Assessor Error in Stratified Evaluation

William Webber
Dept. of Computer Science
and Software Engineering
University of Melbourne, AU

Douglas W. Oard
College of Information
Studies/UMIACS
University of Maryland, US

Falk Scholer
School of Computer Science
and Information Technology
RMIT University, AU

Bruce Hedin
H5, San Francisco, US

ABSTRACT

Several important information retrieval tasks, including those in medicine, law, and patent review, have an authoritative standard of relevance, and are concerned about retrieval completeness. During the evaluation of retrieval effectiveness in these domains, assessors make errors in applying the standard of relevance, and the impact of these errors, particularly on estimates of recall, is of crucial concern. Using data from the interactive task of the TREC Legal Track, this paper investigates how reliably the yield of relevant documents can be estimated from sampled assessments in the presence of assessor error, particularly where sampling is stratified based upon the results of participating retrieval systems. We show that assessor error is in general a greater source of inaccuracy than sampling error. A process of appeal and adjudication, such as used in the interactive task, is found to be effective at locating many assessment errors; but the process is expensive if complete, and biased if incomplete. An unbiased double-sampling method for resolving assessment error is proposed, and shown on representative data to be more efficient and accurate than appeal-based adjudication.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*effectiveness evaluation*.

General Terms

Experimentation, Measurement, Reliability.

Keywords

Estimation theory, e-discovery, recall.

1. INTRODUCTION

Topic 60 of the TREC Legal Track's Ad Hoc Task was estimated in 2007 to have a total of 83 relevant documents in its collection of seven million documents. That estimate was based on a sample of documents, ten of which were assessed relevant. The next year, those ten documents were used to seed a relevance feedback task. A new sample of documents was drawn from the results of that task, and again judged for relevance. This time, more than 36,000 relevant documents were estimated to exist for the same topic.

How can such different estimates be derived for the one document set? The cause is a combination of assessor disagreement and assessment scarcity. The 2007 and 2008 tasks had different assessors; the latter may have had a more expansive conception of topic relevance. Of the fourteen documents judged by both assessors, the earlier assessor found seven relevant, while the latter found ten [5]: a suggestive, though not statistically significant difference. Voorhees [15] reports that on average less than half of documents judged relevant to a brief written topic statement by one assessor are judged relevant by another similarly qualified assessor.

Assessor disagreement is not the sole explanation for the difference between 83 and 36,000 relevant documents. A second factor is how assessments are deployed. Realistic corpora today are too large for exhaustive evaluation by any one person. Either the task must be partitioned, at great cost, amongst multiple assessors (as, for instance, in Roitblat et al. [12]), or estimates must be derived from a sample of documents. For the evaluation of recall-oriented tasks over large collections, sampling and estimation is more rapid and less expensive than exhaustive assessment.

Sampling has an important flaw, however: it is insensitive to the very rare event. In many retrieval tasks, randomly encountering a relevant document would be precisely such a rare event. The rare event problem can be mitigated by arranging the collection so that relevance is sufficiently dense in some sub-population or *stratum*, then sampling the stratum. Fortunately, a tool is at hand for finding likely-relevant documents to put in the stratum: the information retrieval system itself. The union of retrieved results are commonly taken as the assessment pool. But by concentrating relevance in the pool, relevance is made even rarer, and sampling for it even less fruitful, in the remainder of the collection. The response is typically to ignore this vast but sparsely-populated stratum. There may be many relevant documents there, but if no system under evaluation returns them, then they will not affect relative scores. The relative effectiveness of systems can still be evaluated, even if the absolute proportion of relevant documents that systems have uncovered remains uncertain.

Assessor unreliability is generally considered in terms of assessor disagreement, not assessor error, since relevance is taken to be subjective. In any real search scenario, though, some authoritative conception of relevance does exist: that of the person for whom the search is performed. In end-user search, this is the searcher himself or herself. In librarian-mediated search, it is the patron's view of relevance that is ultimately authoritative. In finding evidence to present to a counterparty in civil litigation, or *e-discovery*, the authoritative view is that of the senior attorney who is responsible for certifying in court that the search conducted has, to an extent

CIKM'10, October 25–29, 2010, Toronto, Ontario, Canada.

commensurate with a reasonable good-faith effort, found all relevant documents. The attorney requires a reliable, absolute estimate, not just of how many relevant documents have been found, but of how many might remain to be found—an estimate, that is, of the search's *true recall*.

The interactive task of the Legal Track of TREC embodies the authoritative, recall-focused features of e-discovery. It provides a solid testbed for investigating recall estimation, under an authoritative conception of relevance, in the presence of assessor error. A *topic authority*, generally a practicing attorney, is assigned as the standard of relevance for each topic. The topic authority is available to teams when they are building and tuning their systems, and defines the conception of relevance that assessors must apply. Assessments that diverge from this conception are not just disagreeing: they are in error. The Legal Track uses a run-based stratification method to manage relevance density. But it also takes seriously the task of estimating the true relevance mass or *yield* of the collection. And in so doing, it runs back into the problem of sampling for the rare event. There is one stratum of documents, the *bottom stratum*, that consists of those documents returned by no team. This stratum makes up almost all the collection. It is sampled for assessment, but that sampling is necessarily sparse. If an assessor judges one of the documents in the sample as relevant, its sampling weight can amplify it into an estimate of thousands of relevant documents in the stratum. And assessors make errors—true errors, not just differences of opinion. So is this phalanx of relevant documents, lined up behind its sampled standard-bearer, real, or is it a phantom?

This paper tackles the problems that the interactive task of the Legal Track highlights; problems of sampling and estimation with an authoritative standard of relevance and error, and in which absolute measurements of effectiveness matter as much as relative ones. We begin in Section 2 by surveying related work in the area, and introducing the design and process of evaluation in the interactive task. In Section 3, we lay out an analytical framework for dealing with assessor error, based upon the modeling of measurement error. Using this framework, we demonstrate that, even with quite low rates of assessor error, the bias this error introduces readily dominates sampling error. Taking larger and larger samples gives us more and more precise estimates of the wrong quantity.

There is, therefore, no option but to tackle assessor error directly. The current technique for doing so is an appeals process. We analyze past appeal outcomes, finding that appeals are effective at identifying individual assessor errors. How complete the appeals are, however, remains uncertain. If the appeals are incomplete, their effect is likely to be biased and, whether complete or not, the process is expensive and time-consuming. Therefore, we propose in Section 4 that *double sampling* be applied, in which a sub-sample of the assessed sample is adjudicated by the topic authority, and error rates estimated based on this sub-sample. Double sampling provides an unbiased estimate of relevance yield in a stratum, and is a more efficient and controllable use of the topic authority's time. Finally, in Section 5, we outline future extensions to double sampling, to deal with the bottom stratum and to leverage appeal information.

While the Legal Track of TREC provides the data for our analysis, the issues addressed in this paper are important for tasks that extend beyond review for responsiveness, and beyond the legal domain altogether. Comprehensive search is important in medical literature reviews, evidence-based policy and practice, and patent examination [11]. In addition, as various forms of outsourcing increase the capacity for inexpensive but unreliable first-pass human assessment of retrieval tasks, efficient and reliable methods of expert quality assurance of these assessments become increasingly important.

