Users, Adaptivity, and Bad Abandonment

Alistair Moffat
The University of Melbourne
Melbourne, Australia

Alfan Farizki Wicaksono
The University of Melbourne
Melbourne, Australia

ABSTRACT

We consider two recent proposals for effectiveness metrics that have been argued to be adaptive, those of Moffat *et al.* (ACM TOIS, 2017) and Jiang and Allan (CIKM, 2017), and consider the user interaction models that they give rise to. By categorizing non-relevant documents into those that are *plausibly non-relevant* and those that are *egregiously non-relevant*, we capture all of the attributes incorporated into the two proposals, and hence develop an effectiveness metric that better reflects user behavior when viewing the SERP, including bad abandonment.

ACM Reference Format:
Alistair Moffat and Alfan Farizki Wicaksono. 2018. Users, Adaptivity, and Bad Abandonment. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210075

1 INTRODUCTION

Measurement of the quality of search techniques is usually done in one of two ways: via the application of an *effectiveness metric* to the rankings (SERPs) generated in response to a set of queries, and aggregation or statistical testing over the set of numeric scores that result; or via *user satisfaction* studies that elicit feedback from people using the search service. In the first of these, the choice of metric is a key consideration – even though metrics scores are usually correlated, different metrics reflect different assumptions in regard to user behavior, and hence choosing a metric is tantamount to choosing a *user model*. For example, RR, where the score assigned is the reciprocal of the rank at which the first relevant document appears – corresponds to a model in which the user searches from the top of the SERP, and stops as soon as they find a first relevant document. In this framework, the score assigned by the metric is the average rate at which the user gains satisfaction from the SERP, measured in units of "relevance gain per document inspected".

The duality between metrics and models was noted by Moffat and Zobel [8], who proposed a user model (and corresponding metric *rank-biased precision*, RBP) in which users are assumed to always examine the first document in the SERP, and then to continue from one document to the next with a fixed probability ϕ, regardless of the relevance already accumulated. That is, in RBP the *conditional continuation probability at depth i*, denoted $C(i)$, is taken to be constant as the SERP is examined. Moffat and Zobel [8] also anticipated the possibility of other options, writing (page 17):

> *A further variant is to relax the assumption that* ϕ [...] *is independent of whether or not the document just considered is relevant. An arrangement in which the conditional probability of advancing given a relevant document is* ϕ_1*, and the conditional probability of advancing given an irrelevant document is* ϕ_2 [...] *would lead to another mechanism for scoring runs* [...]

Recent work has taken up that challenge, with Moffat et al. [10] describing a user model (and corresponding metric, INST) in which $C(i)$ is recomputed at each depth i in the ranking, altering the user model in two critical ways. First, in INST $C(i)$ increases as a function of i, reflecting that the longer the examination of the SERP continues, the more likely it is (in a conditional sense) that the next document will also be examined. Second, INST is *adaptive* in that $C(i)$ is also modified as relevance is accumulated, with (when other factors are equal) the computed $C(i)$ at depth i being lower if the SERP has already provided a high degree of relevance, reflecting that the closer the user is to satisfying their information need, the smaller their likelihood of continuing to search.

Scores computed by INST are also in units of relevance gain per document inspected, but now the number of documents examined varies depending on the quality of the SERP. Moffat et al. employ a parameter T, the desired "volume" of relevance that the searcher anticipated when they commenced their search, and show that INST models users as searching to an expected depth of between $T + 0.25$ documents (when every item in the SERP is relevant) and $2T + 0.5$ documents (when nothing in the SERP is relevant). In suggesting this arrangement, Moffat et al. implicitly assert that the user is *more* persistent on low-quality rankings than on high-quality ones. For example, a user who is nominally searching for $T = 2$ units of relevance, and at some depth has accumulated three units of relevance already, is less likely to continue to the next document in the SERP than a user who has not yet accumulated any relevance.

