Impact of Surrogate Assessments on High-Recall Retrieval

Adam Roegiest
University of Waterloo

Gordon V. Cormack
University of Waterloo

Charles L.A. Clarke
University of Waterloo

Maura R. Grossman*
Wachtell, Lipton, Rosen & Katz

ABSTRACT

We are concerned with the effect of using a surrogate assessor to train a passive (i.e., batch) supervised-learning method to rank documents for subsequent review, where the effectiveness of the ranking will be evaluated using a different assessor deemed to be authoritative. Previous studies suggest that surrogate assessments may be a reasonable proxy for authoritative assessments for this task. Nonetheless, concern persists in some application domains—such as electronic discovery—that errors in surrogate training assessments will be amplified by the learning method, materially degrading performance. We demonstrate, through a re-analysis of data used in previous studies, that, with passive supervised-learning methods, using surrogate assessments for training can substantially impair classifier performance, relative to using the same deemed-authoritative assessor for both training and assessment. In particular, using a single surrogate to replace the authoritative assessor for training often yields a ranking that must be traversed much lower to achieve the same level of recall as the ranking that would have resulted had the authoritative assessor been used for training. We also show that steps can be taken to mitigate, and sometimes overcome, the impact of surrogate assessments for training: relevance assessments may be diversified through the use of multiple surrogates; and, a more liberal view of relevance can be adopted by having the surrogate label borderline documents as relevant. By taking these steps, rankings derived from surrogate assessments can match, and sometimes exceed, the performance of the ranking that would have been achieved, had the authority been used for training. Finally, we show that our results still hold when the role of surrogate and authority are interchanged, indicating that the results may simply reflect differing conceptions of relevance between surrogate and authority, as opposed to the authority having special skill or knowledge lacked by the surrogate.

*The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
http://dx.doi.org/10.1145/2766462.2767754.

Categories and Subject Descriptors: H.3.4 Systems and Software Performance evaluation (efficiency and effectiveness).

Keywords: recall; assessor error; evaluation; eDiscovery; electronic discovery; relevance ranking; supervised learning.

1. INTRODUCTION

In high-recall information retrieval tasks—such as electronic discovery ("eDiscovery") in civil litigation [11], systematic review in evidence-based medicine [24], and preparation of test collections for information retrieval research [14]—supervised learning is often used to separate relevant from non-relevant documents [37]. In supervised learning, each of a pre-selected set of documents (the "training set") is labeled as relevant or not by a human assessor, and used to train a machine-learning algorithm, which then classifies or ranks the documents in a corpus (the "evaluation set") according to their likelihood of relevance. We focus here on a type of supervised learning, which, in eDiscovery is referred to as "simple passive learning" [12], to distinguish it from active learning, where the training set is selected incrementally, using feedback from the learning algorithm [38, 11].

To measure the effectiveness of a high-recall retrieval effort, it is necessary to estimate the number r of relevant documents from the evaluation set that are retrieved, as well as the number m that are missed. If r and m were known with certainty, it would be a simple matter to compute an effectiveness measure: For example, recall $= \frac{r}{r+m}$. However, the very notion of relevance is subjective [31, 32, 33], and necessarily relies on imperfect human judgement to determine m and r, and hence recall.

For many high-recall tasks, the opinion of a single subject-matter expert ("the authority") provides the ultimate determination of relevance. In the eDiscovery domain, the authority might be a senior lawyer representing the responding party; in the intellectual property domain, the authority might be a patent examiner; in the medical domain, the authority might be a senior researcher. In all cases, obtaining authoritative opinions for even a small set of documents may be impossible, or may incur unacceptable costs and delays.

Previous studies have investigated the impact of replacing the opinion of the authority with that of a surrogate assessor [41, 46, 28]. The results of those studies suggest that, at least under certain experimental conditions, a surrogate can reasonably replace the authority, thereby increasing the allure of using cheaper, more readily available surrogates as proxies for authorities. On the other hand, some practition-

ers claim that the use of authoritative assessors is of critical importance in training the machine-learning algorithms used for eDiscovery tasks. One commentator [19], writing in a trade magazine, expressed the concern that errors in relevance assessments will be amplified by a chosen learning method to an irrecoverable extent. Others, including one court [2], have referred to this training issue more generally as "garbage in, garbage out."

1.1 Overview of experiments

After a review of related work in Section 2, and a discussion of our general experimental methodology in Section 3, in Section 4, we describe our studies using the TREC-4 test collection and supplementary assessments described by Voorhees [41]. Using the official and supplementary sets of assessments, we trained a support vector machine ("SVM") on each set and ranked the entire corpus through the use of ten-fold cross validation. From these rankings, we determined the depth in the ranking required (i.e., the number of documents reviewed) to achieve a particular level of recall, when each of the three assessors was deemed to be the authority, while the others were treated as surrogates.

Extending these experiments, we explored whether the inclusion of training assessments that reflected greater diversity in the interpretation of relevance would improve ranking performance. We implemented this diversification strategy in two different ways. First, we created three new surrogate-assessment sets, each corresponding to a pair of the three original assessors, which we imagined as working together to judge the training set. We generated the merged set for a surrogate pair by randomly dividing the documents 50/50 between the two surrogates, with each determining relevance for their half. An SVM was then trained on each of these merged surrogates, and the results evaluated using cross-validation, treating the third assessor's opinion as authoritative.

As an alternate means of diversification, we took the union of each surrogate pair, such that a document was deemed relevant if either constituent had designated it as relevant. Again, an SVM was trained on each of these union surrogates and evaluated using cross-validation, treating the third assessor's opinion as authoritative.

In Section 5, we directly explore the impact of a more liberal interpretation of relevance on passive supervised learning. The University of Waterloo, in the course of participating in the TREC-6 adhoc task, created an independent set of relevance assessments that included a third relevance category —"iffy"—denoting documents which they believed to be of borderline relevance. The availability of three relevance categories (i.e., relevant, non-relevant, and iffy) allowed us to take two views of relevance: a "conservative" view, which considered only those documents actually labeled as "relevant" to be relevant; and a "liberal" view, which considered documents labeled as "iffy" to be relevant, in addition to documents actually labeled as "relevant."

