Evaluating Evaluation Measure Stability

Chris Buckley
Sabir Research Inc.
Gaithersburg, MD 20878
chrisb@sabir.com

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
ellen.voorhees@nist.gov

Abstract

This paper presents a novel way of examining the accuracy of the evaluation measures commonly used in information retrieval experiments. It validates several of the rules-of-thumb experimenters use, such as the number of queries needed for a good experiment is at least 25 and 50 is better, while challenging other beliefs, such as the common evaluation measures are equally reliable. As an example, we show that Precision at 30 documents has about twice the average error rate as Average Precision has. These results can help information retrieval researchers design experiments that provide a desired level of confidence in their results. In particular, we suggest researchers using Web measures such as Precision at 10 documents will need to use many more than 50 queries or will have to require two methods to have a very large difference in evaluation scores before concluding that the two methods are actually different.

1 Introduction

Information retrieval has a well-established tradition of performing laboratory experiments on test collections to compare the relative effectiveness of different retrieval approaches. The experimental design specifies the evaluation criterion to be used to determine if one approach is better than another. Because retrieval behavior is sufficiently complex to be difficult to summarize in one number, many different effectiveness measures have been proposed. For example, `trec_eval`, the evaluation software used in the Text REtrieval Conference, reports over 85 numbers based on 20-odd different measures. However, very little attention has been paid to exploring how the properties of a particular evaluation measure support conclusions as to whether one retrieval method is better than another.

In addition to the different evaluation measures, the long tradition of experimentation has evolved a number of rules-of-thumb for acceptable experimental design, including:

- The test collection must have a reasonable number of requests. Sparck Jones and van Rijsbergen suggested a minimum of 75 requests [17], while the TREC program committee has used 25 requests as a minimum and 50 requests as the norm [24]. Five or ten requests is too few [23].

- The experiment must use a reasonable evaluation measure. Average Precision, R-Precision, and Precision at 20 (or 10 or 30) documents retrieved are the most commonly used measures. Measures such as Precision at one document retrieved (i.e., is the first retrieved document relevant?) or the rank of the first relevant document are not usually reasonable evaluation measures [25].

- Conclusions must be based on a reasonable notion of difference. Sparck Jones has suggested that a difference in the scores between two runs that is greater than 5% is noticeable, and a difference that is greater than 10% is material [18].

The reader will notice an overabundance of the terms "reasonable" and "usually" in these rules-of-thumb. While the rules have evolved through the collective experience of the community, they have not been examined in depth, and are seldom explicitly stated. This poses an obstacle for newcomers to the field who are unlikely to know the rules-of-thumb or have sufficient experience to develop a feel for what is "reasonable". It also may be the case that an analysis of the rules will suggest a better experimental methodology for experienced researchers.

This paper examines these three rules-of-thumb and shows how they interact with each other. We present a novel approach for experimentally quantifying the likely error associated with the conclusion "method A is better than method B" given a number of requests, an evaluation measure, and a notion of difference. As expected, the error rate increases as the number of requests decreases. More surprisingly, we demonstrate a striking difference in error rates for various evaluation measures. For example, Precision at 30 documents retrieved has almost twice the error rate of Average Precision. These results do *not* imply that measures with higher error rates should not be used; different evaluation measures evaluate different aspects of retrieval behavior and evaluation measures must be chosen to match the goals of the test. The results do mean that for a researcher to be equally confident in the conclusion that one method is better than another, experiments based on measures with higher error rates require either more requests or larger differences in scores than experiments based on measures with lower error rates.

The paper is organized as follows. The next section provides a very brief summary of previous work in retrieval system evaluation. Section 3 presents the test environment used in this study, defining the evaluation measures examined and describing how the data was collected. The error rates for the evaluation measures are then computed in Section 4. Section 5 examines the implications of the differences in error rates and the limits of this study, while the final section suggests ways to extend the work.

2 Retrieval System Evaluation

Evaluation of retrieval system performance has been an integral part of the field since its beginning, but can be difficult to do well. Tague catalogs dozens of decisions that are required to design and execute a valid, efficient, and reliable retrieval test [19, 20]. A common way of simplifying the experimental process is to perform laboratory tests using test collections, a tradition started by the Cranfield tests [3]. A test collection consists of a set of documents, a set of topics, and a set of relevance judgments. A topic is a description of the information being sought. The relevance judgments specify the documents that should be retrieved in response to each topic. In this paradigm, the effectiveness of different retrieval mechanisms can be directly compared on the common task defined by the test collection [16].

