Hierarchical Topic Detection

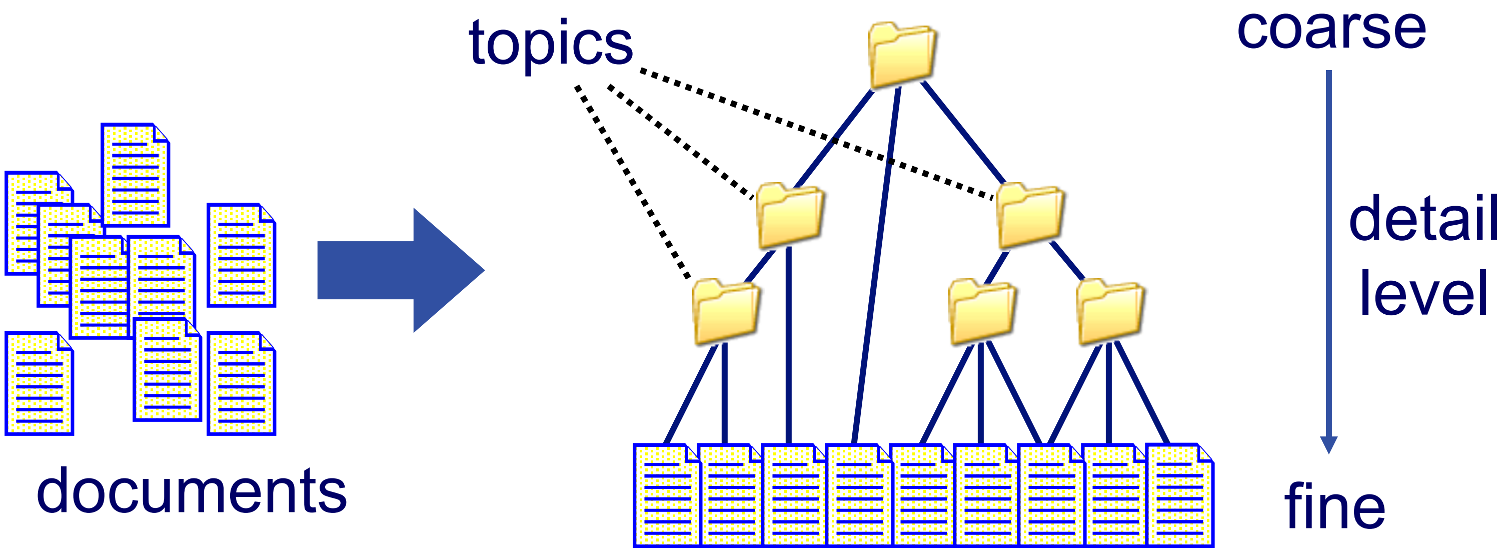
in large digital news archives

Dolf Trieschnigg (UT)
Wessel Kraaij (TNO)

University of Twente
The Netherlands

Hierarchical Topic Detection

- Conveniently group documents in a Yahoo like hierarchy, discussing topics in increasing level of detail:



Overview

- TDT evaluation program and HTD task
- Often used approach
- Our approach
- Experiments & results
- Conclusions & future work

