

Assessor disagreement and document retrieval depth

Author Names Removed for Blind
Review

Undisclosed Location

ABSTRACT

Assessors are well known to disagree frequently on the relevance of documents to a topic, but the factors leading to assessor disagreement are still poorly understood. In this paper, we examine the relationship between the rank at which a document is returned by a set of retrieval systems and the likelihood that assessors will disagree on its relevance, and find that there is a strong and consistent correlation between the two. We adopt a metarank method of summarizing a document's rank across multiple runs, and propose a logistic regression predictive model of assessor disagreement given metarank and initially-assessed relevance. The consistency of the model parameters across different topics, assessor pairs, and collections is considered. The model gives comparatively accurate predictions of absolute scores, but less consistent predictions of relative scores than a simpler rank-insensitive model. We demonstrate that the logistic regression model is robust to using sampled, rather than exhaustive, dual assessment. We demonstrate the use of the sampled predictive model to incorporate assessor disagreement into tests of statistical significance.

Categories and Subject Descriptors

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

Keywords

Retrieval experiment, evaluation, sampling

General Terms

Measurement, performance, experimentation

1. INTRODUCTION

When two humans are asked to assess the one document for relevance to a topic, they are surprisingly likely to disagree. One study finds that even expert assessors with similar backgrounds have only a 60% probability of agreeing that a document is relevant [Voorhees, 2000]. Not just the set of actually relevant documents, but the measured reliability of a document retrieval, can

vary substantially depending upon which assessor is used; and in human-intensive document productions, the results of comparative evaluation can depend on whether you agree with the humans who developed the production, or the humans who assessed it [Oard et al., 2008].

While several studies observe assessor disagreement, and a few have investigated its impact upon evaluation, little work has been done on characterizing, modeling, and predicting disagreement. It is not known what share of disagreement is attributable to assessor inattention, what to differing relevance conceptions, and what to variable thresholds for detecting relevance. Identifying correlates of disagreement will help predict, adjust for, and correct assessor disagreement and its impact upon evaluation; and better determining the nature and causes of disagreement will enable preventative steps to be taken, and improve our understanding of the human perception of relevance.

In this paper, we examine a potential correlate of assessor disagreement, namely the ranks at which a document is retrieved by a set of retrieval systems. We summarize retrieval rank information across the run set using metarank scores. Working with the same TREC 4 and TREC 6 AdHoc datasets as Voorhees [2000], we then estimate the probability of disagreement given meta-rank score, developing separate logistic models for documents assessed relevant and assessed irrelevant by an initial assessor.

The relationship between metarank and assessor disagreement is found to be a strong one; a high metarank document assessed relevant by one assessor is almost twice as likely to be assessed irrelevant by a second assessor as a low metarank one, and the relationship is even stronger for documents assessed as irrelevant by a first assessor. The strength of the relationship, however, varies markedly between different topics; there are clearly other, topic-dependent factors at play in determining assessor disagreement, and a universal model will have limited predictivity.

Models of disagreement by rank can be used to create artificial document assessment sets (or qrels), to simulate and explore the effects of assessor disagreement. Qrels simulated from per-topic rank-sensitive models give much more accurate estimates of absolute scores under alternate assessment than do rank-insensitive flip-rate models. Rank-sensitive simulated qrels, however, provide less stable relative evaluations and system rankings than flip-rate qrels, at least for the meta-rank scoring method we use. Meta-rank score is evidently not independent of system, and (we hypothesize) the same reinforcement of like systems can be observed here as in other simulated relevance methods [Soboroff et al., 2001].

We demonstrate that rank-sensitive models based on sampled dual assessment produce absolute and relative score simulations almost as reliable as those of exhaustive dual assessment. Moreover, sampled rank-sensitive models are more reliable than sam-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'12, October 29–November 2, 2012, Maui, Hawaii, USA.
Copyright 2012 ACM XXXX ...\$10.00.

pled rank-insensitive ones even for relative evaluation. Thus, one can predict the effect of assessor disagree with only a fraction of repeat assessment effort.

Traditional tests of the significance of a comparative retrieval evaluation only consider the topics as variable, holding documents, assessors, and other aspects of the experiment fixed. With reliable, sample-based models of assessor disagreement available, variability in the assessor dimension can be simulated at a fraction of the cost of exhaustive multiple assessment. We demonstrate the use of our rank-sensitive model to determine evaluation significance with tests in which assessors can vary.

The remainder of the paper is structured as follows. Related work is surveyed in Section 2, and our materials and methods in Section 3. Section 4 describes our experiments, and Section 5 summarizes our findings and sketches future work.

2. RELATED WORK

The high level of inter-assessor disagreement on relevance has been noted by a number of studies. Voorhees [2000] examines multiple assessments of the TREC 4 and TREC 6 AdHoc collections (the same datasets used here), finding overlap of between 0.42 and 0.49 on TREC 4. Roitblat et al. [2010] find even lower levels of inter-assessor agreement on an e-discovery production. Bailey et al. [2008] survey earlier studies.

Several studies have concluded that impact of assessor disagreement upon the comparative (rather than absolute) evaluation of automated retrieval systems is minor. Voorhees [2000] finds a mean Kendall’s τ of 0.938 between system rankings produced by different assessors, suggesting a high degree of stability between assessment sets. Trotman and Jenkinson [2007] compare using multiple (non-overlapping) assessors per topic with a single assessor, and find a mean Spearman’s rank correlation coefficient of 0.986. The effect upon relative assessment may be greater for runs created with a large amount of manual involvement, however, including through training classifiers; the correction of assessor errors led to large relative score changes in the interactive task of the TREC Legal Track [Oard et al., 2008].

The interactive task of the TREC Legal Track corrects assessor errors through participant appeals and adjudication by a topic authority [Oard et al., 2008]. Webber et al. [2010] propose instead that assessments be sampled for authoritative verification, and error rates estimated from these samples. Sheng et al. [2008] investigate using multiple overlapping assessors in annotation tasks.

Cuadra and Katter [1967] identify five factors influencing perceptions of relevance: document variables; topic statement variables; judgment conditions; judgment scales; and personal factors. Saracevic [2007] surveys of experiments on these factors. Webber et al. [2010] present a taxonomy of topical grounds for appeal from the appeal documents submitted by a TREC Legal Interactive participant. Re-analyzing appealed assessments from the Interactive task, Grossman and Cormack [2011] conclude that the great majority of assessor error is due to inarguable failure to follow the task’s detailed relevance assessment guidelines; however, their sample is biased by that fact that participants only appealed assessments they felt to be inarguable errors. Webber et al. [2012] find that more detailed instructions do not lead to fewer assessor errors.