2. BACKGROUND

Information is always retrieved as part of some larger task. In TREC, the task most often modeled is end-user search. In the Legal Track, however, the task is finding evidence to produce upon the request of a counterparty in civil litigation, the problem known as *e-discovery*. The process of e-discovery has many aspects. We focus on review for responsiveness: a search through documents in one's possession for those that are responsive to a production request that has been served in a lawsuit. The task can be viewed as identifying documents that merit human review before being produced, or (in a brave new world for which Roitblat et al. [12] argue) as doing the task automatically, with no review.

Information retrieval has a long tradition of experimental evaluation. The dominant approach is the test collection methodology. A test collection consists of queries, documents, and assessments of which documents are relevant to which queries. These relevance assessments are made by human *assessors*. Effectiveness measures are then calculated based on the return of relevant answer documents by systems under evaluation [16]. Several metrics are available, most of which are a function of *recall* (the proportion of all relevant documents retrieved) or *precision* (the proportion of retrieved documents that are relevant). In the Legal Track's interactive task, a system returns a set, rather than a ranking, of documents for each topic. Recall and precision are calculated directly on these sets. The F_1 score is the harmonic mean of precision and recall, and therefore presents a single effectiveness metric that incorporates both aspects of system effectiveness.

Gathering human relevance assessments is one of the most expensive and problematic aspects of test collection formation. Human judgment is subject to various cognitive, perceptual and motivational biases [10]. Cuadra and Katter [4] identify five broad groups of influencing factors: document variables; requirements statement variables; judgment conditions; judgment scales; and personal factors. Saracevic [13] surveys experimental work on these factors. Analysis by Voorhees [15] shows that while absolute effectiveness scores are sensitive to variations in relevance judgments, relative scores remain broadly stable. Wilbur [17] demonstrates that the best possible ranking of documents under the probability ranking principle converges as the number of sets of relevance assessments from independent judges increases, indicating some fundamental level of broad agreement.

The traditional test collection methodology assumes that all documents in a collection are judged in response to every query in the test set. As collection sizes have grown, exhaustive assessment has become infeasible. Evaluation campaigns such as TREC therefore make use of a *pooling* approach, where documents for assessment are taken from the answer lists of participating systems. Zobel [19] finds pooling robust in determining relative system rankings, but incomplete in identifying all relevant documents. Subsequent work has suggested that for very large collections, pooling may be unreliable even for relative comparisons [2].

There has been considerable recent interest in techniques for the efficient estimation of effectiveness metrics. Yilmaz and Aslam [18] introduce *infAP*, a method for estimating average precision using uniform sampling from the set of complete relevance judgments. A refinement is *statAP*, which uses stratified sampling, requiring smaller sample sizes than infAP for the same accuracy [3]. Stratified sampling was also used in the TREC Filtering Track [8]. The "capture-recapture" technique from environmental management is adapted by Kantor et al. [6] for the estimation of recall in large collections. Their approach requires independent document samples, however, which cannot in general be obtained from retrieval systems.

Plaintiffs request that Defendants produce all responsive documents on the following topics:

> 103. All documents which describe, refer to, report on, or mention any "in-store", "on-counter", "point of sale", or other retail marketing campaigns for cigarettes.

Figure 1: Topic 103 from the TREC 2008 Legal Track interactive task.

2.1 The TREC Legal Track interactive task

The interactive task of the TREC Legal Track models the conditions and objectives of document retrieval for purposes of discovery in American civil litigation. The context for the exercise is provided by a mock complaint which details the causes of action that motivate the hypothetical lawsuit. Associated with the complaint are requests for the production of documents that pertain to the practices, processes, or events that are at issue in the lawsuit. Each of these document requests serves as a separate topic for the purpose of the exercise. The production request for Topic 103 is shown in Figure 1; the corresponding complaint runs to 15 pages. For topics numbered 104 and below, the documents are scanned business records in the IIT CDIP 1.0 Test Collection [1]. For topics above 200, the documents are email messages (including all attachments) in the TREC 2009 Legal Track version of the Enron email collection [5].

Any document request, even one formulated with a high degree of specificity, leaves some scope for interpretation. The responding party must therefore make decisions as to what it does and does not consider responsive to the request, and be prepared to defend those decisions in court. The responsibility for making these decisions falls to the senior attorney overseeing the document production. Modeling the part played by this senior attorney, the interactive task includes a *topic authority*, charged with making authoritative decisions as to what will and will not be considered relevant to a topic. For teams participating in the exercise, whose role is that of a vendor or review team charged with executing the senior attorney's instructions, the goal is to retrieve *all* and *only* those documents that match the topic authority's conception of relevance. Teams are permitted to ask for up to ten hours of the topic authority's time for the purposes of clarifying the criteria for relevance to a topic.

A party served with a request for production must make a reasonable good-faith effort to produce all documents in its possession that are responsive to the request; recall is therefore a key measure of effectiveness at executing the task. At the same time, a party will typically want to avoid producing any documents beyond those that are responsive (and, in fact, a party can be sanctioned for grossly "overproducing"); precision is therefore also important.

Estimates of recall and precision are attained by the following protocol. For each topic that a team completes, it submits the set of documents that represents all, and only, those documents that are believed to be relevant to the topic. These sets enable a straightforward stratification of the test collection: one stratum containing the documents all participants deemed relevant (the *top stratum*), another stratum containing all the documents no participant deemed relevant (the *bottom stratum*), and other strata for the other possible combinations of participant assessments. For n participants, there will be a total of 2^n strata. Examples are given in Section 3.

Once the collection has been stratified for a given topic, the evaluation sample can be drawn. Strata are sampled in rough proportion to their size. Smaller strata are given a higher proportional sampling rate, to combat variability, while the bottom stratum, which is by far the largest, is given a lower sampling rate, to prevent it swamping the assessment pool. Within each stratum, documents are selected using uniform random sampling without replacement. Once assessed, the sample serves as the basis both for estimating the proportion of relevant documents in the full population and for estimating the recall and precision achieved by each team.

Team	Appeal Results				Total
	$0 \to 1$	$1 \to 0$	$0 \to 0$	$1 \to 1$	
A	35	0	5	0	40
B	1	1	2	0	4
C	525	216	176	16	933
D	0	0	0	0	0
Gross Total	561	217	183	16	977
Net Total	535	217	182	16	950

Table 1: Number of appeals by team and appeal result for Topic 103 from the TREC 2008 Legal Track, interactive task. The column $0 \to 1$ counts the number of documents that were initially judged not relevant, but were subsequently appealed, with the appeal being upheld by the topic authority, leading to a final judgment of relevant; other column labels can be similarly interpreted. The gross total of appeals by class counts the total number of appeals lodged; the net total counts the distinct number of documents appealed. The latter may be lower than the former, because some teams appealed the same document.

In order to have a document sample assessed, the documents in the sample are randomly assigned to "bins," each holding approximately 500 documents, so that a bin can typically be completed by a single assessor in 20 to 25 hours. These bins are then distributed to volunteer assessors, who, equipped with detailed assessment guidelines (reflecting the guidance the topic authority gave to the teams), assess the contents of their bin(s) for relevance to their topics. Assessors are drawn principally from two populations: review teams from professional document review firms (for Topics 203, 204 and 207), or law school students and legal professionals.

To partially mitigate the effect of assessment errors, teams are given access to all first-pass assessments and invited to appeal to the topic authority any assessments they believe are incorrect. The topic authority then renders a final judgment on all appealed assessments. Once appealed assessments have been adjudicated, final estimates of each team's recall and precision are obtained.