We continue in Section 6 with an experiment investigating the applicability of liberal and diverse interpretations of relevance in the legal domain. The TREC 2009 Legal interactive task used initial assessments generated by volunteer law students and contract attorneys, that were subsequently adjudicated by senior lawyers ("topic authorities"). In addition, the University of Waterloo generated sets of assessments using a high-recall retrieval system. An SVM was trained using each of: the assessments generated by Waterloo; the initial TREC assessments; a combination of Waterloo and the initial TREC assessments; and the final assessments. All classifiers were evaluated with respect to the final assessments. After this, we conducted a brief follow-up experiment to test the hypothesis that adding judgments from a third assessor for documents not originally included in the training set could reduce recall depth.

In Section 7, we discuss our findings and the limitations of our results, and in Section 8, we offer our conclusions.

2. RELATED WORK

Voorhees observed that the retired professional analysts who assessed the TREC-4 adhoc task agreed on relevance, as measured by the Jaccard index, less than 50% of the time [41]. As a consequence, when one assessor's judgements were assumed to be correct and used to evaluate the other's, the other's judgements were found to have both recall and precision on the order of 65%, leading Voorhees to opine, "a practical upper bound on retrieval system performance is 65% precision at 65% recall since that is the level at which humans agree with one another." While Voorhees' primary measure, mean average precision ("MAP"), is a reasonable choice for ad hoc retrieval, it provides little insight into the performance of high-recall retrieval due to the focus on early precision present in the measure. We know of no study that has investigated the applicability of Voorhees' results to a measure suitable for evaluating high-recall retrieval.

Voorhees is not alone in noting that relevance judgements differ for different assessors, and for the same assessor at different times [34, 30, 16, 45, 20, 21], or that differences in assessors, while resulting in different estimates of effectiveness measures, have little impact on determining the relative effectiveness of retrieval methods [3, 26, 7, 9, 40].

Webber and Pickens [46], using the same relevance assessments as Voorhees, deemed Voorhees' "primary" assessor for each topic to be "authoritative." Webber and Pickens reported that, on average, using a non-authoritative assessor for training resulted in a 14% decrease in F_1 and a 24% increase in the number of top-ranked documents that must be retrieved to achieve a recall of 75%.

Cheng et al. [8] and Scholtes et al. [35] have similarly observed that non-authoritative training assessments have a significant but moderate negative impact on high-recall effectiveness. In contrast, Pickens [28] has suggested that non-authoritative training assessments may improve high-recall effectiveness when active (instead of passive) supervised-learning methods are used.

The issue of mitigating training and evaluation error is one of general interest in machine learning [47]. None of these studies specifically considered the interaction between assessor judgements for training and evaluation. A different source of interaction—inclusion bias introduced by the pooling method—is also a subject of current interest [6].

The TREC routing task [29, 44] bears substantial resemblance to high-recall retrieval, but differs by focusing on disjoint training and test collections, and evaluation using precision, rather than recall. Voorhees has suggested that examining the effect of differing relevance assessments on routing-like tasks is an important area of study [41]. To the best of our knowledge, this has yet to be investigated.

The effect of label noise (i.e., incorrect relevance assessments) is an ongoing topic of investigation within the IR

Corpus	Documents	Min. Prevalence	Avg. Prevalence	Max. Prevalence
TREC 4	713,049	0.002%	0.017%	0.028%
TREC 6	556,077	0.0002%	0.0097%	0.057%
Legal 2009	723,386	0.58%	0.72%	0.94%

Table 1: Summary statistics of all three corpora. The official NIST assessor was deemed as the gold standard for these statistics.

community [25, 13, 36], as well as in other communities [5, 17]. A comprehensive study by Frénay and Verleysen [17] outlined research on various facets of label noise (e.g., sources of label noise, its effects on classification, etc). Brodley and Friedl found that removing labels identified as incorrect, primarily by majority or consensus voting filters, improved overall classifier performance [5]. The e-mail spam filtering community has also noticed that label noise can drastically affect spam-filter performance [25, 36, 13].

3. EXPERIMENTAL METHODOLOGY

Our experiments follow the Cranfield paradigm [42], using test collections consisting of documents, topics, and relevance assessments from several TREC tracks. Table 1 offers summary statistics, including the number of documents and the average prevalence of topics, for each of the test collections used in this work; further details on each collection are provided in subsequent sections. In this section, we provide an overview of the general experimental methodology used in all of our experiments.

In these experiments, we used assessments generated by several independent assessors to train our classifiers and evaluate their performance. For the experiments described in Sections 4 and 5, we used assessments only for documents that were assessed by all of the assessors for the particular collection; any document for which there was not a complete set of assessments was treated as non-relevant. This choice was made to control for the fact that some assessors rendered (many) more assessments than others, and using the extra assessments would confound comparison. In addition, using the additional assessments would change the number of training examples and would result in testing a different hypothesis than the one in which we were interested (i.e., the effect of quantity versus quality of assessments).

In contrast, for the experiments reported in Section 6, we used assessments for all documents in the TREC judging pool, while maintaining the roles (initial assessor, topic authority, and independent assessor) established in the original experiment. In cases where there was no independent assessment for a document in the pool, we evaluated two different methods: (i) deeming the assessment to be "not relevant," and (ii) deeming the assessment to be the same as the initial assessment.

The TREC judging pools used for training form convenience samples of the full collections since they are composed of the top-ranked documents from participant submissions. To mitigate any effects from training on a narrowly selected set of documents (i.e., those that appeared relevant to some participant system), we augmented each training set with 1,000 randomly selected documents that were treated as non-relevant. This step broadened the representativeness of documents in each training set, to ensure that the resultant classifier was not focused exclusively on fine-grained distinctions between relevant and non-relevant documents in the judging pool, to the exclusion of non-relevant documents outside the pool.

To rank the documents, we used SVM$^{\text{light}}$[18], with default parameters. The features supplied were tf-idf term scores for all alphabetic words. Scores were generated after the Porter stemmer and case folding were applied.

Because our training and evaluation sets were not disjoint, we used ten-fold cross validation to approximate the effect of an independent evaluation set. Documents appearing in both sets were evenly distributed among 10 splits, as were documents appearing only in the evaluation set. The documents in each split were scored by a classifier whose training set was the union of the other nine splits, and a ranking was formed by sorting documents in the evaluation set according to score. We first tested our experimental methodology by replicating the Webber and Pickens study [46], successfully reproducing their results.

Our primary evaluation measure was *Recall Depth*, which is the size of the shortest prefix of the ranking that achieved a particular level of recall, expressed as percentage of the size of the corpus.

