Assessing the Reliability and Reusability of an E-Discovery Privilege Test Collection

Jyothi K. Vinjumur
iSchool
University of Maryland
College Park, MD USA
jyothikv@umd.edu

Douglas W. Oard
iSchool and UMIACS
University of Maryland
College Park, MD USA
oard@umd.edu

Jiaul H. Paik
UMIACS
University of Maryland
College Park, MD USA
jiaul@umd.edu

ABSTRACT

In some jurisdictions, parties to a lawsuit can request documents from each other, but documents subject to a claim of privilege may be withheld. The TREC 2010 Legal Track developed what is presently the only public test collection for evaluating privilege classification. This paper examines the reliability and reusability of that collection. For reliability, the key question is the extent to which privilege judgments correctly reflect the opinion of the senior litigator whose judgment is authoritative. For reusability, the key question is the degree to which systems whose results contributed to creation of the test collection can be fairly compared with other systems that use those privilege judgments in the future. These correspond to measurement error and sampling error, respectively. The results indicate that measurement error is the larger problem.

Categories and Subject Descriptors

H.3.4 [**Information Storage & Retrieval**]: Systems & Software-performance evaluation

General Terms

Measurement, Performance, Experimentation.

Keywords

Evaluation, sampling, measurement error

1. INTRODUCTION

In civil litigation, the disclosure of documents that could have been withheld on the basis of attorney-client privilege or attorney work-product doctrine is an important concern for litigants. As a result, it is common for every document that is responsive to a counterparty request to be subjected to a manual review to determine whether a claim of privilege can be made. It is not uncommon for there to be tens of thousands of responsive documents, making the cost of this manual review for privilege quite high. Lawyers are naturally reluctant to trust automated techniques because failing to withhold even a small set of privileged documents might be consequential. Thus evaluation techniques that make it possible to compare alternative approaches to privilege classification are needed.

Evaluation of information retrieval systems relies on test collections in which relevance judgments can affordably be created for only a small portion of the collection [4]. One way of selecting these documents is to focus on documents found by the systems that are to be compared. This approach, known as pooling, has been widely used in the Text Retrieval Conference (TREC) and elsewhere. Two concerns arise. The first, which we call reliability, is that different assessors may reach different judgments for the same document. Voorhees has shown that absolute measures of effectiveness are sensitive to this effect but that relative comparisons between systems are relatively insensitive to inter-assessor disagreement [5].

A second concern, reusability, is that new systems will generally find some documents that did not contribute to the pool, and assuming such documents not to be relevant might adversely affect even relative comparisons. Reusability is important because reusable test collections allow the cost of relevance judgments to be amortized over future uses of a test collection. Reusability of pooled judgments was examined by Zobel [10], who found that TREC pooling had likely found no more than half of the relevant documents, but that relative comparisons remained reliable. Buckley et al. [1] later highlighted a key limitation of that conclusion, finding that when distinctive systems had contributed to the pool, removing one such system could yield a substantial adverse effect on measurements of mean average precision. One way to partially address this concern, introduced by Yilmaz and Aslam, is to sample the documents to be judged from the full collection and then to estimate the evaluation measure from the sampled judgments [8, 9].

Random samples drawn from very large collections yield confidence intervals that are so large as to be uninformative, so in this paper we focus on the sampling design used in the interactive task of the TREC Legal Track, in which set intersections were used as a basis for stratification [3]. Between 2006 and 2011, the TREC Legal Track created relevance judgments for more than 100 topics (which in e-discovery are called "production requests"). In 2010, this was augmented by the world's first (and to date only) shared-task evaluation of privilege classification [2]. In this paper, we study the reliability and reusability of the resulting priv-

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.
SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright 2014 ACM 978-1-4503-2257-7/14/07 ...$15.00.

ilege test collection. When working with Legal Track test collections we need to think a bit differently about reliability and reusability. For reliability, we are interested not just in relative comparisons, but also in the reliability of absolute measures of effectiveness, and most particularly in estimates of recall. Point estimates from samples are (in expectation) insensitive to sample size, so characterizing the reusability of stratified samples requires comparing confidence intervals for systems that did and didn't contribute to the stratification.[1] What we call reliability thus corresponds to the statistical concept of measurement error, reusability to the statistical concept of sampling error.

