The TREC-5 Filtering Track

David D. Lewis
AT&T Labs—Research
600 Mountain Avenue, 2A-410
Murray Hill, NJ 07974
lewis@research.att.com
http://www.research.att.com/~lewis

Abstract

The TREC-5 filtering track, an evaluation of binary text classification systems, was a repeat of the filtering evaluation run in a trial version for TREC-4, with only the data set and participants changing. Seven sites took part, submitting a total of ten runs. We review the nature of the task, the effectiveness measures and evaluation methods used, and briefly discuss the results. Some deficiencies in the evaluation are examined, with an eye toward improving future filtering evaluations.

1 Introduction

The goal of the TREC-5 filtering track was to aid research groups in evaluating their approaches to binary text classification. As usual in the TREC context, this was done by making available a large data set and doing blind, impartial evaluation of submitted results. The design used in the TREC-5 filtering track was identical to that tested with four sites in TREC-4, thus much of this paper is identical to the description of the TREC-4 filtering track [10].

We begin by defining binary text classification and presenting some applications of it. We then discuss a particular binary text classification task, filtering, used in TREC-5. The effectiveness of filtering submissions was evaluated using utility as a measure. The several roles that this effectiveness measure played in the evaluation are described.

The large size of the TREC-5 test data set meant that relevance judgments were of necessity incomplete and effectiveness could only be estimated. We describe in detail two approaches that were tested for estimating the utility of filtering submissions. Stratified sampling breaks up each filtering submission into related groups of documents and takes a random sample from each group to be judged. Pooled sampling instead uses all the judged documents from the routing and filtering tracks, including the stratified samples.

We end by briefly discussing the results of the TREC-5 filtering track, with an emphasis on what was learned about the evaluation methods.

2 Binary Text Classification

By binary text classification systems, we mean information retrieval (IR) systems that decide for each document processed whether the document should be accepted or rejected [9]. What it means to be accepted varies between systems. Some applications that make use of binary text classification are:

- A company provides an SDI (selective dissemination of information) service which filters newswire feeds. Relevant articles are faxed each morning to clients.
- A text categorization system assigns controlled vocabulary categories to incoming documents as they are stored in a text database.
- An "agent" program monitors low content text streams (e.g. Usenet newsgroups) and alerts a user when a relevant message appears.

Note that there is no notion of a user choosing how far to go down a ranking in these systems. The system makes Yes/No decisions about documents, and the user only sees the results of those decisions. This affects the kind of evaluation appropriate for the system, as discussed in the next section.

3 Evaluation for Binary Text Classification

We have discussed evaluation for binary classification at length elsewhere [8, 9], and so here will concentrate on how it differs from the evaluation of ranked retrieval in the main TREC-5 tasks.

Effectiveness measures for ranked retrieval typically have two components. The first is a *cutoff*: a specification of how to divide the ranking into a top part and a bottom part. The top part is considered to be the set of documents retrieved by the system. The second component of the overall effectiveness measure is a *set-based effectiveness measure* which is applied to the retrieved set.

Some examples of these two-component measures are:

- Precision at 0.10 recall : Here the cutoff is the highest point in the ranking above which at least 10% of the relevant documents in the test set occur. The set-based effectiveness measure is precision, the proportion of documents in the retrieved set which are relevant.
- Recall at 0.001 fallout : Similar to the above, but the cutoff is based on fallout and the set-based effectiveness measure is recall.
- Precision at 20 documents : Here the cutoff is based on a fixed position in the ranking (20 documents down from the top). The set-based effectiveness measure is precision.
- R-precision : This is precision at R documents, where the cutoff R is the number of relevant documents in the collection. The set-based effectiveness measure is precision.

More complex measures, such as the average of precision over multiple recall cutoffs are also used. A wide variety of such measures with different cutoffs and different set-based effectiveness measures have been applied to rankings in the TREC evaluations. This might seem to make it difficult for a site to decide how rank documents, since there is more than one measure to optimize. In truth, all these measures can be optimized simultaneously by the simple and obvious strategy of ranking documents by their probability of relevance. Doing so will result in an optimal score under essentially any reasonable measure of ranking effectivenes, a property which has been formalized as the Probability Ranking Principle [9, 13].

In contrast, binary classification systems make the separation into accepted and rejected documents themselves, rather than leaving this up to the effectiveness measure used in evaluation. The binary classification system must choose what separation to make in order to optimize the effectiveness measure used. Doing so optimally means that the effectiveness measure must be known in advance. Since binary classification systems do not rank the accepted set, the effectiveness measure should be a set-based one. Since the size of the submitted set is under the control of the system, the effectiveness measure used in evaluation must be able to assign an effectiveness to a set of any size, including the empty set.

It is still desirable in the filtering context to test the ability of systems to satisfy varying user preferences (e.g. high recall vs. high precision), but this should not be done by submitting a single ranking and letting the evaluation program pick cutoffs. Instead, a family of effectiveness measures can be used to capture different user preferences. For each measure used, the filtering system produces a separate set of documents appropriate to that measure, and that same measure is used to evaluate the set of documents.

4 Filtering: A Binary Classification Task for TREC

The mainline tasks for TREC-1 through TREC-5, routing and ad hoc retrieval, require participants to submit ranked lists of documents, which are then evaluated using ranking-oriented effectiveness measures. The number of documents submitted is defined in advance, so the ability of systems to pick the number of documents to submit is not tested.

The TREC-5 filtering track addresses this limitation. This section describes the rationale for the evaluation, the evaluation's structure, and the effectiveness measures used.

4.1 Why Filtering?

The main motivation for the filtering track is the increasing number of IR applications requiring binary text classification (see Section 2). The track should help developers of these applications learn about relevant techniques from the research community, and let researchers compare and evaluate their approaches.

A second motivation is that the demands of the filtering task may encourage the development of IR methods with other desirable properties. For instance, accurately estimating the probability of relevance of documents is useful not only in filtering [7, 9, 5], but also for self-monitoring of effectiveness [9], estimating the number of relevant documents [9], and selection of training data

[11].

Finally, we hope that a binary classification task will attract a broader range of researchers and approaches to TREC. The requirement that TREC results be ranked makes it awkward for approaches that are not ranking-oriented to be tried [6, 14]. These approaches include boolean querying by human experts, as well as the use of binary classifiers (e.g. decision trees) produced by machine learning techniques.

4.2 Structure

The structure of the TREC-5 filtering evaluation was as follows. We combined each of 50 TREC-5 topics with each of 3 set-based effectiveness measures, to produce 150 descriptions of user needs. The topic portion of the user need indicates the kind of information sought by the hypothetical user, while the set-based effectiveness measure captures the hypothetical user's tolerance for different kinds of mistakes. The set-based effectiveness measures were based on a notion of utility, as described in Section 4.4.

