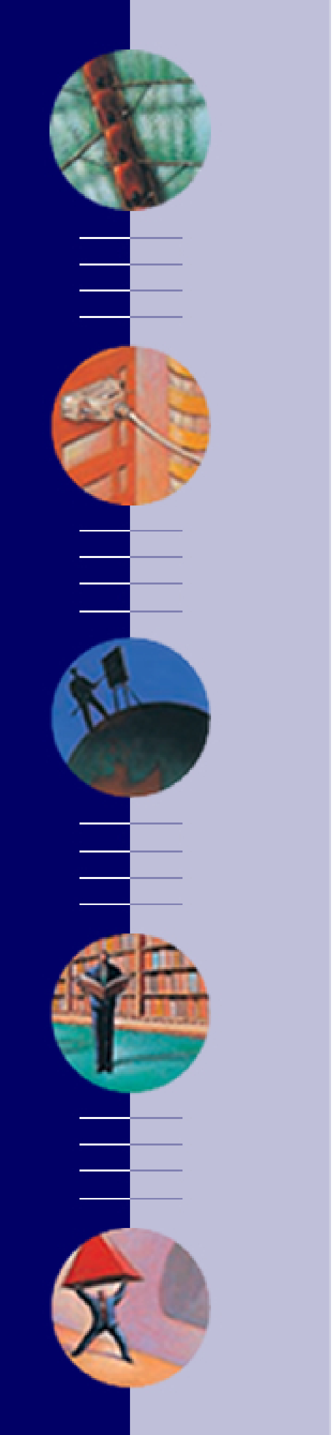
Topic Detection by Clustering Keywords

Christian Wartena
Rogier Brussee

Telematica Instituut
Enschede
The Netherlands

MultimediaN

Telematica Instituut



Overview

- Problem: find the main topics of a collection
- Keyword extraction
- Clustering
- Distance
- Data and evaluation
- Results

Problem: get the topics of a corpus

- Given a collection of texts, can we identify the main topics of this collection?

- Approach
  - Extract meaningful terms (‘keywords’)
  - Cluster these terms
  - Does each cluster represent a topic?

Keywords

- Simple approach for determining meaningful terms:
  - Most frequent nouns, verbs (no auxiliaries) and proper names
  - But no terms that are too general
    - i.e. terms with a distribution of co-occurring terms similar to the background distribution

Clustering (1)

- agglomerative hierarchical
  - single link
  - Worked only well for finding many small clusters

- Density based
  - DBSCAN
  - Almost as good as top-down

- Top down
  - induced bisecting k-means
  - Best results

Clustering (2): Induced bisecting k-means

1. Select two elements *a*,*b* with maximal distance as seed points for two clusters
2. Assign all items to the cluster with the closest seed point
3. Compute the centers *a’* and *b’* of both clusters.
4. Repeat step 2 and 3 starting with *a’* and *b’* as new seed points until the centers become stable.

5. If the diameter of a cluster is larger than a specified threshold value, the whole procedure is applied recursively to that cluster.

