15-826: Multimedia Databases and Data Mining

Lecture #12: Fractals - case studies Part III
(regions, quadtrees, knn queries)

C. Faloutsos

Must-read Material

• Alberto Belussi and Christos Faloutsos,
  Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension
  Proc. of VLDB, p. 299-310, 1995

Optional Material

Optional, but very useful: Manfred Schroeder
Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise
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

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 [Guede+]
  - nn queries [Belussi+]
‘Fat’ fractals & R-tree performance on region data

• Problem [Proietti+, ’99]
• Given
  – N (# of data regions)
• estimate how many of them will qualify for the average range query (q_1 \times q_2 \times \ldots \times q_E)

Of course, we need more info
Q: what?

R-tree performance on region data

A: the distributions of their sizes

Q: do we also need some info about the locations?

R-tree performance on region data

A: the distributions of their sizes

Q: do we also need some info about the locations?
A: no (not for range queries)
R-tree performance on region data
A: the distributions of their sizes
Q: what exactly would we need?
A: for self-similar regions (~‘fat’ fractals), we just need the slope of the Korcak law!
(and the total area) [Proietti+]

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)

B (patchiness exponent)

R-tree performance on regions

- Once we know ‘B’ (and the total area)
- we can second-guess the individual sizes
- and then apply the [Pagel+93] formula
- Bottom line:

R-tree performance on regions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAKES</td>
<td>816</td>
<td>75,910</td>
<td>0.85</td>
</tr>
<tr>
<td>ISLANDS</td>
<td>470</td>
<td>136,893</td>
<td>0.60</td>
</tr>
<tr>
<td>REGIONS</td>
<td>757</td>
<td>190,526</td>
<td>0.70</td>
</tr>
</tbody>
</table>
R-tree performance on regions

sel. error

query side

‘Fat’ fractals - observation

\[ B = \frac{D_{H}}{d} \]

B: patchiness exp; d: embedding dim

\[ D_{H} \]: Hausdorff of periphery

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Exp</th>
<th>B</th>
<th>Exp - 1/B</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAKES</td>
<td>1.74</td>
<td>0.85</td>
<td>0.06</td>
</tr>
<tr>
<td>ISLANDS</td>
<td>1.33</td>
<td>0.45</td>
<td>0.03</td>
</tr>
<tr>
<td>REGIONS</td>
<td>1.48</td>
<td>0.72</td>
<td>0.05</td>
</tr>
<tr>
<td>Angesd, India</td>
<td>1.84</td>
<td>0.57</td>
<td>0.01</td>
</tr>
<tr>
<td>1000 M trees</td>
<td>1.30</td>
<td>0.46</td>
<td>0.06</td>
</tr>
<tr>
<td>Tully phots</td>
<td>1.23</td>
<td>0.62</td>
<td>0.49</td>
</tr>
<tr>
<td>Whole World</td>
<td>1.53</td>
<td>0.65</td>
<td>0.00</td>
</tr>
<tr>
<td>Cypress vegetation</td>
<td>0.97</td>
<td>1.03</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Copyright: C. Faloutsos (2011)
‘Fat’ fractals

• intuition behind $B = D_H / d$?

• A: consider ‘flooding’:

[Graph of 'flooding']
Conclusions

- ‘Fat’ fractals model regions well
- patchiness exp.: $B = D_{H}/ d$
- can help us estimate selectivities

Indexing - Detailed outline

- fractals
  - intro
  - applications
    - disk accesses for R-trees (range queries)
    - dimensionality reduction
    - selectivity in M-trees
    - “fat fractals”
  - quad-tree analysis [Gaede+96]
  - nn queries [Belussi+]

Fractals and Quadtrees

- Problem: how many quadtree nodes will we need, to store a region in some level of approximation? [Gaede+96]
Fractals and Quadtrees

• I.e.:

Datasets:

Franconia

Brain Atlas
Fractals and Quadtrees

- Hint:
  - assume that the boundary is self-similar, with a given fd
  - how will the quad-tree (oct-tree) look like?

Fractals and Quadtrees

Let $p_g(i)$ the prob. to find a gray node at level $i$.
If self-similar, what can we say for $p_g(i)$?
Fractals and Quadtrees

Let $p_g(i)$ the prob. to find a gray node at level $i$.
If self-similar, what can we say for $p_g(i)$?

