Estimating Pool-depth on Per Query Basis

Sukomal Pal
CVPR Unit
Indian Statistical Institute
203 B T Road
Kolkata 700108, India
sukomal_r@isical.ac.in

Mandar Mitra
CVPR Unit
Indian Statistical Institute
203 B T Road
Kolkata 700108, India
mandar@isical.ac.in

Samaresh Maiti
CVPR Unit
Indian Statistical Institute
203 B T Road
Kolkata 700108, India
samaresh_t@isical.ac.in

ABSTRACT

This paper demonstrates a simple and pragmatic approach for the creation of smaller pools for evaluation of ad hoc retrieval systems. Instead of using an apriori-fixed depth, variable pool-depth based pooling is adopted. The pool for each topic is incrementally built and judged interactively. When no new relevant document is found for a reasonably long run of pool-depths, pooling can be stopped for the topic. Based on available effort and required performance level, the proposed approach can be adjusted for optimality. Experiments on TREC-7, TREC-8 and NTCIR-5 data show its efficacy in substantially reducing poolsize without seriously compromising reliability of evaluation.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Systems and Software – performance evaluation

General Terms

Experimentation, Measurement

Keywords

information retrieval, test collection, evaluation, pooling

1. INTRODUCTION

Evaluation of Information Retrieval (IR) systems at standard fora like TREC, CLEF, NTCIR, INEX or FIRE is based on the Cranfield paradigm where a test collection is built with three major components: 1) a set of documents (*corpus*), 2) a set of information need (*topics*) and 3) a set of relevance judgments for each topic (*qrels*). Ideally these qrels should be complete, i.e., each document in the corpus should be judged as relevant or non-relevant with respect to each topic in the topic set. For a large corpus it is infeasible to construct such test collections because of the prohibitive amount of time and effort involved therein. A more practical and efficient approach called *pooling* is used where a subset of documents is chosen for relevance judgment with respect to each topic. The documents within the chosen subset that are judged relevant are assumed to be the only relevant documents for the query, and all unjudged documents are considered non-relevant.

Hence selection of the subset is very crucial as one needs to maximise the number of relevant documents within the judged set. The fact that IR systems rank their retrieved set of documents in decreasing order of *similarity* or expected probability of relevance to the topic, however, facilitates our job. Top ranking documents for any system are more likely to be relevant than the ones at the bottom of the ranked list. Hence, for each query, the union of the set of top-k documents from n independent submitted runs are used to create a pool which is exhaustively judged. If k and n are reasonably large, the set of documents judged relevant may be assumed to be representative of the ideal set and suitable for evaluating the effectiveness of retrieval runs. However, for large k and n, upto kn documents may need to be examined and judged for each query. For example, the TREC-8 ad hoc track [8] used $k = 100$ and $n = 129$ to create a qrels of 86,830 judgements for 50 topics. Though this number was much smaller compared to the entire collection of about 500,000 documents for each of the 50 topics, the effort involved in relevance judgement was considerable. For much higher n (> 100) and an equally high number of topics (~ 100), both the pool-size and associated human effort required for its judgment get enormous. The cost involved in the construction of such a pool becomes a serious constraint.

The issue has attracted the community's concern since the late-nineties and several efforts have been made to reduce this cost ([4], [2], [1], [3], [5]). While Cormack et al. [4] proposed pool reduction by preferentially choosing documents from a few runs, proposals from Aslam et al [1], and Carterette [3] were based on statistical sampling. Guiver et al. [5] again proposed to choose a few good topics for evaluation. While the approaches reduce the cost, sometimes they compromise on the reliability and reusuability of the test collection or introduce bias in the pool [9].

Apart from cost, another issue is that of *recall* estimation. More than a decade ago, Zobel [10] showed that TREC pools are able to identify at most 50%-70% relevant documents at depth 100. Though the pools are reliable enough to judge a "new" system, they can not be used for measuring systems designed to maximise recall. It also pointed out that "In particular, if it has become likely that for a certain query no more relevant documents will be identified, then continuing to judge documents for that query is a waste of resources."