At least two questions remain when constraining retrieval experimentation to laboratory tests using test collections: how to build and validate good test collections, and what measure(s) should be used to assess the effectiveness of retrieval output. The first question was addressed by Sparck Jones and van Rijsbergen who listed a set of criteria that an ideal test collection would meet [17]. The test collections created through the TREC workshops have been validated by demonstrating the stability of relative retrieval scores despite incomplete relevance judgments [8, 28] and different opinions as to what constitutes a relevant document [23]. Zobel [28] and Cormack, Palmer, and Clarke [5] proposed methods for efficiently building large test collections.

The second question—what measures should be used to evaluate retrieval effectiveness—has received enormous attention in the literature. van Rijsbergen [22] contains a good summary, while Keen [11] gives a detailed account on how to present retrieval results. Different evaluation measures have different properties with respect to how closely correlated they are with user satisfaction criteria, how easy they are to interpret, how meaningful aggregates such as as average values are, and how much power they have to discriminate among retrieval results.

Most retrieval evaluation measures are derived in some way from *recall* and *precision*, where precision is the proportion of retrieved documents that are relevant, and recall is the proportion of relevant documents that are retrieved. An exception are measures based on utility-theory [4, 13] for which the quality of retrieval output is measured in terms of its worth to the user. Utility-based measures are frequently used to evaluate set-based retrieval output such as in the TREC filtering task [14].

While many different evaluation measures have been defined and used, differences among measures have almost always been discussed based on their principles. That is, there has been very little empirical examination of the measures themselves. Correlation studies that build equivalence classes of measures based on how similarly they rank systems are one type of empirical study [21, 25]. In this paper we perform a different empirical study to quantify how stable evaluation measures are.

3 Test Environment

Our goal is to help researchers design effective retrieval experiments. In the experimental paradigm assumed in this paper, each retrieval strategy to be compared produces a ranked list of documents for each topic in a test collection, where the list is ordered by decreasing likelihood that the document should be retrieved for that topic. The effectiveness of a strategy for a single topic is computed as a function of the ranks of the relevant documents. The effectiveness of the strategy on the whole is then computed as the average score across the set of topics in the test collection.

The following measures are investigated in this paper:

Prec(λ): Precision at cut-off level λ, for λ = 1, 2, 5, 10, 15, 20, 30, 50, 100, 300, 1000. A *cut-off* level is a rank that defines the retrieved set; for example, a cut-off level of ten defines the retrieved set as the top ten documents in the ranked list.

Recall(1000): Recall after 1000 documents have been retrieved.

Prec at .5 Recall: Precision after half the relevant document have been retrieved. This is the value that would get plotted at .5 recall on a standard recall-precision graph. Precision at .5 recall is an interpolated value since an exact value is undefined for topics that have an odd number of relevant documents [22, 27].

R-Prec: Precision after R documents have been retrieved where R is the number of relevant documents for the current topic.

Average Precision: The mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved.

Since the purpose of the retrieval experiments is to reach general conclusions regarding the relative effectiveness of retrieval strategies, researchers would like to have confidence intervals on the reliability of the conclusions drawn from test scores. For example, if a test result says that method A is better than method B since their scores for measure M differ by 5%, then the researcher would like to be confident that in equivalent environments method B would be better than method A no more than 2% of the time. Unfortunately, there are significant difficulties in establishing such error rates for both IR experiments and evaluation measures. In the physical sciences, a researcher can repeat an experiment and compare the various results. In IR, exactly repeating an experiment is useless since IR systems are deterministic and will thus always produce the same result from the same starting conditions. But changing either the topic set or the document collection is not a viable option because these factors directly affect retrieval behavior. That is, the behavior of retrieval methods—either entire systems or variant algorithms within the same basic system—depends on the number of relevant documents. As an example, methods such as blind feedback work well for broad topics that have many relevant documents but may harm topics with few relevant documents. Evaluation measures are even more dependent on the number of relevant documents. For example, with perfect retrieval Prec(100) changes from 0.2 to 0.3 if the number of relevant documents changes from 20 to 30.

We would have a situation analogous to the physical measurements, and thus could calculate confidence intervals on retrieval evaluation results, if we had multiple expressions of each topic. A *query* is the expression of a topic that is actually processed by a retrieval system. Using different queries does affect retrieval behavior—some queries are better expressions of the topic than others—but the effect of the number of relevant documents is controlled since it remains constant. By varying the expression of a topic and observing how evaluation behavior changes, we can calculate the error associated with an evaluation measure.

