



On Collection Size and Retrieval Effectiveness

DAVID HAWKING
CSIRO Mathematical and Information Sciences, Canberra, Australia

david.hawking@csiro.au

STEPHEN ROBERTSON
Microsoft Research, Cambridge, UK

ser@microsoft.com

Received May 25, 2001; Revised July 17, 2002; Accepted July 17, 2002

Abstract. The relationship between collection size and retrieval effectiveness is particularly important in the context of Web search. We investigate it first analytically and then experimentally, using samples and subsets of test collections. Different retrieval systems vary in how the score assigned to an individual document in a sample collection relates to the score it receives in the full collection; we identify four cases.

We apply signal detection (SD) theory to retrieval from samples, taking into account the four cases and using a variety of shapes for relevant and irrelevant distributions. We note that the SD model subsumes several earlier hypotheses about the causes of the decreased precision in samples. We also discuss other models which contribute to an understanding of the phenomenon, particularly relating to the effects of discreteness. Different models provide complementary insights.

Extensive use is made of test data, some from official submissions to the TREC-6 VLC track and some new, to illustrate the effects and test hypotheses. We empirically confirm predictions, based on SD theory, that $P@n$ should decline when moving to a sample collection and that average precision and R-precision should remain constant. SD theory suggests the use of recall-fallout plots as operating characteristic (OC) curves. We plot OC curves of this type for a real retrieval system and query set and show that curves for sample collections are similar but not identical to the curve for the full collection.

Keywords: text retrieval models, signal detection theory, collection sampling, relevance score distributions

1. Introduction

Hawking et al. (1999) report consistently greater early precision values when text retrieval queries are processed over the TREC¹ Very Large Collection rather than over a representative sample of it. They report that all participants in the TREC-6 (Voorhees and Harman 1997) Very Large Collection track observed at least 28% greater precision at 20 documents retrieved ($P@20$) for the 20 gigabyte full collection compared with a representative 10% sample.

The reported phenomenon is of interest because it appears to contradict the expectations of early workers in the field of information retrieval. For example, Salton and McGill (1983, p. 173) stated that precision can be expected to *decrease* when collection size increases because, “the number of retrieved and relevant documents [i.e., the number of retrieved documents which are actually relevant] is not likely to increase in proportion to the size of the collection.”

An understanding of the effect of collection size on retrieval effectiveness is a key to understanding Web search performance, because the proportion of the Web indexed by different search engines varies greatly (Lawrence and Giles 1999). Our analysis bears upon the trade-off between the economic pressure to reduce the number of documents fetched and indexed and the market pressure to maintain high search effectiveness.

1.1. *Background*

The Text Retrieval Conference (TREC) series (NIST 1997) provides, among other things, both methodology and resources for evaluating the effectiveness of text retrieval systems on an ad hoc retrieval task. In the ad hoc framework, participants are supplied with a static test corpus of electronic documents and with sets of natural language statements of user information need (*topics*). They must convert the topics into *queries*² and process them over the test corpus. Retrieved documents are later judged for relevance by independent human assessors and these judgments form the basis for system comparisons.

The main ad hoc task in TREC traditionally involves a 2 gigabyte corpus but TREC-6 and TREC-7 included a special-interest Very Large Collection (VLC) track in which the corpus is at least an order of magnitude larger. Subsequent references to the VLC track refer to the data, topics, and submissions of TREC-6 VLC.

The present paper analyses results obtained by participants in the VLC track and presents further results obtained using various TREC resources. The reader is referred to Voorhees and Harman (1997) and to Hawking et al. (1999) for fuller explanations of TREC and the VLC track. Corpora and topics employed in the present experiments are documented in Section 3.1.

1.2. *Sampling versus subdivision of collections*

The increase in P@20 reported by VLC participants related to a pair of collections, one of which was a uniform sample of the other. Regardless of topic, this should mean that the expected probability of relevance of a document in the sample is the same as that for a document in the full collection.

By contrast, except for certain subsets of topics, the probability of document relevance is very unlikely to be preserved if the full collection is divided on any other basis, such as source, date, or subject matter. In these cases, the issue of increasing early precision due to collection size is confounded by the changed probability of relevance caused by the selection process. It is possible, however, to correct for the selection bias by averaging. Thus, for example, if a collection is divided into three sub-collections in a biased way, the P@20 values for each of the three sub-collections may differ substantially. It is possible, though not obvious, that P@20 averaged over a query set and all three sub-collections is a fair estimate of P@20 in a “typical” sample collection one-third of the size of the original.

1.3. *How do collections grow?*

Neither the “inverse” of uniform sampling nor the “inverse” of biased sub-division perfectly model the process by which real collections usually grow.³ Even if the mix of new documents

added closely matches the profile of the original collection, the new documents are likely to be selected in some way which affects their relevance to certain topics. For example, they are likely to be more recent. A growing collection of newswire articles in some sense maintains a constant profile but may exhibit dramatic changes in probability of relevance to topics affected by specific events, such as the reactor accident at Chernobyl. In general, the probability of relevance of new documents is likely to be less than or greater than the corresponding figure for documents already in the collection, depending upon the topic.

1.4. The present study

The overall aim of the present study is to investigate the observed phenomenon (variation in P@20 between full collections and samples) in depth. One approach to this investigation is analytical; we do not present a single analytical model of the phenomenon, but rather consider a range of different models which may provide complementary insights. The other major approach is experimental; we conduct a number of new experiments with TREC data, as well as interpreting the results of published experiments.

The development of analytical models provides us with a number of tools which are useful in subsequent discussion. We would like, of course, to reach a single all-encompassing explanation, but the complexities and subtleties of the phenomenon defy that desire. Nevertheless, we believe that the insights achieved are significant.

Hawking et al. (1999) document a number of hypotheses informally put forward by TREC-6 participants to explain the observed increase in P@20. Investigation of these hypotheses both empirically and in the light of analytical models is an important part of our study. We restate the hypotheses here for the convenience of the reader.

1.4.1. The VLC participant hypotheses

Hypothesis 1. Baseline P@20 measurements are necessarily lower because large numbers of topics have too few relevant documents in the baseline sample to achieve the P@20 scores observed on the full VLC. For example, no system can possibly achieve baseline P@20 higher than 0.1 on a topic for which there are only two relevant documents in the baseline collection. However, on the same topic, the full VLC may contain ten times as many relevant documents and P@20 could reach unity for a perfect retrieval system.

Hypothesis 2. The first edition VLC is equivalent to a replicated collection in which each baseline document is repeated ten times. From this, it is expected that the P@20 on the 20 gigabyte collection should equal the P@2 value for the baseline, when averaged over a sufficiently large number of topics.

Hypothesis 3. Precision-recall curves, as plotted for TREC, represent a kind of operating characteristic for the combination of retrieval system, query and spread of documents represented by the VLC. Because the baseline is a representative sample of the VLC, the operating characteristic is the same in both cases. If precision monotonically declines with increasing recall, as is usually the case, the probability that a

particular document in the ranked list returned by a system is relevant also declines with increasing recall. The effect of increasing the collection size (adding both relevant and irrelevant documents) is to change the number of relevant documents represented by a particular recall value. (Recall of 0.1 may represent 10 documents in the baseline and 100 in the VLC.) If the precision-recall curve remains the same for both collections then precision at a fixed number of documents retrieved must increase.

Hypothesis 4. Swets (1963) postulated separate distributions of document scores for relevant and irrelevant documents (for a given combination of query, collection and retrieval system). He assumed that the two distributions were normal and that their means differed by an amount M which could be used to characterise the performance of the combination of query, collection and retrieval system. If the distributions are the same for the VLC and for the sample, then taking either a fixed score cutoff or a fixed recall cutoff will result in the same precision—but very many more documents in the VLC than in the sample. In order to retrieve a fixed *number* of documents, the cutoff score must be set much higher in the VLC than in the sample, resulting in a greater proportion of relevant documents among those retrieved.

Hypothesis 5. The performance of retrieval systems relying on *tf.idf* models may be harmed by anomalous *df* values. (Here, *tf* is the number of times a query term occurs in a document and *idf* is the reciprocal of the number of documents in the collection which contain the term. A simple method for scoring the relevance of a document involves summing *tf.idf* values for each of the query terms, assuming that terms which occur in many documents are less useful indicators of relevance, and that the more occurrences of a query term there are in a document the greater the likelihood of its relevance. All of the systems used in the VLC track, excepting that of the University of Waterloo share these assumptions and make use of *idf* or a similar collection-derived frequency.) It is possible that *df* values obtained from a very large collection are more generally useful. From this, we might expect better retrieval effectiveness over the sample collection if *dfs* from the full collection (scaled down, if necessary) were used. This hypothesis does not explain the difference observed by the University of Waterloo, whose retrieval system did not use *df* information.

1.4.2. Specific aims. Our specific aims may be summarised as follows:

1. To present an analysis of retrieval from samples;
2. To confirm that precision at fixed number of documents retrieved increases with collection size;
3. To investigate whether the effect in question is specific to uniform sampling;
4. To investigate whether other precision measures are subject to similar effects;
5. To explore analytically, and test empirically, the five hypotheses stated above.

1.4.3. Relationship to prior work. Swets (1963, 1969) first cast document retrieval as a signal detection (SD) problem, treating relevant documents as signal and irrelevant as noise, and postulating that the behaviour of a retrieval system may be described by an

Operating Characteristic (OC) curve. Here we extend this work by examining how retrieval characteristics change when the retrieval system is applied to samples or subsets of the original large collection.

Swets considered several possible distribution shapes for the relevant and irrelevant scores. Others have since been proposed (Arampatzis et al. 2000, Baumgarten 1999, Manmatha 2001) in distributed IR and threshold setting applications. We present precision v. sample size results for each of these distribution pairs. Cormack et al. (1999) ignored the distinction between relevant and irrelevant and considered only a single distribution of scores. They used this model to predict precision at a fixed cutoff in sample collections but did not attempt to model the common case where an individual document's score depends upon the collection.

Provost and Fawcett (2001, Section 3.3, last par.) mention the application of a hybrid Receiver Operating Characteristic Convex Hull (ROCCH) classifier to document retrieval and claim that on a retrieval task, their ROCCH-hybrid method will deliver the best N documents regardless of corpus size.

1.4.4. Outline of the remainder of the paper. Section 2 presents a detailed analysis of retrieval from samples, first examining how the retrieval scores of individual documents may be affected by sampling, then considering distributional properties, SD models and their problems, choice of suitable OC curves, and how to predict average precision as well as precision at fixed cutoff. Finally, two alternative discrete models are presented.

Section 3 details the experimental framework within which the empirical investigations of Sections 4 and 5 are carried out. Section 4 confirms the generality of the precision-decrease phenomenon when three different test collections are sampled or sub-divided by source. The effect of these operations on two other precision measures is investigated as is the effect of increasing a collection size by replication. Section 5 analytically examines and empirically tests each of the five hypotheses and Section 6 presents overall discussion and conclusions.

2. Analytical models of sampling and retrieving

We are concerned with the relationship between running a given query or set of queries on a large collection of documents, and running the same query or queries on a sample of that collection. We identify the following elements:

\mathcal{C} , the collection;

\mathcal{S} , the sample;

\mathcal{Q} , the query;

\mathcal{E} , the retrieval engine.

We wish, therefore, to analyse the relationship between $(\mathcal{Q}, \mathcal{E}, \mathcal{C})$ and $(\mathcal{Q}, \mathcal{E}, \mathcal{S})$. A central question is: What aspects or features of this triplet might change or remain the same as we move from the collection to the sample? We need to consider in some detail what these aspects or features might be, and to what extent we can hypothesize concerning their stability or otherwise.

2.1. *What happens to a document?*

One way to consider some relevant aspects is to follow what happens to an individual document as we move from the collection to the sample. Consider therefore a document in the sample (and therefore in the full collection). The combination of query and retrieval engine will assign a score to this document in both the full collection and the sample case. The first question is: will this be the same score?

It is easy to pick out examples of systems in which the score will be identical. In a Boolean system, for example, in which the transformation of the query into the Boolean formulation is made independently of the collection, the “score” is either zero or one, and the document’s score is the same whether we are retrieving from the collection or the sample. The same may be said of weighting schemes in which the weights are determined independently of the collection being searched—proximity-based weighting schemes such as those used by Clarke et al. (1995) or Hawking and Thistlewaite (1996) for example. These cases may be described as “deterministically invariant.”

However, in many systems this will not be the case, and understanding of these cases requires analysis of the various elements that go into the search formulation and scoring process. It may be that the search formulation is essentially independent of the collection (some form of query tokenization for example), but that the weights assigned to terms are collection-dependent. The most obvious example is *idf* in all its variants.

In this case, it is clear that the *idf* weight of a given query term will vary between the collection and the sample. However, since we are concerned with *random* samples, the differences should be of a more-or-less understandable kind associated with the sampling process. Normalised *df* in the sample should be a reasonable and unbiased estimator of normalised *df* in the collection. The complications are that (a) the variance will depend on the sample size, and (b) all *idf* functions in use are non-linear in *df*.

We may surmise that there are other possible scoring functions that use other such parameters. We clearly do not have deterministic invariance in these cases; the best we may hope for is some form of “statistical invariance.”

Finally, there are instances of systems in which basic characteristics of the search formulation vary between the collection and the sample. Two-pass, blind feedback systems fall into this category: if an initial query (which may be deterministically or statistically invariant) is used to select some top-ranked documents, and these are used to expand the query for a second pass, the final search formulation (terms as well as weights) is clearly very dependent on the collection initially searched. If for example this process generates better queries in the full collection than in the sample, this difference confounds any other differences.

Thus we have the following:

Case 1. Deterministic invariance: the score of an individual document remains the same regardless of which collection contains it.

Case 2. Statistical invariance: although a document’s score may depend on the collection from which it is being retrieved, the scoring system will behave in a statistically similar fashion in the full collection and the sample.

Case 3. No invariance: the collection (full or sample) fundamentally affects the behaviour of the scoring system.

Below, we will seek to make the definition of Case 2 more specific, and in doing so we will subdivide it into two cases.

The primary purpose of the rest of this paper is to understand the effect of sampling in the first two cases. Any observed differences in the third case may be partly due to the same causes, but may also contain components which cannot be explained by those causes.

2.2. *Distributions of scores*

Given these considerations, it may be useful to explore the Case 2 scenario by examining *distributions* of scores.

A question we may ask about such distributions is: do the scores from the sample collection look as though they come from the population distribution defined by the scores in the full collection? In other words, can we regard the scores in the sample collection as a random sample of the scores in the population? In Case 1, it is clear that they should be thus: the process of sampling the documents is exactly the same as the process of sampling values from the population distribution of scores.

In Case 2, this may or may not work. The effect of sampling on the scores may simply be to introduce a small amount of noise, so that the score of a document in the sample can be seen as an estimate of its score in the population. If this noise is purely random, unbiased and independent of any important variable, and in particular of relevance, then the distributions should match in this way, at least approximately.⁴

This last qualification reveals the important characteristic here. The real question is not whether sampling leaves the overall distribution the same, but whether it affects the relationship between relevant and non-relevant documents. One can imagine a difference between the full collection and the sample which is systematic but does not affect this relationship (e.g., a change in which all scores are shifted up or down, by a constant or in proportion). On the other hand it might be possible to imagine a more subtle change which leaves the overall distribution apparently unchanged but does alter the relationship.

For this reason, the model to be introduced in the next section separates the documents into relevant and non-relevant first, and then looks at the distributions of scores in these two populations separately. This allows a more formal approach to the problem, and to the definition of the two sub-cases of Case 2.

Cormack et al. (1999) make a related analysis. In their model, they consider just one distribution of scores, for all documents, and then define a form of ranking in a collection-independent way (assuming Case 1). A probability of relevance is then associated with this ranking. The approach is discussed further below.

2.3. *Signal detection theory model*

The signal detection theory (SD) model for information retrieval was first proposed by Swets (1963, 1969). As suggested, it defines two distributions of scores, one for the relevant and one for the non-relevant documents.

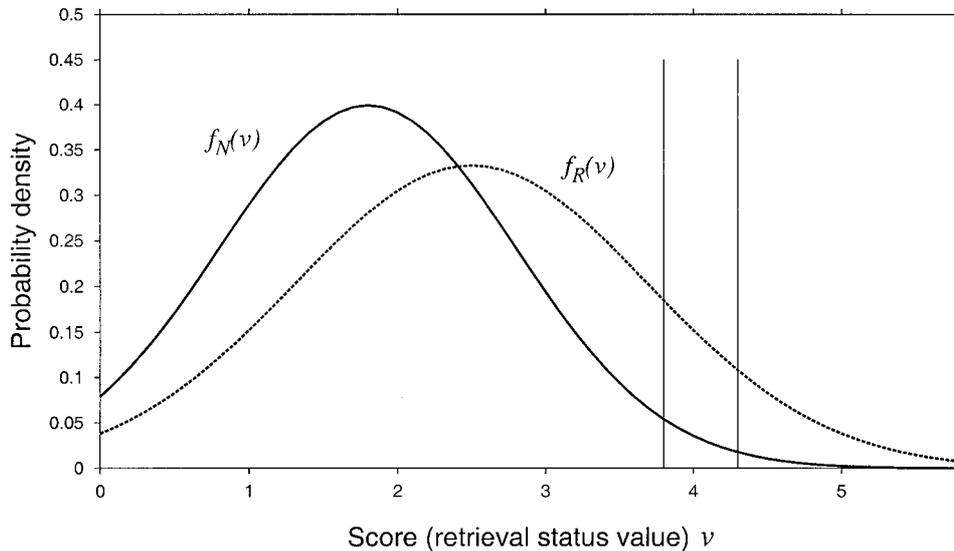


Figure 1. SD model, normal distributions unequal variance: Relevant mean 2.5 variance 1.44 (standard deviation 1.2); non-relevant mean 1.8 variance 1.

In Swets' original papers, he suggested various parametric assumptions about these distributions, including normal-equal-variance, normal-unequal-variance and negative-exponential. Typical observed distributions, particularly the non-relevant distribution, are very far from normality; these distributions have been investigated recently by several authors (Arampatzis et al. 2000, Baumgarten 1999, Manmatha 2001). However, the normal-unequal-variance model is used here to illustrate some of the ideas. We return to the shape of the distribution below. The normality assumptions also take all variables as continuous, and may therefore predict fractions of documents, which is clearly unrealistic (see below).

In figure 1, we see an example of a pair of distributions as assumed in this model. These are shown in the form of density functions (the usual bell curve). The x -axis is the score or retrieval status value, denoted v ; the two distributions are denoted $f_R(v)$ and $f_N(v)$ for relevant and non-relevant documents respectively.

In order to see how these distributions might be reflected in retrieval, we first turn them into cumulative distributions *from the right*—see figure 2. These functions are defined as follows:

$$F_R(t) = \int_{v=t}^{\infty} f_R(v) dv$$

and similarly for F_N .