2. A PRIVILEGE TEST COLLECTION

The privilege task in the 2010 TREC Legal Track[2] requested "all documents or communications that are subject to a claim of attorney-client privilege, work-product, or any other applicable privilege or protection ..." Although privilege classification is normally performed as a second pass after classification for relevance, nothing in the definition of privilege is specific to any litigated matter. The collection to be searched was the EDRM Enron collection, version 2, which is a collection of Enron email messages for which text extracted from attachments is provided with the collection. Following the practice for privilege review in e-discovery, the items to be classified were "document families," in this case a family was defined as an email message together with all of its attachments.[3]

2.1 Stratified Sampling

Two teams (A and H)[4] submitted system results (runs) for the TREC 2010 privilege classification task: Team A submitted four runs ($a1$, $a2$, $a3$, $a4$); Team H submitted one ($h1$). Each run was a binary assignment of families to one of two classes: privileged or not privileged. Following TREC convention, we refer to these five runs as participating systems; each run was produced by people and machines working together (TREC refers to this as interactive task). The collection was partitioned into 32 strata, each defined by a unique 5-bit vector (e.g., 01010 for the stratum containing families runs $a1$, $a3$, and $h1$ classified as not privileged and runs $a2$ and $a4$ classified as privileged) [2]. The 00000 stratum included 398,233 of the 455,249 families (87% of the collection), but only 3,275 of the 6,766 samples (48%) were allocated to that stratum. The resulting sampling rate for the 00000 stratum (0.8%) was far sparser than for any other stratum (which averaged 6.1%). The allocation of samples was a bit denser for smaller strata since a 6.1% sampling rate might otherwise result in very few samples being drawn. Few samples were allocated to these very small strata in aggregate, so the sampling rate remained above 6% for every stratum other than the 00000 stratum.

2.2 Privilege Assessment

First-tier privilege assessors (henceforth, assessors), who were lawyers employed by a firm whose business included provision of document review services for e-discovery, were provided with detailed guidelines written by a senior attorney (the Topic Authority (TA)). Assessors recorded ternary judgments: privileged, not privileged, or unassessable (e.g., for display problems, foreign-language content, or length). As expected, assessors sometimes made judgments that disagreed with the TA's conception of privilege. For other tasks, differing judgments might be treated as equally valid, but in e-discovery the TA's judgments are authoritative (because the TA models the senior attorney who will certify that the review was performed correctly). Judgments that disagree with those of the TA are therefore considered incorrect. In TREC 2010, an assessor's judgment regarding whether a family should be classified as privileged could be escalated to the TA for adjudication in three ways. First, a team might appeal the decision of an assessor to the TA. A total of 237 such appeals were received. Of course, teams might not as easily notice, nor would they rationally appeal, assessor errors that tended to increase their estimated classification accuracy. In particular no team would rationally appeal an erroneous assessor judgment of privileged in the 11111 stratum, nor an assessment of not privileged in the 00000 stratum. The set of appealed judgments is thus biased [7]. To create an unbiased sample, 223 assessor judgments were thus independently drawn using simple random sampling. Since this is a random sample of a stratified sample, it results in a smaller stratified sample of the full collection. To facilitate symmetric comparisons among assessors, a second simple random sample containing 730 families was drawn and each family in that sample was duplicated in the set of families to be assessed. This was done in a manner that had been expected to result in the duplicated families being assigned to different assessors.[5] When conflicting assessments for a family were received, the judgment was adjudicated. Table 1 summarizes the selection process.

Table 1: TA adjudication rates.