Based on Zobel's findings, we revisit the pooling process with TREC-7, TREC-8 and NTCIR-5 data. We observe that for each query, the rate of finding new relevant documents with respect to k is different. Though the rate depends on the number of systems n in the pool; even for a fixed n, it is largely query-specific. However, the rate is generally high at early ranks and it decreases with increase in k. Barring a few cases, the rate drops to zero for most of the queries much before pool-depth $k = 100$. In other words, the pool for a topic gets *saturated* after a certain pool-depth. Even if k is increased thereafter, no more new relevant document is found. This depth, termed as *critical pool-depth* (k_{cr}), is a feature of the topic concerned and significantly varies from one topic to another. The following are some representative examples:

Table 1: Pool saturation at k_{cr}

ad hoc track	topic-id	k_{cr}	*nrels*	pool-size at	
				k_{cr}	$k = 100$
TREC-7	363	20	16	348	1597
	384	76	51	926	1225
TREC-8	403	14	21	148	1382
	410	47	65	943	2183
NTCIR-5	31	25	32	538	1723
	4	20	10	451	1788

Though the number of relevant documents (*nrels*) at k_{cr} remains constant even upto $k = 100$, the pool-size increases monotonically with k. If a fixed k is used across all topics, the assessors are thus burdened with a lot of futile judgments even when there is no hope of finding relevant documents (*reldoc*) any more.

The above table 1 shows that in top-k pooling, it is prudent to use a variable k based on k_{cr} of each topic in the topic set, since assessment effort is greatly reduced if pooling can be stopped at k_{cr}. The only problem is how to estimate k_{cr}. In this paper, a simple approach to estimate k_{cr} is proposed along with some preliminary results.

2. APPROACH

Our approach is inspired from that of Zobel [10]. Zobel started with complete judgment for all topics to some initial depth and then used extrapolation to find the likely number reldocs for each query. He suggested then either the most promising topics should be further judged or the least promising topics should be removed from the pool.

We however consider each of the topics and build the pool based on its k_{cr}. The pool is incrementally built from the runs starting from $k = 1$ to 100, and nrels and pool-size are noted at each k. As k increases, the count of 'new' reldocs found at each k generally decreases. However, the rate of decrement is not uniform. It contains a few irregular bursts in-between. To estimate k_{cr}, thus, one needs to smooth-out the bursts. We find that the irregularity occurs at two levels. First the raw number of reldocs are not uniform. Even when they are smoothed, the rate of finding new reldocs contain a few bursts which need some level of smoothing.Hence 2-stage smoothing is applied. First, the raw nrels is smoothed by a moving average window w (in the smoothed version, raw nrels at any pool-depth i is replaced by the arithmetic mean of w consecutive nrels, starting at $k = i$ to $i + w - 1$). Then the rate of finding new reldocs (count for 'new' reldocs at depth i = [nrels at $k = i + 1$] - [nrels at $k = i$]) is also smoothed with a moving average window W. When this smoothed rate of finding new reldocs remains below a certain threshold t continuously for at least a pre-set number of pool-depths (l), the corresponding depth is estimated as k_{cr} for the topic. A reduced pool is obtained based on k_{cr} of each topic in the topicset. Set-intersection of the reduced pool with original qrels provides a reduced qrels which is a subset of the original qrels. Each reduced qrels corresponds to its parameter setting (w, W, t, l). We consider $w = 6, 8, 10, 12, 14$; $W = 2, 3, 4, 5, 6$; $t = 0.05, 0.10, 0.20, 0.40, 0.80$; $l = 3, 4, 5, 6$. Hence we consider $(5 \times 5 \times 5 \times 4) = 500$ qrels for a collection. We evaluate runs using the qrels and compare their MAP scores with those obtained using original qrels.

3. DATA

The algorithm is tested using the ad hoc test collection of TREC-7 (topics 351-400, 103 runs), TREC-8 (topics 401-450 , 129 runs) and ad hoc cross-lingual test collection of NTCIR-5 (topics 1-50, 67 X-E runs, X stands for Japanese, Chinese, Korean or English and E for English). The original qrels taken from the evaluation fora are used as the baseline judgments.

4. RESULT

The MAP obtained with the most aggressive stopping criteria (small value of k_{cr} ensured by small values of w, W and l and high value of t) are plotted against their original counterparts.

