What Deliberately Degrading Search Quality Tells Us About Discount Functions

Paul Thomas
CSIRO
Canberra, Australia
paul.thomas@csiro.au

Tim Jones
Australian National University
Canberra, Australia
tim.jones@anu.edu.au

David Hawking
Funnelback Pty Ltd and
Australian National University
Canberra, Australia
david.hawking@acm.org

ABSTRACT

Deliberate degradation of search results is a common tool in user experiments. We degrade high-quality search results by inserting non-relevant documents at different ranks. The effect of these manipulations, on a number of commonly-used metrics, is counter-intuitive: the discount functions implicit in P@k, MRR, NDCG, and others do not account for the true relationship between rank and value to the user. We propose an alternative, based on visibility data.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*
General Terms: Measurement
Keywords: Metrics; result set manipulation

1. INTRODUCTION

There is considerable interest in the ability of IR evaluation metrics to predict user performance on search tasks. Similarly, there is interest in rating and comparing retrieval systems on the basis of implicit measures derived from user behaviour. Empirical exploration of these questions would benefit from the ability to generate search result lists with a particular expected score on a chosen measure. For example, it would be useful to produce result lists with expected NDCG scores of 0.2, 0.4, 0.6 and 0.8.

In previous work [5] this has been achieved by artificially constructing result lists from lists of known relevant and irrelevant documents. A major drawback of this approach is the need to constrain subjects to use queries from a predefined set for which judgments are available and to rely on third-party judgments. In our work we wanted to achieve controlled degradation of an initial high quality ranking, in the absence of "canned" queries and judgments.

2. DEGRADATION MODEL

We assume a retrieval system which produces a ranking $\langle d_1, \ldots, d_M \rangle$ to a depth M. These results are presented in pages of p results each, of which the first v are visible without scrolling (*above the fold*). A user of the system views results to a maximum rank V where V is usually less than M, may be less than p and varies with the user, the task and the results viewed earlier in the ranking. We also assume that we have a large set $\mathcal{N}$ of documents which are known (with high probability) to be non-relevant to the present topic.

A document from $\mathcal{N}$ can degrade a ranking $\langle d_1, \ldots, d_M \rangle$ by insertion at position $r \leq M$. As a result of the insertion the lowest ranked document d_M becomes invisible[1].

A *degradation* consists of an ordered set $\mathcal{I} = \{r_1, \ldots, r_m\}$ of insertions, in which the r_i represent the ranks at which insertions are made. The effect of a degradation depends upon both the metric and the ranking.

3. METRIC-DEPENDENT EFFECTS

Clearly, each of the metrics is affected in a different way by a particular degradation applied to a ranking. For example, all degradations which make the same number of relevant documents invisible will cause the same drop in P@k, regardless of where non-relevant documents are inserted. The effect of a degradation on MRR depends upon how many of the $r_1, \ldots, r_m$ are less than or equal to the rank of the first relevant document. If none, there is no drop in score. For P@k, MRR, AP, RBP [4] and the Microsoft variant of NDCG [1] which we use here, it is possible to calculate for a given ranking the score drops associated with all $2^M - 1$ possible degradations.

We took the four top-scoring runs from the TREC-13 web track [2] and all 75 topic distillation queries. We took result lists for each of these 300 run/topic pairs, using TREC qrels, and computed each of the metrics above. Then, letting $M = 10$, we applied all 1023 degradations to all 300 rankings and computed the resulting change in each metric.

The most dramatic $|\mathcal{I}| = 1$ degradation is obviously when we insert a non-relevant document at the highest rank: $\mathcal{I} = \{1\}$. We designate other degradations as *1-equivalent* if they cause the same change in score as insert-at-rank-1-only. Looking at the 300 changes for each of the 1023 degradations for the TREC data, we used a paired t test, $\alpha = 0.05$, to find all those degradations which are 1-equivalent.

By way of example, Figure 1 shows the changes in NDCG scores for a small number of 1-equivalent degradations. The 1-equivalent sets are different for different metrics.

Examining the 1-equivalent degradations in the figure gives pause for thought. Is it really the case that inserting non-relevant documents at ranks 5, 6, 7, 9, and 10 has equivalent effect to inserting a single non-relevant document at rank one? What is NDCG measuring if this is the case? Would

[1] An alternative degradation would be that D could replace d_i but this would sometimes mean that the single best answer for a query (e.g. a homepage) originally ranked 1 might disappear entirely. We consider only insertions.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

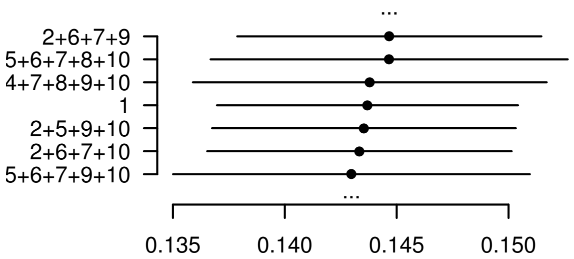

Figure 1: Changes in NDCG for some 1-equivalent degradations. Shown are mean change ± standard error.

it still be equivalent if $v < 5$ as might be the case for search results presented on a mobile phone, i.e. none of the inserted documents are visible without scrolling, or if $p < 10$?

