CMU SCS

15-826: Multimedia Databases and Data Mining

Conclusions

C. Faloutsos

CMU SCS

Outline

Goal: 'Find similar / interesting things'

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

CMU SCS

Indexing - similarity search

CMU SCS

Indexing - similarity search

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

CMU SCS

Multimedia indexing

CMU SCS

Multimedia indexing

- Feature extraction for indexing
- MDS/FastMap

CMU SCS

Time series + forecasting

CMU SCS

Time series + forecasting

- Fourier; Wavelets
- Box/Jenkins and AutoRegression
- non-linear/chaotic forecasting (fractals again)

CMU SCS

Data Mining - stat

CMU SCS

Data Mining - stat

- Chi-square hypothesis testing
- Entropy

CMU SCS

Data Mining - AI

CMU SCS

Data Mining - AI

- Classification trees

CMU SCS

Data Mining - DB

CMU SCS

Data Mining - DB

- OLAP / DataCubes
- Classification & clustering
- Association Rules

CMU SCS

Taking a step back:

We saw some fundamental, recurring concepts and tools:

CMU SCS

Powerful, recurring tools

- SVD (optimal L2 approx)
  - LSI, KL, PCA
- (ICA / Blind Source Separation)

CMU SCS

Powerful, recurring tools

- Discrete Fourier Transform
- Wavelets

CMU SCS

Powerful, recurring tools

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

CMU SCS

Powerful, recurring tools

- Matrix inversion lemma
  - Recursive Least Squares
  - Sherman-Morrison(-Woodbury)
- Approximations
  - Martin-Flajolet, hot-lists, Cohen

CMU SCS

Among them

- **fractals / power laws** probably lead to the most startling discoveries ('the mean may be meaningless')
- **SVD**: behind PageRank/HITS
- **Wavelets**: Nature seems to prefer them
- **RLS; approximations**: seem to achieve the impossible

CMU SCS

Thank you!

- feel free to contact me:
  - christos@cs.cmu.edu
  - WeH 7107
- Have a great summer!