Distance

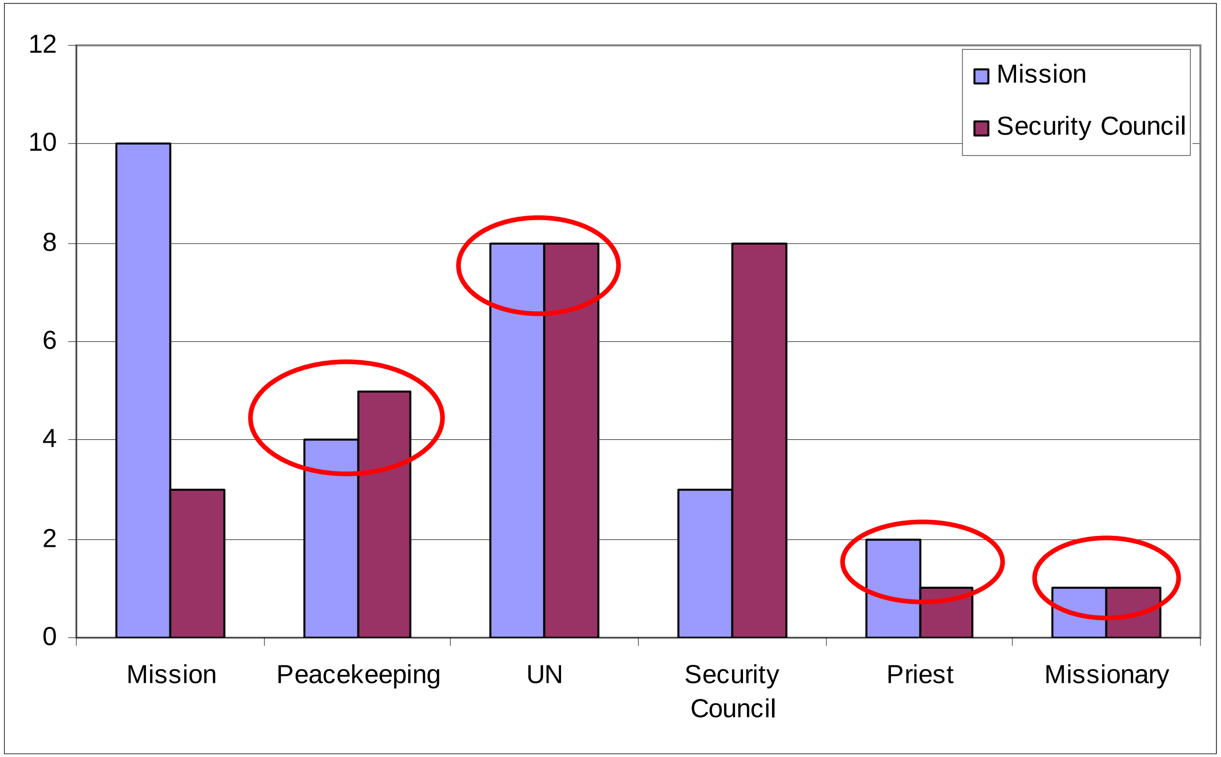
- Two terms are similar if they
  - Have a similar distribution over items
    - Cosine
    - Divergence (relative entropy) of distributions
  - Often co-occur
    - ~~E.g. Jaccard coefficient~~
  - Co-occur with the same other terms
    - New: our approach

