

Authoritative Re-Ranking in Fusing Authorship-Based Subcollection Search Results

Toine Bogers & Antal van den Bosch

Tilburg University

DIR 2006

March 14th, 2006

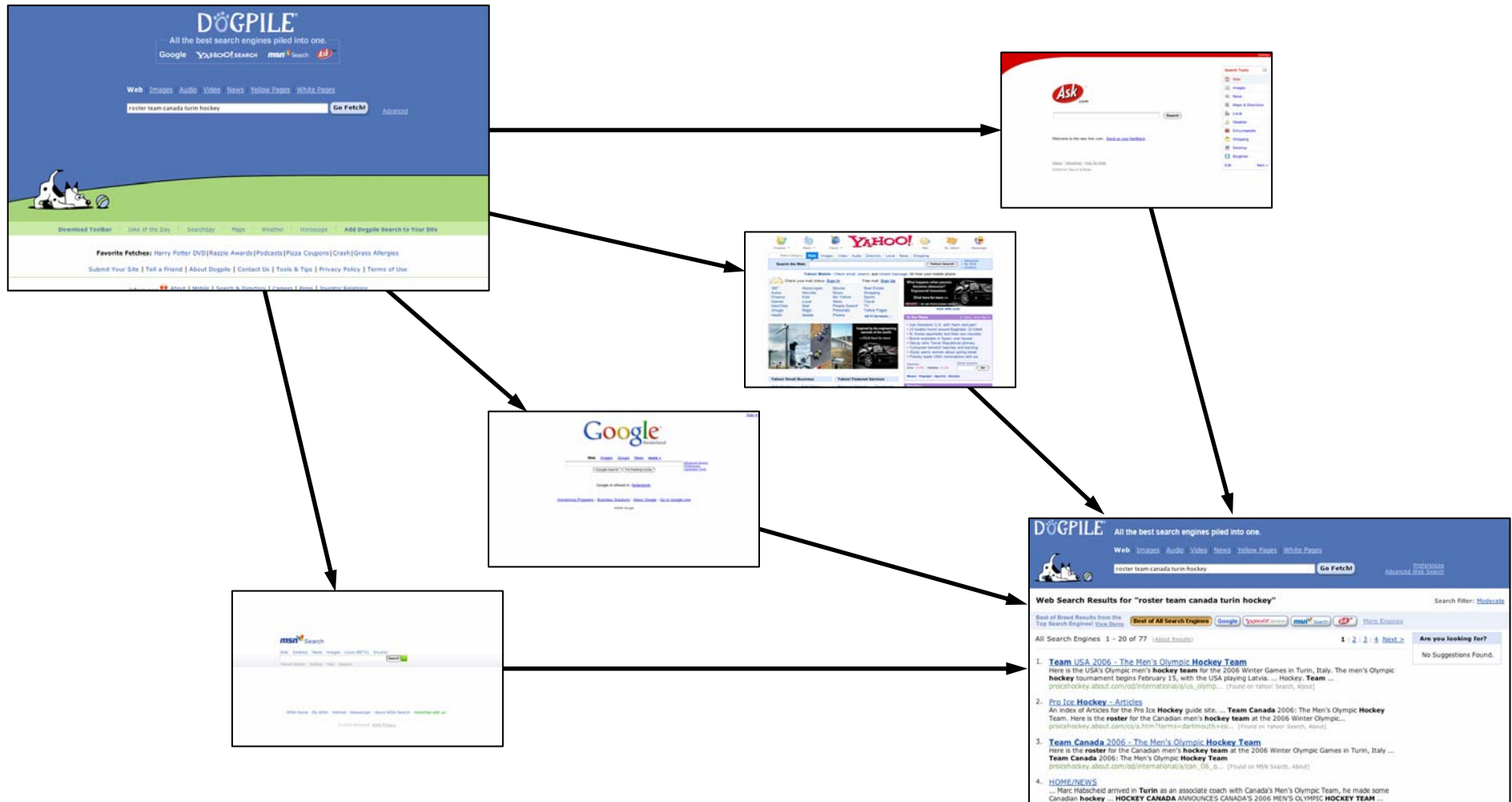


Introduction

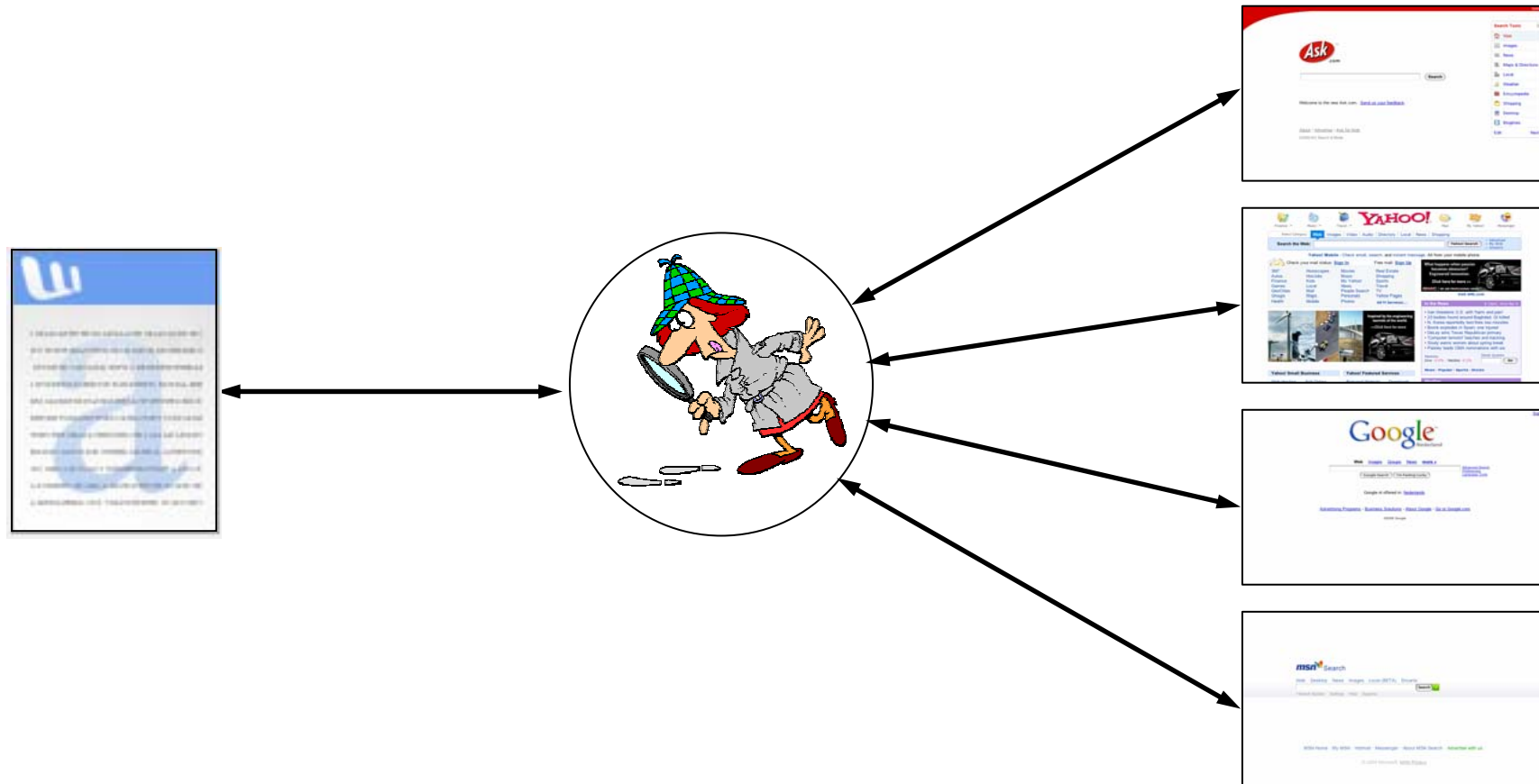
Parsing the title

Authoritative Re-Ranking Fusing Authorship-
Based Subcollection Search Results

Distributed IR



Distributed IR



Distributed IR

Challenge

Approximate the performance of the hypothetical scenario where all the covered documents are in a single collection.

