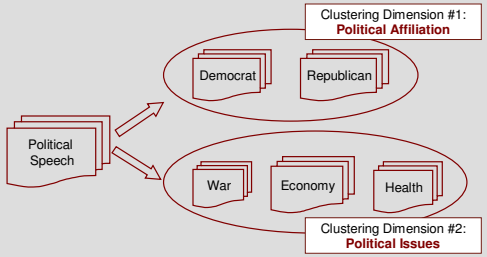
Mining Clustering Dimensions

Sajib Dasgupta and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

Motivation

Same data can be naturally clustered along multiple dimensions



Goals

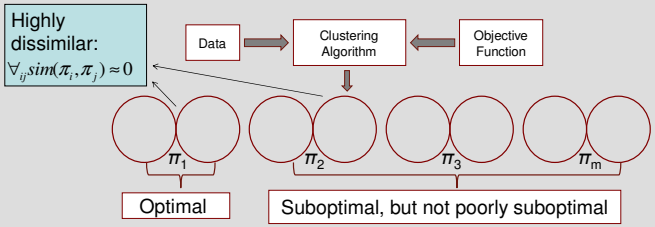
1. Learn the possible clustering dimensions of a dataset
2. Enable a user to visualize the clustering dimensions

Important from the viewpoint of exploratory data analysis:
- User may have no knowledge of the data
- User wants to know how the data can be clustered

Producing Multiple Clusterings

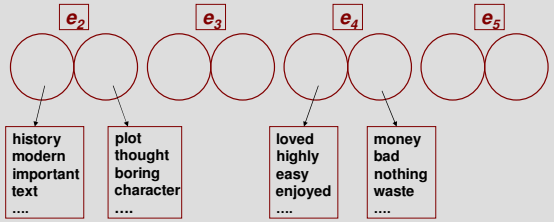
Such an algorithm should possess three desirable properties:

- **Multiplicity:** The clustering algorithm should be able to produce m (m>1) clusterings π_i: i=1:m with a **single feature space** and a **single objective function.**
- **Distinctivity:** The resulting clusterings should be **distinctively different i.e.** $\forall_{ij} sim(\pi_i, \pi_j) \approx 0$
- **Quality:** Each of the clusterings has to be **qualitatively strong (close to optimal)**



Our Algorithm

- **Producing the optimal clustering**
  - Spectral clustering, objective function: normalized cut
  - Optimal partitioning function **f**: $\arg\min_f \sum_{i,j} S_{ij} (\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}})^2$ s.t. $\| f \|^2 = \sum_i d_i$ and $f \perp D^{\frac{1}{2}} 1$
  - **f=e_2**, the second eigenvector of the Laplacian
  - Apply k-means to cluster the data points represented by **e_2**
- **Producing suboptimal clusterings**
  - Solve $\arg\min_f \sum_{i,j} S_{ij} (\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}})^2$ s.t $\| f \|^2 = \sum_i d_i$ and $f \perp D^{\frac{1}{2}} 1$ and $f \perp e_2$
  - **f=e_3**, the third eigenvector of the Laplacian
  - Apply k-means to cluster the data points represented by **e_3**
- **Producing *m* clusterings**
  - Apply k-means to cluster the points represented by **e_2, e_3, .., e_{m+1}** separately
- **The algorithm ensures multiplicity, distinctivity and clustering quality**
  - **Multiplicity:** We don't change the feature space or normalized cut objective
  - **Distinctivity:** The eigenvectors are orthogonal to each other
  - **Quality:** **e_2** achieves the minimum normalized cut, **e_3** achieves the next minimum normalized cut. Each of the eigenvectors is the "next best" orthogonal solution achieved by the spectral system.
- **To help users visualize the induced clustering dimensions, our algorithm**
  - represents each dimension using representative unigrams
  - uses weighted log-likelihood to extract the top unigrams from each partition

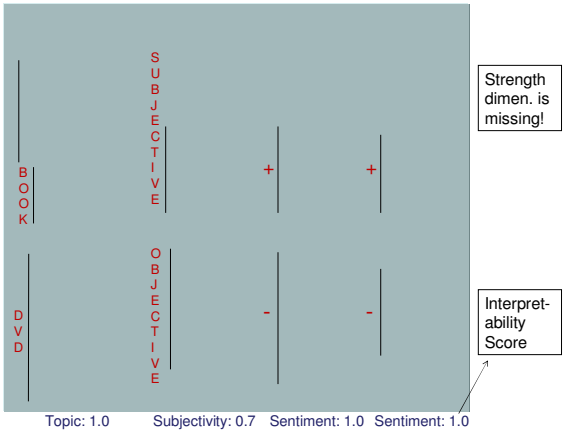


Evaluation

- Document clustering tasks: each dataset has multiple 2-way clustering dimensions
- Feature representation: bag of words; similarity metric: the dot product
- **Book-DVD dataset:** 4000 book and DVD reviews
  - **Dimensions:** Topic (Book vs. DVD), Sentiment (Positive vs. Negative), Subjectivity (Subjective vs. Objective), Strength (Strong vs. Weak)
- **Politics dataset:** 2000 articles written by Democrat and Republican supporters
  - **Dimensions:** Affiliation (Democrat vs. Republican), Policy (Foreign vs. Domestic)
- **Goals:** Determine (1) which of these dimensions our algorithm can recover; (2) whether they are human-interpretable, and (3) how good the clusterings are

Results

- **Interpretability of clustering dimensions**
  - Ask ten humans to independently assign a dimension label to each induced dimension she thinks is interpretable



  - Similar results were obtained for the politics dataset
- **Quality of the clusterings**
  - Compute accuracy against the gold standard clusterings
  - Three **baselines:** Ng et al.'s spectral clustering, meta clustering and iterative feature removals (IFR)

Book-DVD	Topic	Sentiment	Subjectivity	Strength
Spectral	77.9	52.9	68.5	51.8
Meta clustering	50.2	50.2	58.6	50.1
IFR	77.1	50.0	51.0	50.1
Our algorithm	77.1	68.9	59.7	54.2

Politics	Political Affiliation	Policy
Spectral	54.3	67.6
Meta clustering	59.4	61.6
IFR	57.8	61.6
Our algorithm	69.7	70.2

Conclusion and Future Work

- Presented an algorithm that learns and helps users visualize important clustering dimensions of a dataset.
- Future work involves quantifying the *multi-clusterability* and *ambiguity* of a dataset