Cluster Hypothesis in Low-Cost IR Evaluation with Different Document Representations

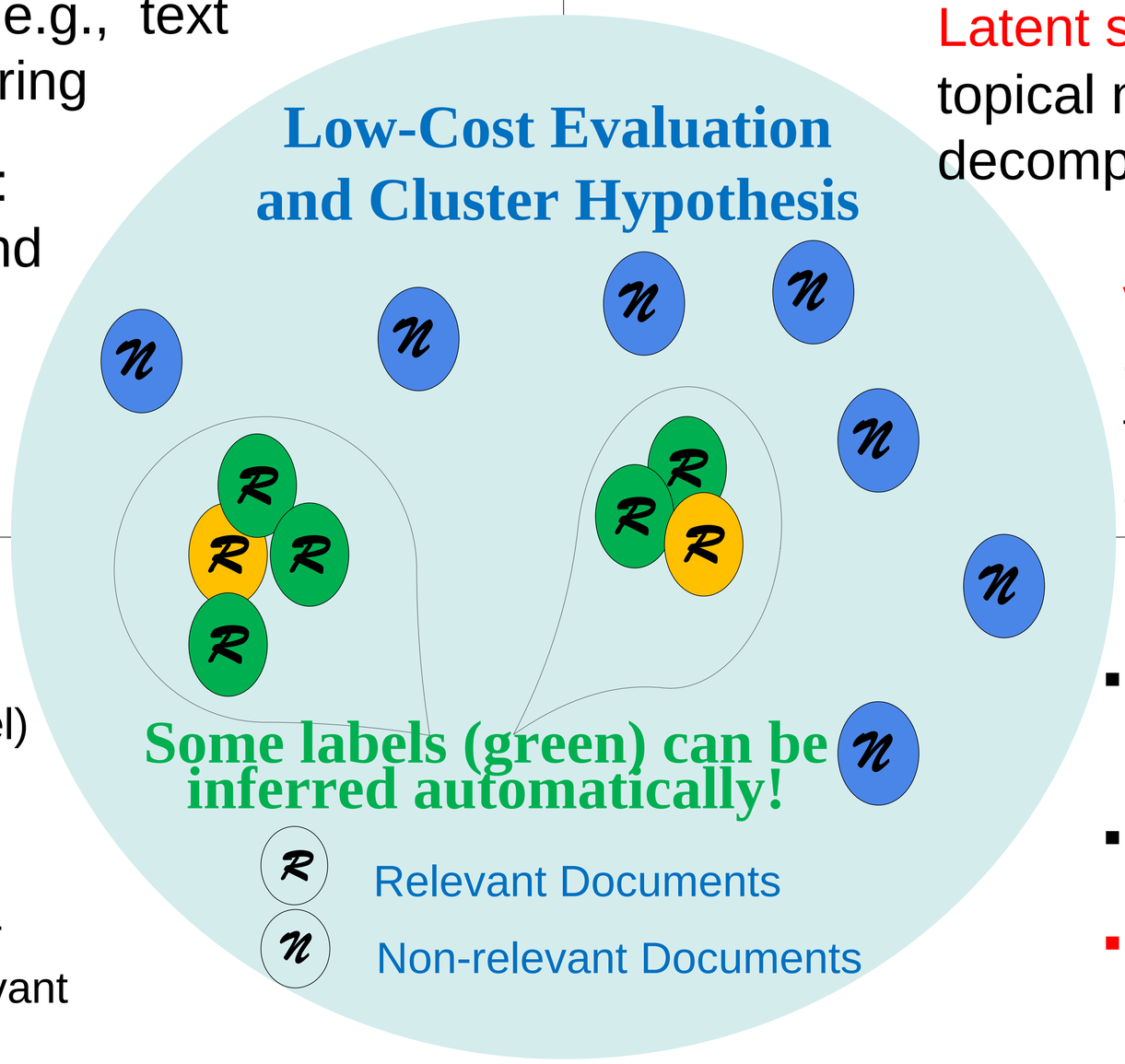
Kai Hui, Klaus Berberich
Max Planck Institute for Informatics

Motivation

Cluster hypothesis: documents that are relevant to the same query should be more similar with each other

Low-cost evaluation with cluster hypothesis: if the cluster hypothesis is satisfied, the manual assessments of document relevance can be done partially by automatic method, e.g., text classification or clustering

Two influential factors: similarity measures and **document representations**



Document Representations

Bag-of-words (*BOW*): with *tf-idf* weight

Expanded BOW with word embedding (*EBOW*): use the similarity matrix among word embedding to expand the document representation

Latent Dirichlet Allocation (*LDA*): generative topical modeling

Latent semantic analysis (*LSA*): topical modeling based on matrix decomposition

Neural network based vectorization (*Para2Vec*): co-train word embedding together with paragraph embedding as memory

Benchmarks

- **Triple Test**: comparison of similarity between (Rel, Rel) against similarity between (Rel, Non-Rel)
- **Knn Test**: the precision of relevance in **k** most similar documents of a given relevant document

Conclusions

- Agreement to the cluster hypothesis is not good enough for low-cost evaluation
- Improvement with word embedding is non-trivial
- **EBOW perform best:** red bold number indicates significance

Task	Benchmark	BOW	EBOW	LDA	LSA	Para2Vec
Adhoc	Triple Test	0.61	**0.62** (1.1%)	0.51 (-17%)	0.47 (-24%)	0.53 (-14%)
	Knn Test @5	0.62	0.62 (0.4%)	0.53 (-15%)	0.58 (-6%)	0.57 (-8%)
	Knn Test @20	0.54	0.54 (0.6%)	0.44 (-18%)	0.47 (-12%)	0.46 (-15%)
Diversity	Triple Test	0.49	**0.51** (4.5%)	0.43 (-13%)	0.44 (-10%)	**0.51** (4.1%)
	Knn Test @5	0.65	0.65 (-0.1%)	0.58 (-11%)	0.64 (-0.8%)	0.63 (-2.9%)
	Knn Test @20	0.56	0.56 (-0.1%)	0.51 (-8%)	0.54 (-3.5%)	0.54 (-4.5%)

MAX-PLANCK-GESELLSCHAFT

max planck institut informatik

contact: khui@mpi-inf.mpg.de