The topics were the same ones used in the main TREC-5 routing evaluation. However, as with the routing evaluation, data for one topic was accidentally omitted from judging. Therefore the filtering track results are based on 49 topics, with three effectiveness measures for each, yielding 147 descriptions of user needs. In the routing evaluation an additional 4 topics for which no relevant documents were found were also omitted when computing effectiveness measures, but these topics were retained for the filtering track evaluation.

For each user need, the system had to make use of the topic and the effectiveness measure to decide whether to accept or reject each test document. The same test documents as the main routing evaluation were used. The submitted set (i.e. the accepted documents) for the user need was then evaluated using the effectiveness measure for that user need. (Actually only a sample from the submitted set was used in the evaluation—see Section 5.)

4.3 Integration with Routing

The TREC-5 filtering and routing evaluations used the same topics and test documents. (See the discussion by Harman and Voorhees elsewhere in the proceedings.) The training data for the topics (i.e. documents judged with respect to the topics in previous TRECs) were also the same in both evaluations.

The evaluations were also similar in that 100 documents from each site's results for a topic went into the pool for judging. In the case of routing, the judged documents were the top 100 from a single ranked list of 1000 documents submitted for the topic. In the case of filtering, the 100 documents were a stratified sample (Section 5.2.1) from the union of the three unranked sets of documents submitted for that topic. Filtering and routing documents were mixed together and treated identically for judging. The expense of running the filtering track was thereby reduced, since there was considerable overlap between filtering and routing submissions.

4.4 The Utility Measures

The family of effectiveness measures used in the filtering track were based on assigning a numeric value or *utility* to each retrieved document [2, 9]. Retrieved relevant documents received a positive utility, and retrieved nonrelevant documents received a negative utility. The total utility of the submitted set for run R_i was:

$$u_i = u_{ai}A_i + u_{bi}B_i$$

where A_i is the number of relevant documents in the submitted set for run R_i and B_i is the number of nonrelevant documents in the submitted set for run R_i. For each run R_i we assume that u_{ai} is the value the user places on receiving a relevant document, while u_{bi} is the value of receiving a nonrelevant document.

Different values for u_{ai} and u_{bi} define different effectiveness measures in the family. The three utility measures used in the filtering evaluation were:

Run	Parameter Values	Effectiveness Measure
R_1	$u_{a1} = 1, u_{b1} = -3$	$u_1 = A_1 - 3B_1$
R_2	$u_{a2} = 1, u_{b2} = -1$	$u_2 = A_2 - B_2$
R_3	$u_{a3} = 3, u_{b3} = -1$	$u_3 = 3A_3 - B_3$

We might imagine run R_1 corresponding to a user who is willing to pay 1 dollar (or pick your favorite currency) to read each relevant document, and loses 3 dollars worth of time if they have to read a nonrelevant document. Therefore, run R_1 requires the filtering system to act in a conservative or high precision fashion. In constrast, run R_3 encourages systems to emphasize recall somewhat more, while run R_2 is in between.

Unlike recall, but like precision, our utility measures take into account only those documents submitted by the system for a run. (It is possible to define utility measures to take into account rejected documents as well [9].) Unlike both recall and precision, the total utility is not normalized to lie between 0 and 1. Indeed, the minimum and maximum achievable utilities can be determined only if the total number of relevant documents in the test set is known. The goal of systems, however, is simply to achieve the highest utility they can.

4.5 Optimizing The Utility Measures

Different effectiveness measures are more or less easy to optimize by different IR techniques. A common approach to binary text classification is to compute a numeric score for a document to be classified, and assign the document to the class if the numeric score is larger than some specified value. Utility measures which are the sum of utilities for individual documents, like the one used in the TREC-5 filtering evaluation, can be optimized by thresholding if the scores computed are monotonic with probability of class membership [9].

In fact, if the document scores are accurate estimates of probability of relevance, the thresholds to use can be derived directly from the effectiveness measure by decision theoretic principles [3, Ch. 2]. For the TREC-5 filtering measures the optimal thresholds on probability of relevance are:

Run	Parameter Values	Threshold On Probability
R_1	$u_{a1} = 1, u_{b1} = -3$	0.75
R_2	$u_{a2} = 1, u_{b2} = -1$	0.50
R_3	$u_{a3} = 3, u_{b3} = -1$	0.25

Of course, these thresholds are optimal only if the probability estimates produced by a system are in fact accurate.

Note that probability of relevance is the same as *instantaneous precision*. That is, if a system estimated a probability of relevance of p for a large set of documents, and those estimates were accurate, then we would expect that set of documents to have a precision of approximately p. This means that the lowest scoring documents submitted for run R_3 by a highly effective filtering system would have an instantaneous precision of around 0.25. Most of the documents in that system's submitted set would be likely to have instantaneous precisions considerably higher than that. The overall precision for a high quality submitted set should therefore typically be much higher than the threshold probability/instantaneous precision. Thus while run R_3 emphasized recall somewhat more than run R_1 or run R_2, all three TREC-5 filtering runs required, by the standards of typical IR systems, high precision results.

5 Estimating Total Utility from a Sample

Computing the exact total utility for a submitted run requires knowing the value of A_i and B_i for that run. This would require assessing the relevance of every document submitted for that run. Because submitted sets can be of any size, this might require too much work from the relevance assessors. For that reason, total utility for filtering runs was estimated using *samples* of the submitted documents.

Two different sampling and estimation methods were used in the TREC-5 filtering track, as described below.

5.1 Pooling

The first approach to sampling was the usual TREC pooling strategy [4]. This approach assumes that some known pool of documents contains all the relevant documents in the test set. The pool for the TREC-5 filtering task consisted of all documents judged for the topic in the main routing task, plus all documents judged for the topic for the filtering task, as chosen by the stratified sampling scheme of Section 5.2.1. Under the pooling assumption an estimate $\hat{u}_i$ of the total utility, u_i, is easily computed.

The total utility computed in this fashion is only an estimate, because the pooling assumption may be wrong. There may be submitted documents that are relevant but were not judged for this topic. An advantage of the pooled estimate, however, is that the same sample is used for all sites, enabling that sample to be large. The use of the same sample for all sites also eliminates sampling variation between sites. A disadvantage is that the sample is not a random sample, meaning that it

is difficult to tell how accurate the estimated utilties are (see Section 7.1). There is also the danger that pooled sampling penalizes sites which submit atypical yet relevant documents.

