16th Conference of the International Federation of Classification Societies
Neutral Benchmarking Studies of Clustering Challenge

Alain LELU - Université de Franche-Comté <alelu@orange.fr>
Martine CADOT - LORIA <Martine.Cadot@loria.fr>

EVALUATION OF TEXT CLUSTERING METHODS AND THEIR DATASPACE EMBEDDINGS: AN EXPLORATION

Motivations and goals

- Our focus: **text** clustering, essential to domain application of bibliometric delineation of scientific fields
- Previous study: *Martine Cadot, Alain Lelu, Michel Zitt. Benchmarking seventeen clustering methods on a text dataset.* (Research Report) LORIA. 2018. 〈hal–01532894v5〉
- As a supplementary material to: *Michel Zitt, Alain Lelu, Martine Cadot, Guillaume Cabanac. "Bibliometric delineation of scientific fields" in Wolfgang Glänzel et al. Handbook of Science and Technology Indicators, Springer International Publishing, 2019*

Lessons learned from this study:

- Separate algorithm in the strict sense and dataspace (raw word count vectors → transformations: tf-idf weights, kernel, spectral, …)
- Affect lower priority to algorithms with 2 parameters (DbScan, Affinity Propagation, Smart Local Moving Algorithm) or 1 parameter with deceptive results on texts (Density Peaks, Independent Component Analysis, Fuzzy c-means, K-Means++) → more tractable number of algorithm for the present Neutral Benchmarking Challenge

Selected options (1): algorithms

- K-Means (best reconstitution error over 20 elementary runs)
- Hierarchical Agglomerative Clustering (average/Ward linkage)
- Spectral clustering in the broad sense (algorithm X in spectral space Y)
- Graph clustering (Louvain, Infomap)
- Kernel clustering (order-2 polynomial kernel)
- Non-negative Matrix Factorization (NMF)
- Latent Dirichlet Allocation (LDA)

Selected options (2): data spaces

- Raw term counts
- Salton's tf-idf → weighting
- Okapi (aka BM25) weighting
- Laplacian spectral spaces (raw, Salton's, Okapi)
- Chi-square distance space
- Correspondance Analysis spectral space (amounts to noise removal in Chi-square space)
- Kernel space (→ square document × document Gram matrix)

Note: cosine distance has proven to be adequate in all spectral and kernel spaces

Selected options (3): evaluation measures

- Normalized Mutual information (NMI)
- Adjusted Rand Index (ARI)
- Mean local class/cluster F-score (F)
- Purity = 1 – global error rate

Note: F needs a 1-to-1 adjustment of clusters to classes, which has been achieved measuring local F-scores

