

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

anonymized

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 in the 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 the best strategy is one based on the Rank-Biased Precision (RBP) scoring strategy and using some relevance information. In the absence of such relevance information, the second best strategy is still based on the RBP scoring strategy, using the sum of weights of all documents provided by the participating systems. These results can inform collection creator 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 [4]. This mechanism 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.

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. [3], aimed at reducing the volume of required relevance assessments and based on RBP.

The generation of pools of limited size is vital to remain within an assessment budget (and thus test collection creation costs). However, 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 cut-off depth ($P@n$) when pooling with fixed-depth, uniform sampling, and stratified pooling strategies [5, 2]. Other work has examined how to correct bias towards unpooled systems [6, 1].

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 pool strategy variants [3], 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 creating 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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WOODSTOCK '97 El Paso, Texas USA

© 2016 ACM. ISBN 123-4567-24-567/08/06...\$15.00

DOI: 10.475/123_4

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. [3]. First, let us fix the notations and the shared elements required among all of the following strategies, R the set of runs to be pooled and N 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 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 , then it continues taking the first N documents with the highest ρ and pool 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 addresses 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 [7]. 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. [3], 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. [3] 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:

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

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), then the *base* RBP is

$$b_r = (1 - p) \sum_{i=1} \bar{u}_i \cdot p^{i-1} \quad (2)$$

where

$$\bar{u}_i = \begin{cases} 0, & \text{if } u_i = ?. \\ u_i, & \text{otherwise.} \end{cases}$$

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

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

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 $r \in R$ 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:

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

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:

$$w_d = \sum_{r \in R} c_{r,d} \quad (5)$$

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, 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:

$$w_d = \sum_{r \in R} c_{r,d} \cdot e_r \quad (6)$$

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 [1, 2], 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 creator 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*. For strategy *S*, 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. [8]).

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 creators require that relevance labels for a document are aggregated across judgements from

multiple assessors, then the use of this pooling strategies puts an additional burden on the collection creators. This is because the strategies 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 across multiple assessors documents that retrieved for the same query to speed up assessment). However, this limitation could be potentially mitigated, till 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 have them presented 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 [3], respectively 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 the limitations introduced by the strategy *C*, we recommend in practice the strategy *A*.

Implementations of the pool strategies examined in this paper are made available at <http://anonymized>, along with the empirical results presented here.

5. REFERENCES

- [1] A. Lipani, M. Lupu, and A. Hanbury. Splitting water: Precision and anti-precision to reduce pool bias. In *Proc. of SIGIR*, pages 103–112, 2015.
- [2] A. Lipani, M. Lupu, and A. Hanbury. The curious incidence of bias corrections in the pool. In *Proc. of ECIR*, pages 267–279, 2016.
- [3] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proc. of SIGIR*, pages 375–382, 2007.

Table 1: Results obtained using only the pooled runs (R_p) of each test collection. Pool bias measures, MAE, SRE and SRE* (the lowest the values, the better; bold values are the best for the test collection), computed on five pooling strategies via leave-one-organization-out. $|R|$ total 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, and; $|Q_n^r|$ number of documents judged relevant for the synthetic pool.