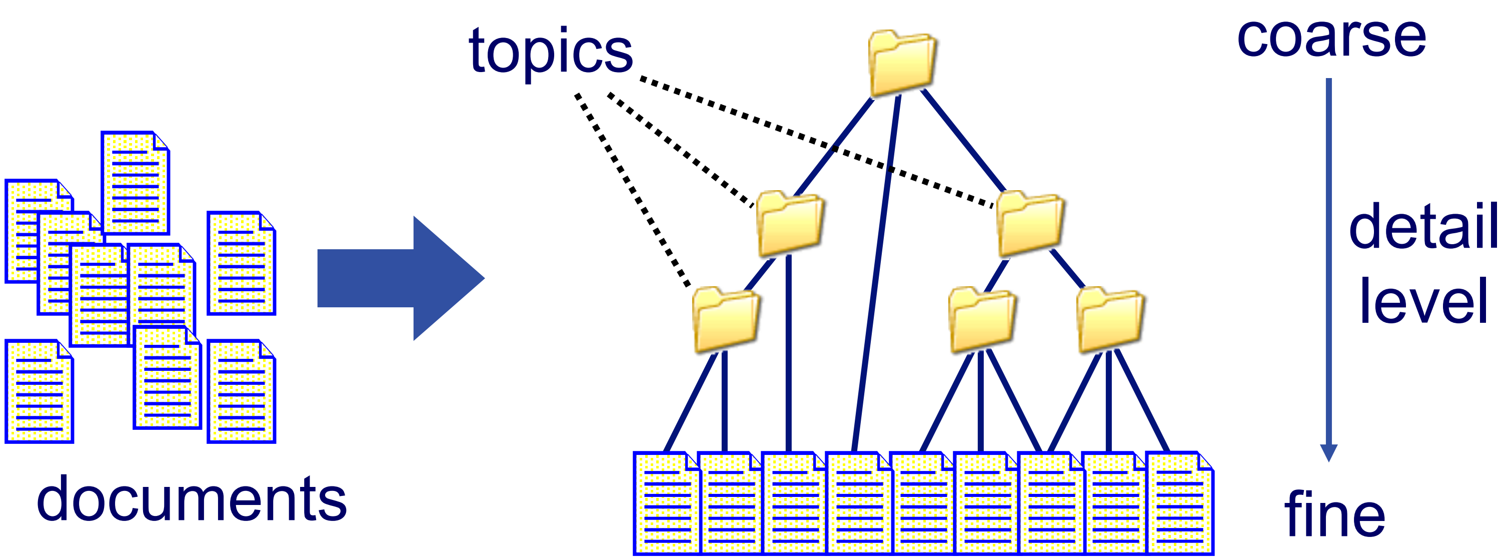
TDT evaluation program

- Discovering and threading together topically related material
- Old topic detection task
  - – hard, flat clustering (partitioning) of corpus
  - – shortcomings:
    - no overlapping clusters/topics
    - only one level of detail makes hard to evaluate: system detail vs. ground truth detail

→ new HTD task in 2004

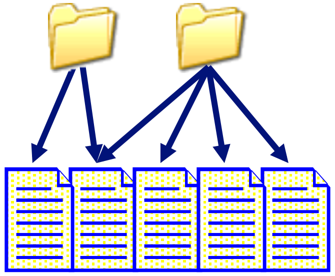
TDT 2004 new HTD task

- Multiple levels of detail
- Fuzzy (overlap between clusters)



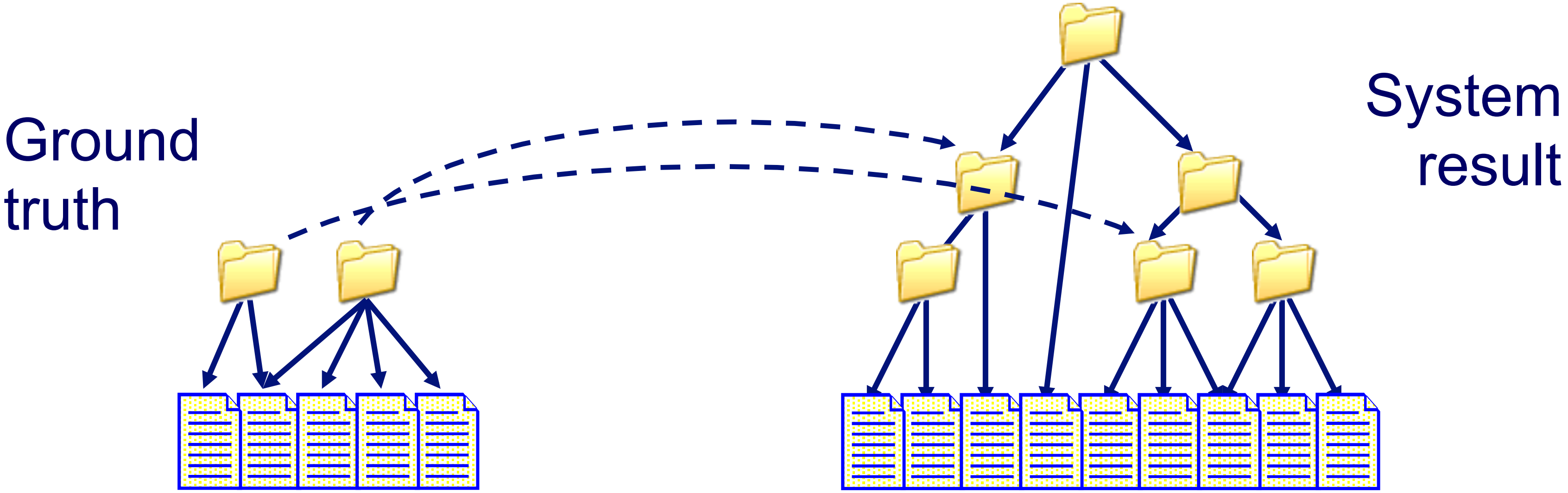
TDT5 corpus & ground truth statistics

- 400,000 multilingual documents
  - English, Arabic and Mandarin news wire
  - English machine translation available
- ground truth: 250 annotated topics
  - involving 9000 documents
  - average topic size: 52 docs (min: 1, max: 809, median: 16)
  - no hierarchy!