Aslam et al. [2005] present a meta-search approach known as meta-AP in which documents are weighted by their implicit average precision scores in each ranking; we adopt meta-AP as a predictor of assessor disagreement in our models. As part of an evaluation score estimation method known as minimal test collection (MTC), Carterette [2007] explicitly builds a multi-level logistic model of probability of relevance based on retrieval rank and system reli-

Orig	Alt 1		Alt 2		Total
	Rel	!Rel	Rel	!Rel	
Rel	4.2%	4.0%	5.1%	3.1%	8.2%
!Rel	4.8%	87.0%	7.5%	84.3%	91.8%
Total	9.0%	91.0%	12.6%	87.4%	

Table 1: Macro-averaged estimated proportional contingency tables between original and two alternatives assessors across the TREC 4 AdHoc topics. Agreement observed on sampled documents extrapolated to the rest of the pool.

ability. The relative values of the estimated scores are a reliable estimate of full assessment, but the absolute estimated values are not, suggesting that absolute probabilities of document relevance are misestimated.

Soboroff et al. [2001] explore randomly assigning relevance assessments to documents, and find that the system ranking that results is moderately correlated with the human-assessed ranking. Carterette and Soboroff [2010] use variable flip rate probabilities to simulate “conservative” and “liberal” assessors, finding that “conservative” assessor maintain stable system rankings, while “liberal” assessor disrupt them.

Voorhees [1998] introduced the use of Kendall’s τ as a measure of the stability of system ranking in the face of changes in the evaluation setup. Savoy [1997] proposes the use of Bootstrap significance tests in information retrieval evaluation. Bodoff and Li [2007] argue that choice of assessor should be included alongside choice of topics in assessing the generalizability of information retrieval evaluation results.

3. MATERIALS AND METHODS

This section describes our data sets and methods. For data (Section 3.1), we use the TREC 4 AdHoc collection, runsets, and qrels, including multiple assessments performed by TREC assessors; we also use the TREC 6 collection, with additional assessments performed by one of the track participants. We model assessor disagreement using logistic regression, with document rank as a predictor (Section 3.2). Our experiments randomly generate qrels following models built from the dataset, and investigate the stability of system evaluation using them (Section 3.3).

3.1 Materials

The TREC 4 AdHoc test collection consists of 49 topics. Documents for assessment were selected by depth-100 pooling. In addition to the 33 systems that ran on the full collection in the ad-hoc task, all of which were pooled, additional pooled documents were drawn from systems that ran on a subset of the collection, or ran in a different modality [Harman, 1995]. Only the 33 adhoc full-collection runs are included in this study. Initial assessment was performed by the author of the topic; we refer to this as the “original” assessor. Each topic was then re-assessed by two other TREC assessors, whom we refer to as the first and second alternative assessors. If there were more than 200 relevant or irrelevant documents in the pool, then 200 were uniform randomly sampled for re-assessment [Voorhees, 2000]. The average number of re-assessed relevant documents is 105, with 12 topics having 200 or more relevant documents; all topics have more than 200 irrelevant documents. Though Voorhees [2000] does not describe a systematic difference in the allocation of assessors as first and second alternate assessors, the first assessor finds an average of 17.3 more

Orig	Alt 1		Total
	Rel	!Rel	
Rel	2.4%	3.3%	5.7%
!Rel	2.1%	92.2%	94.3%
Total	4.5%	95.5%	

Table 2: Macro-averaged estimated proportional contingency tables between original and alternative (Waterloo) assessors across the TREC 6 AdHoc topics.

document relevant than the second assessor ($sd = 51.2$), which is statistically significant ($p = 0.02$ in two-tailed, pair t test). We follow Voorhees [2000] in dropping Topic 214, as the first alternative assessor found no documents relevant for it.

Table 1 gives macro-averaged proportional contingency tables between the original and the two alternative assessors for the TREC 4 AdHoc topics. We estimate agreement on the pool from agreement on the sample. In contrast, Voorhees [2000] calculates statistics on the sample only. Since a higher proportion of originally-relevant documents is sampled than of originally-irrelevant (a macro-average of 91.3% of the former, versus 13.6% of the latter), agreement on the sample tends strongly to overstate agreement on the pool. So, for instance, Voorhees [2000] quotes a (sample) positive overlap with the original assessor of 0.421 for the first and 0.494 for the second alternative assessor, whereas the estimated figures on the population are 0.295 and 0.350, respectively.

The TREC 6 AdHoc multiple-assessment dataset was created as part of a run production by the TREC participant team from the University of Waterloo [Cormack et al., 1997]. The run was produced by interactively developing queries, retrieved ranked results, and judging of top-ranked documents. The documents assessed as part of this process constitute the “alternative” assessments to the (subsequently formed) official assessments of the TREC assessors. The Waterloo assessors used a three-level relevance scale, with a middle grade of marginally relevant; following Voorhees [2000], we collapse the marginally relevant documents to not relevant.

Besides being performed by assessors with a different background and mode of operation from the official assessors, the TREC 6 alternative assessments differ from those from TREC 4 in that they are not a random sample. Rather, they are the documents returned at top ranks by a series of interactive queries. The non-randomness of the selection of the alternative assessments makes it impossible to extrapolate agreement measures to the full TREC pool, and may bias models built upon dual-assessed documents. The proportional macro-averaged contingency table for the TREC 6 dataset is shown in Table 2. The macro-averaged overlap is 0.328.

3.2 Modeling assessor disagreement

We propose to model assessor disagreement as a function of document rank. This requires three components: a statistic for summarizing rank information (Section 3.2.1); a modeling method (Section 3.2.2); and a way of choosing which statistic provides the best fit for the data under the modeling method adopted (Section 3.2.3).

3.2.1 Meta-rank measures

A meta-rank measure summarizes the ranks at which a document is returned across a set of runs for a single topic. Let N be the rank to which meta-rank measures are scored (for instance, the retrieval sizes). We say that document returned in a run s at rank k has an inverse rank I in that run of $N - k$, or 0 if $k > N$. Two simple

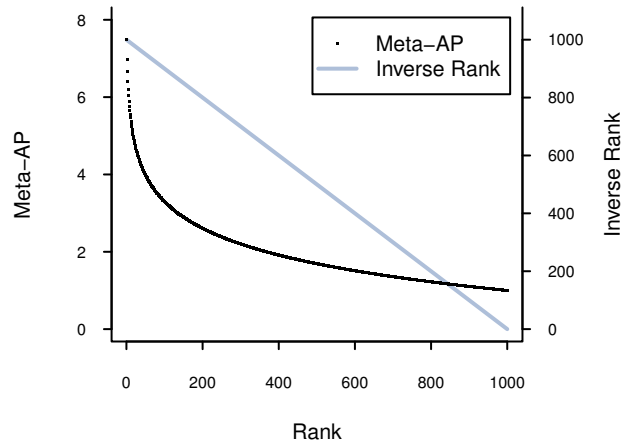


Figure 1: Meta-rank weights for Meta-AP and Inverse Rank weighting schemes, for a single document ranking.

measures are then maximum and mean inverse rank.

