Multi-Element protocol on IR experiments stability: Application to the TREC-COVID test collection

Gabriela Gonzalez-Saez, Philippe Mulhem
Lorraine Goeuriot, Petra Geluscacova

Kodicare ANR PRCI Project

Objectives

- Study the stability of ranking of systems on multiple test collections
  – Check what variations are more impactfull

Protocol proposal

1. From one existing test collection, create multiple controled pairs of sub-collections
2. Study of the stability of the ranking of systems on these pairs

Formalization

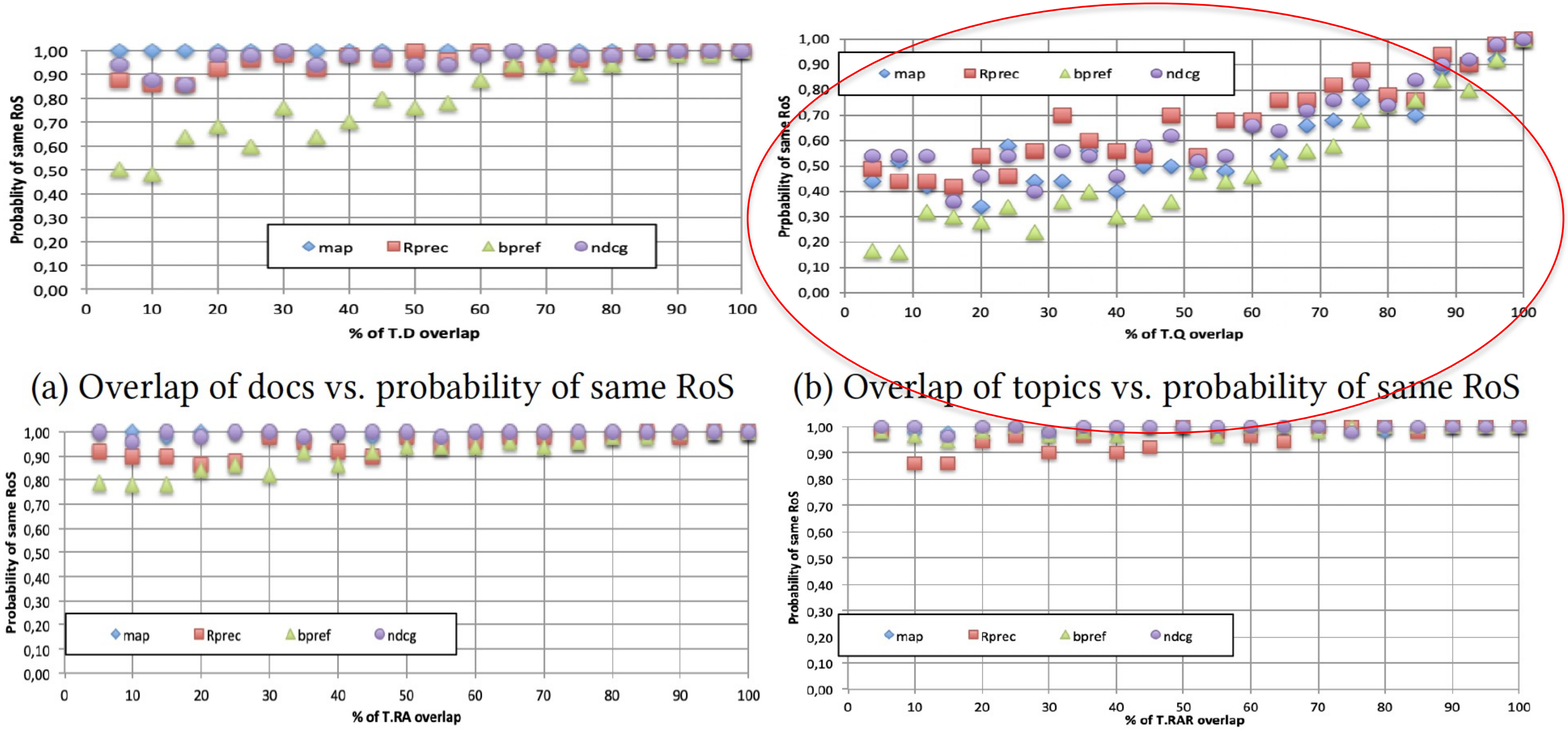
- Two collections T_1 and T_2 are *comparable* according to an eval measure E if a set of systems S are ranked in the same order
- T = (Documents D, Queries Q, Relevance Assessments RA)
- Overlapping of T_1 and T_2 for one element e in {D,Q,RA}:

$$o = \frac{|T_1.e \cap T_2.e|}{|T_1.e|}$$

- Probability of same ranking over overlap values on n runs:

$$p_{e,o}(\Delta(L_{.,1}, L_{.,2}) >= \rho) = \frac{|\{i|i \in [1,n], \Delta(L_{i,1}, L_{i,2}) >= \rho\}|}{n}$$

Results on TREC-COVID collection



(a) Overlap of docs vs. probability of same RoS

(b) Overlap of topics vs. probability of same RoS

(c) Overlap of assessments vs. probability of same RoS

(d) Overlap of positive assessments vs. probability of same RoS

Figure 1: Similarity of sub-test collections on one element in $\{D, Q, RA, RA_R\}$ versus probability of same RoS.

One view results on TREC-COVID

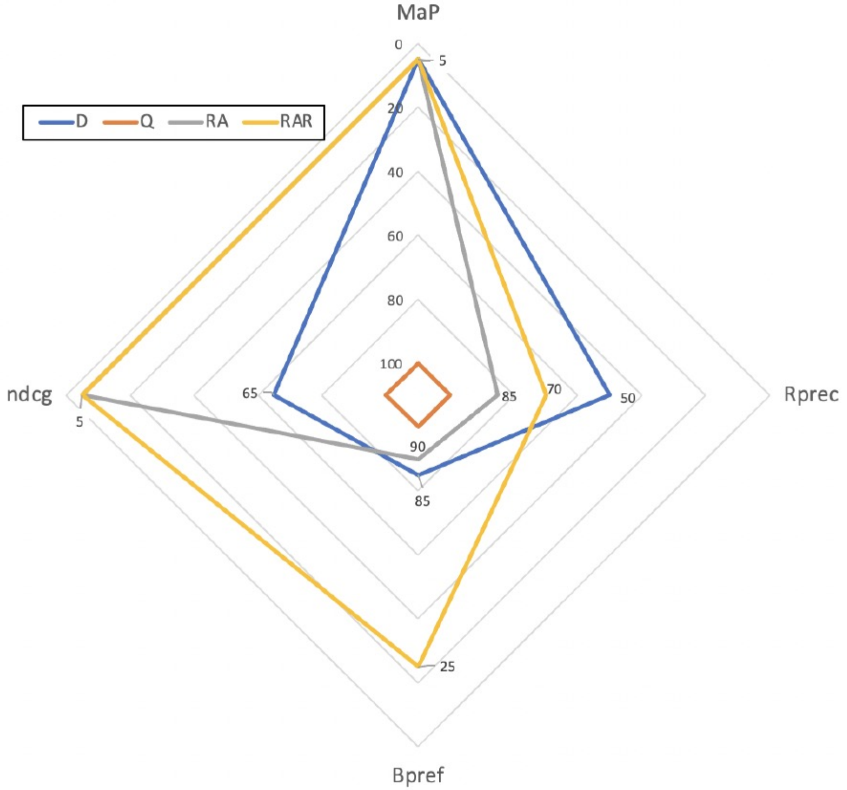


Figure 2: Radar view of $\arg\min_{o \in [5\%,100\%]} [p_{T.e,o}(\Delta(L_{.,1}, L_{.,2}) >= 90\%) = 1]$ for $e \in \{D, Q, RA, RA_R\}$.

Conclusion

- Protocol easy to define

- Elements behave differently
  - More impact of Topics variations: is TREC-COVID a valid test collections ?

- Future
  - Protocol used before delivering a test collection to the community
    - What thresholds ?
  - Study of several elements together
  - Refined overlaps (semantic)