Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me

Gordon V. Cormack
University of Waterloo
gvcormac@uwaterloo.ca

Maura R. Grossman
University of Waterloo
maura.grossman@uwaterloo.ca

ABSTRACT

Technology-assisted review ("TAR") systems seek to achieve "total recall"; that is, to approach, as nearly as possible, the ideal of 100% recall and 100% precision, while minimizing human review effort. The literature reports that TAR methods using relevance feedback can achieve considerably greater than the 65% recall and 65% precision reported by Voorhees as the "practical upper bound on retrieval performance . . . since that is the level at which humans agree with one another" (*Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 2000). This work argues that in order to build—as well as to, evaluate—TAR systems that approach 100% recall and 100% precision, it is necessary to model human assessment, not as absolute ground truth, but as an indirect indicator of the amorphous property known as "relevance." The choice of model impacts both the evaluation of system effectiveness, as well as the simulation of relevance feedback. Models are presented that better fit available data than the infallible ground-truth model. These models suggest ways to improve TAR-system effectiveness so that hybrid human-computer systems can improve on both the accuracy and efficiency of human review alone. This hypothesis is tested by simulating TAR using two datasets: the TREC 4 AdHoc collection, and a dataset consisting of 401,960 email messages that were manually reviewed and classified by a single individual, Roger, in his official capacity as Senior State Records Archivist. The results using the TREC 4 data show that TAR achieves higher recall and higher precision than the assessments by either of two independent NIST assessors, and blind adjudication of the email dataset, conducted by Roger, more than two years after his original review, shows that he could have achieved the same recall and better precision, while reviewing substantially fewer than 401,960 emails, had he employed TAR in place of exhaustive manual review.

1 INTRODUCTION

This study contributes to the body of empirical evidence showing that hybrid human-computer classification systems (known in the legal community as "technology-assisted review" or "TAR") can be more effective and more efficient than exhaustive manual review by experts, where effectiveness is measured with respect to an independent gold standard. The results amplify and extend our 2011 work, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review* [14] in the following ways:

- We rigorously specify and evaluate a semi-automated process for human-in-the-loop classification in which the only human input is an initial query, followed by assessment of the documents selected for review by the system, until the system determines that high recall and precision have been achieved, and that the review process is complete;
- We extend the process with a quality-control ("QC") mechanism, in which the system suggests a subset of the documents for further adjudication, either by the user or another assessor, to mitigate the fallibility of the user's original assessments;
- We present a theory of information retrieval ("IR") system evaluation that extends the Cranfield method [30] to define the end-to-end effectiveness of an interactive IR process, and to model and control for dependencies between the assessments rendered by the human in the loop, and the assessments used to evaluate the result;
- Using the TREC 4 AdHoc collection and the alternate assessments used by Voorhees in *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness* [29], we provide evidence that our proposed TAR method achieves substantially better recall and precision than the the alternate NIST assessors would have achieved, had they reviewed the entire collection, with a small fraction of the effort;
- And finally, using a complete categorization of 401,960 email messages from the administration of Virginia Governor Tim Kaine, which was previously manually reviewed by Senior State Records Archivist Roger Christman ("Roger"), we show, using subsequent assessments rendered by Roger, that Roger could have achieved the same recall and higher precision, for a fraction of the effort, had he employed our TAR method to review the 401,960 email messages.

The following sections develop the theory of how to measure the end-to-end effectiveness of high-recall and high-precision IR efforts, how to simulate a human in the loop, our experimental design, and our results on the TREC 4 and Kaine email datasets.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

ACM ISBN 978-1-4503-5022-8/17/08.
https://doi.org/10.1145/3077136.3080812

2 THEORY

It is well understood that the notion of "relevance" is imprecise, and that different assessors—or even the same assessor at different times—may provide inconsistent relevance determinations for the same document, regardless of their knowledge and expertise, or the specificity with which "relevance" is defined. Nevertheless, it has been observed that relevance determinations by different assessors, while different, are essentially interchangeable as ground truth for the purposes of measuring the relative effectiveness of ad-hoc retrieval systems [3, 29]. In this work, we consider the problem of measuring the end-to-end effectiveness of "total-recall" methods, where the goal is to find substantially all relevant documents, and where the overall accuracy in determining relevance rivals that of the user, or any individual assessor. The model for a user is a "dedicated searcher, not a novice searcher," who is "willing to look at many documents" in order to find as much relevant information as possible (from TREC-1 [16]). This objective is shared by many critical applications, including electronic discovery in civil litigation, archiving of business or government records, patent search, and systematic review in evidence-based medicine. In 2015 and 2016, the TREC Total Recall Track [10, 15] addressed the total-recall problem, providing to participants a "Baseline Model Implementation" ("BMI"),[1] simulating a TAR method known as "continuous active learning" ("CAL") (*cf.* [9, 11]).

The ideal result of a "total-recall" IR effort is to identify *all* and *only* the relevant documents in a collection; that is, to achieve 100% recall and 100% precision. In practice, the ability to reach this goal is limited by the fallibility of human relevance assessment. Even if it were feasible to assess every document in the collection, a certain number of the resulting assessments would be incorrect, yielding less than 100% recall and less than 100% precision. Relevance assessments generated by a learned classifier would also be fallible, likewise falling short of 100% recall and precision. This article addresses the question: Can hybrid human-computer assessments yield higher recall and precision—with less effort—than human assessments alone?

To answer this question, it is necessary to estimate recall and precision, or another measure of how nearly all and only the relevant documents have been identified. The traditional Cranfield method for IR evaluation [30] offers limited insight because it relies on comparison with a "gold standard" for relevance, which itself relies on fallible assessments. At high levels of recall and precision, the Cranfield method tends to measure the ability of the method under test to reproduce the flaws in the gold standard, which—if the flaws are random—is impossible, and if the flaws are systematic—is possible only for methods with similar flaws.

Measuring the effectiveness of total recall is further complicated by the fact that most high-recall methods involve a human in the loop, and are influenced by that user's fallible assessments. In the simplest "ranked-retrieval" scenario, the system orders all documents in the collection by their likelihood of relevance, and the user examines them in order, until a sufficient number of relevant documents have been identified. In the "relevance-feedback" scenario, the user's assessment is communicated to the system, which uses this information to revise the ranking of the yet-to-be-examined documents. The "active-learning" or "uncertainty-sampling" scenario departs from relevance-feedback scenario in that the documents presented to the user are in the order most useful for machine learning, as opposed to likelihood of relevance, with the effect that the user is typically directed to the most marginally relevant documents to examine.