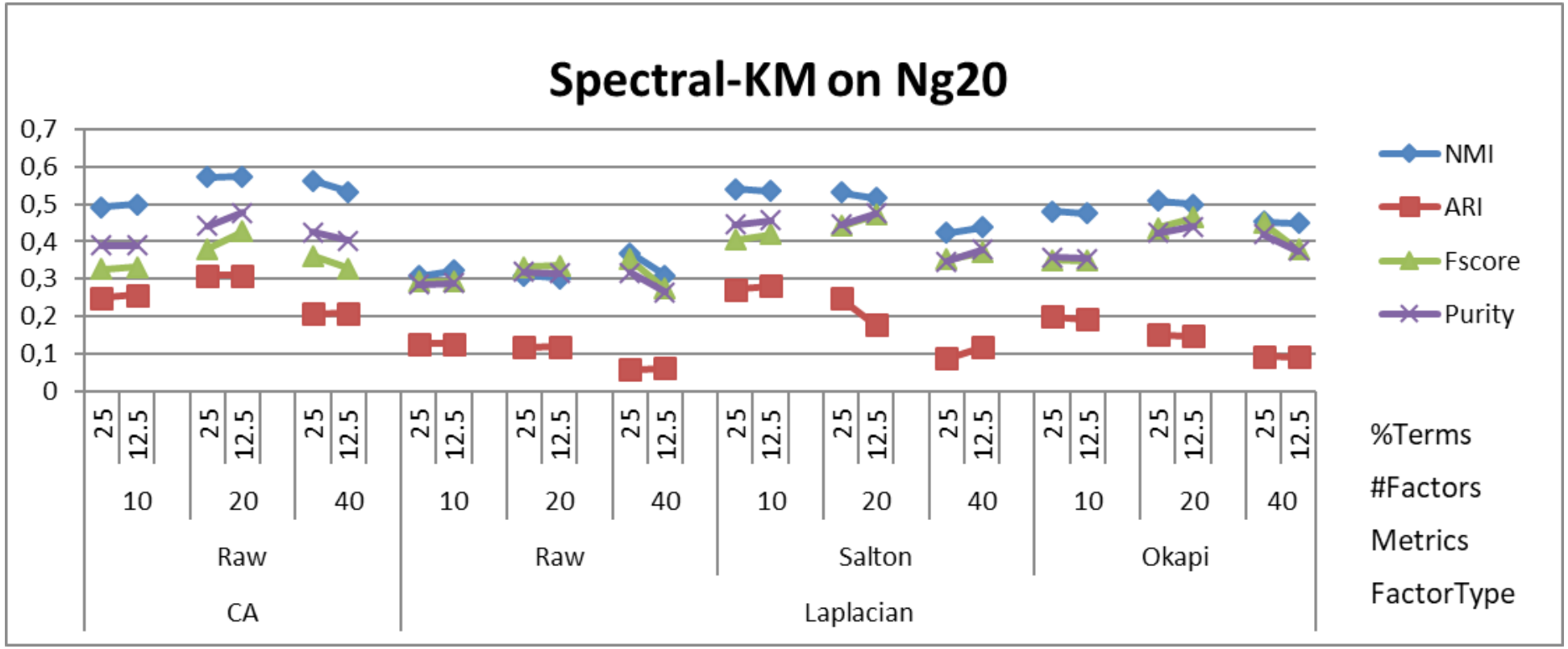
Selected options (4): test corpora

- 3 prototypical public access corpora:
  - ACM proceedings collection: 40 conferences,~3500 full academic papers.
  - 20 Newsgroups: ~19 000 posts in Internet forums, social networks style
  - Reuters'8 classes: ~7700 dispatches, journal style, strong class imbalance
- Homogeneous linguistic processing: public access term-count matrices in http://sites.labic.icmc.usp.br/text_collections/
- Vocabulary « quantile » truncation: keeping 100%, 37.5%, 25%, 12.5% most frequent terms

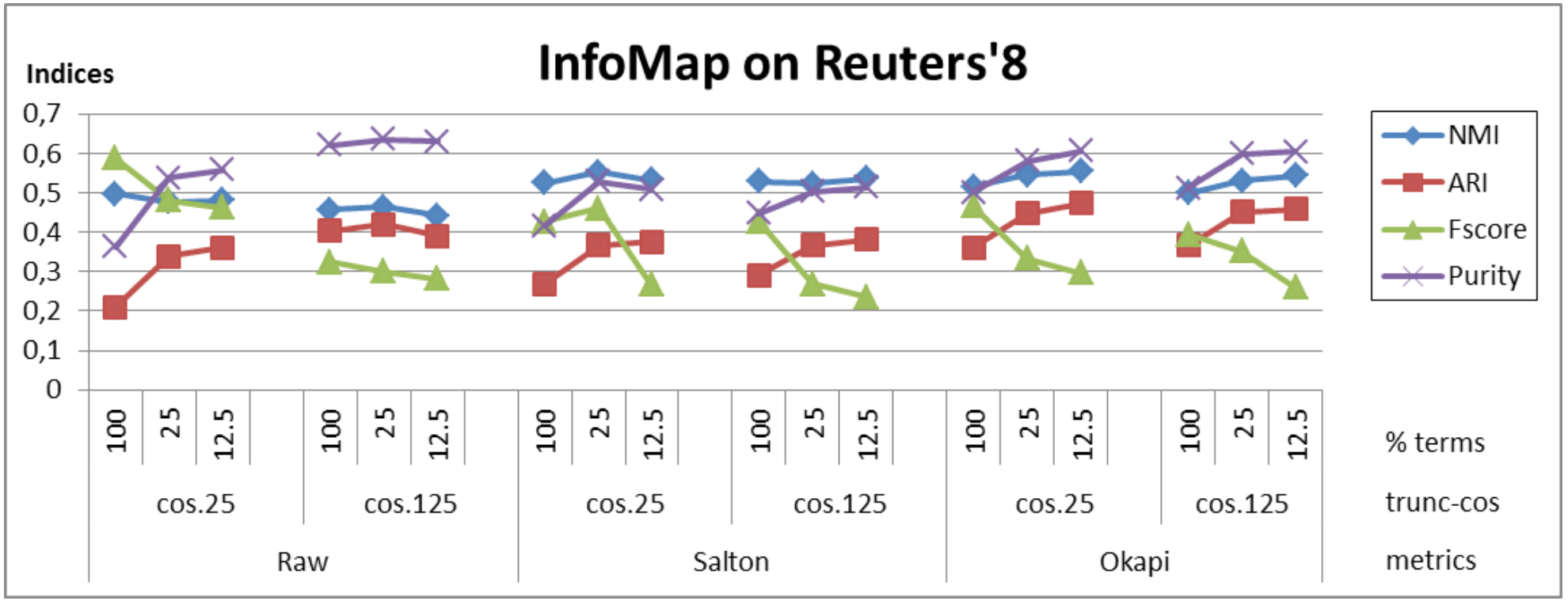
~600 runs (algorithm X in dataspace Y of corpus Z with vocabulary threshold T):