Aslam et al. [2005] propose a meta-rank measure based upon the weighting of the average precision (AP) metric. For AP evaluating to depth N , the implicit weight of a document at rank k is $1 + H_N - H_k$, where H_n is the n 'th harmonic number, or 0 if $k > N$. The mean meta-AP score for a document is its average AP weight across the set of runs.

We take $N = 1,000$, the run depth in the ad-hoc tracks of TREC, as our meta-rank evaluation depth. The meta-AP score for rank 1 in evaluation to this depth is 7.5. The relationship between meta-AP and inverse rank as a function of rank is shown in Figure 1.

3.2.2 Logistic regression

We predict the probability that a document will be judged relevant by assessor B , based upon its meta-rank and the fact that assessor A has judged it either relevant or irrelevant: $p(B = 1|s, A = r)$, building separate models for $A = 1$ and $A = 0$. It is natural to apply a logistic regression to this problem:

$$p(B = 1|s, A = r) = \frac{e^{\beta_0 + \beta_1 s}}{1 + e^{\beta_0 + \beta_1 s}} \quad (1)$$

where the metarank score s is the predictor variable, and the probability p is the predicted value. The probability of disagreement is $p(B = 1)$ if $A = 1$, and $1 - p(B = 1)$ is $A = 0$.

The fitted value β_0 in Equation 1 is the intercept, which gives the log-odds of relevance when the score is 0, while β_1 is the score coefficient, which gives the change or “slope” in log odds of relevance for every one point increase in metarank. The slope gives the strength of the relationship between metarank and probability of relevance, while the intercept shifts the regression curve up or down the score axis. An intercept of -1 means there is a 1 : $e \approx 27\%$ chance of relevance when the score is 0. Conversely, a slope of 1 means that an increase of 1 in the score will take probabilities from 1 : 1 = 50% to $e : 1 \approx 73\%$.

A model can be built for each topic individually, or else alternative assessments and metarank scores from the full collection can be pooled into a single model. The degree to which a per-collection or “universal” model is a good approximation for per-topic models depends upon the strength of per-topic factors in influencing disagreement. The closer the per-collection model is to the per-topic models, the more likely it is that a generalized model can be

built, that is able to predict assessor disagreement on new collections based only on metarank scores.

A simpler, rank-insensitive model of assessor disagreement is the flip rate model, which notes the proportion of originally-relevant documents that the alternative assessor assesses as irrelevant, and vice versa. Taking the original assessor as the objective standard, the flip rate is a pair of proportions, giving the false-positive and false-negative rates for the alternative assessor.

3.2.3 Model goodness-of-fit

There are various ways to measure the goodness-of-fit of a logistic regression. A simple measure, found by [Hosmer et al., 1997] nevertheless to be more powerful than many more complex ones, is the unweighted residual sum-of-squares:

$$\hat{S} = \sum (y_i - \hat{\pi}_i)^2 . \quad (2)$$

We sum across all observations i ; y_i is the actual value of observation i (here, whether document i is relevant or not), and $\hat{\pi}_i$ is the estimated probability of relevance for an item with the predictor variables of item i (here, the meta-rank of document i). Note that Equation 2, applied to the meta-rank predictive model, sums goodness of fit at the metaranks of actual observations, rather than (say) evenly across the metarank range, or with greater weighting in areas that might be more important in particular applications.

3.3 Experimental methods

Our experiments involve generating qrels that simulate the assessments of an alternative assessor, and then examining the stability of system scores and rankings. To do this, we build a model of assessor disagreement from the original assessments, the metarank predictor, and a set of observed alternative assessments. The simulated alternative qrels are generated by applying the predictive model to the original qrels. Each original assessment is plugged into the model with its metarank score to generate a probability of assessor disagreement. An independent uniform random number U in the range $[0, 1]$ is then generated. If U is less than the flip probability, then document relevance is flipped in the simulated qrels; otherwise, the original document relevance is maintained. The same process is used for the metarank and the flip-rate models, except the latter takes no account of document metarank, and also for per-topic and universal models.

We compare the stability of our models when based on exhaustive and on sampled dual assessment. For the sampled dual assessment, n originally-relevant and n originally-irrelevant documents (where $n = 20$ is used in our experiments) are sampled, and the metarank and flip-rate models are built using just the samples. Uniform random sampling is used for both the flip-rate and the metarank models. Additionally, for the metarank model, we consider enforcing some degree of even spacing in the sample as ordered by document metarank. To do this, a stratified sample is drawn: the documents of one class (originally-relevant or originally-irrelevant) are divided into quintile bins, and $n/5 = 4$ documents are sampled from each bin.

In our experiments, we generate $s = 1,000$ simulated alternative qrel sets for each model, and calculate the MAP scores achieved by systems for each alternative qrel set. We report the means and the 2.5% and 97.5% percentiles of these MAP scores. Absolute change in system score is calculated by root mean squared error (RMSE) between the observed and all of the simulated MAP scores. Relative stability is calculated using mean Kendall’s τ between the system ranking under the observed alternative assessments and under the simulated qrels.

	$p(\cdot 1)$	$p(\cdot 0)$
Meta-AP vs Mean inverse rank	-0.055	<i>-0.040</i>
Meta-AP vs Max inverse rank	-0.565	-1.204
Mean inverse rank vs Max inverse rank	-0.510	-1.150

Table 3: Mean difference in sum of squares, across all TREC 4 AdHoc topics, for models of the first and second alternative assessors against the original assessors, between models based on the meta AP, mean rank, and max rank predictors, conditional on first assessor judging a document relevant or irrelevant. Negative values show that the first of the compared models gives a better mean fit than the second. Comparisons that are not statistically significant in a paired t test at level $\alpha = 0.05$ are marked in italics.

4. EXPERIMENTS

In this section we present a series of experiments on various aspects of the model and its use. We start with the choice of predictor (Section 4.1), then evaluate the fidelity of the model of disagreement (Section ??). We next use the model to simulate qrels for use in calculating MAP in TREC evaluation experiments (Section ??). We investigate the use of sampling dual assessments to fit better models (Section ??), and finally investigate the use of simulated qrels to estimate variance due to differences in assessor when testing significance (Section ??).

4.1 Choice of meta-rank measure

Section 3.2.1 described three alternative meta-rank measures: mean meta-AP; mean inverse rank; and maximum inverse rank. Table 3 assesses the goodness of fit of models built using these meta-rank measures, using the unweighted sum of squares (Section 3.2.3). The mean difference in unweighted sum of squares is calculated for the TREC 4 dataset, across models build for all 48 topics and both alternative assessors. The meta-AP and mean inverse rank are both significantly more predictive than maximum rank; this is to be expected, since the means summarize more information than the maximum. The meta-AP measure is slightly more predictive in both the given-relevant and the given-irrelevant conditions, though only for the given-relevant one is the difference statistically significant. We select meta-AP as the meta-rank measure for the remaining experiments.