| C | Stats | M | $ Q_n^r $ | P@10 | | | RBP ($p = 0.80$) | | |
|-----------------|--------------|----------|-------------|---------------|------------|---------------|--------------------|------------|-----------|
| | | | | MAE | SRE | SRE* | MAE | SRE | SRE* |
| Ad Hoc 2 | $ R : 38$ | T | 3634 | 0.0440 | 116 | 5 | 0.0040 | 108 | 8 |
| | | T+ | 3523 | 0.0447 | 102 | 7 | 0.0043 | 103 | 9 |
| | | A^{80} | 3667 | 0.0409 | 107 | 5 | 0.0038 | 102 | 8 |
| | $ O : 22$ | B^{80} | 3642 | 0.0435 | 116 | 5 | 0.0040 | 105 | 9 |
| | $ R_p : 32$ | C^{80} | 3824 | 0.0397 | 100 | 4 | 0.0037 | 95 | 7 |
| | $k : 100$ | A^{73} | 3651 | 0.0423 | 111 | 5 | 0.0039 | 103 | 8 |
| | $ T : 50$ | B^{73} | 3640 | 0.0439 | 117 | 5 | 0.0040 | 109 | 9 |
| $ Q^r : 11645$ | C^{73} | 3793 | 0.0408 | 108 | 4 | 0.0038 | 102 | 7 | |
| $ Q : 62620$ | | | | | | | | | |
| Ad Hoc 3 | $ R : 40$ | T | 2952 | 0.0255 | 22 | 0 | 0.0022 | 19 | 0 |
| | | T+ | 2520 | 0.0331 | 26 | 0 | 0.0029 | 23 | 1 |
| | | A^{80} | 2978 | 0.0248 | 21 | 0 | 0.0022 | 18 | 0 |
| | $ O : 22$ | B^{80} | 2937 | 0.0253 | 21 | 0 | 0.0022 | 19 | 0 |
| | $ R_p : 23$ | C^{80} | 3157 | 0.0222 | 16 | 0 | 0.0019 | 15 | 0 |
| | $k : 200$ | A^{73} | 2971 | 0.0251 | 21 | 0 | 0.0022 | 19 | 0 |
| | $ T : 50$ | B^{73} | 2944 | 0.0254 | 22 | 0 | 0.0022 | 18 | 0 |
| $ Q^r : 9805$ | C^{73} | 3136 | 0.0227 | 17 | 0 | 0.0020 | 17 | 0 | |
| $ Q : 97319$ | | | | | | | | | |
| Ad Hoc 4 | $ R : 33$ | T | 1954 | 0.0373 | 67 | 3 | 0.0035 | 66 | 2 |
| | | T+ | 1920 | 0.0371 | 60 | 3 | 0.0035 | 59 | 1 |
| | | A^{80} | 1979 | 0.0353 | 62 | 2 | 0.0033 | 61 | 1 |
| | $ O : 19$ | B^{80} | 1946 | 0.0369 | 65 | 3 | 0.0034 | 62 | 1 |
| | $ R_p : 32$ | C^{80} | 2224 | 0.0292 | 53 | 0 | 0.0028 | 53 | 0 |
| | $k : 100$ | A^{73} | 1970 | 0.0359 | 64 | 3 | 0.0033 | 62 | 1 |
| | $ T : 50$ | B^{73} | 1946 | 0.0370 | 66 | 3 | 0.0034 | 63 | 1 |
| $ Q^r : 6503$ | C^{73} | 2199 | 0.0296 | 53 | 0 | 0.0028 | 52 | 0 | |
| $ Q : 87069$ | | | | | | | | | |
| Ad Hoc 5 | $ R : 61$ | T | 1487 | 0.0211 | 147 | 19 | 0.0020 | 130 | 15 |
| | | T+ | 1493 | 0.0248 | 176 | 23 | 0.0021 | 139 | 16 |
| | | A^{80} | 1536 | 0.0194 | 139 | 16 | 0.0018 | 133 | 11 |
| | $ O : 21$ | B^{80} | 1472 | 0.0203 | 138 | 16 | 0.0019 | 129 | 15 |
| | $ R_p : 60$ | C^{80} | 1767 | 0.0182 | 124 | 6 | 0.0017 | 115 | 3 |
| | $k : 100$ | A^{73} | 1515 | 0.0198 | 141 | 16 | 0.0019 | 131 | 13 |
| | $ T : 50$ | B^{73} | 1483 | 0.0210 | 144 | 16 | 0.0020 | 136 | 15 |
