Engineering Quality and Reliability in Technology-Assisted Review

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

Maura R. Grossman*
Wachtell, Lipton, Rosen & Katz
mrgrossman@wlrk.com

ABSTRACT

The objective of technology-assisted review ("TAR") is to find as much relevant information as possible with reasonable effort. *Quality* is a measure of the extent to which a TAR method achieves this objective, while *reliability* is a measure of how consistently it achieves an acceptable result. We are concerned with how to define, measure, and achieve high quality and high reliability in TAR. When quality is defined using the traditional goal-post method of specifying a minimum acceptable recall threshold, the quality and reliability of a TAR method are both, by definition, equal to the probability of achieving the threshold. Assuming this definition of quality and reliability, we show how to augment any TAR method to achieve guaranteed reliability, for a quantifiable level of additional review effort. We demonstrate this result by augmenting the TAR method supplied as the baseline model implementation for the TREC 2015 Total Recall Track, measuring reliability and effort for 555 topics from eight test collections. While our empirical results corroborate our claim of guaranteed reliability, we observe that the augmentation strategy may entail disproportionate effort, especially when the number of relevant documents is low. To address this limitation, we propose stopping criteria for the model implementation that may be applied with no additional review effort, while achieving empirical reliability that compares favorably to the provably reliable method. We further argue that optimizing reliability according to the traditional goal-post method is inconsistent with certain subjective aspects of quality, and that optimizing a Taguchi quality loss function may be more apt.

Keywords: Technology-assisted review; predictive coding; electronic discovery; e-discovery; test collections; relevance feedback; continuous active learning; reliability; quality; systematic review.

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

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGIR '16 July 17-21, 2016, Pisa, Italy

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4069-4/16/07.

DOI: http://dx.doi.org/10.1145/2911451.2911510

1. INTRODUCTION

A vexing question that has plagued the use of technology-assisted review ("TAR") is "when to stop"; that is, knowing when as much relevant information as possible has been found, with reasonable effort. We present a provably reliable method to achieve high recall using any search strategy that repeatedly retrieves documents and receives relevance feedback, continuing indefinitely until a decision is made to discontinue the review process. Amenable search strategies include traditional ranked retrieval,[1] interactive searching and judging [8], move-to-front pooling [8], and continuos active learning ("CAL") [5].

For the particular implementation of CAL supplied as the baseline model implementation ("BMI") [7] for the TREC 2015 Total Recall Track [13], we present two stopping procedures that achieve superior empirical reliability for comparable effort, and comparable empirical reliability for less effort, relative to our provably reliable method.

Our primary motivation is to provide quality assurance for TAR applications, including electronic discovery ("eDiscovery") in legal matters [5], systematic review in evidence-based medicine [10], and the creation of test collections for information retrieval ("IR") evaluation [14]. Since these applications generally require that a human review each relevant document, we assume for this study that the effort to provide relevance feedback for relevant documents is a sunk cost. On the other hand, the effort to assess and provide relevance feedback for non-relevant documents is wasted. We measure review effort in terms of the total number of documents reviewed, whether relevant or not. An ideal search would find all of the relevant documents with effort equal to precisely that number. An acceptable search would find most of the relevant documents with minimal wasted effort.

A reliable search method would achieve an acceptable search most of the time. More formally, if S is a random variable representing a search, and $acceptable(s)$ is an indicator function denoting whether a particular search s has an acceptable result, we define:

$$reliability =_{def} \Pr[acceptable(S) = 1]\,.$$

To this end, we define $recall(s)$ and $effort(s)$ to be the recall and effort associated with s. For simplicity, our primary

[1]To be amenable, a retrieval method must be able to rank the entire collection. Incomplete rankings or set-based results may be extended by adding the remaining documents in any order.

Collection	Source	Description	# Docs	# Topics	# Rel (R)
At Home	TREC 2015 Total Recall	Jeb Bush public email	290,000	10	227-17,135
At Home	TREC 2015 Total Recall	Hacker forums	465,147	10	179-9,517
At Home	TREC 2015 Total Recall	Local news	902,434	10	23-2,094
Kaine	TREC 2015 Total Recall	Tim Kaine non-public email	401,953	4	14,341-166,118
MIMIC II	TREC 2015 Total Recall	MIMIC II Clinical Database	31,538	19	180-19,182
RCV1-v2	Reuters	News subject categories	804,414	103	5-381,327
Filtering	TREC 2012 Filtering	NIST topics	804,414	50	12-610
Intersection	TREC 2012 Filtering	Conjunction of RCV1-v2 subject pairs	804,414	50	21-349
Robust-04	TREC 2004 Robust	Amalgam of TREC ad-hoc topics	528,256	249	4-161
Robust-05	TREC 2005 Robust	50 legacy topics, new dataset	1,033,461	50	17-376

Table 1: Ten Evaluation Datasets. In our experiments, the three At Home datasets are treated as a single test collection, for a total of eight test collections.

results use a *goal-post* definition [18] of acceptability:

$$acceptable(s) = \begin{cases} 1 & (recall(s) \geq 0.70) \\ 0 & (recall(s) < 0.70) \end{cases}.$$

Our primary results further assume that 95% reliability is sufficiently high.

The methods and results detailed in this work are:

- The *target method*: a provably reliable method that chooses ten random relevant documents as a target, and employs an independent search method to retrieve documents without knowledge of the target set, until each document in the target set has been retrieved. We prove that this method achieves 95% reliability for a minimum threshold recall of 70%.
- The *knee method*: a geometric stopping procedure, based on the shape of the gain (*i.e.*, recall versus effort) curve, that augments BMI to achieve similar empirical reliability to the target method, with substantially less effort. The knee method, in contrast to the target method, is practical regardless of the number of relevant documents in the collection.
- The *budget method:* a variant of the knee method that achieves superior empirical reliability to the target method, with similar effort.
- *Empirical validation*: we assess the effectiveness and reliability of our methods on eight archival test collections consisting of 555 topics and 4.5 million documents (*see* Table 1).
- *Quality evaluation*: As an alternative to binary relevance and fixed recall and reliability thresholds, we argue and provide evidence that quality loss functions [18] provide more nuanced measures that better reflect the tensions among consistency, effectiveness, and efficiency.

2. BACKGROUND

The modern literature on the effectiveness and reliability of high-recall retrieval is largely confined to the problem of constructing test collections for IR evaluation, and eDiscovery in legal matters. A 1985 study by Blair and Maron [2] showed that teams of lawyers and paralegals, using iterative Boolean searches, believed they had achieved 75% recall, when in fact they had achieved 20%. Blair [3] later described the difficulty of measuring high recall in general, and the use of targeted searching, systematically constructed Boolean queries, and stratified sampling to estimate recall for the Blair and Maron study.

