15-826: Multimedia Databases and Data Mining

Lecture #8: 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:

• Manfred Schroeder *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*
  – 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

• What patterns are in real $k$-dim points?

Conclusions

• What patterns are in real $k$-dim points?
• Self-similarity ( = fractals -> power laws)
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!

Dimensionality??

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:

= \log(n)/\log(f) = \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))!

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: \( \text{nn} ( \leq r ) \sim r^1 \) (‘power law’: \( y = x^a \))

- Q: fd of a plane?
- A: \( \text{nn} ( \leq r ) \sim r^2 \)

\( \text{fd} = \text{slope of (log(nn) vs log(r))} \)

---

EXPLANATIONS

Intrinsic (‘fractal’) dimension

- **Local** fractal dimension of point ‘P’?
- A: \( \text{nn}_P ( \leq r ) \sim r^1 \)

- If this equation holds for several values of \( r \),
- Then, the **local fractal dimension** of point P:

\[ \text{Local \ fd} = \exp = 1 \]
Intrinsic (‘fractal’) dimension

- **Local** fractal dimension of point ‘A’?
- A: \( \text{nn}_P ( \leq r ) \sim r^1 \)
- If this is true for all points of the cloud
- Then the exponent is the global f.d.
- Or simply the f.d.

Intrinsic (‘fractal’) dimension

- **Global** fractal dimension?
- A: if
- \( \sum_{\text{all } P} [ \text{nn}_P ( \leq r ) ] \sim r^1 \)
- Then: \( \exp = \text{global f.d.} \)
- If this is true for all points of the cloud
- Then the exponent is the global f.d.
- Or simply the f.d.
Intrinsic (‘fractal’) dimension

- **Local** fractal dimension for Sierpinski triangle?

- 2x radius, 3x points
- \( n = r^{\log_3\log_2} \)
**Intrinsic (‘fractal’) dimension**

- Algorithm, to estimate it?

Notice
- \( \sum_{all \mathcal{P}} [ \text{nn}_P (\leq r) ] \) is exactly \( \text{tot#pairs}(\leq r) \)
  
  including ‘mirror’ pairs

---

**Sierpinsky triangle**

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

\[ \log( r ) \]

== ‘correlation integral’

1.58
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
Important properties

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

2-d

1-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(#pairs(within <= r))

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

\[ \text{SLOPE} = 1.51 \]
Solution #1

log(#pairs(within $\leq r$))

A: self-similarity
- $\text{avg#neighbors}(\leq r) \sim r^{1.51}$

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])
Solution#2: spatial d.m.

log(#pairs within \(<=r\))

- 1.8 slope
- plateau!
- repulsion!

Spatial d.m.

log(#pairs within \(<=r\))

- 1.8 slope
- plateau!
- repulsion!
Spatial d.m.

Heuristic on choosing # of clusters

Spatial d.m.

log(#pairs within \( \leq r \))

- 1.8 slope
- plateau!
- repulsion!
Spatial d.m.

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

Solution #3: traffic

- disk traces: self-similar:

#bytes
Solution #3: traffic

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

80-20 / multifractals
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/clumped: (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...)

Tape accesses

50-50 = Poisson

# tapes retrieved

# qual. records

real
**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”)

<table>
<thead>
<tr>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg: 3.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg: 3.3</td>
</tr>
</tbody>
</table>
A counter-intuitive example

- avg degree is, say 3.3
- pick a node at random
  - guess its degree, exactly (-> “mode”)
- A: 1!!

Corollary: the mean is meaningless!

- (and std -> infinity (!))
Rank exponent $R$

- Power law in the degree distribution
  [SIGCOMM99]

internet domains

<table>
<thead>
<tr>
<th>log(rank)</th>
<th>log(degree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>1000</td>
<td>10000</td>
</tr>
</tbody>
</table>

- att.com
- ibm.com

$R = -0.82$

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)

log(freq)

"a"

"the"

log(rank)
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

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

\[ \log(\text{#medals}) = -0.9676x + 2.3054 \]

\[ R^2 = 0.9458 \]

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

\[ \log(\text{rank}) = \log(\text{#medals}) \]
TELCO data

Count-frequency plot of real and synthetic data

SALES data – store#96

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

Scandinavian lakes
Any pattern?

Scandinavian lakes
area vs complementary cumulative count (log-log axes)
More power laws: Korcak

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

log(count( >= area)) vs log(area)

(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 (Korčak’s law)
Next:

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

Problem

- What patterns are in real $k$-dim points?
Conclusions

• What patterns are in real $k$-dim points?
• Self-similarity ( = fractals -> power laws)