fractals

Setting: points or objects, w/ some value, eg:
- cities w/ populations
- positions on earth and amount of gold/water/oil underneath
- product ids and sales per product
- people and their salaries
- months and count of accidents

Use of multifractals:

- Estimate tape/disk accesses
  - how many of the 100 tapes contain my 50 phonecall records?
  - how many days without an accident?

Use of multifractals

- how often do we exceed the threshold?
Use of multifractals cont’d

• Extrapolations for/from samples

Use of multifractals cont’d

• How many distinct products account for 90% of the sales?

Road map

• Motivation – 3 problems / case studies
• Definition of fractals and power laws
• Solutions to posed problems
• More examples and tools
• Discussion - putting fractals to work!
• Conclusions – practitioner’s guide
• Appendix: gory details - boxcounting plots
Conclusions

- Real data often disobey textbook assumptions (Gaussian, Poisson, uniformity, independence)

Conclusions - cont’d

Self-similarity & power laws: appear in many cases

Bad news:
lead to skewed distributions
(no Gaussian, Poisson, uniformity, independence, mean, variance)

Good news:
- ‘correlation integral’ for separability
- rank/frequency plots
- 80-20 (multifractals)
- (Hurst exponent, strange attractors, renormalization theory, 116)
- $r = 1$

Conclusions

- tool#1: (for points) ‘correlation integral’:
  (#pairs within $r$) vs (distance $r$)
- tool#2: (for categorical values) rank-frequency plot (a’la Zipf)
- tool#3: (for numerical values) CCDF:
  Complementary cumulative distr. function
  (#of elements with value $\geq a$)
**Practitioner’s guide:**

- **tool#1**: #pairs vs distance, for a set of objects, with a distance function (slope = intrinsic dimensionality)

  ![Graph](image1)

  - **internet**: Log(#pairs) vs log(r) with a slope of 2.8
  - **MGcounty**: Log(#pairs(within <= r)) vs log(r) with a slope of 1.51

- **tool#2**: rank-frequency plot (for **categorical attributes**)

  ![Graph](image2)

  - **internet domains**: Log(degree) vs log(rank) with a slope of -0.82
  - **Bible**: Log(freq) vs log(rank) with a slope of 2.8

- **tool#3**: CCDF, for (skewed) **numerical attributes**, eg. areas of islands/lakes, UNIX jobs...

  ![Graph](image3)

  - **scandinavian lakes**: Log(count(>= area)) vs log(area)
Resources:

- Software for fractal dimension
  - http://www.cs.cmu.edu/~christos
  - christos@cs.cmu.edu

Books

- Strongly recommended intro book:
- Classic book on fractals:

References

- [Broder+00] Andrei Broder, Ravi Kumar, Farzin Maghoul1, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener, Graph structure in the web , WWW’00
- M. Crovella and A. Bestavros, Self similarity in World wide web traffic: Evidence and possible causes , SIGMETRICS ’96.
References


References

- [vldb96] Christos Faloutsos, Yossi Matias and Avi Silberschatz, *Modeling Skewed Distributions Using Multifractals and the '80-20 Law'*, Conf. on Very Large Data Bases (VLDB), Bombay, India, Sept. 1996.

References

References

- [icde99] Guido Proietti and Christos Faloutsos, I/O complexity for range queries on region data stored using an R-tree International Conference on Data Engineering (ICDE), Sydney, Australia, March 23-26, 1999

Appendix - Gory details

• Bad news: There are more than one fractal dimensions
  – Minkowski fd; Hausdorff fd; Correlation fd; Information fd
• Great news:
  – they can all be computed fast!
  – they usually have nearby values
Fast estimation of fd(s):

• How, for the (correlation) fractal dimension?
• A: Box-counting plot:

\[
\begin{align*}
\pi & : \text{the percentage (or count) of points in the } i\text{-th cell} \\
r & : \text{the side of the grid}
\end{align*}
\]

Definitions
Fast estimation of fd(s):

- etc; if the resulting plot has a linear part, its slope is the correlation fractal dimension $D_2$

\[
\log(\text{sum}(\pi^2))
\]

Definitions (cont’d)

- Many more fractal dimensions $D_q$ (related to Renyi entropies):

\[
D_q = \frac{1}{q-1} \frac{\partial \log(\sum p_i)}{\partial \log(r)} \quad q \neq 1
\]

\[
D_1 = \frac{\delta \sum p_i \log(p_i)}{\partial \log(r)}
\]

Hausdorff or box-counting fd:

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

\[
D_0 = -\frac{\partial \log(N(r))}{\partial \log(r)}
\]
**Definitions (cont’d)**

- Hausdorff fd:
  \[
  r \quad \log(\text{#non-empty cells})
  \]

**Observations**

- \(q=0\): Hausdorff fractal dimension
- \(q=2\): Correlation fractal dimension (identical to the exponent of the number of neighbors vs radius)
- \(q=1\): Information fractal dimension

**Observations, cont’d**

- in general, the \(D_q\)’s take similar, but not identical, values.
- except for perfectly self-similar point-sets, where \(D_q=D_{q'}\) for any \(q, q'\)
Examples: MG county

- Montgomery County of MD (road endpoints)

Examples: LB county

- Long Beach county of CA (road endpoints)

Conclusions

- many fractal dimensions, with nearby values
- can be computed quickly
  \(O(N)\) or \(O(N \log(N))\)
- (code: on the web)