The TREC-8 Query Track provides the data needed for such a study. The Query Track was designed as a means for creating a large set of different queries for TREC topics 51–100. Track participants created different query versions using three different query types:

- two or three words extracted from the topic statement;

Label	Organization	Approach
APL	APL at Johns Hopkins U.	APL system
INQa	U. of Massachusetts	INQUERY, words only
INQe	U. of Massachusetts	INQUERY, words with query structure and expansion
INQp	U. of Massachusetts	INQUERY, words with query structure
Saba	Sabir Research	SMART, words only
Sabe	Sabir Research	SMART, words with full expansion
Sabm	Sabir Research	SMART, words with modest expansion
acs	ACSys, Australian National U.	PADRE system
pir	Queens College, CUNY	PIRCS system

Table 1: Retrieval methods used in the TREC-8 Query Track.

- an English sentence based on the topic statement and (possibly) the relevant documents in a training set of documents;
- an English sentence based only on reading 5–10 relevant documents in the training set by someone who didn't know the topic statement.

A *query set* is a collection of 50 queries, one for each topic, all of the same type and developed by the same person. Participants exchanged the query sets they created with all other participants in the track, and all participants ran all query sets. The test set of documents for the track was the documents on TREC disk 1. This disks contains approximately 500,000 documents taken from the *Wall Street Journal*, the *Associated Press* newswire, *Computer Selects* published by Ziff-Davis, the *Federal Register*, and abstracts of U.S. DOE publications.

The track produced 21 different query sets of these types, each of which was run by nine different retrieval methods. A brief description of the retrieval methods is given in Table 1, which also includes the label given to the method and the organization that made the run. Details about the individual methods used in the track can be found in the participants' papers in the TREC-8 proceedings [1, 2, 9, 12, 15]. The results of the track thus provide nine sets of the top 1000 documents retrieved for each of 1050 queries (21 versions of 50 topics), and the list of relevant documents for each of those 50 topics.

4 Computing the Error Rate

Our goal is to use the data from the Query Track to quantify the error rate associated with deciding that one retrieval method is better than another given that the decision is based an experiment with a particular number of topics, a specific evaluation measure, and a particular value used to decide if two scores are different. Our approach is as follows. First, we choose an evaluation measure and a “fuzziness” value. The fuzziness value is the percentage difference between scores such that if the difference is smaller than the fuzziness value the two scores are deemed equivalent. For example, if the fuzziness value is 5%, any scores within 5% of one another are counted as equal. We pick a query set and compute the mean of the evaluation measure over that query set for each of the nine retrieval methods. For each pair of retrieval methods, we compare whether the first method is better than, worse than, or equal to the second method with respect to the fuzziness value. We select another query set and repeat the comparison multiple times. This results in a 9x9 triangular matrix giving the number of times each retrieval method was better than, worse than, and equal to each other retrieval method for each of the query sets used.

Two examples of the matrix are given in Figure 1. Both examples use a fuzziness value of 5%, and the 21 query sets submitted to the Query Track. Each entry in a matrix thus sums to 21. The first matrix in the figure was computed using Average Precision as the evaluation measure; the second matrix was computed using Prec(10). The first number in an entry gives the number of times the retrieval method of the row was better than the method of the column, the second number the number of times the method of the row was worse than the method of the column, and the third number the number of times the methods were equal. For example, the Prec(10) matrix says that the APL method was better than the INQa method 2 times, was worse than the INQa method 12 times, and the two methods were within 5% of one another 7 times. The Average Precision matrix says that APL was better than INQa 18 times, and the two methods were equal 3 times.

Because there are 36 different pairs of retrieval methods and we used 21 different query sets for the examples in Figure 1, the matrix represents 756 decisions regarding the relative effectiveness of retrieval methods. If for each pair of methods we assume that the correct answer is given by the greater of the better-than and worse-than values, then the lesser of those two values is the number of times a test result is misleading or in error. We define the error rate to be the total number of errors across all method pairs divided by the total number of decisions.

$$ErrorRate = \frac{\sum Min(|A > B|, |B > A|)}{\sum(|A > B| + |A < B| + |A == B|)} \quad (1)$$

where $|A > B|$ is the number of times method A is better than method B in an entry. Thus the error rate for the Average Precision matrix in Figure 1 is $16/756 = .021$ or 2.1%. Similarly, the error rate for the Prec(10) matrix is $29/756 = .038$ or 3.8%. Note that the error rate can never be more than 50%, and random effects start dominating the calculation of the error rate if it exceeds approximately 25%.

The number of times methods are deemed to be equivalent is also of interest because it reflects on the power of a measure to discriminate among systems. It is possible for a measure to have a low error rate simply because it rarely concludes that two methods are different. The proportion of ties, defined as the total number of equal-to counts across all method pairs divided by the total number of decisions, quantifies this effect. The proportion of ties for Average Precision in Figure 1 is $98/756 = .130$ and for Prec(10) is $209/756 = .276$.

