# Estimating Precision by Random Sampling

Gordon V. Cormack*, Ondrej Lhotak* and Christopher R. Palmer†

* Department of Computer Science, University of Waterloo
† School of Computer Science, Carnegie Mellon University
cormack@cormack.uwaterloo.ca

## Overview

Precision at $k$ documents retrieved ($P@k$), a common measure of information retrieval performance, is the fraction of relevant documents in the first $k$ returned by an information retrieval system in response to a query. It has been observed that $P@k$ increases with collection size, all other variables being equal. An explanation for this increase derives from the probability ranking principle: information retrieval systems score documents by likelihood of relevance and return the documents in decreasing order of score. In a larger collection there are more high-scoring documents and therefore the average score of the first $k$ will be higher, resulting in higher $P@k$.

We derive and experimentally validate the following equation that quantifies the relationship between $P@k$ and collection size:

$$P_{n_1}@k_1 = P_{n_2}@k_2 \text{ where } \frac{k_1+1}{n_1} = \frac{k_2+1}{n_2}$$

The notation $P_{n_i}@k_i$ indicates $P@k_i$ with respect to a collection of size $n_i$ selected at random from some population of documents.

A significant application of the quantification above is in creating a large archival test collection; that is, a set of documents, a set of queries, and an automatic method of measuring the effectiveness of information retrieval strategies. Current practice, as used for the TREC ad hoc collections, requires manual relevance assessments to identify nearly all the relevant documents in the collection (Voorhees & Harman 1997). In TREC, the pooling method is used to avoid assessing many (mostly non-relevant) documents. Cormack et al (1998) propose a method to reduce further the number of documents assessed. Nevertheless, the effort of manual assessment is formidable for collections of this size (about 500,000 documents); and the effort is proportional to collection size, rendering it prohibitive for much larger collections, like TREC's Very Large Corpus (VLC) with 20,000,000 documents, or the Web with at least an order of magnitude more.

For VLC, it was deemed infeasible to identify most of the relevant documents in the collection. Instead, ad hoc assessments were made only on the first 20 documents from each participating system, and $P@20$ was computed. This method is adequate for comparison among the participating systems, but has limited applicability to new systems − further manual assessments would be required to evaluate any new system.

We argue that arbitrarily large test collections may be constructed as follows: First, identify the set of documents and queries for which the retrieval systems are to be tested. Second, identify a random subset of the documents such that it is feasible to find a near-complete set of relevance judgements using the methods cited above. This subset may be selected a priori but should be unknown to the systems being tested. Third, have the systems retrieve documents in decreasing order of likelihood. Fourth, ignore those documents not in the subset to be judged, and use the standard evaluation measures and the equation above (with linear interpolation as necessary) to estimate $P@k$ for the full set. The expected value of the estimate appears to be very accurate, even for tiny samples of a few thousand documents. The confidence interval for a sample-based estimate decreases as the size of the sample increases − it is not obvious how small the this interval must be in order to be masked by other errors inherent in evaluation, but it appears that a sample of 125,000 documents would be a reasonable sample size for estimating performance on collections the size of TREC's ad hoc collections or larger.

## Derivation

We assume that all documents are drawn randomly from an infinite population $D = \{d_i\}$ .

Each document $d_i$ has a score $S(d_i)$ that increases with likelihood of relevance. For simplicity we assume that $S(d_i) = S(d_j)$ only if $d_i = d_j$.

For each document $d_i$, $R(d_i) = 1$ if $d_i$ is relevant; otherwise $R(d_i) = 0$. While $R$ in the abstract is a total function, we wish to minimize the number of values for which we actually evaluate $R$.

$r$, a collection-size-independent version of ranking is defined as $r(d_i) = Prob(S(d) \geq S(d_i))$; that is, the probability that $d$ randomly selected from $D$ has a score not less than $d_i$.

