The TREC-4 Filtering Track

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

Appeared, with different pagination, in D. K. Harman, ed., *The Fourth Text Retrieval Conference (TREC-4).* National Institute of Standards and Technology, Gaithersburg, MD, 1996, pp. 165–180.

Abstract

The TREC-4 filtering track was an experiment in the evaluation of binary text classification systems. In contrast to ranking systems, binary text classification systems may need to produce result sets of any size, requiring that sampling be used to estimate their effectiveness. We present an effectiveness measure based on utility, and two sampling strategies (pooling and stratified sampling) for estimating utility of submitted sets. An evaluation of four sites was successfully carried out using this approach.

1 Introduction

The goal of the TREC-4 filtering track was to develop methods for evaluating binary text classification systems, and try out those methods on real data. A secondary goal was to give TREC participants their first chance to evaluate approaches to binary text classification.

This paper begins by defining binary text classification and presenting some applications of it. We then discuss a particular binary text classification task, filtering, used in TREC-4. 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-4 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-4 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 [7]. 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 [6, 7], and so here will concentrate on how it differs from the evaluation of ranked retrieval in the main TREC-4 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).
- R-precision : This is precision at R documents, where R is the number of relevant documents in the collection.

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 how likely they are to be relevant. 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 [7, 10].

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-4, 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-4 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 [5, 7], but also for self-monitoring of effectiveness [7] and selection of training data [8].

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 [4, 11]. These approaches include boolean querying by human experts, as well as the use of binary classifiers produced by machine learning techniques.

4.2 Structure

The structure of the filtering evaluation was as follows. Systems were given 150 descriptions of user needs. Each description was made up of a *topic* describing the kind of information sought, and a *set-based effectiveness measure* capturing a hypothetical user's tolerance for different kinds of mistakes. The topics were the same 50 topics used in the main TREC-4 routing evaluation. Three set-based effectiveness measures (see Section 4.4) were combined with each of the 50 topics, to yield the 150 user needs.

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-4 filtering and routing evaluations used the same topics and test documents. (See Harman's discussion of these 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 some filtering and routing submissions overlapped.

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, 7]. 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 is willing to pay 1 dollar (or pick your favorite currency) for each relevant document, and 3 dollars to avoid having 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 high recall, 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 [7].) Unlike both recall and precision, the total utility is not normalized to lie between 0 and 1. Indeed, there is no way to know what the maximum achievable is for any given topic, since it depends on the (unknown) total number of relevant documents in the test set. The goal of systems is simply to achieve the highest utility they can.

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 tried in the TREC-4 filtering track, as described below.

5.1 Pooling

The first approach to sampling was the usual TREC pooling strategy [3]. This approach assumes that some known pool of documents contains all the relevant documents in the test set. The pool for the TREC-4 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:

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

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:

$$\begin{aligned} u_1 &= (4p_1 - 3) N_1 \qquad (2) \\ u_2 &= (2p_2 - 1) N_2 \qquad (3) \\ u_3 &= (4p_3 - 1) N_3 \qquad (4) \\ & \qquad (5) \end{aligned}$$

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:

$$\hat{u}_i = ((u_{ai} - u_{bi}) \hat{p}_i + u_{bi}) N_i \qquad (6)$$

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:

$$\hat{p} = \frac{a}{n} \qquad (7)$$

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. In *stratified sampling* we use additional knowledge about a population to divide the population into groups or *strata* [1, p.

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.

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.