5.2 Random Sampling

To see how random sampling could be used to estimate the total utility it is useful to rewrite the formula for total utility as:

$$\begin{aligned} u_i &= u_{ai} \times A_i + u_{bi} \times B_i \\ &= u_{ai} \times p_i N_i + u_{bi} \times (1 - p_i) N_i \\ &= ((u_{ai} - u_{bi}) p_i + u_{bi}) N_i \end{aligned} \tag{1}$$

where $N_i = A_i + B_i$ is the total number of documents submitted for the run, and $p_i = A_i / N_i$ is the proportion of documents submitted which are relevant (i.e. the precision of the submission). Rewriting the utility measures used in the filtering evaluation in this way gives:

$$u_1 = (4p_1 - 3)N_1 \tag{2}$$

$$u_2 = (2p_2 - 1)N_2 \tag{3}$$

$$u_3 = (4p_3 - 1)N_3 \tag{4}$$

Therefore, if we can produce an estimate $\hat{p}_i$ of the proportion of relevant documents in a submitted set, we can turn that into an estimate, $\hat{u}_i$, of the utility of the submitted set:

$$\hat{u}_i = ((u_{ai} - u_{bi})\hat{p}_i + u_{bi}) N_i \tag{5}$$

One approach to estimating the proportion would be to take a random sample from the submitted set for a topic/run pair, count the number of relevant documents, a, in that sample, and divide by the number of documents, n, in the sample:

$$\hat{p} = \frac{a}{n} \tag{6}$$

This is called *simple random sampling*. A more complex approach to random sampling often has advantages, as described in the next section.

5.2.1 Stratified Random Sampling

Simple random sampling is not the only way to estimate a proportion. The TREC-5 filtering evaluation estimated utilities using *stratified sampling*. In stratified sampling we use additional knowledge about a population to divide the population into groups or *strata* [1, p. 89]. We then take a simple random sample separately from each stratum, estimate the quantity of interest for each stratum, and combine the stratum estimates to get an overall estimate for the population. If the strata are chosen so that items in a stratum are similar to each other, the accuracy of an stratified estimate can be greater than the accuracy of an estimate based on simple random sampling from the whole population.

R_1	R_2	R_3	Stratum Name (h)	Number of Documents in Stratum (N_h)
0	0	0	000	very many
0	0	1	001	many
0	1	0	010	very few or none
0	1	1	011	some
1	0	0	100	very few or none
1	0	1	101	very few or none
1	1	0	110	very few or none
1	1	1	111	few

Figure 1: The test documents submitted by a site can be separated into eight strata, based on which of the three submitted sets, the R_1 set, the R_2 set, and the R_3 set each document appeared in. We indicate presence in the set by a 1, absence by a 0. The comments indicate the relative sizes of the sets in typical filtering track submissions.

For TREC-5, the set of filtering documents submitted by each site for a topic was stratified according to which of the three runs each document was submitted for. By considering all combinations of presence and absence of a document in the three submitted sets, we get 8 strata, as shown in Figure 1.

In most, but not all, cases the submitted sets will be such that the R_1 set is contained in the R_2 set, and the R_2 set is contained in R_3 set. This means that strata 010, 100, 101, and 110 will usually be empty. In general, however, each of the submitted sets can be the union of four strata:

$$\begin{aligned} \text{Set for run } R_1 &: 100, 101, 110, 111 \\ \text{Set for run } R_2 &: 010, 011, 110, 111 \\ \text{Set for run } R_3 &: 001, 011, 101, 111 \end{aligned}$$

To estimate the proportion p_i of relevant documents in set R_i by stratified sampling, we separately estimate the proportion p_h for each stratum h in the R_i set. We then add up the estimated stratum proportions, weighting them by the relative size of their stratum in the submitted set [1, p. 91]:

$$\hat{p}_i = \sum_{h \in R_i} \frac{N_h}{N_i} \hat{p}_h \tag{7}$$

Here h ranges over the strata that make up the R_i set, N_h is the size of stratum h, N_i is the size of the R_i set, and $\hat{p}_h$ is an estimate of the proportion of relevant in stratum h. Expanding this out for runs R_1 to R_3 gives:

$$\hat{p}_1 = \frac{N_{100}}{N_1} \times \hat{p}_{100} + \frac{N_{101}}{N_1} \times \hat{p}_{101} + \frac{N_{110}}{N_1} \times \hat{p}_{110} + \frac{N_{111}}{N_1} \times \hat{p}_{111} \tag{8}$$

$$\hat{p}_2 = \frac{N_{010}}{N_2} \times \hat{p}_{010} + \frac{N_{011}}{N_2} \times \hat{p}_{011} + \frac{N_{110}}{N_2} \times \hat{p}_{110} + \frac{N_{111}}{N_2} \times \hat{p}_{111} \tag{9}$$

$$\hat{p}_3 = \frac{N_{001}}{N_3} \times \hat{p}_{001} + \frac{N_{011}}{N_3} \times \hat{p}_{011} + \frac{N_{101}}{N_3} \times \hat{p}_{101} + \frac{N_{111}}{N_3} \times \hat{p}_{111} \quad (10)$$

These estimates will be unbiased (see Section 7) if the estimate $\hat{p}_h$ of the proportion of relevant documents in stratum h is unbiased for each component stratum h. As our estimate $\hat{p}_h$ we used the proportion of relevant documents found in a simple random sample from stratum h:

$$\hat{p}_h = \frac{a_h}{n_h}$$

where n_h is the size of the simple random sample taken from stratum h and a_h is the number of relevant documents found in that sample. This $\hat{p}_h$ is an unbiased estimate of p_h [1, p. 51].

5.3 Sample Sizes in Stratified Sampling for TREC-5

To be consistent with the TREC-5 routing evaluation, at most 100 documents were judged from the three sets of documents submitted by a site for each filtering topic. These 100 documents had to be allocated to as many as seven strata (all except stratum 000), as described above. This was done by choosing equal sized samples from all nonempty strata. If all documents from a stratum were used up by this procedure the leftover documents were allocated equally among the other strata, until a total of 100 was reached, or until all documents from the three submitted sets were selected.

6 Stratified Sampling: An Example

Figure 2 displays data on the submitted sets from a hypothetical filtering site for a single topic. (The test set size differs from that actually used in TREC-5.) The R_2 set is bigger than the R_1 set, and the R_3 set is bigger than both the R_1 and R_2 sets. One anomaly is that 10 documents are in R_2 set but not in R_3 set. This might happen due to a mistake by the site, or because documents were retrieved by boolean queries which were not in a strict generalization relationship.