4.2 Modeling rank and disagreement

The logistic regression model produces two parameters, the slope and the intercept. A positive slope indicates that higher-ranked documents are more likely to be assessed as relevant by the alternative assessor than lower-ranked ones. If we regard the observed documents as a sample from a larger population, then the model itself estimates the relationship on the population, and needs to be checked for significance. If it happens that all observed documents of an original-relevance class receive the same alternative relevance judgment, or even if all the relevant alternative assessments are to one side of all the irrelevant ones, as ordered by score, no proper model can be constructed, though 0 or 1 probabilities can be assigned to scores.

The number of per-topic models giving slopes of different directions and significances across both datasets is tabulated in Table 4. Over 80% of models show positive slopes (90% if improper models are ignored), and 60% of these are significant. None of the models for the randomly-sampled TREC 4 alternative assessments have a significantly negative slope, and only two of the determinately-sampled TREC 6 models have a significantly negative slope. In

Dataset	Condition	Positive		Improper	Negative	
		Sig	!Sig		!Sig	Sig
T4, alt1	$p(\cdot 1)$	23	19	2	5	0
	$p(\cdot 0)$	24	13	10	2	0
T4, alt2	$p(\cdot 1)$	21	24	0	4	0
	$p(\cdot 0)$	28	10	6	5	0
T6	$p(\cdot 1)$	22	20	3	3	2
	$p(\cdot 0)$	36	7	3	4	0
Total		154	93	24	23	2

Table 4: Number of per-topic models for TREC 4, both alternative assessors, and TREC 6 datasets, giving ($p < 0.05$) significantly and non-significantly positive, improper, and significantly and non-significantly negative slopes.

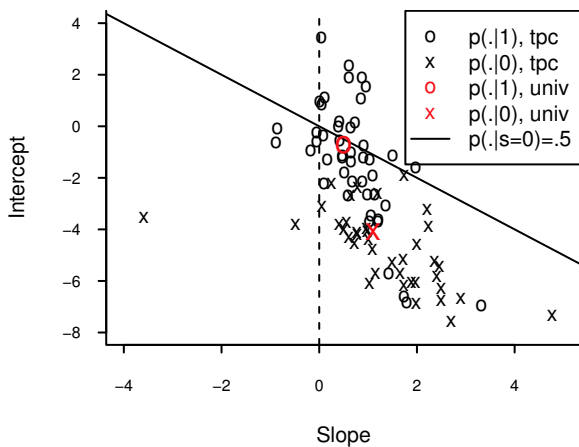


Figure 2: Per-topic and universal logistic regression coefficients for the first alternate assessor on the TREC 4 AdHoc dataset. The diagonal line shows the combination of intercept and slope for which the probability of alternative-assessor relevance at a meta-AP score of 0 is 50%.

summary, there is a strong and consistent relationship between document rank and the probability that an alternative assessor will disagree with an original assessor’s assessment.

The dispersion of slope and intercept coefficients for per-topic and universal models is shown in Figure 2, for the first alternative assessor of TREC 4. To aid interpretation, a diagonal guideline is also shown, above which coefficients assign $p > 0.5$ for a metarank score of 0 (that is to say, for a document ranked in the top N by at least one system). The higher the slope, the stronger the positive rank-sensitivity of alternative-relevance, while the higher the intercept for a given slope, the greater the estimated probability of relevance for each meta-AP score. Thus, model coefficients to the upper-left dominate those to the lower-right.

As a group, models for initial assessments of relevant dominate models for initial assessments of irrelevant, as one would expect: the alternative assessor is more likely to find a document relevant if the initial assessor did than if the initial assessor didn’t. Parameters are highly variable in both slope and intercept between topics, however; the universal model for each condition, though centered amongst the per-topic models, does not well characterize the spread

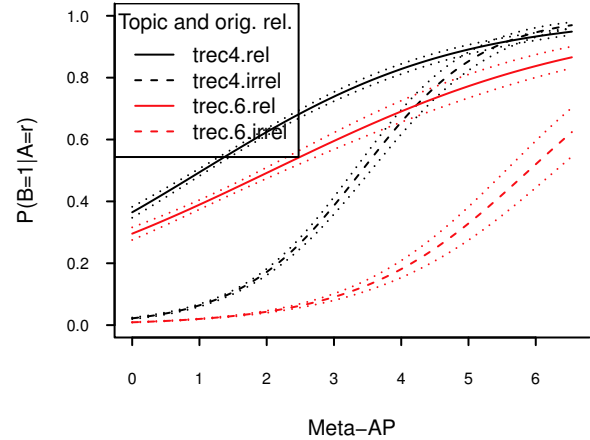


Figure 3: Universal logistic regression models for TREC 4 and TREC 6, with 95% error bounds. The TREC 4 model combines both alternative assessors.

of the per-topic models. Metarank alone (at least as measured by meta-AP) fails to capture all of the variance in assessor disagreement, even for the one collection; other, topic or assessor-pair, features also have an impact.

The universal regression models for TREC 4 and TREC 6 are shown in Figure ???. We have included both the first and the second alternative assessors in the TREC 4 model. A marked difference between TREC 4 and TREC 6 can be observed, particularly for the given-irrelevant model; this is likely due to the different selection and assessment methods for the TREC 6 alternative assessments. For the TREC 4 universal model, we can see that a relevant assessment on a high-ranking document is almost certain to be confirmed by a second assessor, whereas a relevant judgment on a low-ranking document is as likely as not to be overturned. The relationship is even stronger for documents initially assessed as irrelevant: the alternative assessor is very likely to agree with the initial assessment for low-ranked documents, but for the (rare) high-ranked documents initially assessed as relevant, the alternative assessor is very likely to disagree with the initial assessment.

Figure 3 gives regressions for three example topics. For Topic 215, the given-relevant model dominates the given-irrelevant one, and predicts over four times the probability of alternative-relevance for high metarank than for low metarank documents. For Topic 211, the given-irrelevant model has low intercept but high slope, meaning a sudden transition from strongly-irrelevant to strongly-relevant, based on a small number of alternative-relevant documents; smoothing would give a less sudden transition. For Topic 250, the given-positive model has a negative slope, assigning a lower probability of alternative relevance to higher ranked documents; the negative slope is based on a small number of alternatively-irrelevant observations, however, and is not significant.

