Document Clustering

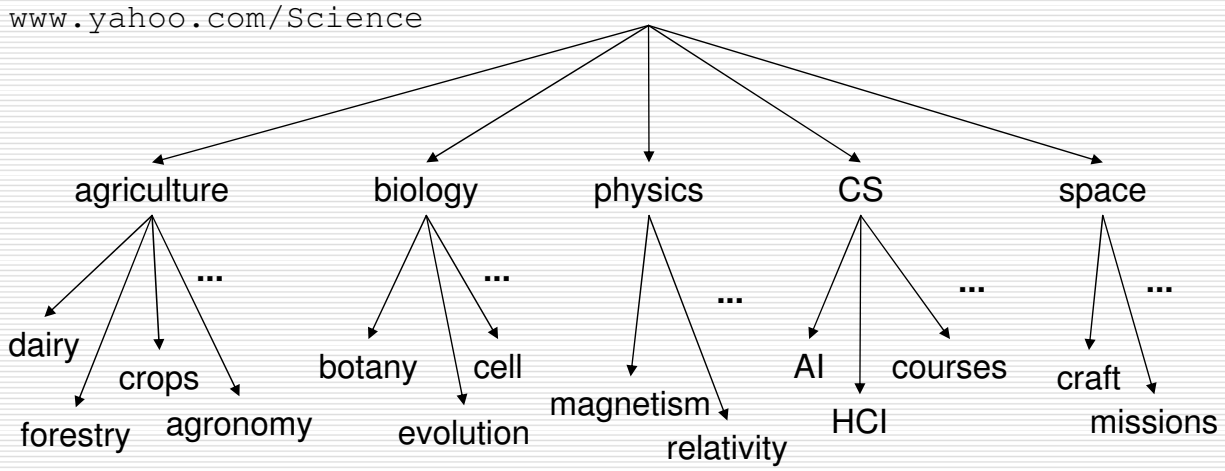
What is clustering?

- Clustering: the process of grouping a set of objects into classes of similar objects
  - The commonest form of *unsupervised learning*
    - Unsupervised learning = learning from raw data, as opposed to supervised data where a classification of examples is given
  - A common and important task that finds many applications in IR and other places
- Not only Document Clustering (e.g. terms)

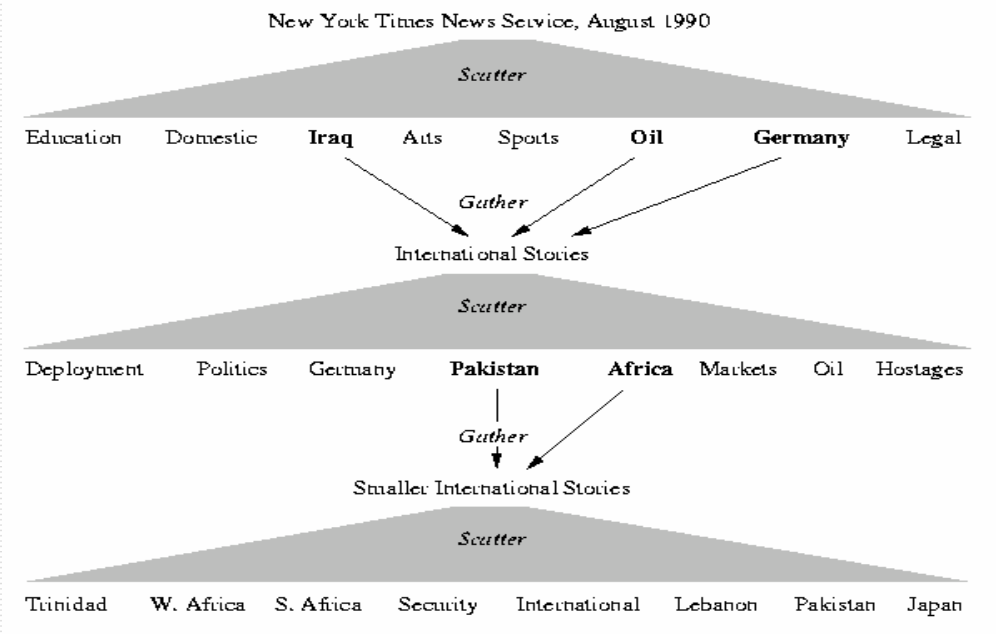
Why cluster documents?

- Whole corpus analysis/navigation
  - Better user interface
- For improving recall in search applications
  - Better search results
- For better navigation of search results
  - Effective "user recall" will be higher
- For speeding up vector space retrieval
  - Faster search