6.1 Estimating Proportion of Relevant Documents

If all submitted documents were judged, then we could compute the true proportion of relevant documents in each submitted set:

$$p_1 = \frac{30}{30+10} = .7500 \quad (11)$$

$$p_2 = \frac{112}{112+138} = .4480 \quad (12)$$

$$p_3 = \frac{130}{130+1110} = .1048 \quad (13)$$

To compute a stratified estimate of these proportions, we assume that simple random samples were drawn from each stratum and judged for relevance, as shown in Figure 2. This gives estimates of the proportion of relevant documents in each stratum, as shown in the last column of Figure 2.

Stratum	Submitted Docs			Samples from Strata		
	Rels	NonRels	True Prop	Rels	NonRels	Est. Prop
000	50	100000	.0005	0	0	–
001	20	980	.0200	1	29	.0333
010	2	8	.2000	2	8	.2000
011	80	120	.4000	10	20	.3333
100	0	0	–	0	0	–
101	0	0	–	0	0	–
110	0	0	–	0	0	–
111	30	10	.7500	23	7	.7667
R1 Total	30	10	.7500	–	–	.7667
R2 Total	112	138	.4480	–	–	.3973
R3 Total	130	1110	.1048	–	–	.1054
Test Set Total	182	101118	.0018	–	–	–

Figure 2: Hypothetical data on a site's submitted sets for a single topic. We show both the true and sampled values of the number of relevant and nonrelevant documents in each stratum and run, and the corresponding proportion of relevant.

We then combine the stratum estimates, using Equations 8 to 10, to get estimates of the proportion of relevant in each submitted set:

$$\hat{p}_1 = \frac{40}{40} \times \frac{23}{23+7} = .7667 \tag{14}$$

$$\hat{p}_2 = \frac{10}{250} \times \frac{2}{2+8} + \frac{200}{250} \times \frac{10}{10+20} + \frac{40}{250} \times \frac{23}{23+7} = .3973 \tag{15}$$

$$\hat{p}_3 = \frac{1000}{1240} \times \frac{1}{1+29} + \frac{200}{1240} \times \frac{10}{10+20} + \frac{40}{1240} \times \frac{23}{23+7} = .1054 \tag{16}$$

6.2 Estimating Utility of a Submitted Set

If we knew the true proportion of relevant documents in each submitted set (Equations 11 to 13), we could compute the true utility of each set, using Equations 2 to 4:

$$u_1 = (4 \times .7500 - 3) \times 40 = 0.0 \tag{17}$$

$$u_2 = (2 \times .4480 - 1) \times 250 = -26.0 \tag{18}$$

$$u_3 = (4 \times .1048 - 1) \times 1240 = -719.2 \tag{19}$$

If we instead have the stratified estimates of the proportions, we use them to get estimates of the total utility:

$$\hat{u}_1 = (4 \times .7667 - 3) \times 40 = 2.672 \tag{20}$$

$$\hat{u}_2 = (2 \times .3973 - 1) \times 250 = -51.35 \tag{21}$$

$$\hat{u}_3 = (4 \times .1054 - 1) \times 1240 = -717.2 \tag{22}$$

The estimates for the R_1 and R_3 sets are close to the true values, while the estimate for the R_2 set is less close. We can see why by comparing in Figure 2 the true and estimated proportion of relevant for each stratum in the R_2 set. Due to bad luck with our random sample from 011, the largest stratum in the R_2 set, we underestimated the proportion of relevant in that stratum. This carried over to our estimate of the overall proportion of relevant for the R_2 set, and thus to the total utility. This raises the question of how much confidence we can have in our estimates of utility, and is the subject of the next section.

7 How Accurate are Our Estimates of Utility?

The pooling and stratified sampling approaches are based on judging only a subset of each submitted set, so in neither case will the estimates of utility be perfect. A common measure of the distance between an estimate, $\hat{\mu}$, and the quantity we want to estimate, μ, is the mean square error (MSE) [1, p. 15]. The MSE of an estimate is the expected value of the square of the difference between the estimate and the true value:

$$\mathrm{MSE}[\hat{\mu}] = \mathrm{E}[(\hat{\mu} - \mu)^2] \tag{23}$$

Letting $m = \mathrm{E}[\hat{\mu}]$, the MSE can be rewritten as the sum of two terms:

$$\begin{aligned}
\mathrm{MSE}[\hat{\mu}] &= \mathrm{E}[(\hat{\mu} - \mu)^2] \\
&= \mathrm{E}[((\hat{\mu} - m) - (\mu - m))^2] \\
&= \mathrm{E}[(\hat{\mu} - m)^2 - 2(\mu - m)(\hat{\mu} - m) + (\mu - m)^2] \\
&= \mathrm{E}[(\hat{\mu} - m)^2] - \mathrm{E}[2(\mu - m)(\hat{\mu} - m)] + \mathrm{E}[(\mu - m)^2] \\
&= \mathrm{E}[(\hat{\mu} - m)^2] - 2 \times 0 \times \mathrm{E}[(\hat{\mu} - m)] + (\mu - m)^2 \\
&= \mathrm{E}[(\hat{\mu} - m)^2] + (\mu - m)^2 \\
&= \mathrm{E}[(\hat{\mu} - \mathrm{E}[\hat{\mu}])^2] + (\mu - \mathrm{E}[\hat{\mu}])^2 \\
&= \mathrm{Var}[\hat{\mu}] + \mathrm{Bias}[\hat{\mu}].
\end{aligned}$$

The first term:

$$\mathrm{Var}[\hat{\mu}] = \mathrm{E}[(\hat{\mu} - \mathrm{E}[\hat{\mu}])^2]$$

is the *variance* of the estimator $\hat{\mu}$ and measures the tendency of the estimator to deviate from its own expected value. The second term:

$$\mathrm{Bias}[\hat{\mu}] = (\mathrm{E}[\hat{\mu}] - \mu)^2$$

is the *bias* of $\hat{\mu}$ and measures the systematic difference between the expected value of the estimator and the value we are trying to estimate. It is often, though not always, desirable to use *unbiased* estimates of a quantity. An estimate is unbiased if $\mathrm{E}[\hat{\mu}] = \mu$, i.e. $\mathrm{Bias}[\hat{\mu}] = 0$.

These two concepts, bias and variance, and their sum the MSE, will be useful in discussing the accuracy of our estimates of utility.

7.1 Accuracy of Pooled Estimates

The MSE of the pooled estimates is difficult to determine, since the pool is not constructed randomly. The variance of a pooled estimate is nonzero, since we do not sample the entire population. However, the variance is likely to be smaller than that of the corresponding stratified estimate, due to the large number of documents judged.

