15-826: Multimedia Databases and Data Mining

Lecture #11: Fractals: M-trees and dim. curse (case studies – Part II)

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

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 [Gaede+]
What else can they solve?

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

Metric trees - analysis

- Problem: How many disk accesses, for an M-tree?
- Given:
  - \( N \) (# of objects)
  - \( C \) (fanout of disk pages)
  - \( r \) (radius of range query - BIASED model)

- NOT ENOUGH - what else do we need?
Metric trees - analysis

• A: something about the distribution

[Ciaccia, Patella, Zezula, PODS98]: assumed that the distance distribution is the same, for every object:

F1(d) = Prob(an object is within d from object #1)
= F2(d) = ... = F(d)
Metric trees - analysis

• A: something about the distribution
• Given our ‘fractal’ tools, we could try them - which one?

Metric trees - analysis

• A: something about the distribution
• Given our ‘fractal’ tools, we could try them - which one?
• A: Correlation integral [Traina+, ICDE2000]

Metric trees - analysis

English dictionary

Portuguese dictionary

log(#pairs) log(#pairs)

log(d) log(d)
Metric trees - analysis

Divina Comedia

Eigenfaces

<table>
<thead>
<tr>
<th>log(#pairs)</th>
<th>log(#pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log(d) )</td>
<td>( \log(d) )</td>
</tr>
</tbody>
</table>

Metric trees - analysis

<table>
<thead>
<tr>
<th>Data Set</th>
<th>N (O. Objects)</th>
<th>Dimension</th>
<th>Distance Method</th>
<th>Distance Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Metric dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goglist</td>
<td>20,453</td>
<td>NA</td>
<td>( L_1 )</td>
<td>4.783</td>
</tr>
<tr>
<td>Divina Comedia</td>
<td>12,706</td>
<td>NA</td>
<td>( L_1 )</td>
<td>4.827</td>
</tr>
<tr>
<td>Descendants</td>
<td>18,719</td>
<td>NA</td>
<td>( L_1 )</td>
<td>3.286</td>
</tr>
<tr>
<td>Portuguese</td>
<td>21,673</td>
<td>NA</td>
<td>( L_1 )</td>
<td>4.688</td>
</tr>
<tr>
<td>Study</td>
<td>1,019</td>
<td>NA</td>
<td>Nat distance</td>
<td>0.621</td>
</tr>
<tr>
<td>Real vector dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MG1</td>
<td>13,550</td>
<td>2</td>
<td>( L_1 )</td>
<td>1.792</td>
</tr>
<tr>
<td>Vigenbae</td>
<td>13,986</td>
<td>10</td>
<td>( L_1 )</td>
<td>5.567</td>
</tr>
<tr>
<td>Synthetic dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berndt</td>
<td>9,841</td>
<td>2</td>
<td>( L_1 )</td>
<td>1.784</td>
</tr>
<tr>
<td>2D Line</td>
<td>20,000</td>
<td>2</td>
<td>( L_1 )</td>
<td>0.980</td>
</tr>
<tr>
<td>Cortada 2D</td>
<td>10,000</td>
<td>2</td>
<td>( L_1 )</td>
<td>1.967</td>
</tr>
</tbody>
</table>
So, what is the # of disk accesses, for a node of radius $r_d$, on a query of radius $r_q$?

$A \sim (r_d + r_q)^D$
Accuracy of selectivity formulas

Fast estimation of \( D \)

- Normally, \( D \) takes \( O(N^2) \) time
- Anything faster? suppose we have already built an M-tree

Fast estimation of \( D \)

- Hint:
Fast estimation of $D$

- Hint:
  
  \[ r_1^D \cdot C = r_2^D \]
  
  \[ D \sim \frac{\log(C)}{\log(r_2/r_1)} \]

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+]

Dim. curse revisited

- (Q: how serious is the dim. curse, e.g.:)
- Q: what is the search effort for k-nn?
  - given N points, in E dimensions, in an R-tree, with k-nn queries ("biased" model)

[Pagel, Korn + ICDE 2000]
(Overview of proofs)

