On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments [*]

Javed A. Aslam
Department of Computer Science
Dartmouth College
jaa@cs.dartmouth.edu

Robert Savell
Department of Computer Science
Dartmouth College
rsavell@cs.dartmouth.edu

ABSTRACT

Soboroff, Nicholas and Cahan recently proposed a method for evaluating the performance of retrieval systems without relevance judgments. They demonstrated that the system evaluations produced by their methodology are correlated with actual evaluations using relevance judgments in the TREC competition. In this work, we propose an explanation for this phenomenon. We devise a simple measure for quantifying the *similarity* of retrieval systems by assessing the similarity of their retrieved results. Then, given a collection of retrieval systems and their retrieved results, we use this measure to assess the *average similarity* of a system to the other systems in the collection. We demonstrate that evaluating retrieval systems according to average similarity yields results quite similar to the methodology proposed by Soboroff et al., and we further demonstrate that these two techniques are in fact highly correlated. Thus, the techniques are effectively evaluating and ranking retrieval systems by "popularity" as opposed to "performance."

Categories and Subject Descriptors:

H.3.3 [**Information Search and Retrieval**]: Retrieval Models.

General Terms: Theory, Experimentation.

Keywords: Ranking Retrieval Systems.

1. INTRODUCTION

Soboroff, Nicholas and Cahan [2] recently proposed a method for evaluating retrieval systems in the absence of relevance judgments by (1) constructing a *pool* from the "top" documents retrieved by the underlying retrieval systems, (2) assigning relevance judgments at random within this pool, and (3) assessing the performance of the underlying systems using these artificially generated relevance judgments. A number of variants on this method for generating relevance judgments were investigated, including (1) constructing pools at varying depths, (2) viewing the pool of documents as sets versus multisets, and (3) assigning relevance judgments in varying ways within the pool. The resulting system assessments and rankings were well correlated with actual TREC rankings in their experiments.

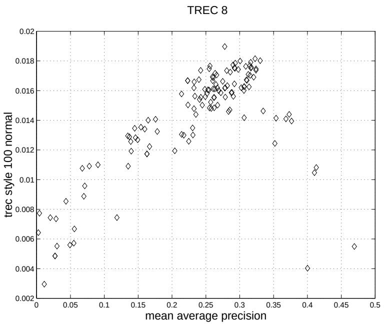

Figure 1: Scatter plot of Soboroff et al.'s "TREC-style100normal" method vs. actual TREC assessments for TREC8. Each point is a system; its x-coordinate is its TREC mean average precision, and its y-coordinate is its assessment by the Soboroff et al. method.

As noted by Soboroff et al., these "pseudo-evaluations" suffered from a common phenomenon. While the bulk of the systems were classified correctly, the best (and therefore most important) systems were consistently ranked with the poor performers (see Figure 1). Our hypothesis for this phenomenon is that these blind methods are suffering from a "tyranny of the masses" effect: the "better" systems are doing something significantly "different" from the more generic systems in the competition and, in the absence of actual relevance judgments, are being punished for this fact. Thus, we believe that the systems are, in fact, being evaluated in terms of "popularity" instead of "performance."

In order to verify this hypothesis, we devise a simple measure for the similarity of two retrieval systems based on the similarity of their retrieved results, and we define the "popularity" of a system by its average similarity to the other systems in the collection. In experiments with TREC data, we demonstrate that assessing "system performance" by popularity yields results strikingly similar to those of the methods of Soboroff et al. Furthermore, we demonstrate that the Soboroff technique and the evaluation of systems by "popularity" are in fact highly correlated. Thus, these techniques are effectively evaluating and ranking retrieval systems by "popularity" as opposed to "performance."

2. RESULTS

Figure 1 shows the results of the Soboroff et al. method in

[*] This work partially supported by NSF Career Grant CCR-0093131 and NSF Grant 5-36955.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.

	System Similarity		*TRECstyle100norm*	
TREC	*K-τ*	*Corr*	*K-τ*	*Corr*
3	0.400000	0.706937	0.407692	0.679134
5	0.578012	0.781613	0.517994	0.683346
6	0.467705	0.623387	0.418506	0.590617
7	0.398286	0.474406	0.392194	0.454289
8	0.464651	0.650457	0.484187	0.647479

Table 1: System Similarity and Soboroff et al.'s "TRECstyle100normal" vs. actual TREC assessments for TRECs 3, 5, 6, 7 and 8.