Our graphical results show, for each method, the average over all topics of (log transformed) recall depth, as a function of recall. In addition, for direct comparison, we show the average of (log transformed) relative recall depth—the ratio of recall depths between pairs of interest. Our tabular results show the same recall depth for 75% recall—a previously reported recall target [12]. We computed the significance of the surrogate-trained classifiers relative to the authority-trained classifier, applying a t-test to the log-transformed difference. In our tables, † denotes $p < 0.05$; ‡ denotes $p < 0.0001$.

4. INDEPENDENT JUDGMENTS

In this section, we describe our experiments using documents, topics, and relevance assessments from the TREC-4 adhoc task [22]. For this test collection, the official relevance assessments were augmented by two independent sets of relevance assessments rendered by different assessors within the course of Voorhees' experiments [41]. We labeled these assessment sets as J1, J2, and J3. While the assessments in J1 were (a subset of) those used for the official TREC evaluation, our experiments treated J1, J2, and J3 equally, treating each in turn as the "authority," and the others as surrogates. We restricted our experiments to topics where all three assessors found at least eight relevant documents, with the intent of reducing variance created by very low prevalence topics, consistent with previous work [46].

J1, J2, and J3 each reflect a single interpretation of relevance. To explore our hypothesis that a more diverse interpretation of relevance derived from several assessors would result in better training, we took each pair of surrogates and merged their assessments by randomly splitting the training set in half and assigning each half to one of the surrogates. The resulting merged-surrogate sets, which we denote J1|J2, J1|J3, and J2|J3, might then be viewed as the result of the two surrogate assessors working together to assess a single set of documents. Classifiers trained using each of the merged-surrogate sets were evaluated using J3, J2, and J1, respectively, as the authoritative assessor. Each merged-surrogate set contains the same documents as the single-surrogate set, reflecting the same amount of training effort.

To explore our hypothesis that a more liberal interpretation of relevance would result in better training, we evaluated training using the union of each pair of surrogates, denoted J1+J2, J1+J3, and J2+J3, in which a document was considered relevant if either of two surrogates considered it relevant. The classifiers constructed using these union surrogates were evaluated using J3, J2, and J1, respectively, as the authoritative assessor. Each of the union-surrogate sets contained the same documents as the single-surrogate and merged-surrogate sets, but reflected twice as much assessment effort. We do not believe that such a practice would necessarily be cost prohibitive, given the assumption that surrogate assessments are substantially less expensive than authoritative assessments.

4.1 Results

Figure 1 shows that the single-surrogate-trained classifiers are generally inferior to the corresponding authority-trained classifiers, requiring greater recall depth to achieve any particular level of recall. This result is reiterated in the relative recall depth plots in Figure 2, and the 75% recall depth values presented in Table 2. The differences among surrogates are most apparent at high levels of recall; as Figure 2 illustrates, some individual surrogates are substantially better than others.

Table 2 shows that with J1 and J3 as the authority, the authority-trained classifiers significantly outperform classifiers trained by individual surrogates. However, with J2 as the authority, the difference is not significant, particularly with respect to the case in which J1 is used as the surrogate assessor. While this reduced difference may be due to chance, it may also be an artifact of the assessment process. J1 corresponds to the official NIST assessments, for which the assessor reviewed the entire TREC-4 pool. This pool was much larger and had a lower prevalence of relevant documents than the pool reviewed by J2. An inverse relationship between prevalence and recall [39] might account for J1's assessments being more liberal than they would have been had J1 assessed only the documents that were assessed by J2.

Figures 1 and 2, as well as Table 2, show that the merged surrogates achieve effectiveness close to the better of the individual surrogates, occasionally exceeding both.

The union surrogates trained substantially and significantly ($p<0.01$) superior classifiers compared to the individual surrogates, as is evident in Figures 1 and 2, as well as Table 2.

Authority	Training		
	J1	J2	J3
J1	0.082% (0.058 - 0.115)	0.542%‡ (0.254 - 1.156)	0.284%‡ (0.139 - 0.584)
J2	0.103% (0.061 - 0.174)	0.087% (0.051 - 0.149)	0.161% (0.083 - 0.312)
J3	0.146%† (0.077 - 0.278)	0.359%‡ (0.162 - 0.797)	0.066% (0.044 - 0.101)
	J1|J2	J1|J3	J2|J3
J1	-	-	0.321%‡ (0.162 - 0.636)
J2	-	0.092% (0.059 - 0.143)	-
J3	0.182%† (0.096 - 0.346)	-	-
	J1+J2	J1+J3	J2+J3
J1	-	-	0.094%† (0.054 - 0.164)
J2	-	0.062%† (0.043 - 0.091)	-
J3	0.094% (0.054 - 0.164)	-	-

Table 2: 75% recall depth values for the TREC-4 experiments, with 95% confidence intervals. Significance is determined by comparing surrogate-trained classifiers to the authority-trained classifier. († denotes $p < 0.05$; ‡ denotes $p < 0.0001$.)

5. LIBERAL ASSESSMENT

In this section, we describe our experiments using documents, topics, and relevance assessments from the TREC-6 adhoc task [43], augmented by assessments rendered independently by the University of Waterloo in the course of their participation in TREC-6, using a process of interactive search and judging [10]. While TREC-6 used binary assessments, Waterloo used three categories of relevance: relevant, not relevant, and "iffy." This "iffy" label was used to identify documents for which the Waterloo assessors were unsure of the true relevance (i.e., they were of borderline relevance).

In Voorhees' study, these "iffy" assessments were treated as non-relevant. One of our hypotheses was that a more liberal interpretation of relevance would result in better classifier performance with respect to an independent third party (i.e., in this case, NIST). To this end, we compared the two classifiers trained by treating these "iffy" documents, alternatively, as non-relevant and as relevant. These different sets of assessments are labeled WaterlooRel and WaterlooRel+Iffy, respectively. In this experiment, WaterlooRel+Iffy represents a "liberal" assessor, while WaterlooRel represents a "conservative" assessor.

We used the NIST assessments to evaluate classifiers trained using each of: the NIST assessments, the WaterlooRel assessments, and the WaterlooRel+Iffy assessments. While our primary interest was in the relative effectiveness of using the liberal versus conservative Waterloo assessments for training, we also reversed the roles of surrogate and authority, as for our previous experiment. We did not investigate the use of the conservative Waterloo assessments as the surrogate and the liberal Waterloo assessments as the authority, or vice versa, as these sets of assessments were not independent.