- assume that your points are uniformly distributed in a $d$-dimensional manifold (= hyper-plane)
- derive the formulas
- substitute $d$ for the fractal dimension

Reminder: Hausdorff Dimension ($D_0$)

- $r =$ side length (each dimension)
- $B(r) =$ # boxes containing points $\propto r^{D_0}$

$$
\begin{align*}
r & = 1/2 & B & = 2 \\
\log r & = -1 & \log B & = 1 \\
r & = 1/4 & B & = 4 \\
\log r & = -2 & \log B & = 2 \\
r & = 1/8 & B & = 8 \\
\log r & = -3 & \log B & = 3
\end{align*}
$$

Reminder: Correlation Dimension ($D_2$)

- $S(r) = \sum p_i^2$ (squared % pts in box) $\propto r^{D_2}$
  $\propto$ #pairs (within $\leq r$)

$$
\begin{align*}
r & = 1/2 & S & = 1/2 \\
\log r & = -1 & \log S & = -1 \\
r & = 1/4 & S & = 1/4 \\
\log r & = -2 & \log S & = -2 \\
r & = 1/8 & S & = 1/8 \\
\log r & = -3 & \log S & = -3
\end{align*}
$$
Observation #1

• How to determine avg MBR side \( l \)?
  – \( N \) = #pts, \( C \) = MBR capacity

Hausdorff dimension: \( B(r) \propto r^{D_0} \)
\[
B(l) = \frac{N}{C} = l^{-D_0} \implies l = \left( \frac{N}{C} \right)^{1/D_0}
\]

Observation #2

• \( k \)-NN query \( \rightarrow \) \( \varepsilon \)-range query
  – For \( k \) pts, what radius \( \varepsilon \) do we expect?

Correlation dimension: \( S(r) \propto r^{D_2} \)
\[
S(\varepsilon) = \frac{k}{N - 1} = (2\varepsilon)^{D_2}
\]

Observation #3

• Estimate avg # query-sensitive anchors:
  – How many expected \( q \) will touch avg page?
  – Page touch: \( q \) stabs \( \varepsilon \)-dilated MBR(\( p \))
Asymptotic Formula

- $k$-NN page accesses as $N \to \infty$
  - $C$ = page capacity
  - $D$ = fractal dimension ($=D_0 \sim D_2$)

$$P_{all}^{L_n}(k) = \sum_{j=0}^{h} \left\{ \frac{1}{C^{h-j}} + \left[ 1 + \left( \frac{k}{C^{h-j}} \right)^{1/D} \right]^D \right\}$$

Asymptotic Formula

$$P_{all}^{L_n}(k) \approx \sum_{j=0}^{h} \left\{ \frac{1}{C^{h-j}} + \left[ 1 + \left( \frac{k}{C^{h-j}} \right)^{1/D} \right]^D \right\}$$

- NO mention of the embedding dimensionality!!
- Still have dim. curse, but on f.d. $D$

Synthetic Data

- plane
  - $D_0 = D_2 = 2$
  - embedded in $E$-space
  - $N = 100K$
- manifold
  - $E = 8$
  - $D_0 = D_2$ varies from 1-6
  - line, plane, etc. (in 8-d)
Accuracy of $L_\infty$ Formula

Embedding Dimension

Intrinsic Dimensionality
Non-Euclidean Data Set

<table>
<thead>
<tr>
<th>E</th>
<th>unif</th>
<th>ind</th>
<th>fractal</th>
<th>leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.49</td>
<td>2.53</td>
<td>4.72±1.81</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>847.26</td>
<td>2.53</td>
<td>6.42±2.11</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>all</td>
<td>2.53</td>
<td>7.76±4.12</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>all</td>
<td>2.53</td>
<td>6.15±2.82</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>all</td>
<td>2.53</td>
<td>5.64±2.32</td>
<td></td>
</tr>
</tbody>
</table>

Conclusions

- Worst-case theory is **over-pessimistic**
- High dimensional data can exhibit good performance if **correlated, non-uniform**
- Many real data sets are **self-similar**
- Determinant is **intrinsic** dimensionality
  - multiple fractal dimensions ($D_0$ and $D_2$)
  - indication of how far one can go

References