

Detailed description of the statistical test

In order to compare results of different strategies statistically, we performed the bootstrap shift precision test (Wilbur 1994). The bootstrap approach was originated by Efron (Efron 1979) and is discussed by Noreen (Noreen 1989). Suppose that we have two methods of ranked retrieval, M_1 and M_2 , and we have performed retrieval by each on a test database D consisting of a set of queries, $Q = \{q_i\}_{i=1}^m$. For any retrieval method M and query q , we let $prec(M, q)$ denote the precision of the retrieval produced by M in answer to the query q . Here the null hypothesis, H_0 , is that M_1 and M_2 have equal performance, in the sense that they have the same expected precision on a random query. If we assumed that the precisions produced by each retrieval method on the individual (randomly sampled) queries were samples from normal distributions, then a paired t -test would be an appropriate parametric test of the null hypothesis. However, we do not make any assumptions about the underlying distributions. Let us set:

$$X_m = \frac{1}{m} \sum_{i=1}^m prec(M_1, q_i) - prec(M_2, q_i) \quad (1.1)$$

Then, X_m may be viewed as one element of a population Π_m of sample means of samples of size m . Further, our null hypothesis states its mean is zero. We now make the bootstrap assumption, namely, we assume that the sample of precision differences which we have,

$\{prec(M_1, q_i) - prec(M_2, q_i)\}_{i=1}^m$, adequately represents the whole population of such differences.

From this list of values, random samples of size m are made by random selection and replacement. This process is repeated a large number of times (in our case 1000) and each time the sample average is computed as in (1.1). The resulting set of sample averages forms a distribution, $\{Y_{km}\}_{k=1}^{1000}$. Let the mean of this distribution be denoted by R . The theoretical aspects of the shift method we use here are described in more detail in (Noreen 1989). The plan is to let the distribution $\{Y_{km}\}_{k=1}^{1000}$ substitute for Π_m . This cannot be done directly, however, because Π_m is assumed to have mean zero, while R may be expected to be very close to X_m which is in general non-zero. In order to correct for this problem, we shift the whole distribution of sample means that we have constructed to the new distribution $\{Y_{km} - R\}_{k=1}^{1000}$ with mean zero. The distribution Π_m is now represented by $\{Y_{km} - R\}_{k=1}^{1000}$.

We determine the critical values that define tails of size $\frac{\alpha}{2}$ on either side of this shifted distribution and reject the null hypothesis if X_m lies outside of these critical values.

- Efron, B. (1979). "Computers and the theory of statistics: thinking the unthinkable." SIAM Review **21**(4): 460-480.
- Noreen, E. W. (1989). Computer Intensive Methods for Testing Hypotheses. New York, John Wiley & Sons.
- Wilbur, W. J. (1994). "Nonparametric significance tests of retrieval performance comparisons." Journal of Information Science **20**(4): 270-284.

Table 1: Statistical test results on the TREC 2006 data

Measure	Method 1	Method 2	Performance Difference	Confidence Interval at the 0.05 level
Mean Average Precision	tf_idf	relemed	0.00919	-0.0171139~0.0185531
	tf_idf	reverse_time	0.032381	-0.0283221~0.0324399
	releme	reverse_time	0.023191	-0.0181612~0.0183158
Precisions at Top 20 Rank	tf_idf	relemed	0.038095	-0.0424882~0.0456078
	tf_idf	reverse_time	0.080952	-0.0498218~0.0549402
	relemed	reverse_time	0.042857	-0.0287616~0.0307614
Precisions at Top 10 Rank	tf_idf	relemed	0.042857	0.0475763~0.0524237
	tf_idf	reverse_time	0.07619	-0.062034~0.0713
	relemed	reverse_time	0.033333	0.0287427~0.0284003
Precisions at Top 5 Rank	tf_idf	relemed	0.047619	-0.0479905~0.0567715
	tf_idf	reverse_time	0.066667	-0.106076~0.112971
	relemed	reverse_time	0.019048	-0.0961805~0.0752475

Table 2: Statistical test results on the TREC 2007 data

Measure	Method 1	Method 2	Performance Difference	Confidence Interval at the 0.05 level
Mean Average Precision	tf_idf	relemed	0.019294	-0.0140324~0.0179976
	tf_idf	reverse_time	0.056177	-0.0233357~0.0251353
	relemed	reverse_time	0.036883	-0.017098~0.0199026
Mean Rank Precisions at Top 20	tf_idf	relemed	0.027942	-0.0265599~0.0278511
	tf_idf	reverse_time	0.089706	-0.0409353~0.0472997
	relemed	reverse_time	0.061764	-0.0393754~0.0444486
Mean Rank Precisions at Top 10	tf_idf	relemed	0	-0.0383978~0.0380742
	tf_idf	reverse_time	0.114706	-0.0533242~0.0554998
	relemed	reverse_time	0.114706	-0.0590445~0.0703675
Mean Rank Precisions at Top 5	tf_idf	relemed	-0.017647	-0.0598651~0.0636649
	tf_idf	reverse_time	0.152941	-0.064265~0.07103
	relemed	reverse_time	0.170588	-0.0808709~0.0838351