The Impact of Fixed-Cost Pooling Strategies on Test Collection Bias

Aldo Lipani[1] Guido Zuccon[2] Mihai Lupu[1] Bevan Koopman[3] Allan Hanbury[1]

[1]Inst. of Software Technology & Interactive Systems, Vienna University of Technology, Vienna, Austria
{lipani, lupu, hanbury}@ifs.tuwien.ac.at

[2]Faculty of Science & Technology, Queensland University of Technology, Brisbane, Australia
g.zuccon@qut.edu.au

[3]Australian e-Health Research Centre, CSIRO, Brisbane, Australia
bevan.koopman@csiro.au

ABSTRACT

In Information Retrieval, test collections are usually built using the pooling method. Many pooling strategies have been developed for the pooling method. Herein, we address the question of identifying the best pooling strategy when evaluating systems using precision-oriented measures in presence of budget constraints on the number of documents to be evaluated. As a quality measurement we use the bias introduced by the pooling strategy, measured both in terms of Mean Absolute Error of the scores and in terms of ranking errors. Based on experiments on 15 test collections, we conclude that, for precision-oriented measures, the best strategies are based on Rank-Biased Precision (RBP). These results can inform collection builders because they suggest that, under fixed assessment budget constraints, RBP-based sampling produces less biased pools than other alternatives.

Keywords

Pooling Method, Pooling Strategies, Pool Bias

1. INTRODUCTION

Traditional evaluation of information retrieval systems relies on the idea of a controlled test collection, comprising of documents, information need statements synthesized into queries, and relevance assessments that capture the relevance relations between documents and information needs. Due to the large scale of modern test collections, it is infeasible to assess every document in the collection for relevance to a query. Instead, documents for which relevance assessments are required are selected from document rankings generated by systems for the queries in the test collection [6]. This method of selection of documents for relevance assessment is called *pooling*. The assumption behind pooling is that if a large enough variety of (good) systems contribute results for pooling and pools are sampled at large enough depths, then the pool should contain the almost totality of relevant documents and thus provide reliable evaluations.

ICTIR '16, September 12 - 16, 2016, Newark, DE, USA

ISBN 978-1-4503-4497-5/16/09...$15.00

DOI: http://dx.doi.org/10.1145/2970398.2970429

Different pooling strategies exist that influence the number of documents that are required to be assessed, the type of measures that can be used to evaluate the systems, the bias of the collection towards specific systems or retrieval techniques and features and thus the reusability of the test collection to evaluate systems that did not contribute to the pool. A standard pooling strategy, called fixed-depth pooling, consists of using all documents retrieved by systems up to a cut-off depth k. Alternative pooling strategies extend the fixed-depth strategy by sampling documents from this pool to create smaller pools, or by combining sampled portions of document rankings (strata) to form stratified pools. Finally, more sophisticated pooling strategies have also been proposed. For example, in this paper we consider the strategies proposed by Moffat et al. [5], aimed at reducing the volume of required relevance assessments and based on RBP.

Pool size affects assessment budget, with larger pools involving larger costs. Under budget constraints, it is vital to limit the size of pools, without compromising the quality of the evaluation and the reusability of the collection. In fact, small pools may be affected by large bias and may render the collection unsuited to evaluate systems that did not participate in the pool. Previous work has investigated the influence of pooling bias on inferred measures and precision at fixed cut-off ($P@n$) when pooling with fixed-depth, uniform sampling, and stratified pooling strategies [7, 4]. Other work has examined alternative pooling methods and how to correct bias towards unpooled systems [1, 2, 3, 5, 8].

In this paper, we focus on controlling the balance between the cost of relevance assessments required by specific pooling strategies (and thus the pool size) and the bias introduced by different pooling strategies (and thus the reliability and reusability of the test collection). Specifically, we consider the setting where a fixed budget for relevance assessments has been given. We then empirically investigate which strategy among (a) a standard constrained pooling strategy and (b) Moffat et al.'s pooling strategies [5], would manifest a lower bias.

The results of our empirical investigation demonstrate that the least bias is obtained when we use information generated in the assessment process itself. This result is of interest for researchers building new test collections because it suggests that, under fixed budget constraint for relevance assessment, they should prefer RBP-based sampling over the alternatives investigated in this work because it returns pools with less bias towards specific systems.

