Precision-At-Ten Considered Redundant

William Webber
Alistair Moffat
Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia
{wew, alistair}@csse.unimelb.edu.au

Justin Zobel
NICTA VRL
The University of Melbourne
Victoria 3010, Australia
jz@csse.unimelb.edu.au

Tetsuya Sakai
Newswatch, Inc., Japan
sakai@newswatch.co.jp

ABSTRACT

Information retrieval systems are compared using evaluation metrics, with researchers commonly reporting results for simple metrics such as precision-at-10 or reciprocal rank together with more complex ones such as average precision or discounted cumulative gain. In this paper, we demonstrate that complex metrics are as good as or better than simple metrics at predicting the performance of the simple metrics on other topics. Therefore, reporting of results from simple metrics alongside complex ones is redundant.

Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and software – *performance evaluation*.

General Terms

Measurement, performance, experimentation

1. INTRODUCTION

The performance of information retrieval systems is assessed using effectiveness metrics. Many metrics have been proposed, amongst which a distinction can be made between relatively simple metrics such as precision at 10 documents retrieved (P@10) and reciprocal rank (RR), and more complex metrics such as average precision (AP), normalized discounted cumulative gain (nDCG), and rank-biased precision (RBP) [Buckley and Voorhees, 2005, Järvelin and Kekäläinen, 2002, Moffat and Zobel, to appear]. The simple metrics, it is sometimes argued, capture some elementary user behavior such as "looks at the first page" or "stops at the first relevant document" [Turpin and Scholer, 2006], while the other metrics make richer use of an arbitrarily long ranking containing (it is presumed) many relevant documents.

While the more complex metrics are recognized as being more stable and discriminative [Buckley and Voorhees, 2000], researchers frequently report several different metrics when giving performance results, and may even base inferences about system properties on inconsistencies between the metrics. However, even if one accepts the premise that a given, simple metric is perfectly representative of user experience (a premise that we do not support), in retrieval experiments what is of interest is not how well a system performs on the particular set of topics included in the test collection, but rather, how reliably these results predict system performance on other topics and other collections.

Here we demonstrate that the more complex metrics are in fact as good as or better than the simple metrics at predicting the performance of the same simple metrics on new topics, presumably because the former incorporate more information about overall system performance than do the latter. Therefore, provided enough information – in particular a sufficient depth of judgments – is available, only the complex metrics should be reported, and conflicting results from the simpler metrics should be either discounted, or interpreted as an indication that no conclusions can be drawn.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. EXPERIMENTAL DATA AND METHOD

We use the submitted runs and relevance judgments from the AdHoc Track of TREC 8 and the Terabyte Track of TREC 2004. The TREC 8 experimental set has 50 topics and 129 systems; the TREC 2004 experimental set has 49 topics and 70 systems.

In a retrieval experiment, systems are ranked by the mean of their per-topic performance measures. Predictive power can be understood as measuring how reliably the ranking based on the experimental topics predicts system ranking on other, untested topics. To estimate this from the TREC experimental collections, we randomly partition the topic set in half, and calculate the Kendall's τ correlation on the system rankings based on one partition and the system rankings based on the other. This random partitioning is performed $2{,}000$ times for each data point. The mean of these Kendall's τ values is then the statistic of predictive power, which we denote as ϕ. (In separate experiments we have explored halving the topic set as an approximation of sampling from a large population, following up the tests of Voorhees and Buckley [2002], Sanderson and Zobel [2005], and Sakai [2005]. Our conclusions are that it is a reliable approach.) Being based on correlation, ϕ is reflexive: $\phi_{A,B} = \phi_{B,A}$. A metric's predictivity, therefore, must be assessed in relation to all other metrics, including its self-predictivity.

In any TREC experiment, there are poor runs that may have some programming bug or use an unsuccessful experimental algorithm. Since such systems are consistently lowly ranked, no matter what the topic, they inflate τ and hence ϕ. To prevent this effect, only the top 75% of systems by AP are included in our experiments. The trends reported below are also observable with the full system sets.

3. EXPERIMENTAL RESULTS

Table 1 gives the predictive power of the simple and complex metrics for the TREC 8 experimental set. The diagonal values indicate that, as expected, the simple metrics are poorer self-predictors,

	P@10	RR	RBP.95	AP	nDCG
P@10	0.50	0.40	0.53	0.51	0.50
RR		0.39	0.41	0.36	0.37
RBP.95			0.58	0.57	0.55
AP				0.63	0.60
nDCG					0.61

Table 1: Predictive power ϕ of different metrics on the top 75% of TREC 8 AdHoc Track systems, calculated from 2,000 random repartitionings of the topic set.

