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
  • dim. curse revisited
  • quad-tree analysis [Gaede+]
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]

Indexing - Detailed outline

- fractals
  - intro
  - applications
    - disk accesses for R-trees (range queries)
    - dimensionality reduction
- dim. curse revisited
- quad-tree analysis [Gaede+]

Dimensionality ‘curse’

- Q: What is the problem in high-d?
Dimensionality ‘curse’

• Q: What is the problem in high-d?
• A: indices do not seem to help, for many queries (eg., k-nn)
  – in high-d (& uniform distributions), most points are equidistant -> k-nn retrieves too many near-neighbors
  – [Yao & Yao, ’85]: search effort ~ O( N^{(1-1/d) } )

Dimensionality ‘curse’

• (counter-intuitive, for db mentality)
• Q: What to do, then?

Dimensionality ‘curse’

• A1: switch to seq. scanning
• A2: dim. reduction
• A3: consider the ‘intrinsic’/fractal dimensionality
• A4: find approximate nn
Dimensionality ‘curse’

- A1: switch to seq. scanning
  - X-trees [Kriegel+, VLDB 96]
  - VA-files [Schek+, VLDB 98], ‘test of time’ award

- A2: dim. reduction

- A3: consider the ‘intrinsic’/fractal dimensionality

- A4: find approximate nn

---

**Dim. reduction**

a.k.a. feature selection/extraction:

- SVD (optimal, to preserve Euclidean distances)
- random projections
- using the fractal dimension [Traina+ SBBD2000]
Singular Value Decomposition (SVD)

- SVD (~LSI ~ KL ~ PCA ~ spectral analysis...)
  - LSI: S. Dumais; M. Berry
  - KL: eg, Duda+Hart
  - PCA: eg., Jolliffe
  - MANY more details: soon

Random projections

- random projections(Johnson-Lindenstrauss thm [Papadimitriou+ PODS98])
  - pick ‘enough’ random directions (will be ~orthogonal, in high-d!!)
  - distances are preserved probabilistically, within epsilon
  - (also, use as a pre-processing step for SVD [Papadimitriou+ PODS98]
Dim. reduction - w/ fractals

- Main idea: drop those attributes that don’t affect the intrinsic (‘fractal’) dimensionality
  [Traina+, SBBD 2000]

Dim. reduction - w/ fractals

global FD=1

Dimensionality ‘curse’

- A1: switch to seq. scanning
- A2: dim. reduction
- A3: consider the ‘intrinsic’/fractal dimensionality
  ➤ A3: consider the ‘intrinsic’/fractal dimensionality
- A4: find approximate nn
**Intrinsic dimensionality**
- before we give up, compute the intrinsic dim.:
- the lower, the better... [Pagel+, ICDE 2000]
- more details: in a few foils

intr. d = 2 [Diagram] intr. d = 1

**Dimensionality ‘curse’**
- A1: switch to seq. scanning
- A2: dim. reduction
- A3: consider the ‘intrinsic’/fractal dimensionality
  - A4: find approximate nn

**Approximate nn**
- [Arya + Mount, SODA93], [Patella+ ICDE 2000]
- Idea: find k neighbors, such that the distance of the k-th one is guaranteed to be within epsilon of the actual.
Dimensionality ‘curse’

- A1: switch to seq. scanning
- A2: dim. reduction
- A3: consider the ‘intrinsic’/fractal dimensionality
- A4: find approximate nn

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 $\propto r^{D_0}$

\[
\begin{align*}
\log r &= -1 \\
\log B &= 1
\end{align*}
\]

\[
\begin{align*}
\log r &= -2 \\
\log B &= 2
\end{align*}
\]

\[
\begin{align*}
\log r &= -3 \\
\log B &= 3
\end{align*}
\]

Reminder: Correlation Dimension ($D_2$)

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

\[
\begin{align*}
\log r &= -1 \\
\log S &= -1
\end{align*}
\]

\[
\begin{align*}
\log r &= -2 \\
\log S &= -2
\end{align*}
\]

\[
\begin{align*}
\log r &= -3 \\
\log S &= -3
\end{align*}
\]

Observation #1

- How to determine avg MBR side $l$?
  - $N$ = #pts, $C$ = MBR capacity

Hausdorff dimension: $B(r) \propto r^{D_0}$

\[
B(l) = \frac{N}{C} = l^{-D_0} \Rightarrow l = (\frac{N}{C})^{\frac{1}{D_0}}
\]
Observation #2

- $k$-NN query $\rightarrow$ $\varepsilon$-range query
  - For $k$ pts, what radius $\varepsilon$ do we expect?

Correlation dimension: $S(r) \propto r^D$

$$S(\varepsilon) = \frac{k}{N - 1} = (2\varepsilon)^D$$

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 $N \rightarrow \infty$
  - $C =$ page capacity
  - $D =$ fractal dimension ($=D_0 \sim D_2$)

$$p_{all}^{L_{\infty}}(k) = \sum_{j=0}^{h} \left\{ \frac{1}{C^{h-j}} + \left[ 1 + \left( \frac{k}{C^{h-j}} \right)^{1/D} \right]^D \right\}$$
Asymptotic Formula

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

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

Embedding Dimension

Conclusions

- Dimensionality ‘curse’:
  – for high-d, indices slow down to \( \sim O(N) \)
- If the intrinsic dim. is low, there is hope
- otherwise, do seq. scan, or sacrifice accuracy (approximate nn)
Conclusions – cont’d

• 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

  ANN library: http://www.cs.umd.edu/~mount/ANN/

• Berchtold, S., D. A. Keim, et al. (1996). The X-tree: An Index Structure for High-Dimensional Data. VLDB, Mumbai (Bombay), India.
References cnt’d


References cnt’d


References cnt’d