For example, four teams submitted runs for Topic 103 at TREC 2008; they are identified in Table 1 as Teams A, B, C, and D. Topic 103 was also included in the Ad Hoc Task for that year, and an additional pseudo-team was created for the interactive task by pooling each of the Ad Hoc submissions (to a maximum depth of 100,000); this pseudo-team is labeled Team E. There are thus 5 teams altogether, making $2^5 = 32$ different strata. A total of 6,500 documents were sampled for assessment. These were divided into 13 bins. There were ten assessors, nine of whom assessed one bin each, while the tenth assessed four. A total of 2,663 documents were assessed as relevant, 3,758 as not relevant, and 79 as unassessable. Some 950 of those assessments were appealed, or nearly 15% of all assessments made. Almost all (98%) of those appeals were made by one team, Team C. Most appeals (72%) were made against initial assessments of not relevant, and most appeals (79%) resulted in a change of the initial relevance assessment.

Stratified sampling provides the basis for score estimation. Let L be the set of strata. The essential quantity for all subsequent calculations is R_l, the number of relevant documents or *yield* of stratum l. Other formulae follow trivially from this. For instance,

the yield of the collection as a whole, R, is:

$$R = \sum_{l \in L} R_l \ . \tag{1}$$

Let L_A be the subset of strata returned by team A. Denote the number of assessable documents in stratum l as N_l (some documents are unassessable due to corruption or excessive length). The total number of assessable documents returned by system A is $N_A = \sum_{l \in L_A} N_l$, and the yield, recall, and precision of A are:

$$R_A = \sum_{l \in L_A} R_l \tag{2}$$

$$\text{Rec}(A) = \frac{R_A}{R} \tag{3}$$

$$\text{Prec}(A) = \frac{R_A}{N_A} \tag{4}$$

The true yield for each stratum is unknown, but it can be estimated from the sample on that stratum. Let n_l be the size of the sample from stratum l, and r_l be the yield of the sample. Then the yield of stratum l can be estimated as:

$$\widehat{R_l} = \frac{r_l}{n_l} \cdot N_l \ . \tag{5}$$

This simple formula forms the basis of the stratified sampling approach; different strata can be sampled at different intensities, with the sample results reweighted to reflect the probability of an item being sampled. The estimate $\widehat{R_l}$ is then used in place of R_l in the above formulae to derive point estimates of recall, precision, and total yield. Expressions for the variance of the resulting estimators are given in Oard et al. [9].

3. ASSESSOR ERRORS

The estimations of the yield of a stratum presented above assume that the relevance assessors make no errors. In reality, though, assessors make errors, returning assessments that disagree with the topic authority's conception of relevance. The estimate of the proportion relevant is, therefore, subject to *measurement error*. Measurement error introduces bias into the estimate, which is not reduced by increasing sample size.

Assessors make two types of error: they assess relevant documents as not relevant, and they assess not relevant documents as relevant. The proportion of not relevant documents assessed as relevant is the *false positive rate*, α, while the proportion of relevant documents assessed as not relevant is the *false negative rate*, β. Different assessors have different error rates; the same assessor can have a different false positive and false negative error rate; and an assessor's rate of each type of error will vary depending upon the nature of the topic and the documents being assessed. Even if two error rates are the same, their effect will differ depending upon the true proportion of documents in a stratum that are relevant.

The interaction between error rates and relevance proportions is illustrated by Figure 2. The proportion π of documents that are assessed as relevant depends not only on the false positive and false negative rates α and β, but also on the proportion of relevant and not relevant documents p and $q = 1 - p$. Specifically:

$$\pi = \alpha q + (1 - \beta)p \ ; \text{and} \tag{6}$$

$$(1 - \pi) = (1 - \alpha)q + \beta p \ . \tag{7}$$

If we estimate the true proportion relevant, p, based upon the assessed proportion relevant π, then the bias of the estimate is:

$$\text{bias} = \pi - p = \alpha q - \beta p \ . \tag{8}$$

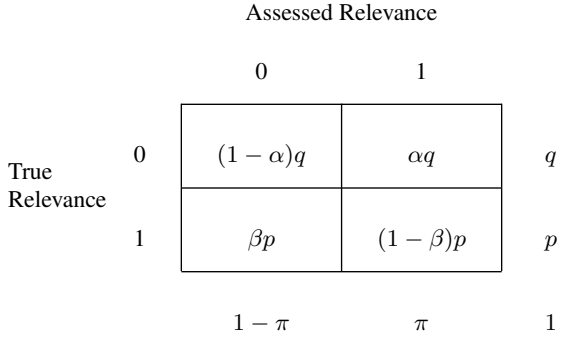


Figure 2: Assessor error types and rates. The value α gives the false positive rate; that is, the proportion of not relevant documents incorrectly assessed as relevant. The value β gives the false negative rate; that is, the proportion of relevant documents incorrectly assessed as not relevant. The true proportion relevant is p, and the true proportion not relevant is $q = 1 - p$, while π is the assessed proportion relevant.

The size and sign of the bias therefore depend on the interaction between error rates and proportions relevant [14].

False positives and false negatives do not simply cancel out if the error rates are the same. Take, for instance, a stratum where $p = 0.01$ (one in a hundred documents relevant—a plausible proportion for the bottom stratum). If error rates on this stratum were $\alpha = \beta = 0.05$, the proportion π fallibly assessed relevant would be 0.049, creating a positive bias of 0.039. Although error rates are the same, the highly unequal proportion of relevance causes the proportion to be overestimated almost five-fold. Indeed, with such a low proportion actually relevant, the false negative rate is all but immaterial. The reverse situation occurs in densely relevant strata, such as the top stratum, returned by all systems. But the overstatement of relevance on the bottom stratum is more serious, since this stratum is generally very large and thus of necessity sparsely sampled. Estimation errors on the bottom stratum therefore have much greater absolute effects than errors on other, smaller strata.

Consider the example in Table 2. The true yield for Team B is the sum of the yields (τ in the table) for the stratum of documents

Stratum	$\lvert l \rvert$	p	τ	α	β	π	$\widehat{\tau}$
b1k1	1.7×10^3	0.93	1569	0.100	0.15	0.80	1346
b1k0	1.7×10^3	0.25	427	0.250	0.10	0.41	711
b0k1	1.3×10^3	0.71	929	0.100	0.50	0.38	503
b0k0	5.6×10^5	0.002	965	0.015	0.25	0.016	9174
Total	5.7×10^5	0.007	3890	0.016	0.25	0.021	11734

Table 2: Worked example of the effect of error rates, based upon Topic 202. Two teams participate, marked B and K. Each stratum l, of size $\lvert l \rvert$, is labeled by the teams that did (1) or did not (0) return documents in the stratum. We assume all documents are assessed. The true proportions relevant, p, produce the true yields, τ, on each stratum. But the false positive (α) and false negative (β) assessment error rates lead to the incorrect relevant proportions π being observed, and hence the incorrect yield estimates $\widehat{\tau}$ for each stratum. The true values from Topic 202 are used for $\lvert l \rvert$; π is as observed on the assessed samples, and $\widehat{\tau}$ is extrapolated from that; and p, τ, α, and β are hypothesized based on the outcome of appeals. The "Total" line shows sums for counts, weighted averages for proportions.

returned by both Team B and Team K (Stratum b1k1), and the stratum returned by Team B but not by Team K (Stratum b1k0). The true recall for a team is the team's yield divided by the total yield of the corpus. True recall is 0.51 for Team B, and 0.64 for Team K. Assessment errors, however, favor Team B: the false positive rate is higher, and the false negative rate lower, for documents returned only by Team B (Stratum b1k0) than for those only by Team K (Stratum b0k1). If the relevance proportion of the bottom stratum, returned by neither team (Stratum b0k0), were correctly estimated, then the error-affected recall scores would be 0.58 for Team B, and 0.52 for Team K, reversing the true ordering of the systems. But the greatest impact upon absolute scores is from errors in the bottom stratum. The false positive rate is only one in sixty—plausible for errors of inattention alone. But the predominance of not relevant documents (99.8%) and the size of the stratum (99% of the collection) causes even so low an error rate to swamp the true relevant count for other strata. The estimated recall for Team B is depressed to 0.18, and for Team K to 0.16. The score ratio is the same with and without errors in the bottom stratum, but absolute scores have fallen by 70%. Both teams have found most of the relevant documents in the collection, but even a small false positive rate in the bottom stratum makes it seem that they have found only a fraction.