Subproblems

- database selection
- query translation
- document selection
- merging results

Distributed IR

Past solutions

- Voorhees et al. (1995)
- Baumgarten (1999)
- Callan et al. (1995)
 - use an inference network to rank the different collections
 - combine collection-weights with the document weights assigned by each separate retrieval engine
 - documents from good collections but also good documents from poor collections are ranked high

Within a single collection

Identifying subcollections

- by topic
 - clustering of documents
 - separate problem
 - clustering not always very reliable

Within a single collection

Old and new challenges

- database selection
 - cost of referencing each subcollection is equal
 - every author in the community should be considered
 - problem of database selection is not relevant
- query translation
- document selection
- merging results

Within a single collection

Old and new challenges

- database selection
- query translation
 - same approach is used for retrieval in all subcollections
 - query translation is not an issue
- document selection
- merging results

Within a single collection

Old and new challenges

- database selection
- query translation
- document selection
 - a document can have multiple authors
 - documents can belong to multiple subcollections
 - document selection needs to be addressed
- merging results

Within a single collection

Old and new challenges

- database selection
- query translation
- document selection
- merging results
 - results of different subcollections need to be combined properly
 - results merging essential as well
 - but: no problems with different or missing rankings

Within a single collection

Old and new challenges

- database selection
- query translation
- document selection
- merging results
- **NEW PROBLEM: no more approximation!**

Authoritative re-ranking

Assumptions

- an author's expertise is implicitly represented by his or her documents
 - expertise based on content & comparison
 - terms that distinguish between authors are representative for their expertise

Authoritative re-ranking

Test collections

- in context of IMAs → collections that represent some kind of community
 - workgroups or scientific communities
- author labels

Authoritative re-ranking

Two components

- document weights (baseline)
 - baseline, based on document-query cosine similarity
- subcollection weights
 - based on author expertise

Document weights

Baseline

- simple Vector Space model based on TF-IDF
 - stopword filtering
 - Porter stemming
 - statistical phrases
 - syntactic phrases (chunking)
 - cosine similarity

Subcollection weights

Informative terms

- how much does encountering term help in predicting the author?
- feature selection metrics
 - Information Gain Chi-Square, Mutual Information the average TF-IDF value
 - one-vs-all datasets
- informativeness weights
for each term:

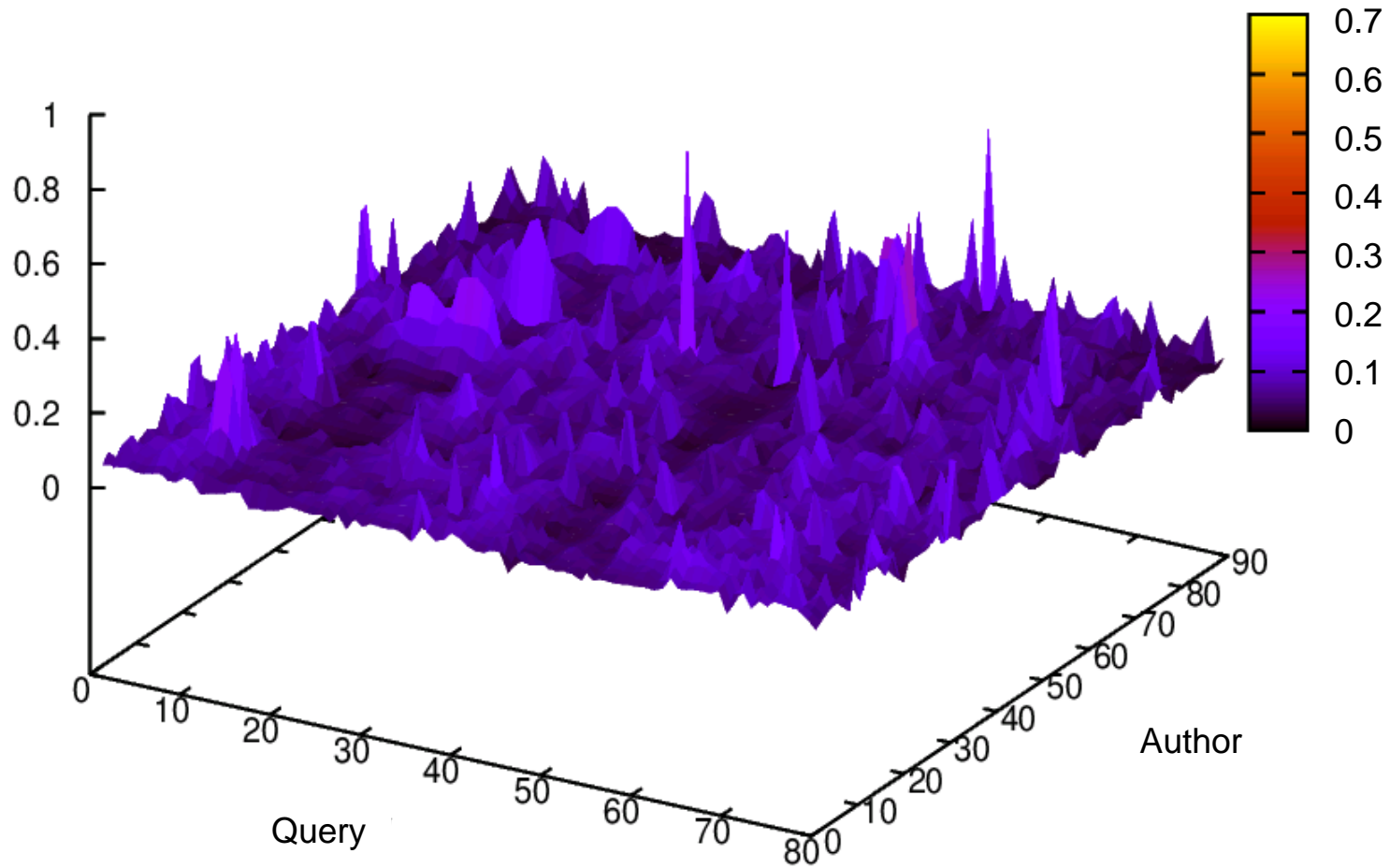
Subcollection weights

Expertise ranking

- calculate informativeness weights for each author-term pair
- collect informativeness weights for each query term
- calculate expert score
 - unweighted average
 - TF-IDF-weighted average
- rank authors on expert score/subcollection weight

Subcollection weights

Expertise



Document selection

De-duplication

- documents with multiple authors
 - in more than one subcollection
 - so more than one expert score
- combine expert scores into a single measure of *suitability*
 - unweighted average
 - weighted average using collection size (= document count)
 - weighted average**
 - natural log
 - square root