TREC	*K-τ*	*Correlation*
3	0.858974	0.980627
5	0.830848	0.949764
6	0.745129	0.916408
7	0.840251	0.966771
8	0.815631	0.961270

Table 2: System Similarity vs. Soboroff et al.'s "TRECstyle100normal" for TRECs 3, 5, 6, 7 and 8.

assessing the performance of a collection of retrieval systems in the absence of relevance judgments. While the inferred system assessments are correlated with actual TREC assessments for the vast majority of systems, the methodology fails to properly identify the best, and thus most important, systems. Our hypothesis is that this is because the methodology is actually evaluating retrieval systems by "popularity" as opposed to "performance."

To isolate this phenomenon, we define a simple measure of system similarity in terms of common returned documents. We use the grossest possible measure. With Ret_i indicating the set of documents returned by system i, the system-to-system similarity score is defined as

$$SysSimilarity(Sys_1, Sys_2) = \frac{|Ret_1 \cap Ret_2|}{|Ret_1 \cup Ret_2|}$$

and the average system similarity score is given by

$$AvgSysSim(S_0) = \frac{1}{n-1} \sum_{(S \neq S_0)} SysSimilarity(S, S_0) \quad (1)$$

where n is the number of systems.

Assuming this quantity to be a measure of system performance, we evaluate the systems accordingly and compare these assessments to those of "TRECstyle100normal" in [2].

Table 1 compares both of these evaluation methods to actual TREC assessments for TRECs 3, 5, 6, 7 and 8. We calculate both the linear correlation coefficient of the actual assessments (method assessment vs. TREC mean average precision) as well as the Kendall's τ measure of the similarity of the system rankings produced by these assessments and those of TREC. Note the striking correspondence between the quality of the methods in terms of both the Kendall's τ and correlation measures.

Table 2 provides a direct comparison between the evaluation methods. Note the near perfect correlation between the assessments produced by the two methods.

Finally, Figure 2 shows scatter plots of the two methods for TRECs 7 and 8. These plots clearly demonstrate that the methods are highly correlated in the aberrant regions associated with the best systems, with increased deviations in the more random regions of the poorly ranked systems.

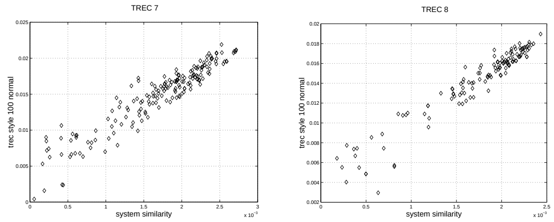

Figure 2: Soboroff et al.'s "TRECstyle100normal" method vs. System Similarity for TRECs 7 and 8. Each point is a system; its x-coordinate is its system similarity assessment, and its y-coordinate is its assessment by the Soboroff et al. method.

(The scatter plots for TRECs 3, 5 and 6 are similar but omitted for space considerations.)

3. CONCLUSIONS

In this paper, we have limited our presentation to the examination of the "TRECstyle100normal" method, but the inability of this method to properly assess the best systems in a collection is a failing common to all "blind" evaluation methods we have encountered. As indicated by the strong correlation of these methods with a simple inter-system similarity measure, we conclude that these methods, in the absence of feedback, are effectively assessing the underlying systems in terms of "popularity" as opposed to "performance." Thus, the positive correlation of these methods with actual system performance assessments is largely due to the justifiable "good faith" assumption that the behavior of the underlying systems is positively correlated with ideal retrieval. Finally, given the aberrant behavior of these methods on the better retrieval systems, it seems appropriate to declare the system similarity scores to be a measure of the aggregate bias which must be overcome to achieve a valid evaluation of system performance. Systematic methods for the minimization of required relevance judgments such as those found in [1] are most useful in correcting for these biases, and we are currently pursuing the application of machine learning techniques to the problem of constructing efficient pools for overcoming this bias.

4. ACKNOWLEDGMENTS

The authors would like to thank Ian Soboroff for providing the data used for the comparisons given above.

5. REFERENCES

[1] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[2] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.