We have defined a query set to be one query for each of the 50 topics such that each query is of the same type and was developed by the same person. For these experiments, however, we want to use a more general notion of a query set: one query for each of the 50 topics. Because each query within a query

	INQa	INQe	INQp	Saba	Sabe	Sabm	acs	pir
APL	18 0 3	2 11 8	19 0 2	11 0 10	0 19 2	3 11 7	21 0 0	0 19 2
INQa		0 21 0	4 6 11	0 14 7	0 21 0	0 21 0	21 0 0	0 21 0
INQe			21 0 0	19 0 2	1 16 4	4 4 13	21 0 0	0 17 4
INQp				0 15 6	0 21 0	0 21 0	21 0 0	0 21 0
Saba					0 21 0	0 21 0	21 0 0	0 21 0
Sabe						21 0 0	21 0 0	2 4 15
Sabm							21 0 0	0 19 2
acs								0 21 0

a) Average Precision

	INQa	INQe	INQp	Saba	Sabe	Sabm	acs	pir
APL	2 12 7	0 19 2	3 9 9	2 11 8	0 20 1	1 14 6	13 1 7	0 19 2
INQa		0 14 7	4 2 15	2 6 13	0 21 0	0 9 12	18 0 3	0 15 6
INQe			20 0 1	16 1 4	4 6 11	14 2 5	21 0 0	6 4 11
INQp				2 5 14	0 20 1	1 12 8	18 0 3	0 19 2
Saba					0 19 2	0 6 15	17 0 4	0 16 5
Sabe						18 0 3	21 0 0	8 1 12
Sabm							19 0 2	1 12 8
acs								0 21 0

b) Prec(10)

Figure 1: Counts of the number of times the retrieval method of the row was better than, worse than, or equal to the method of the column. Counts were computed using a fuzziness factor of 5% and the original 21 query sets.

set submitted to the track is of the same type, using those query sets could bias the error rates if there is an interaction between retrieval method and query type. In addition, some of the query sets were constructed by experts and some by students. To remove these potential biases we randomly permute the queries among 21 new query sets such that each query for a topic is put in a different query set (but different topics use different permutations). The error rate is computed using this new set of query sets. The queries are then re-permuted among a second set of 21 query sets and again the error rate is computed. The permutation process continues until we have 50 different sets of 21 query sets, each of which defines an error rate. By basing the error rate calculation on permutations of the original 21 query sets, we guarantee that each query is represented exactly once in each error rate. This prevents a single query from having too large of an influence on error rate. The design also ensures that any time two methods are compared they are compared on the results of exactly the same queries.

Table 2 lists the mean error rate over the 50 different sets of permuted query sets for a variety of measures using a fuzziness value of 5% and ordered by decreasing error rate. The table also gives the standard deviation of the average error rate and the mean proportion of ties for each measure. The error rates for the different measures are clearly different. Measures that depend on a relatively few highly ranked documents have higher error rates than measures that incorporate more documents. Possible reasons why this dependency exists are given below in Section 5.

The proportion of ties for the various measures also differ substantially. Precision at the various cutoffs fails to distinguish between two systems from 20% to 24% of the time, while Average Precision and Precision at .5 Recall fail to distinguish only 11% to 13% of the time.

4.1 Varying topic set size

One of the IR experimental rules-of-thumb says that experiments must use a reasonable number of topics. In this section we investigate how changing the number of topics used in a test affects the error rate of the evaluation measures.

Measure	Error Rate (%)	Std. Dev. (%)	Ties (%)
Prec(1)	14.3	1.3	23.4
Prec(10)	3.6	0.9	24.3
Prec(30)	2.9	0.8	23.8
Prec at .5 R	2.2	0.5	11.4
Prec(100)	1.8	0.5	20.7
Ave Prec	1.5	0.4	12.8
R-Prec	1.3	0.4	19.1
Prec(1000)	1.0	0.4	22.5
Recall(1000)	0.6	0.2	20.8

Table 2: Average error rate, standard deviation of the average error rate, and average proportion of ties for different evaluation measures. Error rate was computed using a fuzziness factor of 5%. Means were computed from the error rates defined by 50 random permutations of the 21 query sets.

We vary the number of topics used to compute each method's mean score using topic set sizes of 5, 10, 15, 20, 25, 30, 40, and 50. For each topic set size smaller than 50, we randomly choose a set of topics of that size. We restrict the permuted query sets constructed above to just the topics in the selected topic set and compute the average error rate. This is one trial. We continue to pick a new set of topics of the appropriate size and to recompute the error rate until we have a total of 100 trials for each topic set size.