Merging results

Authoritative re-ranking

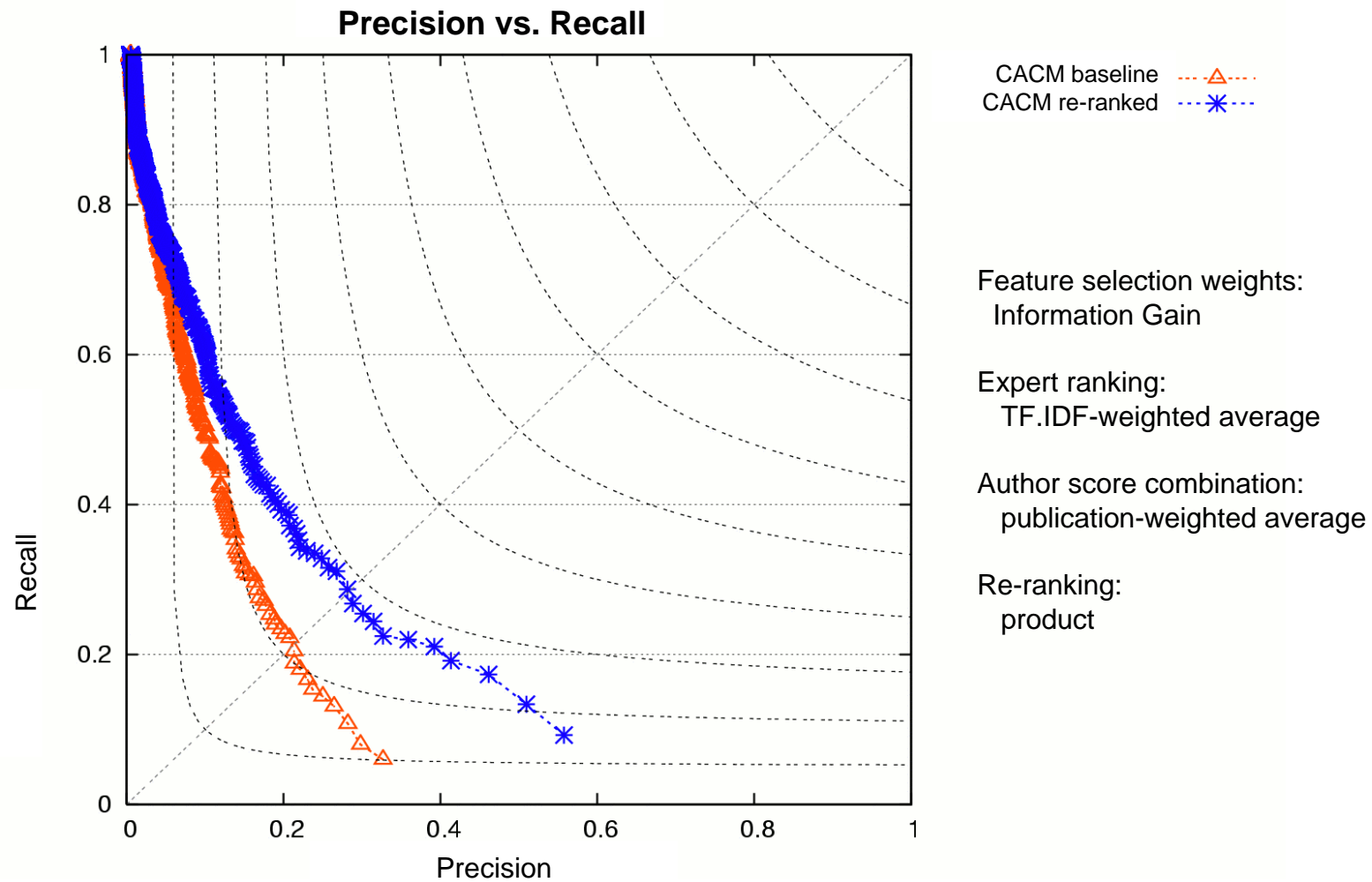
- combine suitability score with baseline similarity scores
 - re-ranking
 - product (similar to Callan et al. (1995))
 - average
 - $(\text{suitability} + 1) \times \text{similarity}$

Results

R-precision scores

collection	re-ranked	baseline	
CACM	0.313	0.233	(+34.3%)
CISI	0.203	0.201	(+1.5%)
ILK	0.649	0.647	(+0.3%)

Results



Results

CACM - Query-by-query improvements



Author rank experiments

Special versions of collections

- small experiments with special versions
 - only the first author (CACM-first, CISI-first, etc.)
 - every author except the last (CACM-m1, etc.)

Author rank experiments

R-precision scores

collection	re-ranked	baseline	
CACM	0.313	0.233	(+34.3%)
CACM-first	0.302		(+20.2%)
CACM-m1	0.304		(+30.5%)
CISI	0.203	0.201	(+1.5%)
CISI-first	0.203		(+1.5%)
CISI-m1	0.203		(+1.5%)
ILK	0.649	0.647	(+0.3%)
ILK-first	0.650		(+0.5%)
ILK-m1	0.656		(+1.4%)

Conclusions

Does authoritative re-ranking work?

- Yes, but...
 - significant improvements on the two largest collections
 - no significant improvements on the other collection
 - improvements highly dependent on collection and collection size

Which authors?

- using all authors appears to yield the best results

Future work

Better expertise identification

- citation analysis to include authority as well as expertise
- gold standard of ILK expertise

Questions?



Example

One particular example from CACM

- query 0006:

Interested in articles on robotics, motion planning particularly the geometric and combinatorial aspects. We are not interested in the dynamics of arm motion.