	P@10	RR	RBP.95	AP	nDCG
P@10	0.64	0.48	0.64	0.64	0.64
RR		0.36	0.48	0.47	0.47
RBP.95			0.68	0.70	0.69
AP				0.80	0.79
nDCG					0.80

Table 2: Predictive power ϕ of different metrics on the top 75% of TREC 2004 Terabyte Track systems.

and therefore less reliable, than the more complex measures, with RR being particularly poor. More significantly, the results show that the complex metrics are as good at predicting the simple measures as the simple measures are at predicting themselves. The results for TREC 2004, in Table 2, tell a similar story for P@10. They also show that the complex measures and even P@10 are better at predicting RR than RR is at predicting itself; despite the emphasis that has been given to RR in some experiments, we infer that it has poor predictivity. Note that on TREC 2004 the RBP metric (with $p = 0.95$) performs relatively poorly, compared to TREC 8, possibly due to the large number of relevant documents in the TREC 2004 Terabyte collection flooding the higher ranks of each run. Other experiments not reported here suggest that $p = 0.98$ yields results consistent with AP and nDCG.

Figure 1 explores the effect on predictive power of increasing the size of the topic sets. When there are only 5 topics in each subset, predictive power in all combinations is low, but the self-predictive powers of nDCG and P@10 are similar, while P@10 is better at predicting itself than nDCG is at predicting P@10. As topic set size increases, nDCG's self-predictive power increases more rapidly than that of P@10, and gradually nDCG becomes a (marginally) better predictor of P@10 than P@10 is of itself. Other complex metrics also demonstrate similar increases in relative predictive power against P@10 as topic subset size increases on the TREC 8 dataset. In contrast, on the TREC 2004 data set, the complex metrics have greater outperformance over P@10 at smaller topic subset sizes, with the gap narrowing as more topics are added. The reasons for this differing behavior merit further investigation.

4. CONCLUSION

Given that the aim of reporting results is to demonstrate that a new system is expected to be better, on unseen data, than a baseline system is according to some measure, and that nDCG and AP are as good at predicting ordering by P@10 as P@10 is, we conclude that reporting P@10 is redundant. For the same reasons, but with even more compelling evidence, we conclude that reporting RR is also unnecessary.

In a comparison of a small number of systems, as is typical of the results section of a research paper, the different metrics can give inconsistent results. A tempting interpretation is that the improvement is good for precision but poor for recall, or similar. Our investigation shows that such inferences are probably wrong, and that a better interpretation is that inconsistency in results means that they cannot be used to draw any substantial conclusions. That is, since experiments show that even significant results are not necessarily predictive, a conflict may mean that the number of topics was insufficient to infer true system behavior. Where such a conflict occurs, the greater overall predictivity of the complex measures observed in Tables 1 and 2 means that they are more likely to be correct than the simple measures that are often given equal prominence.

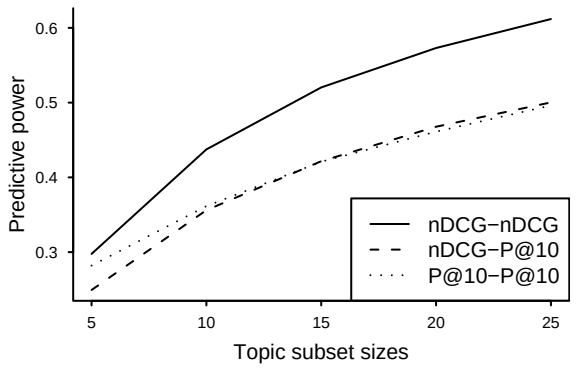

Figure 1: Predictive power ϕ of nDCG and P@10 of themselves, and nDCG of P@10, with different topic subset sizes, on the TREC 8 runs.

We note, as a final remark, that we have not addressed the more complex issue of how best to *design* an experiment so as to obtain the maximum amount of predictivity given a resource cost in terms of judgments to be performed. When all costs are taken into account, it may well be that computing P@10 over a large number of queries is both more economical and more predictive than computing a more intricate measure over a smaller number of queries.

References

C. Buckley and E. Voorhees. Retrieval system evaluation. In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3. The MIT Press, Cambridge, MA, 2005.

C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. 23rd Ann. Int. ACM SIGIR Conf. on Res. and Dev. in Info. Retr.*, pages 33–40, Athens, Greece, 2000.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, to appear.

T. Sakai. The effect of topic sampling on sensitivity comparisons of information retrieval metrics. In *Proc. 5th NTCIR Workshop Meeting on Information Access Technologies*, pages 505–512, Tokyo, Japan, 2005.

M. Sanderson and J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proc. 28th Ann. Int. ACM SIGIR Conf. on Res. and Dev. in Info. Retr.*, pages 162–169, Salvador, Brazil, 2005.

A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. 29th Ann. Int. ACM SIGIR Conf. on Res. and Dev. in Info. Retr.*, pages 11–18, Seattle, WA, 2006.

E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proc. 25th Ann. Int. ACM SIGIR Conf. on Res. and Dev. in Info. Retr.*, pages 316–323, Tampere, Finland, 2002.