CS395T: Large Scale Data-Mining

Lecture I
Introduction

- **Data Mining** is the *automatic discovery* of *interesting patterns and relationships* among *data items.*
- Statistical Pattern Recognition/Machine Learning + huge amounts of data = Data mining.
- This course focus: *text and web mining*

Topics to be covered

1. **Search Engines (Mathematical Aspects)**
   - CLEVER project at IBM ("Hubs" and "Authority" model)
   - Google ("PageRank")
   - Latent Semantic Indexing (LSI) uses SVD for "intelligent query retrieval".
2. **Clustering/Classification**
   - Algorithms
   - Text Clustering (sparse, high-dimensional data)
   - Focussed crawling of the web
   - Classification by Support Vector Machines (SVMs)
3. **Graph Partitioning**
   - Algorithms
   - METIS Graph partitioner for document clustering
   - Spectral partitioning (uses eigenvectors for partitioning)
   - Image segmentation using graph partitioning
4. **Visualization of High-Dimensional Data**
   - CViz Software
   - Visual Text analysis at PNNL (SPIRE)
5. **Other Applications**
   - Face Recognition
   - Bioinformatics/computational biology