Power and Bias of Subset Pooling Strategies

Gordon V. Cormack and Thomas R. Lynam
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
gvcormac@uwaterloo.ca, trlynam@uwaterloo.ca

ABSTRACT

We define a method to estimate the random and systematic errors resulting from incomplete relevance assessments. Mean Average Precision (MAP) computed over a large number of topics with a shallow assessment pool substantially outperforms – for the same adjudication effort – MAP computed over fewer topics with deeper pools, and P@k computed with pools of the same depth. Move-to-front pooling, previously reported to yield substantially better rank correlation, yields similar power, and lower bias, compared to fixed-depth pooling.

Categories and Subject Descriptors

H.3.4 [**Information Search and Retrieval**]: Systems and Software – performance evaluation

General Terms

Experimentation, Measurement

Keywords

significance test, validity, statistical power,pooling methods

1. INTRODUCTION

A number of stategies have been devised to minimize the amount of human adjudicaton involved in IR system evaluation. The TREC pooling method selects for adjudication only the top-ranked k documents from each system under test. The typical value of $k = 100$ appears to work well for test collections with 50 topics and 500,000 documents. However, the effort in conducting an evaluation is substantial for a collection of this size and prohibitive for larger collections. It is not obvious whether a smaller value of k, or some other subset of the pool, might suffice; and it is not obvious that even $k = 100$ is sufficient for larger collections.

Validation of the pooling method, and optimizations of it, have typically been ad hoc and uncalibrated. The commonest approach has been to assume as a gold standard the mean average precision (MAP) for $k = 100$ and to measure the correlation in system rankings (Kendall τ) achieved by the gold standard and the proposed method. Without formal justification, $\tau > .9$ has been taken to be *good agreement*. Furthermore, the qualitative term *bias* has been ascribed to some methods in assessing their validity.

We present a method to estimate the power and bias of pooling methods, and use our method to evaluate the effectiveness of several pooling alternatives as a function of adjudication effort. The alternatives we investigate are: different values of k; different numbers of topics; move-to-front sampling; using precision at cutoff k ($P@k$) as an alternative to MAP.

2. METHOD

Kendall τ simply counts inversions in rank; as such it conflates *random error* – error due to chance – and *systematic error (or bias)* – error due to measuring the wrong quantity. More specifically, it measures errors in the sign of the difference between the MAP scores of pairs of systems. It does not account for the magnitude or significance of the difference. We treat random error and bias separately, using a paired t-test[1] to estimate statistical power, and counting the number of *significant* inversions between the alternative method and the gold standard. Overall, an alternative pooling method is good if it has high power, and if its observed bias is insubstantial relative to random error.

We applied this method to various subsets of the topics and judging pool from the TREC 2004 Robust Retrieval Track [7]. In all cases we computed for all pairs of systems the sign of the difference in MAP (or $P@k$) and also the t-test p-value. We compared the sign of the difference to that yielded by the gold standard, and computed the number of inversions when $p < \alpha$.[2] We compute power as the overall proportion of differences for which $p < \alpha$, and bias as the proportion of differences with $p < \alpha$ whose difference has the opposite sign from the gold standard. If this proportion is substantially less than α, bias is a negligible factor (compared to random error) in the validity of the estimate, and may be discounted.

[1]Although the applicability of the assumptions have been called into question [6], we found the t-test to be very accurate – for all methods and sample sizes presented here – in predicting inversions in rank using the same method on different topics, therefore establishing its validity.

[2]For $\alpha = .05$ and several values not reported here.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

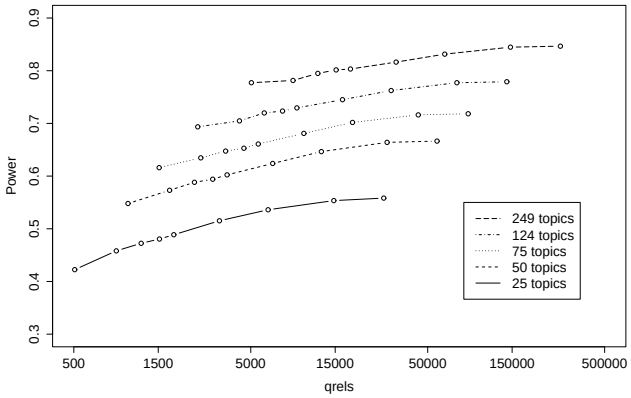

Figure 1: Power vs effort (depth k pool)