We stratified the set of filtering documents submitted by each TREC-4 site for a topic 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]:

$$\hat{p}_i = \sum_{h \in R_i} \frac{N_h}{N_i} \hat{p}_h \tag{8}$$

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:

$$\hat{p}_1 = \frac{N_{100}}{N_1} \times \hat{p}_{100} + \frac{N_{101}}{N_1} \times \hat{p}_{101} + \frac{N_{110}}{N_1} \times \hat{p}_{110} + \frac{N_{111}}{N_1} \times \hat{p}_{111} \tag{9}$$

$$\hat{p}_2 = \frac{N_{010}}{N_2} \times \hat{p}_{010} + \frac{N_{011}}{N_2} \times \hat{p}_{011} + \frac{N_{110}}{N_2} \times \hat{p}_{110} + \frac{N_{111}}{N_2} \times \hat{p}_{111} \tag{10}$$

$$\hat{p}_3 = \frac{N_{001}}{N_3} \times \hat{p}_{001} + \frac{N_{011}}{N_3} \times \hat{p}_{011} + \frac{N_{101}}{N_3} \times \hat{p}_{101} + \frac{N_{111}}{N_3} \times \hat{p}_{111} \tag{11}$$

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-4

To be consistent with the TREC-4 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 TREC-4 filtering site for a single topic. 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:

$$p_1 = \frac{30}{30 + 10} = .7500 \tag{12}$$

$$p_2 = \frac{112}{112 + 138} = .4480 \tag{13}$$

Stratum	Submitted Docs Rels	NonRels	True Prop	Samples from Strata 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 the number of relevant and nonrelevant documents in each stratum and run, and the corresponding proportion of relevant.

$$p_3 = \frac{130}{130+1110} = .1048 \tag{14}$$

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. We then combine the stratum estimates, using Equations 9 to 11, to get estimates of the proportion of relevant in each submitted set:

$$\hat{p}_1 = \frac{40}{40} \times \frac{23}{23+7} = .7667 \tag{15}$$

$$\hat{p}_2 = \frac{10}{250} \times \frac{2}{2+8} + \frac{200}{250} \times \frac{10}{10+20} + \frac{40}{250} \times \frac{23}{23+7} = .3973 \tag{16}$$

$$\hat{p}_3 = \frac{1000}{1240} \times \frac{1}{1+29} + \frac{200}{1240} \times \frac{10}{10+20} + \frac{40}{1240} \times \frac{23}{23+7} = .1054 \tag{17}$$

6.2 Estimating Utility of a Submitted Set

If we knew the true proportion of relevant documents in each submitted set (Equations 12 to 14), we could compute the true utility of each set, using Equations 2 to 4:

$$u_1 = (4 \times .7500 - 3) \times 40 = 0.0 \tag{18}$$

$$u_2 = (2 \times .4480 - 1) \times 250 = -26.0 \tag{19}$$

$$u_3 = (4 \times .1048 - 1) \times 1240 = -719.2 \tag{20}$$

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

$$\hat{u}_1 = (4 \times .7667 - 3) \times 40 = 2.672 \tag{21}$$
$$\hat{u}_2 = (2 \times .3973 - 1) \times 250 = -51.35 \tag{22}$$
$$\hat{u}_3 = (4 \times .1054 - 1) \times 1240 = -717.2 \tag{23}$$

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:

$$\mathrm{MSE}[\hat{\mu}] = \mathrm{E}[(\hat{\mu} - \mu)^2] \tag{24}$$

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 has a nonzero bias as well, 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 the true utility, but the actual value of the pooled estimate is always less than 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:

$$u_i = ((u_{ai} - u_{bi})p_i + u_{bi})N_i \tag{25}$$

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

$$\hat{u}_i = ((u_{ai} - u_{bi})\hat{p}_i + u_{bi})N_i \tag{26}$$

The MSE of such an estimate is

$$\begin{aligned}
\mathrm{MSE}[\hat{u}_i] &= \mathrm{E}[(\hat{u}_i - u_i)^2] \\
&= \mathrm{E}[(((u_{ai} - u_{bi})\hat{p}_i + u_{bi})N_i - ((u_{ai} - u_{bi})p_i + u_{bi})N_i)^2] \\
&= \mathrm{E}[(u_{ai} - u_{bi})^2 {N_i}^2 (\hat{p}_i - p_i)^2] \\
&= (u_{ai} - u_{bi})^2 {N_i}^2 \mathrm{E}[(\hat{p}_i - p_i)^2] \\
&= (u_{ai} - u_{bi})^2 {N_i}^2 \mathrm{MSE}[\hat{p}_i].
\end{aligned} \tag{27}$$