Figure 2 plots the average error rate over 100 trials (where each trial's error rate is the average over the 50 permuted query sets) for each of the topic set sizes smaller than 50. The values plotted for 50 topics are the values shown in Table 2. For all measures the average error rate decreases as the number of topics increases. The ordering of the measures with respect to error rate is stable as topic set size varies, though Prec(10) has a relatively higher error rate at small topic set sizes than at larger

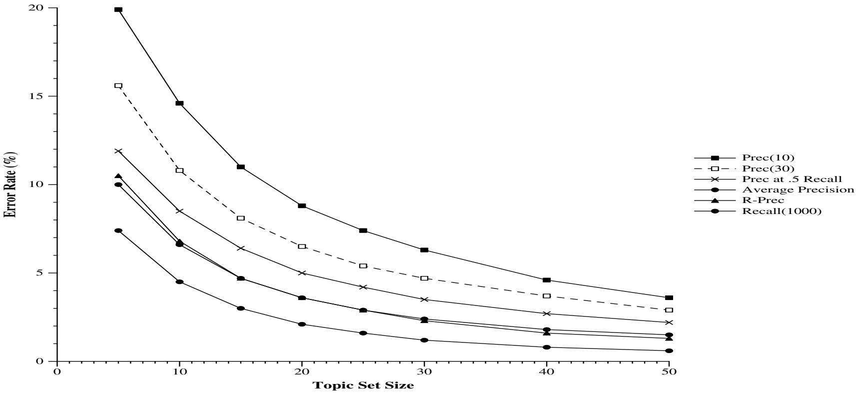


Figure 2: Average error rate of evaluation measures for varying topic set size.

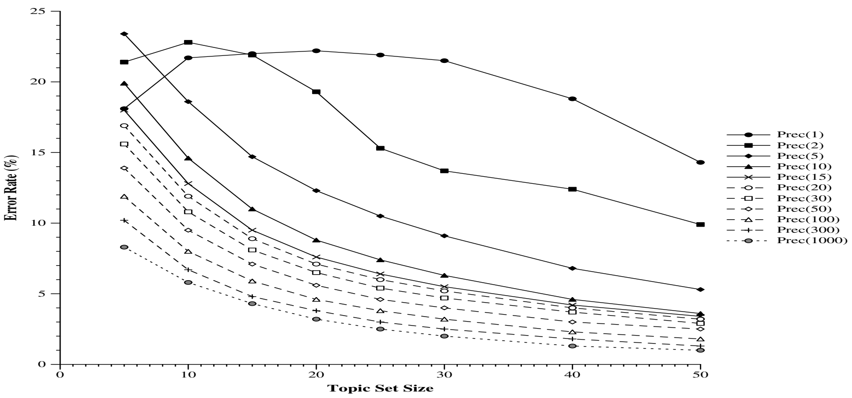


Figure 3: Average error rate of Precision at different cut-off levels for varying topic set size.

topic set sizes.

Figure 3 takes a closer look at the Precision error rate. The average error rates are computed over the same 100 trials as in the previous figure, but in this figure Precision at different cut-off levels is used as the evaluation measure. For cut-off values of 5 or greater, the average error rates show a constant progression of smaller error rates for both larger topic set sizes and larger cut-off levels. The behavior of Prec(1) and Prec(2) is more erratic.

The consistent decrease in error rate for larger topic set sizes demonstrates that one way to increase the confidence in conclusions drawn from measures with relatively large inherent error rates is to use more topics in the experiments. Note, however, that the rate at which the error rate decreases slows as the number of topics increases, so that very large numbers of topics are required to force the error rate below a certain level.

4.2 Varying fuzziness values

The value used for deciding when two runs are sufficiently different—our fuzziness value—is another factor in IR experimental design. Intuitively, larger fuzziness values decrease the error rate but also decrease the discrimination power of the measure. Figure 4 quantifies the effect of the fuzziness value on the error rate. The top graph in the figure plots the average error rate over the 50 sets of permuted query sets for all 50 topics when using fuzziness values between 1% and 10%. The bottom graph is the same except it plots the average error rate over 100 trials for topic sets of size 25.

Once again there is a consistent decrease in error rate as the fuzziness value increases. Thus a second way of increasing the reliability of an experimental conclusion is to increase the amount of difference required between scores to conclude that the methods differ. The cost associated with increasing the difference is that fewer conclusions can be drawn since more methods are considered equal.

