CMU SCS

15-826: Multimedia Databases and Data Mining

Lecture #30: Conclusions

C. Faloutsos

CMU SCS

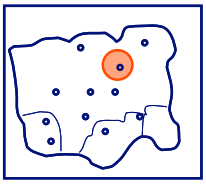
Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - Graphs
- Data Mining

CMU SCS

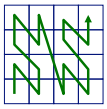
Indexing - similarity search

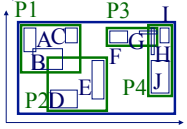


CMU SCS

Indexing - similarity search

- R-trees
- z-ordering / hilbert curves
- M-trees
- (DON'T FORGET …)

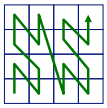


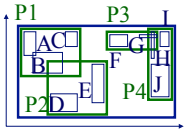


CMU SCS

Indexing - similarity search

- R-trees
- z-ordering / hilbert curves
- M-trees
- **beware of high intrinsic dimensionality**





CMU SCS

Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - Graphs
- Data Mining

CMU SCS

Text searching

- 'find all documents with word *bla*'

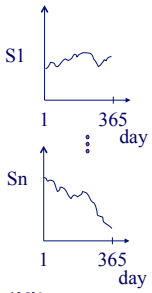
CMU SCS

Text searching

- Full text scanning ('grep')
- Inversion (B-tree or hash index)
- (signature files)
- Vector space model
  - Ranked output
  - Relevance feedback
- String editing distance (-> dynamic prog.)

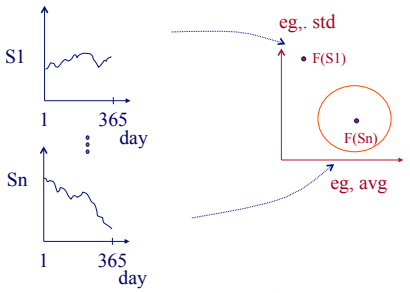
CMU SCS

Multimedia indexing



CMU SCS

'GEMINI' - Pictorially



CMU SCS

Multimedia indexing

- Feature extraction for indexing (GEMINI)
  - Lower-bounding lemma, to guarantee no false alarms
- MDS/FastMap

CMU SCS

Outline

Goal: 'Find similar / interesting things'

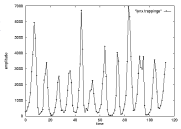
- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - Graphs
- Data Mining

CMU SCS

Time series & forecasting

Goal: given a signal (eg., sales over time and/ or space)

Find: patterns and/or compress

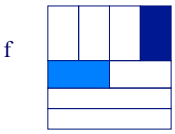
count  lynx caught per year

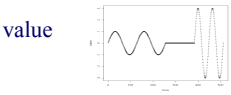
year

CMU SCS

Wavelets

- Q: baritone/silence/soprano - DWT?





CMU SCS

Time series + forecasting

- Fourier; Wavelets
- Box/Jenkins and AutoRegression
- non-linear/chaotic forecasting (fractals again)
  - 'Delayed Coordinate Embedding' ~ nearest neighbors

CMU SCS

Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
  - Points
  - Text
  - Time sequences; images etc
  - Graphs
- Data Mining

CMU SCS

Graphs

- Real graphs: surprising patterns
  - ??

CMU SCS

Graphs

- Real graphs: surprising patterns
  - '**six degrees**'
  - **Skewed** degree distribution ('rich get richer')
  - Super-linearities (2x nodes -> 3x edges)
  - Diameter: **shrinks** (!)
  - Might have **no** good cuts

CMU SCS

Outline

Goal: 'Find similar / interesting things'

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

CMU SCS

Data Mining - DB

CMU SCS

Data Mining - DB

- Association Rules ('diapers' -> 'beer')
- [OLAP (DataCubes, roll-up, drill-down)]
- [Classifiers]

CMU SCS

Taking a step back:

We saw some fundamental, recurring concepts and tools:

CMU SCS

Powerful, recurring tools

- Fractals/ self similarity
  - Zipf, Korcak, Pareto's laws
  - intrinsic dimension (Sierpinski triangle)
  - correlation integral
  - Barnsley's IFS compression
  - (Kronecker graphs)

CMU SCS

Powerful, recurring tools

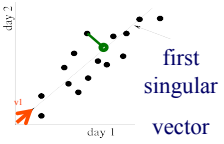
- Fractals/ self similarity
  - Zipf, Korcak, P...
  - ...
  - ...mpression
  - (...ronecker graphs)

•'Take logarithms'
•AVOID 'avg'

CMU SCS

Powerful, recurring tools

- SVD (optimal L2 approx)
  - LSI, KL, PCA, 'eigenSpokes', (& in ICA)
  - HITS (PageRank)



first singular vector

CMU SCS

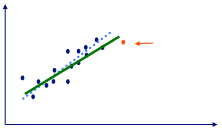
Powerful, recurring tools

- Discrete Fourier Transform
- Wavelets

CMU SCS

Powerful, recurring tools

- Matrix inversion lemma
  - Recursive Least Squares
  - Sherman-Morrison(-Woodbury)



CMU SCS

Summary

- **fractals / power laws** probably lead to the most startling discoveries ('the mean may be meaningless')
- **SVD**: behind PageRank/HITS/tensors/…
- **Wavelets**: Nature seems to prefer them
- **RLS**: seems to achieve the impossible
- (approximate counting – NOT in the exam)

CMU SCS

Thank you!

- Feel free to contact me:
  - christos@cs GHC 8019
- Reminder: faculty course eval's:
  - www.cmu.edu/hub/fce/
- Final: Thu. Dec 13, 5:30-8:30pm @DH A302
  - Double-check at
  - www.cmu.edu/hub/docs/final-exams.pdf
- Have a great break!