A Decade after TREC-4

NTCIR-5 CLIR-J-J Experiments at Yahoo!Japan

9 Dec 2005
Sumio FUJITA
Yahoo Japan Corporation

Introduction

- Automatic feedback from top k documents strategy
  - Dates back to the TREC-2.
  - Was especially successful in the TREC-4.
- As much as +41.7% gain in NTCIR-5 CLIR-J-J
  - our best official TITLE only run vs its no feedback baseline run
  - This is really exceptional!

A Retrospective Study of the top k document feedback strategy

- TREC-2(1993)
  - w*r/R(Okapi), Thesaurus Extaction(Claritech), wpq(UCLA)
- TREC-3(1994)
  - Elimination of “concepts” fields accelerates feedback strategies.
  - Okapi: “unexpectedly successful” improvement of 19.1% by 40terms from 30 documents
- TREC-4(1995)
  - SMART: +27% improvement by 50 single terms and 10 phrases from 20 documents
  - PIRCS: +29% improvement by expanding with 50 terms from 40 subdocuments

Test collection	MAP Rigid	PFB Gain %	MAP Relax	PFB Gain%
NTCIR-1 Adhoc DESC run	0.3596 0.3227	+11.4	- -	-
NTCIR-3 CLIR J-J TITLE query Rigid /Relax	0.3930 0.3292	+19.4	0.4502 0.3830	+17.5
NTCIR-4 CLIR J-J TITLE query Rigid /Relax	0.3801 0.3090	+23.0	0.4711 0.3956	+19.1
NTCIR-3 Patent Desc query A / AB	0.3283 0.2846	+15.4	0.3209 0.2811	+14.2
NTCIR-4 Patent Claim query A / AB	0.2508 0.2290	+9.5	0.1655 0.1549	+6.8
TREC-9 Web Title run	- -	-	0.2028 0.1751	+15.8
TREC-2001 Web Title run	- -	-	0.2060 0.1704	+20.9
TREC 2004 MEDLINE Long query DR /DR+PR	0.3695 0.3526	+4.8	0.4075 0.3915	+4.1

System description

- YLMS evaluation experiment system based on Lemur toolkit 2.0.1[Ogilvie et al. 02] for indexing system
- Indexing language:
  - Chasen version 2.2.9 as Japanese morphological analyzer with IPADIC dictionary version 2.5.1
- Retrieval models:
  - TF*IDF with BM25 TF
  - KL-divergence of probabilistic language models with Dirichlet prior smoothing[Zhai et al. 01]
- Rocchio feedback for TF*IDF and mixture model feedback method for KL-divergence retrieval model [Zhai et al. 01]

Language modeling for IR

$$p(d \mid q) \propto p(d)\,p(q \mid d)$$

$$\log(p(d)\,p(q \mid d)) = \log p(d) + \sum_i \log p(q_i \mid d)$$

$$\sum_{w \in V} p(w \mid q) \log(p(w \mid d))$$

Negative cross entropy between the query language model and a document language model

- Dirichlet-Prior method smoothing methods

$$p_\mu(w \mid d) = \frac{freq(w, d) + \mu p(w \mid C)}{|d| + \mu}$$

CLIR J-J experiments

- Title or Description Only runs: simple TF*IDF with a top k document feedback strategy
- Title and Description runs: Fusion of Title run and Description run
- Post submission: KL-divergence runs(Dirichlet smoothing, KL-Dir) with/without feedback

$$w(d,t,k1,b,k4) = \left(k4 + \log \frac{N}{df(t)}\right) \frac{(k1+1)\,freq(d,t)}{k1\left((1-b) + b\frac{dl_d}{avdl}\right) + freq(d,t)}$$