5.1 Results

Our hypothesis—that surrogate assessors taking a more liberal view of relevance would produce better classifiers—is supported by the results presented in Figures 3a and 4a, where training using the liberal assessor is seen to achieve significantly better recall depth than both the conservative assessor and the NIST assessor. Table 3 shows the difference at 75% recall depth, with 95% confidence intervals. Across all recall levels, the liberally-trained classifier generally performs as well as, or better than, the authority, while the conservatively-trained classifier performs significantly worse.

Table 3, as well as panels (b) and (c) of Figures 3 and 4, show that, consistent with our previous results, classifiers trained using the NIST assessments fall short when evaluated using either the liberal or conservative Waterloo assessments as the authority. It is no surprise that the shortfall is greater with respect to the liberal assessments.

Authority \ Training	NIST	WaterlooRel	WaterlooRel+Iffy
NIST	0.110% (0.065 - 0.185)	0.261%† (0.142 - 0.481)	0.072% (0.049 - 0.105)
WaterlooRel	0.244% (0.130 - 0.458)	0.152% (0.094 - 0.246)	-
WaterlooRel+Iffy	0.882%‡ (0.515 - 1.511)	-	0.129% (0.094 - 0.177)

Table 3: 75% recall depth values for the TREC-6 experiments for Waterloo and NIST-trained classifiers, evaluated using NIST assessments, with 95% confidence intervals. Significance is shown relative to the NIST-trained classifier. († denotes $p < 0.05$; ‡ denotes $p < 0.0001$.)

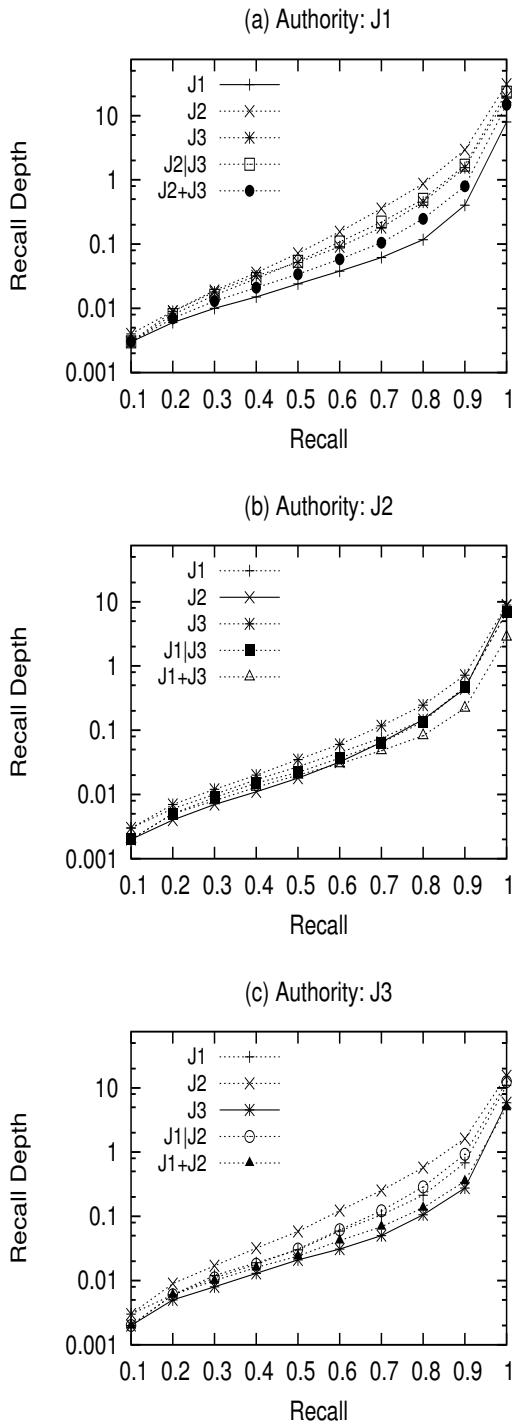


Figure 1: Recall depth plots for the TREC-4 experiments, using (a) JI, (b) J2, and (c) J3 as the authority.

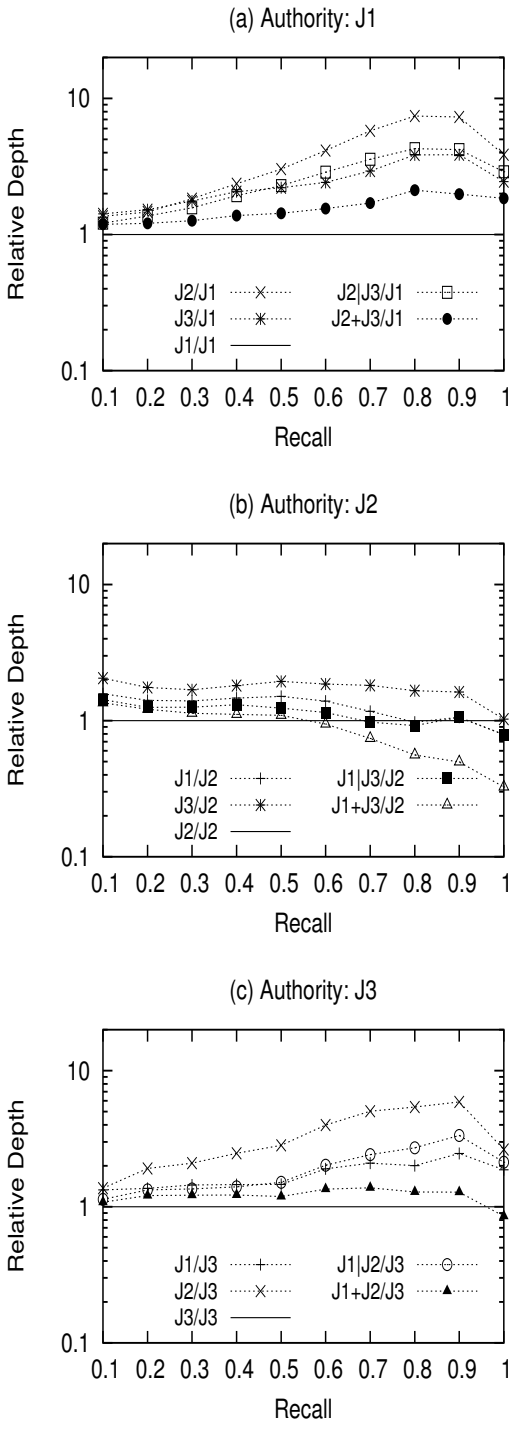
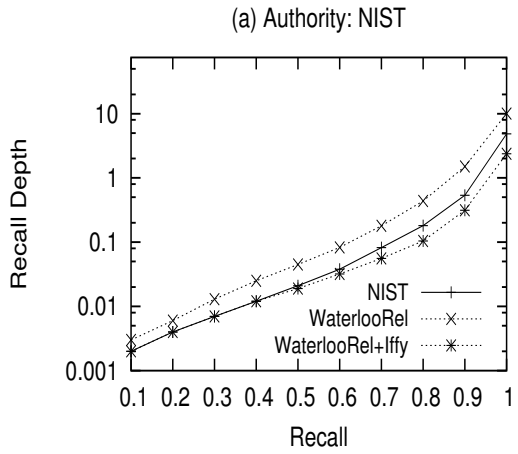
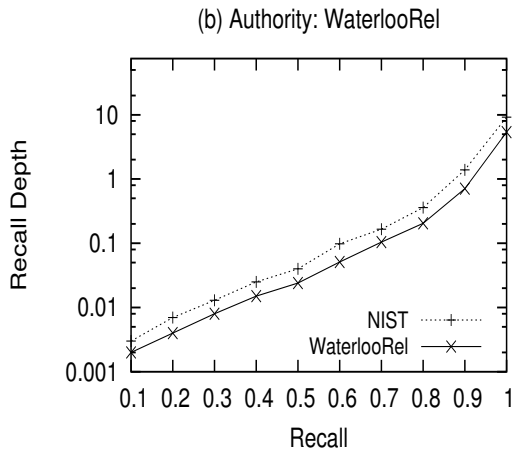