Category	Assessed	Adjudicated	Rate
Random sample	6,766	223	3.3%
Team appeal	6,766	237	3.5%
Assessor disagreement	730	76	10.8%

2.3 Estimation

We focus separately on estimates of recall and precision. To estimate the precision of system S_i we need to estimate the number of privileged families that were correctly found by S_i. Let N^h_{pr} be the number of privileged documents in stratum h. Then

$$\hat{N}^i_{pr} = \sum_{h:S_i \in S^h} \hat{N}^h_{pr} \quad (1)$$

where S^h is the set of systems that retrieved documents in stratum h, $\hat{N}^h_{pr}$ is the unbiased estimator of N^h_{pr} and $\hat{N}^i_{pr}$ is the unbiased estimator of N^i_{pr}. Let the total number of families in stratum h be N_h and let the number of families drawn from stratum h as a simple random sample without replacement be n_h. Then an unbiased estimator of N^h_{pr} is

$$\hat{N}^h_{pr} = \frac{N_h \times n^h_{pr}}{n_h} \quad (2)$$

[1]Thanks to William Webber for pointing this out.
[2]For bookkeeping purposes, the (non-topical!) privilege task was "topic 304."
[3]Use of families is referred to as "message" evaluation in [2].
[4]In [2] Team A was called CB, team H was called IN.
[5]Some pairs may have been judged by the same assessor.

where n_{pr}^h is the number of privileged families in the sample of stratum h. In TREC 2010, the TA judgment was used for each adjudicated family; the assessor's judgment was used for all other sampled families. To estimate recall, an estimate of N_p (total number of privileged families in the collection), is also needed. An unbiased estimator of N_p is

$$\hat{N}_p = \sum_{h=1}^{m} \hat{N}_p^h \tag{3}$$

where N_p^h is the number of privileged documents in stratum h and m is the number of strata. From these estimates, we can derive point estimates for recall and precision. Point estimates alone are of little use, however, since small samples can yield estimates with large variance. We therefore also calculate two-tailed confidence intervals for recall and precision. A two-tailed confidence interval is a range of values $(Recall^{lower}, Recall^{upper})$ within which the recall of system S_i lies with $1 - \alpha$ confidence. Webber observes that the beta-binomial with hyper-parameters $\alpha = \beta = 0.5$ provides accurate estimates of confidence intervals [6]. We therefore compute distributions over the number of privileged families retrieved and not retrieved by S_i in each stratum to obtain the beta-binomial posterior. Once the posterior for each stratum is obtained, the yield N_{pr}^h for each stratum is simulated numerically. A large number of beta-binomial Monte Carlo simulations with 40,000 draws are performed, and the confidence intervals are obtained.

3. RESULTS

Here we analyze the reliability and reusability of the TREC 2010 Legal Track privilege task test collection.

3.1 Analysis of Measurement Error

The use of assessor judgments for families that the TA had not adjudicated would be reasonable if the appeal process had identified most of the assessor errors. This is a testable hypothesis. Although the TA might also make errors, we ignore that factor because we believe its effect to be small. We therefore treat the TA's judgments as a gold standard. As a further simplification, we treat the small handful of unassessable documents (13 families) as not privileged in our analysis. One way of visualizing the effect of assessor errors is to use only some or all of the families that were selected for adjudication, plotting confidence intervals using TA judgments in one case and using assessor judgments in the other. The adjudicated sample is less than 8% of the size of the full set of official judgments, so this yields fairly large confidence intervals, but the comparison does offer useful insights.

Figure 1 compares the (95%) confidence intervals on recall for each participating system using only the families that were selected for adjudication by the simple random sample; Figure 2 shows a similar comparison using all of the adjudicated families. From Figure 1 we can observe that judgments from assessors yield somewhat higher recall estimates than does the TA, but Figure 2 shows the opposite effect. The difference results from some combination of sampling error, appeals that disproportionately benefit participating systems, or systematic biases in the families on which assessors disagree. As the size of the error bars illustrates, we cannot reject sampling error as an explanation. Nonetheless, there is some evidence to support the hypothesis that appeals disproportionately benefit participating systems.

