Metasearch Consistency*

Mark Montague
Department of Computer Science
Dartmouth College
6211 Sudikoff Laboratory
Hanover, NH 03755
montague@cs.dartmouth.edu

Javed A. Aslam
Department of Computer Science
Dartmouth College
6211 Sudikoff Laboratory
Hanover, NH 03755
jaa@cs.dartmouth.edu

ABSTRACT

We investigate the performance of metasearch algorithms in terms of how much they improve *consistency*. We find that three different metasearch algorithms, each over three datasets, usually improve the consistency of search results; sometimes the improvement is dramatic. Furthermore, consistency tends to improve when performance improves.

1. INTRODUCTION

Metasearch, or *ranked list fusion*, is the process of merging the ranked list results of multiple search systems into one ranked list. By doing so, it is possible to significantly outperform the best input search system.

We contend that metasearch can improve upon not only the raw performance of its input retrieval systems (e.g., as measured by *average precision*), but it can also ensure that good results are obtained for *more queries. Metasearch can improve consistency of performance over the queries.* One might expect this to be the case since combining the information gleaned from a number of different sources, each with different strengths, would seem to smooth out performance wrinkles. Our goal is to verify and quantify this effect.

Many researchers have studied the metasearch problem. Most notably, Fox and Shaw [4] developed the simple and effective CombMNZ algorithm (also studied by Lee [5]), and Bartell [2] and Vogt [6] optimize linear combinations of relevance scores. See Croft [3] for a review of the literature. As far as we know, no one has studied specifically the improvment in performance consistency afforded by metasearch.

2. EXPERIMENTS

We present results for three metasearch algorithms, each based on very different models. The fact that they acheive similar improvements in consistency suggests that our findings apply to a large class of metasearch algorithms. For each document, CombMNZ [4] defines the fused relevance score as the sum of the relevance scores given by each system, multiplied by the number of systems that returned the document. Borda-fuse [1] applies the Borda Count optimal voting algorithm to metasearch. Bayes-fuse [1] estimates the log odds of relevance of each document given its rank in each input list, based on simple training data.

We use four data sets: TREC 3, TREC 5, Vogt, and TREC 9, taken from three different adhoc tracks of TREC competitions. They respectively contain 40, 61, 10, and 105 input retrieval systems, with respectively 50, 50, 10, and 50 queries each. The Vogt data set is a subset of the TREC 5 data set, chosen by Vogt [7] to contain retrieval systems that are very different from each other, and queries with a large number of relevant documents.

We perform the following experiment: Randomly select a set of n (for $n \in \{2, 4, \ldots, 12\}$) input systems, fuse them, and report the resulting list's: (1) average precision, (2) standard deviation of average precision over the queries, and (3) *percent standard deviation:* standard deviation as a fraction of average precision. Repeat this experiment 200 times for each n (or the maximum number of times combinatorially possible), and average the results. See Figure 1.

The percent standard deviation measure (known in statistics as the "coefficient of variation") is motivated by the observation that the standard deviation of performance tends to be higher for systems of higher performance. We can more directly compare the variability of two systems by normalizing their performances; e.g., scaling their means to one. This scales their standard deviations appropriately.

3. CONCLUSIONS

The second column of graphs in the figure shows that metasearch consistency is usually between the best input system's consistency and the average system's consistency. The third column shows that after accounting for raw performance by normalizing, fused results are generally more consistent than any of their inputs. Furthermore, as more systems are included in the fusion, consistency generally improves. Finally, when metasearch is able to improve performance over the best input system (see column 1), it is also able to improve consistency (see column 3): TREC 9 and TREC 5 show little or no improvement in both, TREC 3 shows good improvement in both, and the Vogt data set shows large improvement in both.

*This work generously supported by NSF grants EIA-9802068, BCS-9978116, and Career Award CCR 00-93131.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA..