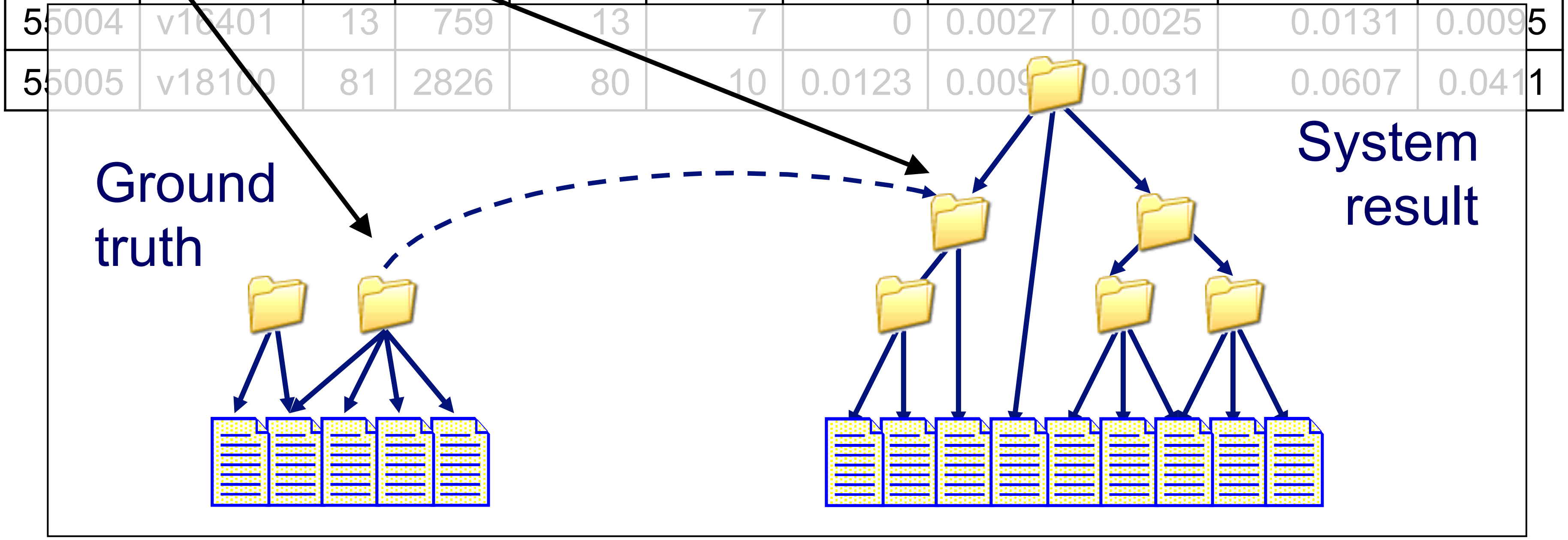
Evaluation

- Find system clusters with *minimal cost*:
  - – Detection cost (false alarms and misses)
  - – New: travel cost (to “find” the best cluster)



Evaluation example

Topic	System Cluster	# Ref	# Sys	# Union	Depth	PMiss	Pfa	Travel Cost	Detection Cost	Min. Cost
55001	v13965	5	261	5	8	0	0.0009	0.0028	0.0045	0.0039
55002	v15445	1	133	1	7	0	0.0005	0.0022	0.0023	0.0023
55003	v14140	27	133	27	11	0	0.0004	0.0035	0.0019	0.0024
55004	v16401	13	759	13	7	0	0.0027	0.0025	0.0131	0.0095
55005	v18100	81	2826	80	10	0.0123	0.009	0.0031	0.0607	0.0411



Often used approach

- Hierarchical agglomerative clustering:
  - Create distance matrix
    - distance metric: cosine, dice, jaccard etc.
    - documents as singleton clusters
  - Do…
    - Join most similar (least dissimilar) clusters
    - Calculate distances between new and existing clusters (different methods for single, complete and average link clustering)