- We need a measure that allows to compute a center of a cluster

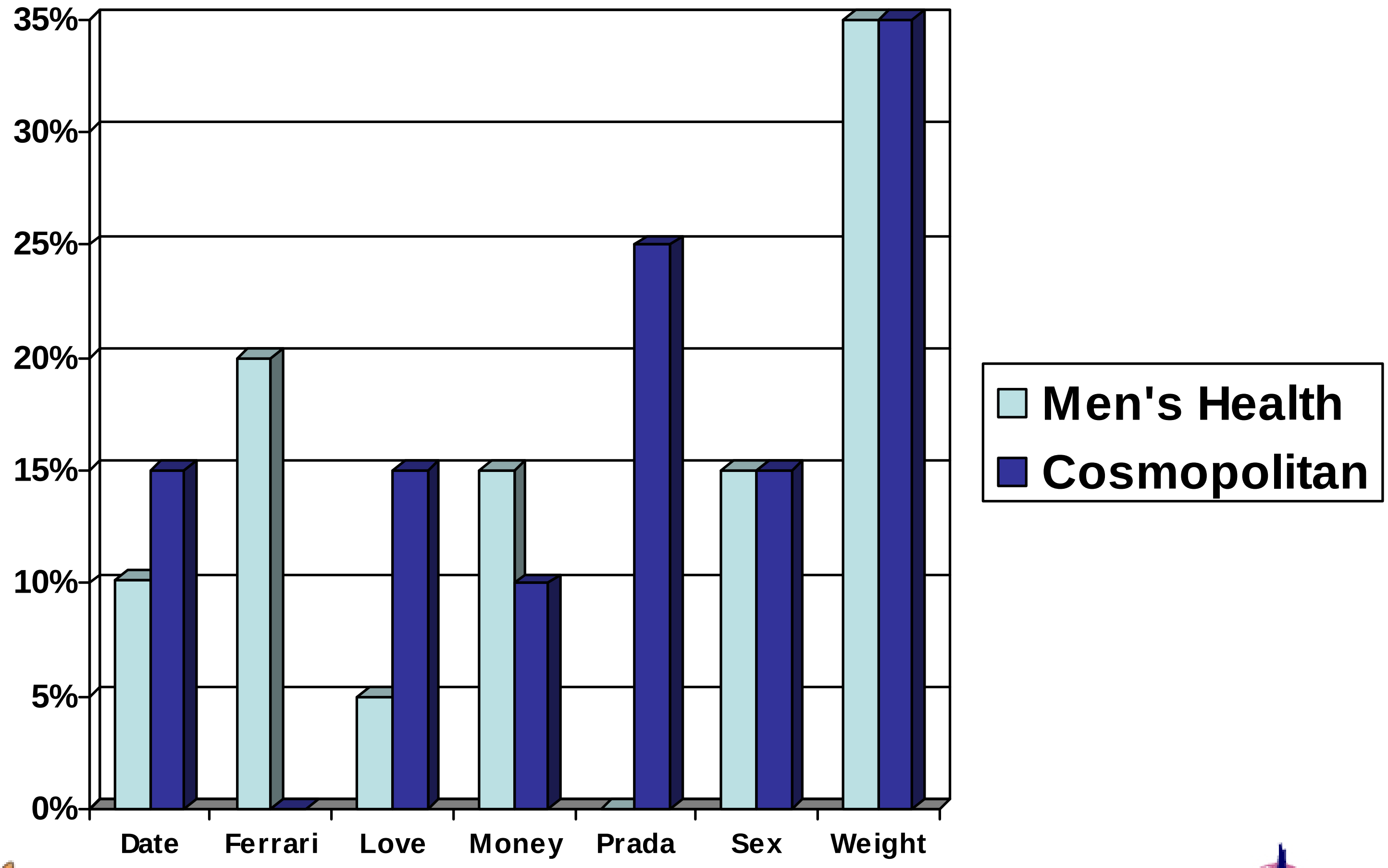
Co-occurence

- Key idea:
- Terms are similar if they have similar co-occurrence patterns
  - Consider the probability distribution that a term co-occurs with other terms
  - Compare these *co-occurence distributions*