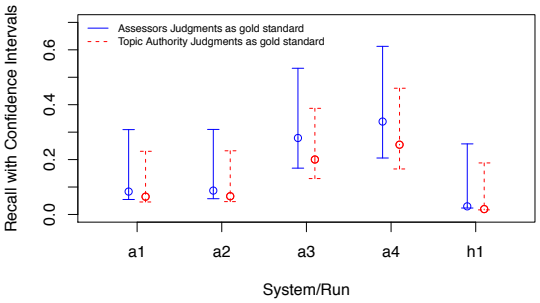

Figure 1: Recall, $a4$ ablated, random adjudication

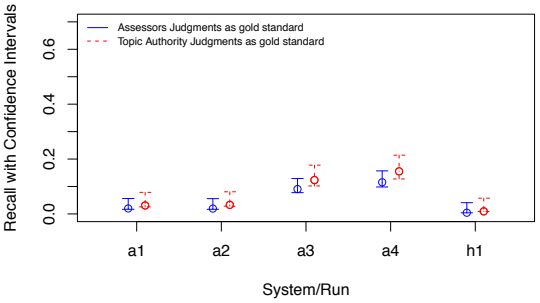

Figure 2: Recall, $a4$ ablated, all adjudication

Table 2 shows how the overturn rate varies with the reason for adjudication and with the original judgment. As the random sampling results show, assessors are more likely to mistakenly judge a family as privileged than as not privileged. Specifically, a z-ratio test for independent proportions finds the $1 \rightarrow 0$ overturn to be significantly more likely than a $0 \rightarrow 1$ overturn ($p < 0.05$). The same is not true for documents appealed by participating teams, however, where the overturn rates in each direction are statistically indistinguishable. Said another way, the increase in total overturn rate from 23% to 36% between randomly sampled adjudications and appealed adjudications (a 58% relative increase) can be largely explained by participating teams being no better than chance at recognizing an assessor's false positive judgments, but by being much better than chance at recognizing an assessor's false negative judgments.

The implications of this for the reliability of the test collection are clear: estimating absolute measures, and particularly absolute estimates of recall, using assessor judgments that exhibit systematic errors results in estimates that are open to question. If uncorrected assessor judgments were a small fraction of the total judgments, this would be a relatively minor concern, but uncorrected judgments are being used for about 92% of the sampled families. On the positive side, the availability of adjudicated random samples offers the potential for modeling differential error rates conditioned on the first-tier assessor's judgment. On the negative side, the inability to associate judgments with individual assessors in TREC 2010 means that such corrections can only be applied on an aggregate basis. We note, however, that relative comparisons between participating systems can still be informative, so long as assessor errors penalize all participating systems similarly.

Table 2: Overturn rates

	Assessor → Topic Authority	
Adjudication Basis	0 → 1	1 → 0
Random sample	31 of 161 (19%)	20 of 62 (32%)
Team appeal	32 of 77 (42%)	54 of 160 (34%)
Disagreement	28 of 49 (57%)	9 of 27 (33%)

