

Early Precision Measures: Implications from the Downside of Blind Feedback

Stephen Tomlinson
Hummingbird
Ottawa, Canada

stephen.tomlinson@hummingbird.com

ABSTRACT

We report the statistically significant mean impacts of blind feedback, as implemented by 7 participants for the 2003 Reliable Information Access (RIA) Workshop, on 30 retrieval measures, including several primary recall measures not originally reported. We find that blind feedback was detrimental to measures focused on the first relevant item even when it boosted “early precision” measures such as mean Precision@10, implying that the conventional reporting of ad hoc precision needs enhancement.

Categories and Subject Descriptors: H.3.4 [Information Storage and Retrieval]: Systems and Software—*performance evaluation*

General Terms: Measurement, Experimentation

Keywords: Early Precision, Blind Feedback, Robust Retrieval

1. INTRODUCTION

Blind feedback (BF), also known as pseudo-relevance feedback or (more ambiguously) query expansion, is a controversial retrieval technique in which terms from the top retrieved items of the original base query are used to expand the query for re-running on the same collection. BF is commonly used by the participants of the ad hoc¹ retrieval tasks of TREC², CLEF and NTCIR, but typically it is not used for the known-item³ tasks of these forums, and the technique’s robustness has been criticized (see Section 5).

Based on our own recent BF experiments in 8 languages [2, 3, 4], we formed the hypothesis that BF impaired measures focused on “primary recall” (retrieval of the first relevant item for a topic) such as mean reciprocal rank (MRR) and Success@10 (S10), while boosting measures with most of their weight on “secondary recall” (retrieval of additional relevant items) such as mean average precision (MAP) and Precision@10 (P10). In this paper, we test this hypothesis on 7 other groups’ BF systems.

2. TESTING THE HYPOTHESIS

Table 1 shows the impact of BF as implemented by 7 other groups for the 2003 RIA Workshop [1]. “Base” gives

¹“Ad hoc” searches seek relevant items for a topic.

² Text REtrieval Conference (<http://trec.nist.gov/>).

³“Known-item” searches seek one particular item.

Table 1: Mean Impact of Blind Feedback on 7 RIA Systems (4 Retrieval Measures)

	MAP	P10	MRR	GS10
albany Base	0.126	0.281	0.463	0.696
BF	0.154	0.300	0.446	0.660
Diff	0.028	0.019	-0.017	-0.036
city Base	0.186	0.393	0.689	0.839
BF	0.216	0.412	0.621	0.797
Diff	0.030	0.019	-0.068	-0.042
clarit Base	0.185	0.373	0.628	0.812
BF	0.210	0.397	0.654	0.798
Diff	0.026	0.024	0.026	-0.014
cmu Base	0.201	0.385	0.640	0.840
BF	0.225	0.391	0.582	0.816
Diff	0.024	0.007	-0.058	-0.024
sabir Base	0.204	0.389	0.611	0.780
BF	0.226	0.393	0.598	0.752
Diff	0.022	0.005	-0.013	-0.028
umass Base	0.203	0.383	0.623	0.829
BF	0.235	0.413	0.608	0.775
Diff	0.032	0.029	-0.015	-0.054
waterloo Base	0.198	0.406	0.662	0.828
BF	0.223	0.422	0.618	0.805
Diff	0.025	0.016	-0.044	-0.023

the mean scores (over the 150 topics of TREC’s 6-8) of the original base query (RIA code bf.0.0⁴). “BF” gives the mean scores of the blind feedback run using the groups’ favored parameters (bf.system). “Diff” is the BF score minus the Base score (rounded to 3 decimal places). The difference is in bold if it is statistically significant (at the 5% level by the *t*-test). Consistent with the hypothesis, the statistically significant differences were positive for MAP and P10 and negative for MRR and GS10 (GS10 is defined below).

Table 2 summarizes the impact of BF for 30 retrieval measures, including the 25 major mean ad hoc measures of the trec.eval 8.0 utility⁵ and the 4 popular Success@*n* measures⁶. “Diff7” is the average difference over the 7 systems. The 5-tuple gives the number of systems for which the mean difference was negative and statistically significant (*n_s*), otherwise negative (*n_o*), zero (*z*), positive but not statistically

⁴from http://ir.nist.gov/ria/experiments/bf_base/

⁵from http://trec.nist.gov/trec_eval/ (our scores may differ slightly because we do not re-order “tied” rows)

⁶“Success@*n*” (for a topic) is 1 if a relevant item is retrieved in the first *n* rows, 0 otherwise.

Table 2: Mean Impact of Blind Feedback on 30 Retrieval Measures (Across 7 RIA Systems)