2. BACKGROUND

In this section we introduce the pooling strategies analyzed in this paper. The first two are variations of the most common fixed-depth at k ($Depth@k$), already described in the introduction. The remaining three pooling strategies were developed by Moffat et al. [5]. First, let us fix the notations and the shared elements required among all of the following strategies, R_p is the set of runs to be pooled and N is the maximum number of documents to be judged.

We start with the variants of $Depth@k$. Despite its common use in practice, $Depth@k$ does not allow to impose a fixed number of documents to be judged. For this reason, we explore two possible variants, $Take@N$ and $Take+@K\&N$ that allows to fix such a constraint:

Take@N (strategy T): This strategy is the most commonly adopted in practice if a fixed pool size is required (e.g., due to budget constraints) and is based on the rank at which documents have been retrieved. It starts assigning to every retrieved document d the highest rank ρ to which d has been retrieved by R_p, then it continues taking the first N documents with the highest ρ and pools them. The main drawback of this strategy with respect to $Depth@k$ is that it does not guarantee fairness among all the pooled runs. In $Depth@k$ all the runs contribute equally to the pool with their first k documents, here some runs can get more judged documents than others.

Take+@K&N (strategy $T+$): This strategy aims to address the lack of fairness of the previous strategy by introducing a non-deterministic selection of the documents to be judged. It may be thought of as an application of the $Stratified$ pooling strategy [9]. The $Stratified$ strategy defines multiple strata, each characterized by a depth and a sample rate. Each document is assigned to a stratum based on the rank at which it has been retrieved first, then sampled based on the sample rate of the stratum. By definition, this strategy guarantees fairness because it forces a constant number of pooled documents per run. The $Take+@K\&N$ strategy defines a stratification of two strata, where K is its maximum depth. For the first stratum we fix a sample rate $r_1 = 1.0$ and depth k_1 as deep as the number of documents to be judged does not overcome the imposed limit of N judgments. If we call N^{k_1} and N^K the estimated pooled documents for a $Depth@k$ strategy at depth k_1 and K, the second stratum is characterized by $k_2 = K - k_1$ and sample rate $r_2 = (N - N^{k_1})/(N^K - N^{k_1})$, which makes the stratification able to reach, in expectation, N, the number of documents to be judged.

In the above strategies, documents that have been retrieved on top by at least one run are more likely to be pooled, but no distinction is made if a document has been retrieved by multiple runs or just by one. If we were to apply this last distinction as a mere pooling strategy we would count how many runs have retrieved a certain document, and pool the most retrieved ones. Such a strategy would have the effect of over-emphasizing the importance of highly retrieved documents with respect to the uniquely retrieved ones, thereby rewarding conformity and creating collections with the strongest pool bias possible due to the systematic avoidance of documents that have been uniquely retrieved by a single run. But if this distinction is combined with information about the rank at which the documents were retrieved, as pointed out by Moffat et al. [5], it may lead to pooling strategies that are more effective in selecting documents to pool, because it would select documents that provide more benefit to the final result.

Before describing the three variants of this strategy, let us review some terminology. Moffat et al. [5] defined *base* RBP as the RBP calculated on the assumption that unjudged documents are non-relevant. It is a lower threshold to the real RBP value which would have been obtained had all documents been judged. For instance, if RBP is defined as:

$$RBP = (1-p)\sum_{i=1} u_i \cdot p^{i-1} \tag{1}$$

where $u_i \in [0,1] \cup \{?\}$ denotes the relevance judgment of the document at position i (taking values between 0 and 1 if judged and ? if unjudged) and $p \in]0,1[$ is a constant, then the *base* RBP is:

$$b_r = (1-p)\sum_{i:u_i \neq ?} u_i \cdot p^{i-1} \tag{2}$$

As a complement, the paper also defined *residual* RBP as:

$$e_r = (1-p)\sum_{i:u_i = ?} p^{i-1} \tag{3}$$

For a run r, $b_r + e_r$ is therefore the maximum RBP value that could be obtained by that run (i.e., if all unjudged documents are actually relevant). Given these definitions, the following pooling strategies can be formalized:

RBPBasedA@N&p (strategy A): To every retrieved d in every pooled run $r \in R_p$ is associated a score equal to the RBP residual, in Eq. (4), which is a function of the rank ρ to which d has been retrieved in r and a parameter p, fixed in advance:

$$c_{r,d} = (1-p)p^{\rho(r,d)-1} \tag{4}$$

Then, a weight is assigned to every retrieved d, which is calculated by summing up all the scores obtained in the pooled runs R_p, as follows:

$$w_d = \sum_{r \in R_p} c_{r,d} \tag{5}$$

Finally, the first N documents with highest weight are included in the pool. This strategy rewards documents that have been retrieved by multiple runs at high ranks, but it has the drawback of leaving runs' residuals free (unconstrained), which may be undesirable.

RBPBasedB@N&p (strategy B): Like strategy A, but the weighting function is calculated by multiplying the RBP gain-based score with the current residual of r from which d comes:

$$w_d = \sum_{r \in R_p} c_{r,d} \cdot e_r \tag{6}$$

This pooling strategy, with respect to the others, adapts the weights w_d at every pooled document, due to the inclusion in the weighing schema of the residuals, which are in function of the current set of judged documents. This strategy is characterized by N sequential re-weighting stages in which at each stage the document with the highest w_d is pooled. It is worth mentioning that the re-weighting is independent of the actual judgment of the document.

RBPBasedC@N&p (strategy C): Like strategy B, but the weighting function is calculated taking into account also the current RBP score of the run r:

$$w_d = \sum_{r \in R_p} c_{r,d} \cdot e_r \cdot (b_r + e_r/2)^3 \quad (7)$$

This strategy rewards the runs that have retrieved more relevant documents.

Note that pooling strategies B and C are adaptive, in that they select the next document to pool according to the relevance assessments done up to that point.

3. EXPERIMENTS & RESULTS

To determine which of the examined pooling strategies produces a test collection with lower bias we use a set of 15 test collections selected from TREC: 7 test collections from the Ad-Hoc track, 3 from the Web track, and 5 from more domain specific IR tracks: Genomics, Robust, Legal, Medical and Microblog. Each collection was created using a *Depth@k* pooling strategy; this allows us to create synthetic pools with the pooling strategies examined here. In particular this is done for the only pooled runs $R_p \subseteq R$ of the test collections, for which the pooling strategy guarantees the complete judgment of the first k documents. To measure the bias that a new run would have observed if it had not been part of the construction of the pool, we simulate its absence performing a leave-one-organization-out approach for which at each iteration a new synthetic pool is generated excluding all the runs submitted by the organization for which the selected run belongs to. The bias measures used, as in a previous study [3, 4], are: 1) Mean Absolute Error (MAE), 2) System Rank Error (SRE), and 3) System Rank Error with Statistical Significance (SRE*). MAE is the mean, over all the runs, of the absolute difference between the score of a run when it is and is not part of the pool. SRE is the sum, over all the runs, of the difference in ranks between the run when it is and is not part of the pool. SRE* is like SRE, but the difference is counted only if the runs in between are statistically different (Tukey's test, $p < 0.05$).

Herein, we experiment with a real case study where we have a budget that allows the collection builder to judge a maximum of 10,000 documents (N). We also assume that the measures targeted by the evaluation task are P@10 and RBP with $p = 0.80$ (this is a common setting for p). We test all the pooling strategies with fixed $N = 10,000$. For strategy $T+$, we fixed K to 20. For strategies A, B and C, two instantiations of p have been tested, $p = 0.80$ (in line with the setting for the RBP evaluation measure), indicated by A^{80}, B^{80} and C^{80} in Tab. 1 and $p = 0.73$, indicated by A^{73}, B^{73} and C^{73} in Tab. 1 (this value was found to correlate best with observed user behavior, see Zhang et al. [10]).