Regardless of the scenario, it is important to define precisely the circumstance under which a document is considered to be "identified" by the method. In the "*system-recall*" scenario, a document is deemed to be identified when it is presented to the user, regardless of the user's ultimate relevance assessment. In the "*end-to-end-recall*" scenario, a document is deemed to be identified only when it is presented to the user *and* the user judges it to be relevant. Where the user is fallible, system recall will generally be higher than end-to-end recall, while system precision will be lower. Which scenario is more apt depends on whether the role of the user is simply to provide guidance to the system, or to make the ultimate determination of whether a document is relevant or not.

Quality-control ("QC") procedures seek to mitigate the impact of fallible relevance assessments, using one or more supplemental assessments for some or all of the documents. In perhaps the easiest case, a second assessment might be rendered for each document. Where the assessments agree, it would be reasonable to assume that they are likely (but not certainly) both correct; where the assessments disagree, one is likely correct and the other is likely not—but how do we know which is which? One might defer to the second assessment, if it could be ascertained that it was more likely to be correct than the first, perhaps due to the application of additional care, or greater knowledge and skill on the part of the second assessor. One might defer to the "relevant" assessment if high recall were particularly important, and to the "not relevant" assessment if high precision were important. Alternatively, one might defer to a third assessment, effectively deeming the "majority vote" to be correct.

Majority-vote QC incurs overhead of $(1 + d)N$ additional assessments, where N is the number of documents in the collection, and d is the rate of discord between the first and second assessments. This overhead may be reduced by selecting a subset of $n \ll N$ documents for supplemental assessment. If the subset is a statistical sample, it is possible to quantify, but not to substantially mitigate, the fallibility of the first assessment. If a subset can be identified that includes many of the documents with discordant assessments, deferring to a third assessment for those particular documents can provide mitigation approaching that of majority vote, with considerably lower overhead.

This work distinguishes between *total-recall methods* and *search tools*. At the outset, TREC sought to measure the ease

[1] http://cormack.uwaterloo.ca/trecvm/.

with which *search tools* might be used within the context of a total-recall effort [16]:

> It should be assumed that the users need the ability to do both high precision and high recall searches, and are willing to look at many documents and repeatedly modify queries in order to get high recall. Obviously they would like a system that makes this as easy as possible.

To this end, relevance-based measures of search-tool effectiveness—notably, rank-based measures such as (mean) average precision ("(M)AP"), precision at a fixed cutoff ("P@k"), and R-precision ("P@R," where R is the number of relevant documents in the collection)—were used as proxies for ease of use [18]. With certain exceptions, primarily in the legal, intellectual property, and medical domains, interest within TREC and the IR community has generally shifted to more user-centric contexts, where the goal is to satisfy an ephemeral and user-specific information need, and search-tool effectiveness is quantified by proxy measures for user satisfaction, *cf.* [1].

Regardless of the context or proxy measure, evaluation efforts like those characterized by Voorhees [29, 30] have focused on search-tool (*i.e.*, "system") effectiveness, not the overall effectiveness of the user at using the tool to identify as nearly as practicable *all* and *only* the relevant documents, where "relevance" is defined by extrinsic criteria (*i.e.*, "end-to-end" effectiveness). The method of repeated ad-hoc search envisioned by TREC is commonly used, but is far from the only—or necessarily the most effective—total-recall method. In some domains, such as the curation of government archival records, exhaustive manual review by an expert constitutes the *de facto* standard of acceptable practice. In many contexts, a single query is used to identify the subset of the collection for manual review. In Boolean retrieval, the query specifies precisely the subset to be reviewed; in ranked retrieval, the query suggests the nature of relevance, the search tool ranks the documents by their likelihood of relevance, and the user assesses some number of the top-ranked documents. Traditionally, relevance feedback has been construed as a method to automate the query-formulation task envisioned by TREC: The user's assessment of the results from an initial query are provided to the search tool, which reformulates the query and presents a new set of results to the user, and so on. More recently, supervised machine-learning methods have been used to harness relevance feedback, with reported effectiveness apparently exceeding Voorhees' "practical upper bound," *see e.g.*, [6, 14, 25].

3 MODELING ASSESSMENT ERROR

3.1 Assessment Error in Measurement

For the purposes of this study, we assume that every document d is either "relevant" or "not relevant" in its own right ($r(d) \in \{rel, nrel\}$), but its relevance can be observed only indirectly by an assessment under conditions c yielding a positive or negative judgment ($j(c,d) \in \{+,-\}$). We use the abbreviations rel_d, $nrel_d$, $+_{cd}$, and $-_{cd}$ to denote $r(d) = rel$, $r(d) = nrel$, $j(c,d) = +$, and $j(c,d) = -$, respectively. We assume that for a random document D, a positive judgment is evidence of relevance: $\Pr[rel_D|+_{cD}] > \Pr[rel_D]$. It follows that a negative judgment is evidence of non-relevance: $\Pr[nrel_D|-_{cD}] > \Pr[nrel_D]$. One of the principal questions to be addressed by a model is: How strong is this evidence?

The Cranfield method generally assumes for the purpose of evaluation that human assessments are infallible ($\Pr[rel_D|+_{cD}] = 1$, $\Pr[rel_D|-_{cD}]) = 0$), where c is chosen carefully, considering the myriad of factors that influence relevance assessment.

Biased sampling and/or statistical sampling may be used to to reduce the cost of assessment. The pooling method [18] is the most prominent of a family of biased sampling methods that identify a subset of documents for human assessment, and render automatic judgments for the remaining documents. In the pooling method, each document d is either in the pool or not ($p(d) \in \{judged, unjudged\}$); documents in the pool are assessed, while documents not in the pool are summarily deemed not relevant. The pooling method can be viewed as a semi-automated assessment under conditions c' where $j(c',d) = \begin{cases} j(c,d) & (judged_d) \\ - & (unjudged_d) \end{cases}$. Biased sampling methods place further stress on the Cranfield assumption that $j(c',D)$ is infallible.

