Unsupervised Annotation and Exploratory Search
Daniel Tunkelang – Chief Scientist, Endeca

ACM digital library
1 million documents; some with author-supplied tags

SOLUTION
Bootstrap on the author tags!
Prune 600K+ tag set to 10K:
- frequency threshold (freq > 10);
- normalization (case/number) (task = Task = tasks);
- subphrase elimination (feature > feature extraction).

RESULTS
- order of magnitude increase in relative recall;
- no discernible decrease in precision;
- increase in recall is sufficient to make the tags useful in supporting exploratory search.

PROBLEM
- tags: high precision, low recall
- 90% false negatives
- poor vocabulary reuse
- most tags unique
- useless for exploratory search

Tag the articles with this new vocabulary, using TF*IDF scores for matching terms (score > 90% of median).

old tags → controlled vocabulary → apply to all records → new tags

Need to increase recall without sacrificing precision!