A: $p_g(i) = p_g = \text{constant}$

Fractals and Quadtrees

Assume only ‘gray’ and ‘white’ nodes (ie., no volume’)
Assume that $p_g$ is given - how many gray nodes at level $i$?

A: 1 at level 0;

$4p_g$
$(4p_g)^2$ $(4p_g)^3$

... ...

$(4p_g)^i$
Fractals and Quadtrees

• I.e.:

\[
\text{log}(\text{# of quadtree 'blocks')} \quad ? \quad \log[(4*p_g)^i]
\]

level of quadtree

Fractals and Quadtrees

• Conclusion: Self-similarity leads to easy and accurate estimation

\[
\text{log}_5(\text{#blocks}) \quad \text{level}
\]
Fractals and Quadtrees

- Conclusion: Self-similarity leads to easy and accurate estimation

\[ \log_2(\#\text{blocks}) \]

level
Fractals and Quadtrees

- Final observation: relationship between $p_g$ and fractal dimension?

- A: very close:
  
  $(4^i p_g)^r = \# \text{ of gray nodes at level } i = \# \text{ of Hausdorff grid-cells of side } (1/2)^i = r$

  Eventually: $D_H = 2 + \log_2(4^i p_g)$

  and, for E-d spaces: $D_H = E + \log_2(p_g)$

Sanity check:
- point in 2-d: $D_H = 0 \quad p_g = ??$
- line in 2-d: $D_H = 1 \quad p_g = ??$
- plane in 2-d: $D_H = 2 \quad p_g = ??$
- point in 3-d: $D_H = 0 \quad p_g = ??$
Fractals and Quadtrees

for E-d spaces: \( D_H = E + \log_2(p_g) \)

Sanity check:
- point in 2-d: \( D_H = 0 \) \( p_g = 1/4 \)
- line in 2-d: \( D_H = 1 \) \( p_g = 1/2 \)
- plane in 2-d: \( D_H = 2 \) \( p_g = 1 \)
- point in 3-d: \( D_H = 0 \) \( p_g = 1/8 \)

Final conclusions:
- self-similarity leads to estimates for # of z-values = # of quadtree/oct-tree blocks
- close dependence on the Hausdorff fractal dimension of the boundary

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+]
- nn queries [Belussi+]
NN queries

- Q: in NN queries, what is the effect of the shape of the query region? [Belussi+95]

- that is, for $L_2$, and self-similar data:

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

- $L_1$, $L_{\infty}$?

\[
\log(\text{#pairs-within}(\leq d))
\]
NN queries

- Q: What about $L_1$, $L_{\infty}$?
- A: Same slope, different intercept

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

$log(d)$  $L_2$

$D_2$

NN queries

- Q: What about $L_1$, $L_{\infty}$?
- A: Same slope, different intercept

$log(\#\text{neighbors})$

$log(d)$

NN queries

- Q: What about the intercept? I.e., what can we say about $N_2$ and $N_{\infty}$

$N_2$ neighbors

$N_{\infty}$ neighbors

$L_2$

$L_{\infty}$

volume: $V_2$

volume: $V_{\infty}$

SKIP
NN queries

- Consider sphere with volume $V_{\text{inf}}$ and $r'$ radius

$$N_2' = N_{\text{inf}}$$ (since shape does not matter)

and finally:

$$\left( \frac{N_2}{N_{\text{inf}}} \right)^{1/D_2} = \left( \frac{V_2}{V_{\text{inf}}} \right)^{1/E}$$
NN queries
Conclusions: for self-similar datasets
• Avg # neighbors: grows like $(\text{distance})^{D_x}$, regardless of query shape (circle, diamond, square, e.t.c.)

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+]
    • nn queries [Belussi+]
  – Conclusions

Fractals - overall conclusions
• self-similar datasets: appear often
• powerful tools: correlation integral, NCDF, rank-frequency plot
• intrinsic/fractal dimension helps in
  – estimations (selectivities, quadtrees, etc)
  – dim. reduction / dim. curse
• (later: can help in image compression...)
References