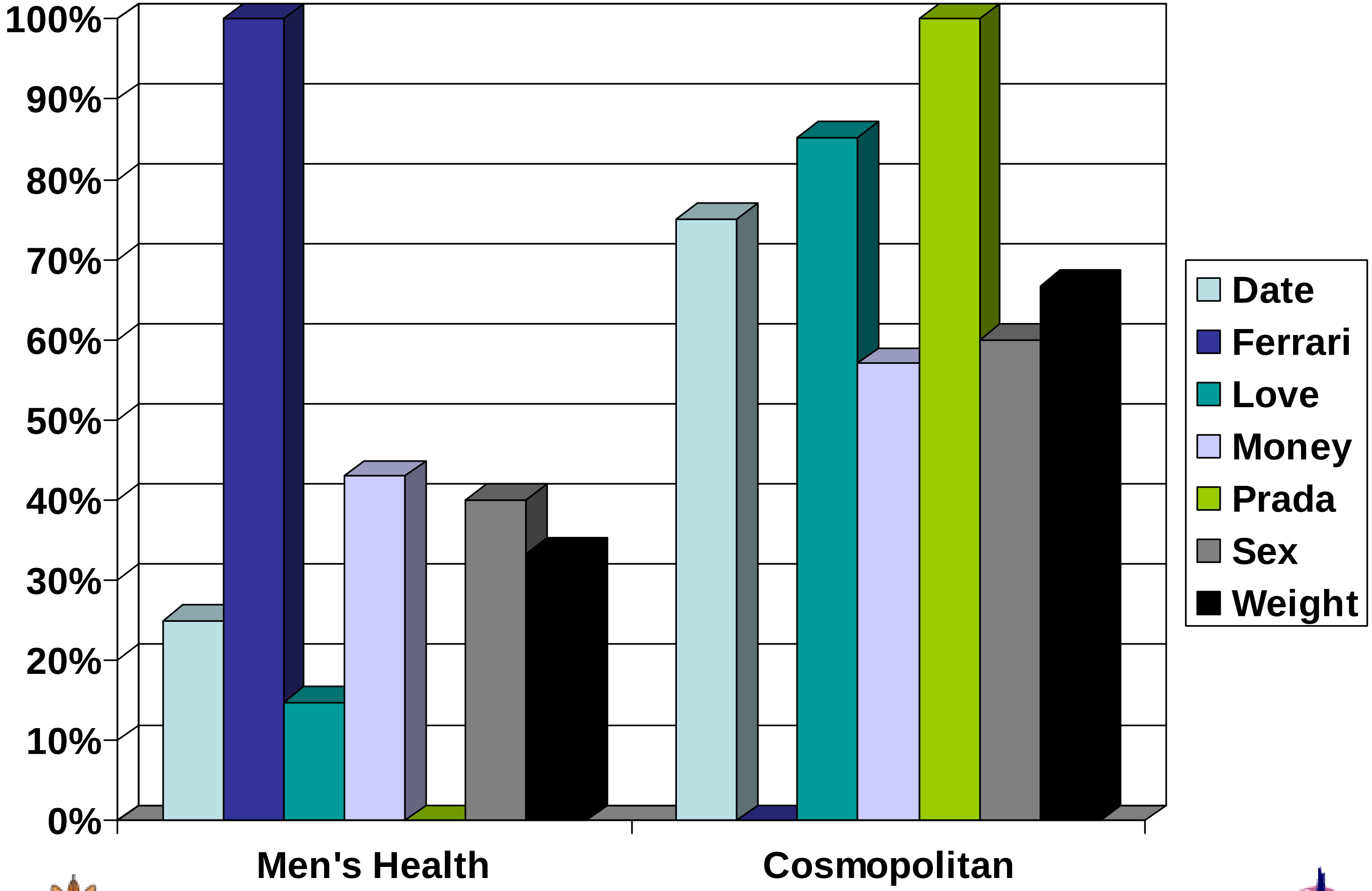
	Mission	*Peacekeeping*	*UN*	*Security Council*	*Priest*	*Missionary*
Mission	10	4	8	3	2	1
Peacekeeping	4	7	4	5	0	0
UN	8	4	14	8	1	0
Security Council	3	5	8	8	1	1
Priest	2	0	1	1	6	4
Missionary	1	0	0	1	4	8



Term Distribution for source: $q(t|d)$



Document Distribution for a term: $Q(d|t)$



Distribution of co-occurring terms

- $$\overline{p_z}(t) = \sum_d q(t \mid d) Q(d \mid z)$$

- where
  - *q(t|d)* is the term distribution of d
  - *Q(d|z)* is the document distribution of *z*
    - “The fraction of *z*’s that is found in *d*”

- Weighted average of the term distributions of documents
  - The weight is the relevance of d for z given by the probability *Q(d|z)*

Distance of terms

- Jensen-Shannon divergence of distributions of co-occurring terms

- Kullback-Leibler divergence:

$$D(p\|q) = \sum_t p(t) \log\left(\frac{p(t)}{q(t)}\right)$$

- Jensen-Shannon divergence:

$$JSD(p\|q) = \tfrac{1}{2} D(p\|m) + \tfrac{1}{2} D(q\|m)$$

- Mean distribution: $m = \tfrac{1}{2}(p+q)$

Evaluation

- Data
  - 758 Wikipedia articles from 8 categories
  - Categories:
    - pop music
    - painting
    - architecture
    - trees
    - monocots
    - charadriiformes
    - aviation
    - space flight
  - 118.099 words
  - 27.373 unique terms

Task

- 160 keywords selected
  - Most frequent
  - $D(\overline{p}_t \| q) > 1$
- Cluster keywords into disjoint sets
- Keep keywords and clustering method constant
- Vary distance measure and number of clusters