- increase in R-precision: $0.0 \rightarrow 0.67$
- 3 relevant documents \rightarrow R-precision calculated over first 3 returned documents

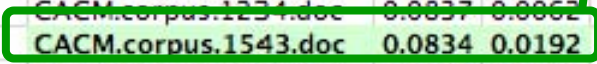
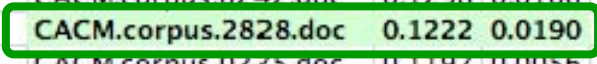
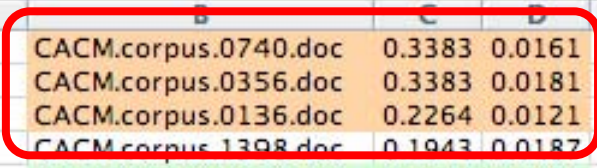
CACM.query-0006-comparison.xls

	A	B	C	D	E	F	G	H	I	J
1	CACM.corpus.0006.query	CACM.corpus.0740.doc	0.3383	0.0161		CACM.corpus.0695.doc	0.1727	0.0352		
2	CACM.corpus.0006.query	CACM.corpus.0356.doc	0.3383	0.0181		CACM.corpus.1543.doc	0.0834	0.0192		
3	CACM.corpus.0006.query	CACM.corpus.0136.doc	0.2264	0.0121		CACM.corpus.2828.doc	0.1222	0.0190		
4	CACM.corpus.0006.query	CACM.corpus.1398.doc	0.1943	0.0187		CACM.corpus.1398.doc	0.1943	0.0187		
5	CACM.corpus.0006.query	CACM.corpus.0695.doc	0.1727	0.0352		CACM.corpus.0356.doc	0.3383	0.0181		
6	CACM.corpus.0006.query	CACM.corpus.1148.doc	0.1373	0.0086		CACM.corpus.1664.doc	0.0802	0.0162		
7	CACM.corpus.0006.query	CACM.corpus.0279.doc	0.1321	0.0062		CACM.corpus.0740.doc	0.3383	0.0161		
8	CACM.corpus.0006.query	CACM.corpus.2804.doc	0.1304	0.0124		CACM.corpus.2826.doc	0.1062	0.0135		
9	CACM.corpus.0006.query	CACM.corpus.1256.doc	0.1301	0.0039		CACM.corpus.2804.doc	0.1304	0.0124		
10	CACM.corpus.0006.query	CACM.corpus.0242.doc	0.1298	0.0106		CACM.corpus.0136.doc	0.2264	0.0121		
11	CACM.corpus.0006.query	CACM.corpus.2828.doc	0.1222	0.0190		CACM.corpus.0242.doc	0.1298	0.0106		
12	CACM.corpus.0006.query	CACM.corpus.0275.doc	0.1197	0.0056		CACM.corpus.2873.doc	0.0875	0.0099		
13	CACM.corpus.0006.query	CACM.corpus.0402.doc	0.1175	0.0063		CACM.corpus.1148.doc	0.1373	0.0086		
14	CACM.corpus.0006.query	CACM.corpus.0438.doc	0.1118	0.0060		CACM.corpus.0888.doc	0.0886	0.0085		
15	CACM.corpus.0006.query	CACM.corpus.2826.doc	0.1062	0.0135		CACM.corpus.0887.doc	0.0882	0.0085		
16	CACM.corpus.0006.query	CACM.corpus.3061.doc	0.1051	0.0032		CACM.corpus.1113.doc	0.0530	0.0067		
17	CACM.corpus.0006.query	CACM.corpus.2553.doc	0.1008	0.0032		CACM.corpus.2386.doc	0.0650	0.0065		
18	CACM.corpus.0006.query	CACM.corpus.0317.doc	0.0944	0.0027		CACM.corpus.0605.doc	0.0667	0.0065		
19	CACM.corpus.0006.query	CACM.corpus.1541.doc	0.0934	0.0057		CACM.corpus.0402.doc	0.1175	0.0063		
20	CACM.corpus.0006.query	CACM.corpus.2502.doc	0.0926	0.0030		CACM.corpus.1234.doc	0.0837	0.0062		
21	CACM.corpus.0006.query	CACM.corpus.0888.doc	0.0886	0.0085		CACM.corpus.0279.doc	0.1321	0.0062		
22	CACM.corpus.0006.query	CACM.corpus.0705.doc	0.0886	0.0043		CACM.corpus.0438.doc	0.1118	0.0060		
23	CACM.corpus.0006.query	CACM.corpus.0887.doc	0.0882	0.0085		CACM.corpus.1541.doc	0.0934	0.0057		
24	CACM.corpus.0006.query	CACM.corpus.0704.doc	0.0882	0.0042		CACM.corpus.0275.doc	0.1197	0.0056		
25	CACM.corpus.0006.query	CACM.corpus.2873.doc	0.0875	0.0099		CACM.corpus.1046.doc	0.0326	0.0044		
26	CACM.corpus.0006.query	CACM.corpus.1234.doc	0.0837	0.0062		CACM.corpus.2254.doc	0.0467	0.0044		
27	CACM.corpus.0006.query	CACM.corpus.1543.doc	0.0834	0.0192		CACM.corpus.2667.doc	0.0692	0.0043		
28	CACM.corpus.0006.query	CACM.corpus.1664.doc	0.0802	0.0162		CACM.corpus.0705.doc	0.0886	0.0043		
29	CACM.corpus.0006.query	CACM.corpus.2514.doc	0.0792	0.0037		CACM.corpus.0704.doc	0.0882	0.0042		
30	CACM.corpus.0006.query	CACM.corpus.2836.doc	0.0713	0.0034		CACM.corpus.1939.doc	0.0390	0.0041		
31	CACM.corpus.0006.query	CACM.corpus.2667.doc	0.0692	0.0043		CACM.corpus.2786.doc	0.0545	0.0040		
32	CACM.corpus.0006.query	CACM.corpus.0605.doc	0.0667	0.0065		CACM.corpus.1554.doc	0.0541	0.0040		
33	CACM.corpus.0006.query	CACM.corpus.2386.doc	0.0650	0.0065		CACM.corpus.1256.doc	0.1301	0.0039		
34	CACM.corpus.0006.query	CACM.corpus.0859.doc	0.0637	0.0028		CACM.corpus.0971.doc	0.0473	0.0037		
35	CACM.corpus.0006.query	CACM.corpus.1213.doc	0.0630	0.0019		CACM.corpus.2514.doc	0.0792	0.0037		
36	CACM.corpus.0006.query	CACM.corpus.1026.doc	0.0590	0.0016		CACM.corpus.1365.doc	0.0494	0.0037		