Statistical sampling may be used to estimate the proportion of relevant documents in particular subsets of the collection, as necessary to compute summary measures of effectiveness [2].

Multiple assessments per document may be used in place of a single assessment. The majority judgment of a n assessments under conditions c_1 ... c_n will more closely approximate an infallible assessor, under the assumption that there is greater than 50% conditional probability that each judgment will be positive for a relevant document, and negative for a non-relevant document, notwithstanding the other judgments.

Where multiple fallible assessments are available, latent class analysis [21] may be used to infer the true positive rate $\Pr[+_{c_iD}|rel_D]$ and false positive rate $\Pr[+_{c_iD}|nrel_D]$ for each of the assessment conditions, as well as prevalence $\Pr[rel_D]$, under the assumption of pairwise conditional independence: $\Pr[+_{c_iD}|r(D)] = \Pr[+_{c_iD}|r(D), j(c_j, D)]$ for all $c_i \neq c_j$.

3.2 Assessment Error in Simulation

Total-recall methods may require relevance assessment for three purposes: (i) To train the system to rank or classify the remaining documents; (ii) to determine the ultimate disposition of each document presented to the user by the system (*e.g.*, produce or withhold in the context of electronic discovery in civil litigation, include or exclude in the context of systematic review in evidence-based medicine); and (iii) to inform the cost-benefit analysis inherent in determining when to stop the review process. For the purposes of this work, we aspire to emulate a user whose fallible assessments are conditionally independent of those used for evaluation. The

fallible assessments used for evaluation, although they may closely emulate those of a user, would confound evaluation were they also to be used to simulate feedback, as they are not conditionally independent. Infallible assessments—if they existed—would be conditionally independent, but a poor emulation of a real user's feedback. In either case, it is desirable to use a separate set of assessments to emulate user feedback. Even so, it is well known that the order of presentation and the proportion of relevant documents can influence human assessment [24, 27, 28]; such influences are not easily controlled when simulating different total-recall methods. The quest for better models to emulate human assessment is met with a triple challenge: (i) determining the true relevance of a document; (ii) aptly modeling the user's response; and (iii) ensuring the model is conditionally independent of the model used for evaluation.

3.3 When to Stop?

An important but rarely studied issue in achieving total recall is when to stop. Blair and Maron [4] reported that users who terminated their searches when they believed they had achieved at least 75% recall, had in fact achieved 20% recall. Eliciting from the user a reliable judgment of when high recall has been achieved remains a vexing problem. Evaluations styled after the TREC AdHoc task have largely finessed this issue, reporting rank-based measures under the assumption that the user would know when to stop, perhaps after reading a fixed number of documents.

Automated methods show some promise, but have not previously been evaluated in terms of end-to-end recall with a human in the loop, where user feedback is independent of the assessments used for evaluation. Cormack and Grossman [6] have reported statistical and non-statistical methods for ensuring, with high probability, that their continuous active learning ("CAL") method achieves very high recall, at the expense of precision. The non-statistical "knee method" searches for an inflection point in the recall-effort curve—the "knee"—and continues well beyond that point. Empirical evidence, based on an assumption of infallible user feedback, suggests that their knee method can achieve system-level recall of over 90%, with more than 95% probability, with precision considerably less than 50%.

4 EXPERIMENTAL DESIGN

We conducted two experiments to test the hypothesis that total-recall systems with human assessors in the loop could achieve comparable—or higher—recall and precision, with a small fraction of the effort, than an expert assessor who examined every document in the collection. For this effect to be observable, it is necessary to depart from a model assuming infallible assessment, at least with respect to human assessors in the loop. For evaluation, it is necessary to have a source of reasonably authoritative assessments separate from those used for relevance feedback.

4.1 Datasets, Topics, and Assessments

Our first experiment simulated participation in the TREC 4 AdHoc Task, using the TREC 4 test collection consisting of 567,528 documents, 49 topics, and the official NIST gold standard of relevance [17]. For user feedback and QC, we used two alternate sets of judgments obtained by NIST, for the same topics, using assessors distinct from those who created the gold standard—the same set of alternate assessments studied by Voorhees [29].

Our second experiment reprised the exhaustive manual review of 401,960 email messages from the administration of former Virginia Governor Tim Kaine, which was undertaken by Senior State Archivist Roger Christman, prior to the publication, in 2014, of those he deemed to be "open records."[2] To simulate user feedback in our experiment, we used Roger's original assessments (the "Roger I" assessments). To simulate QC, Roger re-reviewed blind a stratified sample of 2,798 documents (the "Roger II" assessments). As the ultimate arbiter of truth, Roger reviewed blind, for a third time, all 901 cases of disagreement between Roger I and Roger II (the "Roger III" assessments).

4.2 Total-Recall Methods

Our simulation used the same TREC Total Recall Baseline Model Implementation ("BMI"), referenced above in Section 2, modified to read the dataset, topics, and simulated relevance assessments from local files, instead of a server, and to implement Cormack and Grossman's "knee-method" stopping criterion [6]. BMI runs autonomously and has no tunable parameters: Our input consisted of the datasets, topics, relevance assessments, and the knee method. Output from the BMI runs consisted of: a ranked list of documents, in the order presented to the simulated user, ending where the knee-method stopping criterion was met; and the inflection point ("knee") in the gain curve, determined retrospectively by the knee method.

The output from BMI was further manipulated to simulate three different result-selection strategies: (i) System-Determined, (ii) User-Determined, and (iii) Adjudicated. The end result of the *System-Determined* strategy was the entire ranked list returned by BMI. The end result of the *User-Determined* strategy was the subset of documents in the ranked list that the user judged to be relevant. The end result of the *Adjudicated* strategy was the subset of the ranked list consisting of those documents that the user *and* the knee method agreed were relevant, or, where the user and knee method disagreed, a second, auxiliary assessment deemed to be relevant. For the purposes of this work, we deemed the knee method to judge all documents in the ranked list before the knee to be relevant, and all documents after the knee to be non-relevant.

[2] *See* http://www.virginiamemory.com/collections/kaine/.