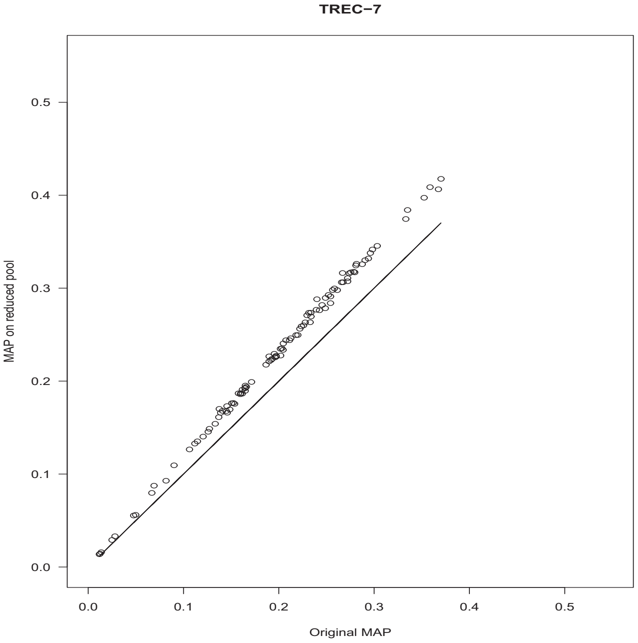

Figure 1: Similarity between MAP-values for TREC-7 dataset ($w = 6$, $W = 2$, $t = 0.80$, $l = 3$)

As evident in Figure 1 - 3 MAP values are in close agreement, although a-bit overestimated in general compared to

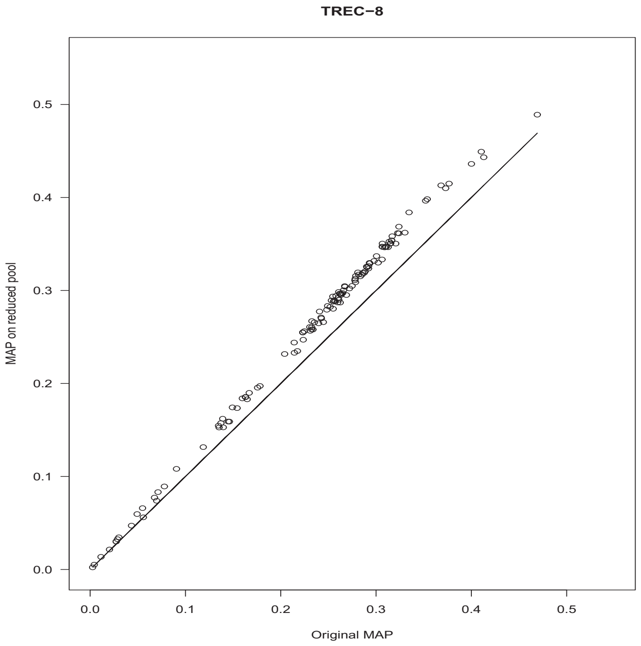

Figure 2: Similarity between MAP-values for TREC-8 dataset ($w = 6$, $W = 2$, $t = 0.80$, $l = 3$)

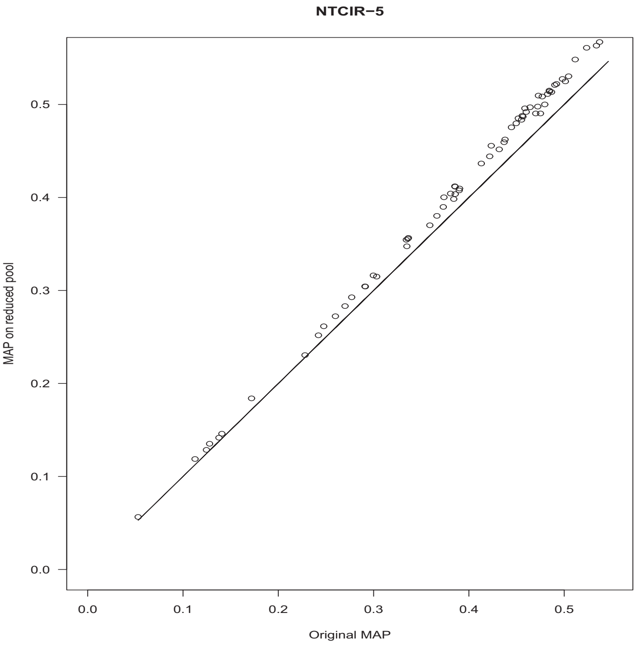

Figure 3: Similarity between MAP-values for NTCIR-5 dataset ($w = 6$, $W = 4$, $t = 0.80$, $l = 3$)

their original counterparts. Aggressive estimation of k_{cr} causes slight underestimation of *recall*-base or total number of relevant documents for each topic which eventually leads overestimation in MAP-score in the reduced settings. However this over-estimation is not alarmingly high (small RMS error in MAP values) even with the most aggressive stopping criteria of pooling (2). Moreover, during evaluation one is more concerned in relative rankings of the systems, which again in fact is of negligible variation (high Kendall's τ as shown in 2).