More formally, the error of an estimator can be expressed in terms of its mean squared error (MSE). For an unbiased estimator, MSE equals variance; for a biased one, MSE is variance plus the square of the bias. In the absence of measurement error, the sample proportion is an unbiased estimator $\widehat{p}_T$ of the population proportion p, having an MSE for sample size N of:

$$\text{MSE}(\widehat{p}_T) = \text{var}(\widehat{p}) = \frac{p(1-p)}{N} \ . \tag{9}$$

In the presence of measurement error, the sample proportion provides a biased estimator $\widehat{p}_F$, whose MSE is:

$$\text{MSE}(\widehat{p}_F) = \frac{\pi(1-\pi)}{N} + (\alpha q - \beta p)^2 \ , \tag{10}$$

consisting on the left of the variance of sampling from the fallible assessments on the population, plus on the right the squared bias of the fallible from the true proportion. The variance term can be reduced by taking a larger sample, but bias is unaffected.

An estimator's bias can be said to dominate its sampling error when the bias is twice the sample standard deviation ($\sqrt{\text{var}}$), since the bias then places the expected value outside the sampling error's 95% confidence interval. Beyond this point, increasing sample size loses traction: meaningful improvements in accuracy can only be achieved by reducing measurement error. Figure 3 shows this threshold for decreasingly small true proportions relevant p, such as might be found in the bottom stratum, and varying but low false positive error rates α, such as might occur from assessor inattention. If one in ten documents is relevant, then the sample size can usefully be increased to 2,000 or so, even with a false positive rate of one in twenty. But if the true proportion of relevance is one in a thousand, then even a small false positive error rate of 0.01 renders sampling beyond a few hundred ineffective. And if the bottom stratum is sufficiently large, the (biased) yield estimates from this stratum will dominate the recall calculation.

3.1 Estimating errors without appeals

An aspect of the Legal Track setup that assists in the detection of assessor error is its *parallel assessment*. An equal number of documents from each stratum are randomly assigned to each bin, and the bins are independently assessed by different assessors. There are also *reserve bins*, used for supplementary assessment tasks, which are not assigned in the same way; the current analysis excludes

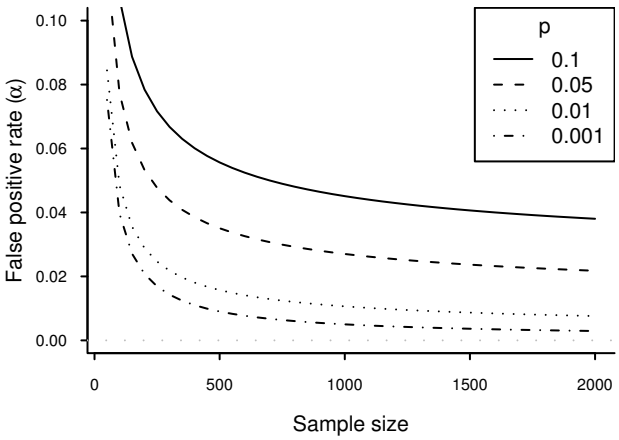


Figure 3: The threshold at which bias dominates sampling error, for various true proportions relevant p. The line shows the combination of false positive rate and sample size at which bias is twice the sampling standard error; increasing sample size beyond this point shrinks the 95% confidence interval such that it no longer includes the expected value of the (biased) estimator.

these, and works solely with the *core bins*. Parallel assessment means that each bin should have randomly the same characteristics. In particular, each bin should hold approximately the same true proportion of relevant documents. If the proportion assessed as relevant differs significantly between bins, this indicates both the existence, and the (net) scale and direction, of assessor error.

The presence of assessment errors can be probed using a statistical test on proportions. The proportion relevant is found significantly different between core bins for nine of the ten TREC 2008 and 2009 topics (using a χ^2 test on proportions; achieved significance level $\alpha = 0.001$), demonstrating the presence of assessment errors with different net biases. The exception is Topic 207, whose proportions are not significantly different. Topic 207 was assessed by a commercial provider of legal document review services; the provider's internal processes may have enforced a consistency of interpretation, or at least a random assignment of assessment duties. This does not mean that there are no assessment errors in Topic 207; it just means that the errors are (statistically) consistent between bins.

The proportion assessed relevant per bin reveals not only the presence of errors, but also their net magnitude and direction. This is illustrated for Topic 103 in Figure 4. The proportions are highly unequal, ranging from 0.21 to 0.62. The relative bias of each bin compared to the others can readily be observed. For instance, the assessor of Bin 5 is the most *conservative* in the interpretation of relevance, while the assessor of Bin 9 is the most *liberal*. The net bias of the former, relative to the mean proportion assessed relevant, is -0.21, while that of the latter is $+0.20$.

While randomly parallel assessment provides some data about assessor bias, the information is of limited use in compensating for measurement error. First, Figure 4 only reveals relative measurement biases, not absolute ones. The mean proportion assessed relevant amongst bins is not necessarily the true proportion relevant for the full sample. It could be that all assessors are more conservative than the topic authority, or more liberal. Second, even if the true mean were known, the assessed proportion relevant only reveals net bias, not gross error rates. A bin that hit the true mean might contain no errors, but it might contain many errors in both directions that simply canceled out.

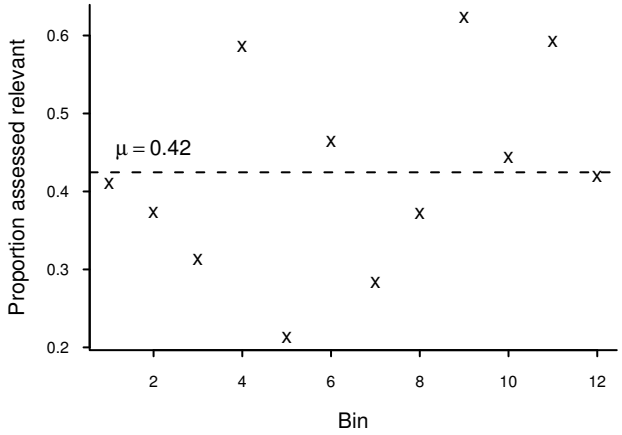


Figure 4: The proportion of sampled documents assessed as relevant for each of the twelve core bins of Topic 103, from the TREC 2008 Legal Track interactive task, along with the mean number assessed relevant across these bins. The proportions are significantly different in a χ^2 test at level $\alpha = 0.0001$; their standard deviation is 0.13.