Reference Clustering

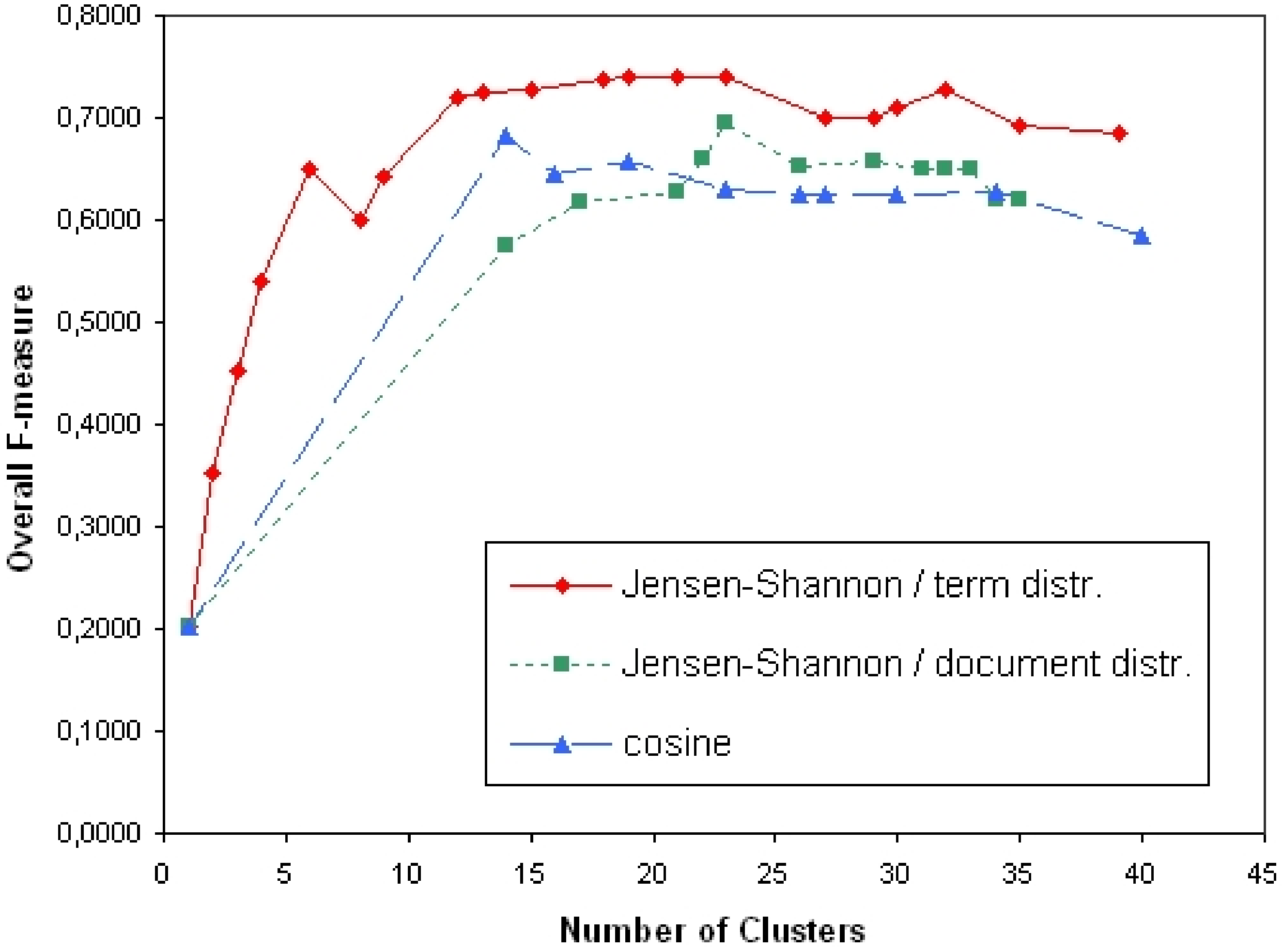
- **1:** Define a cluster for each category
  - Compute term distribution q_c for each category
  - Assign each term *t* to a cluster *c* such that $JSD(\bar{p}_t \| q_c)$ is minimal
- **2:** As 1 but with one additional cluster defined by the term distribution of the whole collection