Table 2: Guaranteed Performance in reduced pool

track	Kendall's τ			RMS error(ϵ)		
	τ_{min}	E	R	ϵ_{max}	E	R
TREC-7	0.979	0.381	0.847	0.033	0.379	0.846
TREC-8	0.967	0.368	0.821	0.030	0.369	0.821
NTCIR-5	0.970	0.341	0.850	0.026	0.331	0.846

E denotes the fraction of assessment effort with respect to the original pool-size and R, denotes the ratio of nrels in reduced qrels to that in original qrels in the table above.

Note that with less than 40% of the original effort, one can identify more than 80% reldocs, with guaranteed reliable system rankings (Kendall's $\tau > 0.96$ when compared to the baseline) and less than 3.3% RMS error in MAP among n systems. Moreover, the above table exhibits the worst-case scenario since it shows the minimum τ and maximum ϵ (or ϵ_{max}) among 500 cases considered in each task. Needless to say, τ in general is much above this minimum (τ_{min}) (Ranges are, TREC-7: [0.979, 0.996], TREC-8: [0.967, 0.999], NTCIR-5: [0.970, 0.999]) while RMS error is much smaller than ϵ_{max} (ranges are TREC- 7: [0.006, 0.033], TREC-8: [0.0009, 0.030], NTCIR-5: [0.002,0.026]).

In general, RMS error is found to be inversely proportional to the assessment effort and Kendall's τ is proportional to the assessment effort as depicted in the curves (Fig 4 - Fig 9). Correlation-values are obtained based on individual system MAPs among 500 cases we considered.

For all collections, assessment effort monotonically increases with each of w, W and l, when the other 3 parameters are constant. When either w, W or l is increased, we opt for a higher level of smoothing, leading to higher estimates of k_{cr}. Thus, the reduced pool gets closer to the original pool resulting in higher assessment effort, higher values of Kendall's τ and lower RMS errors. However, the relation between the threshold t (acceptable change in the rate of finding 'new' reldocs) and assessment effort is inversely proportional. As t is increased, so is tolerance to the rate of change in nrels, leading to early assumption of pool saturation; hence the overall effort decreases.

5. CONCLUSION AND FUTURE WORK

Unlike other low-cost evaluation proposals, our method is not statistical sampling based, nor does it look for a few good topics. Within the traditional framework of the Cranfield paradigm, it offers an interactive pooling approach based on variable pool-depth per query. The approach reduces assessment effort to a great extent for most of the queries where the pool saturates quickly. Again, for the queries where the rate of finding new reldocs is quite high, better estimates of recall can be ensured by going deeper in the pool ($k > 100$). That k is determined dynamically per query requiring more assessor responsibility can be a potential criticism. But tuning 4 parameters w, W, t and l suitably, an optimal tradeoff can be achieved between the cost of evaluation and its