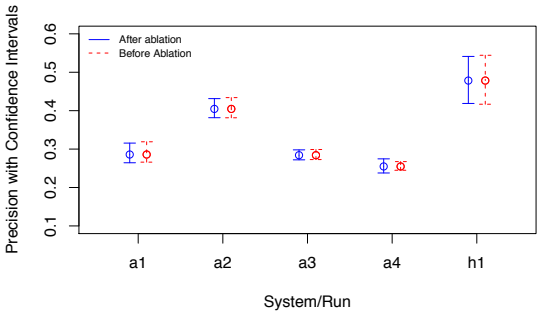

Figure 3: Precision, $a4$ ablated, all adjudication

3.2 Analysis of Sampling Error

To assess reusability, we need to assess the comparability of evaluation results for systems that did and did not contribute to the development of the test collection. A standard way of performing such analyses is through system ablation [10]: removing a system that in fact did participate in the stratification and then rescoring all systems, including the ablated system, and observing the effect on system comparisons. With pooling, ablation results in removing judgments for documents that were uniquely found by one system. With stratified sampling, by contrast, ablation results in re-stratification. For example, when system $a4$ (the participating system with the highest recall) is ablated, the 00000 stratum and the 00010 stratum become merged into a 000?0 stratum (where ? indicates a don't-care condition), the 11001 stratum gets merged with the 11011 stratum to form a 110?1 stratum, and similarly for each other stratum pair that is differentiated only by the ablated system. If we then reapply the process for deciding on the number of families to sample from each merged stratum, we will see little effect on the sampling rate for most strata. The one important exception is the 000?0 stratum (continuing with our example of ablating system $a4$), where we are merging large strata with quite different sampling rates (very small strata can also see substantial changes in their sampling rate, but their effect on the overall estimate will be small). We therefore model the effect of ablation by allocating all of the samples in each pair of strata to the corresponding merged stratum, adjusting the contributions of each sample to the estimate of the yield for the merged stratum to be equal.

To generalize, let a refer to the stratum in the pair including families classified as privileged by the ablated run, b to the corresponding stratum containing families classified as not privileged by the ablated run, and c to the merged stratum. We assume that the merged stratum would include the same number of samples that the two original strata contained separately; that is $n_c = n_a + n_b$ and the sampling rate for merged stratum c is $p_c = n_c/N_c$, where $N_c = N_a + N_b$.

We performed three ablation experiments, in each case ablating one system with high, medium or low recall and then recalculating point estimates and confidence intervals for every system. Comparing post-ablation to pre-ablation results, we see that point estimates are unchanged, as expected, but as Figure 3 shows confidence intervals for precision increase for the ablated system (system $a4$ in this figure). We attribute this to the reductions in the sampling rate for the 00010 stratum (from merging with the 00000 stratum, which results in documents in the former 00010 stratum being sampled at a far lower rate), since we expect families classified uniquely by any reasonable system as privileged to more often actually be privileged than families that no system classified as privileged. The same pattern is evident in our other two ablation studies (ablating systems $a2$ or $h1$; not shown). No similar effect was observed for confidence intervals on recall, however, perhaps because the estimates for the retrieved set contribute to both the numerator and the denominator of the recall computation.

4. CONCLUSION

We have explored set-based evaluation for privilege classification using stratified sampling, with strata defined by the overlapping classification results from different participating systems. We have characterized reliability by examining the impact of unmodeled assessor errors on evaluation results, and we have characterized reusability by showing that confidence intervals are affected when we reconstruct the test collection in a way that does not rely on the contributions of one participating system. We found that assessor errors do adversely affect absolute estimates of recall, and we have suggested future work on statistical correction for the effect of those errors. Confidence intervals for precision increased noticeably when we ablate a system, but no comparable effect was noted for recall. Since recall is the more important measure in e-discovery, this is a promising result.

5. ACKNOWLEDGMENTS

This work has been supported in part by NSF award 1065250. Opinions, findings, conclusions & recommendations are those of the authors and may not reflect NSF views.

6. REFERENCES

[1] C. Buckley et al. Bias and the limits of pooling for large collections. *Information Retrieval*, 10(6), 2007.
[2] G. Cormack et al. Overview of the TREC 2010 legal track. In *TREC*, 2010.
[3] D. Oard et al. Overview of the TREC 2008 legal track. In *TREC*, 2008.
[4] K. Spärck Jones et al. Information retrieval test collections. *Journal of Documentation*, 32(1), 1976.
[5] E. Voorhees. Variations in relevance judgments & the measurement of retrieval effectiveness. *IP&M*, 2000.
[6] W. Webber. Approximate recall confidence intervals. *Transactions on Information Systems*, 31(1), 2013.
[7] W. Webber et al. Assessor error in stratified evaluation. In *CIKM*, 2010.
[8] E. Yilmaz et al. Estimating average precision with incomplete and imperfect judgments. In *CIKM*, 2006.
[9] E. Yilmaz et al. A simple and efficient sampling method for estimating AP & NDCG. In *SIGIR*, 2008.
[10] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *SIGIR*, 1998.