At any given cut-off or threshold t (examples shown in the form of vertical lines), the cumulative distributions give the probability of retrieving a relevant or non-relevant document

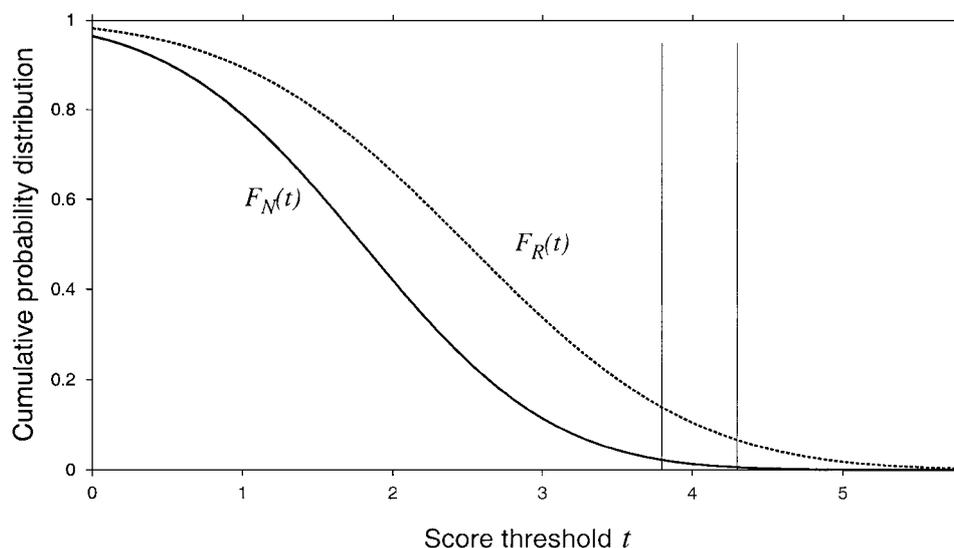


Figure 2. SD model, normal distributions unequal variance, as in figure 1, cumulative form.

respectively at or above that threshold score. These two probabilities may be equated with the traditional measures of recall and fallout respectively (fallout is the proportion of non-relevant documents retrieved).⁵ That is, the probabilities can be used as *definitions* of recall and fallout, and observed recall and fallout values are then estimates of these measures:

$$\begin{aligned} \text{Recall at threshold } t &= Pr(\text{Document } d \text{ retrieved at threshold } t \mid d \text{ relevant}) \\ &= Pr(v(d) \geq t \mid d \text{ relevant}) \\ &= F_R(t) \end{aligned}$$

and similarly for fallout and F_N .

Next we assume specific total numbers of relevant and non-relevant documents, and estimate the number of each retrieved at or above each threshold score. From this we can also estimate the total number of documents retrieved and the precision. These quantities are shown in figure 3, based on 10,000 documents in the collection of which 100 are relevant.

Given these data, we can eliminate the score itself, and simply plot the precision against the estimated number of documents retrieved—see figure 4.

As we shall see, figure 4 is not strongly dependent on the distributional assumptions of the previous figures. Any sensible SD model will produce very similar behaviour.

2.4. OC curves

The step taken between figures 3 and 4, that of eliminating the actual document scores, reveals an important property of the situation. A scoring method in information retrieval is

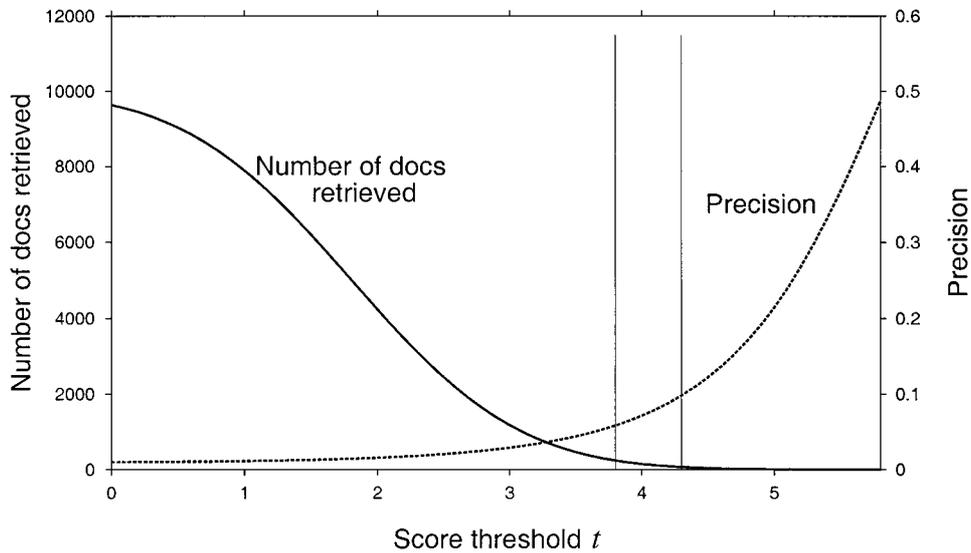


Figure 3. SD model, normal distributions, unequal variance, as in figure 1: Total number of documents retrieved and precision, plotted against score threshold, assuming 10,000 documents in the collection, of which 100 are relevant.

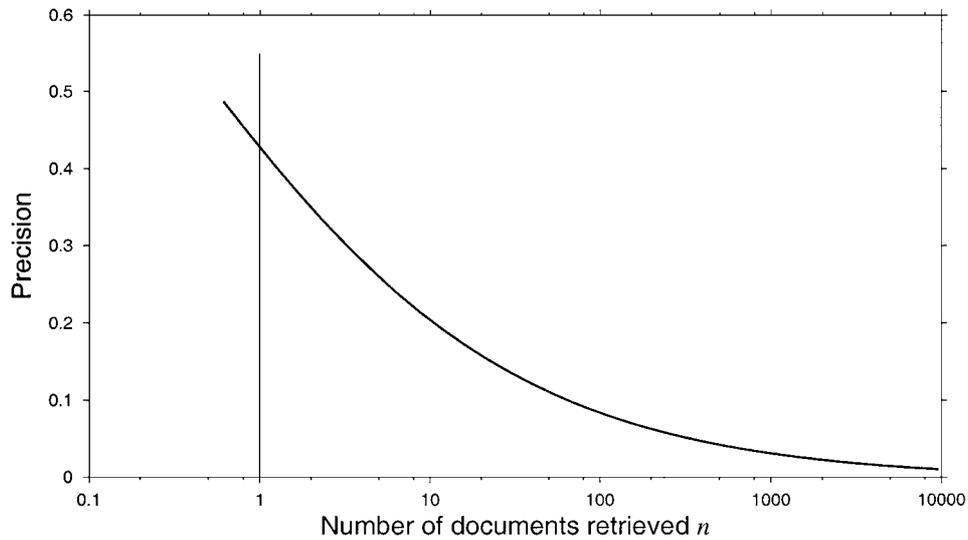


Figure 4. SD model, as for figure 3. Precision is plotted against total number of documents retrieved.

not generally of interest in itself, but only as a mechanism for ranking documents. Figure 4 is a function of the ranking induced by the SD model distributions on the collection of documents, rather than a function of the distributions themselves. In fact, any monotonic transformation of the document scores will induce the same ranking.

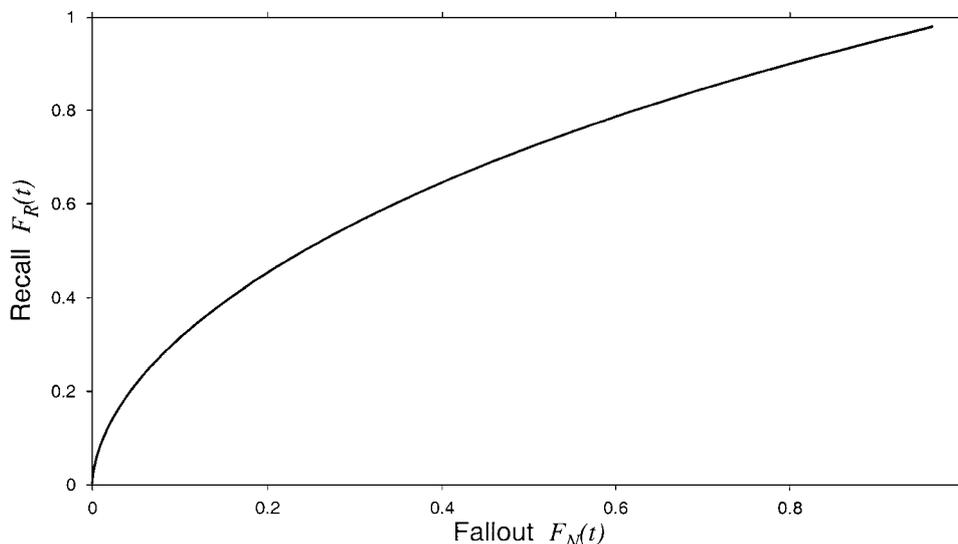


Figure 5. Operating characteristic (OC) curve for the distributions used in figures 1–4.

In signal detection theory, a similar operation to eliminate the document scores or their equivalents is normally performed by taking the operating characteristic (OC) curve of the system. The usual form of the operating characteristic curve involves plotting cumulative distributions (equated to probabilities). We may consider the distributions used above, $F_R(t)$ and $F_N(t)$, cumulated from the right, which we equated with recall and fallout. These correspond to true-positive-rate and false-positive-rate; a standard OC curve would plot one against the other for different values of t . Figure 5 presents such a curve, based on the distributions used above.

While the recall-fallout curve seems not to be easy to understand intuitively in information retrieval, we might see the recall-precision curve as performing a similar function, or even perhaps the curve of precision against number of documents retrieved. Problems with using both of these suggestions are identified in Section 2.9.

2.5. The SD model and determinate invariance

We are now in a position to indicate, using the above example, what the SD model has to tell us about the sampling question, in Case 1 (determinate invariance) as defined above.

Our assumption for this case was that a document score remains the same whether it is being retrieved by the retrieval engine from the whole collection or from the sample. In this case, the process of randomly sampling from the collection of documents is actually identical to the process of randomly sampling from the mixture of SD model distributions. The sampling distributions for relevant and non-relevant documents are exactly the distributions as given, and of course the proportion of the mixture of relevant and non-relevant documents in the sample is also an unbiased estimate of the proportion in the collection.

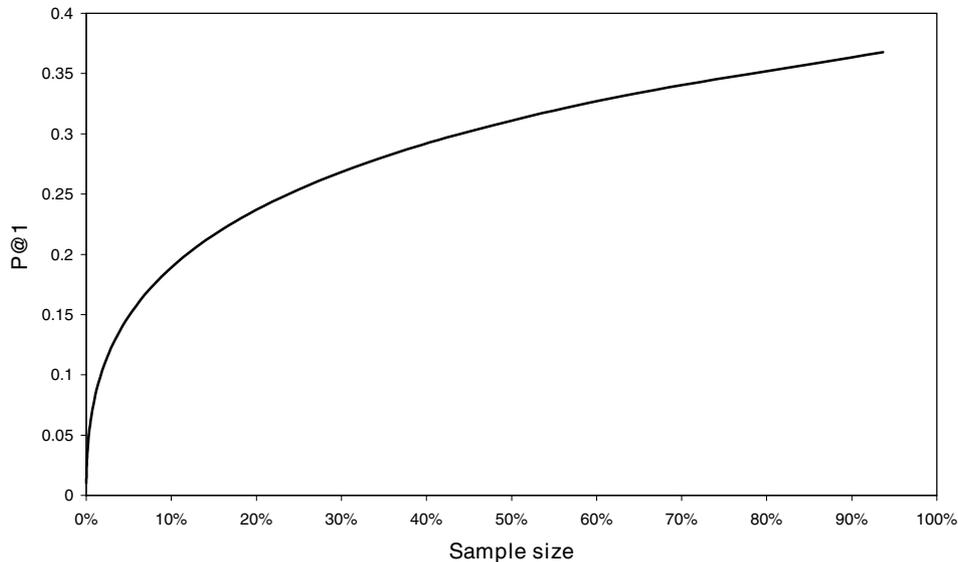


Figure 6. SD model, as for figure 3. P@1 plotted against sample size.

If we consider the graph in figure 4, we see that sampling the collection corresponds to shifting the x -axis to the right (the smaller the sample the further the shift). Thus we can estimate precision at any rank position under any size sample. Figure 6 shows the relation between P@1 and sample size, for the situation of figure 4.

In order to demonstrate the relative independence of the SD model predictions from the specific distributional assumptions used, we repeat the entire analysis under different assumptions. First we use two exponential distributions (this was one of the suggestions made by Swets (1969)). The shapes of the distributions are shown in figure 7. The corresponding predicted graph of P@1 against sample size is in figure 8. Next we assume two gamma distributions (as proposed by Baumgarten (1999)). The distributions are shown in figure 9. P@1 against sample size is in figure 10. Finally, we assume an exponential for non-relevants and a normal for relevants, as used by Arampatzis and Hameren (2001) and Manmatha et al. (2001). The distributions are shown in figure 11. P@1 against sample size is in figure 12.

We see that despite the very different distribution shapes, the basic predicted relationship between P@1 and sample size is very similar indeed. The reasonableness of these assumptions is further discussed in Section 2.8.

This example suggests that a reasonable SD model, coupled with the determinate invariance assumption, would predict a decline of P@ n with sampling. Furthermore, it would predict that P@ n in the whole collection would be the same as P@ $\frac{nS}{N}$ in the sample. What constitutes a reasonable SD model is discussed further below.

Cormack et al. (1999) conclude on the basis of their model (see next section) that P@ $(n-1)$ in the whole collection would be the same as P@ $\frac{(n-1)S}{N}$ in the sample. However, the -1 is an artifact of their method of approximating an integral.

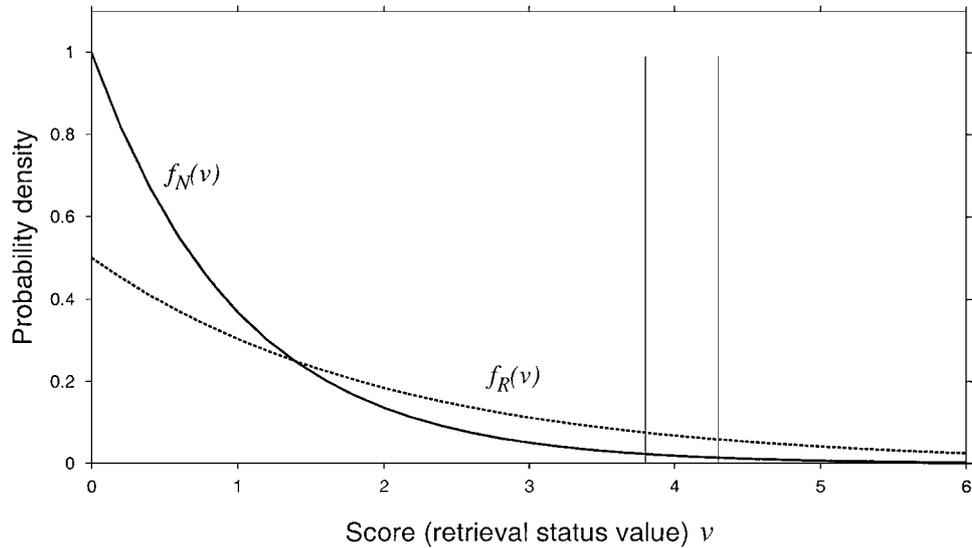


Figure 7. SD model, exponential distributions: relevant mean 2; non-relevant mean 1.

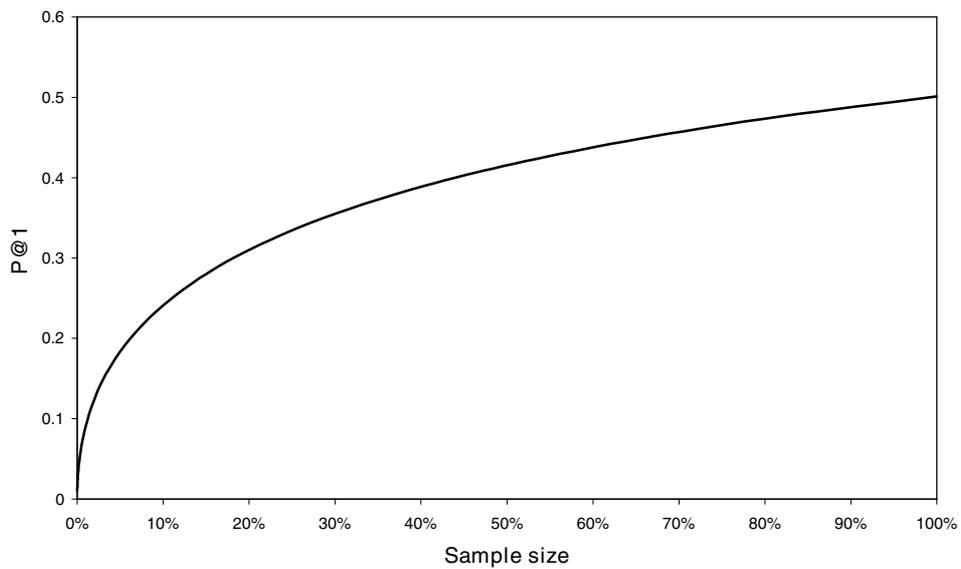


Figure 8. As for figure 6 but based on the exponential distributions shown in figure 7.

2.6. The Cormack model

The Cormack model starts with the distribution of scores of all documents, without initially dividing them into relevant and non-relevant. In terms of our previous notation, we may

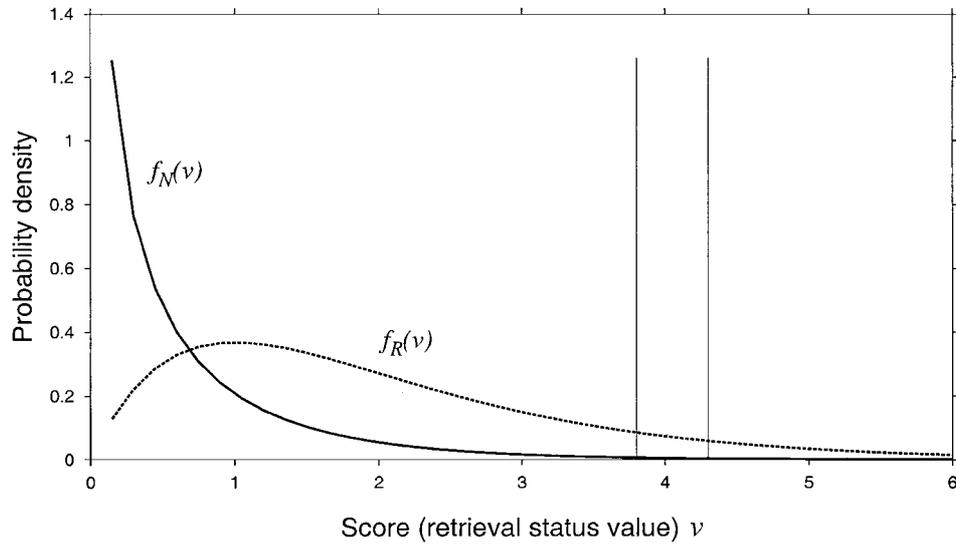


Figure 9. SD model, gamma distributions: Shape parameters 0.5 and 2, scale parameters 1 and 1, for non-relevant and relevant respectively.