Strategy	Recall	Precision	F_1	Effort
Manual	0.57	0.69	0.63	567,528
System	0.94	0.06	0.10	**22,911**
User	0.55	0.81	0.62	**22,911**
Adjudicated	**0.64**	**0.82**	**0.69**	23,662
Adj. − Man.	0.07	0.13	0.06	543,866
p-value	0.0001	0.0002	0.0001	-

Table 1: Average Effectiveness Measures Over 98 Combinations of 49 TREC 4 Topics and Two Simulated Users. p was computed using a paired t-test.

4.3 Evaluation

For all strategies, we report recall, precision, and F_1 , as well as effort, as measured by the number of documents presented to the user for assessment. As a baseline, we used an *Exhaustive Manual Review* ("Manual Review") strategy, for which effort is simply the number of documents in the collection. For the System-Determined strategy, effort depends on the number of *relevant* documents in the collection, as well as the recall and precision achieved: For a given topic and recall level, effort is inversely proportional to precision. For the User-Determined and Adjudicated strategies, there is no direct relationship between effort and precision.

Only the System-Determined strategy returns a ranked list from which we can evaluate the recall-precision tradeoff. We can, however, plot recall and precision as a function of effort throughout the progress of a review. These curves illustrate the result that might have been achieved, had a different stopping criterion been applied. They also illustrate how quickly a substantial fraction of the relevant documents can be discovered and forwarded for further analysis or release, while the total-recall effort is still in progress.

4.4 Prediction and Rationale

Previously published results report recall-precision break-even scores on the order of 80% for BMI and related methods [7, 8, 11, 25]. With one notable exception (discussed below), these results were derived using simulated feedback from an assumed-infallible user, and evaluated with respect to the same assumed-infallible gold standard. On the one hand, the simulated feedback was conditionally dependent on the evaluation standard, and therefore possibly "too good to be true." On the other hand, the evaluation standard was assumed to be perfect, offering BMI no opportunity to better it. The experiments, by design, could not show whether or not BMI could achieve better recall and/or better precision than a fallible user.

There is no basis to assume that a hybrid human-computer system cannot exceed both the recall and precision of its human operator. The literature reports inter-assessor agreement results that are consistent with the hypothesis that a human assessor can achieve on the order of 70% recall and 70% precision [26, 29]. Are the higher results reported for BMI an artifact of too-perfect training, or is a system involving BMI and human assessment, combined, superior to human assessment alone?

Achieving a high recall-precision break-even score is irrelevant to the success of a total-recall effort, if the point at which this score is achieved is unknown to the user. Cormack and Grossman's knee-method stopping criterion sacrifices (System-Determined) precision to achieve very high recall, under the assumption that an infallible assessor would screen the results, and the only consequence of low precision would be increased effort. The User-Determined strategy has the (fallible) user act in this capacity.

The Adjudicated strategy has BMI and the user share the role of screening, deferring to a second (fallible) assessor the adjudication of cases of disagreement. We assume the inflection point calculated by the knee method to be a good approximation of the recall-precision break-even point; that documents before the knee are more likely to be relevant than not, and that documents after the knee are less likely to be so. This assumption motivates our choice to defer to a second assessor any document before the knee that is judged non-relevant by the user, and any document after the knee that is judged relevant by the user. This strategy makes no assumption that the second assessor is "better" than the user. As long as the second assessor is usually correct (as would certainly be the case for an assessor capable of achieving 70% recall and 70% precision), the Adjudicated strategy should achieve higher recall *and* higher precision than the User-Determined strategy.

As noted above, one strand of research has evaluated total-recall methods in the face of fallible users. The TREC Legal Track Interactive Task (*see*, *e.g.*, [19, 23]) assigned participating teams the task of finding all and only the relevant documents that were responsive to requests for production in a mock civil litigation. A subject matter expert (the "Topic Authority") was made available for consultation while the teams were conducting their reviews; the same Topic Authority adjudicated cases of disagreement, after the fact, between the teams and the human assessors who had created a provisional gold standard for evaluation. Teams were allowed to use any method of their choosing, with no restriction on the nature or quantity of human input. Two TAR methods—one rule-based and one substantially similar to BMI—achieved on the order of 80% recall and 80% precision [19]. In a subsequent analysis, Grossman and Cormack [14] estimated the recall and precision of the human assessments that comprised the provisional gold standard, on average, to have been 59.3% and 31.7%, respectively. While deferring a fuller discussion of these results to Section 5, we note that this work generated some criticism, *e.g.*, [12, 13, 31]; most notably, claims that: (i) the assessors were unskilled, poorly trained, poorly vetted, or poorly supervised; (ii) the assessors had a different "conception of relevance" from the Topic Authority; (iii) the participating teams devoted extraordinary skill or extraordinary resources to accomplishing the task; (iv) the gold standard, by virtue of the reconsideration process, was biased in favor of the participants; and (v) the gold standard, by