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:

$$\mathrm{MSE}[\hat{u}_i] = (u_{ai} - u_{bi})^2 {N_i}^2 \mathrm{Var}[\hat{p}_i]. \tag{28}$$

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:

$$\hat{p} = \frac{a}{n} \tag{29}$$

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]:

$$\begin{aligned}\widehat{\mathrm{Var}[\hat{p}]} &= \frac{N-n}{(n-1)N} \times \hat{p} \times (1-\hat{p}) \\ &= \frac{N-n}{(n-1)N} \times \frac{a}{n} \times \frac{n-a}{n} \\ &= \frac{(N-n)a(n-a)}{n^2(n-1)N} \end{aligned} \tag{30}$$

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:

$$\begin{aligned}\mathrm{MSE}[\hat{u}_i] &= (u_{ai} - u_{bi})^2 {N_i}^2 \frac{(N_i - n_i)a_i(n_i - a_i)}{{n_i}^2(n_i - 1)N_i} \\ &= (u_{ai} - u_{bi})^2 N_i \frac{(N_i - n_i)a_i(n_i - a_i)}{{n_i}^2(n_i - 1)} \end{aligned} \tag{31}$$

NOTE: *The published 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:

$$\hat{p}_i = \sum_{h \in R_i} \frac{N_h}{N_i} \hat{p}_h \tag{32}$$

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]:

$$\mathrm{Var}[\hat{p}_i] = \sum_{h \in R_i} \frac{{N_h}^2}{{N_i}^2} \mathrm{Var}[\hat{p}_h] \tag{33}$$

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:

$$\widehat{\mathrm{Var}}[\hat{p}_h] = \frac{(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)N_h} \tag{34}$$

Substituting Equation 34 into Equation 33 then gives us an unbiased estimate of the variance of our stratified estimate of the proportion of relevant in the R_i set:

$$\begin{aligned} \widehat{\mathrm{Var}}[\hat{p}_i] &= \textstyle\sum_{h \in R_i} \frac{{N_h}^2}{{N_i}^2} \frac{(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)N_h} \\ &= \frac{1}{{N_i}^2} \textstyle\sum_{h \in R_i} \frac{N_h(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)} \end{aligned} \tag{35}$$

Further substituting Equation 35 into Equation 28 gives us the MSE for the estimate $\hat{u}_i$ (Equation 26) based on the stratified estimate of $\hat{p}_i$:

$$\begin{aligned} \mathrm{MSE}[\hat{u}_i] &= (u_{ai} - u_{bi})^2 {N_i}^2 \mathrm{Var}[\hat{p}_i] \\ &= (u_{ai} - u_{bi})^2 {N_i}^2 \frac{1}{{N_i}^2} \textstyle\sum_{h \in R_i} \frac{N_h(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)} \\ &= (u_{ai} - u_{bi})^2 \textstyle\sum_{h \in R_i} \frac{N_h(N_h - n_h)a_h(n_h - a_h)}{{n_h}^2(n_h - 1)} \end{aligned} \tag{36}$$

If we compare Equation 36 to Equation 31, we see that the stratified estimate has a smaller MSE than an estimate based on simple random sampling when:

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

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

NOTE: *The published version of this paper had, instead of the above, the incorrect equation:*

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

8 Stratified Sampling: An Example (Part II)

Returning to our example, we can use Equation 36 to give the MSE's of the utility estimates in Equations 21–23:

$$\mathrm{MSE}[\hat{u}_1] = 4^2(0 + 0 + 0 + \frac{64400}{26100}) = 39.5 \tag{37}$$

$$\mathrm{MSE}[\hat{u}_2] = 2^2(0 + \frac{6800000}{26100} + 0 + \frac{64400}{26100}) = 1052.0 \tag{38}$$

$$\mathrm{MSE}[\hat{u}_3] = 4^2(\frac{28130000}{26100} + \frac{6800000}{26100} + 0 + \frac{64400}{26100}) = 21452.5 \tag{39}$$