4.3 Simulating qrels

A probabilistic model of a phenomenon can be tested by seeing how well a simulation based on the model reproduces the phenomenon. The assessments in our datasets were made to form test collection qrels, used in calculating effectiveness metrics on system retrievals. If the qrels of an alternative assessor are used, then different effectiveness metric values will result. How well do probabilistic models based upon the observed assessor disagreement sim-

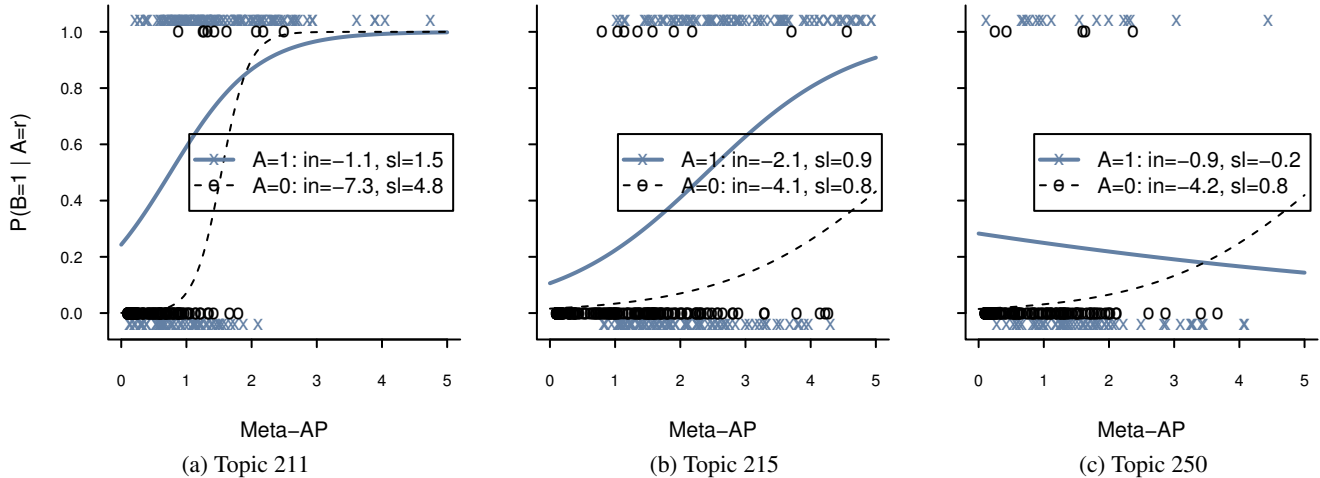


Figure 4: Logistic regression models for example topics from TREC 4, first alternative assessor.

ulate the actual change in absolute and relative system scores?

We examine model accuracy in predicting system scores in Figure 4. Mean AP scores for each TREC 4 system are shown, for the official qrels and for the qrels produced by the first alternative assessor. Simulated qrels are generated using the method described in Section 3.3, and the range of scores induced on each system by the randomly generated qrels are shown. The models we consider are a rank-insensitive universal flip rate based on disagreements micro-averaged across all topics (Figure 4 (a)); a separate flip-rate model built for each topic (Figure 4 (b)); a universal logistic regression model based upon meta-AP scores, pooling assessments across all topics (Figure 4 (c)); and a different logistic model for each topic (Figure 4 (d)). We also show the per-topic logistic regression model on the second alternate assessors for TREC 4 (Figure 4 (e)), and the single alternative assessor for TREC 6 (Figure 4 (f)).

The metarank models (Figures 4 (c) and (d)) achieve much more accurate absolute scores than the flip-rate models (Figures 4 (a) and (b)), with the flip-rate models generating MAP scores that are half or less of the true MAP scores. There is little difference between the universal and per-topic flip-rate models. There is, however, a noticeable difference between the universal and per-topic rank-sensitive models, with the per-topic models generating more accurate simulations of alternative-assessor scores; the universal model, in contrast, tends systematically to underestimate MAP scores. This result is not surprising, given the variance in per-topic model coefficients (Figure 2).

For comparative scores, however, the accuracy of the metarank and flip-rate models is reversed. Although the flip-rate models grossly understate true MAP, they do so by similar amounts for each system, leading to a ranking that is relatively consistent with the original. In contrast, while the average change in absolute scores is much smaller for the metarank models, the scores of different systems change by different amounts and directions, leading to instability in relative ranking. These observations are summarized by the RMSE and τ scores reported in Table 5.

The explanation for the better absolute but worse relative score stability of the metarank model, compared to the flip-rate model, is that while the metarank model is much more accurate model of probability of disagreement, it is less independent of systems. Since the overwhelming majority of per-topic models have positive slopes (Table 4), documents with higher meta-scores have

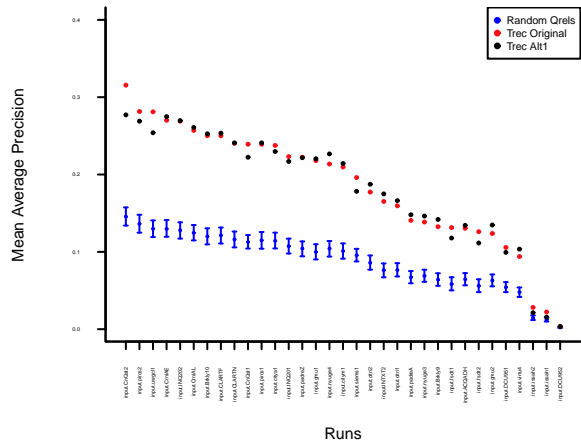
Model	Context	RMSE	$\bar{\tau}$
Flip-rate	Universal	0.102	0.911
	Per-topic	0.114	0.919
Meta-score	Universal	0.035	0.851
	Per-topic	0.015	0.867
	TREC 4 alt2	0.015	0.857
	TREC 6	0.080	0.815

Table 5: Root mean squared error and Kendall’s τ , averaged across simulations, between MAP scores from alternative-assessor and from model-generated qrels; summarizing the information in Figure 4.

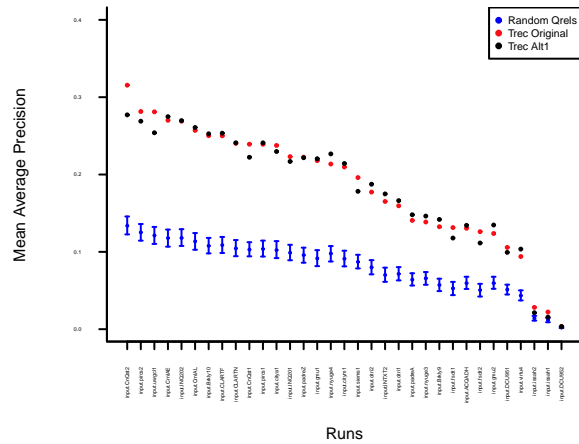
more chance of remaining (if originally assessed relevant), or flipping to (if originally irrelevant), relevant. Systems that return more documents with higher metarank will more positive score changes under the model than system that return fewer such documents. Metarank, however, is an average of ranking position across different runs, so the more systems that return a document, the higher its metarank. This behaviour is by design; but it does mean that documents returned by similar systems get higher metaranks, and hence the systems bigger boosts, than documents returned by dissimilar systems. We have returned to a common problem with these predictive models, that they favour conformist systems (say, automated methods using a standard document similarity measure) over non-conformist ones (say, hand-crafted manual runs).