4. DISCUSSION & CONCLUSION

In Tab. 1 we observe that the best performing pooling strategy is $RBPBasedC@N\&p$ with $p = 0.80$ (C^{80}), which sometimes matches the performance of the same strategy but with $p = 0.73$ (C^{73}). This is an expected result since this strategy uses the most information. However, although this is the best performing strategy, it presents several limitations when used for building test collections in practice:

1. If the test collection builders require that relevance labels for a document are aggregated across judgements from multiple assessors, then the use of this pooling strategy puts an additional burden on the collection builders. This is because the strategy requires information about the relevance of documents already assessed to decide which documents to pool next (i.e., the strategy is adaptive). This in turns requires that the assessment process is coordinated such that the selection and assessment of the next document to assess cannot start until all assessors have judged the current document: this may happen at different times due to different assessor cognitive abilities, workload, and work scheduling.

2. A per-topic parallelization of the assessment exercise is not possible (i.e., the practice of distributing documents that are retrieved for the same topic across multiple assessors to speed up assessment). However, this limitation could be potentially mitigated, to a certain extent, by parallelizing the process across topics (i.e., exclusively assign each topic to an assessor, but assigning different topics to different assessors).

3. The third limitation is introduced by the impossibility of randomizing the pooled documents in order to mitigate assessment bias coming from the judgment of documents in order of their predicted relevance. This bias is usually overcome by the standard pooling strategies by randomizing the pooled documents before presenting them to the assessors.

These limitations make the $RBPBasedC@N\&p$ strategy of difficult practical application.

The second best strategy is $RBPBasedA@N\&p$ (A^{80}) with $p = 0.80$. This strategy does not present the previous limitations since all the documents to be assessed are pooled with no required human intervention. The peculiarity of this strategy is that it rewards documents that have been retrieved by multiple runs at high ranks, yet it does minimize pool bias.

In summary, this paper examined a number of strategies aimed at selecting documents for relevance assessment under fixed budget constraints while minimizing pool bias. The empirical results demonstrate that variants A and C of Moffat et al.'s strategies [5], a static and an adaptive strategy based on RBP, should be preferred over traditional fixed-depth or stratified pooling when deciding upon the pooling strategy to be used to form a new test collection under fixed assessment budget constraints. However, due to the limitations introduced by the strategy C, we recommend in practice the strategy A.

The software used to create and analyze the pooling strategies examined in this paper are made available on the website of the first author.

Acknowledgments

This research was partly supported by the Austrian Science Fund (FWF) project number P25905-N23 (ADmIRE).

5. REFERENCES

[1] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *Proc. of SIGIR*, 2006.

[2] G. V. Cormack, C. R. Palmer, and C. L. Clarke. Efficient construction of large test collections. In *Proc. of SIGIR*, 1998.

Table 1: Results obtained using only the pooled runs (R_p) of each test collection (C). Pool bias measures, MAE, SRE and SRE* (the lower the values, the better; bold values are the best for the test collection), computed on five pooling strategies (S) via leave-one-organization-out. $|R|$ total number of run of the test collection; $|O|$ number of organizations; k depth of the original pool; $|T|$ number of topics; Q number of judged documents; $|Q^r|$ number of documents judged relevant; $|Q_n^r|$ number of documents judged relevant for the synthetic pool, and; $|Q_n^?|$ number of documents non-judged for the synthetic pool.