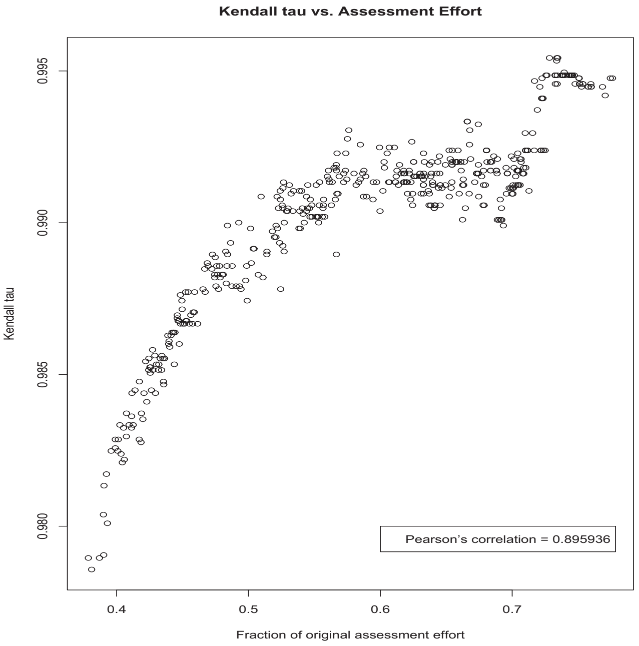

Figure 4: TREC 7 - Relation between Kendall's τ and assessment effort

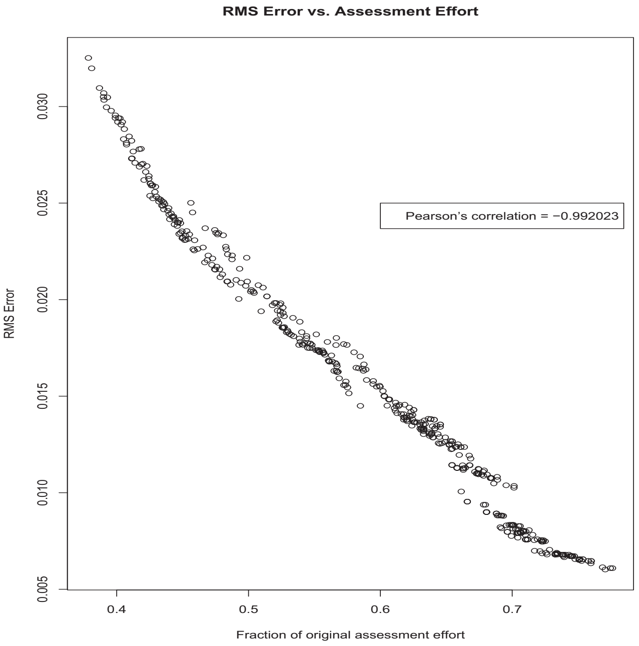

Figure 5: TREC 7 - Relation between RMS error in MAP and assessment effort

reliability. Even with the most aggressive stopping criteria (worst-case scenario with small w, W, l and high t), performance shown is reasonably good. To build a large test collection based on the Cranfield methodology, our simple approach can be cost-effective yet reliable. Our results conform to the findings of the NTCIR CLIA and IR4QA tasks ([7], [6]) that popular documents (documents retrieved by many systems at high ranks) are more likely to be relevant. However, compared to other low-cost evaluation methodologies, specifically that of Aslam et al [1] and Carterette [3] we have not checked how much reusable our method is, i.e. how accurately the collection built based on our proposal can evaluate a 'new system'. The study will certainly be one direction of our future work.

6. REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 541–548, New York, NY, USA, 2006. ACM.

[2] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, New York, NY, USA, 2004. ACM.

[3] B. Carterette. Robust test collections for retrieval evaluation. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 55–62, New York, NY, USA, 2007. ACM.

[4] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 282–289, New York, NY, USA, 1998. ACM.

[5] J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Trans. Inf. Syst.*, 27:21:1–21:26, November 2009.

[6] T. Sakai and N. Kando. Are popular documents more likely to be relevant? a dive into the aclia ir4qa pools. In *EVIA '08: Proceedings of the Second International Workshop on Evaluating Information Access*, pages 8–9, 2008.

[7] T. Sakai, N. Kando, C.-J. Lin, T. Mitamura, H. Shima, D. Ji, K.-H. Chen, and E. Nyberg. Overview of the ntcir-7 aclia ir4qa task. In *NTCIR '08: Proceedings of the 7th NTCIR Workshop Meeting*, pages 77–114, 2008.

[8] E. Voorhees and D. Harman. Overview of the seventh text retrieval conference. In *Seventh Text REtrieval Conference (TREC-7)*, pages 1–24. National Institute of Standards and Technology, Gaithersburg, Maryland, USA., 1999.

[9] E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF '01: Revised Papers*, pages 355–370, London, UK, 2002. Springer-Verlag.

[10] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314, New York, NY, USA, 1998. ACM.