The rank-sensitive meta-score model applied to the second alternative assessor for TREC 4 (Figure 4(e)) gives similar results as for the first. Interestingly, although the absolute scores under the second assessor differ from those for the first, the relationships between the observed and the simulated scores for each system are similar. For instance, the simulated scores for CLARTF and CLARTN (seventh and eighth systems from the left) both fall below the observed alternative scores by a similar amount, even though the observed scores are higher for the second alternative assessor than for the first. This consistency in error reinforces the hypothesis that the errors are due to system-dependent mutual reinforcement, rather than assessor-dependent or purely random factors.

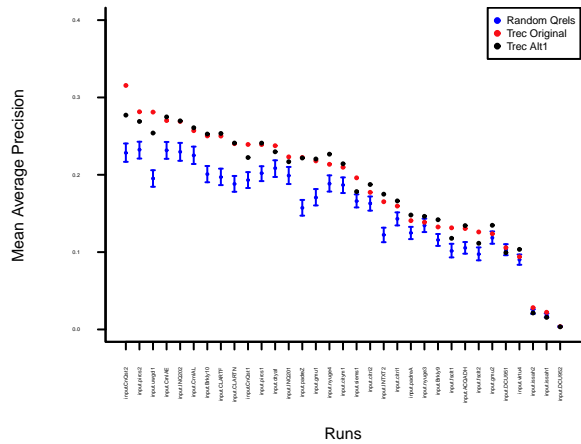
Finally, Figure 4 (f) shows the per-topic metarank model applied



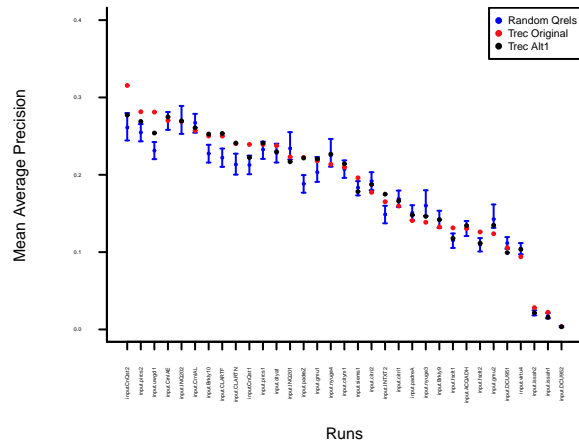
(a) FNR/FPR universal



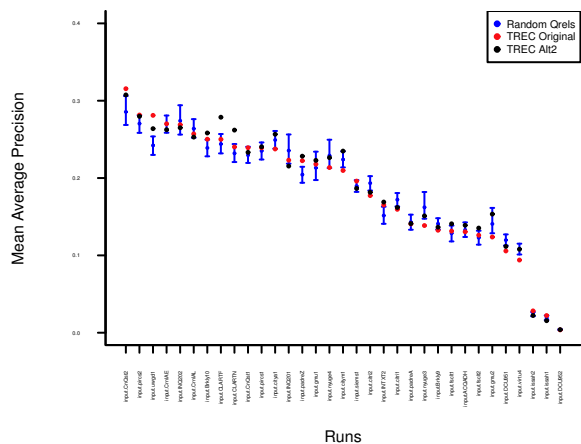
(b) FNR/FPR per-topic



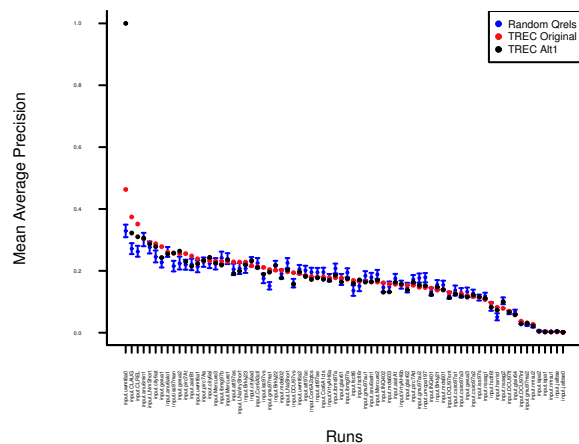
(c) Meta-score universal



(d) Meta-score per-topic



(e) Meta-score per topic, second assessor



(f) Meta-score per topic, TREC 6

Figure 5: System MAP scores under original and first alternative qrels, and 95% MAP score ranges under qrels randomly generated from original using different models of assessor disagreement. 1,000 qrels were simulated for each model.

Model	Sampling	RMSE	$\bar{\tau}$
Flip-rate	Uniform	0.162	0.779
Meta-score	Uniform	0.018	0.867
	Even	0.018	0.848

Table 6: Sampled dual assessment: root mean squared error and Kendall’s τ , averaged across simulations, between MAP scores from alternative-assessor and from model-generated qrels; summarizing the information in Figure 5.

to the alternative assessor for TREC 6. As described in Section 3.1, the dual assessments were not randomly sampled, but were made by one of the participants in the course of run development. Perhaps as a result of this, the MAP scores on the observed alternative assessments differ from the official ones more than for TREC 4. The simulated alternative assessments induce MAP scores that also depart further from the observed than for TREC 4. This may be a sign that the non-random selection of alternative assessments have biased the model.

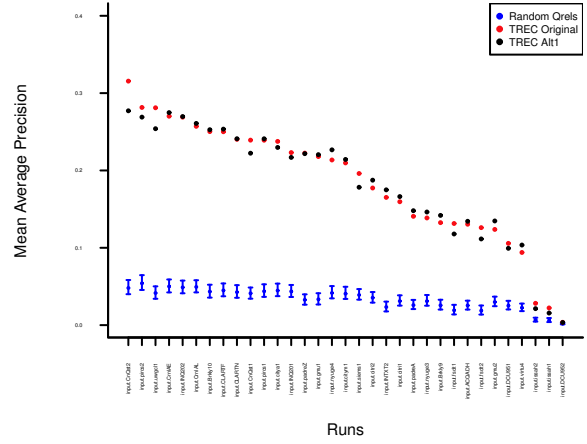
4.4 Sampled models

Exhaustive dual assessment is expensive. An alternative is to perform second assessments on a sample of documents, and estimate a model of assessor disagreement on the sample. A flip-rate model is estimated by observing the flip rate on the sample; a logistic metarank model by fitting the regression curve to the sampled documents. Which model is more robust to sampling?