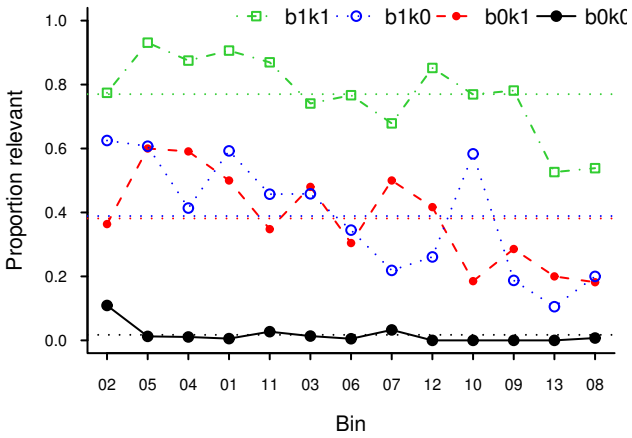
Another limitation of the data shown in Figure 4 is that it reveals the biases of bins, whereas what is wanted is the error rates of strata. Figure 5a breaks down the bin figures per stratum, this time for Topic 202 from TREC 2009. Parallel assessment means that for each bin across a stratum, the proportions should be randomly equal; marked divergences indicate relative biases. Some assessors, for instance, are markedly more inclined to agreed with Team B, others with Team K. However, these results again do little to help correct for assessment error, since they only reveal net, relative bias. We might be tempted to take the mean per-stratum proportion relevant as an estimate of the true proportion, and adjust strata yields accordingly, but this adjustment would be to no effect, since the mean per-stratum relevance proportion is precisely the overall proportion assessed relevant for that stratum.

3.2 Grounds for appeal

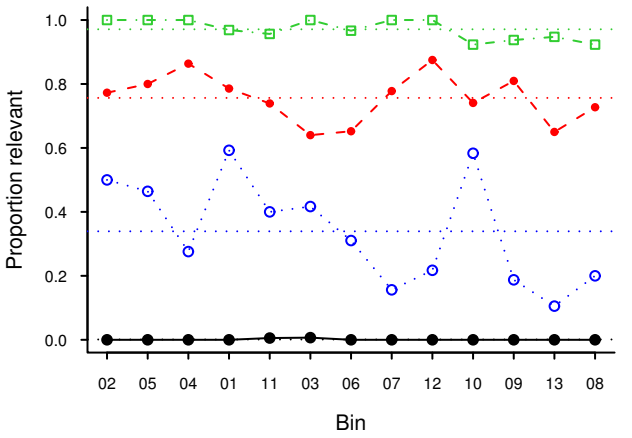
A team's interaction with the topic authority, combined with their own analysis and expertise, means that they may well develop a stronger understanding of the authoritative conception of relevance than the assessors. In Topic 103, for instance, Team C has a lower post-appeal false positive rate than 5 of the 13 bins (corresponding to 5 of the 10 unique assessors), and a lower false negative rate than 10 of the bins (corresponding to 7 unique assessors). In addition, teams are a source of evidence about relevance that is independent of the assessors.

As a corrective to assessor error, teams may appeal assessments to the topic authority. The topic authority adjudicates all appealed assessments, and only those appealed assessments. The appeals process is highly effective at finding errors. From 70% to 90% of appeals are successful, depending on the topic. Where teams are active in appealing, the scale of appeals can be quite substantial, with well over 10% of assessments appealed for several topics.

Appealing teams must lodge a document setting out their grounds for appealing each assessment. These appeal documents provide information on the different types of assessor error. For Topic 103 in the 2008 Track, Team C identified 17 grounds for appeals; of these, 5 were to challenge perceived false negatives, and 12 were for perceived false positives. Each of the 17 grounds or *aspects* is therefore one-directional in terms of the error that it seeks to address. We have coded these aspects numerically, with aspects 1 through 5 being false negatives, and 6 through 17 false positives.



(a) Pre-appeal



(b) Post-appeal

Figure 5: Proportion assessed relevant, pre- and post-appeal, for each core bin on every stratum, for Topic 202 from the TREC 2009 Legal Track interactive task. Each line represents the proportions for one stratum. The bins are ordered by decreasing total proportion assessed relevant. The mean assessed relevant for each stratum is shown.

Table 3 shows the upheld false positive appeal counts for selected bins. Different assessors are subject to different kinds of false positive assessments; over all bins, a χ^2 test finds variations between assessor and aspect significant at level $\alpha = 0.01$. For instance, the assessor of Bin 103.011 is particularly subject to errors on Aspect 12 (mistaking a program to discourage smoking for a marketing campaign to sell cigarettes). Interestingly, for false negatives, the interaction is not significant. One hypothesis is that false positives are errors of interpretation (the assessor mistakes text they see

Bin	6	7	8	9	10	11	12	13	14	15	16	17
103.004	1	1	2	4	3	1	6	5	1	0	2	9
103.009	3	3	5	1	0	1	3	1	3	2	0	15
103.011	3	2	2	3	1	2	14	3	1	0	1	15

Table 3: Count of upheld appeals for false positives by aspect, for selected bins. The appeals are from Team C for Topic 103 of the TREC 2008 Legal Track interactive task.

Stratum	6	7	8	9	10	11	12	13	14	15	16	17
a0b0c0d0e0	3	7	9	5	3	3	2	12	1	0	2	4
a0b0c0d0e1	7	0	3	5	5	2	22	11	8	2	2	22
aXbXc0dXeN	4	8	1	3	1	4	11	0	7	5	2	30

Table 4: Count of upheld appeals for false positives by stratum and selected aspect. Stratum `aXbXc0dXeN` is aggregated from the documents not returned by Team C but returned by at least one of Teams A, B, or D.

Topic	χ^2 p-value on prop. rel.		% assmnt.
	Assessed	Appealed	appealed
t102	< 0.01	< 0.01	0.2
t103	< 0.01	< 0.01	14.7
t104	< 0.01	< 0.01	0.2
t201	< 0.01	0.71	11.5
t202	< 0.01	0.22	11.8
t203	< 0.01	0.40	8.0
t204	< 0.01	0.51	6.3
t205	< 0.01	< 0.01	22.2
t206	< 0.01	< 0.01	1.3
t207	0.60	0.75	4.3

Table 5: The p values of a χ^2 test on the proportion of assessable sampled documents assessed relevant, before and after appeals, per core bin for each of the TREC 2008 and TREC 2009 Legal Track interactive task topics. The proportion of core bin assessments that are appealed is shown in the sixth column.

as evidence of relevance), and thus are dependent on the assessor's conception of relevance, while false negatives are largely errors of inattention (the assessor fails to see valid evidence of relevance), and hence are independent of conceptions of relevance.

Table 4 shows the upheld false positive appeal counts for the strata that Team C did not return. (Team C naturally did not appeal false positives for strata they did return.) The interaction here between stratum and aspect is even stronger than for bins, being significant at level $\alpha = 0.001$. Different systems return documents containing different specious grounds for being judged relevant. For instance, none of the actual teams (Team E being a composite of Ad Hoc Task runs) returned documents that embody errors on Aspect 13 (mistaking market research or product testing for marketing campaigns). Conversely, the most common false positive errors, Aspect 17 (assessing as relevant documents about anti-smoking and advertising legislation in general, when they had to refer to a particular advertising campaign) and Aspect 12 (described above), are relatively rare in the bottom stratum. There is also a strong stratum interaction for false negatives, significant at the level $\alpha = 0.01$.