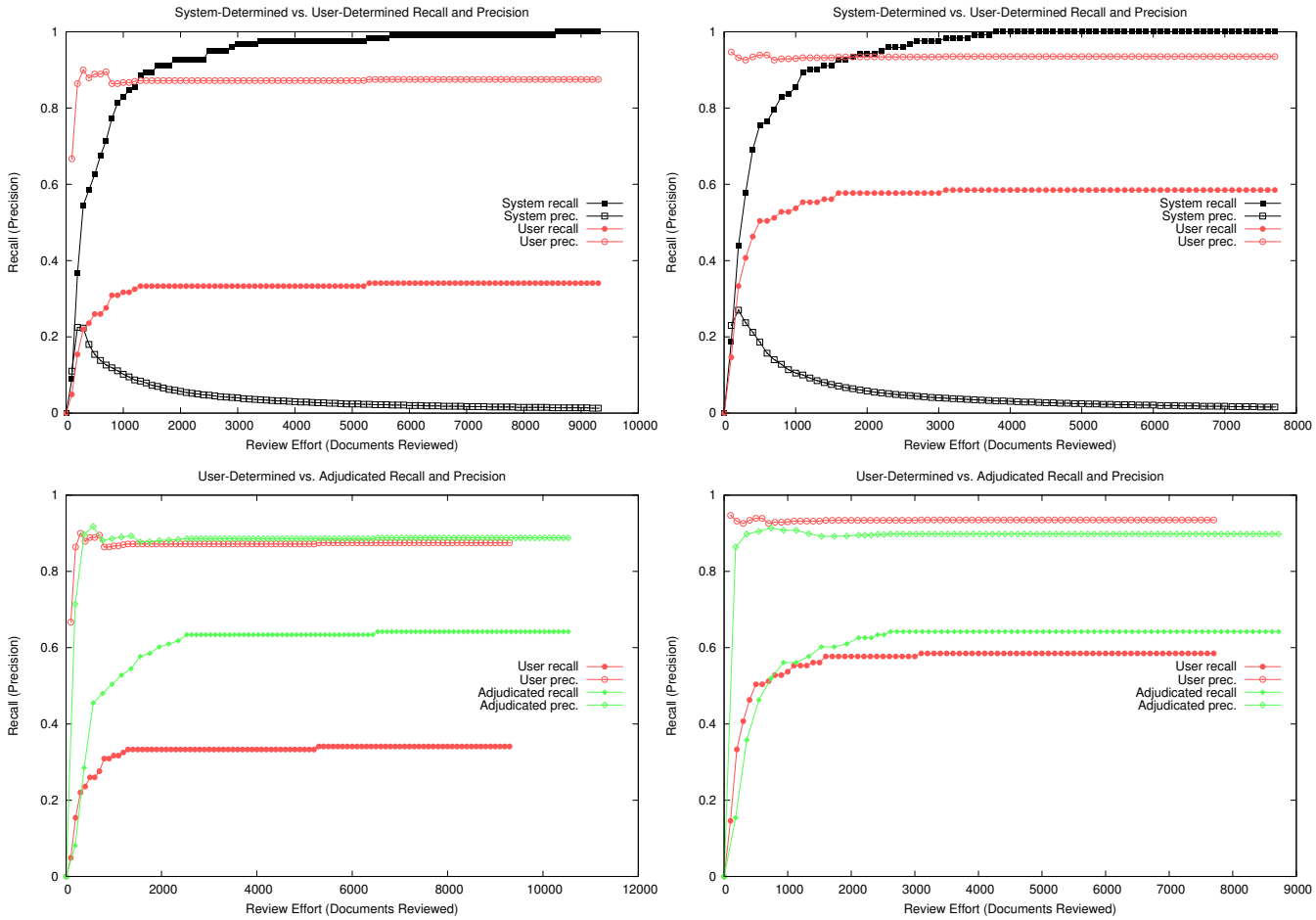

Figure 1: TREC 4 Topic 239 – Tradeoff Between Recall, Precision, and Effort. The top panels compare the System-Determined vs. User-Determined strategies; the bottom panels compare the User-Determined vs. Adjudicated strategies. For the left panels, the first alternate TREC assessor was the user, and the second alternate assessor was the adjudicator; for the right panels, the roles were reversed.

virtue of bias on the part of the Topic Authority and the lack of blinding of his or her review, was biased in favor of the participants.

This study controls for the skill, effort, and motivation of both users and assessors. The gold standard for the TREC 4 collection, as well as the alternate assessments, were fixed more than two decades ago. All of the NIST assessors were clearly skilled in their craft; most were former NSA analysts. The alternate assessors had no direct knowledge of the primary assessor's judgments. On the other hand, the simulated assessments for the Kaine email dataset were derived from an assessment of 401,960 documents, by the Virginia Senior State Records Archivist, in his official capacity. In forming the gold standard, the same archivist reviewed and then re-reviewed some of his own previous assessments, after wash-out periods of two years and then two months, respectively.

Our rationale predicted that: (i) BMI alone (the System-Determined strategy) would achieve superior recall to Manual Review, but inferior precision, for substantially less effort; (ii) the User-Determined strategy would achieve inferior recall, but superior precision, to the System-Determined strategy, for the same effort; and, (iii) the Adjudicated strategy would achieve superior recall and precision to all other strategies, for moderately higher effort than the System-Determined and User-Determined strategies, but still substantially less than Manual Review.

5 RESULTS

Figure 1 plots recall and precision for a representative TREC 4 topic as a function of effort for each of the BMI-derived methods, using each of the alternate assessors as the user, and the other assessor, as occasioned, for adjudication.[3] In comparison, the recall and precision of the Manual Review by

[3]Plots and raw results for the other 48 topics are available on request from the authors.

	Manual Review				Adjudicated			
Topic	Recall	Precision	F_1	Effort	Recall	Precision	F_1	Effort
Legal Hold	**0.97**	0.91	0.94	401,960	0.96	**0.96**	**0.96**	**40,522**
Archival	0.89	0.84	0.86	381,819	**0.90**	**0.89**	**0.89**	**332,410**
Restricted	**0.98**	0.75	0.84	146,594	0.95	**0.80**	**0.87**	**38,048**

Table 3: Individual Topic Effectiveness for the Kaine Email Dataset.

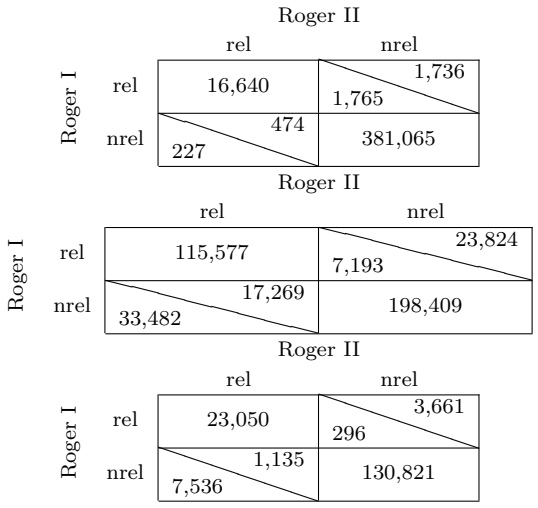

Table 2: Agreement Among Roger I, Roger II, and Roger III on the Virginia Tech Legal Hold Emails. In split cells, numbers below the diagonal show agreement between Roger I and Roger III; numbers above the diagonal show agreement between Roger II and Roger III. The top panel shows Virginia Tech legal hold identification; the middle panel shows archival record identification; the bottom panel shows restricted record identification.