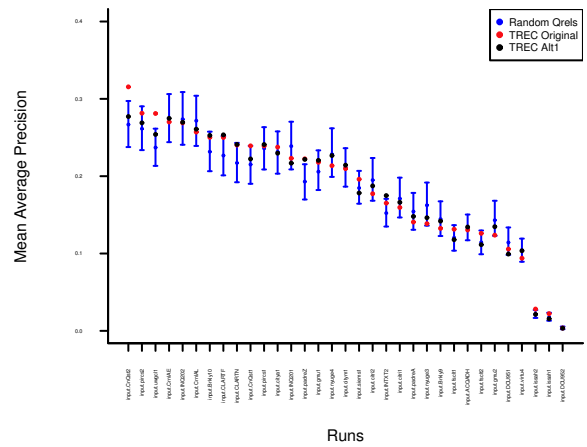
Figure 5 compares the reliability of the flip-rate and metarank models for sampled dual assessment. Only per-topic models are examined. Twenty relevant and twenty irrelevant documents are sampling, make up an average of 19% of the former and precisely 10% of the latter. The comparisons are summarized statistically in Table 6. (These results should be compared to the full dual-assessment reported in Figure 4 and Table 5). As before, the rank-sensitive meta-score models produce much more accurate absolute score estimates than the rank-insensitive flip-rate models. Indeed, while the RMSE of the flip-rate model increases by 0.060 or 60% with sampling, while the metarank model RMSE is only 0.003 or 20% higher. Unlike for full dual-assessment, however, the ranking and relative scores of the metarank models for sampled assessment are more stable than for the flip-rate models. Flip rate τ falls from 0.919 to 0.779 with sampling, whereas meta-score τ remains the same on 0.867.

The results in Figure 5 and Table 6 demonstrate that the metarank logistic models are highly robust to sampling: with only 10% to 20% of the pool dual-assessed, absolute and relative scores are almost as reliable as with full dual-assessment. Again, this demonstrates the accuracy of the model as a predictor of disagreement (the dependence between metarank score and system conformity aside). In contrast, the flip-rate model degrades badly with sampling, giving absolute scores as much as 80% below the correct values, and introducing sufficient noise that relative scores, and hence system ranking, become unreliable.

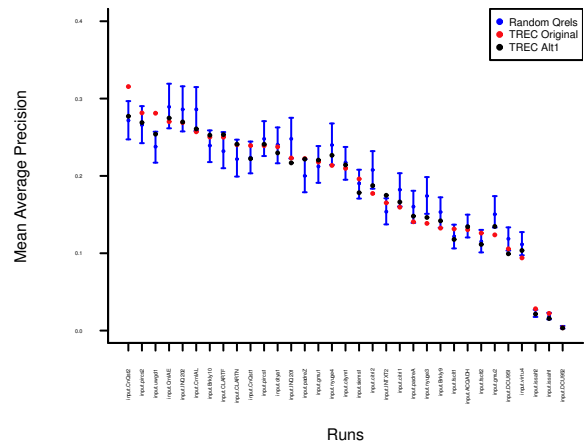
The availability of metarank scores allows for more advanced sampling regimes than simple uniform random sampling. In particular, documents can be ranked by metascore, and sampling performed to ensure representative items from each section (a form of stratified sampling). Intuitively, this would seem likely to provide more stable estimates, since we are sampling from along the full regression curve, and avoiding the risk of all the samples bunching up in one location. However, as the results in the final row



(a) Flip-rate



(b) Meta-score, uniform sampling



(c) Meta-score, even sampling

Figure 6: System MAP scores for random qrels generated from sampled dual-assessment. For each topic, twenty originally-relevant and twenty originally-irrelevant documents were sampled for re-assessment by the first alternative assessor. For (a) and (b), uniform random sampling was performed. For (c), 4 documents were sampled from each metarank quintile. Per-topic models are used.

of Table 6 (Figure 5) show, even sampling appears if anything to harm the model, leading to lower mean τ scores. Why this is so is unclear, and requires further investigation.

4.5 Effect of disagreement on significance

The purpose of significance testing is to determine whether a measured difference in the effectiveness of two retrieval algorithms is due to “chance”. In practice, this almost always means determining the extent to which variance due to the topic sample overrides the difference in effectiveness. There are other sources of variance apart from the topic sample, however; variance due to disagreement between assessors is one potentially important source. If two systems are significantly different for a sample of topics but *not* significantly different once assessor disagreement is modeled, the strength of the conclusion is reduced.

To incorporate disagreement into a significance test, we use our simulated qrels in a bootstrap-style significance procedure. For each pair of systems in the TREC-4 set, we sample a qrels produced by one of the models described above. We compute the significance between the two systems for that qrels. We can then compare this to the significance between the same two systems for the original qrels. Over many trials, we obtain a sense of the variance in significance due to disagreement as modeled by the simulated assessors.

The above procedure only models variance due to assessors—it no longer models variance due to the topic sample because the topic sample is held fixed in every experiment. To model both sources of variance, when we select two systems to compare, we also obtain a bootstrap sample of 48 topics over which to compare them; since the 48 topics and the qrels will vary with each experiment, this will produce a bootstrap distribution that incorporates variance due to both topic and assessor.

We compare the results of significance tests using the original assessments using three measures: precision (the proportion of system pairs found significant using simulated assessments that are also significant with the original assessments), recall (the proportion of system pairs significant with the original assessments that are also significant with the simulated assessments), and accuracy (the agreement between original and simulated assessments on both significance and non-significance). Note that higher values of these measures indicate greater agreement with the original assessors, but they are not necessarily *better*. If the values were 1.0 across all three, it would suggest that variance due to assessors has no effect whatsoever on significance.

Table 7 summarizes results for four models of disagreement using both procedures. The first set of results, with variance due only to assessor disagreement, has very high recall but fairly low precision and accuracy. Using simulated assessments without varying the topic sample results in significance being found at a higher rate than it is when varying the topic sample without varying assessments. This suggests that variance due to assessments is generally lower than variance due to the topic sample. Within these results there is not a great deal of difference due to the choice of model for simulation.

In the second set of results, varying both assessments and topics, we see decreases in recall but increases in both precision and accuracy; all three are very close in value. Disagreements now tend to be “symmetric”: a roughly equal number of pairs go from significant to non-significant as do the other direction. The flip-rate model has slightly higher numbers than the metarank model, suggesting it is more conservative in its modeling of assessor disagreement.

The results are somewhat unexpected in that when incorporating more variance into an experiment, one would generally expect that fewer pairs would be found significant, i.e. that recall would de-

Test	Model	Context	Prec.	Rec.	Acc.
Assessor only	Flip-rate	Universal	0.794	0.995	0.813
		Per-topic	0.813	0.992	0.831
	Meta-score	Universal	0.802	0.989	0.818
		Per-topic	0.802	0.992	0.820
Assessor + topic	Flip-rate	Universal	0.944	0.942	0.919
		Per-topic	0.943	0.928	0.909
	Meta-score	Universal	0.880	0.934	0.862
		Per-topic	0.921	0.929	0.892

Table 7: Summary of comparisons between statistical significance with the original TREC-4 assessors and significance with simulated assessors.

crease while precision would remain high. In our experiment, some pairs that had not been significantly different became significantly different with simulated assessments. One possible reason is bias introduced by the simulation. The models are imperfect, and may assign relevance in such a way that works in favor of certain systems that are less-favored by human assessors. Nevertheless, these results suggest that significance in IR is fairly robust to assessor disagreement.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have examined the relationship between the rank at which documents are returned and the probability that assessors will disagree about their relevance. Logistic regression has been deployed to test and model this relationship. The meta-AP metarank scoring method has been shown to be the best predictor of three metarank methods considered. We have found that there is a strong and consistent relationship between rank and probability of disagreement. An alternative assessor is much more likely to disagree with an original assessor’s relevant judgment if a document has low rank and is returned by few systems. Conversely, a high-ranked document that the original assessor found irrelevant is more likely to produce disagreement from an alternative assessor than is a low-ranked one.