$P@\rho$, a collection-size-independent version of $P@k$, is defined as

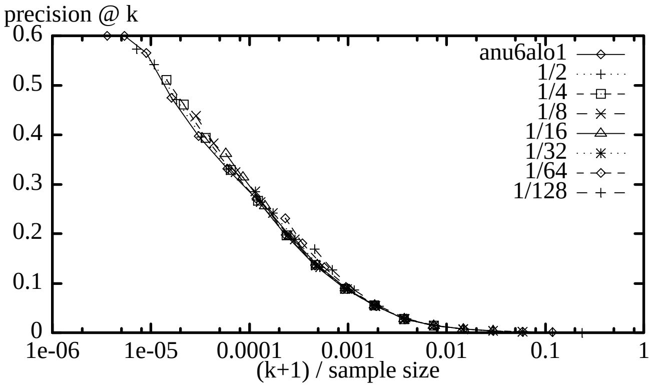
$$\begin{aligned} P@\rho &= Prob(R(d) = 1 \mid r(d) \leq \rho) \\ &= \frac{1}{\rho} \int_{x=0}^{\rho} Prob(R(d) = 1)\, Prob(r(d) = x)\; dx \end{aligned}$$

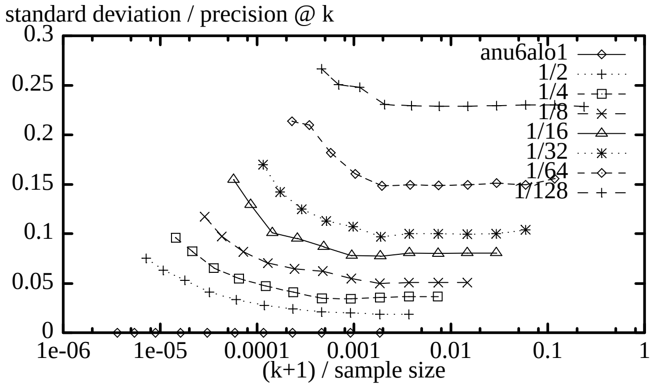
That is, $P@\rho$ is the conditional probability that a document is relevant, given that its rank is less than or equal to $\rho$.

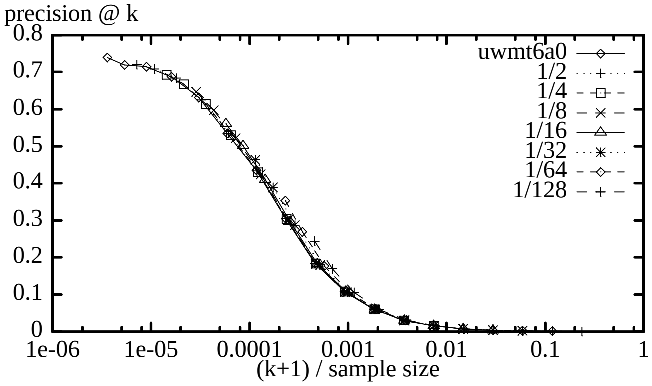
Now consider $d_i$, the $k$th-ranked document from a sample of size $n$. The expected value of $r(d_i)$ is $k/n$, and therefore $R(d_i)$ approximates the probability density function at $\rho = k/n$. This value is used to estimate (using the rectangle rule) an interval in the probability distribution centred at $k/n$; that is,

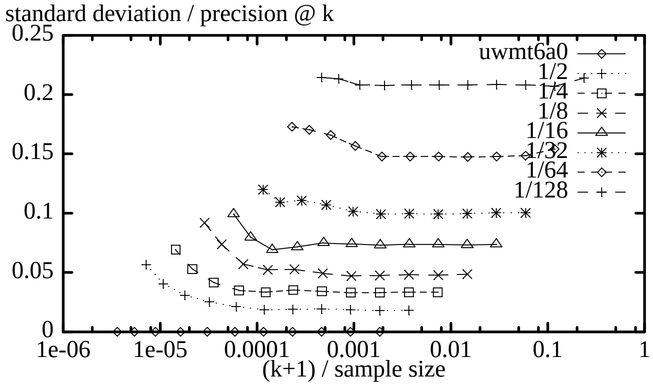
$$\frac{1}{\rho} \int_{x=k/n-\Delta}^{k/n+\Delta} Prob(R(d) = 1)\; Prob(r(d) = x)\; dx\ .$$

