15-826: Multimedia Databases and Data Mining

Lecture #9: Fractals – examples & algo’s

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:

  – 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

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

More power laws on the Internet

log(rank) vs log(degree)

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

More power laws - internet

- pdf of degrees: (slope: 2.2 )

Log(count)

Log(degree)
Even more power laws on the Internet

Scree plot for Internet domains (log-log) [sigcomm99]

log(1-th eigenvalue)

log(i)

Fractals & power laws:

appear in numerous settings:

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

More apps: Brain scans

- Oct-trees; brain-scans

log(#octants)

2.63 = fd

More apps: Brain scans

- Oct-trees; brain-scans

Log(#octants)
More apps: Medical images

[Burdett et al, SPIE ‘93]:
- benign tumors: fd ~ 2.37
- malignant: fd ~ 2.56

Fractals & power laws:

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

More fractals:

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

More fractals:

- Coastlines: 1.2-1.58

1 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:
- 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]

from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]

“Foiled by power law”

- [Broder+, WWW’00]
“Foiled by power law”

- [Broder+, WWW’00]

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

- length of file transfers [Crovella+Bestavros ‘96]

- duration of UNIX jobs [Harchol-Balter]

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

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:

- 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 (e.g., 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

#bytes

time

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

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

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 >= a )
Practitioner’s guide:

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

  - Log(#pairs) vs log(#pairs(within <= r))

  - **log(#pairs)** vs **log(hops)**: 2.8

  - **log(#pairs(within <= r))** vs **log(r)**: 1.51

- **internet**

- **MGcounty**

- **internet domains**

- **Bible**

Practitioner’s guide:

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

  - Log(rank) vs Log(degree)

  - **log(rank)** vs **log(freq)**: -0.82

- **internet domains**

- **Bible**

Practitioner’s guide:

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

  - Log(count( >= area)) vs Log(area)

  - **area distribution of regions (SL dataset)**: -0.85*x + 10.8

- **scandinavian lakes**

Resources:

- Software for fractal dimension
  - [www.es.cmu.edu/~christos/software.html](http://www.es.cmu.edu/~christos/software.html)
  - And specifically ‘fdnq_h’:
  - [www.es.cmu.edu/~christos/SRC/fdnq_h.zip](http://www.es.cmu.edu/~christos/SRC/fdnq_h.zip)

- Also, in ‘R’: ‘fdim’ package
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.

- [pods94] Christos Faloutsos and Ibrahim Kamel, Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension, PODS, Minneapolis, MN, May 24-26, 1994, pp. 4-13
- [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

References


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 \( \text{fd(s)} \):

- How, for the (correlation) fractal dimension?
- A: Box-counting plot:
  \[
  \log(\text{sum}(\pi^2))
  \]

Definitions

- \( \pi \): the percentage (or count) of points in the \( i \)-th cell
- \( r \): the side of the grid

Fast estimation of \( \text{fd(s)} \):

- compute \( \text{sum}(\pi^2) \) for another grid side, \( r' \)

- etc; if the resulting plot has a linear part, its slope is the correlation fractal dimension \( D_2 \)
Definitions (cont’d)

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

$$D_q = \frac{1}{q-1} \frac{\partial \log \left( \sum p_i^q \right)}{\partial \log(r)} \quad q \neq 1$$

$$D_1 = \frac{\partial \sum p_i \log(p_i)}{\partial \log(r)}$$

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 Dq’s take similar, but not identical, values.
- except for perfectly self-similar point-sets, where Dq=Dq’ 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:
  - www.cs.cmu.edu/~christos/SRC/fdnq_h.zip
  - Or `R` (‘fdim’ package)