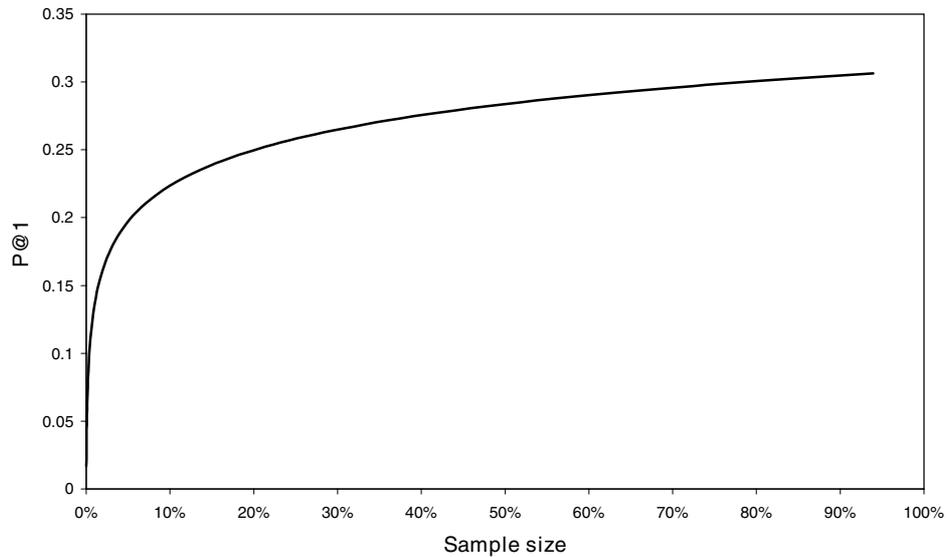


Figure 10. As for figure 6 but based on the gamma distributions shown in figure 9.

define the density as $f_A(v)$ (“All documents”) and the cumulative distribution from the right as $F_A(t)$. The document population is assumed to be infinite (all finite collections are taken to be samples from this infinite population), and Case 1 is assumed between this infinite population and all samples.

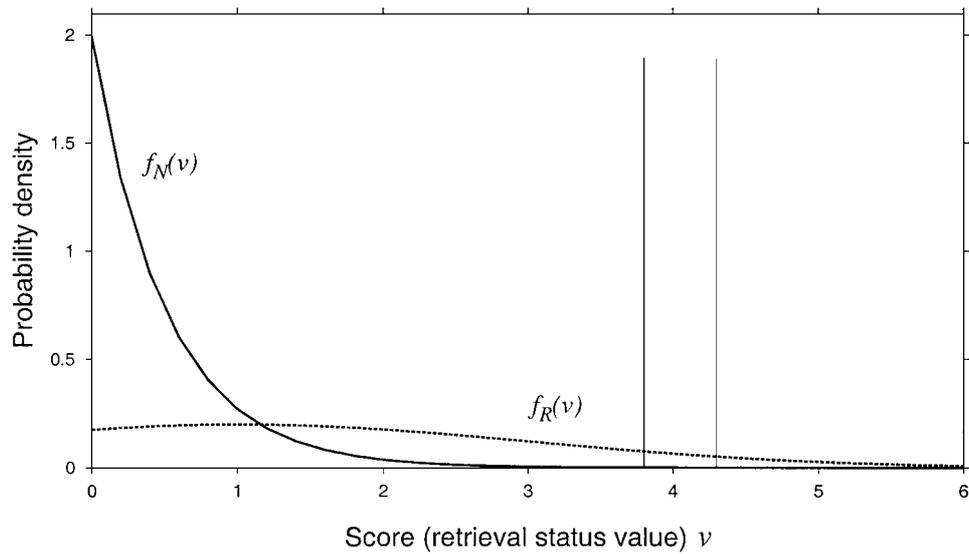


Figure 11. SD model, normal-exponential distributions: Normal relevant mean 1, standard deviation 2; exponential non-relevant mean 0.5.

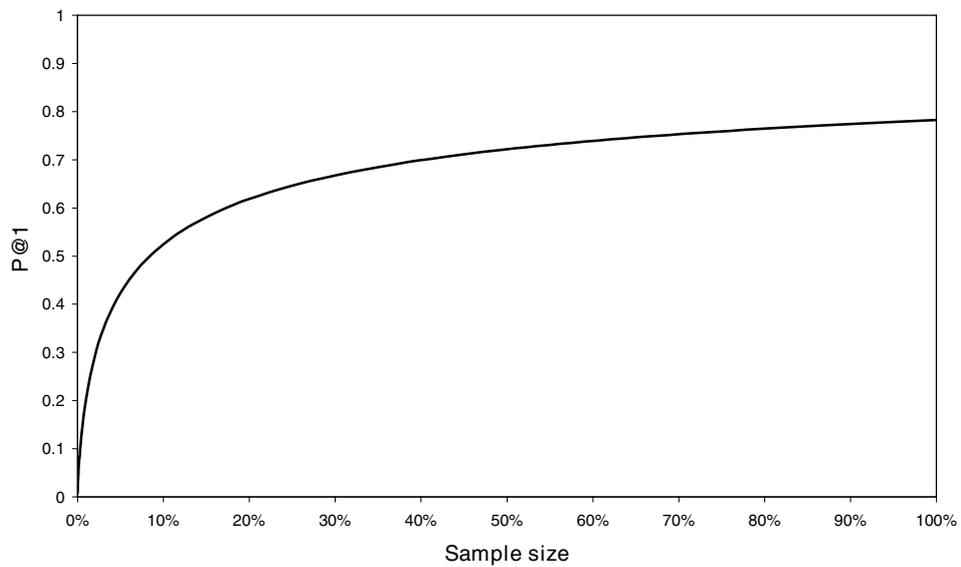


Figure 12. As for figure 6 but based on the normal and exponential distributions shown in figure 11.

The cumulative distribution function is now interpreted as a generalised ranking function, independent of the collection. We can see this as a monotonic transformation of the original scores, so that $v' = F_A(v)$. Each document has a new score or generalised ranking $v'(d) = F_A(v(d))$, ranging from 0 (top ranking) to 1 (bottom), and retains this new score/ranking as

we go from a full collection to a sample. The characteristic of this new scoring is that the distribution of score values over all documents is uniform.

We now associate a probability of relevance with each new score value (they could equally well be associated with the old scores and transferred). This in turn enables us to define a generalised P@ n function:

$$\begin{aligned} P@ \rho &= Pr(d \text{ relevant} \mid v'(d) \leq \rho) \\ &= \int_{v'=0}^{\rho} Pr(d \text{ relevant} \mid v') dv' \end{aligned}$$

for some ρ , a threshold generalised rank.

We now observe that for a sample of size S , if document d has rank $k(d, S)$, then the expected value of the generalised rank for d is

$$E(v'(d)) = \frac{k(d, S)}{S}$$

Therefore we can take the expected value of P@ k in the sample as

$$E(P@k)(S) = \int_{v'=0}^{\frac{k}{S}} Pr(d \text{ relevant} \mid v') dv'$$

This result applies to any sample size, so in particular,

$$E(P@k_1)(S_1) = E(P@k_2)(S_2) \quad \text{if} \quad \frac{k_1}{S_1} = \frac{k_2}{S_2}$$

This is essentially the same result as was obtained with the SD model, but it applies only to Case 1.

2.7. *The SD model and statistical invariance*

The discussion of statistical invariance above left its definition a little vague. We may use the SD way of looking at retrieval to provide a more formal definition.

We suggested that the differences between full collection and sample scores for individual documents might be a small amount of noise, independent of relevance. In this case, we could still see the distribution of scores for the relevant documents in the sample as a random sample of the corresponding full-collection distribution, and similarly for the non-relevant. Then we have:

Case 2a. Statistical invariance of scores: the SD distributions for the sample are random samples of the SD distributions for the full collection.

On the other hand, we may have some systematic variation in the scores, which nevertheless leaves the relationship between relevant and non-relevant the same. We have seen that this relationship is represented in the SD model by the OC curve.

Case 2b. Statistical invariance of OC curves: the OC curve for the sample is the same (give or take some random noise) as the OC curve for the full collection.

Case 2a is clearly a special case of 2b. However, in Case 2a it should be fairly obvious how to test for statistical significance of any difference between the collection and the sample. The “random noise” element of Case 2b is a little more problematic.

We may illustrate these cases with example scoring methods which may be seen to have the required characteristics. These example methods are not in actual use, and are clearly unlikely candidates for use—they are intended for illustration only.

First, consider a simplified *idf* (inverse document frequency) weighting. Assume that *df* is normalised as a proportion or probability: the number of documents containing the term divided by the total number of documents in the collection. But suppose that instead of the usual *idf* formula of a multiplicative factor of $(-\log df)$, we have an additive factor of $-df$: in other words the document score is of the form $\sum_i (f(d, t) - df_i)$ (we assume $f(d, t)$ to depend only on the term and document, not on the other documents in the collection). In this case, df_i in the sample will not be exactly the same as in the collection, but will nevertheless (for a random sample) be an unbiased estimate of the collection value. Therefore the *expected* value of the score for a given document will be the same in the sample as in the collection, although the actual values may vary. This example will satisfy Case 2a provided only that the sample size is not too small (a small sample would imply more variance in the SD distributions in the sample than in the population). Note that the argument fails with the usual *idf* formula because of $(-\log df)$ being non-linear: the expected value is not necessarily the same.

Now consider a system similar to the above, except that document scores are normalised by the score of the highest-ranked document for this query (so that after normalisation the highest-ranked document always scores 1). Typically the highest-ranked document in a sample scores less than the highest-ranked document in the collection, so typically the resulting score for any document in the sample will be higher than it was in the collection (though it may be the same). But (other things being equal) the rank order of documents, and thus the OC curve, will remain unchanged. Thus this example satisfies Case 2b but not 2a.

2.8. Reasonable SD models

We have seen that any monotonic transformation of the document score, which would also transform the distributions of scores, produces the same ranking of documents and therefore leaves the OC curve unchanged. Since one could transform a normal distribution into a very large variety of different shapes by means of monotonic but non-linear transformations, the assumption of normality is actually much weaker than it seems at first. In effect, the assumption would be not that the distributions are actually normal, but that they could be transformed into normal distributions by some single but arbitrary monotonic transformation of the scores.

However, actual observed distributions often look as though they would defy transformation into normal form. (see figures 13 and 14 for some actual examples.) In particular, most

scoring methods assign zero score to any document not sharing any feature with the query. Since this state is normally enjoyed by the vast majority of the documents in the collection, the non-relevant distribution at least usually has what amounts to a singularity at that point.

What then would constitute a reasonable mode? We may seek a definition in the well-known property of information retrieval systems, the inverse relationship of recall and precision. Any model which predicts any systematic violation of this relationship seems perverse, and likely to fall at the first empirical hurdle.

Because $F_R(t)$ is cumulative from the right, it decreases monotonically with increasing t (as indeed recall must do). $F_N(t)$ has the same property, so for given numbers of relevant and non-relevant documents, the total number of documents retrieved must also be monotonic with t in the same direction.

If we then assume the inverse relationship of recall and precision, the model must have precision as a monotonic function of t in the other direction (as for instance in figure 3). So we may define:

A reasonable SD model is one which predicts increasing precision with increasing threshold t .

If we were to discover a violation of this property, we could simultaneously restore the property and improve the performance of the system by a simple re-ordering of the values of the scoring function (Robertson 1969).

But as we have seen from the example, given a graph of precision against total number of documents retrieved, sampling corresponds to shifting the x -axis values to the right. Thus if precision decreases monotonically with increasing total documents retrieved, sampling implies decreasing $P@n$ for smaller samples. This result applies under both deterministic and statistical invariance, given a reasonable SD model.

Strictly speaking, these monotonicities are not strict, and therefore allow equality. A region in which the graph of precision against total number of documents retrieved was flat would not be inconsistent with the above argument. The argument of Robertson (1969) is actually a little stronger, and implies that such a flat region could only occur at the left-hand end of the curve: as soon as it starts dropping, it must continue to do so. Furthermore, it could not be flat over the whole range, except in the trivial case of a random retrieval system. In fact, one of the sets of results examined below does exhibit a flat region of the sort described. On topics with at least 13 relevant documents (see Section 5.2 for both some results and the reason for this selection), the University of Waterloo system exhibits an almost flat graph for $n \leq 20$ in the VLC collection. In the 10% sample collection, it appears to start tailing off around $n = 6$.

It is appropriate to consider whether the various distributional assumptions made in Section 2.5 constitute reasonable SD models according to the above definition. In fact in some conditions they do not, though probably all the violations of the reasonableness assumption occur outside the effective practical range of the model. For example, in the SD model with two normal distributions of unequal variance, there will be a discrepancy at one end of the score scale. In the example given, this occurs in the negative range of scores. Since most scoring systems never produce negative scores, we might regard this as not a practical problem, although clearly it represents an anomaly. The problem becomes more apparent

when considering the mixture of normal for relevants and exponential for irrelevant. Because the right-hand tail of the normal distribution decays much faster ($\approx \exp(-x^2)$) than the tail of the exponential ($\approx \exp(-x)$), this model predicts that a very high-scoring document is less likely to be relevant than a lower one. It may be that in realistic situations such a high-scoring document is unlikely to occur, given the likely parameter values for the distributions; however, there are many combinations of parameter values for which this problem is all too evident. For example, significantly reducing the standard deviation of the relevant distribution in figure 11 brings the problem well into the operational range.

It seems likely that these anomalies are artifacts of the models rather than representing real properties of the scoring functions. If any scoring function does indeed exhibit such a property, then it is clearly a poor function!

2.9. OC curves for retrieval

We have seen that there is some advantage to moving from an explicit examination of distributions, to an examination of the important characteristic of these distributions from the point of view of retrieval, namely the OC curve. The traditional OC curve in signal detection theory corresponds to the recall-fallout curve in retrieval. However, in principle we could consider any suitable pair of variables which reflect the two characteristics of performance for each score threshold. The question arises: which forms of OC curve are likely to be useful for examining the particular situation of this paper, namely the taking of a sample from a larger collection? Ideally, we want a form of OC curve which we could reasonably expect to be the same in the population and in the sample.

The argument for the recall-fallout pair is that the estimation of these parameters is in some sense straightforward. Given a threshold score, and assuming deterministic invariance, recall at that score in the sample is a simple, unbiased estimate of recall at that score in the full collection; the same statement may be made about fallout. Furthermore, under the above definition of Case 2a, the same statement may be made here. Thus we may reasonably expect that the recall-fallout graph should be the same in the sample and in the population in these cases. This in turn suggests that we should take such invariance as the definition of Case 2b.

Precision is more problematic. It may be expressed as a function of recall, fallout and *generality* (the proportion of documents in the collection that are relevant (Salton and McGill 1983):

$$\text{Prec} = \frac{\text{Gen} * \text{Rec}}{\text{Gen} * \text{Rec} + (1 - \text{Gen}) * \text{Fall}} \quad (1)$$

Thus, at least on a superficial level, the assumption that the recall-fallout curve is the same in the sample as in the collection transfers to the recall-precision curve if we assume generality also remains the same—which is a reasonable assumption for a random sample. However, the problem arises because precision is a *non-linear* function of these variables. This means that the statement above about estimating the population value from the sample value does not apply to precision. It is not at all clear that precision at a given score in the sample is a good estimate of precision at that score in the full collection.

The number of documents retrieved depends (for given recall and fallout) on the absolute numbers of relevant and non-relevant documents in the collection. This clearly changes between the full collection and the sample. So it is very clear that the $P@n$ against n graph is not suitable directly for use as an OC curve in this context. $P@ \frac{n}{N}$ against $\frac{n}{N}$ might possibly be, or precision against recall, depending on the stability of precision.

Even in the “suitable” cases (recall v. fallout, recall v. precision, $P@ \frac{n}{N}$ v. $\frac{n}{N}$), other problems arise, associated with averaging over topics. The description above of the OC curve suggested that one should use the threshold score as the “control” (that is, to define successive points on the graph). In retrieval studies it is usual to use one of the variables as control, e.g., to associate a precision value with each recall level. Then averaging across topics means fixing the recall level and averaging precision values. The corresponding score-based method would be to fix the score and average *both recall and precision values* for that score.

Both methods have problems. The score-based method could only make sense if the score values are in some sense comparable across topics; this is not normally the case. The other method essentially uses recall to identify “comparable” points on the scale. The question of whether fallout for fixed recall (in the sample) is a good estimate of something in the population is hard to answer.

2.10. Average precision and R-precision in the SD model

We may ask what the SD model has to tell us about the effects of sampling on other measures of retrieval performance. The obvious one to consider (since it is used extensively in TREC and elsewhere) is average precision. Average precision is defined (in some variant ways) in terms of an actual set of documents and relevance judgements, and converting it to an abstract measure on the distributions is not entirely straightforward. However, we may attempt to define a measure which is a reasonably close analogy.

The simplest definition is to take the precision value at every relevant document in the ranking and average these values. The important point here is that the distribution of *relevant* documents determines the values to be averaged. We may indeed do something analogous in a continuous SD model—but first we must define some parameters. Precision as a function of the threshold score value t may be represented thus (this is essentially the same as Eq. (1) but using some notation defined earlier):

$$P(t) = \frac{GF_R(t)}{GF_R(t) + (1 - G)F_N(t)}$$

(P is precision and G is generality). Averaging may be equated with integrating in a continuous model:

$$P^* = \int_{t=-\infty}^{\infty} \frac{GF_R(t)}{GF_R(t) + (1 - G)F_N(t)} f_R(t) dt \quad (2)$$

$$= \int_{F_R=0}^1 \frac{GF_R}{GF_R + (1 - G)F_N} dF_R \quad (3)$$

Here P^* is the analogue of average precision. The f_R in 2 uses the distribution of relevant documents to control what is averaged. In 3, the score threshold t has disappeared, and F_N (fallout) is now regarded as a function of F_R (recall). Thus P^* is defined purely in terms of the recall-fallout curve. If we can assume that the recall-fallout curve remains invariant between full collection and sample (as argued above for Cases 1, 2a and 2b) then we can make the same assumption for average precision, or at least for its analogue P^* .