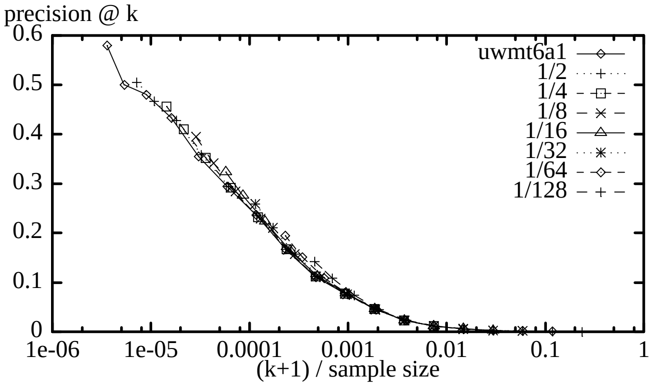
Consider, for example, the case of $k = 1$. $P_n@1$ approximates $\int_{x=0}^{2/n} Prob(R(d) = 1)\; Prob(r(d) = x)\; dx$ or $P@\rho$ where $\rho = 2/n$. $P@k$ is, in general, the sum of $k$ such values which provide a piecewise approximation to $P@\rho$ where $\rho = (k+1)/n$. Equating the approximations for different $n_i$ and $k_i$ gives the formula given in the overview.

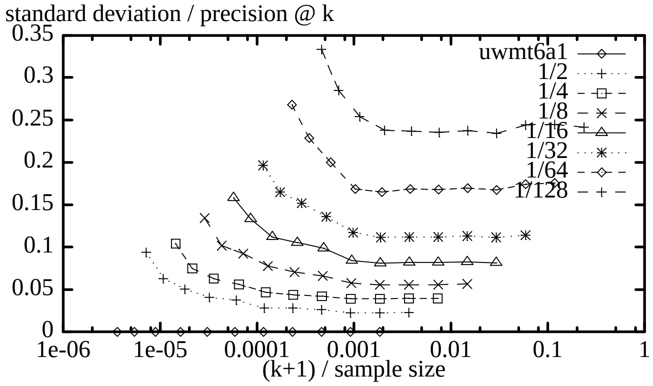












**Experimental Results**

The accuracy of our derivation was tested using samples from the TREC 6 ad hoc collection with eight sizes ranging from 1/128 (about 4200 documents) to the full collection (about 540,000 documents). We randomly selected 200 subsets of each size, and computed the mean and standard deviation of the predictions over these 200 subsets. The first three figures show $P@k$ as a function of $(k+1)/n$ for three participating runs: anu6alo1 by Australian National University − the best automatic run; uwmt6a0 by the Univesity of Waterloo − the best manual run; uwmt6a1 by the University of Waterloo − also an automatic run. The last three figures show the standard deviation of these values over the 200 samples, as a fraction of the $P@k$ value.

We see that the curves for the various sample sizes are very similar, with a small systematic increase for the smallest $P@k$ values (corresponding to estimates based on the value of $P@1$ alone). The standard deviation, as one expects, decreases as sample size increases until the pathological case of the full collection, which has a standard deviation of 0. In addition, the standard deviation increases in predicting $P@k$ for smaller $k$.

From these results we extrapolate that it should be possible, by rendering assessments on a subset comparable in size to the TREC ad hoc collections, to estimate precision on arbitrarily large collections. There is, however, a lower limit to the value of $k$ for which $P@k$ may be estimated by sampling. From a sample of size $n_1$ the measurement of $P_{n_1}@1$ predicts in a larger database of size $n_2$ the value of $P_{n_2}@k_2$ where $k_2 = \frac{2n_2}{n_1} - 1$. It is not possible to predict $P@k_2$ for smaller values of $k_2$; there is insufficient information in the sample to do so. One might attempt to extrapolate these values from the values of $P@k$ for higher $k$ or one might consider a hybrid approach in which these values are measured by assessing the first $k$ documents of each run while the values for larger $k$ are estimated by sampling.

**References**

Voorhees E.M. and Harman D.K (eds), *The Sixth Text REtrieval Conference (TREC-6)*, NIST Publication 500-200, Gaithersburg, MD, November 1997.

Cormack G.V., Palmer C.R. and Clarke C.L.A, *Efficient Construction of Large Test Collections*, Proceedings of SIGIR 98, Melbourne, August 1998.