d : document

t : term

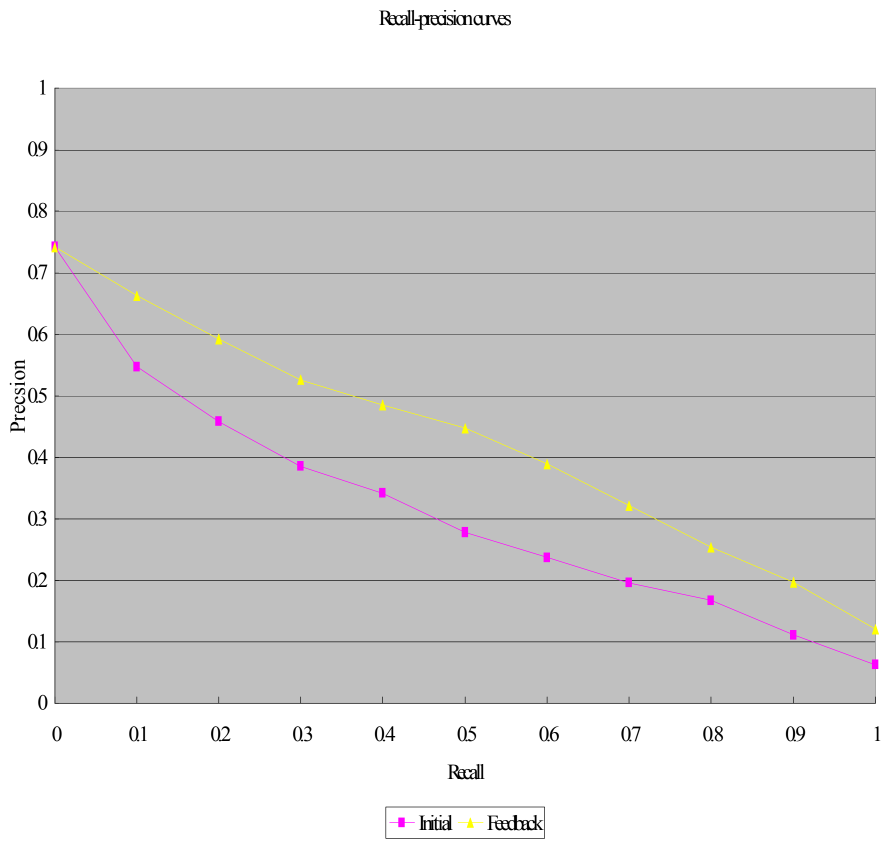
N : total number of documents in the collection

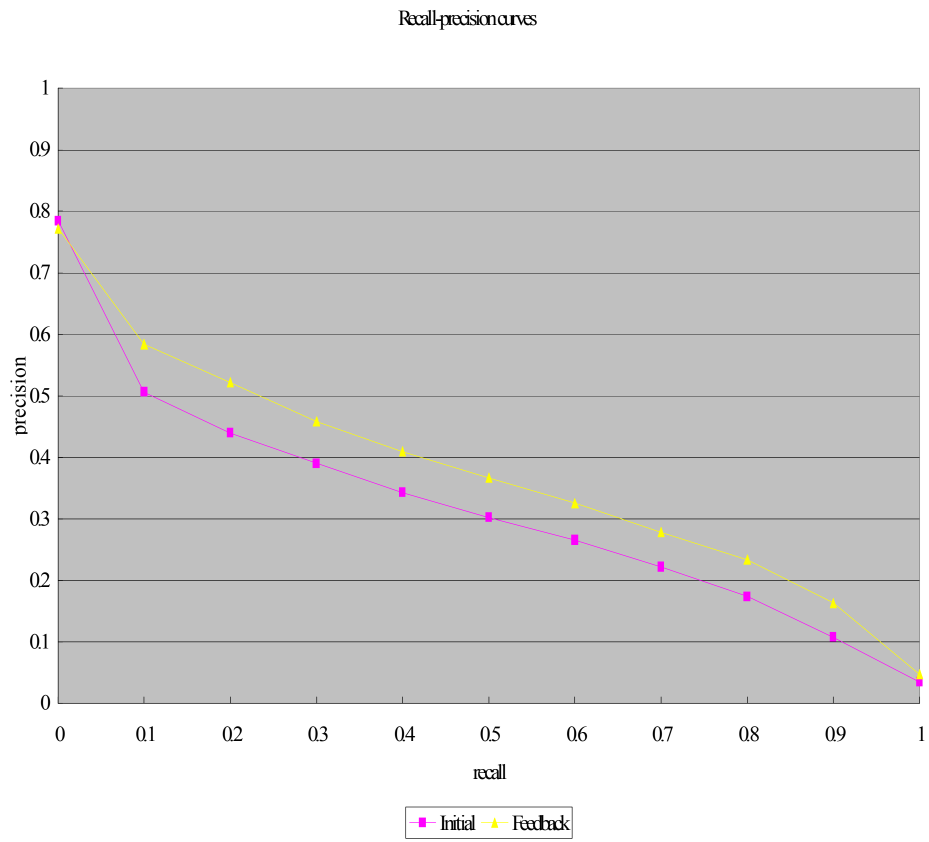
$df(t)$: number of documents where t appears