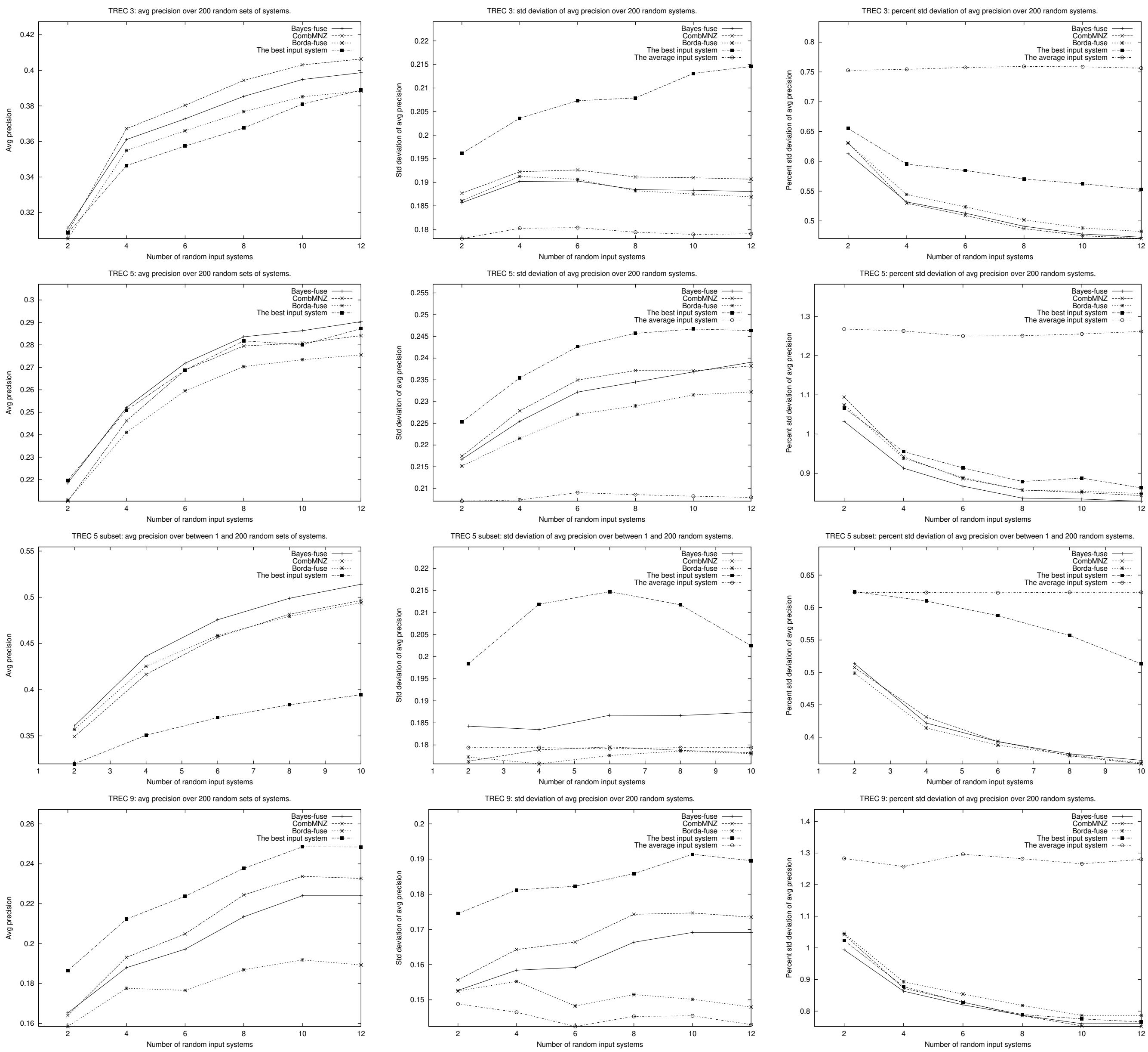

Figure 1: Each row of graphs is over one data set: TREC 3, TREC 5, and Vogt. Each column shows one statistic: raw performance, standard deviation, and percent standard deviation. The horizontal axis of each graph is the number of systems being fused; the vertical axis is the statistic of interest.

4. REFERENCES

[1] J. Aslam and M. Montague. Models for metasearch. In *ACM SIGIR '01*, New Orleans, Louisiana, USA, Sept. 2001. ACM Press, New York.

[2] B. T. Bartell. *Optimizing Ranking Functions: A Connectionist Approach to Adaptive Information Retrieval*. PhD thesis, University of California, San Diego, 1994.

[3] W. B. Croft. Combining approaches to information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, chapter 1. Kluwer Academic Publishers, 2000.

[4] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *TREC-2*, pages 243–249, Gaithersburg, MD, USA, Mar. 1994. U.S. Government Printing Office, Washington D.C.

[5] J. H. Lee. Analyses of multiple evidence combination. In *ACM SIGIR '97*, pages 267–275, Philadelphia, Pennsylvania, USA, July 1997. ACM Press, New York.

[6] C. C. Vogt. *Adaptive Combination of Evidence for Information Retrieval*. PhD thesis, University of California, San Diego, 1999.

[7] C. C. Vogt. How much more is better? Characterizing the effects of adding more IR systems to a combination. In *RIAO*, pages 457–475, Paris, France, Apr. 2000.