Models of assessor disagreement allow for the simulation of the phenomenon. We have found that rank-sensitive models produce more accurate predictions of absolute scores than rank-insensitive ones, but less accurate relative scores, suggesting a dependence between systems and metarank scores. However, metarank models based upon sample dual assessment are more reliable than rank-insensitive ones for both absolute and relative measures. Indeed, sampled metarank models are almost as reliable as exhaustively-assessed ones, at a fraction of the assessment cost. Finally, we have demonstrated the use of such sampled models in adding assessor variability to test the significance of retrieval evaluation results, finding that while the rate of significance overall does not change, around a tenth of system pairs switch from being significantly to non-significantly different, or vice versa.

We have observed that the dependence between systems and the meta-AP metarank measure produces simulated qrel sets that favour conformist systems. Finding a metarank score that avoids this bias, or developing a method that corrects for it if it occurs, is future work. For instance, metarank contributions from similar systems could be downweighted. Also, the distribution of meta-AP scores is dependent on both assessment depth and the number of systems in the pool; it is desirable to find a metarank measure less dependent on these factors.

Finally, we observed in Figure 3 that models based even on exhaustive assessment can be sensitive to the assessment and metas-

core of a small number of documents, making them prone to anomalous behaviour (inverse regression, for instance, or sudden transitions in probability, or again models that give zero probability to alternative relevance no matter what the metarank score). This sensitivity will be heightened for sample-based models, and higher still as sample size decreases. A multi-level Bayesian model would help alleviate this problem, by smoothing the logistic models for one topic based upon results on other topics.

Acknowledgments

Ellen Voorhees provided, and helped with the interpretation of, the TREC 4 and TREC 6 multiple-assessment data sets.

References

- Javed Aslam, Virgiliu Pavlu, and Emine Yilmaz. Measure-based metasearch. In Gary Marchionini, Alistair Moffat, John Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 571–572, Salvador, Brazil, August 2005. doi: <http://doi.acm.org/10.1145/1076034.1076133>.
- Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, A. de Vries, and Emine Yilmaz. Relevance assessment: are judges exchangeable and does it matter? In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–674, Singapore, Singapore, July 2008.
- D. Bodoff and P. Li. Test theory for assessing IR test collections. In Clarke et al. [2007], pages 367–374.
- Ben Carterette. Robust test collections for retrieval evaluation. In Clarke et al. [2007], pages 55–62.
- Ben Carterette and Ian Soboroff. The effect of assessor errors on IR system evaluation. In Hsin-Hsi Chen, Efthimis N. Efthimiadis, Jacques Savoy, Fabio Crestani, and Stephanie Marchand-Maillet, editors, *Proc. 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–546, Geneva, Switzerland, July 2010.
- Charles L. A. Clarke, Norbert Fuhr, Noriko Kando, W. Kraaij, and A. de Vries, editors. *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, the Netherlands, July 2007.
- Gordon V. Cormack, Charles L. A. Clarke, Christopher R. Palmer, and Samuel S. L. To. Passage-based refinement (MultiText experiments for TREC-6). In Ellen Voorhees and Donna Harman, editors, *Proc. 6th Text REtrieval Conference*, pages 303–320, Gaithersburg, Maryland, USA, November 1997. NIST Special Publication 500-240.
- Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In Croft et al. [1998], pages 282–289.
- W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors. *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998.
- C. Cuadra and R. Katter. The relevance of relevance assessment. *Proc. American Documentation Institute*, 4:95–99, 1967.
- Maura R. Grossman and Gordon V. Cormack. Inconsistent assessment of responsiveness in e-discovery: difference of opinion or human error? In *DESI IV: The ICAIL Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*, pages 1–11, Pittsburgh, PA, USA, June 2011.
- Donna Harman. Overview of the fourth text REtrieval conference (TREC-4). In Donna Harman, editor, *Proc. 4th Text REtrieval Conference*, pages 1–23, Gaithersburg, Maryland, USA, November 1995. NIST Special Publication 500-236.
- D. W. Hosmer, T. Hosmer, S. le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16:965–980, 1997.
- Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. Overview of the TREC 2008 legal track. In Ellen Voorhees and Lori P. Buckland, editors, *Proc. 17th Text REtrieval Conference*, pages 3:1–45, Gaithersburg, Maryland, USA, November 2008. NIST Special Publication 500-277.
- Herbert L. Roitblat, Anne Kershaw, and Patrick 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.
- Tetsuya Sakai. Evaluating evaluation metrics based on the bootstrap. In Susan Dumais, E. Efthimiadis, David Hawking, and Kalervo Järvelin, editors, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 525–532, Seattle, Washington, USA, August 2006. doi: <http://doi.acm.org/10.1145/1148170.1148261>.
- T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, 2007.
- Jacques Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4):495–512, 1997.
- Victor S. Sheng, F. Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622, Las Vegas, Nevada, USA, August 2008.
- Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In Mario J. Silvaa, Alberto A. F. Laender, R. Baeza-Yates, Deborah L. McGuinness, Bjorn Olstad, Øystein Haug Olsen, and Andre O. Falcão, editors, *Proc. 16th ACM International Conference on Information and Knowledge Management*, pages 623–632, Lisboa, Portugal, November 2007.
- Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking retrieval systems without relevance judgments. In W. Bruce Croft, D. J. Harper, D. H. Kraft, and Justin Zobel, editors, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 66–73, New Orleans, Louisiana, USA, September 2001.
- Andrew Trotman and Dylan Jenkinson. IR evaluation using multiple assessors per topic. In MingFang Wu, Andrew Turpin, and Amanda Spink, editors, *Proc. 12th Australasian Document Computing Symposium*, pages 9–16, Melbourne, December 2007.
- Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In Croft et al. [1998], pages 315–323.
- Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5): 697–716, September 2000.
- William Webber, Douglas W. Oard, Falk Scholer, and Bruce Hedin. Assessor error in stratified evaluation. In *Proc. 19th ACM International Conference on Information and Knowledge Management*, pages 539–548, Toronto, Canada, October 2010.
- William Webber, Bryan Toth, and Marjorie Desamito. Effect of written instructions on assessor agreement. In William Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *Proc. 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, Oregon, USA, August 2012. to appear.