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 reasonable assumption that $\hat{u}_i$ has a roughly normal distribution, then a 95% confidence interval around $\hat{u}_i$ is [9, ch. 7]:

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

Then combining Equations 21–23 with Equations 37–39, 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 18 to 20), 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 Discussion

The results of the TREC-4 filtering evaluation appear in an appendix to these proceedings. Table 1 for each site shows the raw data used in computing utility estimates. For each topic 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.

One reassuring observation is that the pooled estimates in Table 1 for each site are almost always within the confidence intervals for the stratified estimates. This suggests that the bias of the pooled estimates is not large in a statistical sense. Since the pooled utility provides a lower bound on the true utility, combining the two estimates can be useful. For instance, Topic 45, Run 2 for the *pircs* (Queen's) system has a stratified estimate of 16.8 ± 12.8, but the pooled estimate of 16.0 shows that the true utility is not in the lower part of this range.

In many cases the confidence intervals have an estimated margin of error [9, p. 451] of ± 0.0. This can occur for two very different reasons. First, the submitted set may have been small enough that the entire set could be included in the stratified sample. In this case, we have an exact value for that set's utility, not an estimate, and so there is no sampling error.

In contrast, a margin of error of ± 0.0 can also arise because the variances which appear in the confidence intervals are themselves only estimates. In particular, if the random sample from a stratum contains no relevant documents, or only relevant documents, the *estimated* variance will be 0.0, even though the true variance due to sampling, if known, would be higher than this.

For example, Topic 30, Run 3 for Xerox had a submitted set of size 561 and an estimated utility of -235.3 ± 144.5. The margin of error is large because relatively few documents from the set could be judged. Topic 50, Run 3 for Xerox also had a large submitted set. Its margin of error is given as ± 0.0 because no relevant documents appeared in the stratified sample. While 0.0 is an unbiased estimate of the variance, we can rightly be skeptical of this value. A more extreme case is Topic 142, Run 3 for HNC where the confidence interval for the stratified estimate is -47.0 ± 0.0. However, the pooled estimate, which is a lower bound on the true utility, is 33.0, well outside the (degenerate) confidence interval.

In most cases, however, the margins of error are both small and believable. The largest submitted set was 809 documents, putting to rest worries that submitted sets with sizes in the thousands would cause both statistical and practical problems for the evaluation.

All sites attempted to adjust the size of submitted sets in response to the utility measures for the three runs. As the papers from the sites discuss, doing this in an optimal fashion is difficult and is likely to be the focus of future research. Achieving good results in run R_1, where a precision of over 0.75 was required to achieve a positive utility, was particularly difficult. All sites had many topics where they would have been better off submitting no documents than submitting their run R_1 set.

Our policy of drawing the stratified sample evenly from all strata (see Section 5.3) resulted in relatively tight confidence intervals on the run R_1 and R_2 stratified estimates, and relatively loose intervals on run R_3. In future evaluations, it may be desirable to draw more documents from the run R_3 strata, to make the confidence intervals somewhat more balanced.

10 Summary

The TREC-4 filtering track broke new ground for TREC by introducing a binary classification task and a set of new evaluation procedures. The sites who participated in the track (HNC, NSA, Queen's, and Xerox), had to tolerate considerable uncertainty and are to be congratulated for their perseverance. The filtering track will be run again in TREC-5, with much the same structure as in TREC-4, and we encourage sites to take part.

11 Acknowledgments

I am greatly appreciative of Donna Harman and her team at NIST, for both their work in conducting the filtering evaluation and their suggestions for improving it. Gavin O'Brien in particular put great effort into generating and presenting the filtering results. The ideas presented here were developed in extensive discussions with the members of the TREC-1 through TREC-4 program committees, and with TREC-4 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

[11] 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.