The pooled estimate also has a nonzero bias, since if there are any relevant documents in the submitted set which were not judged, the estimated utility will be lower than the true utility. In fact, not only is the expected value of the pooled estimate always less than or equal to the true utility, but the actual value of the pooled estimate is always less than or equal to the true utility. So the pooled estimate is a lower bound on the true utility.

7.2 Accuracy of Estimates Based on Random Sampling

Recall that the utility of a submitted set can be expressed in terms of the proportion of relevant documents in that set:

$$u_i = ((u_{ai} - u_{bi})p_i + u_{bi})N_i \tag{24}$$

Similarly, we can estimate the utility of a submitted set based on an estimate of the proportion of relevant documents in that set:

$$\hat{u}_i = ((u_{ai} - u_{bi})\hat{p}_i + u_{bi})N_i \tag{25}$$

The MSE of such an estimate is

$$\begin{aligned}
\mathrm{MSE}[\hat{u}_i] &= \mathrm{E}[(\hat{u}_i - u_i)^2] \\
&= \mathrm{E}[(((u_{ai} - u_{bi})\hat{p}_i + u_{bi})N_i - ((u_{ai} - u_{bi})p_i + u_{bi})N_i)^2] \\
&= \mathrm{E}[(u_{ai} - u_{bi})^2 {N_i}^2 (\hat{p}_i - p_i)^2] \\
&= (u_{ai} - u_{bi})^2 {N_i}^2 \mathrm{E}[(\hat{p}_i - p_i)^2] \\
&= (u_{ai} - u_{bi})^2 {N_i}^2 \mathrm{MSE}[\hat{p}_i].
\end{aligned} \tag{26}$$

So the MSE of $\hat{u}_i$ is a simple function of the MSE of our estimate of the proportion of relevant documents. For simple random sampling and stratified sampling, the estimates of the proportion are unbiased, that is $\mathrm{E}[\hat{p}_i] = p_i$. Therefore, the MSE of $\hat{p}_i$ results solely from its variance, and we have:

$$\mathrm{MSE}[\hat{u}_i] = (u_{ai} - u_{bi})^2 {N_i}^2 \mathrm{Var}[\hat{p}_i]. \tag{27}$$

Also note that $\hat{u}_i$ is unbiased as well, so its MSE consists solely of variance.

In the rest of this section we will look at what $\hat{p}_i$'s variance is under different sampling techniques.

7.2.1 Variance of Proportions Estimated by Simple Random Sampling

We begin with the estimate produced by simple random sampling, as this is both a component of, and a point of comparison with, the stratified sampling method used for the filtering evaluation. Recall that our estimator of the proportion of relevant documents, based on a simple random sample from a set, is:

$$\hat{p} = \frac{a}{n} \tag{28}$$

where n is the size of the simple random sample, and a is the number of relevant documents in the sample. We cannot know the exact variance of this estimate without knowing the actual value of p, which is of course what we are trying to estimate in the first place. However, an unbiased estimate of the variance of our estimate of the proportion is [1, p. 52]:

$$\begin{aligned} \widehat{\mathrm{Var}[\hat{p}]} &= \frac{N-n}{(n-1)N} \times \hat{p} \times (1-\hat{p}) \\ &= \frac{N-n}{(n-1)N} \times \frac{a}{n} \times \frac{n-a}{n} \\ &= \frac{(N-n)a(n-a)}{n^2(n-1)N} \end{aligned} \tag{29}$$

Suppose we used a simple random sample of size n_i from set R_i to estimate the utility of set R_i. Then the MSE of the resulting utility estimate for set R_i would have been:

$$\begin{aligned} \mathrm{MSE}[\hat{u}_i] &= (u_{ai}-u_{bi})^2 {N_i}^2 \frac{(N_i-n_i)a_i(n_i-a_i)}{{n_i}^2(n_i-1)N_i} \\ &= (u_{ai}-u_{bi})^2 N_i \frac{(N_i-n_i)a_i(n_i-a_i)}{{n_i}^2(n_i-1)} \end{aligned} \tag{30}$$

NOTE: *The TREC-4 version of this paper had, instead of the above, the incorrect equation:*

$$\mathrm{MSE}[\hat{u}_i] = (u_{ai}-u_{bi})^2 \frac{(N_i-n_i)a_i(n_i-a_i)}{{n_i}^2(n_i-1)N_i}$$

7.2.2 Variance of Proportions Estimated by Stratified Sampling

In stratified sampling we separately estimate the proportion of relevant in each stratum and combine these estimates to get an estimate of the overall proportion:

$$\hat{p}_i = \sum_{h \in R_i} \frac{N_h}{N_i} \hat{p}_h \tag{31}$$

By the properties of the variance of linear combinations of random variables, and the fact that our samples from the strata are independent, we have [1, p. 92]:

$$\mathrm{Var}[\hat{p}_i] = \sum_{h \in R_i} \frac{{N_h}^2}{{N_i}^2} \mathrm{Var}[\hat{p}_h] \tag{32}$$

Each $\hat{p}_h$ is an estimate of the proportion of relevant in a stratum, based on a simple random sample from the stratum. Therefore, the results of the previous section tell us that an unbiased estimate of the variance of $\hat{p}_h$ is:

$$\widehat{\text{Var}}[\hat{p}_h] = \frac{(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)N_h} \tag{33}$$

Substituting Equation 33 into Equation 32 then gives us an unbiased estimate of the variance of our stratified estimate of the proportion of relevant in the R_i set:

$$\begin{aligned} \widehat{\text{Var}}[\hat{p}_i] &= \sum_{h \in R_i} \frac{{N_h}^2}{{N_i}^2} \frac{(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)N_h} \\ &= \frac{1}{{N_i}^2} \sum_{h \in R_i} \frac{N_h(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)} \end{aligned} \tag{34}$$

Further substituting Equation 34 into Equation 27 gives us the MSE for the estimate $\hat{u}_i$ (Equation 25) based on the stratified estimate of $\hat{p}_i$:

$$\begin{aligned} \text{MSE}[\hat{u}_i] &= (u_{ai} - u_{bi})^2 {N_i}^2 \text{Var}[\hat{p}_i] \\ &= (u_{ai} - u_{bi})^2 {N_i}^2 \frac{1}{{N_i}^2} \sum_{h \in R_i} \frac{N_h(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)} \\ &= (u_{ai} - u_{bi})^2 \sum_{h \in R_i} \frac{N_h(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)} \end{aligned} \tag{35}$$

If we compare Equation 35 to Equation 30, we see that the stratified estimate has a smaller MSE than an estimate based on simple random sampling when:

$$\sum_{h \in R_i} \frac{N_h(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)} < \frac{N(N - n)a(n - a)}{n^2(n - 1)}$$

This is almost always true when reasonable strata are defined and appropriately sized samples are chosen from those strata [1, p. 99].