$freq(d,t)$: number of occurrence of t in d

$k1, k4, b$: parameters

Recall-precision curves
NTCIR-5 CLIR-J-J VS NTCIR-4 CLIR-J-J





Title only run feedback / baseline effectiveness BM25 TF*IDF / KL-Dir

	MAP-Rigid	RP-Rigid	Rel-Ret	P@10	P@20	MAP-Relax	RP-relax	Rel-Ret	P@10	P@20
YLMS-J-J-T-03	0.4193	0.4250	1959	0.5277	0.4309	0.5028	0.4911	3844	0.6915	0.6128
%gain	+41.7	+29.2	+12.3	+29.8	+29.0	+32.9	+24.6	+11.6	+18.6	+19.8
YLMS-J-J-T-03 No FB	0.2960	0.3289	1745	0.4064	0.3340	0.3782	0.3940	3444	0.5830	0.5117
KL-Dir Mix FB	0.4134	0.4174	1902	0.5128	0.4277	0.4874	0.4811	3744	0.6702	0.5926
%gain	+40.4	+33.0	+11.3	+28.9	+25.6	+29.0	+21.8	+10.2	+15.0	+18.3
KL-Dir No FB	0.2944	0.3139	1709	0.3979	0.3404	0.3779	0.3951	3396	0.5830	0.5011

- Some correlation factors between measures on topic by topic basis

Initial AP vs Feedabck AP: 0.778

Initial AP vs Feedback gain: -0.434

Feedback AP vs Feedback gain: 0.019

Initial 5-precision vs Feedback gain: -0.139

Our hypotheses: Top k document feedback strategy is especially successful when:

- Short query
  - Feedback gain is emphasized when the original queries are short and terminologically not so rich.
- Terminologically controlled and “clean” document collections such as newspapers or newswires
  - The strategy is not straightforwardly applicable to web documents, where the gain is smaller.
- The document collections are repeatedly used in the preceding workshops.
  - The repeated use of the document collections or similar collections uncovers the collection characteristics and the task practitioners can afford to take an aggressive strategy.
- Sufficient number of relevant documents
  - In order to achieve improvements, there should be some relevant documents to be promoted, which have retrieved at lower ranks in the pilot search.

Our hypotheses presumably hold true because:

- Our NTCIR-1, TREC-9 TREC-2004 experiments show that the k document feedback strategy gets more improvements when the initial query is short and poor.
- It seems to be more effective with clean documents:
  - Newspaper collections (TREC-3,4 NTCIR-3,4,5) vs Web collection (TREC-9, 2001)
- Presumably it is more effective when the document collection is repeatedly used.
  - TREC-2, TREC-3 < TREC-4
  - NTCIR-3,4 CLIR-J-J < NTCIR-5 CLIR-J-J
- But the number of relevant documents does not seem to affect the % gain?
  - NTCIR-3 CLIR J-J: 19.4% 1654 rel docs
  - NTCIR-4 CLIR J-J: 23.0% 7137 rel docs
  - NTCIR-5 CLIR J-J: 41.7% 2112 rel docs

