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.

Outline

Goal: ‘Find similar / interesting things’

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

Recommended Material

optional, but very useful:

  - Chapter 10: boxcounting method
  - Chapter 1: Sierpinski triangle
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.

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, …)

What is a fractal?

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

Dimensionality??

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

Definitions (cont’d)

- Paradox: Infinite perimeter ; Zero area!
- ‘Dimensionality’: between 1 and 2
- actually: \( \log(3)/\log(2) = 1.58... \)
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: dfn for a given set of points?

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
**Intrinsic (‘fractal’) dimension**

- **Q: fractal dimension of a line?**
  - **A:** nn \((<= r) \sim r^1\)
  ('power law': \(y=x^a\))

- **Q: fd of a plane?**
  - **A:** nn \((<= r) \sim r^2\)
  fd = slope of \((\log(nn) vs \log(r))\)

---

**Local fractal dimension of point ‘P’?**

- **A:** nn\(_P\) \((<= r) \sim r^1\)
  - If this equation holds for several values of \(r\),
  - Then, the **local fractal dimension** of point P:
    - Local fd = exp = 1

---

**Global fractal dimension?**

- **A:** if
  - sum\(\sum_{all P} [\text{nn}\(_P\) (<= r) \sim r^1]\)
  - Then: exp = global fd.
  - Or simply the f.d.
Intrinsic (‘fractal’) dimension

• Algorithm, to estimate it?

Notice

• $\sum_{all \ P} \{ nn_{P} (\leq r) \}$ is exactly $\text{tot}\#\text{pairs}(\leq r)$

including ‘mirror’ pairs

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 = embedding dimension -> uniform pointset
• a point set may have several fd, depending on scale

Sierpinsky triangle

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

$\log( r )$

1.58

== ‘correlation integral’
Important properties

• \( fd = \) embedding dimension \( \rightarrow \) uniform pointset
• a point set may have several \( fd \), depending on scale

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

A: self-similarity ->
• <= fractals
• <= scale-free
• <= power-laws
  \(y = x^a, F = C r^{-2}\)
• avg#neighbors(<= r )
  \(= r^D\)

Solution #1

log(#pairs(within <= r ))

A: self-similarity

• avg#neighbors(<= r )
  \(~ r^{1.51}\)

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])

Spatial d.m.

- 1.8 slope
- plateau!
- repulsion!

log(#pairs within ≤r)

log(r)
Spatial d.m.

Heuristic on choosing # of clusters

log(#pairs within <=r)

- 1.8 slope
- plateau!
- repulsion!

Solution #3: traffic

- disk traces: self-similar:
- #bytes vs. time
Solution #3: traffic

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

\[ \begin{array}{ccc}
\text{time} & \text{20\%} & \text{80\%} \\
\text{#bytes} & & \\
\end{array} \]

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

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...)

Tape accesses

# tapes retrieved

50-50 = Poisson

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 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]

  internet domains

  log(rank) vs. log(degree)

  att.com
  ibm.com

  -0.82

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

A famous power law: Zipf’s law

- 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: freq = 1/ rank
- generalized Zipf: freq = 1 / (rank)^a
Olympic medals (Sydney):

\[ \log(\text{#medals}) \]

\[ y = 0.3676x + 2.69 \]
\[ R^2 = 0.9458 \]

Olympic medals (Sydney’00, Athens’04):

\[ \log(\text{rank}) \]

\[ \log(\text{#medals}) \]

TELCO data

count of customers

'best customer'

Count-frequency plot of real and synthetic data

SALES data – store#96

count of products

"aspirin"

# units sold

Count-frequency plot for store no. 96.
More power laws: areas – Korcak’s law

Scandinavian lakes
Any pattern?

More power laws: areas – Korcak’s law

log(count( >= area))

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

More power laws: Korcak

log(count( >= area))

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

(Korcak’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)
Next:

- More examples / applications
- Practitioner’s guide
- Box-counting: fast estimation of correlation integral