A second commonly-used measure is R-precision, which may be defined in two ways. It is the precision (or recall) at the point at which the number of documents retrieved is equal to the number of relevant documents (R) for this topic. Alternatively, it may be defined by the property that it is the precision (or recall) at the point in the ranking at which recall = precision $\neq 0$; it is also referred to as the break-even point (we will denote it B). The two definitions are equivalent; provided that at least one relevant document has been retrieved, precision can only equal recall when the number of documents retrieved is R .

We may combine the above equation for precision in terms of F_R , F_N and G , with the second definition to come up with an analogous measure to B (i.e. R-precision). A small amount of algebra gives the following rule: we choose the score t_{B^*} at which

$$F_R(t_{B^*}) = 1 - \frac{1 - G}{G} F_N(t_{B^*})$$

and the measure is

$$B^* = F_R(t_{B^*})$$

Because both $F_R(t)$ and $F_N(t)$ decline monotonically with t from 1 to zero, it is easy to show that this rule gives a well-defined single value for B^* .⁶ Although it still involves t (because it relates to a single point rather than the whole curve), B^* could be read off the recall-fallout graph without reference to the score. Thus if we can assume that the recall-fallout curve remains invariant between full collection and sample, the same statement may be made of B^* , the analogue of R-precision.

2.11. Problems with the SD model

One aspect of the distributional analysis which we have not discussed is that of discreteness. Documents are discrete objects, and even supposing that the distributions were well describable in terms of continuous functions, some problems would arise from this discreteness. For example, $P@n$ for a given topic can take on a maximum of $(n + 1)$ values, less if the total number of relevant documents is less than n . If we wish to compare $P@20$ in a collection with $P@2$ in a 10% sample, this will clearly be a significant effect. Similarly, the total number R of relevant documents is almost always small compared to N , and often small in absolute terms, and the number in a sample even smaller.

A model which addresses this aspect of discreteness is discussed below.

We may also have another form of discreteness: any scoring function which is a function of the usual parameters (term occurrence, tf , idf etc. for a finite, possibly small set of query terms) is also discrete in the sense that it can take on only certain discrete values (at least

if we consider one topic at a time). This discreteness may be coarse or fine-grained—the old quorum matching function (number of terms matching) is exceedingly coarse, whereas one that includes all the above plus document length is probably fine enough that no two documents score the same. Nevertheless, it is still present.

The idea of modelling the relevant document scores as a continuous distribution is clearly unsatisfactory in these circumstances. The basic arguments should hold with a discrete distribution, but some other effects may be present.

Another problem hinted at in the previous section is that of estimation. We equated recall and fallout with probabilities; however, recall and fallout from a sample must represent *estimates* of probabilities. What they represent in the full collection is questionable. Our model may address some abstract notion of what documents of this kind might look like, or equivalently some notional infinite population, in which case the full collection must be seen as a sample from this notional population. On the other hand, we might regard the full collection as the entire population, so that recall and fallout on the full collection are actual probabilities. Both views have disadvantages.

The estimation problem is also reflected in the fact that in empirical studies, we must normally take averages over topics. Since each topic may have (e.g.) a different number of relevant documents, the exact effect on the observations of either discreteness or estimation from samples is extremely hard to disentangle.

In general, these problems seem analytically somewhat intractable. In the empirical studies below, a pragmatic view is generally taken.

2.12. *SD model: Conclusions*

The problem has been analysed in terms of two features of the situation:

- I. The relation between the score of a document in the sample and the score of the same document in the population; and
- II. The distributions of scores of relevant and non-relevant documents.

The first item leads to a classification of conditions into:

- Case 1.* Deterministic invariance;
- Case 2a.* Statistical invariance of scores;
- Case 2b.* Statistical invariance of OC curves; and
- Case 3.* All other conditions.

We see that in the first three cases, a decline in $P@n$ in samples is a natural consequence of the logic of the situation.

The problems associated with the SD model, of discreteness, sampling, and identifying a suitable form of the OC curve for analysis, limit its formal application. The insights it provides must be qualified by our inability to resolve these problems; nevertheless, these insights seem substantive and valuable. We now attempt to develop discrete models which may provide complementary insights.

2.13. A discrete approach: The hypergeometric model

As before, a random sample \mathcal{S} of S documents is drawn from a collection \mathcal{C} of N documents, and in response to a query \mathcal{Q} a retrieval engine \mathcal{E} ranks the S documents and selects the $n \leq S$ top-ranked documents. We also assume that \mathcal{C} contains R documents relevant to \mathcal{Q} . Within the sample, s documents are relevant, of which $r \leq s, n$ are retrieved. The P@ n measure is thus given by r/n .

Samples used in experiments reported here sometimes constitute a large proportion of the full collection. We therefore start by analysing the problem as sampling *without replacement*. The distribution of values of s is thus hypergeometric (Hays 1963) and the probability of a particular value of s is given by:

$$\Pr(s) = \text{hyperg}(N, R, S, s) = \frac{\binom{R}{s} \cdot \binom{N-R}{S-s}}{\binom{N}{S}} \quad (4)$$

for $0 \leq s \leq \min(R, S)$, where $\binom{n}{r}$ represents the number of combinations of n things taken r at a time, equal to $\frac{n!}{r!(n-r)!}$. This is referred to below as the HG (hypergeometric) model.

Note that the expected value of s , that is the average of a large number of trials, is

$$\bar{s} = \frac{SR}{N} \quad (5)$$

regardless of whether sampling is with or without replacement.

Note also that although we do not assume that S is small compared to N , we can reasonably make this assumption in respect of R , the total number of relevant documents. The hypergeometric distribution is symmetrical in R and S , and we can approximate it by the binomial distribution:

$$\Pr(s) \approx \text{binomial}(R, S/N, s) = \binom{R}{s} \left(\frac{S}{N}\right)^s \left(1 - \frac{S}{N}\right)^{R-s} \quad (6)$$

We apply the HG model to an ideal case, where the documents in the sample are assumed to be perfectly ranked.

2.14. Expected P@ n values, assuming perfect ranking

The assumption of perfect ranking, applied to any sample, is clearly a simple form of invariance assumption about samples. In the SD model, perfect ranking implies that the two distributions have no overlap at all. Here we pursue a discrete analysis.

If the ranking mechanism were perfect, all s relevant documents in the sample would occupy the top s spots in the ranking and the precision measure would be determined by s

and n . That is, $P@n$ regarded as a function of s is:

$$P@n(s) = \frac{\min(s, n)}{n}$$

Assuming perfect ranking, the value of $P@n$ averaged across all possible S -document samples depends upon S as follows. Consider the simplest case, that of $P@1$. $P@1$ is constrained to be either 0/1 or 1/1 for any particular sample but account must be taken of all cases where $1 \leq s \leq \min(R, S)$. Its average over all samples is given by $\Pr(s \geq 1)$, or equivalently $1 - \Pr(s = 0)$.

In general, $\bar{P}@n$ (the expected average value of $P@n$ over a large number of samples) is a weighted sum of probabilities:

$$\bar{P}@n = \sum_{i=1}^S P@n(i) \cdot P_r(s = i)$$

In the HG model, this becomes

$$\bar{P}@n = \sum_{i=1}^{\min(R, S)} \frac{\min(i, n)}{n} \text{hyperg}(N, R, S, i) \quad (7)$$

$$= 1 - \sum_{i=0}^{n-1} \frac{n-i}{n} \text{hyperg}(N, R, S, i) \quad (8)$$

We now show that, under the assumption of perfect ranking, $\bar{P}@n$ is monotonic with both n and S , as we have already established for the SD model.

2.14.1. $\bar{P}@n$ is monotonic with n . Specifically, $\forall n \geq 1, \bar{P}@n \leq \bar{P}@n - 1$.

This may be proved as follows. By inspection of Eq. (7):

1. the number of terms in the summation is independent of n ,
2. the probability factor in each term is independent of n ,
3. the weight of the n th and all subsequent terms is 1,
4. considering the $n - 1$ terms whose weights are less than 1, the weight decreases strictly monotonically with increasing n .

Since probabilities cannot be negative, the above implies that each term in the summation for $\bar{P}@n$ is either equal to or less than the corresponding term in the summation for $\bar{P}@n - 1$.

2.14.2. $\bar{P}@n$ is monotonic with S . Denoting by $\bar{P}@n(S)$ the average considered as a function of S and by $P@n(S)$ the specific value for a set \mathcal{S} ,

$$\forall S > n \geq 1, \quad \bar{P}@n(S) \geq \bar{P}@n(S - 1)$$

This may be proved as follows. Consider first a sample \mathcal{S}^- of size $S - 1$, and one extra document, giving a sample \mathcal{S} of size S . The extra document may be relevant and in the top

n , or it may not. If it is not relevant, the perfect ranking assumption implies that it cannot displace any of the relevant documents already in \mathcal{S}^- . Thus

$$P@n(\mathcal{S}) \geq P@n(\mathcal{S}^-)$$

Consider now the class of all such pairs $\{\mathcal{S}^-, \mathcal{S}\}$, that is, every pair of samples \mathcal{S}^- and \mathcal{S} such that $\mathcal{S}^- \subset \mathcal{S}$, $|\mathcal{S}^-| = S - 1$ and $|\mathcal{S}| = S$. In this class, every possible sample of size S occurs exactly S times, and every possible sample of size $S - 1$ occurs exactly $N - S + 1$ times. Therefore the average $P@n(\mathcal{S})$ for this class is exactly $\bar{P}@n(S)$, and similarly for \mathcal{S}^- and $\bar{P}@n(S - 1)$. Since the above inequality is true for every such pair, the result follows.

Furthermore, equality is excluded whenever the following condition is met: there exists a sample \mathcal{S}^- containing less than n relevant documents, and there exists also an additional relevant document. A sufficient condition to guarantee this possibility (that is, its truth for some of the pairs) is that $(R > 0) \wedge (R + S \leq N + n)$. This condition is certainly true in all interesting cases.

Thus in the case of perfect ranking, some degradation of $\bar{P}@n$ is a necessary consequence of sampling and of the limited number of relevant documents available.

2.14.3. Perfect ranking examples. Figures 15 and 16 show how, under the assumption of perfect ranking and in a hypothetical collection, the relationship between $\bar{P}@n$ and sample size S is affected by variations in n and R .

Note that, even with perfect ranking, $P@n$ will be less than unity for a collection whenever $R < n$. Table 1 shows a similar effect, this time with a real collection: a perfect \mathcal{Q}, \mathcal{E}

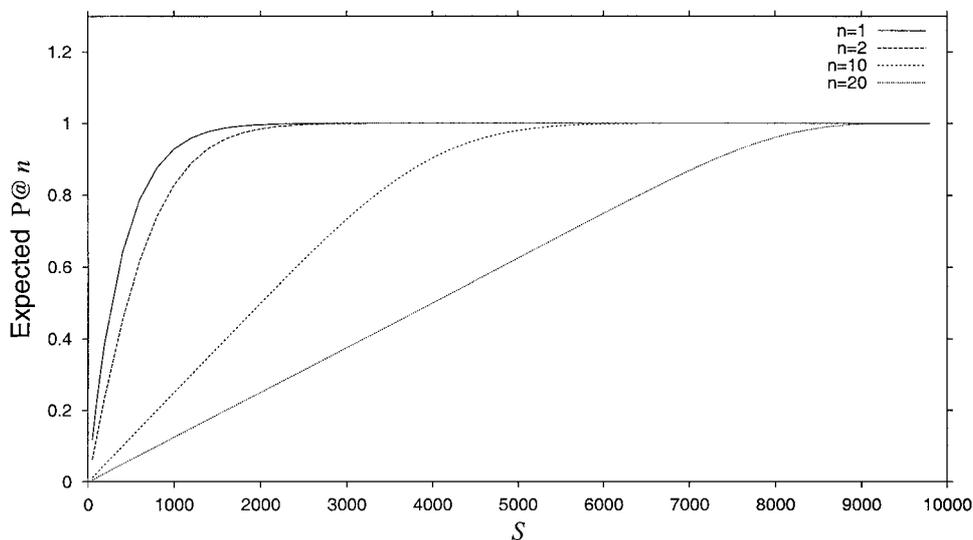


Figure 15. Perfect ranking. Variation of $\bar{P}@n$ with sample size S for a number of values of n using a hypothetical collection of $N = 10,000$ documents of which $R = 25$ are relevant.

Table 1. Perfect ranking.

Measure	Full collection	10% Sample	Ratio
$\bar{P}@1$	1	0.8712	1.148
$\bar{P}@20$	0.8430	0.3858	2.19

Comparison of $P@n$ values for the full TREC-6 collection with expected values for a 10% random sample based on actual R values for each TREC-6 Ad hoc topic.

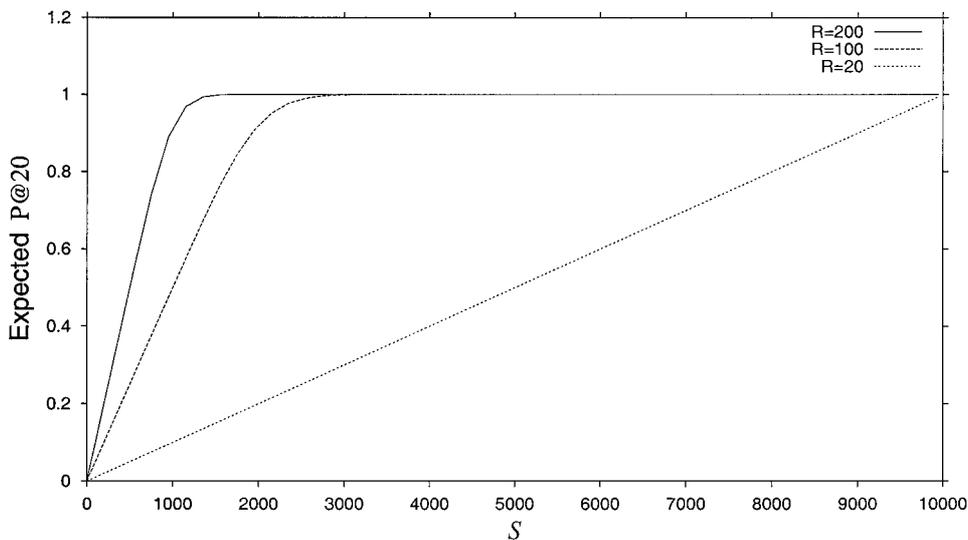


Figure 16. Perfect ranking. Variation $\bar{P}@20$ with sample size S using a hypothetical collection of $N = 10,000$ documents of which R are relevant, for a number of values of R .

combination, operating on a random 10% sample of the TREC-6 data and using TREC-6 ad hoc topics, would on average achieve $P@20$ scores less than half of those obtainable on the full collection. The effect is markedly less for the $P@1$ measure.

2.15. Another discrete model: Sampling after ranking

An alternative view of the effect of sampling the collection on the output of a search is to consider the two processes in the reverse order (as in the Cormack model described earlier). That is, instead of sampling the collection and repeating the search, we take the ranked output of the search in the full collection and sample that. This analysis applies strictly only to the case of deterministic invariance—we assume that the ranking (at least) stays exactly the same in the sample as it was in the collection. However, in contrast to the previous analysis based on the HG model, we do not assume perfect ranking. In contrast to the SD and Cormack models, we try to take account of discreteness.

If we start with the whole-collection ranking from a search, and consider sampling of the collection, we can generate the sample by taking each document in turn down the ranking, and including it in the sample or otherwise, independently of the decisions on previous documents, on the basis of any chosen sampling probability. This is not quite identical to the process previously assumed, because of the problem of replacement; in theory, the probability of choosing the next document is marginally affected by the previous choices. However, in realistic circumstances (we are only interested in the top-ranked documents, a tiny number compared to the size of the collection or sample), this effect will be so small as to be negligible. Thus for example if we wish to generate a 10% sample, rolling a ten-sided die for each ranked document in turn will give us, to all intents and purposes, the same kind of sample as we would have obtained by other means—at least as regards its effect on the top of the ranking.

Such a model might be the basis for a simulation experiment, starting with some real or imagined output (specific documents with specific relevance values at each rank position) and conducting Monte-Carlo type sampling many times, to assess the relationship between $\bar{P}@n$ and sample size, where $P@n$ is averaged over the samples. We might also attempt something more analytical, starting from a probability of relevance associated with each individual rank. Here we pursue the second line a short distance only.

Suppose that the probability of relevance of a document ranked i in the whole collection is p_i . We might imagine this probability arising in at least two possible ways: either we are in some way averaging over queries, or the whole collection is itself a sample from some even larger population of potential documents (as in the Cormack model). There is, however, a problem with these probabilities, discussed below.

If we nevertheless assume for the moment that the set of p_i s is a reasonable representation of the situation, then we can further reasonably assume that $p_{i+1} \leq p_i$. This is very similar to the argument concerning ‘reasonable’ SD models in Section 2.8 (and that in Robertson 1969)—if we had a system for which, say, $p_{12} > p_{11}$, then the trivial modification to the system of always interchanging the eleventh and twelfth documents in the output ranking would improve the performance of the system. As in the SD analysis, it would be compatible with this idea that the first few probabilities were the same, but at some point there must be a decline; once having declined, the probabilities might remain at the same level or decline further, but cannot increase again.

Given these probabilities, we can specify an expected value for $P@n$ in the collection:

$$E(P@n) = \frac{1}{n} \sum_{i=1}^n p_i$$

Declining p_i as we go down the ranking then predicts declining $P@n$ with increasing n .

Sampling now involves choosing some of the ranked documents and excluding others. Each ranked document chosen for the sample carries with it an associated p_i , but in order to get a fixed number n we must take them from further down the ranking, the p_i s are in general lower than their corresponding ones in the full collection. So again a reduction in $P@n$ in the sample is predicted.

The problem with this model is as follows. There is an implicit assumption (nevertheless very necessary for the equation given for $E(P@n)$) that the relevance of the $i + 1$ th ranked

item is independent of the relevance of the i th item. In fact there may be a rather subtle, second-order dependence, which may be seen if we consider the matter in terms of scores (as in the SD model). Assuming the scores have some correlation with an absolute probability of relevance, if we discover that the i th item is not relevant, it suggests we are now (at this point in the ranking) rather lower down the scale than if it were relevant. Hence p_{i+1} may be lower.

This problem seems to make any further development of this model (with explicit relevance probabilities) highly problematic. We therefore pursue it no further here. However, putting aside relevance for the moment, a similar approach with sampling after ranking can give us insight into another aspect of discreteness, as discussed in the next section.

2.16. *The granularity of n*

In the discussion of the continuous SD model, we suggested that $P@n$ in the collection should be the same as $P@\frac{nS}{N}$ in the sample. The discreteness of n may however disturb this result. In order to make this argument easier to follow, we use a specific example: we compare $P@20$ in the full collection with $P@2$ in a 10% sample.