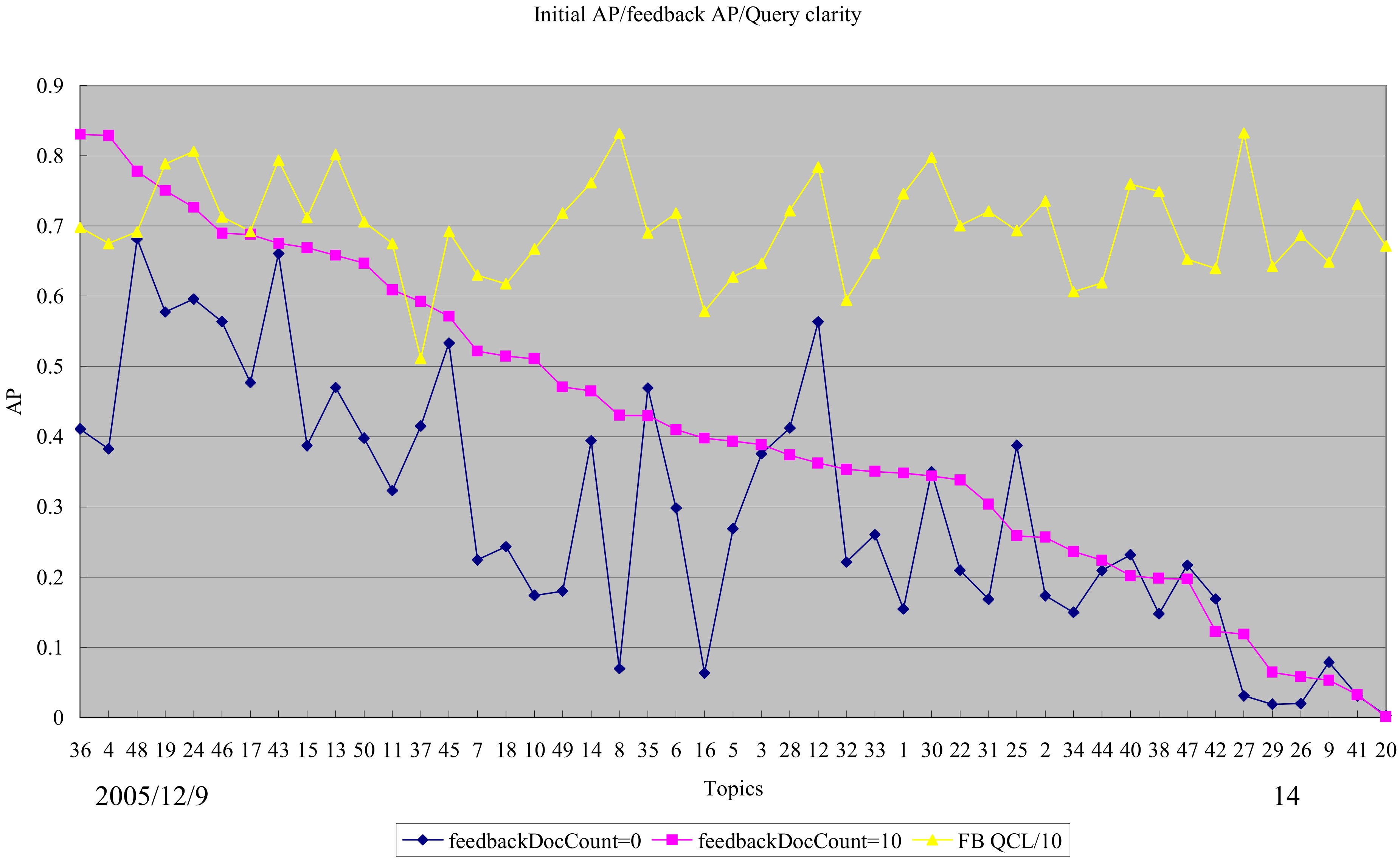
Too many relevant documents cause topic divergence

- Eguchi et al (2002) showed different behaviors of search engines according to the topic difficulty.
  - Our aggressive feedback run performed better with difficult topics.
- Does an aggressive feedback strategy perform better in difficult topics?
  - No correlation between feedback AP and % feedback gain: 0.019
- Through NTCIR-3 to NTCIR5 CLIR J-J, feedback gains are larger when evaluated by rigid relevance criteria.
- What does this mean?
  - Certain levels of term cohesion is necessary among relevant documents for feedback improvement.
  - Relax relevant documents are topically too diverse to achieve improvement by a feedback while the feedback narrows down the query topics adding more terms.