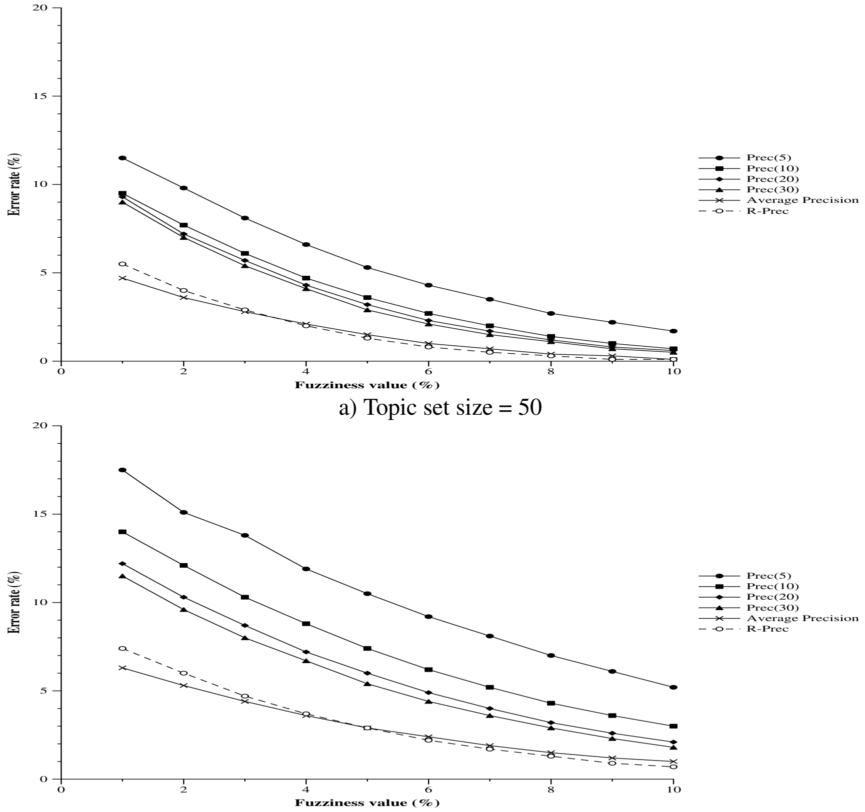
5 Discussion

In this section we discuss the implications of different error rates on the design of retrieval experiments and make recommendations for common scenarios. We begin by noting the limits of this study.

5.1 Limits of this study

Our method for computing the error rate of evaluation measures depends on having retrieval results from a variety of retrieval methods for multiple queries for each of many topics. The study is thus constrained by the available data. In particular, we have computed the error rates using only one collection and a limited variety of retrieval methods.

There is a strong interaction between retrieval methods and the topics and documents that make up a test collection. Our experimental methodology controlled for this interaction as much as possible—both by using large numbers of random samples and by using paired comparisons between methods with each method evaluated over exactly the same topics and queries (and relevance judgments). Nevertheless, it is impossible to know



a) Topic set size = 50

b) Topic set size = 25

Figure 4: The effect of fuzziness value on average error rate.

for sure how results might change on another collection with different characteristics. A particular concern is the number of relevant documents per query, which averages about 209. The *value* of the various evaluation measures is certainly affected by the large number of relevant documents, but it is unknown how the *error rate* is affected.

It is also unclear how the retrieval methods used affect the computed error rate. For instance, the error rate is likely to be quite low if we evaluate nearly identical methods. The particular methods in this study vary substantially in their performance on individual queries, but all of them are automatic statistical systems that concentrate on matching individual terms or phrases in the documents. When data from more retrieval methods becomes available we can compare the error rates computed from similar and dissimilar approaches, but with only the nine methods that is infeasible.

Because we were restricted to one document collection and nine retrieval methods, the error rates computed in this paper must be regarded as lower bounds for the true error rates. We've presented proof that the error rates can be at least this large; they may in fact be larger.

5.2 Individual measures

As mentioned above, there is a strong (inverse) correlation between the size of the error rate and the number of documents used to determine the value of the measure. There are at least two reasons for such a correlation. For measures that depend on very few documents, there simply aren't enough documents to get a stable evaluation score. Precision at one or two documents retrieved is always going to be unstable, though using very large numbers of topics will mitigate the effect somewhat.

For measures that depend on very large numbers of documents, the increased stability is likely to be the result of query expansion. Five of the nine retrieval methods used in the Query Track expand the query substantially either implicitly or explicitly. The effect of expansion on the top retrieved documents depends on how good the expansion is. But any decent expansion strategy will bring more relevant documents into the retrieved set. By rank 1000, the expansion methods are very consistently better than the non-expansion methods, and the computed error rate is therefore low. The very low error rate for Recall(1000) is an example of this effect. The errors in the Recall(1000) matrix occur only between pairs of expansion methods or pairs of non-expansion methods; there are no errors between an expansion method and a non-expansion method.

Precision at various cut-off levels had the highest error rates of any of the measures we tested. Precision is inherently less stable than other measures for two related reasons. First, Precision does not average well. The meaning and value of Prec(10) is very different for a topic with 8 relevant documents as compared to a topic with 300 relevant documents. If the topic with few relevant documents has a high Prec(10) score, then nearly every relevant document was retrieved by rank ten and the retrieval result represents a point close to $Prec = 1, Recall = 1$. An equivalent value for Prec(10) for the topic with many relevant documents represents a point close to $Prec = 1, Recall = 0$.