Yahoo! Hierarchy

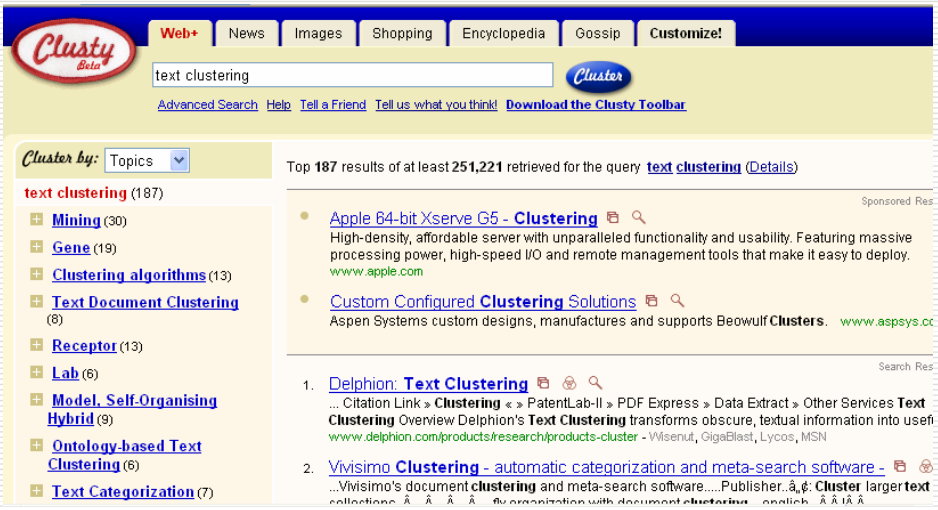


Scatter/Gather: Cutting, Karger, and Pedersen



For better navigation of search results

- For grouping search results thematically
  - clusty.com / Vivisimo



For improving search recall

- *Cluster hypothesis* - Documents with similar text are related
- Therefore, to improve search recall:
  - Cluster docs in corpus a priori
  - When a query matches a doc *D*, also return other docs in the cluster containing *D*
- Hope if we do this: The query "car" will also return docs containing *automobile*
  - Because clustering grouped together docs containing *car* with those containing *automobile.*

The Clustering Problem

Given:

- A set of docuements D={d_1,..d_n}
- A similarity measure (or distance metric)
- A partitioning criterion
- A desired number of clusters K

Compute:

- An assignment function $\gamma : D \rightarrow \{1,..,K\}$
  - None of the clusters is empty
  - Satisfies the partitioning criterion w.r.t. the similarity measure

Issues for clustering

- Representation for clustering
  - Document representation
    - Vector space? Normalization?
  - Need a notion of similarity/distance
- How many clusters?
  - Fixed a priori?
  - Completely data driven?
    - Avoid "trivial" clusters - too large or small
      - In an application, if a cluster's too large, then for navigation purposes you've wasted an extra user click without whittling down the set of documents much.

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.
  - Any kernel function can be used

Objective Functions

- Often, the goal of a clustering algorithm is to optimize an objective function
- In this cases, clustering is a search (optimization) problem
- K^N / K! different clustering available
- Most partitioning algorithms start from a guess and then refine the partition
- Many local minima in the objective function implies that different starting point may lead to very different (and unoptimal) final partitions

What Is A Good Clustering?

- Internal criterion: A good clustering will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the document representation and the similarity measure used

External criteria for clustering quality

- Quality measured by its ability to discover some or all of the hidden patterns or latent classes in gold standard data
- Assesses a clustering with respect to ground truth
- Assume documents with C gold standard classes, while our clustering algorithms produce K clusters, $\omega_1,..,\omega_k$ with n_i members.

External Evaluation of Cluster Quality

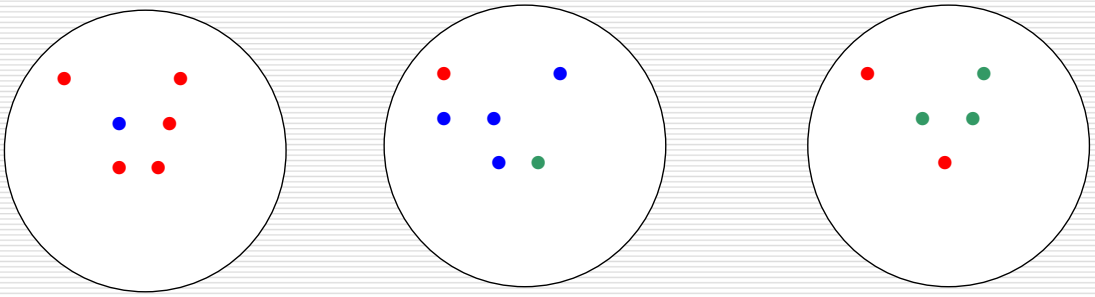
- Simple measure: purity, the ratio between the dominant class in the cluster π_i and the size of cluster ω_i

$$Purity(\omega_k) = \frac{1}{|\omega_k|} \max_j n_{kj}, \; n_{kj} = |\omega_k \cap c_j|$$

- Others are entropy of classes in clusters (or mutual information between classes and clusters)

$$I(\Omega, C) = \sum_k \sum_j P(\omega_k c_j) \log \frac{P(\omega_k c_j)}{P(\omega_k)P(c_j)}$$

Purity example



Cluster I | Cluster II | Cluster III

Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6

Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

Rand Index

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	A	C
Different classes in ground truth	B	D

Rand index: symmetric version

$$RI = \frac{A + D}{A + B + C + D}$$

Compare with standard Precision and Recall.

$$P = \frac{A}{A + B} \qquad R = \frac{A}{A + C}$$

Rand Index example: 0.68

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	20	24
Different classes in ground truth	20	72