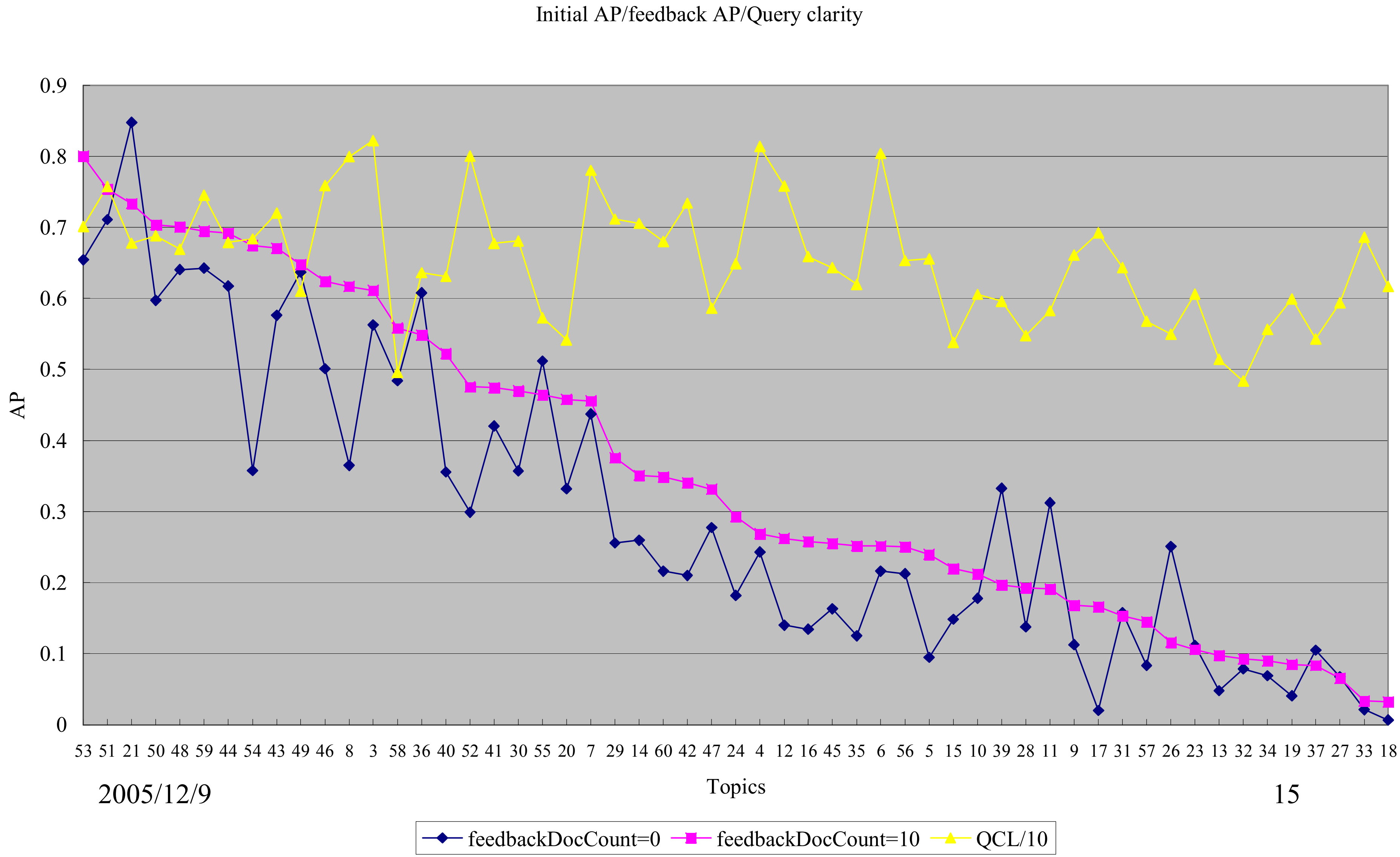
Feedback Document Clarity Test

- Query Clarity measure by Cronen-Townsend et al.(2002)
- KL-Divergence between the query and the collection language model
- We computes KL-Divergence between feedback documents models and the collection language model
- This may indicate the topic cohesion, but….
- Very weak or no correlation on a topic by topic basis in NTCIR-5 CLIR J-J
  - Query clarity vs Feedback AP : 0.117
  - Query clarity vs Feedback Gain : 0.006
- Moderate correlation on a topic by topic basis in NTCIR-4 CLIR J-J
  - Query clarity vs Feedback AP : 0.46
  - Query clarity vs Feedback Gain : 0.057

Average precision of initial (pilot search) run / feedback run / Query clarity of feedback language models of NTCIR-5 CLIR J-J



