CS347

Lecture 8

May 7, 2001

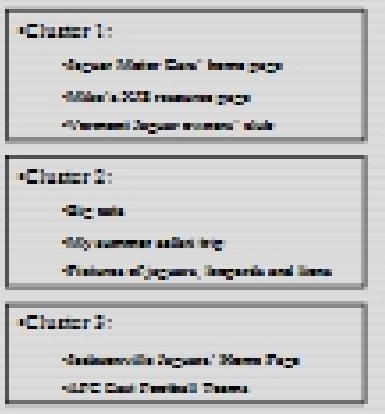
Today's topic

- Clustering documents

Why cluster documents

- Given a corpus, partition it into groups of related docs
  - Recursively, can induce a tree of topics
- Given the set of docs from the results of a search (say *jaguar*), partition into groups of related docs
  - semantic disambiguation

Results list clustering example



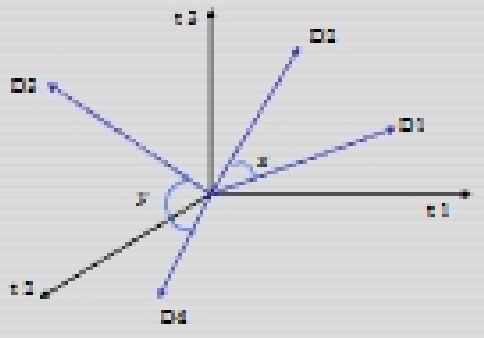
What makes docs "related"?

- Ideal: semantic similarity.
- Practical: statistical similarity
  - We will use cosine similarity.
  - Docs as vectors.
  - For many algorithms, easier to think in terms of a *distance* (rather than similarity) between docs.
  - We will describe algorithms in terms of cosine distance

Recall doc as vector

- Each doc *j* is a vector of *tf*×*idf* values, one component for each term.
- Can normalize to unit length.
- So we have a vector space
  - terms are axes
  - *n* docs live in this space
  - even with stemming, may have 10000+ dimensions

Intuition



Postulate: Documents that are "close together" in vector space talk about the same things.

Cosine similarity

Cosine similarity of D_j, D_k:

$$sim(D_j, D_k) = \sum_{i=1}^{n} w_{ij} \times w_{ik}$$

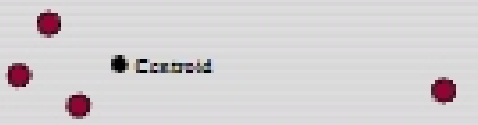
Aka normalized inner product.

Two flavors of clustering

- Given n docs and a positive integer k, partition docs into k (disjoint) subsets.
- Given docs, partition into an "appropriate" number of subsets.
  - E.g., for query results - ideal value of k not known up front.
- Can usually take an algorithm for one flavor and convert to the other.

Cluster centroid

- Centroid of a cluster = average of vectors in a cluster - is a vector.
  - Need not be a doc.
- Centroid of (1,2,3); (4,5,6); (7,2,6) is (4,3,5).



Outliers in centroid computation

- Ignore outliers when computing centroid.
  - What is an outlier?
  - Distance to centroid > M × average.
    - Say 10.

Agglomerative clustering

- Given target number of clusters k.
- Initially, each doc viewed as a cluster
  - start with n clusters;
- Repeat:
  - while there are > k clusters, find the "closest pair" of clusters and merge them.