Strategy	Recall	Precision	F_1	Effort
Manual	**0.95**	0.83	0.88	310,196
Adjud.	0.93	**0.88**	**0.91**	**136,993**
Δ	-0.01	+0.05	+0.02	173,203
p-value	0.4	0.006	0.03	
95% *c.i.*	(−.05, .03)	(.03, .07)	(.004, .04)	

Table 4: Average Effectiveness Measures Over Three Topics for the Kaine Email Dataset. p was computed using a paired t-test.

	Legal Hold	Archival	Restricted
Roger I & Roger II	**80.6%**	60.2%	64.2%
System & Roger I	79.1%	**70.2%**	**67.9%**
System & Roger II	79.9%	62.1%	55.8%

Table 5: Pairwise Overlap (*i.e.*, Jaccard Index) Between the System, Roger I, and Roger II.

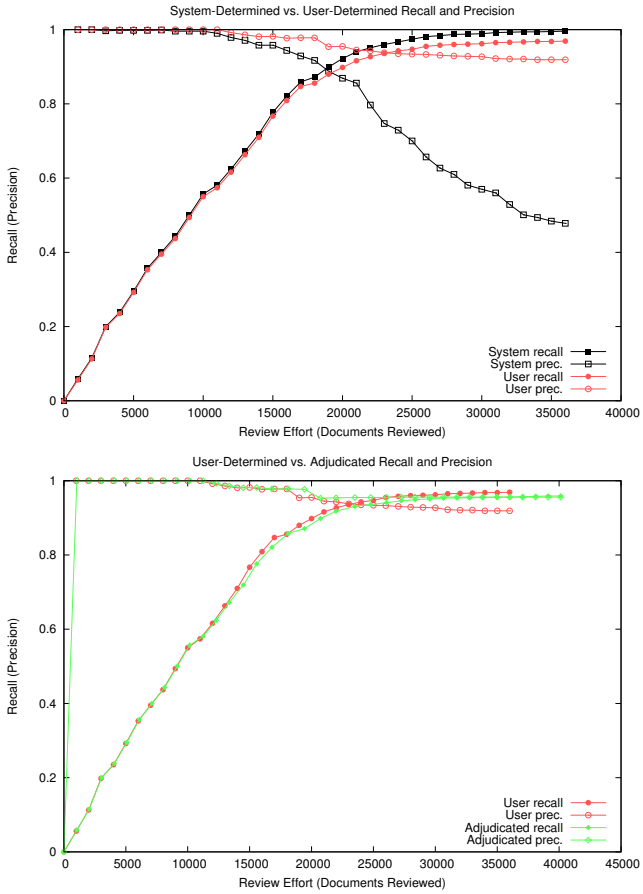

Figure 2: Identification of Kaine Administration Email Pertaining to the Virginia Tech Shooting for Legal Hold – Tradeoff Among Recall, Precision, and Effort. The top panel compares the System-Determined vs. User-Determined strategies; the bottom panel compares the User-Determined vs. Adjudicated strategies.

the two assessors were 0.34 and 0.88, and 0.59 and 0.94, respectively. Although the first assessor has substantially lower recall and precision than the second, the System-Determined recall curves are remarkably similar. The User-Determined recall curves are, as predicted, bounded by the recall of the respective users. On the other hand, the Adjudicated recall curves, like the System-Determined curves, are remarkably

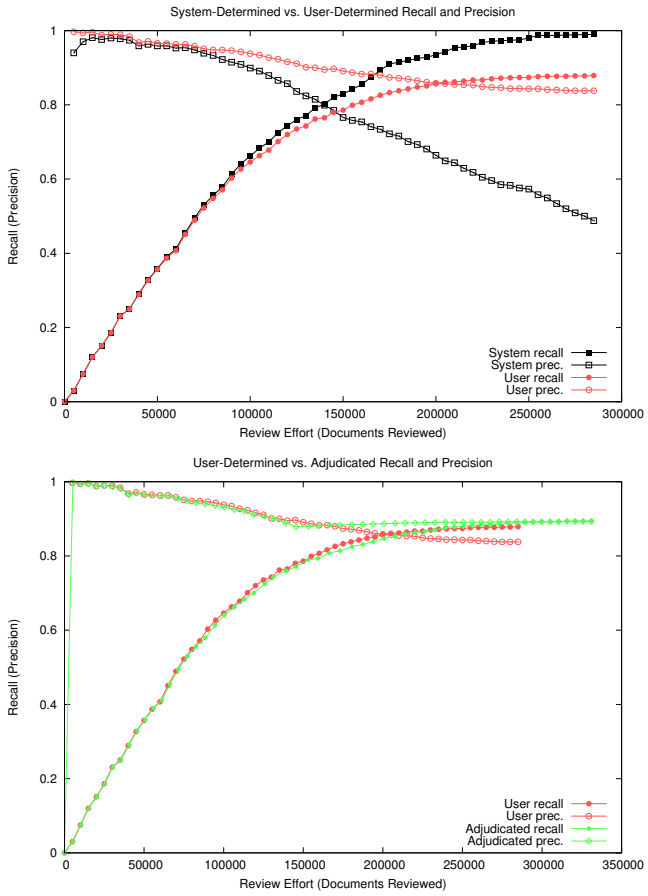

Figure 3: Identification of Archival Records from the Kaine Administration – Tradeoff Among Recall, Precision, and Effort. The top panel compares the System-Determined vs. User-Determined strategies; the bottom panel compares the User-Determined vs. Adjudicated strategies.