The Text Retrieval Conference ("TREC") [21] first addressed the problem of IR evaluation for "large" datasets, which at the time of TREC's inception in 1992, contained on the order of 500,000 documents. TREC follows the Cranfield paradigm [20], which evaluates the results of subject systems against a gold standard that identifies every relevant document. For large datasets, the effort to render a human assessment for each document is prohibitive, thus occasioning the use of automated or semi-automated methods to limit the human review effort required to label the dataset. TREC saw the introduction of the "pooling method," which selects the top-ranked documents from a number of independent retrieval efforts for assessment, and deems all other documents to be non-relevant. A number of studies (*see* [19]) indicate that this method fails to identify a substantial number of documents, but even so, the resulting gold standard yields a stable evaluation of the relative effectiveness of candidate systems, as measured by Kendall's τ rank correlation. We are unaware of any studies that address the effectiveness of pool-based gold standards for evaluating high-recall retrieval, or for simulating interactive relevance feedback. Studies suggest that greedy or machine-learning methods to select the pool yield a more nearly complete gold standard [8, 14].

Interactive searching and judging ("ISJ"), in which a searcher repeatedly formulates queries and examines the top results from a relevance-ranking search engine, has been shown to yield gold standards with comparable quality to the pooling method, with considerably less effort [8]. Continuous active learning ("CAL") [5] is essentially the same as ISJ, but uses machine learning instead of, or in addition to, manually formulated queries to rank the documents for review. An approach similar to CAL was used in the TREC 2012 Filtering Track (*see* [17]) to construct the gold standard that was used for evaluation, and also to simulate relevance feedback. A subsequent study based on pooling showed that the CAL-like approach had achieved high recall, and high effectiveness, as measured by Kendall's τ [17]. CAL achieved superior results at the TREC 2009 Legal Track [4], and remains state of the art for eDiscovery.

The TREC 2015 Total Recall Track [13] represents the first study of high-recall human-in-the-loop retrieval in which all aspects of human intervention are simulated, and

hence controlled. Fully automated or semi-automated retrieval systems were tested through their interaction with an evaluation server. At the outset, the evaluation server provided a document collection and a topic description, after which the system under test submitted potentially relevant documents from the collection to the evaluation server. In response, the evaluation server provided an assessment (derived from a pre-computed gold standard) for the submitted documents, and the process continued until the documents were exhausted or the system chose to stop.

Participants in the Total Recall Track were supplied with a CAL baseline model implementation[2] ("BMI") that, when connected to the evaluation server, performed all aspects of the task—other than deciding when to stop—without human intervention. Participating systems were allowed to run indefinitely, and were evaluated (primarily) on the quality of the ranking determined by the order in which the system presented documents to the server. Instead of actually terminating when they thought an acceptable result had been achieved, participants were invited to "call their shot" by indicating, in real time, when they would have stopped, had they been required to balance benefit with cost. The current study considers the addition of a call-your-shot mechanism to BMI, and, more generally, to any ranking system.

The TREC 2015 Total Recall Track contributed five fully labeled archival datasets. The *Jeb Bush*, *Hacker Forums*, and *Local News* datasets were used for the *At Home* task, in which participants ran their systems on their own platforms, connecting via the internet to the evaluation server, which was run by the track coordinators. The *Kaine* and *MIMIC II* datasets were used for the *Sandbox* task, in which participants encapsulated their systems as a virtual machine, which was run by the track coordinators, along with the evaluation server, isolated from the internet.

The reliability of methods for constructing gold standards for IR evaluation has typically been evaluated by how well the resulting gold standard ranks the relative effectiveness of precision-oriented retrieval systems, where the objective is to find as much relevant information as possible at low rank. For this purpose, a calibrated estimate is not required; it is sufficient to determine whether one system achieves higher recall than another, and the actual numerical value is ascribed little meaning. A number of studies (*see* [23]) eschew recall altogether, assuming that the user's information need will be satisfied by a tiny fraction of a vast sea of relevant documents. Zobel et al. [23] suggest that recall is a poor effectiveness measure, even for the "high-recall applications" where the user seeks "total recall," and that only an extensive ad-hoc effort using multiple queries and tools will satisfy the user that their information need has been met.

The reliability and effectiveness of TAR (also known as "predictive coding") is the subject of much interest in the legal community [9, 16]. A number of approaches to TAR, to deciding when to stop, and to quality assurance have been advanced, but no stopping procedure has previously been shown to be mathematically or empirically reliable. Perhaps the most commonly used approach to TAR involves the use of a supervised machine-learning algorithm trained using a set of documents from the collection (typically referred to as a "seed set") to partition the collection into a "review set," which is subject to human review, and a "null set," which is not. This approach is referred to as either simple passive learning ("SPL") or simple active learning ("SAL"), depending on whether or not the learning algorithm is involved in selecting the training documents [5]. Recently, CAL has been advanced as a superior alternative [5, 7].

Regardless of the TAR method used, the question remains of when to stop. For SPL and SAL, two questions must be answered: when to stop training; and how many documents should be included in the review set. For CAL, the sole question is when to stop. One approach that has been advanced is to draw a random hold-out set (referred to as a "control set") that is used to measure the effectiveness of the classifier, in order to determine when to stop training, and then to measure recall, so as to determine how many documents should comprise the review set. The control set must be large enough to contain a sufficient number of relevant documents to yield a precise estimate. Bagdouri et al. [1] note that the use of a control set constitutes sequential sampling, with the net effect that it yields a biased estimate of recall, and cannot be used for quality assurance. As an alternative, they propose "certified text classification," in which part of the review budget is set aside to conduct a frequentist acceptance test that will accept or reject the classifier. Bagdouri et al. are concerned with the problem of testing whether the classifier has achieved a threshold level of F_1; they do not consider recall, or how to proceed in the event that the classifier is rejected by the test.

The limitations of binary relevance may be of particular importance in evaluating the effectiveness and reliability of TAR systems. Binary relevance does not account for the differential importance of relevant documents, and there will necessarily be documents near the threshold about which competent assessors will disagree (*see* [19]). In evaluating the recall of a system against a gold standard, there will necessarily be uncertainty for some documents as to whether the system is correct, the gold standard is correct, or reasonable minds could disagree. If a system fails to meet a target recall threshold, is it because the system has missed important documents, because it has missed marginal documents about which reasonable minds could differ, or because it has missed documents that are incorrectly coded relevant in the gold standard? And, is the effort to remedy the shortfall proportionate to the importance of the missed documents?