This analysis only scratches the surface of what is a rich source for exploring (mis)conceptions of relevance. Its importance for us here is to indicate the variegated nature of errors. This should caution against too simple an error model, which attempts to apportion overall error rates amongst bins and strata, without appreciating the interactions between these two factors and the possible aspects of assessor error for a topic. We cannot simply conclude that if assessor A has twice the error rate of assessor B, then assessor A will have twice the error rate on every stratum.

3.3 Estimating errors from appeals

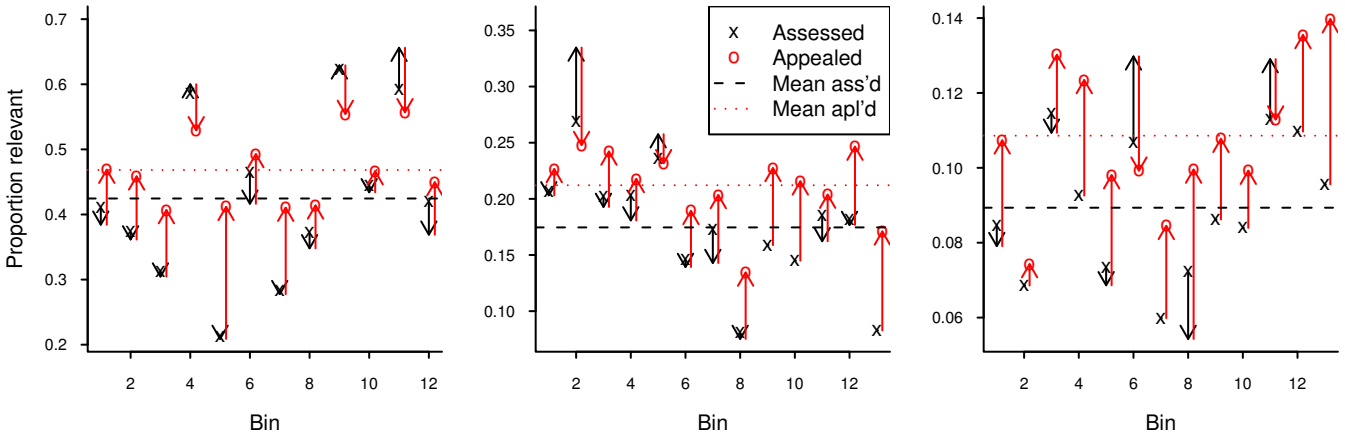
The appeals process relies upon teams to locate and object to assessment errors; if teams fail to do so thoroughly (due to a lack of resources to check assessments and write out appeals documents, for instance), assessment errors will remain. The completeness of appeals can be tested by examining post-appeal proportions relevant. If the proportion is significantly different between bins, then a sizeable number of assessment errors are yet to be found. Table 5 shows that the number of significantly uneven topics falls from nine prior to appeals to five post appeals, and of these still uneven five, three have been very sparsely appealed. Thus, in most (but not all) cases where teams are active in appealing, proportions relevant are largely evened out amongst bins. But a lack of significant differences between bins does not prove that all, or even most, errors have been found, merely that the errors remaining do not result in substantially uneven net biases. A high appeal rate is no guarantee of resolving all errors.

The outcome of the appeals process also provides information about the nature and distribution of errors. With (randomly parallel) assessments alone, only relative, net biases can be inferred. But appeals uncover specific false negatives and false positives, and so provide evidence about these errors rates. Figure 6 displays the per-bin outcomes of the appeals process for Topics 103, 202, and 207. The first two topics have high appeal rates (15% and 12% of core bin assessments appealed), while the last is lower but still substantial (4%). The topics also have different relevance densities. In Topic 207, 10% of the sampled documents are assessed relevant (post-appeal), implying 1.5% of the document collection (adjusting for different strata sampling rates). In Topic 103, 46% of the sampled documents are assessed relevant (post-appeal), making 11% of the document collection. At least half of the bins in Figure 6 have both false positives and false negatives; assessors are in general not simply liberal or conservative, but make a mixture of errors. Some assessors demonstrate high gross error rates in both directions, which largely cancel each other out to a small net bias. This could, of course, simply be a sign of genuine, countervailing errors of interpretation; but such assessors are also natural starting points for investigating assessor inattention and unreliability.

If the appeals process is complete, or very nearly complete, then both the revealed error rates and the final evaluation will be reliable, as the great majority of errors will have been found—assuming, as we do throughout, that the topic authority itself does not make errors of inattention.[1] But there is no sure way of knowing that the process is complete, since balanced relevance proportions can occur even with incomplete appeals. And if the appeals process is incomplete, then the question of bias arises. A team has no incentive or occasion to appeal judgments that do not go against it. Effort put into appealing by any team can only help that team's score, and so the most energetic appealers stand to gain the largest benefit. Figure 7 indicates that there is indeed a strong correlation between the number of appeals a team lodges and the improvement in that team's score. One interpretation would be that good teams are energetic in appealing; an alternative explanation would be that energetic appealers are rewarded with good scores.

The rewards for active appealing can be observed in the post-appeal proportions relevant for each bin and stratum, displayed for Topic 202 in Figure 5b; this figure should be compared with the pre-appeal rates in Figure 5a. Team K lodged four times as many appeals as Team B; and Team B's appeals are almost entirely a subset of Team K's, leaving only a handful that solely benefited

[1] Of course, the topic authority might indeed make random errors of inattention, and moreover if their conception of relevance were to change over time then additional systematic inconsistencies could arise. We leave accommodation for such effects to future work, but observe that Lam and Stork suggest ways of accommodating estimated topic authority error rates when system and topic authority errors are independent [7].



(a) Topic 103 (b) Topic 202 (c) Topic 207

Figure 6: The outcome of the appeals process for three topics. The topics are Topic 103 from the TREC 2008 Legal Track interactive task (left), and Topics 202 and 207 from TREC 2009 interactive task (middle and right). The arrows show the smaller correction first, followed by the larger correction. Note differences in the y scale.

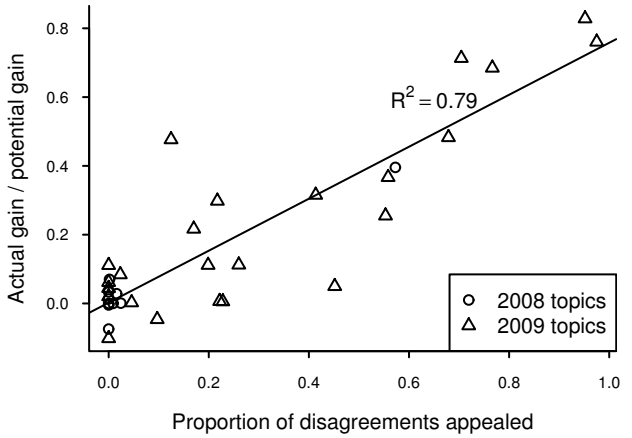


Figure 7: Relationship between number of appeals a team lodges and the improvement in that team's score, for all teams and topics in the TREC 2008 and TREC 2009 Legal Track interactive task. The x axis shows the proportion of overall assessments that the team appealed. The y axis shows the proportion of the potential improvement in a team's F1 score (that is, the difference between the team's pre-appeal F1 score and a perfect score of 1.0) that was achieved post-appeal. Each point in the figure represents a team. Some team scores are lower post-appeal, due to the negative affect of competing teams' appeals. The line of best fit for the combined years' points is shown.

it. The appeals process has raised the mean proportion relevant in Stratum `b0k1`, returned only by Team K, from 39% to 77%, while depressing that of Stratum `b1k0`, returned only by Team B, from 42% to 36%. The relevance proportions for Stratum `b0k1` are substantially evened out, while those for Stratum `b1k0` remain significantly uneven (at level $\alpha = 0.001$). False positives may remain unfound in Stratum `b0k1`, though, since almost no relevant assessments have been challenged in this stratum; and it is almost certain that there are unfound false negatives in Stratum `b1k0`, based on both the lack of challenges and the bins' significant unevenness. Also, while the top stratum is showing 97% relevant, and the bottom stratum 0.1% (only 3 documents remain assessed relevant), both these figures may be questioned as exaggerations, since in neither stratum are there any appeals against the majority assessment. The caveat about these two strata holds for every topic, as no team has incentive or occasion to appeal not relevant assessments in the bottom stratum, or relevant ones in the top.