C	Stats	S	$\|Q_n^r\|$	$\|Q_n^?\|$	P@10			RBP ($p = 0.80$)		
					MAE	SRE	SRE*	MAE	SRE	SRE*
Ad Hoc 2	$\|R\|$: 38 $\|O\|$: 22 $\|R_p\|$: 30 k : 100 $\|T\|$: 50 $\|Q^r\|$: 11645 $\|Q\|$: 62620	T	3627	0	0.0422	106	9	0.0441	113	5
		T+	3523	0	0.0436	104	10	0.0447	99	7
		A^{80}	3651	0	0.0398	103	**8**	0.0410	105	5
		B^{80}	3626	0	0.0418	106	9	0.0436	113	5
		C^{80}	**3818**	0	**0.0392**	**100**	**8**	**0.0401**	**97**	**4**
		A^{73}	3641	0	0.0407	106	9	0.0424	108	5
		B^{73}	3628	0	0.0422	106	9	0.0441	115	5
		C^{73}	3788	0	0.0398	103	**8**	0.0415	109	**4**
Ad Hoc 3	$\|R\|$: 40 $\|O\|$: 22 $\|R_p\|$: 21 k : 200 $\|T\|$: 50 $\|Q^r\|$: 9805 $\|Q\|$: 97319	T	2946	0	0.0234	20	**0**	0.0258	22	0
		T+	2520	0	0.0301	25	1	0.0331	26	0
		A^{80}	2971	0	0.0226	19	**0**	0.0250	21	0
		B^{80}	2931	0	0.0229	20	**0**	0.0254	21	0
		C^{80}	**3143**	0	**0.0199**	**15**	**0**	**0.0225**	**17**	0
		A^{73}	2963	0	0.0228	19	**0**	0.0251	21	0
		B^{73}	2939	0	0.0232	21	**0**	0.0258	22	0
		C^{73}	3128	0	0.0206	17	**0**	0.0230	**17**	0
Ad Hoc 4	$\|R\|$: 33 $\|O\|$: 19 $\|R_p\|$: 24 k : 100 $\|T\|$: 50 $\|Q^r\|$: 6503 $\|Q\|$: 87069	T	1954	0	0.0361	68	2	0.0373	68	3
		T+	1921	0	0.0355	62	1	0.0371	60	3
		A^{80}	1979	0	0.0342	65	2	0.0354	62	2
		B^{80}	1946	0	0.0357	66	2	0.0369	66	3
		C^{80}	**2224**	0	**0.0287**	**54**	**0**	**0.0293**	**52**	**0**
		A^{73}	1970	0	0.0347	66	2	0.0359	64	3
		B^{73}	1946	0	0.0357	67	2	0.0371	66	3
		C^{73}	2199	0	0.0289	**54**	**0**	0.0298	**52**	**0**
Ad Hoc 5	$\|R\|$: 61 $\|O\|$: 21 $\|R_p\|$: 53 k : 100 $\|T\|$: 50 $\|Q^r\|$: 5524 $\|Q\|$: 133681	T	1482	0	0.0234	154	17	0.0242	167	21
		T+	1495	0	0.0239	161	20	0.0248	175	23
		A^{80}	1532	0	0.0217	148	14	0.0224	154	16
		B^{80}	1468	0	0.0231	156	16	0.0235	157	17
		C^{80}	**1762**	0	**0.0189**	**122**	**3**	**0.0198**	**132**	**6**
		A^{73}	1506	0	0.0220	146	15	0.0227	154	16
		B^{73}	1479	0	0.0236	158	16	0.0241	165	19
		C^{73}	1736	0	0.0192	124	4	0.0200	137	7
Ad Hoc 6	$\|R\|$: 74 $\|O\|$: 29 $\|R_p\|$: 25 k : 100 $\|T\|$: 50 $\|Q^r\|$: 4611 $\|Q\|$: 72270	T	1544	0	0.0304	35	2	0.0312	37	2
		T+	1605	0	0.0289	35	2	0.0290	30	**0**
		A^{80}	1561	0	0.0294	34	**1**	0.0302	35	**0**
		B^{80}	1532	0	0.0301	35	2	0.0309	37	2
		C^{80}	**1762**	0	**0.0254**	**29**	**1**	**0.0258**	**23**	**0**
		A^{73}	1553	0	0.0296	35	2	0.0305	35	**0**
		B^{73}	1538	0	0.0306	35	2	0.0315	38	2
		C^{73}	1739	0	0.0258	**29**	**1**	0.0264	25	**0**
Ad Hoc 7	$\|R\|$: 103 $\|O\|$: 42 $\|R_p\|$: 63 k : 100 $\|T\|$: 50 $\|Q^r\|$: 4674 $\|Q\|$: 80344	T	1697	0	0.0149	132	25	0.0160	114	15
		T+	1694	0	0.0141	129	25	0.0149	104	15
		A^{80}	1728	0	0.0133	105	**12**	0.0142	86	10
