Observing Users to Validate Models

(Extended Abstract)

Falk Scholer
School of Computer Science and Information Technology
RMIT University
falk.scholer@rmit.edu.au

Paul Thomas
ICT Centre
CSIRO
paul.thomas@csiro.au

Alistair Moffat
Department of Computing and Information Systems
The University of Melbourne
ammoffat@unimelb.edu.au

ABSTRACT

User models serve two purposes: to help us understand users, and hence determine how to supply them with effective search services; and as a framework against which to evaluate the quality of those services once they have been developed. In this extended abstract we describe an experiment we have undertaken in which we observe user behaviours, and try to determine whether these behaviours can be connected to search quality metrics via an existing or novel user model. We provide summary evidence that suggests that the answer is a qualified "yes".

Categories and Subject Descriptors

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

General Terms

Experimentation, measurement.

Keywords

Retrieval experiment, evaluation, system measurement.

1. METRICS, MODELS, AND BEHAVIOUR

Every model embodies certain assumptions about the system it represents, and is used to make predictions about the future behaviour of that system. The usefulness of these predictions depends both on the veracity of those initial assumptions, and on the fidelity of the model. In retrieval, metrics such as DCG, RBP, or Prec@k are built atop explicit or implicit models of user behaviour and preferences. The usefulness of these metrics again depends on their assumptions and fidelity.

In their description of the discounted cumulative gain (DCG), effectiveness metric, Järvelin and Kekäläinen [2002] observe that the discounting of relevance contributions is in no small part a response to user behaviour. They argue that for a variety of reasons a user is less likely to gain benefit from a relevant document deep in a ranking than they would if they observed the same document earlier, and propose that the relevance of the document at depth i in the ranking be discounted by a factor of $1/\log_2(i+1)$ (in the Microsoft version of the formulation). Moffat and Zobel [2008] suggest a different discounting function but retain the same underlying philosophy. Their rank-biased precision (RBP) metric assumes that the user is likely to abandon their review of a search ranking at each presented document with some fixed probability $(1-p)$; they then derive a geometric discounting function.

Both DCG and RBP presume that user behaviour can be anticipated via a probabilistic one-state model with just three significant transitions: *enter the reading state*, taken with probability 1.0; *remain in the reading state*, taken with probability p at each trial; and *exit from the reading state*, taken with probability $1-p$. With such a probabilistic model, weighted effectiveness metrics (of which RBP is an example) can then be seen as being a calculation of the rate at which the user gains utility from their searching action. Zhang et al. [2010] used click-log data to compare the user behaviour predicted by the models embedded in DCG and RBP, and found that $p = 0.73$ was a good fit with RBP. Researchers have also measured user behaviour by monitoring gaze locations while they view search rankings [Joachims et al., 2005]. Both types of study provide support for models of behaviour in which documents near the top of the ranking are more likely to be accessed than ones further down.

Other user models have also emerged. The expected reciprocal rank (ERR) metric introduces the notion of *adaptivity*; that is, that the user's behaviour will be affected by the relevance of the documents that they encounter in the ranking. In contrast, DCG and RBP are *static* – the user is predicted to act in a certain way, regardless of whether they are overwhelmed by relevant documents or see none at all. Other "knobs" have also been added by researchers seeking to make their models more accurate or to derive a given metric [Clarke et al., 2008, Dupret and Piwowarski, 2010]. Most recently, Smucker and Clarke [2012] have described a metric they call *time-biased gain*, which models utility not as something that accrues over documents, but over units of seconds or minutes. That is, a long (and hence slow to read) document contributes a lower score than does a shorter one that is equally relevant.

2. AN EXPERIMENT

If metrics are bound with user models, and if we accept that a model is only as useful as its predictions are reliable, there are clearly questions to be asked. Is user behaviour predicted more precisely by some user models than others? Is more complexity in a model (and metric) justified by greater fidelity? Or is user behaviour so varied that there is no such thing as a model user at all?

To investigate this, we used an instrumented search interface built on the anonymised API of a commercial search service to monitor a group of 34 users, while they undertook a set of six search tasks of differing complexity. Each user was asked to note any "useful" pages that they found while searching, so as to answer the information need. At the same time, all user actions – including gaze position, via eye-tracking hardware – were monitored. In half of the user-topic combinations we showed deliberately degraded

SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013), August 1, 2013, Dublin, Ireland.