Figure 2: Relative recall depth plots for the TREC-4 experiments, using (a) J1, (b) J2, and (c) J3, as the authority.





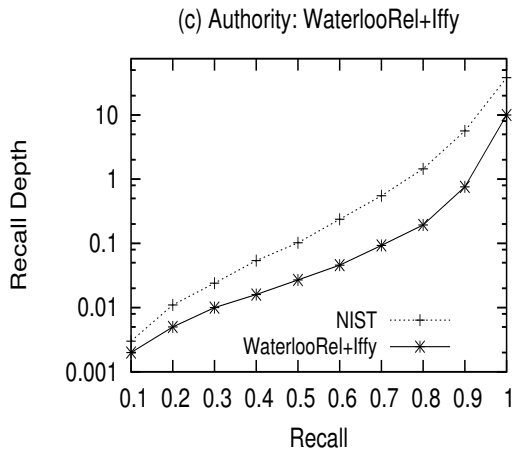
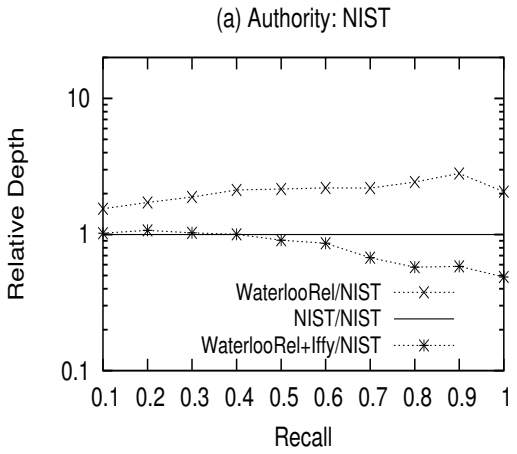
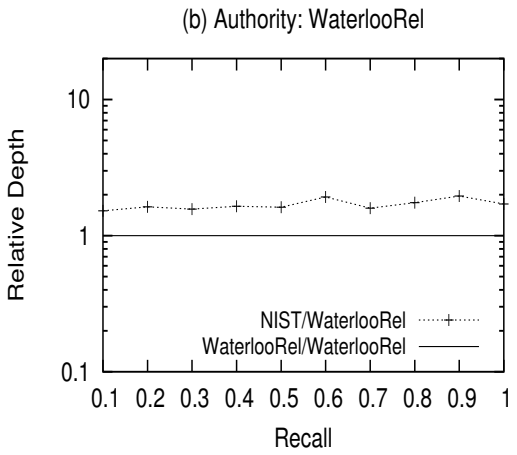


Figure 3: Recall depth plots for the TREC-6 experiments, using classifiers trained by each surrogate, and evaluated by each authority.





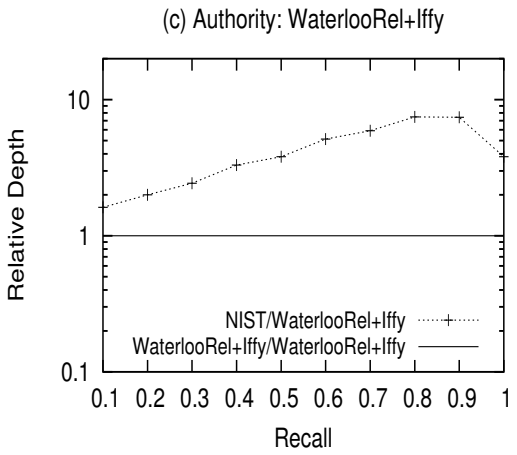


Figure 4: Relative recall depth plots for TREC-6 experiments, using classifiers trained by each surrogate, and evaluated by each authority.

Topic	Initial	Waterloo	Waterloo w/ Initial	Final
201	1.056%	0.214%	0.254%	0.215%
202	1.005%	0.977%	0.993%	0.936%
203	6.542%	0.955%	0.816%	0.456%
207	1.314%	1.401%	1.324%	1.236%

Table 4: 75% recall depth values for the TREC 2009 Legal experiments, using classifiers trained by Waterloo and initial assessments, and evaluated using final assessments.

	Judging Pool		Full Corpus	
Topic	Precision	Recall	Precision	Recall
201	0.40	0.70	0.05	0.76
202	0.93	0.93	0.27	0.80
203	0.47	0.25	0.13	0.25
207	0.98	0.88	0.89	0.79

Table 5: Recall and Precision of initial assessments in the TREC 2009 Legal Track judging pool versus the full corpus.