Binary relevance and fixed recall targets are examples of traditional goal-post methods in quality engineering ([18]), where success or failure is a binary quantity, and reliability is the probability of success. In quality engineering, a quadratic loss function blends reliability and effectiveness into a single quality measure, with targets, but no arbitrary thresholds [18].

3. GUARANTEED RELIABILITY

Our *target method* involves drawing a target set T of k random relevant documents from the collection; for simplicity we fix $k = 10$, but a different number could be chosen. In order to draw T, we retrieve and review documents selected at random until k relevant documents are found, or the collection is exhausted. The underlying search strategy retrieves documents for review without knowledge of T, until every document in T has been found. This method achieves 95% reliability, as shown below.

Consider a document collection C and a function $rel(d)$ indicating binary relevance. The number of relevant docu-

[2] *See* `http://plg.uwaterloo.ca/~gvcormac/trecvm/`.

ments in the collection is $R = |\{d \in C | rel(d)\}|$. A search strategy is a ranking on C where $1 \leq rank(d) \leq |C|$ denotes the position of d in the ranking. It is important to note that the following argument holds for any such ranking, provided it is independent of T.

The retrieved set of the target method is the shortest prefix P of the ranking that contains T:

$$P = \{d | rank(d) \leq \max_{d' \in T} rank(d')\} .$$

Now consider the ranking $relrank(d)$ of only relevant documents:

$$relrank(d) = \left|\left\{d' \in C | rel(d) \wedge rank(d') \leq rank(d)\right\}\right| .$$

The last retrieved document d_{last} is necessarily in T:

$$d_{last} = \arg\max_{d \in T} rank(d) = \arg\max_{d \in T} relrank(d) .$$

The recall of our method is:

$$recall = \frac{relrank(d_{last})}{R} .$$

Taking T to be a random variable, the method is reliable if:

$$R \leq 10 \vee \Pr[\frac{relrank(d_{last})}{R} \geq 0.7] \geq 0.95 .$$

Assuming large R, consider the problem of determining a cutoff c such that:

$$\Pr[\frac{R - relrank(d_{last})}{R} > c] = 0.05 \quad (1)$$

$$\Pr[R - relrank(d_{last}) > cR] = 0.05 \quad (2)$$

For the condition in Equation 2 to hold, it must be the case that the [numerically] top-ranked cR documents are absent from T, which occurs with probability

$$\left(1 - \frac{10}{R}\right)^{cR} = 0.05 .$$

It follows that:

$$c = \frac{\log 0.05}{R \log\left(1 - \frac{10}{R}\right)} .$$

For all $R > 10$,

$$c < \lim_{R \to \infty} \frac{\log 0.05}{R \log\left(1 - \frac{10}{R}\right)} = 0.299573 < 0.3 . \quad (3)$$

Combining (1) and (3), we have:

$$\Pr[\frac{R - relrank(d_{last})}{R} > 0.3] < 0.05$$

$$R \leq 10 \vee \Pr[\frac{relrank(d_{last})}{R} \geq 0.7] \geq 0.95 \; \square.$$

Reliability is obtained at the cost of supplemental review effort inversely proportional to R, the number of relevant documents. The number of randomly selected documents that need to be reviewed to find k relevant ones is $k\frac{|C|}{R}$, on average for $R \ll |C|$. For $k = 10$ and prevalence $\frac{R}{|C|} \approx 1\%$, the target method entails a review overhead of about 1,000 additional documents. Lower prevalence entails substantially more overhead, while higher prevalence entails less.

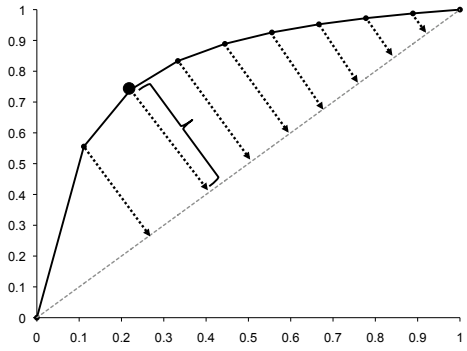

Figure 1: Knee Detection [15].

4. EMPIRICAL RELIABILITY

Our *knee method* relies on the assumption that CAL, in accordance with the probability-ranking principle, succeeds in ranking more-likely relevant documents before less-likely relevant documents. As a consequence, the gain curve plotting recall versus rank is assumed to be generally convex, with high slope (*i.e.*, "marginal precision") at the outset, and near-zero slope once nearly all relevant documents have been retrieved.

An ideal gain curve would have slope 1.0 until an inflection point corresponding to the rank at which all documents had been retrieved, and slope 0.0 thereafter. An actual gain curve typically diverges from the ideal due to limitations in probability ranking, random factors, and a noisy gold standard. Suppose that the retrieval method were able to achieve 70% recall and 70% precision at some rank r, as is typical for modern classifiers [11], or as might be achieved by exhaustive manual review [19]. The slope, up to that rank ($slope_{<r}$), would be 0.7, and the slope after that rank ($slope_{>r}$) would approach, but not equal 0. For small values of $R \ll |C|$, we would expect $slope \approx 0.0$, and for all R we would expect the "slope ratio" $\rho = \frac{slope_{<r}}{slope_{>r}} \gg 1$.

Based on our experience with non-public datasets, we observed that for $R \gtrsim 100$, $\rho \gg 6.0$ (with suitable smoothing) was a good indicator of high recall, achieving recall and reliability that compared favorably to that achieved by the target method. We formed the hypothesis that these thresholds were universal; that the same threshold would work for a wide variety datasets, including the ten that we subsequently used for our empirical evaluation (*see* Table 1).

4.1 Noise Abatement

If we were to stop at the minimum rank s, such that there exists an inflection point $1 < i < s$ such that $\rho \geq 6$, we would almost certainly stop prematurely due to chance. Moreover, this naïve approach would entail quadratic computational effort as a function of the size of the collection. To avoid both eventualities, we evaluate ρ only at values of s arising from the batches of documents selected by BMI. The number of batches is proportional to $\log |C|$, as the values of s are separated by an exponentially increasing interval. Relatively few of the candidate values for s will be viable, even by chance. Any residual sequential-testing bias is offset by a conservative choice of threshold for ρ.