As a further example, under MAP the degradations $\{1\}$ and $\{4, 5, 6, 7, 8\}$ produce equivalent effects; under MRR there are 151 1-equivalent degradations (of 1023 possible degradations) including those with the insertion of only two documents and those which insert eight, leaving only two documents from the original ranking. Should a manipulation which leaves original documents only at ranks two and nine, say, be considered equivalent to one which includes a single insertion at high rank?

4. DISCOUNT FUNCTIONS

From the definition of a particular metric we can infer a user model, if we assume that scores are intended to reflect user satisfaction or performance. Of particular interest here is the relationship between the value (gain) attributed to a retrieved relevant document and the rank at which it is retrieved: Järvelin and Kekäläinen's "discount function" [3]. Figure 2 shows the striking differences between the discount functions inferred from several commonly used metrics.

Considering the effect of insertions of non-relevant documents on user performance or behaviour, it is clear that insertions beyond rank V can have no effect at all on user behaviour because the user never sees them. Insertions beyond rank v on a displayed page have less effect on users than those above because there is a less than unity probability that the user scrolls to see them. Finally, the effect of insertions beyond rank p is further diminished by the low probability that the user will click on the next-page button.

It is clear that if an IR metric is to predict user satisfaction or performance, the discount function implicit in the metric must attempt to more accurately model the dependence of view-probability on rank. Rank-biased precision (RBP) [4] explicitly models probability-of-view but does not take into account the discontinuities in probability which surely occur after ranks $p, 2p, 3p, \ldots$ and at $v, p+v, 2p+v, 3p+v, \ldots$.

Discount functions can easily be extended to accommodate these discontinuities. The plot labelled "RPB-visible" in Figure 2 illustrates such a modification, here of RBP, where the base probability of viewing the next result is normally 0.75 but drops to 0.5 at each fold ($v = 5$) and after each page ($p = 10$). The parameter values in the plot were arbitrarily chosen, but browser interaction logging and/or eye-gaze tracking can determine appropriate values for the constants in particular search scenarios. We are presently carrying out such studies.

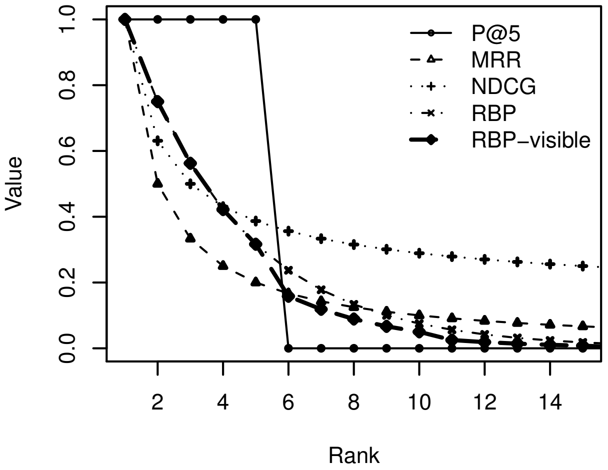

Figure 2: Discount functions implicit in some illustrative evaluation metrics. "RPB-visible" is discussed in the text. (Connecting lines shown only for clarity.)

5. SUMMARY AND CONCLUSIONS

In our user experiments, our attempts to artificially degrade search results to achieve specified quality levels have encountered a number of difficulties. Since we could not assume canned queries or prior relevance judgments, we inserted known non-relevant documents at chosen points in the ranking. The effect of degradation operations consisting of such insertions is highly dependent upon the metric chosen but all commonly used metrics showed counter-intuitive properties. We concluded that the discount functions implicit in metrics such as P@k, AP, MRR, NDCG and RBP do not model the true relationship between rank and visibility/value to the user. We propose an alternative shape of discount function for empirical confirmation, which is introduced for RBP but perfectly consistent with multiple relevance levels and NDCG. We suggest determining parameter values for this shape using browser instrumentation and eye-gaze tracking.

6. REFERENCES

[1] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proc. Int. Conf. on Machine Learning*, 2005.

[2] N. Craswell and D. Hawking. Overview of the TREC-2004 Web Track. In *Proc. TREC*, 2004.

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4), 2002.

[4] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1), 2008.

[5] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. SIGIR*, 2006.