Second, as the cut-off level used to define the retrieved set increases, the properties of Precision at that cut-off change. As an illustration of this effect, consider the comparison of the APL and INQa methods. In the Prec(10) matrix of Figure 1, INQa (a non-expansion method) was better than APL (an expansion method) by a 10 to 2 margin. The two methods have an equal chance of being better than the other for Prec(50), and APL is the better run by a 15 to 0 margin by Prec(1000). Obviously even Prec(50) is measuring something quite different from Prec(10). Where this change-over occurs is unknown, but it is highly unlikely that it occurs at the same cut-off for all topics (or queries). These effects help explain why the error rate for Precision is substantially higher than the error rate for measures such as R-precision and Average Precision that do not have these problems. Note that even Prec(1000) has a considerably higher relative error rate than Recall(1000) (see Table 2).

R-Precision and Average Precision have been shown to behave similarly in previous examinations of evaluation measures [21, 25]. This is actually quite remarkable given that R-precision evaluates at exactly one point in a retrieval ranking and Average Precision represents the entire area underneath the recall-precision curve. Yet once again in this study the two measures behave almost identically in terms of accuracy (see Figures 2 and 4). Both measures have a noticeably smaller error rate than any of the Prec(λ) measures tested except Prec(1000). However, R-Precision does not have as much discrimination power as Average Precision has; Table 2 shows that R-precision has one and a half times as many ties as Average Precision. This is unlikely to be strictly the result of the number of points precision is being calculated at since Precision at .5 Recall has slightly more discrimination power than Average Precision while also being calculated at only one point.

Comparisons among TREC systems are most often made in terms of Average Precision, R-Precision, or Precision(30). Of these, Average Precision and R-Precision have low error rates, though R-precision has less discrimination power than Average Precision. Precision(30) is clearly a less powerful measure: it has both twice the error rate and twice the number of ties as Average Precision.

5.3 Acceptable error rates

The notion of an "acceptable" error rate will probably never be well defined. However, the experiences of TREC can be used to infer an operational definition of acceptable error rate. One of the functions of the TREC conferences is to be a venue for discussions of what constitutes good IR experimental methodology. Simplifying enormously, the general consensus within TREC has been that Average Precision is a suitable evaluation measure for general-purpose retrieval; that 25 topics is just barely enough for an experiment but that 50 topics is stable; and that 5% differences are worth noting. Given this, the error rate of Average Precision with a fuzziness of 5% and using 25 topics might be considered an upper bound for marginal acceptability. The error rate computed in this paper for that combination is 2.9% with a standard deviation of .8%. Using the same reasoning, the error rate computed using 50 topics should be definitely acceptable; that value is 1.5% with standard deviation of .4%.

If we use 2.9% as a minimally acceptable error rate, then in this study Prec(λ) for $\lambda < 30$ had too high of an error rate even when using 50 topics. Prec(30) had an error rate of 2.9%. Prec(20) was close at 3.2%, but these results suggest that a good experiment should use more queries. All other evaluation measures tested (including several minor measures not presented in this paper) had error rates below 2.9% so should be fine using 50 topics.

5.4 Recommendations

The error rate of an evaluation measure is only one of a measure's properties. Since different measures evaluate different aspects of retrieval behavior, it would be foolish to select an evaluation measure based on error rate alone. For example, Recall(1000) is very stable, but it is an appropriate evaluation measure only for environments such as legal case law or patent searching where finding all relevant documents is of primary importance. The evaluation measure to be used in a retrieval experiment should be selected based on the particular aspect of retrieval behavior that is of interest. The results in this paper provide the means for manipulating other parts of the experimental design to obtain a desired level of confidence in the conclusions drawn from the experiment.

For general purpose retrieval, Average Precision seems to be a reasonably stable and discriminating choice.

In environments such as the Web where it is very difficult to know how many relevant documents exist for a query, precision at a cut-off level of 10 or 20 is an appropriate evaluation measure. However, the results here show that many more queries need to be evaluated in such an experiment in order to show that Method A is better than Method B, as compared to a test collection environment evaluated with Average Precision. Extrapolating the results here (always dangerous, but there is no other guidance available), doubling the number of queries should suffice. This suggests that 100 queries is a good target number for an experiment measuring Precision(20).