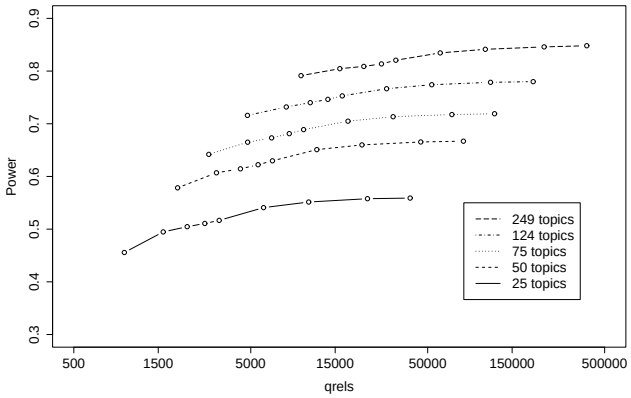

Figure 2: Power vs effort (move-to-front)

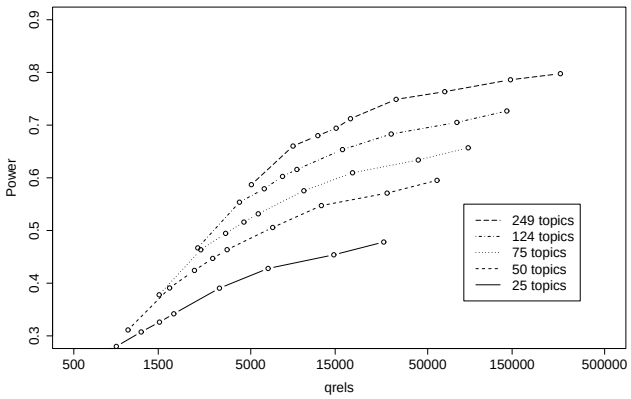

Figure 3: Power vs effort (P@k)

3. RESULTS

Figure 1 shows the effect of varying k (pool depth) and n (number of topics) on adjudication effort and power ($\alpha = .05$) for the standard TREC pooling method. The y-axis is power and the x-axis is the number of relevance assessments necessary to achieve that power for a given n. Each point represents a different value of k. Figure 2 shows that move-to-front pooling [4] yields insubstantially different results. Figure 3, on the other hand, shows that substituting $P@k$ for MAP yields inferior power.

Figure 4 shows the observed bias for each of the methods, as a function of judging effort. We observe that move-to-front exhibits substantially less bias than methods commonly taken to be more *fair*.

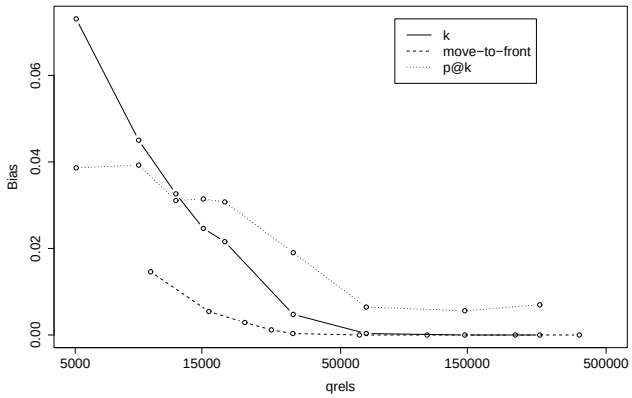

Figure 4: Bias vs effort (249 topics)

Our results support the suggestion that an experimental design using more topics and fewer judgements is more efficient [5], but not the assertion that more regimented selection techniques yield lower bias. We advance power and bias analysis as a methodology to supplant rank correlation in assessing new pooling strategies and evaluation measures (e.g. [2, 3, 1]).

4. REFERENCES

[1] Aslam, J. A., Pavlu, V., and Yilmaz, E. A statistical method for system evaluation using incomplete judgments. In *SIGIR '06* (2006), pp. 541–548.

[2] Buckley, C., and Voorhees, E. M. Retrieval evaluation with incomplete information. In *SIGIR '04* (2004), pp. 25–32.

[3] Carterette, B., Allan, J., and Sitaraman, R. Minimal test collections for retrieval evaluation. In *SIGIR '06* (2006), pp. 268–275.

[4] Cormack, G. V., Palmer, C. R., and Clarke, C. L. A. Efficient construction of large test collections. In *SIGIR Conference 1998* (Melbourne, Australia, 1998).

[5] Sanderson, M., and Zobel, J. Information retrieval evaluation: Effort, sensitivity, and reliability. In *SIGIR Conference 2005* (Salvador, Brazil, 2005).

[6] Van Rijsbergen, C. J. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.

[7] Voorhees, E. M. Overview of the TREC-2004 robust track. In *13th Text REtrieval Conference* (Gaithersburg, MD, 2004).