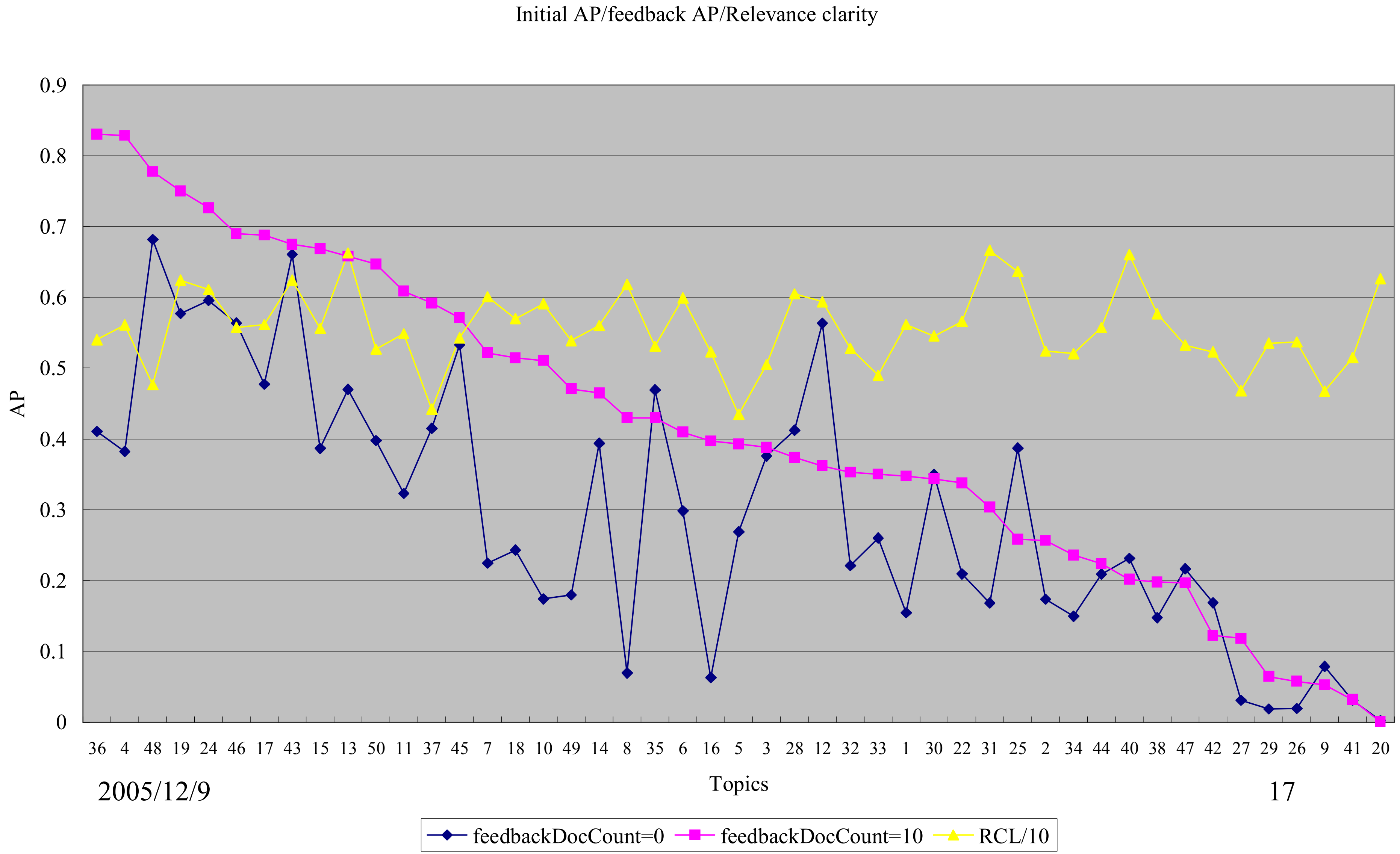
Average precision of initial (pilot search) run / feedback run / Query clarity of feedback language models of NTCIR-4 CLIR J-J



