UniNE at CLEF 2016: Author Clustering

Mirco Kocher

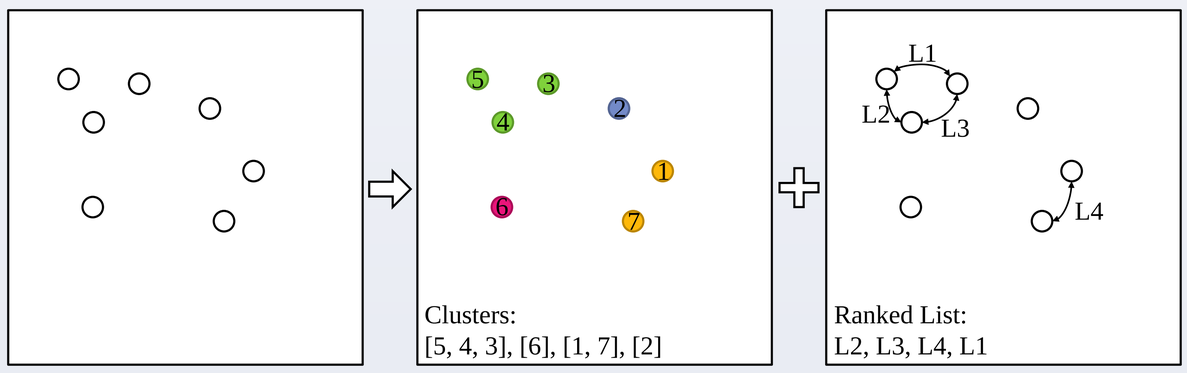
University of Neuchâtel, Switzerland

Swiss National Science Foundation

Task

Given a collection of up to 100 documents, identify authorship links and groups of documents by the same author.

- Non-overlapping clusters
- Document pairs with probability of having same author



- Different datasets
  - Languages: Dutch, English, and Greek
  - Genres: newspapers and reviews
- Total 7 different participants

Strategy

- Select 200 most frequent terms from the query text

	Term	Q	cand1	cand2	…	candN
≤200	the	0.09	0.09	0.10		0.06
	of	0.06	0.06	0.06		0.04
	to	0.03	0.04	0.00		0.05
	and	0.02	0.03	0.04		0.00
	…					

Number of samples in training set

- L_1-norm as distance measure
  - $\Delta(Q, cand1)$: 0.02
  - $\Delta(Q, cand2)$: 0.06
  - …
  - $\Delta(Q, candN)$: 0.09
- Mean and standard deviation of new text to all other
  - mean(Q, X): 0.08
  - SD(Q, X): 0.02
- Indication if distance is more than 2 SD below mean
  - mean(Q, X) – 2.0 * SD(Q, X) = 0.08 – 2.0*0.02 = 0.04
  - $\Delta(Q, cand1) = 0.02 \rightarrow$ 3 SD below mean
  - High probability of authorship link
- Cluster pairs according to links with transitivity rule
  - If A&B are linked then cluster A&B
  - If A&B are linked and B&C are linked, then cluster A&B&C

Results

Language	F-Score	Precision	Recall	MAP
Dutch Newspaper	0.8230	0.9942	0.7180	0.1210
Dutch Reviews	0.8201	0.9900	0.7133	0.0369
English Newspaper	0.7915	0.9600	0.6867	0.0317
English Reviews	0.8036	0.9792	0.6944	0.0252
Greek Newspaper	0.8239	0.9919	0.7172	0.1368
Greek Reviews	0.8480	1.000	0.7515	0.2700
Overall	**0.8184**	**0.9859**	**0.7135**	**0.1036**

Evaluation

- High precision, not a lot of noise in the clusters
- Good recall, many complete clusters
- Balanced clustering performance, independent of text
- Low MAP, some wrong links in high positions
- Limit of 2.0 SD
  - Lower: link and cluster with less evidence
  - Higher: clustering and linking only for extreme cases
- Explanation of proposed assignment (*e.g.* in English corpus)
  - Usually, the relative frequency differences with very frequent words such as *when*, *is*, *in*, *that*, *to*, or *it* can explain the decision

Conclusion

- Simple unsupervised approach solves the clustering problem
- Reduced set of comprehensible features explains the decision
- No language or genre dependent adjustments necessary
- Most frequent terms tend to select most discriminative features
- Biased towards many authors and small clusters

Acknowledgments

This research was supported, in part, by the NSF under Grant #200021_149665/1.

Mirco Kocher – University of Neuchâtel – Rue Emile-Argand 11 – 2000 Neuchâtel, Switzerland – Mirco.Kocher@unine.ch

http://style.unine.ch/