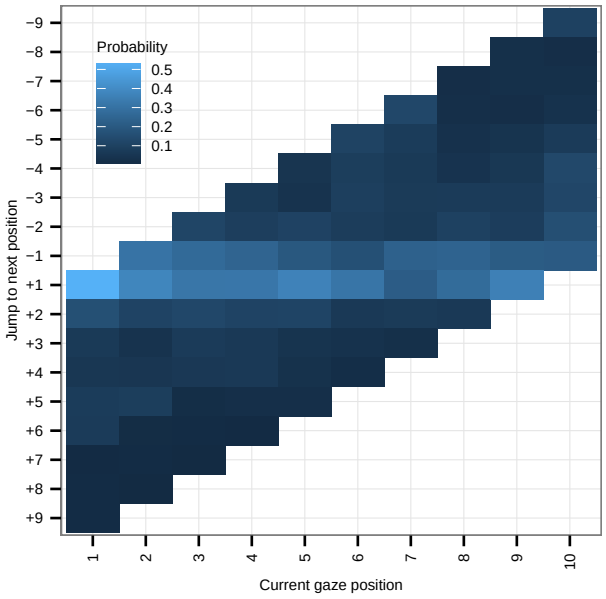

Figure 1: Gaze transitions as ranked answer lists are viewed. The horizontal scale shows current rank position of gaze point; the vertical scale shows the proportion of next gaze points that are at positions $\ldots, -2, -1, 0, +1, +2, \ldots$ relative to that starting position. Higher probabilities are shown by brighter colors. Except at rank one, jumps of -1 (that is, to the snippet above this one on the result page) are only slightly less frequent than jumps of $+1$ to the next snippet down the page. The "fold" in the results page occurred between snippets 6 and 7.

result pages, with every second result one that matched some of the query terms, but was nevertheless clearly not relevant. This allows us to gauge the extent to which users are sensitive to the appearance of relevant documents in the ranking. Details of the experimental structure are provided in the full paper.

3. SUMMARY OF RESULTS

Our preliminary investigations have focused on two facets of user behaviour. First, most of the user models embedded in quality metrics assume that users read result lists top-to-bottom. We sought to understand how users actually progress through rankings. Figure 1 illustrates some of the data collected, after processing sequences of gaze points into fixations and snippet views. In this heatmap, each column represents the observed probability distribution of users' next gaze positions, conditioned by their current gaze location. Jumps of $+1$ (to the next document down the ranking) are common, but so too are jumps of -1. Jumps of $+2$ and -2 also take place. Jumps down the list are more likely than not to be followed by jumps back up, and vice versa, although with an overall downward trend. It seems that in a sense users do progress linearly down the ranking, but also frequently compare the snippet just inspected with one(s) seen earlier, perhaps maintaining a mental "best candidate so far" until they reach enough confidence to go ahead and click.

We also asked whether users behave differently when presented with diluted rankings that contain fewer relevant documents. Static effectiveness metrics such as DCG and RBP predict that there should be no difference in behaviour; adaptive models such as ERR suggest we should see a difference. Our preliminary analysis of the data suggests that users are certainly aware of the inserted snippets,

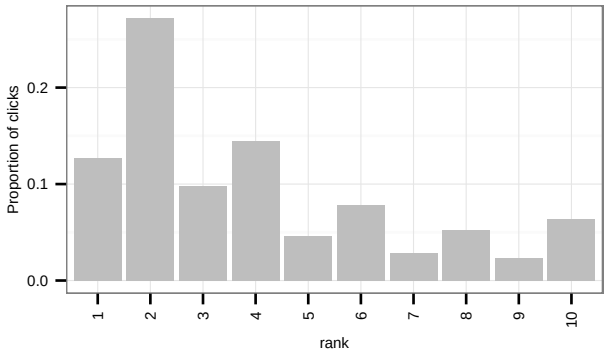

Figure 2: Clickthroughs by depth, for degraded rankings only, with the inserted irrelevant snippets at ranks 1, 3, 5, and so on. A further 6.6% of clicks were past rank 10.

because they do not click on them (Figure 2). But nor have we found any real difference in any of their other behaviours that might have been sensitive to snippet quality – there is no increased depth of viewing in the ranked list, and there is no difference in second-page requests or query reformulation [Thomas et al., 2013]. Hence, it is as yet unclear that the additional complexity of this sort of adaptive model can be recouped by more precise prediction of user behaviour.

4. ONGOING WORK

We have constructed an empirical "best fit to data" model that chooses from amongst a wide range of possible factors when seeking to predict the point at which a user abandons a result listing, and takes a different action. Results to date suggest that rather than "relevance of the most recently viewed document" as a predictor of exit, a more potent factor is "proportion of anticipated total relevance that has been accrued until this moment", which automatically folds in task type. In the full paper we embed this concept into a model of user behaviour, and hence an alternative effectiveness metric.

Acknowledgment. This work was supported by the Australian Research Council.

References

C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proc. SIGIR*, pages 659–666, Singapore, 2008.

G. Dupret and B. Piwowarski. A user behavior model for average precision and its generalization to graded judgments. In *Proc. SIGIR*, pages 531–538, Geneva, Switzerland, 2010.

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

T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. SIGIR*, pages 154–161, Salvador, Brazil, 2005.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2:1–2:27, 2008.

M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, Portland, Oregon, 2012.

P. Thomas, F. Scholer, and A. Moffat. Fading away: Dilution and user behaviour. In *Proc. EuroHCIR*, Dublin, 2013. To appear.

Y. Zhang, L. A. F. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1):46–69, 2010.