	Measure	Diff7	(n_s, n_o, z, p_o, p_s)
1	GS10	-0.032	(5, 2, 0, 0, 0)
2	S10	-0.030	(3, 2, 2, 0, 0)
3	MRR	-0.027	(2, 4, 0, 1, 0)
4 *	I0	-0.025	(1, 5, 0, 1, 0)
5	S1	-0.020	(1, 2, 0, 4, 0)
6	S5	-0.023	(0, 4, 1, 2, 0)
7	S1000	-0.004	(0, 2, 5, 0, 0)
8 *	I100	0.002	(0, 0, 0, 6, 1)
9 *	P5	0.019	(0, 0, 0, 6, 1)
10 *	P10	0.017	(0, 0, 0, 5, 2)
11	GMAP	0.006	(0, 2, 0, 2, 3)
12 *	I90	0.009	(0, 0, 0, 4, 3)
13 *	I10	0.029	(0, 0, 0, 4, 3)
14 *	I20	0.035	(0, 0, 0, 2, 5)
15 *	R-Prec	0.019	(0, 0, 0, 1, 6)
16 *	I80	0.022	(0, 0, 0, 1, 6)
17 *	P15	0.023	(0, 0, 0, 1, 6)
18 *	I70	0.029	(0, 0, 0, 1, 6)
19 *	P1000	0.005	(0, 0, 0, 0, 7)
20 *	P500	0.009	(0, 0, 0, 0, 7)
21 *	P200	0.016	(0, 0, 0, 0, 7)
22	bpref	0.018	(0, 0, 0, 0, 7)
23 *	P100	0.021	(0, 0, 0, 0, 7)
24 *	P30	0.023	(0, 0, 0, 0, 7)
25 *	P20	0.025	(0, 0, 0, 0, 7)
26 *	MAP	0.027	(0, 0, 0, 0, 7)
27 *	I60	0.036	(0, 0, 0, 0, 7)
28 *	I50	0.038	(0, 0, 0, 0, 7)
29 *	I30	0.038	(0, 0, 0, 0, 7)
30 *	I40	0.041	(0, 0, 0, 0, 7)

significant (p_o), or positive and statistically significant (p_s). The measures are in ascending order by $p_s - n_s$ and then Diff7. Of the 22 mean measures reported on the RIA web site⁷ (marked with an asterisk (*)), 21 favored BF, typical of the lack of coverage of BF’s downside.

3. MEASURING EARLY PRECISION

“Early precision” is usually evaluated as precision after the first 5-30 documents retrieved. In Table 2, the listed Precision@ n measures⁸ (P5, P10, P15, P20, P30, etc.) were all higher with BF, including at least one statistically significant increase for each measure.

But if one eliminates any possible influence of secondary recall either by evaluating precision after just 1 item is retrieved, i.e. Success@1 (S1), or after just the first retrieved relevant item, i.e. mean reciprocal rank (MRR), one finds that BF impaired early precision (the few statistically significant differences for S1 and MRR were negative).

Similarly, the earliest of the “interpolated” precision measures (I0) declined with BF, but when even 10% recall was required before measuring precision (I10), BF boosted it.

The situation for measuring early precision is unsatisfactory. The common choice, Precision@ n , is apparently com-

⁷bf_base/runs/*/eval files viewed December 15, 2005

⁸“Precision@ n ” (for a topic) is x/n where x is the number of relevant items retrieved in the first n rows.

promised by secondary recall even for n of 5 or 10. Meanwhile, the pure early precision measures, S1 and MRR, are overly sensitive to small changes in the first few ranks, not accurately reflecting user benefit and making statistical significance harder to achieve.

4. ESTIMATING READING SAVED

An alternative to precision is to estimate “reading saved” to the first relevant item, such as by using the Generalized Success@10 (GS10) measure (introduced as “First Relevant Score” in [2]). GS10 is 1.08^{1-r} where r is the rank of the first relevant item retrieved (GS10 is 0 if no relevant item is retrieved). Compared to reciprocal rank ($\frac{1}{r}$), GS10 falls less sharply in the early ranks; GS10 is 1.0 at rank 1, 0.93 at rank 2, 0.86 at rank 3, etc. GS10 is considered a generalization of Success@10 because it rounds to 1 for $r \leq 10$ and to 0 for $r > 10$. GS10 drops by less and less as r increases, reflecting that users are less and less likely to read deeper into the result list.

GS10 produced more negative statistically significant differences for BF ($n_s=5$ in Table 2) than any of the precision-based measures.

5. IMPLICATIONS FOR ROBUSTNESS

The robustness of blind feedback has been criticized because of its tendency to “not help (and frequently hurt) the worst performing topics” [5]. However, the experimental “Geometric MAP” measure of the TREC Robust Track [6] still favored BF ($p_s=3$ in Table 2). Primary recall measures are more consistent with the expectations of a robustness measure ($n_s \geq 1$), and GS10 particularly so.

This result suggests that robustness is essentially about getting at least one relevant item near the top of the list for every topic, intuitively a clean fit with the Robust Track goal of avoiding “abject failures” [5].

6. CONCLUSIONS

Blind feedback is detrimental if seeking just one item. This downside is not reflected by most of the commonly reported early precision measures. A more careful focus on the first relevant item is needed.

7. ACKNOWLEDGMENTS

Thanks to Donna Harman for suggesting that we test our hypothesis on the RIA archive.

8. REFERENCES

- [1] D. Harman and C. Buckley. The NRRC Reliable Information Access (RIA) Workshop. In *Proceedings of ACM SIGIR 2004*, pp. 528-529, 2004.
- [2] S. Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServerTM at CLEF 2005. In *Working Notes of CLEF 2005*.
- [3] S. Tomlinson. CJK Experiments with Hummingbird SearchServerTM at NTCIR-5. In *Proceedings of NTCIR-5*, pp. 156-163, 2005.
- [4] S. Tomlinson. Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServerTM at TREC 2005. In *Proceedings of TREC 2005*.
- [5] E. M. Voorhees. Overview of the TREC 2003 Robust Retrieval Track. In *Proc. of TREC 2003*, pp. 69-77, 2004.
- [6] E. M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. In *Proc. of TREC 2004*, pp. 70-79, 2005.