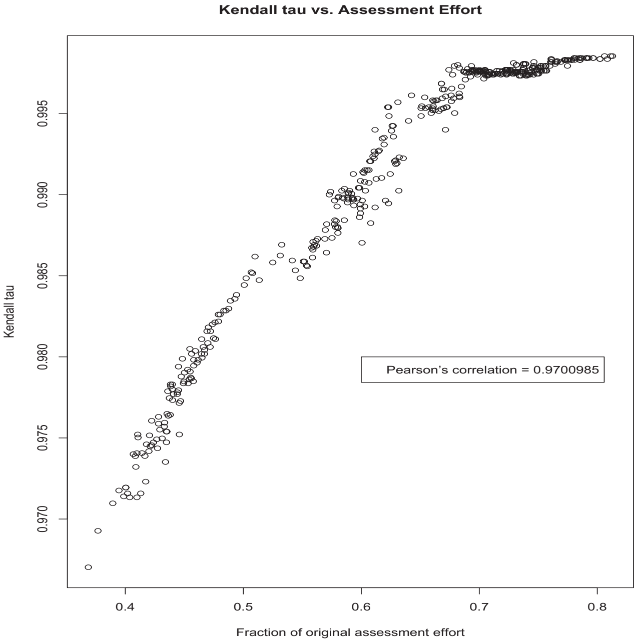

Figure 6: TREC 8 - Relation between Kendall's τ and assessment effort

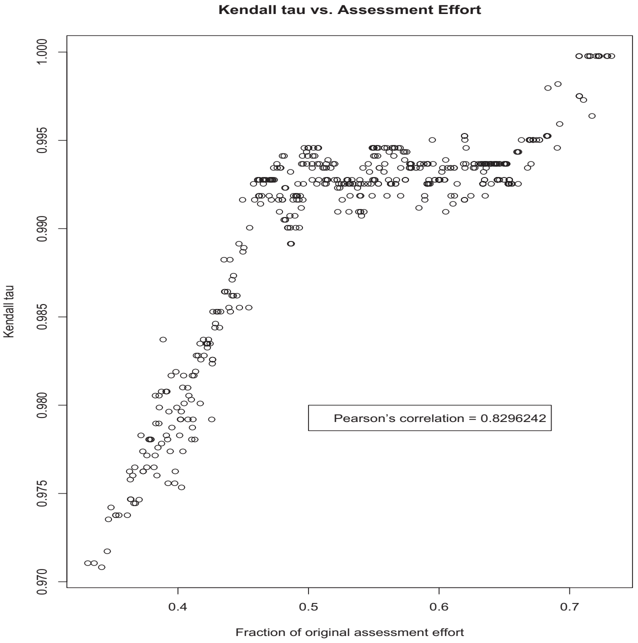

Figure 8: NTCIR 5 - Relation between Kendall's τ and assessment effort

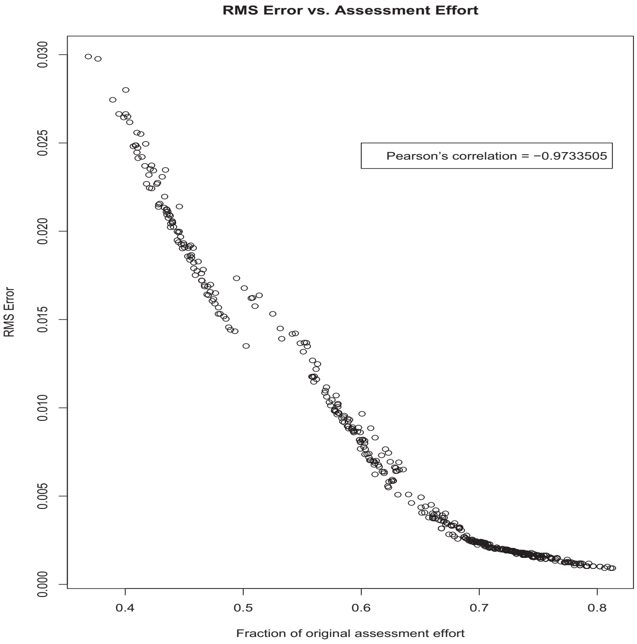

Figure 7: TREC 8 - Relation between RMS error in MAP and assessment effort

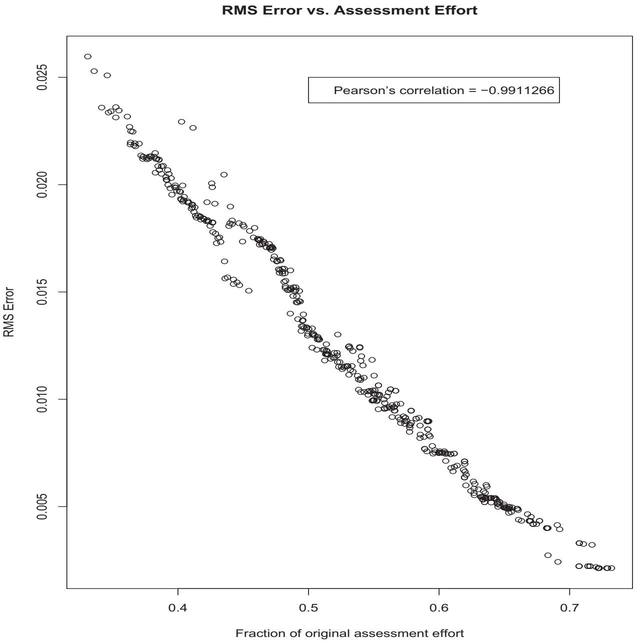

Figure 9: NTCIR 5 - Relation between RMS error in MAP and assessment effort