NOTE: *The TREC-4 version of this paper had, instead of the above, the incorrect inequality:*

$$\sum_{h \in R_i} \frac{N_h(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)} < \frac{(N - n)a(n - a)}{n^2(n - 1)N}$$

8 Stratified Sampling: An Example (Part II)

Returning to our example, we can use Equation 35 to give the MSE's of the utility estimates in Equations 20–22:

$$\text{MSE}[\hat{u}_1] = 4^2(0 + 0 + 0 + \frac{64400}{26100}) = 39.5 \tag{36}$$

$$\text{MSE}[\hat{u}_2] = 2^2(0 + \frac{6800000}{26100} + 0 + \frac{64400}{26100}) = 1052.0 \quad (37)$$

$$\text{MSE}[\hat{u}_3] = 4^2(\frac{28130000}{26100} + \frac{6800000}{26100} + 0 + \frac{64400}{26100}) = 21452.5 \quad (38)$$

Recall that the $\hat{u}_i$ are unbiased, so the MSE of each estimate is just its variance, i.e. MSE[$\hat{u}_1$] = Var[$\hat{u}_i$]. Making the usually reasonable assumption that $\hat{u}_i$ has a roughly normal distribution, then a 95% confidence interval around $\hat{u}_i$ is [12, ch. 7]:

$$\hat{u}_i \pm 1.96\sqrt{\text{Var}[\hat{u}_i]}.$$

Then combining Equations 20–22 with Equations 36–38, and using the above expression for the confidence interval gives:

$$\begin{aligned} u_1 &= 2.672 \pm 12.3 \\ u_2 &= -51.35 \pm 63.6 \\ u_3 &= -717.2 \pm 287.1 \end{aligned}$$

We of course arranged this example so that the true utilities (Equations 17 to 19), which are known in our example but which would not be known in general, fell within the 95% confidence intervals. This would usually be the case in practice.

9 TREC-5 Results

The results of the TREC-5 filtering evaluation appear in an Appendix to these proceedings. The seven sites that participated in the filtering evaluation, and the code names for their ten filtering system submissions were City Univ. (*city96f*), ITI (*iti96f*), Intext (*INTXA* and *INTXM*), U Mass (*INR3*), U Illinois (*ispF*), Queens (*pircs96f*), and Xerox (*xerox.f1*, *xerox.f2*, and *xerox.f3*). In the Appendix, Table 1 for each site shows the raw data used in computing utility estimates. For each of the forty-nine topics and each of the three runs, we see the number of documents submitted and the pooled and stratified estimates of the utility of those submitted sets. Additional tables provide both summary and graphical presentations of this data.

Harman's summary of TREC-5 elsewhere in these proceedings discusses the nature of the topic set, training set, and test set used for the routing and filtering evaluations. From the standpoint of the filtering evaluation, the most notable characteristic of the data was that the distribution of relevant documents was more skewed than in TREC-4. There were more topics with few or no relevant documents found in the test set, and a few topics with a very large number of relevant documents. This skewing, combined with the variety of training data sets, most of unclear relation to the test data, made the filtering task quite challenging.

In the rest of this section we make a few observations on the approaches taken by TREC-5 filtering sites, briefly consider the effectiveness of various systems, and discuss how the properties of the TREC-5 data affected the methods used to estimate utility.

9.1 Approaches

As mentioned earlier, the TREC-5 filtering track used the same training and test data as the TREC-5 routing evaluation. Six of the seven filtering sites took advantage of this by basing their filtering runs on one of their routing runs. These sites produced routing queries by whatever means and used them to generate a score (implicitly or explicitly) for each of the test documents. The top 1000 scoring test documents were submitted (sorted by scores) for the routing evaluation, while some other processing of the scores was used to choose the test documents submitted in the site's filtering runs. The exception to this general strategy was Intext, which did not take part in the routing evaluation. However, it too computed numeric scores for all test documents. No purely boolean or other non-numeric methods for filtering were tried at TREC-5.

All sites except U Illinois applied machine learning techniques to the known relevant and non-relevant training documents in an attempt to produce a query better than the original routing topic description. Most of the algorithms used gave more weight to the training documents than to the original topic description, a reasonable strategy given the large amount of training data available.

The machine learning strategies used varied widely, and details can be found in the individual sites' papers. One point of interest was that three sites (City, Queens, and U Mass) explicitly tuned their routing queries to optimize average precision, a measure of ranking effectiveness that is the focus of the routing evaluation. Despite this emphasis on the routing evaluation, these sites also turned in three of the four best performances on the filtering evaluation.

While simply computing scores for documents was sufficient to rank them, additional processing was necessary to produce the submitted sets of documents required for filtering. Four sites (City, ITI, Queens, and U Mass) used the training data to set a threshold on their raw scores. Each routing query was applied to some or all of the training documents, and the documents were sorted by the resulting score. The sorted list was then scanned to find the score that, if used as a threshold, optimized the appropriate utility measure on the training documents. This procedure was repeated for each of the three utility measures to find a threshold for each of the filtering runs. The query was then used to score test documents, and all test documents exceeding the appropriate threshold went into the submitted set for that filtering run.

Intext and Xerox also generated their submitted sets by thresholding, but chose thresholds in different fashion. Xerox's logistic regression approach produces raw scores which are estimates of the probability of relevance of a document. The Xerox runs assumed these probability estimates were in fact correct, in which case the optimal thresholds follow from the definition of the utility measures (see Section 4.5). Intext found the highest scoring training document for each of their filtering queries and set thresholds at fixed percentages of this score. Finally, U Illinois did not use thresholding at all, but simply submitted a fixed number of documents for each utility measure, regardless of the topic.