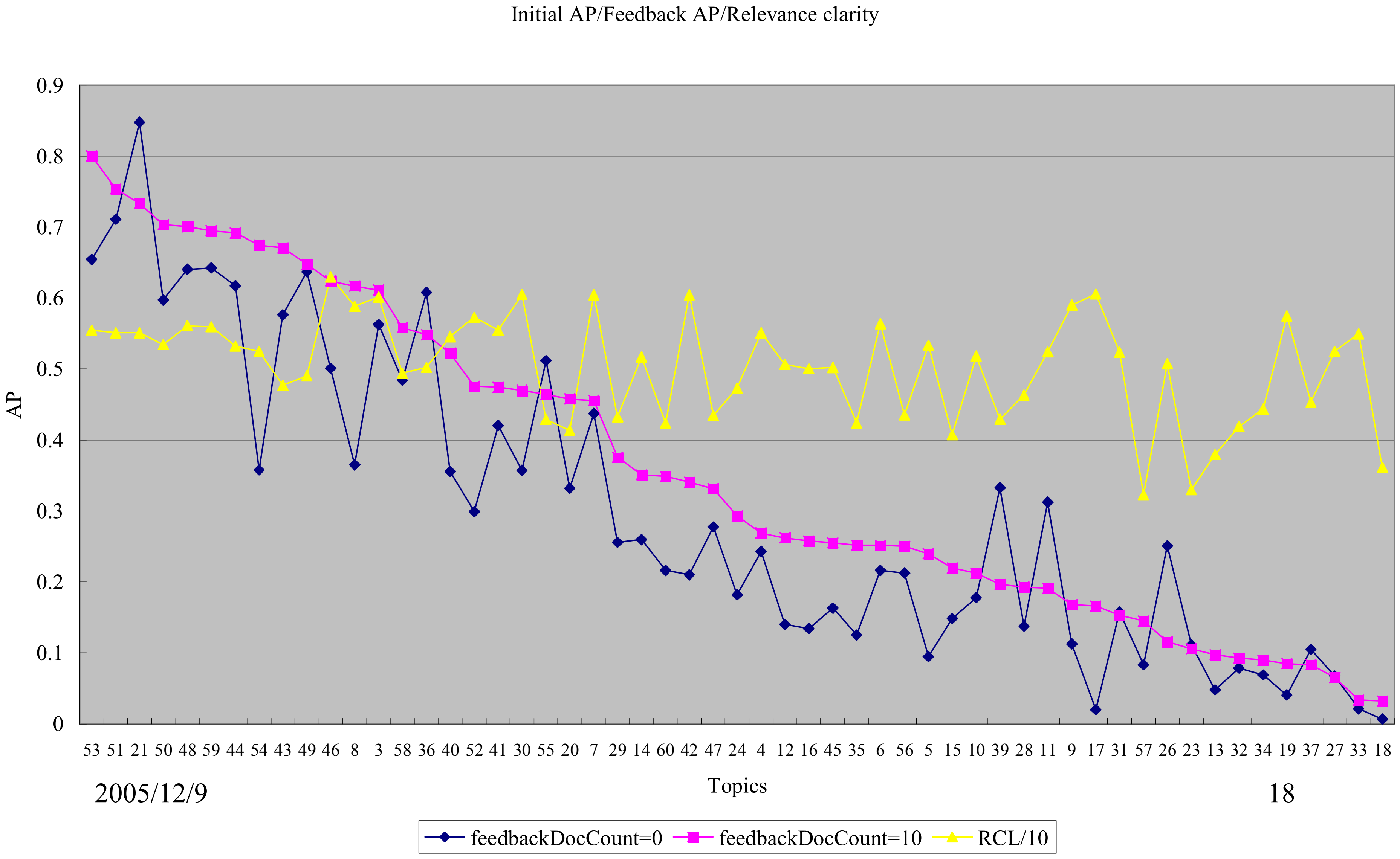
Relevance Clarity Test

- KL-Divergence between the relevant documents and collection language models
- Very weak or no correlation with NTCIR-5 CLIR-J-J
  - Relevance clarity vs Feedback AP : 0.155
  - Relevance clarity vs Feedback Gain : -0.058
- Moderate correlation in NTCIR-4 CLIR J-J
  - Query clarity vs Feedback AP : 0.411
  - Query clarity vs Feedback Gain : 0.037

Average precision of initial (pilot search) run / feedback run / Relevance clarity of feedback language models of NTCIR-5 CLIR J-J



Average precision of initial (pilot search) run / feedback run / Relevance clarity of feedback language models of NTCIR-4 CLIR J-J