For each value of s, we evaluate ρ at only one i, chosen using a geometric "knee-detection" algorithm [15], illustrated

Collection	Target Method			Knee Method			Budget Method		
	Reliability	Recall	*Effort*	Reliability	Recall	*Effort*	Reliability	Recall	*Effort*
At Home	1.00	0.91	*44,079*	0.93	0.93	*5,244*	0.97	0.97	*43,896*
Kaine	1.00	0.92	*119,644*	1.00	0.98	*172,774*	1.00	0.98	*172,774*
MIMIC II	1.00	0.89	*14,440*	1.00	0.97	*19,387*	1.00	0.97	*19,418*
RCV1-v2	0.96	0.88	*83,412*	0.99	0.94	*60,645*	0.99	0.94	*70,601*
Filtering	0.98	0.93	*133,788*	1.00	0.99	*3,857*	1.00	1.00	*143,798*
Intersection	1.00	0.92	*174,415*	0.98	0.96	*153,638*	1.00	0.99	*190,671*
Robust-04	0.89	0.94	*169,907*	1.00	1.00	*7,575*	1.00	1.00	*162,673*
Robust-05	1.00	0.92	*155,405*	1.00	0.96	*8,444*	1.00	1.00	*134,719*

Table 2: Reliability, Mean Recall, and Mean Effort for the Target, Knee, and Budget Methods.

in Figure 1. We draw a line from the origin to the recall achieved at rank s, and compute the maximum perpendicular distance from this line to the gain curve. Our candidate value of i is the projection to the x-axis of the intersection between the perpendicular and the gain curve. Our rationale in choosing this point was that it would correctly choose the inflection point for an ideal curve, and would avoid anomalies associated with points very close to the origin or to rank s, while capturing our intuitive notion of a genuine tipping point.

We calculated the slope ratio as:

$$\rho = \frac{\frac{|\{d|rank(d)\leq i \wedge rel(d)\}|}{i}}{\frac{1+|\{d|i<rank(d)\leq s\}|}{s-i}} .$$

Smoothing was accomplished by adding 1 to the number of relevant documents beyond the knee. This choice avoided the singularity of no relevant documents beyond the knee, and generally penalized situations in which the chosen inflection point was close to s. No smoothing was applied to the numerator, as we were not concerned with occasional underestimates.

4.2 Adjustment for Low Prevalence

The case of $R \lesssim 100$ is more problematic. Any correction for small R faces a dual problem:

1. the stopping procedure has no knowledge of the value of R, other than what can be estimated through relevance feedback from retrieved documents; and,

2. even if it were known that R was small, the sparsity of relevant documents compromises the reliability of our slope-ratio calculation.

The knee method relies entirely on the slope-ratio test, adjusted to compensate for low R. Initial tuning on the training collections from the TREC 2015 Total Recall Track indicated that a fixed lower bound β on the rank at which to stop, might be effective. For our submission to the TREC 2015 Total Recall At Home task, we conducted a parameter sweep of six combinations of $\beta \in \{100, 1000\}$ and $\rho \in \{3, 6, 10\}$. Our results showed that combinations involving $\beta = 100$ or $\rho = 3$ were unreliable, and we eliminated them from further consideration. Unsurprisingly, the combination of $\beta = 1000$, $\rho = 10$ proved most reliable, achieving the recall target for 29 of 30 topics (*reliability* = 0.97 [0.830 − 0.999 95% c.i.]).

We observed that recall and reliability appeared to be lower for smaller R, while effort (especially for $\rho = 10$) appeared to be disproportionately higher for large R. This observation led us to seek more reliable methods for small R, and to choose $\rho = 6$ for large R. To aid in this endeavor, we used a non-public dataset consisting of about 300,000 documents reviewed by attorneys and labeled according to 63 criteria, with R ranging from 5 to 164,000 (median 431). Based on tuning experiments using this dataset, we calibrated the slope-ratio cutoff as a function of *relret*, the number of relevant documents retrieved at any given rank:

$$\rho = 156 - \min(relret, 150) .$$

In other words, we set the threshold for the slope ratio to be 150 when no relevant documents have been retrieved, 6 whenever at least 150 relevant documents have been retrieved, and use linear interpolation between these values.

We further observed that with this adjustment, the choice of $\beta = 100$ versus $\beta = 1000$ became less critical. The lower value occasionally achieved lower effort than the higher value, and occasionally failed when the higher value did not. We chose to retain the value of $\beta = 1000$ from our earlier experiments.

4.3 Effort Adjustment

A variant of our knee method—the *budget method*—adjusts for small R by stopping only when a review budget comparable to that of the target method has been expended, and the slope-ratio test $\rho \geq 6$ is also satisfied. This adjustment substantially delays termination for small R, thus ensuring reliability.

The approach is predicated on the hypothesis that the supplemental review effort entailed by the target method would be better spent reviewing more documents retrieved by CAL. The target method entails the supplemental review of about $\frac{10|C|}{R}$ documents in order to find 10 relevant ones. According to the probability-ranking principle, we would expect CAL to find more relevant documents than random selection, for any level of effort, up to and beyond $\frac{10|C|}{R}$.

While the supplemental documents retrieved by the target method provide a statistical estimate of R, the documents retrieved by CAL provide a lower bound for R, and therefore an upper bound for the expected effort entailed by the target method. At the outset, this upper bound is loose, but as the review progresses, it tightens. The budget method retrieves documents using CAL until review effort exceeds this upper bound *and* $\rho \geq 6.0$, or until $0.75|C|$ documents are retrieved.

For small R, the budget determines the stopping point. For large R, enough relevant documents will likely be discovered to bound the review budget to an insubstantial fraction of R, and the slope-ratio test will determine when to stop. In any event, the review stops at $0.75|C|$. This final cutoff is predicated on the probability-ranking principle: random

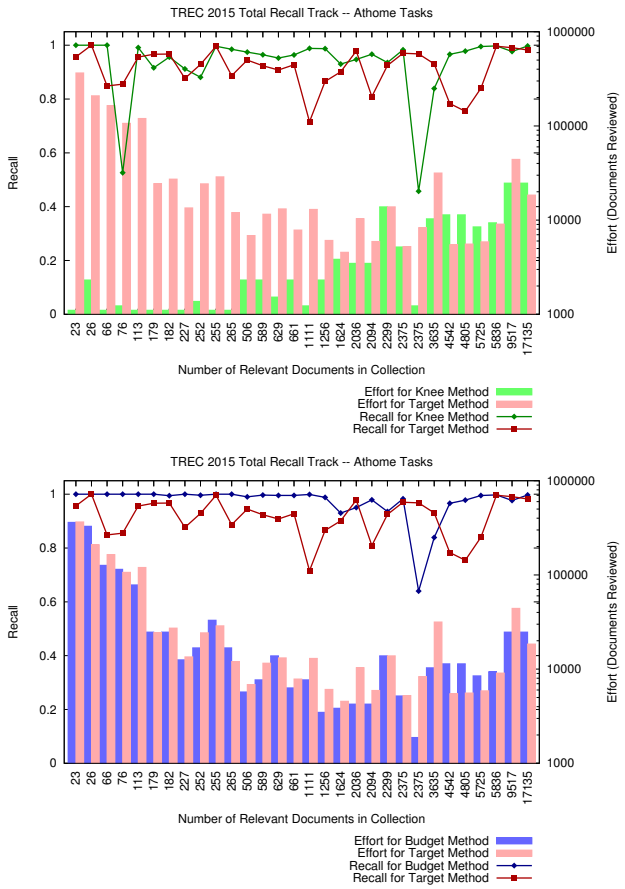

Figure 2: TREC 2015 Total Recall At Home Collection.