6 Conclusion

This paper presents a method for quantifying how the number of requests, the evaluation measure, and the notion of difference used in an information retrieval experiment affect the confidence that can be placed in the conclusions drawn from the experiment. We show that some evaluation measures are inherently more stable than others. For example, Precision after 10 documents are retrieved has more than twice the error rate associated with it than the error rate associated with Average Precision. We confirm that conclusions drawn from experiments using more requests are more reliable than conclusions drawn from experiments using fewer requests. We also show that requiring a larger difference between scores before considering the respective retrieval methods to be truly different increases reliability, but at a cost of not being able to discriminate between as many methods.

For most measures, 50 topics is sufficient to give an error rate less than 2 or 3% in these experiments. Precision at low cut-off values (30 documents or less), however, has substantially higher error rates.

We show that in the particular environment tested here, if the Average Precision score for Method A is more than 5% greater than the score for Method B in a single test, there is less than a 2% chance of being wrong. However, the chance of being wrong may easily be higher if the experiment is run on a different collection with different information requests. Many modern IR evaluation papers present results on several collections, and we suggest this practice should be strongly encouraged.

This is an initial attack at evaluating evaluation measures. An obvious extension in the short term is to look at other measures. A second area to explore is using the methodology described here to examine how stable and accurate significance tests are in practice [10]. Our long term goal is to be able to statistically model the effects of topics and queries on evaluation measures and results. While we do not yet have enough data to to do this reliably, we hope to develop a firm, theoretical foundation for this approach as we gain more information from the TREC Query Track and other sources.

References

[1] James Allan, Jamie Callan, Fang-Fang Feng, and Daniella Malin. INQUERY and TREC-8. In Voorhees and Harman [26].

[2] Chris Buckley and Janet Walz. SMART in TREC 8. In Voorhees and Harman [26].

[3] C. W. Cleverdon, J. Mills, and E. M. Keen. Factors determining the performance of indexing systems. Two volumes, Cranfield, England, 1968.

[4] William S. Cooper. On selecting a measure of retrieval effectiveness. Part I. In Karen Sparck Jones and Peter Willett, editors, *Readings in Information Retrieval*, pages 191–204. Morgan Kaufmann, 1997.

[5] Gordon V. Cormack, Christopher R. Palmer, and Charles L.A. Clarke. Efficient construction of large test collections. In Croft et al. [6], pages 282–289.

[6] W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998. ACM Press, New York.

[7] D. K. Harman, editor. *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, October 1996. NIST Special Publication 500-236.

[8] Donna Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In Harman [7], pages 1–23. NIST Special Publication 500-236.

[9] David Hawking, Peter Bailey, and Nick Craswell. ACSys TREC-8 experiments. In Voorhees and Harman [26].

[10] David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 329–338, 1993.

[11] E. Michael Keen. Presenting results of experimental retrieval comparisons. *Information Processing and Management*, 28(4):491–502, 1992.

[12] K.L. Kwok, L. Grunfeld, and M. Chan. TREC-8 ad-hoc, query and filtering track experiments using PIRCS. In Voorhees and Harman [26].

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

[14] David D. Lewis. The TREC-4 filtering track. In Harman [7], pages 165–180. NIST Special Publication 500-236.

[15] J. Mayfiled, P. McNamee, and C. Piatko. The JHU/APL HAIRCUT system at TREC-8. In Voorhees and Harman [26].

[16] Gerard Salton. The state of retrieval system evaluation. *Information Processing and Management*, 28(4):441–449, 1992.

[17] K. Sparck Jones and C.J. van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32(1):59–75, 1976.

[18] Karen Sparck Jones. Automatic indexing. *Journal of Documentation*, 30:393–432, 1974.

[19] Jean M. Tague. The pragmatics of information retrieval experimentation. In Karen Sparck Jones, editor, *Information Retrieval Experiment*, pages 59–102. Butterworths, 1981.

[20] Jean Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, 28(4):467–490, 1992.

[21] Jean Tague-Sutcliffe and James Blustein. A statistical analysis of the TREC-3 data. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3) [Proceedings of TREC-3.]*, pages 385–398, April 1995. NIST Special Publication 500-225.

[22] C.J. van Rijsbergen. *Information Retrieval*, chapter 7. Butterworths, 2 edition, 1979.

[23] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In Croft et al. [6], pages 315–323.

[24] Ellen M. Voorhees. Special issue: The sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1), January 2000.

[25] Ellen M. Voorhees and Donna Harman. Overview of the seventh Text REtrieval Conference (TREC-7). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 1–23, August 1999. NIST Special Publication 500-242. Electronic version available at http://trec.nist.gov/pubs.html.

[26] E.M. Voorhees and D.K. Harman, editors. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. Electronic version available at http://trec.nist.gov/pubs.html, 2000.

[27] D. Williamson, R. Williamson, and M. Lesk. The Cornell implementation of the Smart system. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 2, pages 43–44. Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1971.

[28] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In Croft et al. [6], pages 307–314.