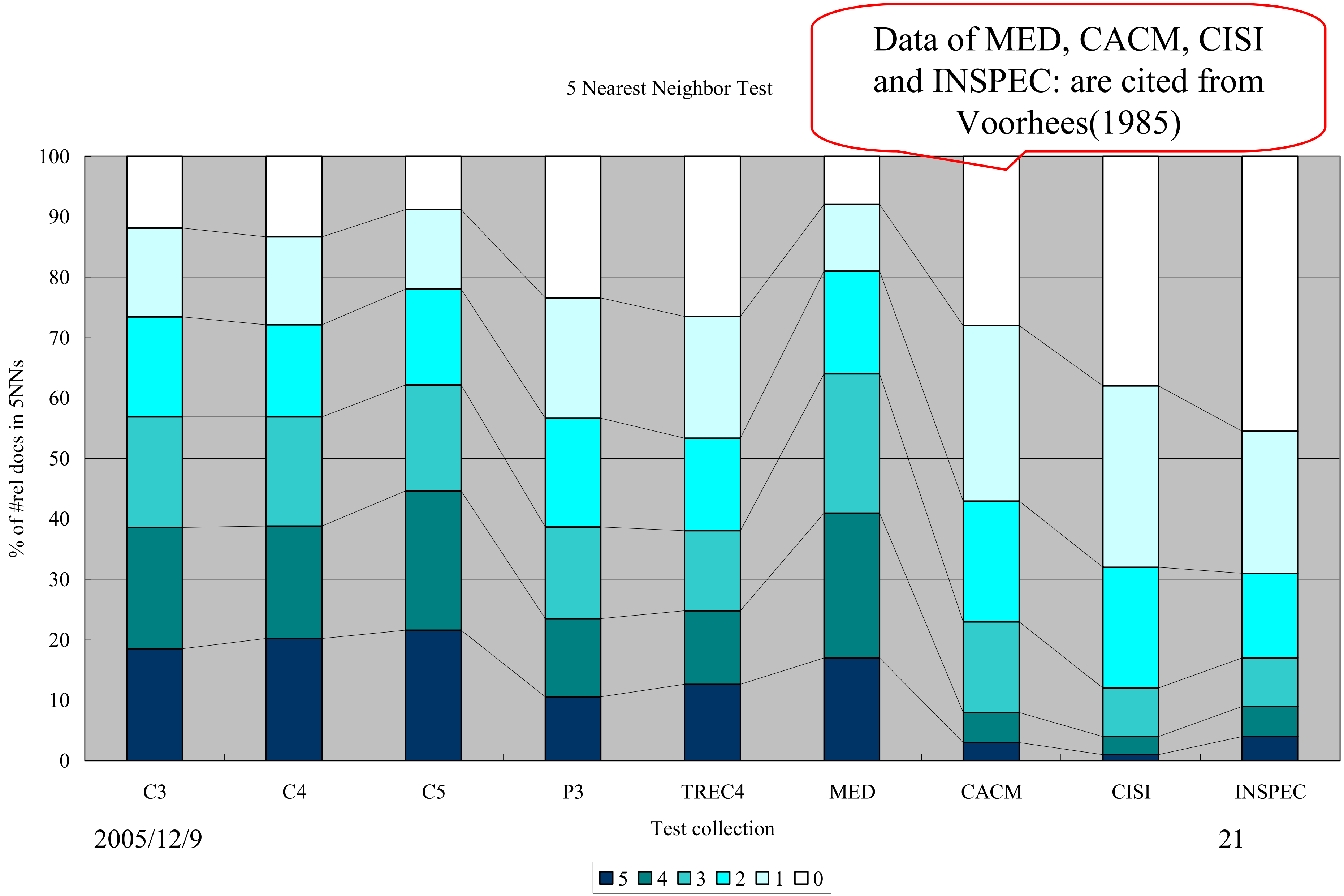
Correlation on a run by run basis

- Strong correlation between feedback gain and average query clarity: 0.824
  - NTCIR-3 CLIR J-J: 19.4% 4.614
  - NTCIR-4 CLIR J-J: 23.0% 6.543
  - NTCIR-5 CLIR J-J: 41.7% 6.985
  - NTCIR-3 Patent : 15.4% 4.731
- Strong correlation between feedback gain and average relevance clarity: 0.807
  - NTCIR-3 CLIR J-J: 19.4% 3.342
  - NTCIR-4 CLIR J-J: 23.0% 5.038
  - NTCIR-5 CLIR J-J: 41.7% 5.563
  - NTCIR-3 Patent : 15.4% 3.094
  - NTCIR-3 CLIR J-J Relax: 17.5% 2.964
  - NTCIR-4 CLIR J-J Relax: 19.0% 4.957
  - NTCIR-5 CLIR J-J Relax: 32.9% 5.302
  - NTCIR-3 Patent Relax: 14.2% 2.937

5 Nearest Neighbor Test

- Voorhees(1985)
  - The n nearest neighbors of a document d are the n documents that are the most similar to d.
  - %of 0,1,2,3,4 or 5 relevant documents in the 5 nearest neighbors of each relevant document
  - It might indicate how similar relevant documents are to each other.
  - An alternative way to test the cluster hypothesis
- A 5 nearest neighbor test may indicate how relevant documents are similar to each other.
  - This measure may indicate topic cohesion as well.

5-Nearest Neighbor test of NTCIR-3 to 5 CLIR-J-J, NTCRIR-3 Patent, TREC-4 adhoc and other test collections



Conclusions

- An automatic feedback strategy from top k documents is exceptionally effective in the NTCIR-5 CLIR J-J test collection (as much as 41.7% gain of MAP).
- Conditions where such an automatic feedback strategy is effective are hypothesized.
- In order to test the topic cohesion, feedback document clarity and relevance clarity tests are carried out.
  - Strong correlation between feedback gain and clarity scores on a run-by-run basis
- Relax relevant documents are topically too diverse to achieve improvements by an automatic feedback.
- 5 nearest neighbor test was carried out.
- Results were consistent with NTCIR CLIR J-J collections but with the TREC-4 collection.