9.2 Effectiveness

We leave the analysis of TREC-5 filtering results largely up to the track participants, and make only a few observations here. Table 1 shows the mean rank of each of the 10 systems for the three runs and two effectiveness estimates. The ordering of the systems by mean rank is for the most

System	Run 1 (thr. 0.75) Pooled	Run 1 (thr. 0.75) Strat	Run 2 (thr. 0.5) Pooled	Run 2 (thr. 0.5) Strat	Run 3 (thr. 0.25) Pooled	Run 3 (thr. 0.25) Strat	Best Routing Ave. Precision
INR3	3.61	3.49	3.55	3.31	2.71	2.84	0.3359
city96f	3.71	3.53	3.02	3.08	3.16	3.35	0.3475
xerox.f1	3.76	3.59	3.55	3.61	4.35	4.39	0.1223
pircs96f	4.10	4.12	3.24	3.33	3.88	4.02	0.3402
xerox.f2	4.14	4.10	3.59	3.71	4.12	4.29	0.1223
xerox.f3	5.14	5.14	4.78	4.92	5.02	5.16	0.1223
iti96f	6.37	6.45	6.82	6.86	7.14	6.65	0.1657
INTXA	6.53	6.59	6.92	6.92	6.80	6.65	-
INTXM	7.55	7.69	8.61	8.37	7.67	7.43	-
ispF	8.88	9.10	9.45	9.45	9.24	9.35	0.0196

Table 1: Mean rank of the 10 filtering systems under 6 evaluation conditions produced by pairing 3 runs with 2 estimation methods. (We show the optimal threshold on probability of relevance/instantaneous precision p for each run.) For each condition, the ten systems were sorted by utility and given a rank between 1 and 10 for each of the 49 topics. (If a group of systems had identical utilities for a topic their ranks were replaced by the mean rank for the group.) The mean of those ranks, over the 49 topics, is shown in this table. Systems are sorted by the mean over the resulting six means. The last column shows that the best average precision achieved by the site in the routing evaluation is, except for Xerox, roughly correlated with the effectiveness of the site's filtering runs.

part consistent across runs and effectiveness measures. The last column of Table 1 shows the best average precision (over 45 topics) for the routing runs submitted by each filtering site that did the routing task. (We did not attempt to establish a correspondence between particular routing submissions and particular filtering submissions.) Roughly speaking, the sites that did well at routing did well at filtering, and vice versa. Xerox was a bit of an exception to this, raising the interesting possibility that probability estimates that aren't good enough for ranking may be good enough to do reasonable binary classification.

9.3 Estimating Utilities

The widely varying number of relevant documents for different filtering topics stressed our estimation methods in interesting ways, as reflected in the relatively large number of cases where the pooled estimate of utility falls outside the 95% confidence interval for the corresponding stratified estimate of utility. In Table 2, we show the number of times the pooled estimate is above or below the outer limit of confidence interval on the stratified estimate. (The raw data is taken from Table 1 for each site in the filtering Appendix.)

These large disagreements between the pooled and stratified estimates occur in two very different situations. First, there are 48 cases where the pooled estimate is higher than the upper end of the confidence interval for the stratified estimate. At first glance, this seems quite surprising, since the pooled estimate is guaranteed to be an *under*estimate of the true utility, while the stratified estimate is an unbiased estimate of true utility.

The fault here is actually in the confidence interval estimation. As the parenthesized values in Table 2 show, in 32 of these 48 cases the estimated standard deviation is 0, so the confidence interval degenerates to a point. One way we can get a degenerate confidence interval is when the submitted set has 100 or fewer items. In this case, all documents in the set are judged and there is no sampling error. However, in this case the stratified and pooled estimates will always agree exactly.

The other, pernicious, way we can get a degenerate confidence interval is when no relevant documents are found in the submitted set. This leads to a statistically unbiased, but usually wrong, estimate that the sampling error is 0. Table 3 shows that when the pooled estimate is higher than the top of confidence interval, the submitted set tends to be large (median size of 256.0), and the number of relevant in the submitted set (as estimated by pooling) tends to be small (median 6.5). This is the worst case for our sampling method: attempting to estimate a small value (the proportion of relevant) by drawing a small sample from a large set.

In this situation, the likely result is that we will see no relevant documents in some or all strata. When no relevant documents are found in any strata we get a degenerate confidence interval, as seen in the 32 cases mentioned above. While we have not analyzed the remaining 16 cases, it is likely that many of them result from finding relevant documents in some strata but not others. This would result in an estimated standard deviation which is nonzero, but usually too low. In other words, our stratified estimate is less accurate than it appears. The pooled estimate, which uses a much larger sample size, is likely to be more accurate. This is likely to be true for some cases where the pooled estimate is within the confidence interval for the stratified estimate as well.

On the other hand, the 62 cases where the pooled estimates are lower than the bottom of the

	Run 1 (thr. 0.75)	Run 2 (thr. 0.5)	Run 3 (thr. 0.25)	Totals
$\hat{u}_{i,p} < \hat{u}_i - 1.96\sqrt{\text{Var}[\hat{u}_{i,s}]}$	11 (0)	20 (0)	31 (0)	62 (0)
$\hat{u}_i - 1.96\sqrt{\text{Var}[\hat{u}_{i,s}]} <= \hat{u}_{i,p} < \hat{u}_{i,s}$	17 (0)	49 (0)	56 (0)	122 (0)
$\hat{u}_{i,s} = \hat{u}_{i,p} = \hat{u}_{i,s}$	447 (447)	364 (364)	344 (344)	1155 (1155)
$\hat{u}_{i,s} < \hat{u}_{i,p} <= \hat{u}_i + 1.96\sqrt{\text{Var}[\hat{u}_{i,s}]}$	13 (0)	33 (0)	37 (0)	83 (0)
$\hat{u}_i + 1.96\sqrt{\text{Var}[\hat{u}_{i,s}]} < \hat{u}_{i,p}$	2 (2)	24 (16)	22 (14)	48 (32)
Totals	490 (449)	490 (380)	490 (358)	1470 (1187)

Table 2: Comparison of pooled estimate of utility, $\hat{u}_{i,p}$, with stratified estimate of utility, $\hat{u}_{i,s}$ for filtering submissions. There are 10 systems, 3 runs (column header shows threshold on probability of relevance for run), and 49 topics for a total of 1470 pairs of estimates. We break this total down by runs and by the relationship of $\hat{u}_{i,p}$ to $\hat{u}_{i,s}$ and to the bounds of a 95% confidence interval on $\hat{u}_{i,s}$. In parentheses we show the number of times the confidence interval on $\hat{u}_{i,s}$ degenerates to a point.

confidence interval on the stratified estimates exhibit a strength of the stratified approach. As Table 3 shows, the submitted sets are typically large here also (median size 562.0), but so is the number of relevant in the submitted set (median value 282.5 as estimated by pooling). These 62 cases involve only 7 topics, which are 7 of the 9 topics with the highest number of known relevant documents. (In fact, 58 of the cases occur on the four topics where the pooled sample reveals there are at least 500 relevant documents.) This is a situation where the stratified estimates of both utility and sample variance will be quite accurate. In contrast, the pooling assumption (that unjudged documents are non-relevant) is most likely to be wrong in these cases. For instance, our stratified sample lets us say with high confidence that the set submitted by system *iti96f* for Run 3 on Topic 111 included 2254 ± 441.1 relevant documents. Since the entire judged pool for this topic contained only 1303 documents (887 of which were found to be relevant) the pooled estimate cannot help but substantially underestimate the true utility.