The SD model would predict, more specifically, that if we take a score somewhere between the twentieth and the twenty-first ranked documents in the full collection, then that score will occur (on average, in some sense) between the second and third documents in the sample; therefore giving the same precision at 20 documents in the collection and 2 in the sample. However, it does not predict that the exact score of the 20th document in the collection is the same as the 2nd in the sample. Indeed, since it is likely that the gap between the scores of the 20th and 21st in the collection is smaller than that between the second and third in the sample, it also seems likely that the score of the 2nd document in the sample is higher than that of the 20th in the collection. In this case, assuming that $P@n$ declines with n , $P@20$ in the collection might be expected to come between $P@2$ and $P@3$ in the sample.

This prediction is investigated below.

2.17. *$P@n$ for very small samples*

Calculating $P@n$ for an empty sample is not very meaningful—indeed it does not make sense to talk about $P@n$ if $S < n$. However, it is useful to have some form of limiting value for $\bar{P}@n$ as S becomes small.

In fact, for any S , $\bar{P}@S = R/N$ (from Eq. (5)). Thus we can define a minimum point on the curve for any given n as the point ($S = n$, $\bar{P}@n = R/N$). Again, for the data analysed below, which represents topics with varying R , we use average R and set this point to

$$(S = n, \bar{P}@n = \bar{R}/N). \quad (9)$$

3. Experimental framework

This section describes the conditions under which the experiments described in Sections 4 and 5 were conducted.

3.1. Data sets, queries and measures

Some of the data analyses included here are based on evaluations of official submissions by participants in the VLC track. Submissions consisted of the top 20 documents returned by a system in response to a query.

The VLC (like the TREC collections) is distributed in the form of hierarchical directories of compressed document bundles. Each bundle contains one or more (sometimes many) individual documents. The VLC sample was created by traversing the directory structures and selecting every tenth bundle encountered.⁷

Unfortunately, only the first twenty documents retrieved by each run were judged for relevance. This certainly means that only a small proportion of the relevant documents in the VLC have been identified and consequently limits the type of follow-up study which can be performed using this data. Accordingly, many of the experiments reported here make use of other sets of TREC data for which relatively complete judgements are available.

Tables 2–4 summarise the document collections, the query sets and the measures used in the experiments. In each of the runs performed specifically for this paper, queries were evaluated using the PADRE retrieval system and the Okapi BM25 weighting function

Table 2. Data collections used in experiments.

Name	Size	No. of documents	Remarks
VLC	20.14 gB	7,492,048	Includes all five TREC CDs
TREC-6	2.09	556,077	CD4 plus CD5
TREC-3	2.02	741,856	CD1 plus CD2

Table 3. Query sets used in experiments.

Name	Topics	Generation method	Average no. of terms per topic
Q3A	151–200	Automatic, title plus description	11.8
Q6A	301–350	Automatic, title-only	3.7
Q6M	301–350	Manual, interactive	24.5

Table 4. Precision measures used in experiments.

Measure	Explanation
$P@n$	Precision at n documents retrieved.
R -precision	Precision at the (non-zero) point where precision = recall.
Average precision	$\frac{r_0}{n_0} + \frac{r_1}{n_1} + \dots$

Average precision is computed as in TREC, i.e. precision is calculated at each point in the ranking where a relevant document was retrieved. These precision values are then summed and divided by the total number of relevant documents (whether retrieved or not).

(Hawking et al. 1997, Robertson et al. 1994). The official TREC evaluation package `trec_eval` was used throughout after applying some minor modifications⁸ to allow precision to be reported at finer intervals for both the interpolated precision-recall and precision at fixed cutoff.

3.2. Method for creating sample collections

The VLC sampling method was generalised to allow creation of sample collections comprising documents from every n th bundle, starting with the i th, $0 \leq i < n$. It was thus possible to create n disjoint samples, each comprising close to $\frac{1}{n}$ of the original collection.⁹ Retrieval runs could then be performed independently over all the samples and the precision results averaged to eliminate biases resulting from unrepresentative sampling.

In order to investigate the way precision varies with increasing collection size, composite $\frac{2}{n}$, $\frac{3}{n}$, \dots , and $\frac{n-1}{n}$ samples were created by merging combinations of primary samples. In general, the number of possible combinations of n items taken r at a time may be too large to permit exhaustive testing. Therefore, n composite samples of each size were examined (where possible). Thus, the results reported for $\frac{3}{7}$ of a collection are the average of the results obtained using the composite samples:

(0, 1, 2), (1, 2, 3), (2, 3, 4), (3, 4, 5), (4, 5, 6), (5, 6, 0), and (6, 0, 1)

Note that this approach ensures that:

1. The average measures reported for a particular nominal sample size take into account all the data and weight it uniformly; and
2. Although the actual sizes of samples of the same nominal size may vary, the average of their sizes is exactly the desired proportion of the full collection.

For each primary and composite sample collection the appropriate subset of the official TREC relevance judgments file was extracted and used with the `trec_eval` program when evaluating runs.

3.3. Samples actually employed

Using the methods described above, ten disjoint uniform $\frac{1}{10}$ primary samples and eighty-one¹⁰ composite samples were made of the TREC-6 and TREC-3 collections. Table 5 shows the distribution of judged and relevant documents across the TREC-6 data set. The number of documents judged per sample averages 7,251 and varies by only $\pm 4\%$.

3.4. Method for creating replicated collections

Replicated collections were constructed by taking each $\frac{1}{10}$ primary sample of the TREC-6 data and duplicating them the required number of times, up to ten-fold, resulting in one hundred different composite collections:

(0), (0, 0), \dots , (0, 0, 0, 0, 0, 0, 0, 0, 0, 0), (1), \dots , (9, 9, 9, 9, 9, 9, 9, 9, 9, 9)

Table 5. Details of the primary 10% samples of the TREC-6 data.

Sample	Judged				Relevant			
	Total	Least/top.	Most/top.	Zeroes	Total	Least/top.	Most/top.	Zeroes
0	7548	82	204	0	461	0	44	3
1	7068	74	219	0	504	0	91	7
2	7340	89	213	0	471	0	54	5
3	7156	83	210	0	443	0	50	9
4	7482	77	205	0	469	0	51	10
5	7046	80	252	0	485	0	58	8
6	7568	96	213	0	460	0	69	7
7	7316	75	221	0	455	0	47	7
8	7052	85	220	0	450	0	59	7
9	6930	78	195	0	433	0	41	8
Full coll.	72,270	912	1902	0	4611	3	474	0

For each sample, four pieces of information are given about the judged set and about the relevant set: the number of documents in the set, the least number of documents on any one topic, the greatest number of documents on any one topic and the number of topics for which this sample contained no documents in the set.

3.5. Method for creating biased sub-collections

The TREC-6 data was divided into five disjoint sub-collections on the basis of source: Congressional Record (235.4 MB), Federal Register (394.9 MB), Financial Times (564.1 MB), Foreign Broadcast Information Service (470.2 MB) and Los Angeles Times (475.3 MB). These documents vary considerably in probability of document relevance. Unfortunately, they also vary considerably in size and number of documents.

4. Confirmation of precision increase

The experimental environment of the VLC track comprises seven different query sets, each with its own specific query processing method, but only one sample collection and only one reported precision measure (P@20).

Here, a wider range of measures is computed for large numbers of samples of both the TREC-6 and TREC-3 data and for biased subsets of the TREC-6 data.

4.1. Uniform samples

A range of precision measures were computed for several query sets processed over primary and composite samples of both TREC-6 and TREC-3 data.

Table 6 shows TREC-ad hoc-collection values for the ratio used in the VLC track, namely the ratio of precision at 20 documents retrieved for the full collection and for a 10% sample.

Table 6. Ratios of P@20 between full collections and averaged 10% samples for different combinations of collection and query set. (Using the PADRE(Okapi) retrieval system.)

Collection	Query set	P@20 (sample)	P@20 (full)	P@20 ratio
TREC-6	Q6M	0.169	0.467	2.76
TREC-6	Q6A	0.144	0.294	2.04
TREC-3	Q3A	0.235	0.436	1.86

In Section 2.14.3 it was shown that the ratios which would be observed for a perfect Q, \mathcal{E} combination on the TREC-6 data and topics would be 2.19. Of the two experimentally observed TREC-6 ratios, one is considerably higher than this value and the other lower. It is interesting that the better-performing query set is associated with the higher of the two ratios. The observed differences between the ratios must be due to the different queries, as the same collection and the same engine was used in both cases.

The experimental ratios are much higher than were observed in the VLC track but it is believed that this is due to low numbers of relevant documents on some topics. Table 5 shows that, for each 10% sample of the TREC-6 data, somewhere between 3 and 10 of the 50 TREC-6 topics were not associated with any relevant documents. Because the VLC contains 13.5 times as many documents as the TREC-6 data, 10% samples are less likely to contain few or no relevant documents, even though the probability of individual document relevance is somewhat reduced (Hawking et al. 1999).

Figures 17–19 show how the precision at fixed cutoff (5, 10 and 20 documents retrieved) varies with the size of the sample collection. Each of the nine curves shows a characteristic shape which can be approximated quite closely as a combination of a linear function and an exponential:

$$c_0 + c_1x + c_2e^{-kx}$$

Note that the minimum point on the graphs is as established in Section 2.17.

Figures 20 and 21 show R-precision and average precision measures for Q6A and Q6M over the TREC-6 data. It is notable that both measures are quite constant despite large variation in sample size.

4.2. Replicated collections

Figure 22 shows that the P@20 results obtained for an $f\%$ subset of a collection are, on average, remarkably independent of whether the subset was obtained by combining disjoint uniform samples or by replicating one such sample. The P@5 plots shown in figure 23 are not as similar, due to the fact that P@5 outcomes for the 50% and bigger replicated collections are completely determined by the first document returned from the basic sample. If it is relevant then the five largest replicated collections constructed from it achieve $P@5 = 1$ otherwise they score $P@5 = 0$.

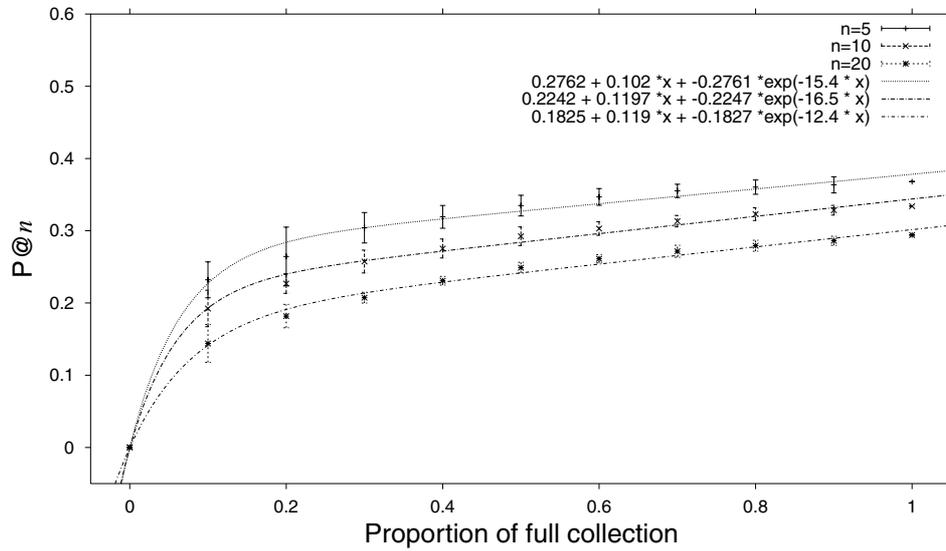


Figure 17. Variation of precision at fixed number of documents retrieved with sample size, using uniform samples of the TREC-6 data and the Q6A queries. The data show the mean and standard deviation of the values for each of the samples of the particular size. The general form of the best-fit curves is given in the text.

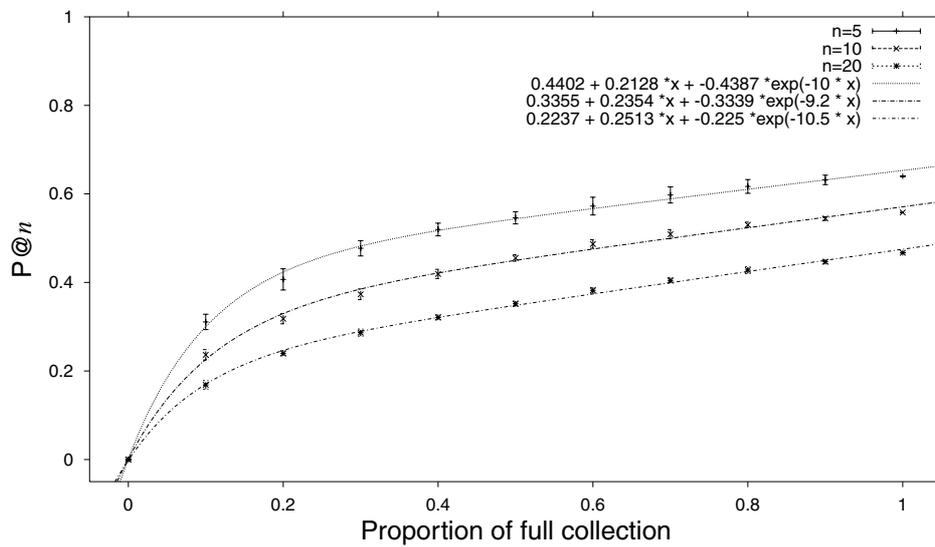


Figure 18. As for figure 9 but using the Q6M queries.

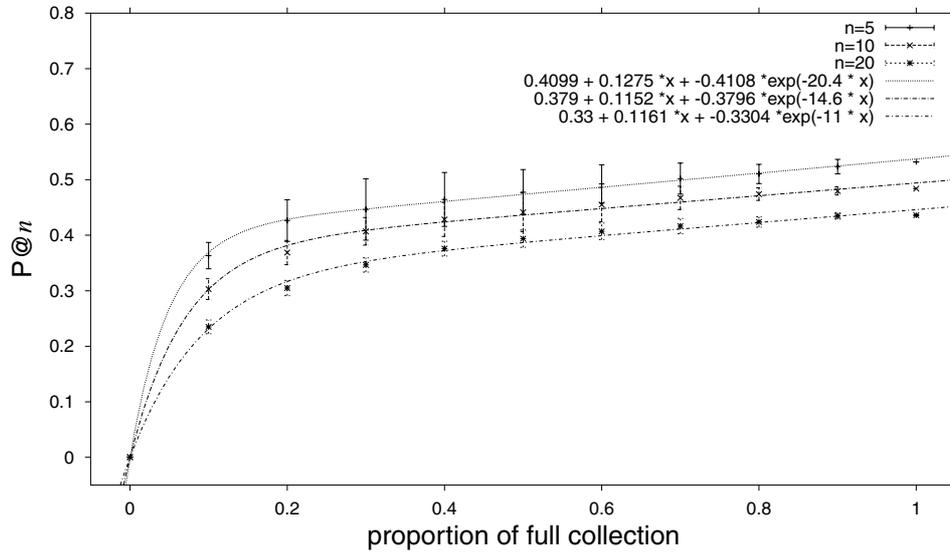


Figure 19. As for figures 9 and 10, but using uniform samples of the TREC-3 data and the Q3A queries.

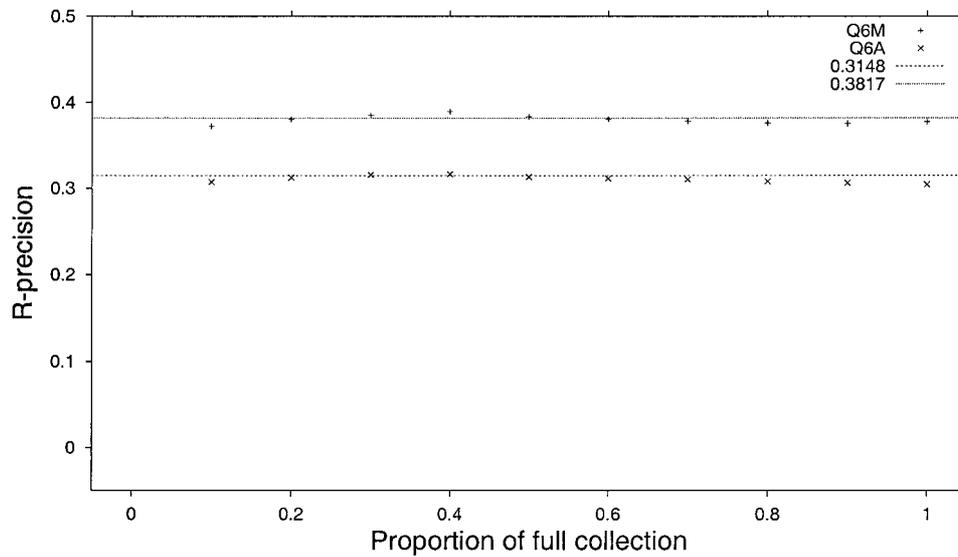


Figure 20. Variation of R-precision (precision at the point where (non-zero) precision and recall are equal) with sample size, using uniform samples of the TREC-6 data and query sets Q6A and Q6M. The mean and standard deviation of the values for each of the samples of the particular size are shown.

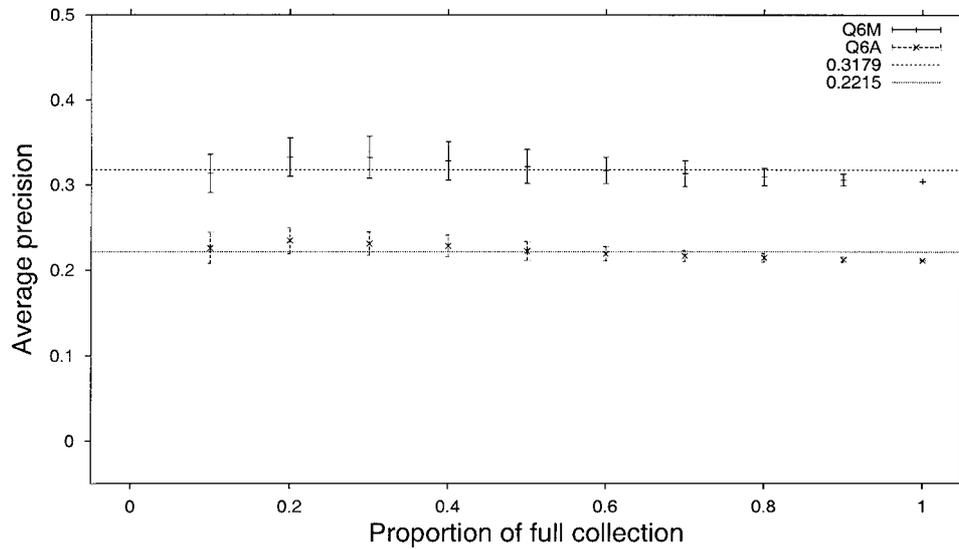


Figure 21. Variation of average precision (as defined in TREC) with sample size, using uniform samples of the TREC-6 data and query sets Q6A and Q6M. The mean and standard deviation of the values for each of the samples of the particular size are shown.