		B^{80}	1682	0	0.0151	132	25	0.0162	115	15
		C^{80}	**1921**	0	**0.0116**	**92**	**12**	**0.0123**	**74**	**7**
		A^{73}	1719	0	0.0136	107	13	0.0147	88	10
		B^{73}	1682	0	0.0149	131	25	0.0160	111	15
		C^{73}	1900	0	0.0123	94	**12**	0.0129	81	**7**
Ad Hoc 8	$\|R\|$: 129 $\|O\|$: 41 $\|R_p\|$: 66 k : 100 $\|T\|$: 50 $\|Q^r\|$: 4728 $\|Q\|$: 86830	T	1566	0	0.0199	130	20	0.0201	142	15
		T+	1611	0	0.0191	117	19	0.0190	125	11
		A^{80}	1653	0	0.0168	101	15	0.0168	106	12
		B^{80}	1559	0	0.0193	121	17	0.0194	135	15
		C^{80}	**1874**	0	**0.0148**	80	**8**	**0.0146**	**80**	**6**
		A^{73}	1619	0	0.0178	107	17	0.0178	122	14
		B^{73}	1559	0	0.0200	131	19	0.0202	142	15
		C^{73}	1850	0	0.0152	**79**	**8**	0.0152	84	**6**
Web 9	$\|R\|$: 104 $\|O\|$: 23 $\|R_p\|$: 39 k : 100 $\|T\|$: 50 $\|Q^r\|$: 2617 $\|Q\|$: 70070	T	912	0	0.0136	74	0	0.0145	78	0
		T+	953	0	0.0127	71	0	0.0135	79	0
		A^{80}	942	0	0.0134	75	0	0.0141	80	0
		B^{80}	916	0	0.0137	74	0	0.0146	80	0
		C^{80}	**1062**	0	**0.0105**	**54**	0	**0.0112**	**67**	0
		A^{73}	932	0	0.0132	73	0	0.0139	77	0
		B^{73}	919	0	0.0136	75	0	0.0147	80	0
		C^{73}	1048	0	0.0106	55	0	0.0115	69	0
Web 2001	$\|R\|$: 97 $\|O\|$: 29 $\|R_p\|$: 35 k : 100 $\|T\|$: 50 $\|Q^r\|$: 3363 $\|Q\|$: 70400	T	1272	0	0.0111	43	0	0.0125	67	0
		T+	1221	0	0.0105	43	0	0.0118	63	0
		A^{80}	1286	0	0.0100	40	0	0.0112	55	0
		B^{80}	1260	0	0.0109	42	0	0.0121	62	0
		C^{80}	**1383**	0	**0.0087**	**33**	0	**0.0097**	**49**	0
		A^{73}	1280	0	0.0100	39	0	0.0114	59	0
		B^{73}	1260	0	0.0112	43	0	0.0125	65	0
		C^{73}	1370	0	0.0088	35	0	0.0099	52	0
Web 2002	$\|R\|$: 69 $\|O\|$: 16 $\|R_p\|$: 60 k : 50 $\|T\|$: 50 $\|Q^r\|$: 1574 $\|Q\|$: 56650	T	588	0	0.0185	296	0	0.0195	330	0
		T+	580	0	0.0184	297	0	0.0196	326	0
		A^{80}	626	0	0.0170	264	0	0.0179	297	0
		B^{80}	600	0	0.0184	292	0	0.0194	326	0
		C^{80}	645	0	**0.0159**	**253**	0	**0.0165**	**279**	0
		A^{73}	618	0	0.0179	284	0	0.0188	314	0
		B^{73}	595	0	0.0185	292	0	0.0195	329	0
		C^{73}	**649**	0	0.0161	255	0	0.0169	288	0
Genomics 2005	$\|R\|$: 62 $\|O\|$: 32 $\|R_p\|$: 46 k : 60 $\|T\|$: 49 $\|Q^r\|$: 4584 $\|Q\|$: 39958	T	1905	0	0.0218	265	0	0.0236	326	**2**
		T+	1919	0	0.0193	**230**	0	0.0215	308	3
		A^{80}	1924	0	0.0203	244	0	0.0219	304	**2**
		B^{80}	1894	0	0.0215	260	0	0.0229	314	**2**
		C^{80}	**2046**	0	0.0193	239	0	**0.0206**	**298**	**2**
		A^{73}	1912	0	0.0209	254	0	0.0225	309	**2**
		B^{73}	1898	0	0.0217	264	0	0.0232	321	**2**
		C^{73}	2029	0	**0.0192**	239	0	0.0208	300	**2**
Robust 2005	$\|R\|$: 74 $\|O\|$: 17 $\|R_p\|$: 18 k : 55 $\|T\|$: 50 $\|Q^r\|$: 6561 $\|Q\|$: 37798	T	2624	0	0.0422	35	11	0.0446	35	**10**
		T+	2325	0	0.0488	43	13	0.0511	43	13