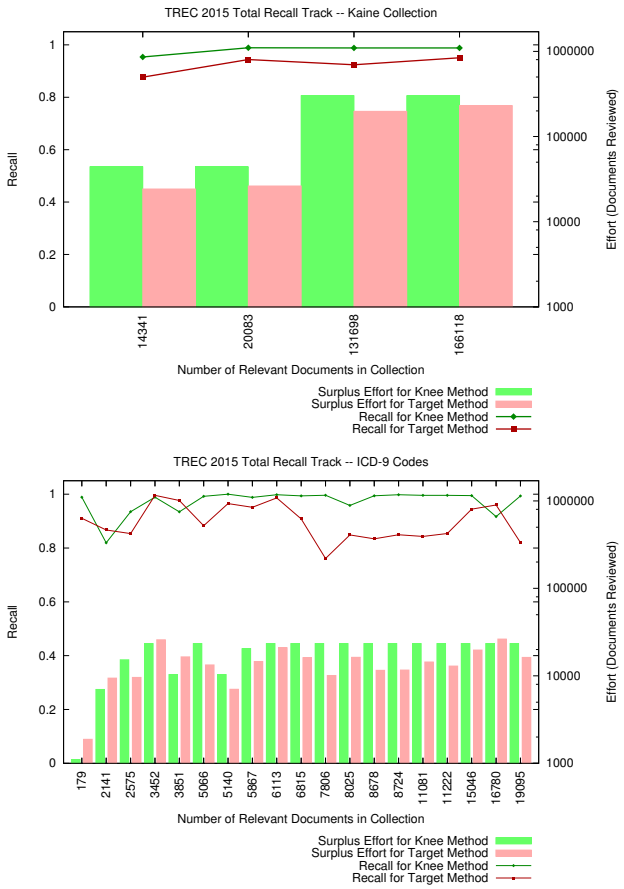

Figure 3: TREC 2015 Total Recall Sandbox Collections.

selection of 75% of the collection would, with high probability, achieve 70% recall; the top-ranked 75% should achieve even higher recall.

5. EXPERIMENTS

Testing the reliability of our stopping methods occasioned the use of "fully assessed" test collections, with a large number of topics and documents, where by "fully assessed," we mean that the pooling method, ISJ, or a rule base was used, and the resulting documents were labeled by a human assessor. From the limited number of collections that met these criteria, we selected the TREC 2015 Total Recall Track collections, the Reuters RCV1-v2 news dataset, the TREC 2002 Filtering Track collections, and the TREC 2004 and 2005 Robust Track collections, as detailed in Table 1. We used our Total Recall At Home participation to conduct an initial parameter sweep with six combinations, as well as final testing; the other datasets were used solely for testing.

The first phase of our experiments took place within the context of the TREC 2015 Total Recall Track, which had three distinct phases: training, At Home, and Sandbox. We conducted our initial development and tuning during the training phase, and submitted the knee method for evaluation in the At Home phase, but not the Sandbox phase. We captured the sequence of documents retrieved by BMI in both the At Home and Sandbox phases, and later used them to simulate the effect of the stopping methods whose results are presented here. After conducting further tuning on our non-public collection of 300,000 documents with 63 topics, we froze all parameters, and ran BMI on the other evaluation datasets, capturing the order in which the documents were retrieved. We then simulated our stopping methods by applying them to the ranking.

Summary results showing reliability, average recall, and average effort for all collections are shown in Table 2. The overall reliability of the target method, the knee method, and the budget method are substantially higher than the target of 0.95. Considering reliability, alone, there is little to choose among the methods; but the recall achieved by the knee and budget methods is substantially higher, while the effort expended by the knee method is, for some datasets, dramatically lower.

As illustrated in Figures 2 through 6, R (the number of relevant documents) appears to be the principal determinant of effort. For small R, effort for the target and budget methods approaches the size of the collection, while effort for the knee method, with one notable exception, generally diminishes with R, approaching the floor of $\beta = 1000$ that we chose for this study. On the other hand, for large R, the effort for all methods appears proportional to R.

The top panel of Figure 2 compares recall and effort for the knee and target methods, for each topic in the At Home col-

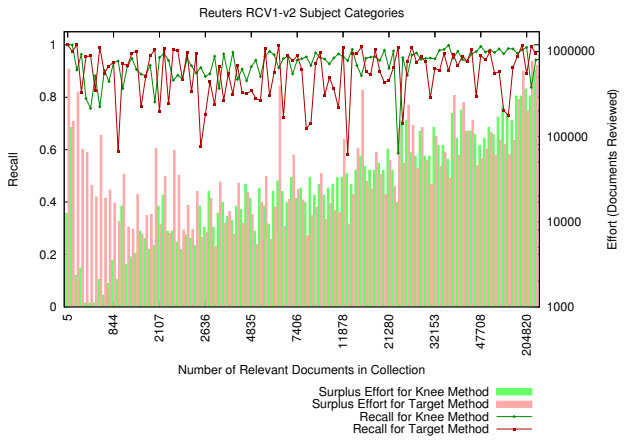

Figure 4: Reuters RCV1-v2 Subject Codes.

lection, ordered by R. We see that 28 of the 30 recall points for the knee method (shown by the green curve) fall above 0.7, indicating reliability of 0.93, while all of the points for the target method (shown by the red curve) fall above 0.7, indicating reliability of 1.00 for this collection. We also see that the most of the recall points for the knee method fall above those for the target method, indicating higher median recall, and the (signed) area between the curves is positive, indicating higher mean recall. Per-topic effort is shown as a bar graph on a logarithmic scale spanning three orders of magnitude. For small R, the knee method entails about 100 times less effort than the target method, while for large R, the effort is comparable.

The bottom panel of Figure 2 follows the same format, comparing the budget method (shown in blue) to the target method (shown in red). While the budget method achieves higher recall than the target method for nearly all topics, that superiority is not reflected in higher reliability. Effort for the two methods is very similar. The same observations apply to the results for the other collections: For low R, recall for the budget method exceeds that of the target method, while effort is indistinguishable; for large R, recall and effort are indistinguishable from the knee method. Both methods are reliable.