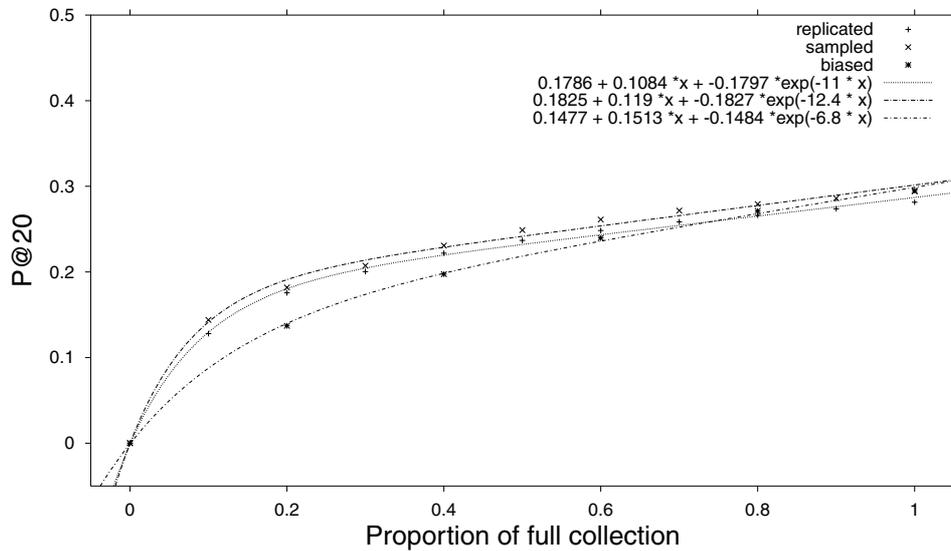


Figure 22. Variation of P@20 with sample size, using subsets of the TREC-6 data and Q6A queries created by replication, by sampling and by biased selection. The sampling line corresponds to the P@20 line in figure 9. The same minimum point (derived as per Section 2.17) is used for all three curves.

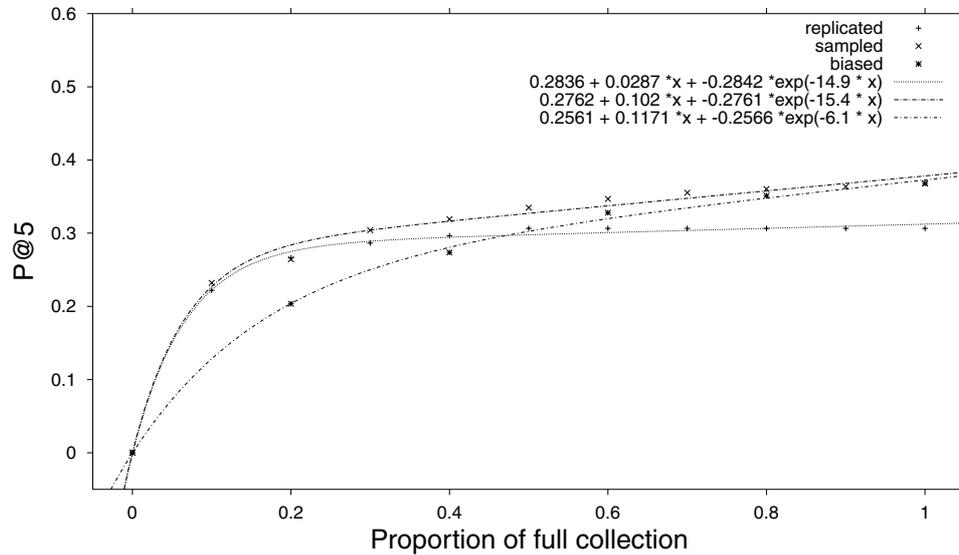


Figure 23. Variation of P@5 with sample size, using subsets of the TREC-6 data and Q6A queries created by combining disjoint samples, by replication and by biased selection.

4.3. Biased sub-collections

Figure 22 shows that the P@20 results obtained for biased sub-collections of a collection also decline with diminishing average sub-collection size. Indeed, the decay is considerably faster than is the case for uniform sampling.

5. Hypothesis testing

In this section, the five informal hypotheses listed in Section 1.4.1 are analytically examined and empirically tested.

5.1. Hypothesis 1

“Sample P@20 measurements are necessarily lower because large numbers of topics have insufficient relevant documents in the baseline sample to achieve the P@20 scores observed on the full VLC.”

5.1.1. Analytical. Table 1 shows that P@n values are inevitably reduced when the size of a sample collection is reduced to the point where it contains only a small number of relevant documents. It is therefore certain that Hypothesis 1 is capable of explaining reduced precision at fixed number of documents retrieved for a sample collection.

Table 7. P@20 results recorded by participants in the VLC track, averaged over topics for which there are known to be at least 20 relevant documents in the baseline sample.

Group	Baseline	VLC	Ratio
City	0.4979	0.6937	1.39 (1.61)
ATT	0.5375	0.7125	1.33 (1.52)
ANU/ACSys	0.5667	0.6708	1.18 (1.43)
UMass	0.5854	0.7042	1.20 (1.31)
IBMg	0.4479	0.5500*	1.23 (1.31)*
IBMs	0.4292	0.5375	1.25 (1.28)
U Waterloo	0.7375	0.8458	1.14 (1.29)

There were 24 such topics. The ratios in parentheses are those obtained using all 50 topics.

*IBMg (Brown) are based on approximately 85% of the VLC data.

5.1.2. Empirical. Examination of the VLC track judgments suggests that Hypothesis 1 was responsible for at least a part of the observed change in early precision, as there are seven topics for which there were fewer than five known relevant baseline documents (baseline $P@20 < 0.25$) and thirteen with less than ten baseline relevants (baseline $P@20 < 0.5$). However, it is possible that other factors also contributed to the observed phenomenon.

This possibility was investigated by considering only VLC topics known to have at least 20 relevant documents in the baseline sample. For these topics, any reduction in $P@20$ cannot be due to an insufficiency of relevant documents.

Table 7 shows $P@20$ figures calculated on this basis for each of the VLC track runs. As may be seen the VLC: baseline ratios are reduced but are still consistently above unity, suggesting that Hypothesis 1 is not, by itself, sufficient to explain the observation. Even when there were sufficient relevant documents in the sample to achieve $P@20 = 1.0$, all of the retrieval systems in the VLC track were able to achieve appreciably better $P@20$ scores using the larger collection.

It is concluded that Hypothesis 1 specifies an absolute upper bound on sample $P@20$ for some topics and thus contributes significantly to the observed precision decline but that it is not a comprehensive explanation.

5.2. Hypothesis 2

“The first edition VLC is approximately equivalent to a replicated collection in which each baseline document is repeated ten times. From this, it is expected that the $P@20$ value for a particular Q, \mathcal{E} combination over the 20 gigabyte collection should equal the corresponding $P@2$ value for the baseline, when averaged over a sufficiently large number of topics.”

5.2.1. Analytical. The replication argument is in some sense the reverse of the analytical approach taken throughout the analytical models section (Section 2). It might be taken as a simple model of collection growth, though it is clearly a very unsubtle one.

The prediction that $P@20$ in the full collection should equal $P@2$ in the sample is consistent with arguments presented in the analytical models section. However, again the prediction is unsubtle, and does not appear to allow room for arguments such as those presented in Section 2.16, to the effect that the match might not be exact. In particular, if the situation is exactly one of replication, the two precision values should match exactly.

5.2.2. Empirical. The results for replicated collections presented in Section 4.2 above, using the ANU/ACSys retrieval system and TREC-6 data, suggest that a complete collection does tend to behave like multiple replication of a uniform sample. (Obviously, this cannot apply if the initial sample is too small.) Hypothesis 2 was further tested by comparing the two precision values for each pair of VLC track runs. Table 8 shows the VLC $P@20$ value for each run and the approximate intersection point with the $P@n$ curve for the corresponding baseline run. The process is illustrated graphically for the ANU/ACSys runs in figure 24, where the horizontal line extended from the 20 documents retrieved point on the VLC line intersects the baseline graph at about $n = 2.9$ documents retrieved. No intersection occurred for City or for IBMg (see figures 25 and 26).¹¹

These results do not fit very well with the exact prediction of Hypothesis 2. They do however shed some interesting light on the less exact predictions made in Section 2. Concerning the cases where the intersection is non-existent (City, IBMg) and the cases where it is far from 2 (IBMs, UMass), the strong suggestion is that these are Case 3 systems, with fundamental effects of the collection on the behaviour of the scoring system. In at least City's case, this is correct: the queries were expanded using pseudo relevance feedback over the test document set, leading to very different final queries in the two runs.

Concerning the other three cases, ANU/ACSys and ATT show the effect predicted in Section 2.16: the closest baseline n figure is between 2 and 3. The Waterloo result does not. However, it has already been observed that the Waterloo $P@n$ for the full collection is more-or-less flat over the range $n = 1, \dots, 20+$; the Section 2.16 result assumed a declining $P@n$.

Table 8. Intersection of VLC $P@20$ with baseline precision at n documents retrieved curves.

Group	All topics		Topics w. > 12 baseline rel.	
	VLC $P@20$	Closest baseline n	VLC $P@20$	Closest baseline n
ANU/ACSys	.509	3.4	.6409	2.9
ATT	.530	2.7	.6712	2.6
City	.515	–	.6682	–
IBMg	.361	–	.4606	–
IBMs	.348	7.1	.4561	5.7
UMass	.505	4.8	.6515	5.2
U Waterloo	.643	4.5	.8121	1.4

Tabulated for all topics and also for the 33 topics for which there were at least 13 relevant documents.

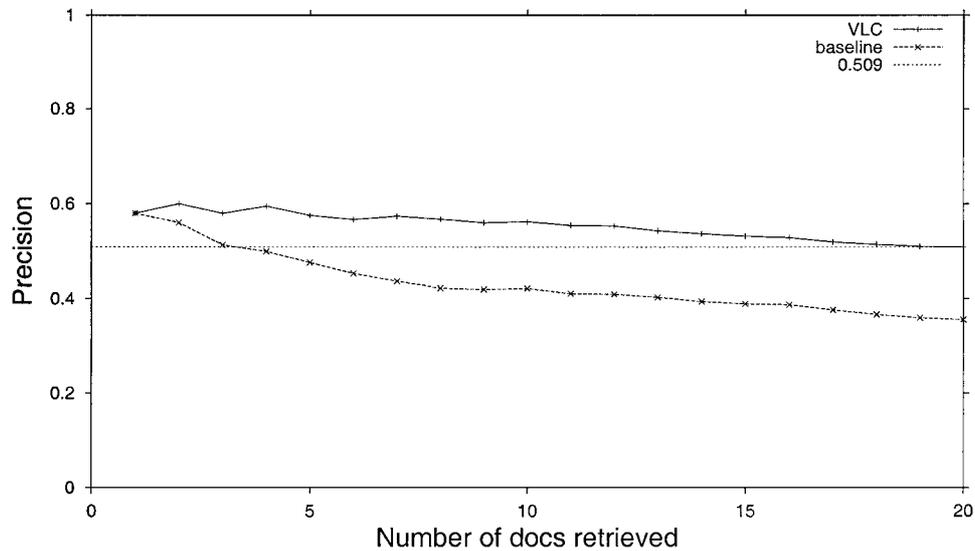


Figure 24. Comparison of precision at n documents retrieved for ANU/ACSys baseline and VLC runs. Averaged over all 50 topics.

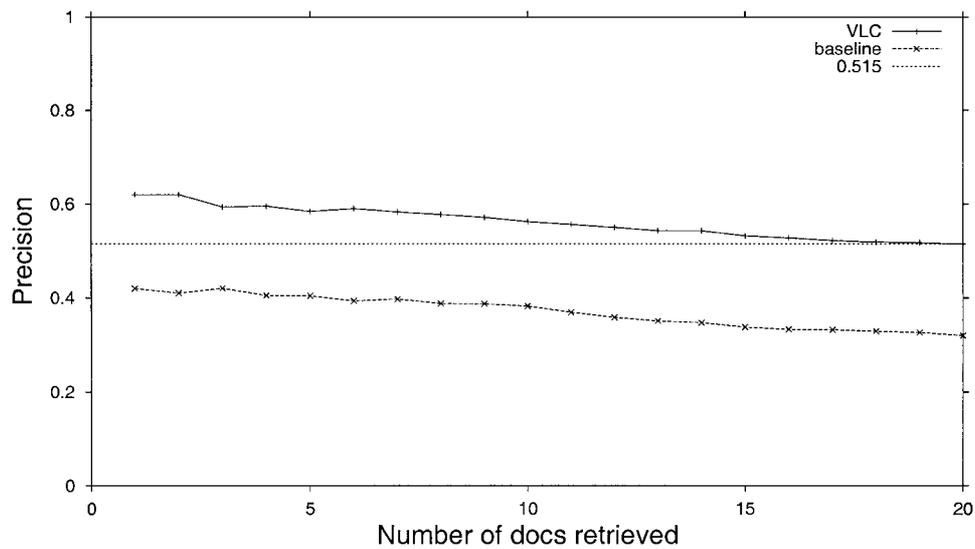


Figure 25. Comparison of precision at n documents retrieved for City baseline and VLC runs. Averaged over all 50 topics.

It is concluded that the replication hypothesis is not a generally useful one. Although it appears to have some validity for some systems, it is not really subtle enough to address the various complications.

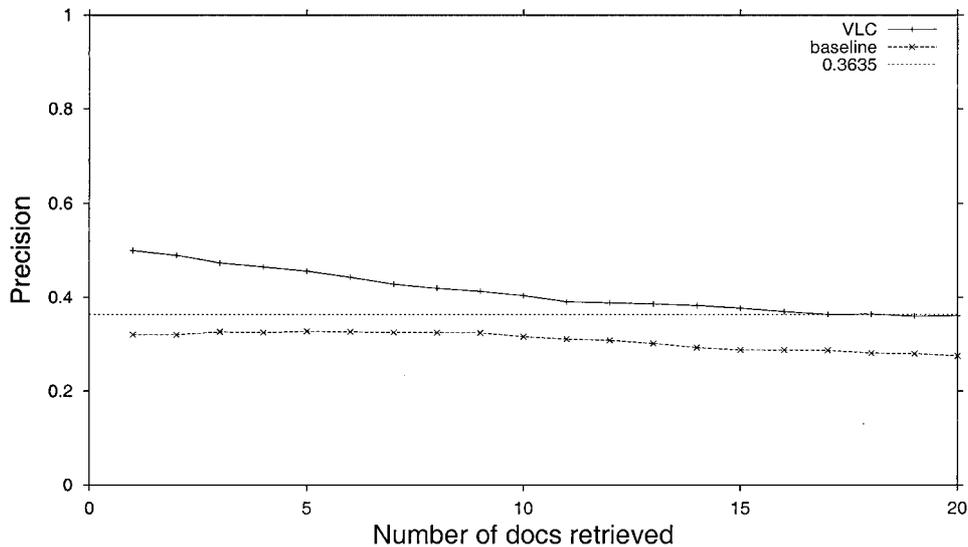


Figure 26. Comparison of precision at n documents retrieved for IBMg baseline and VLC runs. Averaged over all 50 topics.

5.3. Hypothesis 3

“Precision-recall curves, as plotted for TREC, represent an operating characteristic for the Q , \mathcal{E} , \mathcal{C} combination represented by the VLC runs. Because the baseline is a uniform sample of the VLC, the operating characteristic is the same in both cases. If precision monotonically declines with increasing recall, as is usually the case, the probability that a particular document in the ranked list returned by a system is relevant also declines with increasing recall. The effect of increasing the collection size (adding both relevant and irrelevant documents) is to change the number of relevant documents represented by each recall value. (For example, recall of 0.1 may represent 10 documents in the baseline and 100 in the VLC.) If the precision-recall curve is constant for both collections then precision at a fixed number of documents retrieved must increase.”

5.3.1. Analytical. The notion of a precision-recall curve as an OC curve was discussed at some length in Section 2.9. Although the hypothesis is to some degree consistent with that discussion, various problems were raised there. In particular, the average of the precision-recall curves for a set of topics cannot usefully predict average $P@n$ values for that set of topics because the function which maps the number of relevant documents in the collection and the precision-recall curve to a $P@n$ value is not linear.

Consequently, the analysis presented here and the experiments immediately following relate, not to precision-recall curves, but to the type of OC curve recommended in Section 2.9, namely recall-fallout. Fallout is the better measure to take as the control variable, because it is likely to be more stable than recall (in other words, an OC curve should indicate average recall for each value of fallout, not vice-versa).

The total number of relevant documents may be safely assumed to be very much smaller than the number of documents N in the collection, hence fallout for a given collection size is well approximated by the number of irrelevant documents retrieved. Accordingly, it may be useful to plot $R@n_i$ (recall at n_i irrelevant documents retrieved) against n_i (number of irrelevant documents retrieved).

When comparing different collection sizes, it seems appropriate to consider proportional numbers of irrelevant documents retrieved. Our hypothesis is that $R@(f \times n)_i$ for a full collection would be predicted by $R@n_i$ for a $1/f$ sample collection. We further hypothesize that if such a prediction works for a single topic, it should work for macro-average recall, because of the apparent linearity of the operations involved.

5.3.2. Empirical: Is (Q6A, PADRE(Okapi)) an example of Case 2b? We processed the Q6A query set (without relevance feedback) over both the TREC-6 collection and samples of it using the PADRE retrieval system. The relevance weighting model was the Okapi BM25 (2.0, 0.0, inf, 0.75) function (Robertson et al. 1994)

$$w_t = tf_q \times tf_d \times \frac{\log\left(\frac{N-n+0.5}{n+0.5}\right)}{2 \times \left(0.25 + 0.75 \times \frac{dl}{avdl}\right) + tf_d} \quad (10)$$

where w_t is the relevance weight assigned to a document due to query term t , tf_q is the frequency of t in the query, tf_d is the number of times t occurs in the document, N is the total number of documents, n is the number of documents containing at least one occurrence of t , dl is the length of the document and $avdl$ is the average document length.

Document scores are thus dependent on collection parameters. Consequently the (Q6A, PADRE(Okapi)) combination is not an example of Case 1.

We computed OC curves for (Q6A, PADRE(Okapi)) over the full collection and over samples of it. We approximated fallout by the number of irrelevant documents retrieved, scaled to reflect relative collection size. If the (Q6A, PADRE(Okapi)) combination were an example of Case 2b, the OC curves should be similar.

Figure 27 shows how OC curves for (Q6A, PADRE(Okapi)) compare for collections which differ by a factor of ten in size. As can be seen, the shapes of the curves are quite similar. However, one or two of the disjoint 10% samples overestimate the full collection curve and most slightly underestimate it. The two samples shown are those whose curves are furthest above and furthest below the full collection curve.