Bias is one objection to the appeals process; an equally strong one is cost. All appealed documents go to the topic authority for adjudication. The topic authority is generally a senior attorney, and although their time is given to the task *pro bono*, it remains a limited resource. In a real e-discovery procedure, such as the interactive task is simulating (or indeed in other expert-assured retrieval processes), time spent in post-retrieval review by the senior attorney attracts real and considerable costs. The topic authority is also a bottleneck, since the work cannot be shared, and the adjudication of appeals substantially delays the reporting of final results. A more efficient method is desirable to save both expended and elapsed time. The appeals process also leaves the thoroughness and cost of adjudication in the hands of the teams: their willingness and eagerness to lodge appeals, rather than the objective evaluation requirements of the task, determines how much and what adjudication is performed. In the next section, we propose a method that avoids the bias and expense of appeal and adjudication.

4. SAMPLING THE SAMPLES

The appeals process has two main problems. First, it is too expensive: every document appealed is adjudicated, when only a proportion of them might suffice. And second, it is subject to bias: appeals can only help the team appealing, and some types of errors may not be adequately characterized. Both problems suggest the same solution: sampling. Sampling the appealed assessments will reduce the expense of adjudicating them; and sampling the unappealed assessments will provide evidence for estimating their accuracy. And above all, sampling will make our estimates unbiased.

A sampling-based approach to mitigating measurement error is *double sampling*. Double sampling is of use where there are two types of measurement: an inexpensive but fallible one, and a more expensive but authoritative (or "true") one. In our case, these correspond to the assessments of the assessors and of the topic authority.

Under double sampling, a large initial random sample is taken of the population, and measured using the fallible standard. Then, a subsample of this initial sample is measured using the authoritative standard. The error rates of the fallible standard, relative to the authoritative one, are estimated on this subsample, and adjustments are applied to the estimates derived from the full, initial sample.

Put more formally, let the size of the initial sample (of documents judged by the assessors) be N, and of the subsample (of documents adjudicated by the topic authority) be n. Denote the number of the n subsampled items that are classified t by the true classifier, and f by the fallible classifier, as n_{tf} ($t, f \in \{0, 1\}$). So, for instance, n_{10} is the number of documents in a given stratum assessed as irrelevant by the assessor, but adjudicated as relevant by the topic authority; that is, the count of false negatives. Further, let $n_{\cdot f}$ denote all the subsampled documents classified as f by the fallible classifier, regardless of the ruling of the true classifier. There remain another $N - n$ items for which we only have the fallible classification. Let Y be the number of these $N - n$ items classified as 0 (irrelevant), and $X = N - n - Y$ be the number classified as 1 (relevant). Then the maximum-likelihood estimators for the true proportion of relevant documents p, the false positive error rate α, and the false negative error rate β, are:

$$\widehat{p} = \frac{n_{11}}{n_{\cdot 1}} \frac{X + n_{\cdot 1}}{N} + \frac{n_{10}}{n_{\cdot 0}} \frac{Y + n_{\cdot 0}}{N} \quad (11)$$

$$\widehat{\alpha} = \left(\frac{n_{01}}{n_{\cdot 1}} \frac{X + n_{\cdot 1}}{N} \right) / (1 - \widehat{p}) \quad (12)$$

$$\widehat{\beta} = \left(\frac{n_{10}}{n_{\cdot 0}} \frac{Y + n_{\cdot 0}}{N} \right) / \widehat{p} \quad (13)$$

Write q for $1 - p$, the true proportion irrelevant; and recall from Equation (6) that π is the proportion of documents classified as relevant by the fallible classifier (on the population, and in expectation on the sample). The asymptotic variance of the estimator of the proportion relevant is then given as [14]:

$$\text{var}(\widehat{p}) = \frac{pq}{n} \left[1 - \frac{pq(1 - \alpha - \beta)^2}{\pi(1 - \pi)} \right] + \frac{p^2 q^2 (1 - \alpha - \beta)^2}{N \pi (1 - \pi)} . \quad (14)$$

For $n = N$, Equation (14) simplifies to Equation (9), as expected.

The estimate of p in Equation (11) is an unbiased one. The crucial question then is to compare the (asymptotic) variance of double-sampling, given in Equation (14), with the mean squared error of an incomplete appeals process (one in which some assessment errors are unappealed and undiscovered); the latter MSE value can be calculated from Equation (10). We make this comparison with illustrative data, taken from Stratum `h0i0k1` of Topic 203.

Consider a stratum with $p = 0.61$, $\alpha = 0.16$, and $\beta = 0.83$ (the observed post-appeal values for the sample on Stratum `h0i0k1`). Let 113 documents be sampled from this stratum for assessment. If the full sample were re-sampled, then assessor error would be eliminated. The resulting proportion relevant in the sample serves as an unbiased estimator of the proportion relevant in the population. The standard deviation (SD) of this estimator is given by the square root of Equation (9), and is 0.046. Under the normal approximation, the value of the estimator will fall in the range 0.61 ± 0.090 roughly 95% of the time. As we reduce the re-sample size, the SD of the estimator increases, as specified in Equation (14), and shown in Figure 8. When $n \ll N$, the first term of Equation (14) dominates; therefore, quartering the re-sample size roughly doubles the standard deviation of the estimator, as Figure 8 confirms.

In the actual evaluation of Stratum `h0i0k1`, some 57 false negatives were located, all through the appeals of Team K. To illustrate the MSE of an incomplete appeals process, imagine that Team K had been less thorough with appealing. Every false negative they miss increases the false negative rate, β, of the post-appeal assessment. The resulting MSE can be calculated from Equation (10). Imagine that Team K missed 10 false negatives in their appeals, or 18% of all false negatives in the stratum; hardly a great omission on the team's part. The resulting RMSE is 0.10, compared to the sampling SD of 0.046; bias exceeds standard deviation. And some 64 documents would still have been appealed and adjudicated (even assuming teams only appealed true errors). An equivalent (unbiased) RMSE is achieved by re-sampling and adjudicating just 20% of the full sample, or 23 documents; in addition, the latter method requires no effort by teams in writing up appeal documents. One has to be quite confident in the thoroughness of all participants in appealing (and the patience of the topic authority in adjudicating) before one should expect the appeals process to be more accurate than one-in-five double-sampling.