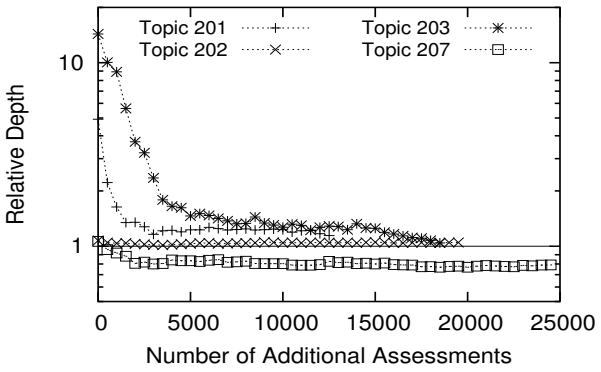


Figure 6: Per-topic 75% relative recall depth plots for the retrospective TREC 2009 Legal experiment, using classifiers trained on initial assessments, progressively augmented by Waterloo assessments, and evaluated using final assessments.

6. TREC 2009 LEGAL TRACK

The TREC 2009 Legal interactive task [23], simulated a high-recall eDiscovery task. For each topic in this task, the judging pool was a stratified sample of the document collection. An initial assessment of the judging pool was rendered using volunteer law students or contract attorneys. The initial assessments were provided to participating teams, who were invited to appeal assessments with which they disagreed. The appealed assessments were adjudicated by a TREC-designated *Topic Authority*, a senior lawyer who rendered the final, authoritative relevance assessments that were used to evaluate submissions. During the course of their participation in TREC 2009, the University of Waterloo developed an independent set of assessments using their own interactive, high-recall retrieval system for four of the task topics (Topics 201, 202, 203, and 207) [15].

Because the TREC judging pool included a large random sample of the document population, only a relatively small fraction (17.7%) of the documents in the pool were included in the Waterloo assessments; the rest were excluded by Waterloo's search method, as unlikely to be relevant. We investigated two approaches to determine the relevance of these documents for training purposes. The "Waterloo" surrogate assessments deemed the excluded documents to be "not relevant" for the purpose of training, and used the final TREC assessments as the sole authority for evaluation. The "Waterloo w/Initial" surrogate assessments used the initial NIST assessment for each excluded document. Thus, the "Waterloo" assessments were fully independent of the initial assessments, whereas the "Waterloo w/Initial" assessments, while not fully independent, might better model the situation in which the excluded documents had been manually assessed.

We evaluated the effect of training using the two sets of Waterloo surrogate assessments, as well as the initial and final assessments, using the final assessments as authoritative.

6.1 Results

Figure 5 shows relative recall depth plots, with respect to the final assessments, for each of the four topics. Table 4 shows 75% recall depth values for the same four topics. Overall, the surrogate-trained results appear inferior for high recall, but substantially so only when using the initial assessments as surrogate, and only for Topics 202 and 207. We note that Topics 202 and 207 have much higher prevalence than Topics 201 and 203, and due to the stratified sampling used to select the judging pool, the initial assessments achieved much higher precision and recall within the pool than in the collection at large, as illustrated in Table 5.

Figure 5 and Table 4 further indicate that the combination of assessments generally yields results as good as, and often superior to, the better of the individual surrogates.

6.2 Interactive Training

Retrospectively, we conducted one final supplemental experiment in an effort to shed some light on the applicability of our results to a more interactive, high-recall retrieval effort. Our supplemental experiment tracked improvement in relative recall depth as the initial assessments were supplemented incrementally with batches of 500 Waterloo assessments.

Figure 6 shows 75% relative recall depth, as a function of the number of Waterloo assessments. For Topics 201, 203, and 207, we see a dramatic gain from supplementing the assessments with a small fraction of the Waterloo assessments. For Topic 202, we see little improvement over the near-perfect initial assessments. The result suggests that having an independent assessor judge a fairly small fraction of the documents can result in a dramatic improvement in effectiveness, but further study is needed.

7. DISCUSSION

Our results show that it matters who assesses relevance; in particular, it matters whether the assessors whose judgements are used to train the system are the same as those whose judgements are used to evaluate the result. A statistic like "75% recall" conveys little meaning without considering, "according to whom?"

In one of the first cases where a court ruled in favor of the responding party's use of machine learning for eDiscovery, over the requesting party's objection, the responding party's brief [1] asserts:

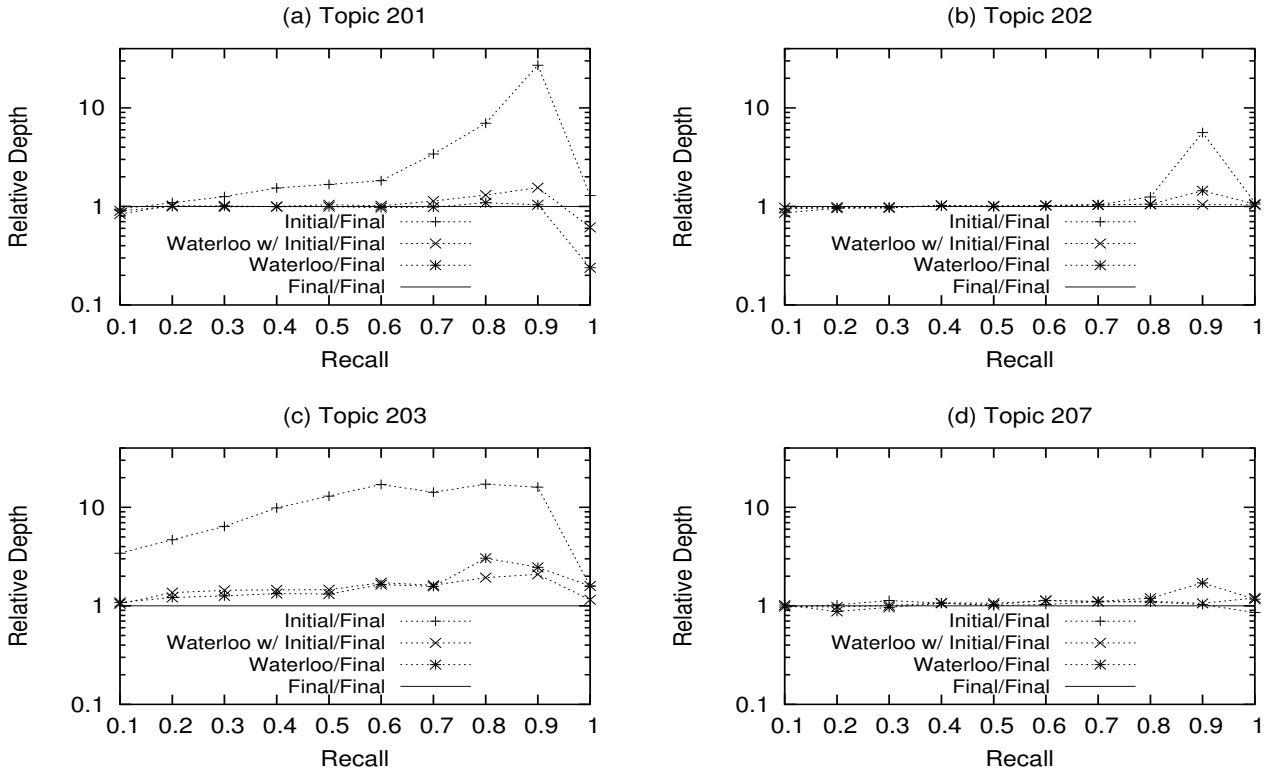


Figure 5: Relative recall depth plots for the TREC 2009 Legal experiments, using classifiers trained by each surrogate, and evaluated by the final assessments.