Evaluation measure

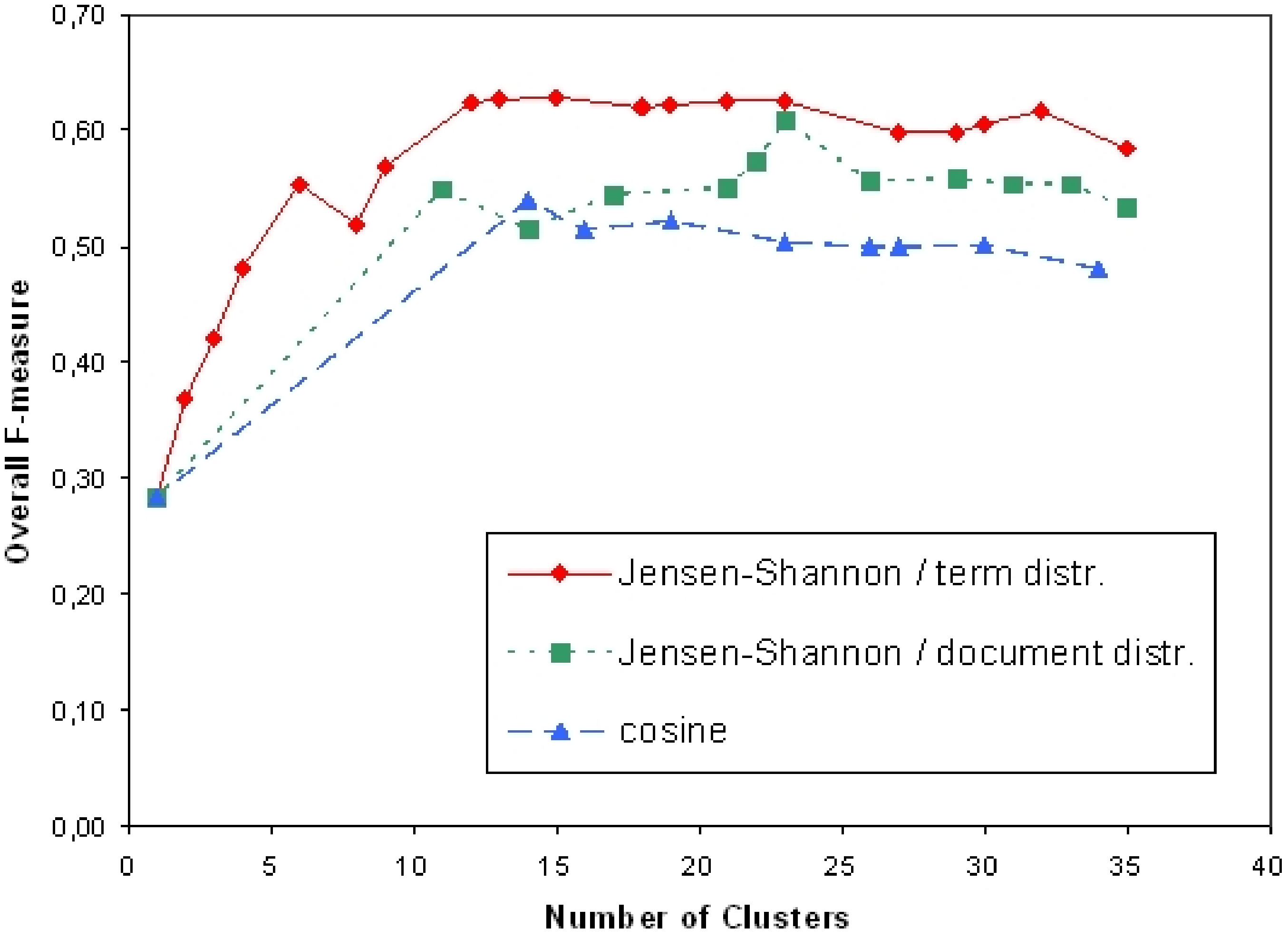
- For each reference cluster
  - find the best fitting cluster
  - compute the F-value for that cluster

- Compute the weighted average of all 8 (9 resp) F-values.
  - Weighted by the size of the reference cluster

Results (8 categories)



Results (9 categories)



Summary

- Selection and clustering of most meaningful terms seems to be a good method to identify topics
- Divergence of co-occurring terms distributions is an interesting measure for similarity of terms in a collection of documents