RELEVANT

RELEVANT

RELEVANT



Introduction

Affiliations

- member of ILK, Tilburg University
 - machine learning for NLP
 - information extraction & retrieval
- PhD student in À Propos project
 - collaboration between Tilburg University, Radboud University (Nijmegen) and 5 companies
 - create the ultimate research assistant (precision & presentation)

Results

cutoff	CACM baseline			CACM re-ranked		
	P	R	F	P	R	F
1	0.32692	0.05985	0.10118	0.55769	0.09230	0.15839
2	0.29808	0.07976	0.12585	0.50962	0.13360	0.21170
3	0.28205	0.10769	0.15587	0.46154	0.17334	0.25203
4	0.26442	0.13069	0.17492	0.41346	0.19177	0.26201
5	0.25000	0.14405	0.18278	0.39231	0.21059	0.27406
6	0.23718	0.15319	0.18615	0.35897	0.21972	0.27259
7	0.23077	0.16619	0.19323	0.32692	0.22447	0.26618
8	0.22115	0.17967	0.19826	0.31490	0.24411	0.27502
9	0.21367	0.18810	0.20007	0.30128	0.25452	0.27593
10	0.21346	0.20426	0.20876	0.28846	0.26787	0.27778
11	0.20804	0.22220	0.21489	0.28147	0.28679	0.28411
12	0.20192	0.22734	0.21388	0.26763	0.31106	0.28772
13	0.19675	0.23388	0.21371	0.25740	0.31581	0.28363
...						
R	0.23291	0.23291	0.23291	0.31275	0.31275	0.31275

Distributed IR

Past solutions

- Voorhees et al. (1995)
 - cluster a set of training queries
 - clustering on overlap in relevant retrieved documents
 - new queries are matched to cluster centroids
 - weights of best match are used to determine how many documents are retrieved from each collection
- Callan et al. (1995)
- Baumgarten (1999)

Distributed IR

Past solutions

- Voorhees et al. (1995)
- Callan et al. (1995)
- Baumgarten (1999)
 - completely probabilistic approach
 - less dependent on heuristics