> Given that recall for linear review averages only 59.3%, [the responding party] proposes an acceptable recall criterion of 75%. In other words, predictive coding will conclude once the sampling program establishes that at least 75% of the relevant documents have been retrieved from the [responding party's electronically stored information] and are available to [them] for discovery purposes.

The 59.3% recall average was derived from Grossman and Cormack's analysis of the TREC 2009 Legal Track results [20], calculating the precision and recall of the initial assessments, evaluated with respect to the final (i.e., independent) authoritative assessments. The acceptance criterion of 75% recall, however, was established using the responding party's own reviewers, who were, we presume, also involved in training the system. Our results suggest that, had recall been evaluated using an independent assessor, the calculated recall value might have been considerably lower.

The designation of a particular assessor to be "authoritative" is, in many ways, an artificial construct designed to sidestep well-known uncertainties in the definition of relevance, and hence recall (see Webber et al. [45]). While some assessors may be more knowledgeable or skillfull than others, it is well known that even expert assessors will disagree on a substantial number of assessments [3]. The IR literature suggests than an exhaustive assessment effort by one such expert would be unlikely to achieve more than 65% recall and 65% precision in the eyes of another, equally skilled and knowledgeable expert [41]. The surrogates used for our TREC-4 and TREC-6 experiments achieved comparable recall and precision levels: When assessed by J1, J2 achieved recall of 52.9% and precision of 80.8%, and J3 achieved recall of 63.1% and precision of 78.1%; when assessed by NIST, the conservative Waterloo surrogate achieved recall of 62.8% and precision of 65.2%, while the liberal Waterloo surrogate achieved recall of 86.6% and precision of 50.0%.

Our results do not support the proposition that the use of machine learning "amplifies" inconsistencies between the surrogate and authority. The classifier trained using a surrogate's assessments achieves higher recall—at a recall depth corresponding to a fraction of the corpus—than the surrogate would achieve by assessing the entire corpus. For example, Figure 1 shows that a classifier trained using J2's assessments achieves 55% recall at a recall depth corresponding to 0.1% of the corpus, while a classifier trained using J2's assessments achieves 75% recall at a recall depth corresponding to 0.542% of the corpus.

Notwithstanding the discussion above, it is a worthwhile objective to try to maximize recall with respect to an authoritative assessor, either because that assessor has been stipulated to be the purveyor of true relevance, or because that assessor acts as a proxy for an as-yet-unavailable authority, such as a judge or regulator. Presumably, if a system achieves high recall at low recall depth with respect to one reasonable independent assessor, it is likely to achieve similar results with respect to another.

The results from our three experiments indicate that using a surrogate assessor instead of the authoritative assessor for

training can dramatically increase recall depth. While the effect is not universally large, the instances in which it is, cannot be attributed to chance ($p<0.05$, corrected for multiple hypothesis testing). On the other hand, there are several instances where surrogate training appears to be as good as, or better than, authoritative training. The same general effect is observed, regardless of which assessor is deemed to be authoritative.

Our TREC-4 experiments suggest that randomly intermingling the assessments of two surrogates achieves recall depth similar to that of the better surrogate, while using the union of the two surrogates' assessments improves on both, approaching the effectiveness of using the authority's assessments.

Our TREC-6 experiments show that, when the surrogate assessor makes the deliberate choice to label marginally relevant documents as relevant, recall depth is substantially and significantly reduced, relative to the case in which such documents are labeled as non-relevant. Furthermore, training using this liberal assessment strategy yields a materially lower 75% recall depth than authoritative training. While caution must be exercised in extrapolating this result to other assessors' efforts, it strongly suggests that a surrogate can, at least in this instance, train a classifier as effectively as the authority, even when the authority's assessments are deemed to be the gold standard.

Our TREC 2009 Legal experiments show the same pattern as the others: For two topics, training using the initial assessor yields substantially inferior results to training using the authority; for the other two topics, the difference was small and not significant. This apparent dissonance might be explained by the fact that, for the latter two topics, the recall and precision of the surrogate—with respect to the judging pool—were exceptionally high. For the former two topics, they were substantially inferior.

7.1 Limitations

Our experiments study only the case of simple passive learning, where a fixed training set is to used to train a learning method to rank the entire corpus, and the top-ranked documents are reviewed until high recall is achieved. Although this practice appears to be widely employed in eDiscovery today, the state of the art is perhaps better represented by interactive, active-learning approaches [12, 27]. Accordingly, our results are applicable only to the former method; their utility in guiding individual stages of an interactive or active approach has not been established.

Our "convenience sample" of available assessments and collections may not be representative of a typical application of passive supervised machine learning. In our experiments, the judging pool was far from a random sample of the collection, as evidenced in Table 5. Nor was it the result of uncertainty sampling as commonly used in active learning. The judging pools for the TREC-4 and TREC-6 experiments might be construed to be representative of relevance feedback, because the documents were those ranked highly by TREC submissions. The judging pool for the TREC 2009 Legal experiments might be construed to represent query-by-committee, as it was constructed using strata to illuminate disagreements among the TREC submissions.

While our findings indicate quite strongly that some combinations of surrogates and authorities fare poorly, while others fare very well, more research is needed—with a larger population of assessors—to gain a thorough understanding of the causal factors. That said, our results clearly suggest that, when other factors are held constant, increasing the diversity or liberality of training assessments increases the quality of ranking, relative to that produced by a single surrogate.

8. CONCLUSIONS

Situations where assessments from a single authoritative assessor are unavailable, expensive, or limited, may occasion the use of a surrogate assessor to train a passive supervised-learning method to rank documents. In many situations, the resulting ranking is significantly and substantially inferior to that which would have occurred, had the authoritative assessor been used for training. Our experiments indicate that this effect can be mitigated, and sometimes overcome, by merging the assessments of multiple surrogates, or by instructing the surrogate to use a more liberal interpretation of relevance.