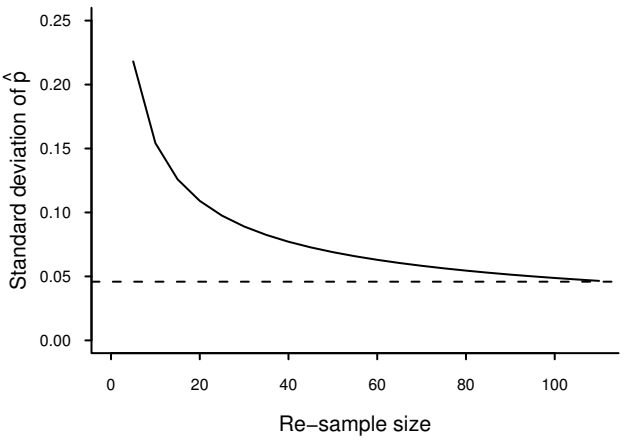


Figure 8: The standard deviation of a double-sampled estimate for varying n, with $p = 0.61$, $\alpha = 0.16$, $\beta = 0.83$, and $N = 113$ (values derived from Stratum `h0i0k1` of Topic 203 in the TREC 2009 Legal Track interactive task). The SD with full adjudication is shown by the dashed line.

Double sampling, combined with error modeling, provides unbiased estimates of the proportion relevant, and through that of yield, precision, and recall. The amount of double sampling and adjudication can be varied between strata, and between topics, depending upon resources and conditions. An appeals process is not required—although where appeals information is available, it is attractive to integrate this into the double sampling method, something which will be outlined in the concluding section.

5. CONCLUSION AND FUTURE WORK

The interactive task of the Legal Track of TREC must confront two problems that some (but not all) evaluation domains in IR can evade: an authoritative conception of relevance, and hence genuine assessor errors; and absolute measures of retrieval performance, founded upon this authoritative conception. Absolute measures in turn require that the problem of sampling for the rare event be tackled, whether that rare event is relevance or error. To date, the track has used two essential tools. First is stratified sampling of assessments, with an allocation of assessments to all strata, including the bottom stratum of unreturned documents; and second is an appeals process, to highlight likely errors by the numerous volunteer assessors, and have them adjudicated by the topic authority.

The technical insight put into designing, and the effort and resources into implementing, the task's assessment protocol, has re-

sulted in a degree of accuracy exceptional for retrieval tasks of this scale. Sampling the bottom stratum is already an advance on much standard practice. Yet even stratified sampling is insufficient without a method for tackling assessor error, such as the task's appeal process provides. This paper has identified a number of remaining problems. First, the appeals process is expensive and time-consuming, requiring professional adjudication of many hundreds of assessments. Second, the appeals process is dependent upon the engagement of participating teams. If none are active appealers, then few errors are caught; if only some actively appeal, the results are likely to favor them. And finally, even where the appeals process is extensive, there is no way of proving that it is complete.

As well as describing and analyzing these problems, this paper has suggested an interlinked pair of solutions to them. The first is an error model, which decomposes assessor error into false positive and false negative rates, and accounts for the interaction of these rates with the relevance densities of different strata. The second is a double sampling approach, where rather than adjudicating all and only appealed assessments, the assessments for adjudication are sub-sampled, to provide unbiased and cost-effective estimates of error rates.

There are a number of specializations that can be made to the double sampling approach. In particular, it has been noted previously that an appeals process is very effective at identifying individual assessor errors, even if adjudicating all appeals is expensive and potentially biased. If appeal information is available, then it can be incorporated into the double sampling method, by sub-stratifying the stratum into appealed and unappealed documents. Determining optimal sampling rates for such an approach, and deriving estimators and variance expressions, remains as future work.

A particular challenge of stratified evaluation is posed by the bottom stratum, where relevant documents, and therefore both true positive assessments and false negative errors, are rare events, reducing the effectiveness of double sampling. A pre-stratification of the bottom stratum into appealed and unappealed documents helps concentrate potential errors. And since all relevant assessments in this bottom stratum are suspect, and high-impact, the manual appeals process could be skipped altogether for this stratum, and all judgments of relevant auto-appealed (which, under double-sampling, does not mean they must all be adjudicated). This addresses true and false positives, but what of false negatives? The solution would appear to lie, not in a double sampling of the not relevant assessments in this bottom stratum, which would be sampling for the (very) rare event; but rather by making use of our error model, and observing that false negatives are the complement of true positives, and true positives would turn up in the auto-appealed relevant assessments. The development of this solution is also left for future work.

Finally, and perhaps most importantly of all, the double sampling method, with the above improvements, needs to be subject to a rigorous empirical evaluation. Some tentative results could be obtained by separate laboratory work, but a full evaluation of the double sampling approach requires deployment in the crucible of a live evaluation task, such as that of the Legal Track. We hope to have the opportunity to submit our methods to precisely this test.

Acknowledgements

This work was supported in part by the Australian Research Council and by DARPA contract HR-0011-06-2-0001. The authors thank Stephen Tomlinson, Maura R. Grossman, Gordon V. Cormack, and Jason R. Baron for their suggestions.

References

[1] J. Baron, D. Lewis, and D. Oard. TREC 2006 legal track overview. In *Proc. 15th Text REtrieval Conference*, pages 79–98, 2006.

[2] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10(6): 491–508, 2007.

[3] B. Carterette, V. Pavlu, E. Kanoulas, J. Aslam, and J. Allan. Evaluation over thousands of queries. In *Proc. 31st ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 651–658, Singapore, Singapore, 2008.

[4] C. Cuadra and R. Katter. The relevance of relevance assessment. In *Proc. American Documentation Institute*, volume 4, pages 95–99, 1967.

[5] B. Hedin, S. Tomlinson, J. Baron, and D. Oard. Overview of the TREC 2009 legal track. In *Proc. 18th Text REtrieval Conference*, pages 4:1–40, 2009.

[6] P. Kantor, M. Kim, U. Ibraev, and K. Atasoy. Estimating the number of relevant documents in enormous collections. In *Proc. 62nd Annual Meeting of the American Society for Information Science*, pages 507–514, 1999.

[7] C. Lam and D. Stork. Evaluating classifiers by means of test data with noisy labels. In *Proc. 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 2003.

[8] D. Lewis. The TREC-4 filtering track. In *Proc. 4th Text REtrieval Conference*, 1995.

[9] D. Oard, B. Hedin, S. Tomlinson, and J. Baron. Overview of the TREC 2008 legal track. In *Proc. 17th Text REtrieval Conference*, pages 3:1–45, 2008.

[10] E. Pronin. Perception and misperception of bias in human judgment. *Trends in cognitive sciences*, 11(1):37–43, 2007.

[11] G. Roda, J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. In *Proc. 10th CLEF workshop*, 2009.

[12] H. Roitblat, A. Kershaw, and P. Oot. Document categorization in legal electronic discovery: computer classification vs. manual review. *JASIST*, 61(1):70–80, 2010.

[13] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, 2007.

[14] A. Tenenbein. A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, 65(331):1350–1361, September 1970.

[15] E. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. In *Proc. 21st ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 315–323, Melbourne, Australia, 1998.

[16] E. Voorhees and D. Harman. *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[17] W. Wilbur. Human subjectivity and performance limits in document retrieval. *Information Processing & Management*, 32(5):515–527, 1996.

[18] E. Yilmaz and J. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. 15th ACM International Conference on Information and Knowledge Management*, pages 102–111, Arlington, Virginia, USA, 2006.

[19] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. 21st ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, 1998.