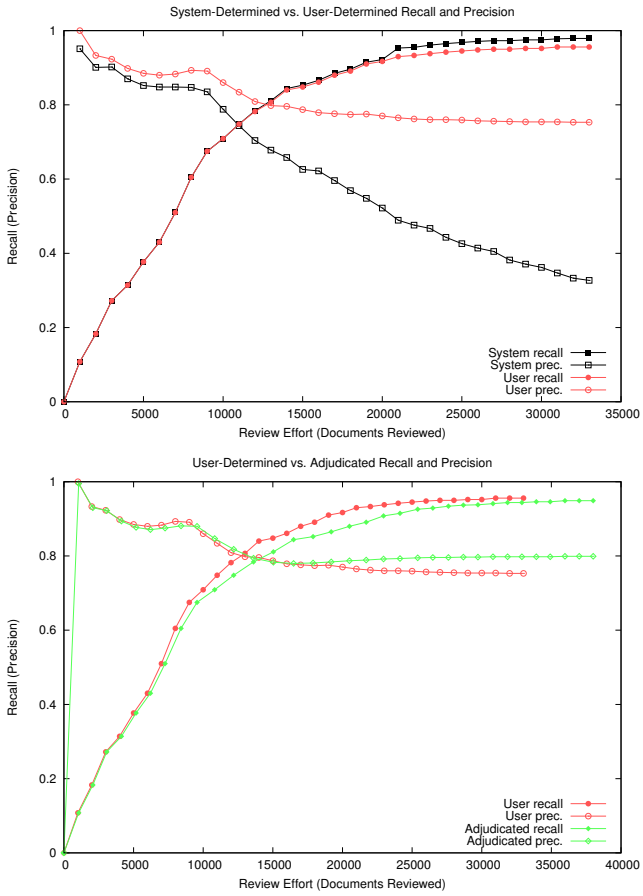

Figure 4: Identification of Restricted Archival Records from the Kaine Administration – Tradeoff Among Recall, Precision, and Effort. The top panel compares the System-Determined vs. User-Determined strategies; the bottom panel compares the User-Determined vs. Adjudicated strategies.

similar, but are superior to both the User-Determined curves. The System-Determined precision curve initially climbs and then declines with increased effort, as expected, while the User-Determined and Adjudicated precision curves are remarkably flat.

Table 1 shows average effectiveness and effort measures over 98 runs, comprising 49 topics and two simulated users. As predicted, the System-Determined strategy achieves very high recall on average, with 4% of the effort of Manual Review. The User-Determined strategy achieves slightly lower recall, but substantially higher precision than Manual Review, also with 4% of the effort. The Adjudicated strategy achieves substantially and significantly higher recall, and precision than Manual Review, with 4.2% of the effort.

The agreement between Roger I and Roger II, and Roger III's adjudication of their disagreements, is shown in Table 2.

Overall, Roger I and Roger II have overlap (*i.e.*, Jaccard index) of 80.6%, 60.2%, and 64.2% on each of three topics—higher than most reported results for separate assessors, but far from perfect. Roger III generally splits the difference between Roger I and Roger II, with a propensity to agree with the negative assessment. Roger, on completing the Roger III assessments, volunteered that "this was a challenging review," suggesting that the adjudication process had identified many hard-to-classify, as opposed to randomly misclassified, documents.

Roger I rendered these decisions for each of the three topics seriatim as follows: First, the Virginia Tech documents subject to a legal hold were identified; second, documents not subject to the hold were classified as either archival records or non-records; and finally, documents classified as archival records were categorized as restricted or open records. As a consequence, the document collection diminished for each

subsequent topic. Roger II and Roger III employed the same protocol, resulting in a handful of anomalous judgments. For example, Roger I classified some records as records or non-records not subject to legal hold, while Roger II classified them as subject to legal hold. For these documents, we recorded the disagreement with respect to legal hold, and asked Roger II to specify whether the document would be an archival record or a non-record, were it not subject to legal hold. Roger III was asked in advance to consider all six combinations of: subject to legal hold or not, and open record, restricted record, or non-record. The appropriate hypothetical judgments were used as the gold standard for each topic.

The documents reviewed by Roger II formed a stratified sample of the dataset; measures using Roger II or Roger III were estimated using the Horvitz-Thompson estimator [20]. The strata were selected as follows: For each of the three topics and each of the four possible modes of disagreement, 200 documents were selected independently, at random. Because the documents were selected independently, there was some overlap among these strata, and the total number of unique documents was 2,398. After Roger II had commenced his review, it was discovered that one stratum had been repeated and one had been omitted due to a clerical error, so 200 documents from the omitted stratum were added, along with 200 randomly selected documents from outside the stratum, for a total sample size of 2,798. Inclusion probabilities were adjusted to account for the overlapping strata.

Figures 2, 3, and 4 show effectiveness versus effort for the Adjudication strategy, while Table 3 shows set-based measures for the three Kaine email topics, and Table 4 shows averages over the topics. The summary measures suggest that the Adjudication strategy achieves significantly better precision and F_1 than Roger I ($p < 0.05$), and no significant difference in recall. The gain in F_1 appears to come primarily from balancing recall and precision, which is consistent with the purpose of the Adjudication strategy. None of the differences is larger than 0.5%, suggesting that there is little to choose (in terms of effectiveness) between the Adjudication strategy and Manual Review. On the other hand, the Adjudication strategy offers a huge advantage in terms of efficiency.

Table 5 shows the pairwise overlap between the system's assessment (that was compared to Roger I's assessment in the Adjudication strategy), and Roger I and Roger II themselves. Collectively, the results indicate that, for all intents and purposes, there is little to choose between the system's judgments and Roger's, and that a second opinion—whether by a human or a bionic assessor—can be helpful.

The bootstrap method was used to determine the variance due to sampling in the Roger II and Roger III datasets, which showed all per-topic differences between the Adjudication strategy and Manual Review to be significant (with respect to sampling uncertainty).

DISCUSSION AND CONCLUSIONS

Effectiveness measures for total recall depend on who you ask, when you ask, and how often you ask. Even a small amount of error in gold-standard assessments can substantially depress recall, as evidenced by the provisional versus final recall estimates of the TREC 2009 Legal Track, where the estimated recall of the best submissions for four topics rose from less than 20% in the Notebook Draft, to about 80% in the Final Overview Paper [19, appendix]. While Webber et al. [32] attribute this difference to bias on the part of the assessors, it is equally well explained by a typical true positive rate $\Pr[+_D|rel_D > 0.7]$, and a typical false positive rate $\Pr[+_{cD}|nrel_D] \ll 0.01$. The difference between the Legal Track assessment and other TREC efforts is that the assessors reviewed a statistical sample of the entire collection, not just the pool of documents identified by the systems. As a consequence, a sample representing more than 700,000 non-relevant documents was reviewed; it is no surprise that examples representing several thousand of these non-relevant documents were incorrectly judged as relevant. Most of those false positives were identified by a process similar to the Adjudication strategy we evaluated, in which disagreements between the participating systems and the first assessor were adjudicated by a second assessor, the Topic Authority. It is not necessary to assume that the first assessor was incompetent, or that the Topic Authority was more competent than the first assessor, to explain the TREC 2009 results.