We question whether it is possible to sweep away uncertainties in relevance determination simply by arbitrarily deeming relevance to be the judgment of a single authoritative assessor. It is well known that informed, qualified assessors disagree, and even the same assessor will disagree with him or herself, at different times and in different circumstances. We wonder whether it is useful to expend heroic efforts to anticipate the judgments of one particular assessor, and posit, instead, that it might be better to target a hypothetical "reasonable authority," selected from a pool of equally competent choices. In any event, it is important when evaluating the recall of a retrieval effort, to ask, "according to whom?" 75% recall measured through independent assessment is a formidable achievement, but the same 75% recall measured through self-assessment is unremarkable.

9. REFERENCES

[1] Memorandum in Support of Motion for Protective Order Approving the Use of Predictive Coding, *Global Aerospace v. Landow Aviation*, No. CL 61040, 2012 WL 1419842 (Va. Cir. Ct., Loudoun Cnty., Apr. 9, 2012).

[2] *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182 (S.D.N.Y., 2012).

[3] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter? In *Proc. SIGIR*, 2008.

[4] T. Barnett, S. Godjevac, J.-M. Renders, C. Privault, J. Schneider, and R. Wickstrom. Machine learning classification for document review. In *ICAIL DESI III Workshop*, 2009.

[5] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *J. of A.I. Research*, 11, 2011.

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

[7] R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing & Management*, 28(5), 1992.

[8] J. Cheng, A. Jones, C. Privault, and J.-M. Renders. Soft labeling for multi-pass document review. In *ICAIL DESI V Workshop*, 2013.

[9] C. W. Cleverdon. *The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages.* Cranfield Library, 3, 1970.

[10] G. V. Cormack, C. L. A. Clarke, C. R. Palmer, and S. S. L. To. passage-based refinement (MultiText experiments for TREC-6). In *Proc. TREC-6*, 1997.

[11] G. V. Cormack and M. R. Grossman. The Grossman-Cormack glossary of technology-assisted review with foreword by John M. Facciola, Magistrate Judge. *Fed. Courts Law Rev.*, 7(1), 2013.

[12] G. V. Cormack and M. R. Grossman. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proc. SIGIR*, 2014.

[13] G. V. Cormack and A. Kolcz. Spam filter evaluation with imprecise ground truth. In *Proc. SIGIR*, 2009.

[14] G. V. Cormack and T. R. Lynam. Spam corpus creation for TREC. In *Proc. 2nd CEAS*, 2005.

[15] G. V. Cormack and M. Mojdeh. Machine learning for information retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks. In *Proc. TREC-18*, 2009.

[16] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 14(5), 2011.

[17] B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE Trans. Neural Networks and Learning Systems*, 25(5), 2013.

[18] T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

[19] D. Gonsowski. A look into the e-discovery crystal ball. *Inside Counsel*, Dec. 2, 2011.

[20] M. R. Grossman and G. V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Rich. J. Law & Tech.*, 17, 2011.

[21] M. R. Grossman and G. V. Cormack. Inconsistent responsiveness determination in document review: Difference of opinion or human error? *Pace Law Rev.*, 32, 2012.

[22] D. Harman. Overview of the Fourth Text REtrieval Conference (TREC-4). In *Proc. TREC-4*, 1995.

[23] B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard. Overview of the TREC 2009 Legal Track. In *Proc. TREC-18*, 2009.

[24] J. P. Higgins, S. Green, eds. *Cochrane Handbook for Systematic Reviews of Interventions*, Wiley Online Library, 2008.

[25] A. Kolcz and G. V. Cormack. Genre-based decomposition of email class noise. In *Proc. KDD*, 2009.

[26] M. E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4(4), 1968.

[27] C. Li, Y. Wang, P. Resnick, and Q. Mei. ReQ-ReC: High Recall Retrieval with Query Pooling and Interactive Classification. In *Proc. SIGIR*, 2014.

[28] J. Pickens. In TAR, wrong decisions can lead to the right documents (a response to Ralph Losey). http://web.archive.org/save/http://www.catalystsecure.com/blog/2014/02/in-tar-wrong-decisions-can-lead-to-the-right-documents-a-response-to-ralph-losey/.

[29] S. Robertson and I. Soboroff. The TREC 2002 Filtering Track report. In *Proc. TREC-11*, 2002.

[30] H. L. Roitblat, A. Kershaw, and P. Oot. Document categorization in legal electronic discovery: Computer classification vs. manual review. *J. Am. Soc. for Info. Sci. and Tech.*, 61(1), 2010.

[31] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. *J. Am. Soc. Info. Sci.*, 26(6), 1975.

[32] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *J. Am. Soc. for Info. Sci. and Tech.*, 58(13), 2007.

[33] 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. *J. Am. Soc. for Info. Sci. and Tech.*, 58(13), 2007.

[34] L. Schamber. Relevance and information behavior. *Annual Rev. Info. Sci. and Tech. (ARIST)*, 29, 1994.

[35] J. C. Scholtes, T. van Cann, and M. Mack. The impact of incorrect training sets and rolling collection on technology assisted review. In *ICAIL DESI V Workshop*, 2013.

[36] D. Sculley and G. V. Cormack. Filtering email spam in the presence of noisy user feedback. In *Proc. CEAS*, 2008.

[37] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1), 2002.

[38] B. Settles. *Active learning literature survey.* TR 1648, University of Wisconsin, Madison, 2010.

[39] M. D. Smucker and C. P. Jethani. Human performance and retrieval precision revisited. In *Proc. SIGIR*, 2010.

[40] A. Trotman and D. Jenkinson. IR evaluation using multiple assessors per topic. In *Proc. ADCS*, 2007.

[41] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5), 2000.

[42] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*. Springer, 2002.

[43] E. M. Voorhees and D. Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). In *Proc. TREC-6*, 1997.

[44] E. M. Voorhees and D. K. Harman, eds. *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, 2005.

[45] W. Webber, D. W. Oard, F. Scholer, and B. Hedin. Assessor error in stratified evaluation. In *Proc. CIKM*, 2010.

[46] W. Webber and J. Pickens. Assessor disagreement and text classifier accuracy. In *Proc. SIGIR*, 2013.

[47] X. Zhu and X. Wu. Class noise vs. attribute noise: A quantitative study. *A. I. Rev.*, 22(3), 2004.