| $ Q^r : 5524$ | C^{73} | 1741 | 0.0184 | 122 | 6 | 0.0017 | 113 | 3 | |
| $ Q : 133681$ | | | | | | | | | |
| Ad Hoc 6 | $ R : 74$ | T | 1538 | 0.0314 | 37 | 2 | 0.0029 | 33 | 2 |
| | | T+ | 1595 | 0.0292 | 31 | 0 | 0.0028 | 31 | 1 |
| | | A^{80} | 1557 | 0.0304 | 35 | 0 | 0.0028 | 34 | 1 |
| | $ O : 29$ | B^{80} | 1526 | 0.0311 | 37 | 2 | 0.0029 | 35 | 2 |
| | $ R_p : 28$ | C^{80} | 1759 | 0.0255 | 25 | 0 | 0.0025 | 27 | 1 |
| | $k : 100$ | A^{73} | 1551 | 0.0305 | 35 | 0 | 0.0028 | 35 | 2 |
| | $ T : 50$ | B^{73} | 1533 | 0.0314 | 38 | 2 | 0.0029 | 35 | 2 |
| $ Q^r : 4611$ | C^{73} | 1736 | 0.0261 | 27 | 0 | 0.0025 | 29 | 1 | |
| $ Q : 72270$ | | | | | | | | | |
| Ad Hoc 7 | $ R : 103$ | T | 1697 | 0.0138 | 100 | 15 | 0.0012 | 114 | 22 |
| | | T+ | 1697 | 0.0148 | 103 | 15 | 0.0013 | 116 | 25 |
| | | A^{80} | 1731 | 0.0124 | 79 | 10 | 0.0011 | 86 | 12 |
| | $ O : 42$ | B^{80} | 1681 | 0.0139 | 101 | 15 | 0.0012 | 108 | 25 |
| | $ R_p : 74$ | C^{80} | 1924 | 0.0114 | 70 | 7 | 0.0010 | 89 | 11 |
| | $k : 100$ | A^{73} | 1726 | 0.0126 | 82 | 10 | 0.0011 | 91 | 12 |
| | $ T : 50$ | B^{73} | 1688 | 0.0138 | 99 | 15 | 0.0012 | 114 | 25 |
| $ Q^r : 4674$ | C^{73} | 1899 | 0.0122 | 81 | 7 | 0.0011 | 92 | 11 | |
| $ Q : 80344$ | | | | | | | | | |
| Ad Hoc 8 | $ R : 129$ | T | 1560 | 0.0172 | 114 | 13 | 0.0016 | 109 | 19 |
| | | T+ | 1609 | 0.0189 | 128 | 12 | 0.0017 | 106 | 18 |
| | | A^{80} | 1649 | 0.0154 | 100 | 7 | 0.0014 | 84 | 15 |
| | $ O : 41$ | B^{80} | 1554 | 0.0166 | 110 | 9 | 0.0016 | 87 | 15 |
| | $ R_p : 74$ | C^{80} | 1873 | 0.0140 | 81 | 5 | 0.0014 | 74 | 8 |
| | $k : 100$ | A^{73} | 1612 | 0.0163 | 109 | 11 | 0.0015 | 87 | 15 |
| | $ T : 50$ | B^{73} | 1553 | 0.0170 | 116 | 11 | 0.0016 | 95 | 17 |
| $ Q^r : 4728$ | C^{73} | 1852 | 0.0146 | 84 | 5 | 0.0014 | 72 | 8 | |
| $ Q : 86830$ | | | | | | | | | |
| Web 9 | $ R : 104$ | T | 911 | 0.0146 | 78 | 0 | 0.0012 | 71 | 0 |
| | | T+ | 952 | 0.0132 | 79 | 0 | 0.0012 | 67 | 0 |
| | | A^{80} | 936 | 0.0142 | 80 | 0 | 0.0013 | 70 | 0 |
| | $ O : 23$ | B^{80} | 915 | 0.0146 | 79 | 0 | 0.0013 | 69 | 0 |
| | $ R_p : 62$ | C^{80} | 1054 | 0.0112 | 67 | 0 | 0.0010 | 56 | 0 |
| | $k : 100$ | A^{73} | 929 | 0.0141 | 77 | 0 | 0.0012 | 71 | 0 |
| | $ T : 50$ | B^{73} | 913 | 0.0147 | 80 | 0 | 0.0013 | 71 | 0 |
| $ Q^r : 2617$ | C^{73} | 1041 | 0.0114 | 69 | 0 | 0.0010 | 56 | 0 | |
| $ Q : 70070$ | | | | | | | | | |