For brevity, we show graphical results comparing only the knee and target methods for the other collections. Tabular results for all methods are presented in Table 2.

Figure 3 shows results for the Sandbox task of the TREC 2015 Total Recall Track, which was notable in that participants had no prior access to the datasets or the topics, and their retrieval systems had to run fully autonomously. The top panel shows our results for the Kaine collection, which consisted of about 400,000 documents from Tim Kaine's eight-year tenure as Governor of Virginia. These documents had been previously reviewed and labeled by the archivist at the Library of Virginia according to four statutory categories: "record" (versus "non-record"), "open record," "restricted record," and "pertaining to the Virginia Tech shooting." Two of the topics had moderately high $R \approx 10^4$, and two had very high $R \approx 10^5$. For all topics, the knee method achieved higher recall at the expense of somewhat higher effort. The bottom panel shows our results for the MIMIC II collection, which consisted of about 30,000 medical records

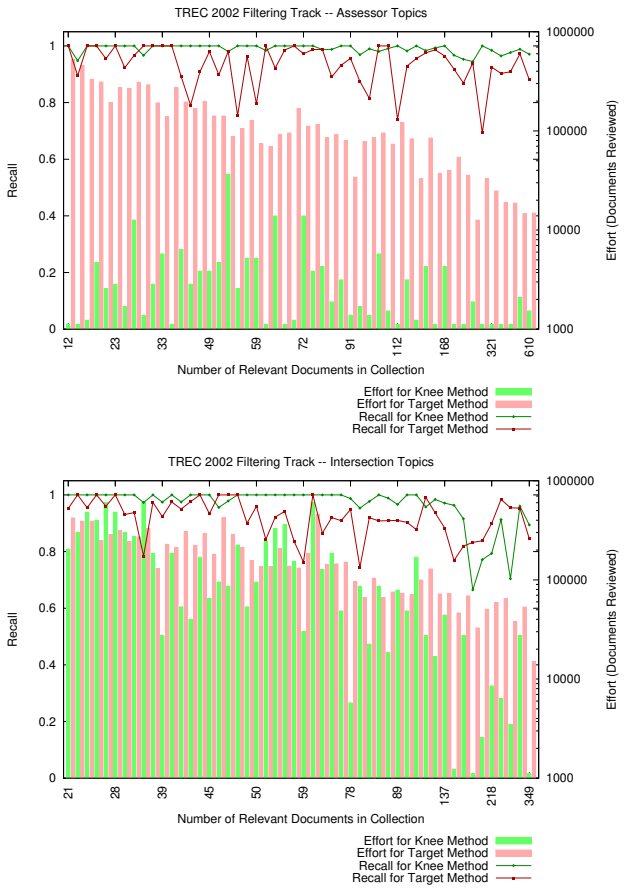

Figure 5: TREC 2002 Filtering Track Collections.

collected from a hospital intensive care unit. The documents consisted of nurses' notes, radiology reports, and discharge summaries. The "topics" consisted of ICD-9 diagnostic codes extracted from non-textual database records. With one exception ($R = 179$), all topics had moderately high R. The knee method generally achieved higher recall than the target method, at the expense of somewhat higher effort for most topics.

Figure 4 shows the results for the RCV1-v2 dataset, using the subject categories and descriptions published with the dataset as topics [11]. Over a very wide range $10^1 \lesssim R \lesssim 10^5$, we observe a familiar pattern: The knee method has somewhat higher recall and lower variance, with dramatically lower effort, for small R.

Figure 5 shows results for two sets of topics created for the TREC 2002 Filtering Track. The top panel shows results for topics that were created and assessed by NIST for the track. All topics had low $R \leq 610$; the majority had very low $R \leq 100$. For all topics, including those with the lowest $R \ll 100$, the knee method achieved near-perfect recall. Recall for the target method showed much higher variance, suggesting that its reliability is actually lower. The knee method entails order(s) of magnitude less effort. The lower panel shows results for intersection topics, each of which was the conjunction of two RCV1-v2 subject categories. If $rel_1(d)$ and $rel_2(d)$ indicate relevance for two RCV1-v2 topics, $rel_1(d) \wedge rel_2(d)$ indicates relevance for the intersection

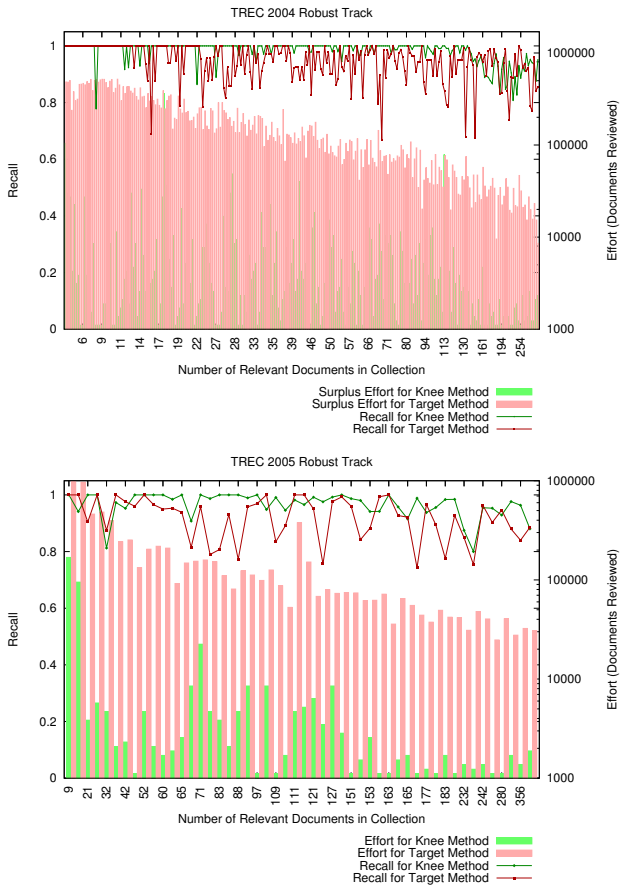

Figure 6: TREC 2004-2005 Robust Track Collections.

topic. The intersection topics were reported as a failed experiment [17], since no system achieved reasonable results on them. The results show that, while the effort to achieve high recall for these anomalous topics is inordinately large, our stopping methods are reliable.

Figure 6 shows results for the TREC 2004 and 2005 Robust tracks. In 2004, the Robust Track aggregated 150 topics developed for the TREC 6, TREC 7, and TREC 8 Ad-Hoc tasks, 50 topics developed for the 2003 Robust Track, and 49 new topics, for a total of 249 topics. For 2005, 50 of these topics—those deemed to be "difficult"—were reprised with a new dataset. The top panel reports our results for 2004; the bottom for 2005. The results further confirm that the target and knee methods both achieve high reliability, while the knee method entails dramatically less effort.