		A^{80}	2644	0	0.0408	32	11	0.0430	34	**10**
		B^{80}	2607	0	0.0423	35	11	0.0446	37	**10**
		C^{80}	**3033**	0	**0.0358**	**28**	**10**	**0.0374**	**31**	**10**
		A^{73}	2639	0	0.0418	35	11	0.0440	35	**10**
		B^{73}	2611	0	0.0424	35	11	0.0448	36	**10**
		C^{73}	3002	0	0.0360	29	**10**	0.0376	**31**	**10**
Legal 2006	$\|R\|$: 34 $\|O\|$: 8 $\|R_p\|$: 14 k : 10 $\|T\|$: 39 $\|Q^r\|$: 4323 $\|Q\|$: 31739	T	539	7181	0.0794	53	6	0.0947	59	8
		T+	539	2650	0.1032	63	13	0.1227	77	26
		A^{80}	539	7181	0.0789	51	5	0.0936	59	8
		B^{80}	539	7181	0.0792	52	6	0.0945	59	8
		C^{80}	539	7181	**0.0742**	**47**	**3**	**0.0876**	**55**	**6**
		A^{73}	539	7181	0.0792	51	5	0.0944	59	8
		B^{73}	539	7181	0.0792	52	6	0.0945	59	8
		C^{73}	539	7181	0.0744	**47**	**3**	0.0879	**55**	**6**
Medical 2011	$\|R\|$: 127 $\|O\|$: 29 $\|R_p\|$: 46 k : 10 $\|T\|$: 34 $\|Q^r\|$: 1765 $\|Q\|$: 8865	T	**1443**	4909	0.0193	100	0	0.0258	126	**0**
		T+	**1443**	4019	0.0219	107	0	0.0293	137	2
		A^{80}	**1443**	4909	0.0186	97	0	0.0248	116	**0**
		B^{80}	**1443**	4909	0.0194	101	0	0.0258	126	**0**
		C^{80}	1442	4942	**0.0157**	**86**	0	**0.0202**	**92**	**0**
		A^{73}	**1443**	4909	0.0190	99	0	0.0251	118	**0**
		B^{73}	**1443**	4909	0.0197	102	0	0.0262	129	**0**
		C^{73}	1442	4941	0.0163	90	0	0.0214	97	**0**
Microblog 2011	$\|R\|$: 184 $\|O\|$: 58 $\|R_p\|$: 49 k : 30 $\|T\|$: 49 $\|Q^r\|$: 2965 $\|Q\|$: 60129	T	1760	42	0.0179	100	0	0.0181	109	29
		T+	1707	15	0.0174	91	0	0.0179	100	26
		A^{80}	1790	47	0.0159	89	0	0.0164	96	26
		B^{80}	1767	44	0.0179	102	0	0.0184	109	29
		C^{80}	**1912**	114	**0.0132**	**68**	0	**0.0137**	**75**	**18**
		A^{73}	1788	44	0.0168	94	0	0.0171	106	29
		B^{73}	1756	42	0.0181	102	0	0.0184	109	29
		C^{73}	1887	95	0.0138	71	0	0.0142	77	20

[3] A. Lipani, M. Lupu, and A. Hanbury. Splitting water: Precision and anti-precision to reduce pool bias. In *Proc. of SIGIR*, 2015.

[4] A. Lipani, M. Lupu, and A. Hanbury. The curious incidence of bias corrections in the pool. In *Proc. of ECIR*, 2016.

[5] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proc. of SIGIR*, 2007.

[6] K. Spärck Jones and C. J. van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. *British Library Research and Development Report No. 5266*, 1975.

[7] E. M. Voorhees. The effect of sampling strategy on inferred measures. In *Proc. of SIGIR*, 2014.

[8] W. Webber and L. A. Park. Score adjustment for correction of pooling bias. In *Proc. of SIGIR*, 2009.

[9] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proc. of SIGIR*, 2008.

[10] Y. Zhang, L. A. F. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 2009.