10 Future Filtering Tracks

The TREC-5 filtering track was an advance over the TREC-4 filtering track from the standpoint of planning (the evaluation procedures were finalized well before participants retrieved the data sets, unlike in TREC-4) and participation (seven sites participating vs. four sites for TREC-4). However, much remains to be improved. First, as pointed out in Section 4.5, all the filtering track runs required high precision by the standards of current IR systems. This fact, combined with the small number of relevant documents available for many topics, led to a situation where the optimal behavior for most systems on many topics was to submit a set of size 0. This is obviously not conducive to either comparing systems or understanding the behavior of a single system. In future filtering evaluations it will be desirable to adjust the data sets and/or the effectiveness measures to

	Median Submitted Set Size	Median Relevant in Submitted Set (pooled estimate)	Median Relevant in Pool
$\hat{u}_{i,p} < \hat{u}_i - 1.96\sqrt{\text{Var}[\hat{u}_{i,s}]}$	562.0	282.5	808.0
all runs	10.0	2.0	42.0
$\hat{u}_i + 1.96\sqrt{\text{Var}[\hat{u}_{i,s}]} < \hat{u}_{i,p}$	256.0	6.5	20.0

Table 3: Statistics on submitted sets where the pooled estimate of utility ($\hat{u}_{i,p}$) is outside the 95% confidence interval for the stratified estimate of utility ($\hat{u}_{i,s}$). We compare with corresponding values for the collection of all submitted sets for all topics, systems, and the three runs. We show median values for the size of the submitted set, the number of relevant in submitted set (estimated using the pooling assumption), and the number of relevant in the judged pool for the corresponding topic.

avoid this situation. In any case, utility-based measures make it difficult to compute the average effectiveness of systems across topics, and it is unclear how well they capture user needs. So non-utility measures need to be investigated.

Our estimation procedures could be improved as well. A more careful choice of sample sizes in stratified sampling could improve our estimates there. It might also be possible to increase the number of documents judged for filtering submissions, since there is substantial redundancy between them and routing submissions from the same sites.

Since most sites have chosen to base their filtering systems very closely on their routing systems, it seems worth investigating more closely the relationship between the two. For instance, Chris Buckley has suggested examining the full rankings produced by filtering systems to find the optimal score that *could* have been produced, as a measure of how well systems are setting their thresholds.

Finally, the TREC filtering and routing evaluations have been criticized as being unrealistic, particularly with respect to the training data supplied. On the one hand, the task is made too easy, by supplying more training data, with a higher density of relevant documents, than would be available in most real routing or binary classification applications. On the other hand, the task is too difficult, in that training data is drawn from sources that have little or no relation to the test data, and the procedures used to obtain the training data (pooling of previous years ad hoc runs) are complex and poorly understood. One goal for future evaluations will be to provide more realistic training and test data, ideally drawn from a chronologically ordered stream of documents.

11 Summary

The TREC-5 filtering track solidified the role of an evaluation of binary text classification in TREC. Seven sites participated in the track, producing data both on their effectiveness of their systems and on the appropriateness of evaluation strategies. We encourage all TREC participants to consider taking part in future TREC filtering evaluations.

12 Acknowledgments

I am greatly appreciative of Donna Harman, Ellen Voorhees, and the rest of the team at NIST, for both their work in conducting the filtering evaluation and their suggestions for improving it. The ideas presented here were developed in extensive discussions with the members of the TREC-1 through TREC-5 program committees (particularly Chris Buckley, David Hull, and Karen Sparck Jones), and with TREC-4 and TREC-5 participants (particularly David Hull, Paul Kantor, K. L. Kwok, and Julian Yochum).

References

[1] William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, 3rd edition, 1977.

[2] William S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24:87–100, March–April 1973.

[3] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience, New York, 1973.

[4] Donna Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman, editor, *The Fourth Text REtrieval Conference (TREC-4)*, Gaithersburg, MD, 1996. U. S. Dept. of Commerce, National Institute of Standards and Technology.

[5] Marti Hearst, Jan Pedersen, Peter Pirolli, Hinrich Schütze, Gregory Grefenstette, and David Hull. Xerox site report: Four TREC-4 tracks. In D. K. Harman, editor, *The Fourth Text REtrieval Conference (TREC-4)*, pages 97–119, Gaithersburg, MD, 1996. U. S. Dept. of Commerce, National Institute of Standards and Technology.

[6] Paul S. Jacobs. GE in TREC-2: Results of a boolean approximation method for routing and retrieval. In D. K. Harman, editor, *The Second Text Retrieval Conference (TREC-2)*, pages 191–199, Gaithersburg, MD, March 1994. U. S. Dept. of Commerce, National Institute of Standards and Technology. NIST Special Publication 500-215.

[7] K. L. Kwok, L. Grunfeld, and D. D. Lewis. TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 247–255, Gaithersburg, MD, April 1995. U. S. Dept. of Commerce, National Institute of Standards and Technology.

[8] David D. Lewis. Evaluating text categorization. In *Proceedings of Speech and Natural Language Workshop*, pages 312–318. Defense Advanced Research Projects Agency, Morgan Kaufmann, February 1991.

[9] David D. Lewis. Evaluating and optimizing autonomous text classification systems. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 246–254, New York, 1995. Association for Computing Machinery.

[10] David D. Lewis. The TREC-4 filtering track. In D. K. Harman, editor, *The Fourth Text REtrieval Conference (TREC-4)*, pages 165–180, Gaithersburg, MD, 1996. U. S. Dept. of Commerce, National Institute of Standards and Technology.

[11] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In W. Bruce Croft and C. J. van Rijsbergen, editors, *SIGIR 94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, London, 1994. Springer-Verlag.

[12] David S. Moore and George P. McCabe. *Introduction to the Practice of Statistics*. W. H. Freeman, New York, 1989.

[13] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, December 1977.

[14] Richard M. Tong and Lee A. Appelbaum. Machine learning for knowledge-based document routing (a report on the TREC-2 experiment). In D. K. Harman, editor, *The Second Text Retrieval Conference (TREC-2)*, pages 253–264, Gaithersburg, MD, March 1994. U. S. Dept. of Commerce, National Institute of Standards and Technology. NIST Special Publication 500-215.