6. DIMINISHING LOSS

As evidenced by the results above, reliability does not capture certain important aspects of effectiveness or efficiency. Moreover, empirical measurements of reliability lack statistical power, while parametric estimates depend on assumptions regarding the distribution of recall values. Since the choices of acceptable recall and acceptable reliability are both somewhat arbitrary, bias due to incorrect distributional assumptions may be of little consequence. We suggest that reporting the mean μ and standard deviation σ of recall conveys more useful information, if not a provably accurate estimate of reliability. Such an estimate would have to be compared to one or more tacit thresholds to determine the reliability of the method; for example, assuming normality, any pair of μ and σ such that $\mu - 1.64\sigma \geq 0.70$ would be 95% reliable. More generally, the value of $Q = \mu - 1.64\sigma$ is a quantitative measure of quality, which may be used to determine the threshold level of acceptable recall for which 95% reliability may be obtained. Alternatively, by substituting the appropriate z-score in place of 1.64, a threshold of reliability different from 95% may be tested.

We suggest that reliability and recall should be supplanted by quality estimates based on loss functions, of which recall and reliability are special cases. We define $Q = 1 - \overline{loss}$, where $\overline{loss}$ is the mean value of a loss function over all topics. If

$$loss = 1 - recall\,,\ Q = \overline{recall}\,;\ \text{if,}$$

$$loss = \begin{cases} 0 & (recall \geq 0.7) \\ 1 & (recall < 0.7) \end{cases}\,,\ Q = reliability\,.$$

A quadratic loss function such as:

$$loss_r = (1 - recall)^2 \tag{4}$$

captures the desirability of consistently high recall, subsuming the roles of μ and σ in the previous discussion. Our aspirational goal is to achieve 100% recall. Any shortfall is penalized, and larger shortfalls are penalized more heavily.

Quadratic loss further generalizes to other aspects of quality, such as graded relevance, facet relevance [6], and efficiency. For example, let a_1, a_2, ..., a_n be categories of relevance, and $rel_a(d)$ be the indicator function for category a. Define:

$$recall_a = \frac{|\{d \in C | relret(d) \wedge rel_a(d)\}|}{|\{d \in C | rel_a(d)\}|}$$

$$loss_a = (1 - recall_a)^2$$

$$loss = \sum_{i=i}^{n} \alpha_i loss_{a_i}\,,\ \text{where } 1 = \sum_{i=1}^{n} \alpha_i\,. \tag{5}$$

The choice of weights α_i is not critical; the value $\alpha_i = \frac{1}{n}$ for all α_i will often suffice, as it rewards consistent recall over all categories, with the effect that documents belonging to rarer categories are afforded more influence.

Review effort may also be modeled as a category of loss, thus quantifying the notion of "reasonable effort." For the problem as we have framed it, an ideal method would entail $effort = R$. From the presentation of results in the TREC 2015 Total Recall Track Overview [13], we observe that a reasonable effort might entail $effort = aR + b$, where $a \approx 1$ represents effort proportional to sunk review cost, and $b \approx 0$ represents fixed overhead. We suggest that $a \leq 2$ and $b \leq 1000$ represent reasonable effort to achieve $recall \geq 0.70$ with 95% reliability. The use of a quadratic loss replaces the a and b thresholds by a soft target:

$$loss_e = \left(\frac{b}{|C|}\right)^2 \left(\frac{effort}{R + b}\right)^2\,. \tag{6}$$

$loss_e$ may be used to measure efficiency in its own right, or treated as a category loss in (5).

Collection	Target Method			Knee Method			Budget Method		
	$\sqrt{loss_r}$	$\sqrt{loss_e}$	$\sqrt{loss_{re}}$	$\sqrt{loss_r}$	$\sqrt{loss_e}$	$\sqrt{loss_{re}}$	$\sqrt{loss_r}$	$\sqrt{loss_e}$	$\sqrt{loss_{re}}$
At Home	0.0132	0.0090	0.0111	0.0197	**0.0000**	0.0099	**0.0056**	0.0108	**0.0082**
Kaine	0.0815	**0.0016**	0.0577	**0.0252**	0.0025	**0.0179**	**0.0252**	0.0025	**0.0179**
MIMIC II	0.1229	**0.0734**	0.1012	**0.0516**	0.0862	**0.0710**	**0.0516**	0.0866	0.0712
RCV1-v2	0.1475	0.0883	0.1216	0.0947	**0.0154**	**0.0678**	**0.0824**	0.0795	0.0809
Filtering	0.1011	0.2110	0.1654	0.0181	**0.0079**	**0.0140**	**0.0015**	0.2278	0.1611
Intersection	0.1057	**0.2499**	**0.1919**	0.0818	0.2740	0.2022	**0.0159**	0.2947	0.2087
Robust-04	0.0870	0.4136	0.2989	0.0430	**0.0481**	**0.0456**	**0.0025**	0.3865	0.2733
Robust-05	0.1141	0.2368	0.1858	0.0570	**0.0265**	**0.0445**	**0.0087**	0.1843	0.1305

Table 3: Root Mean Loss for Relevance, Effort, and Combined.

Target Method		Knee Method		Budget Method	
$\sqrt{loss_r}$	$\sqrt{loss_h}$	$\sqrt{loss_r}$	$\sqrt{loss_h}$	$\sqrt{loss_r}$	$\sqrt{loss_h}$
0.0837	0.0504	0.0134	0.0021	0.0007	0.0011

Table 4: Root Mean Loss for Relevance and High Relevance.

In Table 3, we report, for each collection, the root mean loss ($\sqrt{loss}$) over all topics for relevance loss as defined in (4); effort loss as defined in (6); as well as their unweighted average, $loss_{re} = 0.5 \cdot loss_r + 0.5 \cdot loss_e$. The results show conclusively the superiority of the budget method in terms of $loss_r$. They show the general superiority of the knee method in terms of $loss_e$, while calling to our attention three collections where the target method is more efficient. On inspection, we see that two of the three collections have exclusively or nearly exclusively topics with high prevalence. We further see that that the the target method's narrow margin of superiority in terms of $loss_r$ is offset by a wide margin of inferiority in $loss_e$, as reflected in $loss_{re}$. For the intersection collection, no system achieved acceptable $loss_e$.

The bottom line is that the quality loss results support our qualitative observation that the knee method affords the best balance between consistently high recall and consistently low effort; the budget method provides consistently higher recall at the expense of disproportionate effort for topics with few relevant documents; the target method, while provably reliable, yields empirical results that are generally inferior to the knee and budget methods.