- High variability of « best methods » depending on corpora, no clear winner
- Good results of classic methods and combinations (Hierarchical Clustering, K-Means, Spectral K-means)…
- … but best results with unexpected combinations, such as spectral NMF in Okapi Laplacian space !
- Rich material online in <https://hal.archives-ouvertes.fr/hal-02116493v2>

Results (2): homogeneous behaviour of the 4 indices in « balanced » corpora



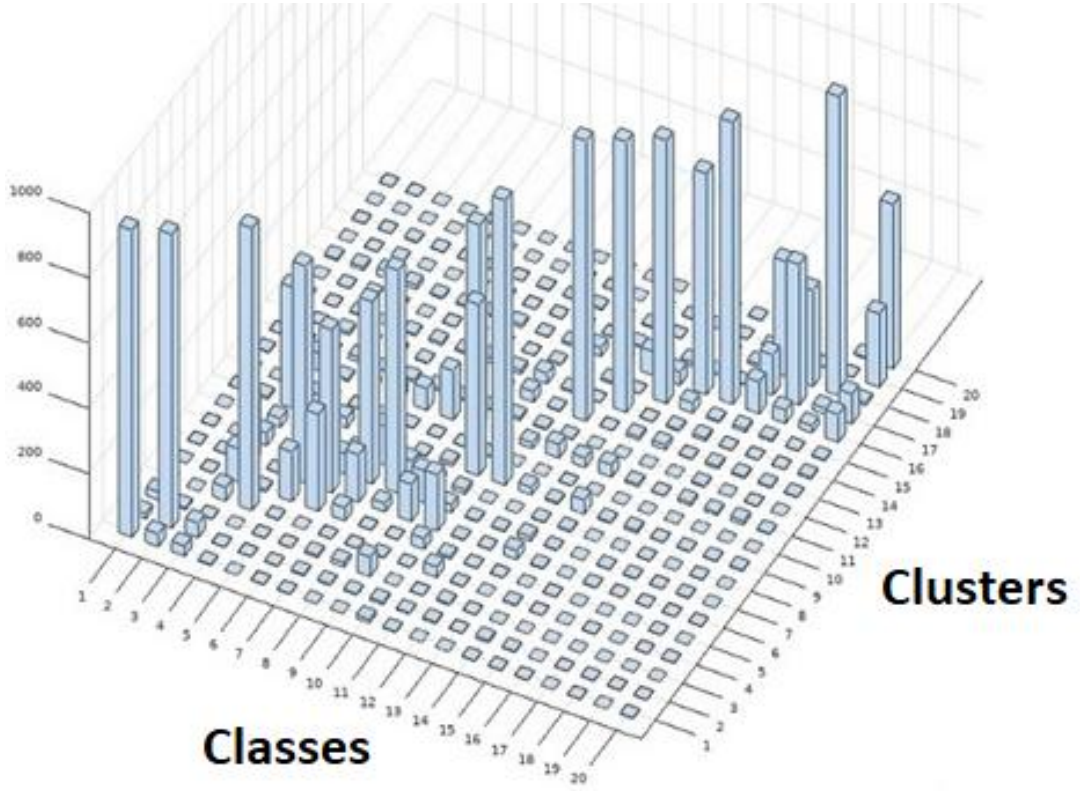
Results (3): heterogeneous behaviour of the 4 indices in « unbalanced » corpus Reuters'8



Winning combinations for each corpus:

- ACM: Spectral HC Ward in Okapi-weighted dataspace – NMI = 0.698
- Re8: Kernel HC Ward – NMI = 0.625
- Ng20: NMF in Okapi-weighted dataspace – NMI = 0.625

Best results are not so good (e.g. best Ng20):



- « snake » structure
- Class#3 and class#9 are dispersed through many clusters

In search of inter-corpus commonalities:

- **Spectral HC Ward in CA dataspace:**

ACM .631/.698 - Re8 .604/.625 – Ng20 .622/.625 (**95% / NMI max**)

- Spectral K-Means in CA dataspace:

ACM .592/.698 - Re8 .515/.625 – Ng20 .574/.625 (87% / NMI max)

- Spectral HC Ward in Okapi dataspace:

ACM .698/.698 - Re8 .523/.625 – Ng20 .550/.625 (91% / NMI max)

Conclusions, perspectives

- Interactions between algorithm, type of dataspace, class imbalance, linguistic pre-processing, text styles, and ultimately classification viewpoint, need further investigations
- In the meantime Spectral Hierarchical Clustering (and spectral K-Means to a lesser extent) in Correspondence Analysis space (in Okapi-weighted Laplacian space to a lesser extent) seem(s) the least corpus-dependant way(s) to stick to the human categorization process.
- More details in https://hal.archives-ouvertes.fr/hal-02116493v3

Thank you for your attention !