  … until one cluster remains

Example: complete link

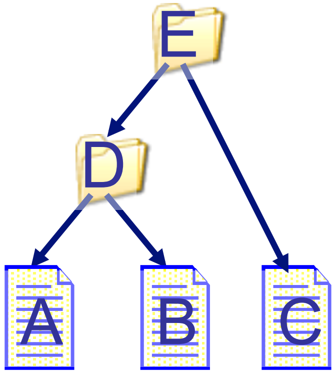
Symmetric dissimilarity matrix

	A	B
B	0.6	
C	0.8	0.7

complete: max

single: min

	D
C	0.8



Often used approach approach (cont’d)

- Hierarchical agglomerative clustering
  - Results in binary tree
  - Difficulties:
    - time complexity > $O(N^2)$
    - space complexity $O(N^2)$

→ unmodified not applicable for 400,000 document set

Our approach

- Cluster sample (20.000 documents)
  - O(N^2) still feasible

  → binary unbalanced cluster tree
- Optimize for cost metric
  - Rebranching the tree

  → more balanced cluster tree
- Assign remaining 380,000 documents to clusters obtained from sample

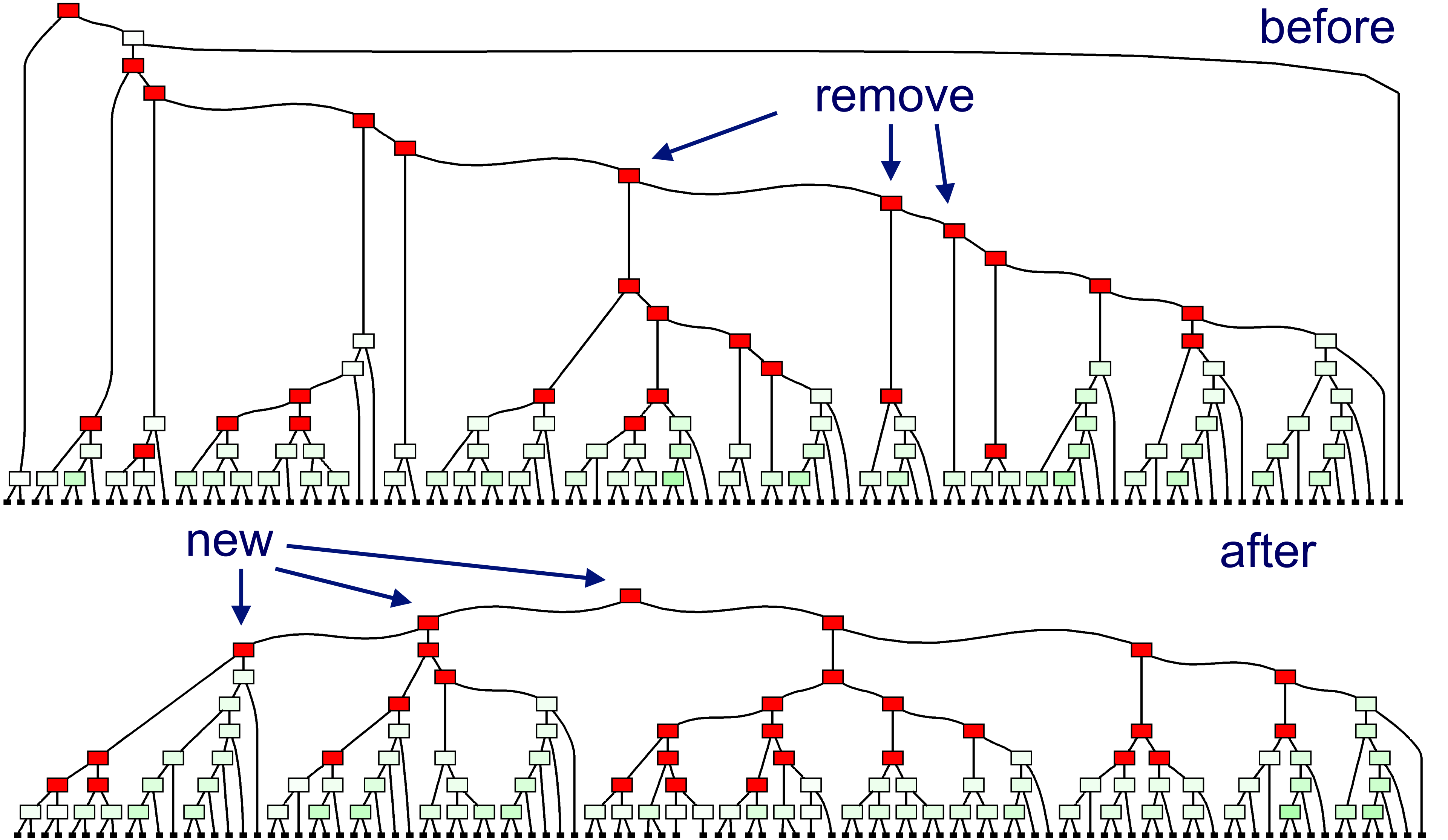
  → fuzzy cluster tree

Cluster sample

- Distance metric
  - Cross entropy reduction using background model of document collection
- Agglomerative hierarchical clustering
  - Experiments with complete, single and average linkage
- Results in a binary unbalanced tree

Optimize for cost metric

- Reduce travel cost without increasing detection cost
- Rebranch unbalanced tree:
  - remove clusters with dissimilarity value above certain threshold
  - combine “branches” of clusters in a better balanced tree with optimal (metric) branching factor



before

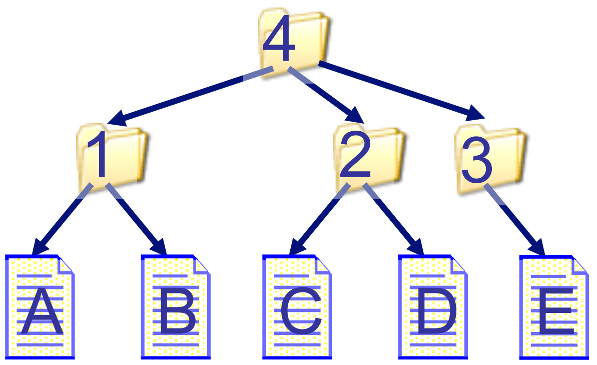
remove

new

after

Assigning remaining documents

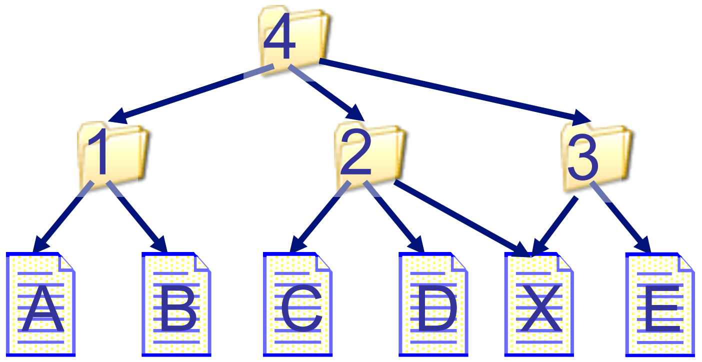
- Index sample
- Use remaining documents as queries
- Assign to clusters of best document-likelihood matches.
- Results in fuzzy cluster result