To illustrate the use of quality loss for graded relevance, we used a subset of 84 topics from Robust-04, for which relevance assessments were available for the categories "highly relevant" and "relevant." Table 4 shows $\sqrt{loss_r}$ and $\sqrt{loss_h}$ for these categories, respectively. The knee and target methods have lower $loss_h$, than $loss_r$, indicating they retrieve highly relevant documents more consistently than merely relevant documents. The budget method shows the opposite effect, but even so, is markedly superior to the target and knee methods. While we cannot draw any firm conclusions from this small experiment, the results do not support the proposition that TAR methods achieve high recall by "bulking up" on marginal documents at the expense of important ones (*cf.* [12]).

7. DISCUSSION

To our knowledge, the target method is the first provably reliable method for TAR. The commonly used frequentist acceptance test (*see* [1, 9]) offers a p-value or confidence level which is a measure of the reliability of the *test*, not the reliability of the TAR method, not the probability that a given result is acceptable, and not the probability that a TAR method will pass the acceptance test. In eDiscovery, it is common to calculate a frequentist recall estimate, with a 5% margin of error and 95% confidence, and deem the result acceptable if the estimate exceeds 75%. Calculating such an estimate requires a sample of about 385 random relevant documents, entailing 38.5 times as much surplus effort as the target method.

Our proof of reliability does not require that the target sample T be chosen at the outset, as long as it is independent of the retrieval method. The target method could be used as an acceptance test, such that the consequence of failing the test would be to continue to retrieve documents without knowledge of T, until all the documents in T are retrieved.

Over test collections like the ones used in this study, there can be little doubt ($p \approx 0.00$) that the knee and budget methods are reliable, that the budget method is more reliable than both the knee and target methods, and that the knee method is the most efficient. As with any empirical work, the test collections constitute convenience samples and ongoing research is necessary to characterize the scope of TAR tasks to which our results may be generalized.

The target method is reliable regardless of the underlying review method; however, if the underlying method uses a human in the loop to formulate queries or to influence the selection of documents in any way, that human must be isolated from any knowledge of T. The simplest approach to accomplish this goal might be to complete all such human intervention before drawing T, and to rely on fully automated document selection thereafter. An alternative would be to establish an "information barrier" between those who draw T and those who conduct the search.

This work establishes the reliability of the knee and budget methods as applied to BMI. It remains to be determined how well these approaches would work—possibly with different tuning parameters—for other CAL methods, including hybrid systems in which a human is afforded influence in the selection of documents for review. It is not obvious how to adapt the knee or budget method to SPL or SAL, for which an essential question is when to stop training.

The target method, by design, targets less than 100% recall. It could be modified to continue past the point at which

the last document in T is retrieved, thereby expending additional effort to increase the probability of achieving 100% recall. One might, for example, extrapolate from the distribution of $rank(d \in T)$. The knee method, on the other hand, does target 100% recall, and only incidentally optimizes reliability. It appears that loss functions better characterize the tension among consistency, effectiveness, and efficiency, as compared to goal-post methods. Regardless of which measure is chosen for evaluation, systems should be tuned to optimize their suitability for their intended purpose, not the measure itself (*cf.* [22]).

8. CONCLUSIONS

Reservations about the effectiveness and reliability of TAR have impeded its adoption for eDiscovery and other high-recall retrieval tasks. A primary area of concern has centered on the issue of "when to stop," or knowing with reasonable certainty—and being able to show an adversary or the court—that a particular TAR effort has identified an acceptable amount of relevant information. Many approaches to validation in common use today are simply invalid, or require disproportionate effort compared to the information they yield, and are often misunderstood and misapplied [9, 16].

We offer a method to determine when to stop that is guaranteed to be reliable, for the price of reviewing a number of random documents that is an order of magnitude less than acceptance tests that estimate recall, but neither determine when to stop nor guarantee reliability. We provide two other methods that entail no effort beyond that required by the underlying TAR method and, while not providing a guarantee of reliability, consistently demonstrate better reliability, and better recall, when evaluated on eight test collections, comprising 555 topics and 4.5M documents. Of particular interest is the knee method which, in contrast to the other methods, is demonstrated to be reliable and efficient when the collection contains few relevant documents.

While our primary results are demonstrated using measures derived from traditional goal-post methods—binary relevance, a recall threshold, and a reliability floor—we describe how loss functions may be formulated to capture the tension among consistency, degrees of relevance, facets of relevance, and efficiency. We apply these formulae to show insights into our results that might not have been readily apparent from the goal-post measures.

9. REFERENCES

[1] M. Bagdouri, W. Webber, D. D. Lewis, and D. W. Oard. Towards minimizing the annotation cost of certified text classification. In *SIGIR 2013*.

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

[3] D. C. Blair. Stairs redux: Thoughts on the stairs evaluation, ten years after. *J. Am. Soc. Inf. Sci.*, 47(1):4–22, Jan. 1996.

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

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

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

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

[8] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *SIGIR 1998*.

[9] M. R. Grossman and G. V. Cormack. Comments on "The implications of Rule 26(g) on the use of technology-assisted review". *Fed. Cts. L. Rev.*, 7:285–312, 2014.

[10] C. Lefebvre, E. Manheimer, and J. Glanville. Searching for studies. *Cochrane Handbook for Systematic Reviews of Interventions*, 2008.

[11] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, 2004.

[12] D. Remus and F. S. Levy. Can robots be lawyers? Computers, lawyers, and the practice of law. http://dx.doi.org/10.2139/ssrn.2701092, 2015.

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

[14] M. Sanderson and H. Joho. Forming test collections with no system pooling. In *SIGIR 2004*.

[15] V. Satopää, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *ICDCSW 2011*.

[16] K. Schieneman and T. Gricks. The implications of Rule 26(g) on the use of technology-assisted review. *Fed. Cts. L. Rev.*, 7:239–274, 2013.

[17] I. Soboroff and S. Robertson. Building a filtering test collection for TREC 2002. In *SIGIR 2003*.

[18] G. Taguchi. *Introduction to Quality Engineering: Designing Quality Into Products and Processes*. 1986.

[19] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.

[20] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pages 143–170. Springer, 2002.

[21] E. M. Voorhees and D. K. Harman. The Text REtrieval Conference. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, pages 3–19. MIT Press, 2005.

[22] E. Yilmaz and S. Robertson. On the choice of effectiveness measures for learning to rank. *Information Retrieval*, 13(3):271–290, 2010.

[23] J. Zobel, A. Moffat, and L. A. Park. Against recall: Is it persistence, cardinality, density, coverage, or totality? In *ACM SIGIR Forum*, volume 43, pages 3–8. ACM, 2009.