Our Adjudication strategy results on the Kaine email dataset are consistent with the superior results reported by Cormack and Mojdeh at TREC 2009 [5]; they "[re-]examined documents with high scores that were marked 'not relevant' and documents with low scores that were marked 'relevant.'" Our results are also entirely consistent with the results reported in our original 2011 study [14]. It is important to bear in mind that results measuring *system* recall, rather than *end-to-end* recall, cannot be compared to the results reported here, to those reported in the TREC 2009 Legal Track Overview [19], or to those reported by Grossman and Cormack [14]. Neither can such results be compared when the user assessments are the same as the evaluation assessments.

Our results do not support the mantra of "garbage in, garbage out," or that errors in user feedback are "amplified" by the use of a TAR method, as opposed to manual review. To the contrary, our results show that system effectiveness is hardly affected by inferior feedback, and that certain TAR methods can *mitigate* rather than *amplify* user error.

This study raises a number of questions that may be addressed by future work. It is well known—and reconfirmed by this study—that humans judge the same document differently under different circumstances, including the order of presentation. The effect of using a dynamically learned ranking on user feedback has yet to be studied. On the one hand, studies suggest that when assessors review a higher proportion of relevant documents, they are less likely to judge them relevant (*see*, *e.g.*, Roegiest [24]). Is this explained by a higher error rate, or by the general observation that higher

prevalence or more experience with a review set can lead assessors to become more discriminating? Our results show that Roger III was less likely to judge documents relevant than Roger I or Roger II, perhaps because he exercised greater diligence, or maybe because he had become better informed through the course of examining borderline documents.

Roger II and Roger III were blind to the previous Rogers' assessments. While Roger II commented that he recalled several of the themes in the documents, it had been at least two years since he had previously reviewed them. Roger III, on the other hand, had seen the documents two months prior, and in the interim, had reviewed more than ten thousand emails from a different department of the Kaine administration. It is not apparent what effect Roger's memory may have had on the results. The impact of blind review with a wash-out period has yet to be studied; indeed, it is not clear whether the user should be blind, so as to reduce bias, or informed, so as to aid in deliberation, *cf.* [22]. At TREC 2009, Cormack and Mojdeh [5] appear to have achieved superior results without blinding and no discernible wash-out period, but more study is necessary to arrive at a definitive answer.

Overall, our results reconfirm the thesis that hybrid human-computer classification (*i.e.*, TAR) methods can achieve recall and precision that compare favorably with exhaustive manual review by experts, for much less effort. Where higher recall and precision is desired, additional resources are better spent re-reviewing documents that may have been misjudged by the user, than examining the ranked list to extraordinary depths, or sampling low-ranked documents. When recall and precision values approach 100%, it is essential to consider carefully both the accuracy and independence of the gold standard used for evaluation.

ACKNOWLEDGEMENT

We are very grateful for the enthusiasm and support we received from our colleagues at the Library of Virginia, most notably, Roger Christman, Susan Gray Page, Rebecca Morgan, and Kathy Jordan; without them, this work would not have been possible.

REFERENCES

[1] A. Al-Maskari and M. Sanderson. A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology*, 61(5):859–868, 2010.

[2] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *SIGIR 2006*.

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

[4] D. Blair and M. E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289–299, 1985.

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

[6] G. V. Cormack and M. R. Grossman. Engineering quality and reliability in technology-assisted review. In *SIGIR 2016*.

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

[8] G. V. Cormack and M. R. Grossman. Multi-faceted recall of continuous active learning for technology-assisted review. In *SIGIR 2015*.

[9] G. V. Cormack and M. R. Grossman. Scalability of continuous active learning for reliable high-recall text classification. In *CIKM 2016*.

[10] G. V. Cormack and M. R. Grossman. Waterloo (Cormack) participation in the TREC 2015 Total Recall Track. In *TREC 2015*.

[11] G. V. Cormack and M. R. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv:1504.06868*, 2015.

[12] G. V. Cormack, M. D. Smucker, and C. L. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 14(5):441–465, 2011.

[13] S. Green and M. Yacano. Computers vs. humans? Putting the TREC 2009 study in perspective. *New York Law Journal*, (Oct. 1), 2012.

[14] M. R. Grossman and G. V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law & Technology*, 17(3), 2011.

[15] M. R. Grossman, G. V. Cormack, and A. Roegiest. TREC 2016 Total Recall Track Overview. In *TREC 2016*.

[16] D. Harman. Overview of the First Text REtrieval Conference (TREC-1). In *TREC 1*, 1992.

[17] D. Harman. Overview of the fourth text retrieval conference (trec-4). In *TREC 4*, 1996.

[18] D. K. Harman. The TREC ad hoc experiments. In E. M. Voorhees and D. K. Harman, editors, *TREC - Experiment and Evaluation in Information Retrieval*, chapter 4. MIT Press, 2005.

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

[20] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[21] K. Krstovski. *Efficient Inference, Search and Evaluation for Latent Variable Models of Text with Applications to Information Retrieval and Machine Translation*. University of Massachusetts, 2016.

[22] T. McDonnell, M. Lease, T. Elsayad, and M. Kutlu. Why is that relevant? Collecting annotator rationales for relevance judgments. In *4th HCOMP*, 2016.

[23] D. W. Oard, B. Hedin, S. Tomlinson, and J. R. Baron. Overview of the TREC 2008 Legal Track. In *TREC 2008*.

[24] A. Roegiest and G. V. Cormack. Impact of review-set selection on human assessment for text classification. In *SIGIR 2016*.

[25] A. Roegiest, G. V. Cormack, M. R. Grossman, and C. L. A. Clarke. TREC 2015 Total Recall Track Overview. In *TREC 2015*.

[26] H. L. Roitblat, A. Kershaw, and P. Oot. Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.

[27] M. Sanderson et al. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[28] L. Schamber. Relevance and information behavior. *Annual review of information science and technology (ARIST)*, 29:3–48, 1994.

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

[30] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 355–370. Springer, 2001.

[31] W. Webber. Re-examining the effectiveness of manual review. In *Proc. SIGIR Information Retrieval for E-Discovery Workshop*, 2011.

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