DISTANCE-BASED CLUSTERING

BUCKSHOT CLUSTERING ALGORITHM

Created by Jan Ferko / @janferko

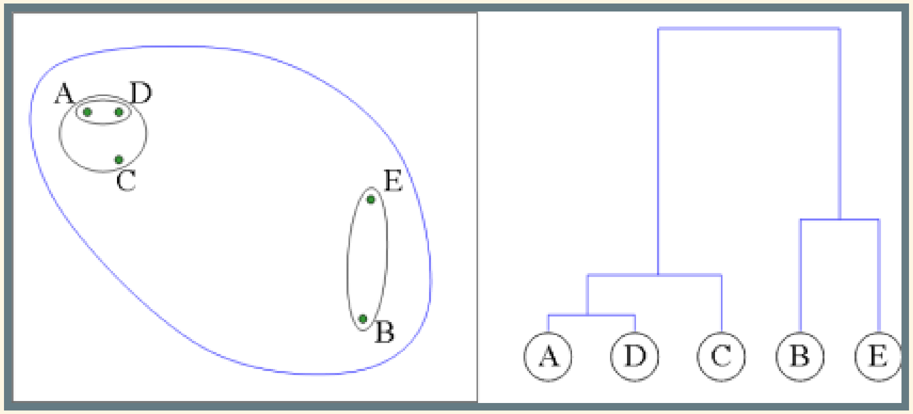
DISTANCE-BASED CLUSTERING

- Computes clusters based on closeness between text documents
- Uses cosine similarity function

$$cosine(U, V) = \frac{\sum_{i=1}^{k} f(u_i) \cdot f(v_i)}{\sqrt{\sum_{i=1}^{k} f(u_i)^2} \cdot \sqrt{\sum_{i=1}^{k} f(v_i)^2}}$$

HIERARCHICAL CLUSTERING

- creates cluster hierarchy (dendogram)
- merges clusters with the best pairwise similarity
- complexity - $O(n^2)$
- merging methods - single link, complete link, group average



PARTITION CLUSTERING

```
let d be distance between two instance
select k random seeds
until convergence or stop condition
  for each x in nodes
    assign x to cluster c_j with d(x, c_j) is the smallest

  for each c_i in clusters
    s_i = centroid(c_i)
```

- set fixed number of clusters before algorithm starts
- different cluster assignment for different runs
- faster than HAC

HOW TO GET THE BEST OF BOTH WORLDS?

Buckshot algorithm

- Use HAC to select k clusters
- Select $\sqrt{k \cdot n}$ documents
- Apply HAC until we get k clusters (complexity $O(k \cdot n)$)
- Apply K-means to create document classifier
- Much more robust initial set of seeds
- Linear complexity

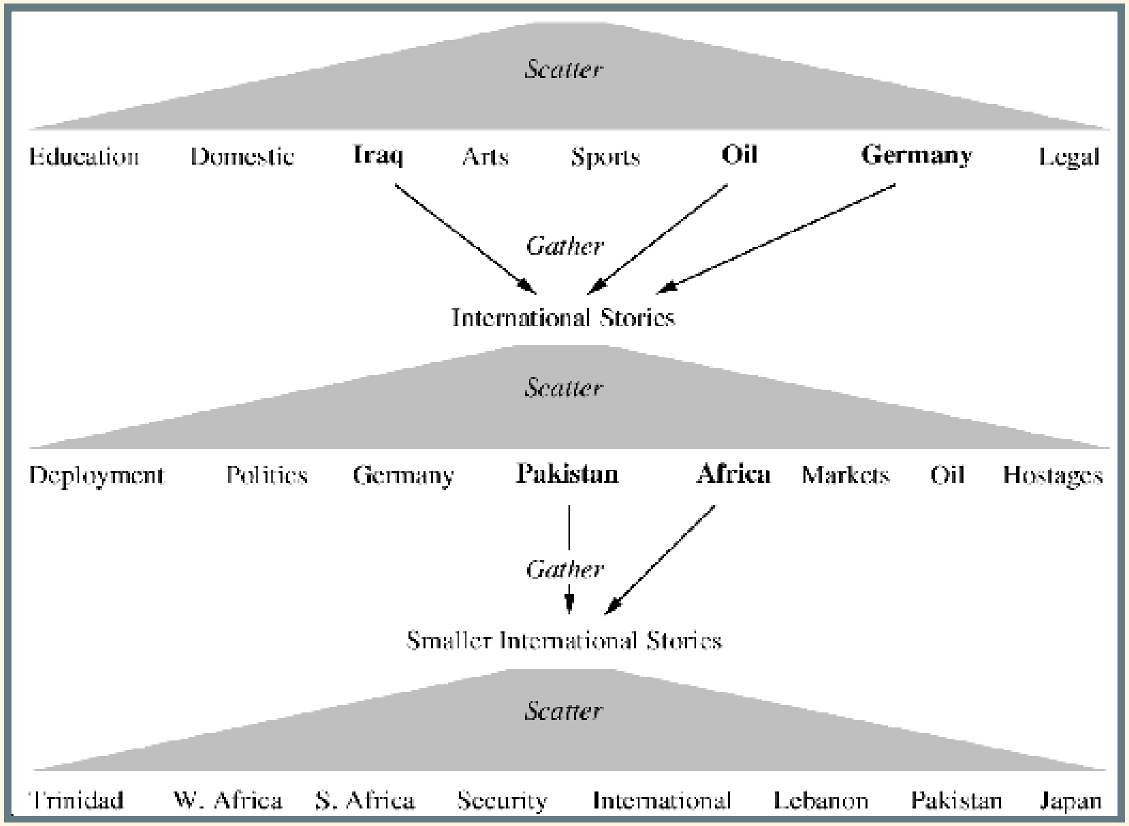
ONLINE DOCUMENT SEARCH

- Scatter / Gather
- Needs fast clustering algorithm to categorize documents in user friendly time
- Separate documents to initial clusters (Scatter)
- use words with highest weight in whole group to create description
- User selects clusters she is interested in
- Group clusters together (Gather)
- Do reclustering on merged groups (Scatter)

AN ILLUSTRATION

- Search 5000 articles from New York Times collected during August 1990
- User wants to find out what happened in August?
- Conventional algorithms don't work because of:
-
  - Vague search query, e.g. What happened in August?
  - User does not know words to describe topic
  - Words to describe topic might not be used to discuss topic, e.g. international event
  - Usage of synonyms in documents

ALGORITHM WORKFLOW EXAMPLE



RESOURCES

- Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey: Scatter/Gather - A Cluster-based Approach to Browsing Large Document Collections

THE END

BY JAN FERKO