(sample documents)

X best matches D E

Result (add to cluster 2 and 3):



Experiments & results

- Experimented with cluster method
  - average link method gave best results
  - single link suffered from chaining
  - complete link suffered from “chaining”
    - rebranching improved results
- Adding documents to multiple clusters pays off: false alarm relatively cheap
- System performed best in TDT 2004

Results (sample)

Topic	System Cluster	# Ref	# Sys	# Union	Depth	PMiss	Pfa	Travel Cost	Detection Cost	Min. Cost
55001	v13965	5	261	5	8	0	0.0009	0.0028	0.0045	0.0039
55002	v15445	1	133	1	7	0	0.0005	0.0022	0.0023	0.0023
55003	v14140	27	133	27	11	0	0.0004	0.0035	0.0019	0.0024
55004	v16401	13	759	13	7	0	0.0027	0.0025	0.0131	0.0095
55005	v18100	81	2826	80	10	0.0123	0.0099	0.0031	0.0607	0.0411

Low precision
(not in cost)

High recall

Discussion

- Metric intuitive?
  - Travel cost not working out properly
    - Preferring balanced hierarchies
    - Preferring certain branching factor
    - Not discouraging fuzzy (powerset) clusters enough
    - How to judge hierarchy using non-hierarchical ground truth?:
  - Precision not important enough
- Is such a large hierarchy usable?
  - for cluster based retrieval?
  - for browsing and navigation of a large unlabelled dataset?

Conclusions and future work

- Sample based clustering method looks promising
  - How to improve precision?
  - Samples of different size: scalable?
  - Influence of distance metric?
- Evaluation metric should be improved
  - discouraging scattering documents
- How can it be made useful for browsing?

Questions?