| C | Stats | M | $ Q_n^r $ | P@10 | | | RBP ($p = 0.80$) | | |
|----------------|--------------|-------------|-------------|---------------|------------|---------------|--------------------|-----------|-----------|
| | | | | MAE | SRE | SRE* | MAE | SRE | SRE* |
| Web 2001 | $ R : 97$ | T | 1269 | 0.0122 | 64 | 0 | 0.0010 | 43 | 0 |
| | | T+ | 1211 | 0.0117 | 64 | 0 | 0.0010 | 46 | 0 |
| | | A^{80} | 1285 | 0.0110 | 56 | 0 | 0.0009 | 40 | 0 |
| | $ O : 29$ | B^{80} | 1255 | 0.0117 | 61 | 0 | 0.0010 | 41 | 0 |
| | $ R_p : 59$ | C^{80} | 1380 | 0.0092 | 48 | 0 | 0.0008 | 33 | 0 |
| | $k : 100$ | A^{73} | 1279 | 0.0113 | 58 | 0 | 0.0009 | 45 | 0 |
| | $ T : 50$ | B^{73} | 1256 | 0.0119 | 64 | 0 | 0.0010 | 41 | 0 |
| $ Q^r : 3363$ | C^{73} | 1367 | 0.0094 | 50 | 0 | 0.0008 | 35 | 0 | |
| $ Q : 70400$ | | | | | | | | | |
| Web 2002 | $ R : 69$ | T | 589 | 0.0151 | 266 | 0 | 0.0015 | 240 | 0 |
| | | T+ | 584 | 0.0143 | 243 | 0 | 0.0014 | 224 | 0 |
| | | A^{80} | 627 | 0.0157 | 262 | 0 | 0.0015 | 226 | 0 |
| | $ O : 16$ | B^{80} | 602 | 0.0161 | 276 | 0 | 0.0016 | 255 | 0 |
| | $ R_p : 69$ | C^{80} | 651 | 0.0135 | 220 | 0 | 0.0013 | 199 | 0 |
| | $k : 50$ | A^{73} | 618 | 0.0165 | 280 | 0 | 0.0016 | 250 | 0 |
| | $ T : 50$ | B^{73} | 596 | 0.0159 | 282 | 0 | 0.0015 | 249 | 0 |
| $ Q^r : 1574$ | C^{73} | 649 | 0.0137 | 219 | 0 | 0.0013 | 209 | 0 | |
| $ Q : 56650$ | | | | | | | | | |
| Genomics 2005 | $ R : 62$ | T | 1905 | 0.0226 | 322 | 2 | 0.0019 | 243 | 0 |
| | | T+ | 1925 | 0.0213 | 306 | 2 | 0.0018 | 229 | 0 |
| | | A^{80} | 1924 | 0.0207 | 294 | 2 | 0.0018 | 233 | 0 |
| | $ O : 32$ | B^{80} | 1894 | 0.0218 | 307 | 2 | 0.0019 | 233 | 0 |
| | $ R_p : 55$ | C^{80} | 2046 | 0.0196 | 290 | 2 | 0.0017 | 221 | 0 |
| | $k : 60$ | A^{73} | 1912 | 0.0214 | 300 | 2 | 0.0018 | 238 | 0 |
| | $ T : 49$ | B^{73} | 1898 | 0.0222 | 314 | 2 | 0.0019 | 243 | 0 |
| $ Q^r : 4584$ | C^{73} | 2029 | 0.0199 | 292 | 2 | 0.0017 | 219 | 0 | |
| $ Q : 39958$ | | | | | | | | | |
| Robust 2005 | $ R : 74$ | T | 2626 | 0.0447 | 35 | 10 | 0.0041 | 34 | 11 |
| | | T+ | 2325 | 0.0511 | 43 | 13 | 0.0048 | 41 | 13 |
| | | A^{80} | 2645 | 0.0428 | 34 | 10 | 0.0040 | 32 | 11 |
| | $ O : 17$ | B^{80} | 2610 | 0.0444 | 36 | 10 | 0.0041 | 35 | 11 |
| | $ R_p : 18$ | C^{80} | 3030 | 0.0378 | 31 | 10 | 0.0036 | 26 | 10 |
| | $k : 55$ | A^{73} | 2640 | 0.0436 | 34 | 10 | 0.0041 | 33 | 11 |
| | $ T : 50$ | B^{73} | 2615 | 0.0448 | 36 | 10 | 0.0042 | 35 | 11 |
| $ Q^r : 6561$ | C^{73} | 2998 | 0.0379 | 31 | 10 | 0.0036 | 28 | 10 | |
| $ Q : 37798$ | | | | | | | | | |
| Legal 2006 | $ R : 34$ | T | 539 | 0.0803 | 59 | 8 | 0.0071 | 53 | 6 |
| | | T+ | 539 | 0.1040 | 77 | 26 | 0.0092 | 65 | 15 |
| | | A^{80} | 539 | 0.0793 | 59 | 8 | 0.0071 | 52 | 6 |
| | $ O : 8$ | B^{80} | 539 | 0.0801 | 59 | 8 | 0.0071 | 52 | 6 |
| | $ R_p : 14$ | C^{80} | 539 | 0.0742 | 55 | 6 | 0.0067 | 49 | 3 |
| | $k : 10$ | A^{73} | 539 | 0.0799 | 59 | 8 | 0.0071 | 52 | 6 |
| | $ T : 39$ | B^{73} | 539 | 0.0801 | 59 | 8 | 0.0071 | 52 | 6 |
| $ Q^r : 4323$ | C^{73} | 539 | 0.0745 | 55 | 6 | 0.0067 | 49 | 3 | |
| $ Q : 31739$ | | | | | | | | | |
| Medical 2011 | $ R : 127$ | T | 1443 | 0.0250 | 126 | 0 | 0.0020 | 103 | 0 |
| | | T+ | 1443 | 0.0285 | 137 | 2 | 0.0022 | | |