It is clear that some problems arise with this method in the process of averaging over topics. It is possible that some better method of averaging might improve the fit. However, it is also possible that the use of fallout as the control does not provide sufficient stability.

Figure 28 shows that predictions from 50% sample collections are subject to less variation and less error. Visually, one would say that the curves are effectively the same, for the collection and for a range of different 50% samples.

This suggests that the approximated recall-fallout curve describes the behaviour of this Q, \mathcal{E} combination well for a reasonably wide range of collection sizes.

It is concluded that Hypothesis 3 (modified to use a more tractable form of OC curve) may explain the part of phenomenon which is not due to Hypothesis 1. This would be

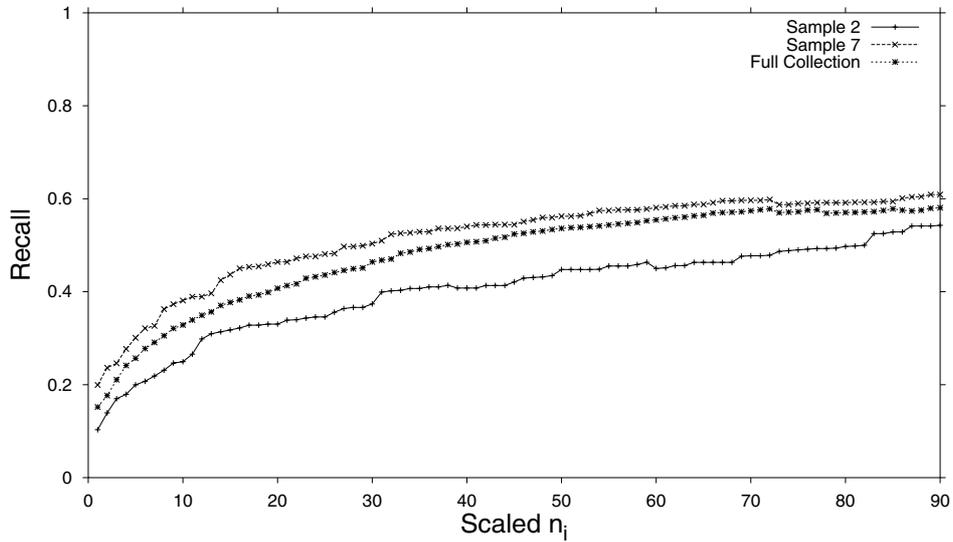


Figure 27. OC curves for (Q6A, PADRE(Okapi)), plotting recall against scaled number of irrelevant documents retrieved (n_i) for a full collection and two 10% samples of it. The samples chosen for display represent the limits of the range of curves for ten disjoint 10% samples. Scaling values are 1 for the samples and 10 for the collection, i.e. the point on the horizontal axis labelled 90 corresponds to 90 irrelevant documents retrieved in the sample and 900 in the full collection. Some points represent the average of fewer than 50 topics because: (1) sometimes a sample contained no relevant documents for a topic, and (2) sometimes too few documents achieved non-zero scores.

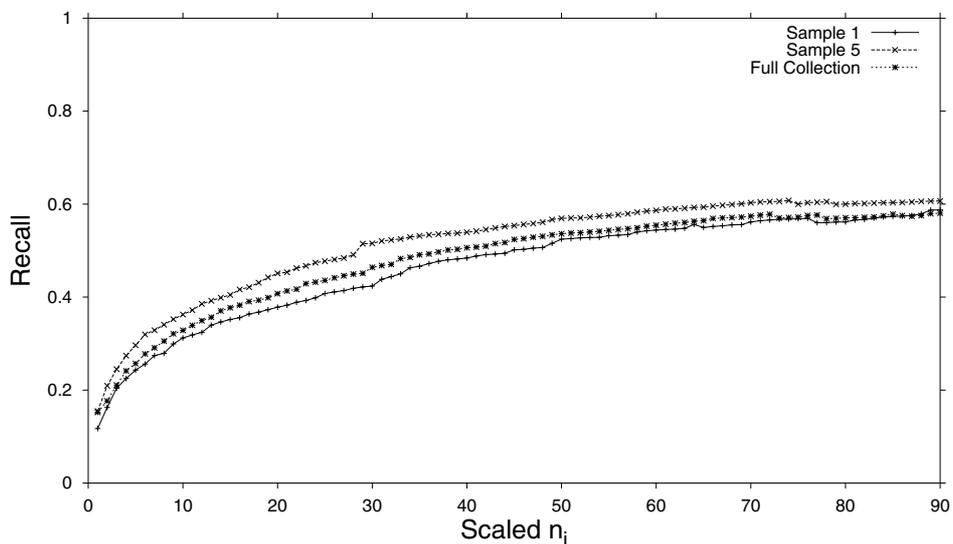


Figure 28. As for figure 27 but using 50% samples. Scaling values are 1 for the samples and 2 for the collection. Again, the two sample curves chosen for display represent extremes observed among ten samples.

consistent with other explanations arising from the various analytical models in Section 2, for Cases 1, 2a and 2b.

5.4. Hypothesis 4

“Swets (1963) postulated separate distributions of document scores for relevant and irrelevant documents (for a given \mathcal{Q} , \mathcal{E} , \mathcal{C} combination). He assumed that the two distributions were normal and that their means differed by an amount M which could be used to characterise the performance of the combination of query, collection and retrieval system. Robertson¹² suggested that, if the distributions are the same for the VLC and for the sample then taking a fixed number of documents to retrieve will result in a greater proportion of relevant documents among those retrieved. In order to retrieve a fixed *number* of documents, the cutoff score must be set much higher in the VLC than in the sample, resulting in a greater proportion of relevant documents among those retrieved.”

5.4.1. Analytical. The SD model used by Swets was the basis for much of the analytical argument in Section 2. Thus it might be seen as subsuming Hypothesis 3, and also replacing the replication idea of Hypothesis 2 with a more subtle analysis.

The SD analysis prompts a number of questions which might be answered experimentally. The most obvious one concerns the normality assumptions originally made by Swets; however, these are unlikely to be satisfied, and in any case are not crucial to the argument. Other investigations may focus on whether the system can be classified according to the cases described. Case 1 instances should be identifiable logically, without experiment; however, the distinction between Cases 2a, 2b and 3 may well require experimental evidence. We concern ourselves here with tests for these cases, in instances which do not appear to fall into Case 1.

The SD analysis prompts a number of questions deserving of experimental investigation. Empirical confirmation of the existence of systems conforming to Cases 2a, 2b and 3 could be sought in the following manner:

1. For a variety of systems \mathcal{E} (with suitable query sets \mathcal{Q}), compute relevant and irrelevant score distributions for both a collection \mathcal{C} and a sample \mathcal{S} .
2. Divide the systems into two classes on the basis of the results of a statistical test to determine whether the corresponding relevant and irrelevant distributions are the same. Those systems for which the two distributions are the same may be classified as Case 2a.
3. Those systems which fail the Case 2a test may still satisfy the test for Case 2b. This should be tested by comparing OC curves, as shown in Section 5.3. Systems which fail this test also must be regarded as Case 3.

Detailed experiments along these lines are left for future work. Here we restrict ourselves to demonstrating for the (Q6A, PADRE(Okapi)) combination, that, while the relevant and irrelevant distributions are far from normal in shape, they nevertheless overlap in such a way that the probability of relevance increases with increasing score. In other words, as more documents are retrieved, precision decreases.

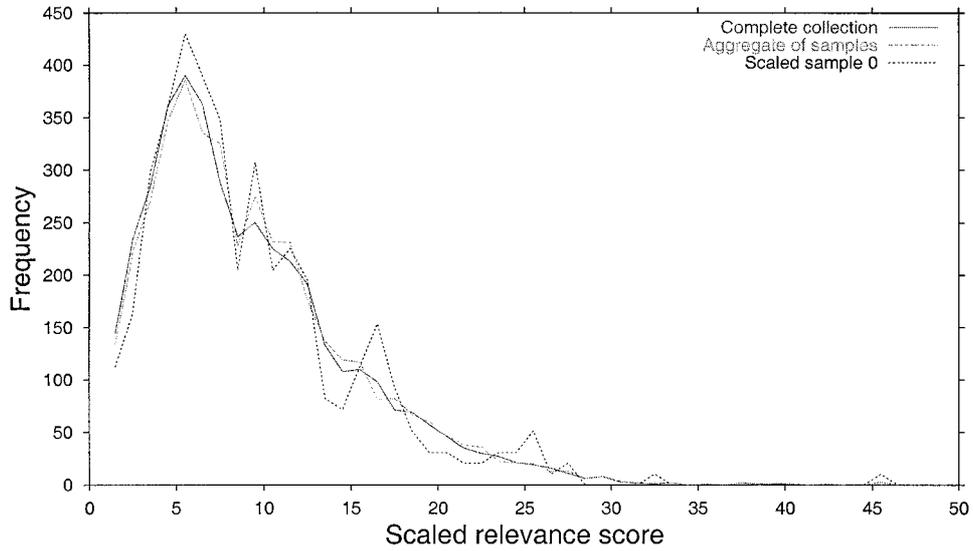


Figure 13. As for figure 14 but showing score distributions for relevant documents only. The means of the three distributions are: Full collection—8.99; sample 0—9.07; aggregate of samples—9.13.

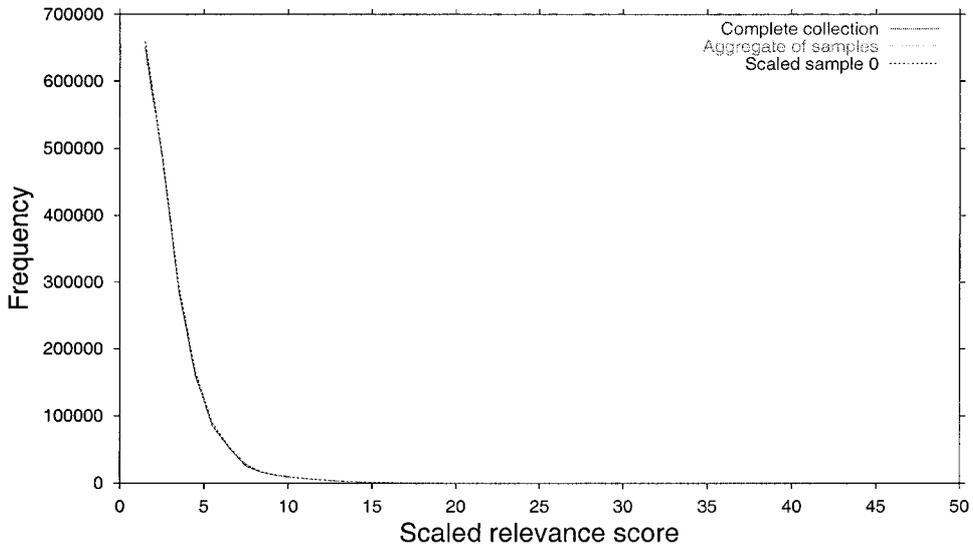


Figure 14. Distributions of relevance scores (Q6A, PADRE(Okapi)) for irrelevant documents only. All documents achieving a non-zero score are included. Scores are scaled relative to the theoretical maximum possible score for each query and assigned to one of 50 buckets. Distributions are shown for the full collection, for the aggregate of ten disjoint 10% samples and for an arbitrarily chosen 10% sample. For case of visual comparison, frequencies for the single sample are scaled up so that the scaled total for the sample matches the total for the full collection. The scaled means of the three distributions are: Full collection—2.49; Sample 0—2.55; aggregate of samples—2.53.

5.4.2. Empirical: Is (Q6A, PADRE(Okapi)) an example of Case 2a? We calculated relevant and irrelevant distributions for the Q6A queries against the TREC-6 dataset. We then generated corresponding distributions for the same queries over ten disjoint 10% samples of that dataset.

To achieve comparability of scores across different collections, we scaled the scores relative to a hypothetical maximum possible Okapi score achievable by any document within that collection. This would correspond to a zero-length document containing an infinite number of each of the query terms. Such a document might score differently in different collections (including samples) because of different n and N values.

Figures 14 and 13 show the relevant and irrelevant score distributions for the (Q6A, PADRE(Okapi)) combination. Visual inspection of them lends some support to the suggestion that relevant and irrelevant distributions may be modeled with exponential and Gaussian curves respectively (Arampatzis et al. 2000, Manmatha 2001). However the observed relevant distribution is significantly skewed and, when zero-scoring documents are included, shows a second peak at the extreme left.

Visually at least, the distributions of scores from the aggregate of samples are very close to the corresponding distributions for the full collection. This indicates that the (Q6A, PADRE(Okapi)) combination may indeed satisfy the Case 2a definition. A fuller investigation would require a significance analysis.

It is also the case that there is a distinct difference between the relevant and the irrelevant distributions. Consequently, there is a marked increase in probability (see figure 29) of relevance with increasing normalised relevance score. A larger collection implies a higher

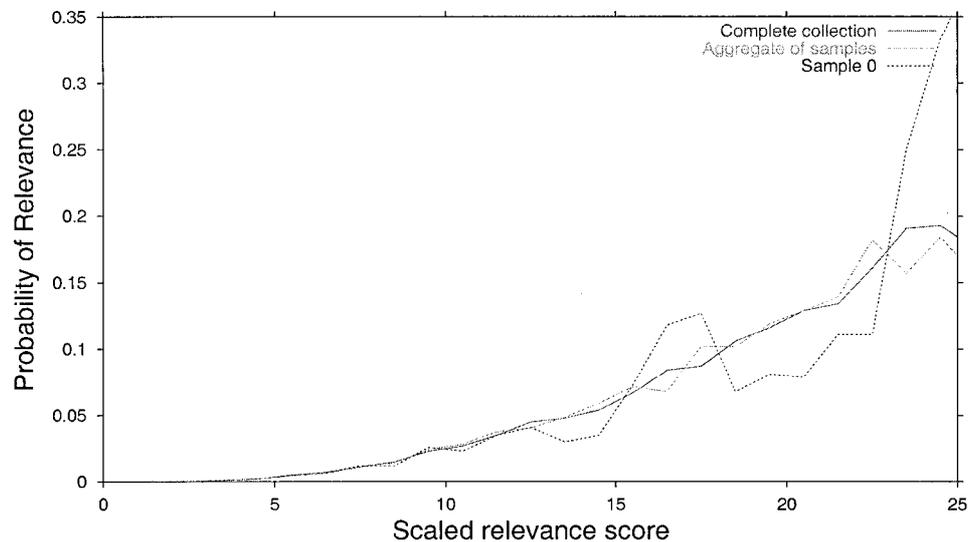


Figure 29. The probability that a scaled score lying in a particular range implies document relevance (using the same data as in figures 13 and 14). The high-end of the score range has been suppressed because it is based on very few data points and is consequently very noisy.

number of high-scoring documents and therefore a higher proportion relevant above a fixed cutoff in the ranking.

5.4.3. Empirical followup: Different collections. We have provisionally classified the (Q6A, PADRE(Okapi)) combination as a possible example of Case 2a, but we should confirm this on collections which are not uniform samples of each other. Accordingly, we plotted (Q6A, PADRE(Okapi)) score distributions and recall-fallout curves for the five biased sub-collections of the TREC-6 data.

Figure 30 shows that the degree of agreement between the full-collection OC curve and those of the biased sub-collections is rather lower than was the case with uniform collection samples. In the worst case, the agreement is quite poor.

Differences between the irrelevant document score distributions and that of the full collection are relatively small (e.g., figure 31). However, some of the relevant score distributions differ quite markedly from that of the full collection. The LATimes sub-collection (figure 32) differs least and the Federal Register 94 sub-collection (figure 33) differs most.

Not surprisingly, when the (Q6A, PADRE(Okapi)) combination is applied to a pair of collections in which the smaller is not a uniform sample of the larger, Case 2a does not, in general apply—the relevant distributions may differ considerably. The observation that Case 2b does not apply either, was less predictable. It implies that the behaviour of PADRE(Okapi) depends both on the query and on the collection.

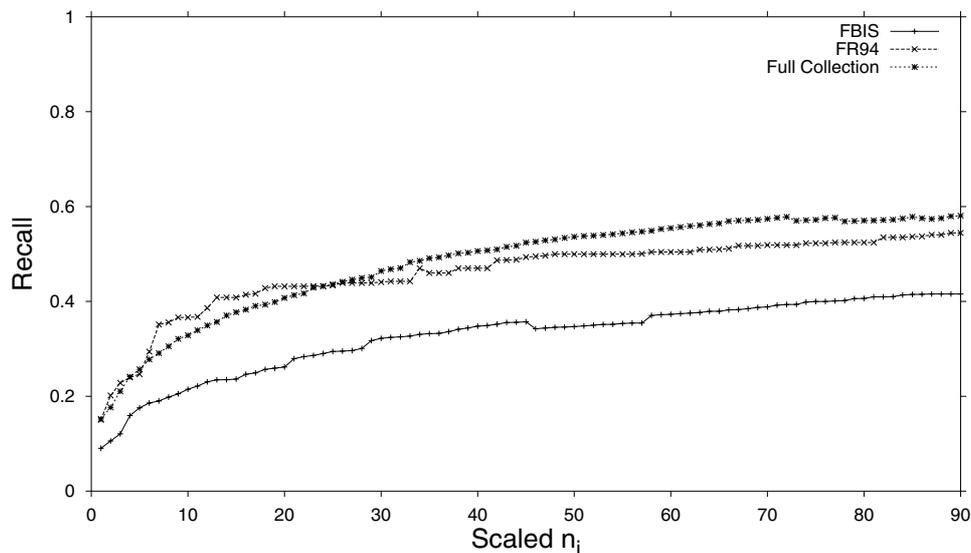


Figure 30. OC curves for (Q6A, PADRE(Okapi)), plotting recall against scaled number of irrelevant documents retrieved (n_i) for a full collection and two biased sub-collections of it. The samples chosen for display represent the limits of the range of curves for the five biased sub-collections. Scaling values are in approximate inverse proportion to the size of the collection/sub-collection. Some points represent the average of fewer than 50 topics because: (1) sometimes a sample contained no relevant documents for a topic, and (2) sometimes too few documents achieved non-zero scores.

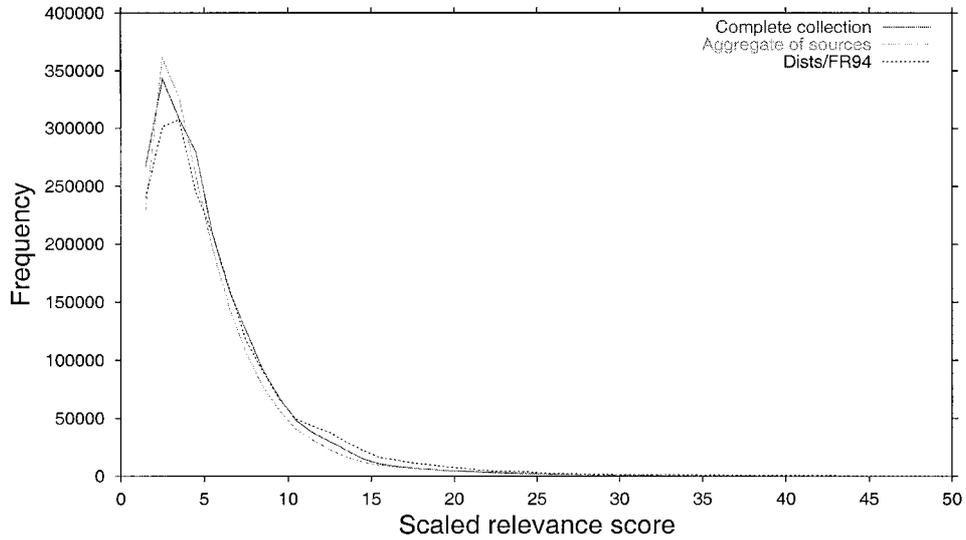


Figure 31. The (Q6A, PADRE(Okapi)) irrelevant distribution for the Federal Register 94 sub-collection.

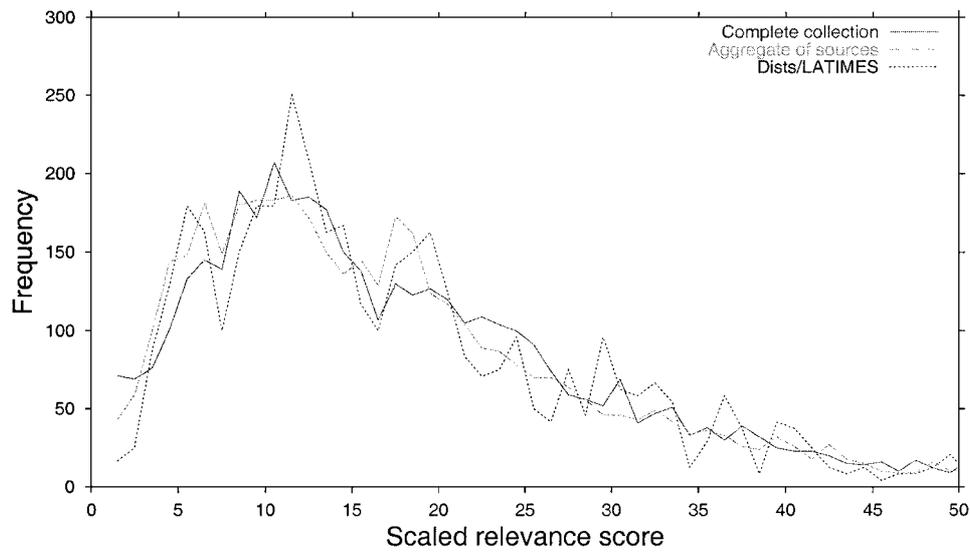


Figure 32. The (Q6A, PADRE(Okapi)) relevant distribution for the LA times sub-collection.

Despite the lack of normality of the relevant and irrelevant score distributions, the signal detection model provides a useful way of describing the observed behaviour and explaining the decline in P@20 in sample collections. However, both the score distributions and the OC curves derived from them, appear to depend on the Collection as well as the Query and the Engine.

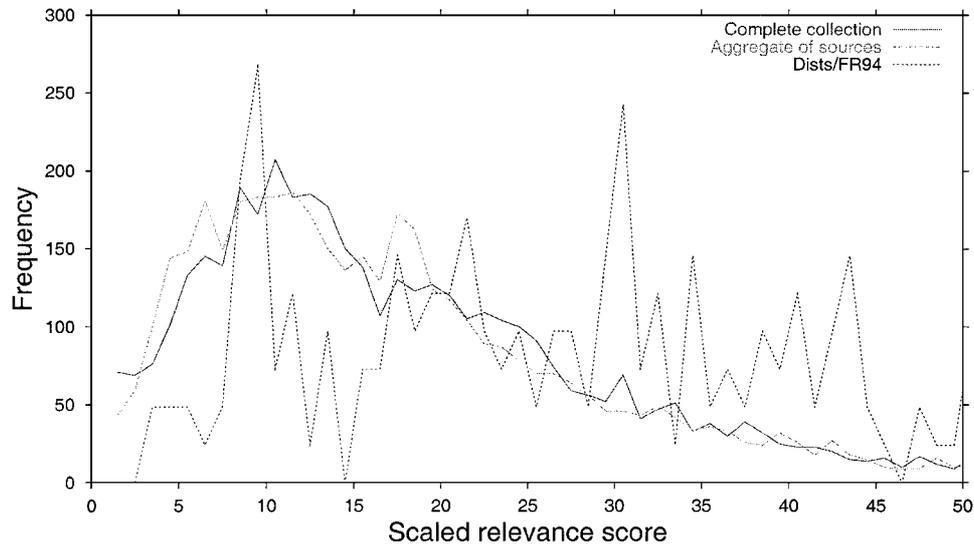


Figure 33. The (Q6A, PADRE(Okapi)) relevant distribution for the Federal Register 94 sub-collection.

5.5. Hypothesis 5

“The performance of retrieval systems relying on *tf.idf* models may be harmed by anomalous *df* values. It is possible that *df* values obtained from a very large collection are more generally useful. From this, we might expect better retrieval effectiveness over the sample collection if *dfs* from the full collection (scaled down, if necessary) were used.”

This hypothesis cannot explain the difference observed by the University of Waterloo, whose retrieval system did not use *df* information.

5.5.1. Analytical. This hypothesis might be regarded as a specific instance of a variation which might force a system from Case 2 (a or b) into Case 3. As discussed in Section 2, we might reasonably expect sample *df* values to be unbiased estimates of population *df* values for the same terms; however, two features might cause them not to work so well. First, *df* values are often very small, so that the sample estimates may be poor ones; second, all *idf* weighting formulae are non-linear in *df*, which destroys the unbiasedness property.

Table 9. Comparative P@20 performance for global and local *df* information.

Collection	Query set	P@20 (local)	P@20 (global)
10% samples of TREC-6	Q6A	0.1440	0.1278
10% samples of TREC-6	Q6M	0.1689	0.1686
Biased sub-collections of TREC-6	Q6A	0.1371	0.1411

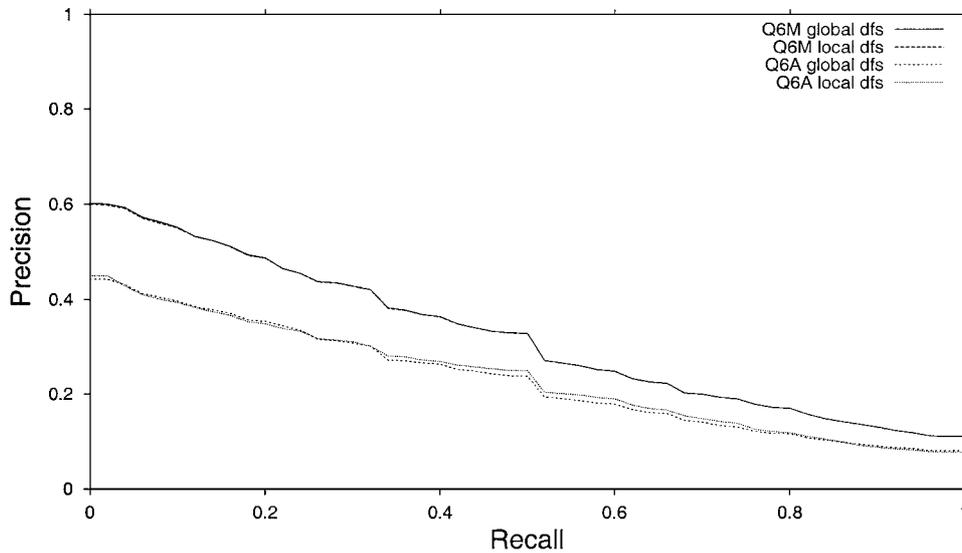


Figure 34. Comparison of precision-recall performance using global vs. local *df* information using 10% uniform samples of the TREC-6 collection and query sets Q6A and Q6M. The two lines for Q6M coincide almost exactly.

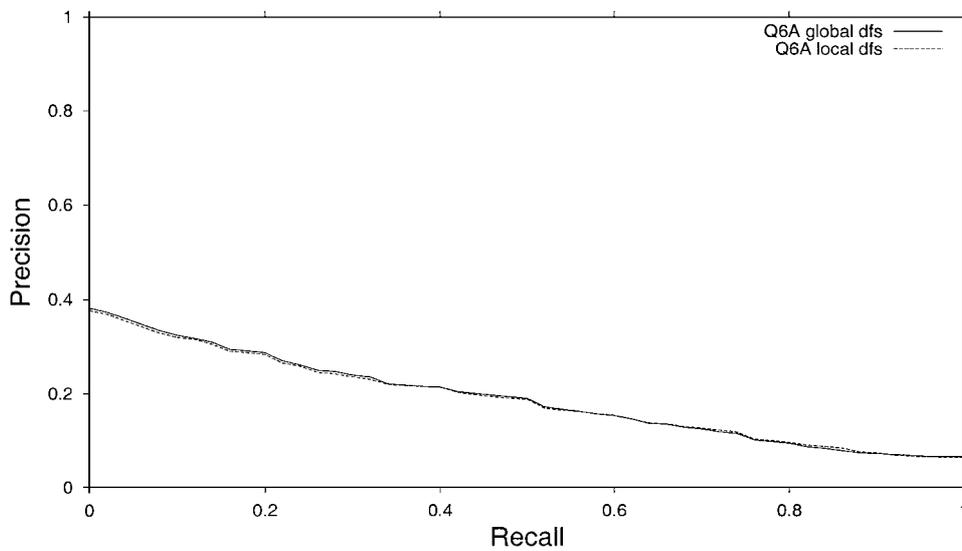


Figure 35. Comparison of precision-recall performance using global vs. local *df* information using biased sub-collections of the TREC-6 collection and query set Q6A.

5.5.2. Empirical. Hypothesis 5 was tested for a *df*-based method by modifying the PADRE system to allow relevance to be scored using *df* values read from a file. Document frequency information for the whole TREC-6 collection was collected and scaled down to match the

target sample size. Runs were then performed using Q6A and Q6M query sets over each of the 10% samples and the averaged precision values were compared with those obtained in corresponding earlier runs using “local” *dfs*. P@20 values are compared in Table 9 and comparative interpolated precision-recall graphs are shown in figure 34. As may be seen, the differences between the two are small and tend to favour the runs using local *dfs*. Inspection of samples of output from the runs showed that the scaled global *dfs* were in fact very similar to the local ones.

Global *dfs* may be more likely to prove their worth in the case of a sub-divided rather than sampled collection. Accordingly, the Q6A queries were run over the five biased sub-collections referred to above (e.g., figure 22). In the event the differences shown in Table 9 and figure 35 are negligible. However, the effect could be larger for certain topics.

It is concluded that Hypothesis 5 does not contribute appreciably to the phenomenon which is the subject of this paper.

6. Discussion and conclusions

The preceding analysis and experiment leads us to claim that signal detection (SD) theory, with suitable allowance for discreteness, provides a good way of understanding the variation in effectiveness of retrieval systems due to collection size. The distribution of scores assigned to relevant documents is the signal distribution and the irrelevant distribution is the noise. Based on the SD model we have outlined four types of relationship between collection size and retrieval effectiveness.

Although relevant and irrelevant score distributions generally deviate markedly from assumptions of normality, we have found that they do overlap and that their means differ substantially. We also found, for a particular combination of query set and retrieval system (\mathcal{Q}, \mathcal{E}), that the relevant and irrelevant score distributions in sample collections are very similar in shape to those of the full collection.

For reasons outlined above, we recommend the use of approximated recall-fallout curves as Operating Characteristic curves for particular (\mathcal{Q}, \mathcal{E}) combinations. For one such combination, we have shown that the OC curves are relatively constant despite considerable change in collection size.

This work was originally initiated in order to test a number of hypotheses proposed to explain an observed increase in P@*n* scores with larger collections. We conclude that the observed increase is due to two main factors.

The first is simply the number of relevant documents in the collection. ($\forall (\mathcal{Q}, \mathcal{E}) P@n \leq R/n$). One of the major explanations of the decrease in P@*n* with decreasing collection size is the reduction in the number of relevant documents (Hypothesis 1). It was shown in Section 2.14.3 that the maximum possible P@20 score for a 10% sample of the TREC-6 data and the TREC-6 ad hoc topics is 0.3858, averaged over all possible samples, whereas the corresponding figure for the full collection is 2.19 times as high.

The second factor is the ability of the \mathcal{Q}, \mathcal{E} combination to rank relevant documents ahead of irrelevant ones. It appears to be the case that P@*n* increases with collection size (for an imperfect retrieval system) even if the measure is not limited by the lack of relevant

documents in the sample. We believe that this behaviour is best modelled in terms of SD theory, which more or less subsumes Hypotheses 2–4.

Section 5.5 shows that Hypothesis 5 (global dfs) does not measurably contribute to the observed phenomenon.

We have found that biased sub-collections exhibit similar behaviour with respect to $P@n$ as do uniform samples. However, $P@n$ decays more sharply with decreasing sub-collection size. It is likely that this is because of an interplay between both of the factors listed above. Relevant documents may be very unevenly distributed across the primary sub-collections. $P@n$ for ill-favoured sub-collections is limited by lack of relevant documents, but additional relevant documents in the more fortunate sub-collections may be “wasted” because of limitations in Q , \mathcal{E} performance.

Our analysis using the SD model predicts that, for a retrieval system exhibiting statistical invariance, average precision and R-precision measures should remain constant regardless of sample size. Our experiments have confirmed this prediction for both measures using the PADRE(Okapi) retrieval system on two sets of queries and a broad range of sample sizes.

Space and time have not permitted the analysis of an extensive range of Q , \mathcal{E} , \mathcal{C} combinations which would confirm or otherwise the general applicability of the model proposed and the use of recall-fallout curves as Operating Characteristic curves for particular retrieval systems.

Acknowledgments

We are grateful to the late Paul Thistlewaite, and to Donna Harman, Nick Craswell and Natasa Milic-Frayling for their suggestions and background information. Chris Buckley and Cornell University kindly made available the TREC evaluation software and gave permission for it to be modified. Thanks are due to NIST in the USA and various copyright holders for access to the TREC and TREC VLC data collections and also to the TREC-6 VLC track participants.

The authors wish to acknowledge that this work was partly carried out within the Co-operative Research Centre for Advanced Computational Systems established under the Australian Government’s Co-operative Research Centres Program.

Notes

1. Text Retrieval Conference.
2. Instructions to the particular retrieval system.
3. A replicated collection may be considered a very special case of inverse sampling because the probability of document relevance remains constant for all topics.
4. A large amount of noise, even purely random, would probably affect the shape of the distribution.
5. Recall maps to true-positive-rate and fallout to false-positive-rate as usually defined.
6. There may be a range of scores t_{B^*} which satisfy the condition, but every value in the range gives the same $B^* = F_R(t_{B^*})$.
7. A small number of bundles was manually removed from the result to more closely achieve the desired size.
8. Modified with permission from Chris Buckley, then of Cornell University.
9. In other words, collection bundles were striped across the samples.

10. The eighty-first is the full collection.
11. The horizontal line is extended from the P@17 point to compensate for the use of 17 rather than 20 gigabytes of data. In actual fact, there is very little difference between the two values.
12. Oral presentation given at the TREC-6 conference. The version here corrects an unfortunate error in the version in Hawking, Thistlewaite and Harman.

References

- Arampatzis A, Beney J, Koster CHA and van der Weide TP (2000) Incrementality, half-life and threshold optimisation for adaptive document filtering. In: Voorhees EM and Harman DK, Eds., Proceedings of TREC-9, Gaithersburg MD. NIST special publication 500-249, trec.nist.gov.
- Arampatzis A and van Hameren A (2001) The score-distributional threshold optimization for adaptive binary classification tasks. In: Proceedings of ACM SIGIR'2001 Conference. ACM Press, New York, pp. 285–293.
- Baumgarten C (1999) A probabilistic solution to the fusion problem in distributed information retrieval. In: Proceedings of ACM SIGIR'99 Conference. ACM Press, New York, pp. 246–253.
- Clarke CLA, Cormack GV and Burkowski FJ (1995) Shortest substring ranking (MultiText experiments for TREC-4). In: Harman DK, Ed., Proceedings of TREC-4. Gaithersburg, MD, pp. 295–304, NIST special publication 500-236.
- Cormack GV, Lhotak O and Palmén CR (1999) Estimating precision by random sampling. In: Hearst M, Gey F and Tong R, Eds., Proceedings of SIGIR'99, Berkeley, CA. ACM Press, New York, Poster, pp. 273–274.
- Hawking D and Thistlewaite P (1996) Relevance weighting using distance between term occurrences. Technical Report TR-CS-96-08, Department of Computer Science, The Australian National University, cs.anu.edu.au/techreports/1996/index.html.
- Hawking D, Thistlewaite P and Craswell N (1997) ANU/ACSys TREC-6 experiments. In: Voorhees E and Harman D, Eds., Proceedings of the TREC-6 Conference. Gaithersburg, MD, pp. 275–290, NIST. trec.nist.gov/pubs/trec6/papers/anu.ps.gz.
- Hawking D, Thistlewaite P and Harman D (1999) Scaling up the TREC collection. *Information Retrieval*, 1(1): 115–137.
- Hays WL (1963) *Statistics*. Holt, Rinehart and Winston, London.
- Lawrence S and Giles CL (1999) Accessibility of information on the web. *Nature*, 400:107–109.
- Manmatha R, Rath T and Feng F (2001) Modeling score distributions for combining the outputs of search engines. In: Proceedings of SIGIR'2001 Conference. ACM Press.
- National Institute of Standards and Technology (1997) TREC home page. trec.nist.gov/.
- Provost F and Fawcett T (2001) Robust classification for imprecise environments. *Machine Learning Journal*, 42(3):203–231. www.stern.nyu.edu/~fprovost/Papers/rocch-mlj.pdf.
- Robertson SE (1969) The parametric description of retrieval tests. Part 1: The basic parameters. *Journal of Documentation*, 25(1):1–27.
- Robertson SE, Walker S, Hancock-Beaulieu MM and Gatford M (1994) Okapi at TREC-3. In: Harman DK, Ed., Proceedings of TREC-3, Gaithersburg MD, November 1994. NIST special publication 500-225.
- Salton G and McGill MJ (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.
- Swets JA (1963) Information retrieval systems. *Science*, 141(3577):245–250.
- Swets JA (1969) Effectiveness of information retrieval methods. *American Documentation*, 20:72–89.
- Voorhees EM and Harman DK (1997), Eds. Proceedings of TREC-6, Gaithersburg MD. NIST special publication 500-240, trec.nist.gov.