In other recent work, Jiang and Allan [7] develop a different approach. They also suggest that the user's reaction to each SERP is influenced by the documents that appear in it, and seek to modify the user model embedded in RBP (and other weighted-precision metrics) to account for that behavior. But (in the context of RBP) their approach makes use of a constant $C(i) = \phi$ function, with ϕ computed not according to characteristics of the user and their inherent understanding of what they are seeking, but instead based on characteristics of the documents in the SERP. By fitting coefficients to eye-tracking and click-through data derived from user interactions, Jiang and Allan suggest that ϕ (and similarly, the parameter k that governs the evaluation depth of the SDCG@k metric) be adjusted so that users faced with a low quality SERP examine *fewer* documents, thereby accounting for *bad abandonment* [5, 6].

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-5657-2/18/07...$15.00
https://doi.org/10.1145/3209978.3210075

That is, Jiang and Allan require that a smaller value of ϕ be employed when the SERP contains few or no relevant documents, and a larger one be used when the SERP contains a high fraction of relevant documents. Note that this corresponds to viewing behavior that is the *opposite* of what is suggested by Moffat et al. [10].

Our contribution. We first examine the implications of the Jiang and Allan approach, examining the scores that are generated for a range of SERPs. We then draw these two different models of user behavior into a cohesive whole that captures both the user behaviors observed by Jiang and Allan and also embeds the fully adaptive approach espoused by Moffat et al. In particular, we provide an enhanced mechanism with the following properties:

- All other aspects being equal, $C(i)$ increases with i.
- All other aspects being equal, $C(i)$ is smaller at any depth i if a larger volume of relevance has been accumulated from documents 1 to i.
- All other aspects being equal, $C(i)$ is smaller at any depth i if a larger number of *egregiously non-relevant* responses have been encountered from documents 1 to i.

As is suggested by the third factor, we achieve this blend by partitioning non-relevant documents into two categories: those that are *plausibly non-relevant*, and those that are *egregiously non-relevant*. For example, if the query were "*the melting point of lead*", we would regard a document that described the "*melting point of* tin" as being plausibly non-relevant, and would feel encouraged that a relevant document might appear soon in the ranking. Conversely, a document that described the role of "Freddie Mercury as the *lead* singer of the band Queen" could suggest abandoning that particular SERP, and indicate that reformulation might be required – it would fall into the "discouraging", or egregiously non-relevant category.

That is, we propose a categorization in which non-relevant documents are of two distinct types: those that are regarded as being *en*couraging of eventual success, and those that are *dis*couraging.

2 BACKGROUND

This section describes the context in which we work. The assumption throughout is that the user examines documents in the order in which they appear in the SERP, continuing from one to the next until they stop, an approach that leads to *weighted-precision* metrics and is sometimes also called the *cascade model* [1, 3, 4]. This simplification ignores the time taken to read each document, and ignores the additional variability introduced by the presentation of the initial snippet, and other non-document cards on the SERP. Other work has explored some of these factors [11].

Weighted-precision metrics. Table 1 lists several effectiveness metrics that yield scores that have units of "expected gain per document inspected", and summarizes the user models that are associated with them. The metric Precision@k reflects an evaluation model in which the user is equally likely to select any document from amongst the first k documents in the SERP (and never selects beyond the first k); and scores the ranking as a whole by computing the expected gain per document inspected. The cascade view of the same metric presumes that the user starts reviewing documents at the top of the ranking, and then continues from the i th document to the next with a *conditional continuation probability* $C(i)$ that has two values, $C(i) = 1$ when $i < k$, and $C(i) = 0$ when $i \geq k$.

In RBP [8], a user-persistence parameter is used to control the evaluation, rather than a depth limit k, and the value of the metric is computed as the expected relevance gained per document inspected when it is assumed that the first document in the ranking is always viewed, and thereafter, $C(i) = \phi$. Moffat and Zobel also introduce the notion of a *residual*, as a useful way of quantifying the extent of the measurement uncertainty.

INSQ. The third row of Table 1 describes an RBP-like approach that adjusts the weightings so that $C(i)$ increases with i rather than remaining constant [9]. Precisely, $C(i) = (2T + i - 1)^2/(2T + i)^2$, which increases towards 1 as i increases. The motivation for this change is one of "sunk cost": a user who has looked at the 20th (or 200th) document is more likely to then go on and look at the 21st (or 201st) than they are to go from the 2nd to the 3rd document. The parameter T is the user's expectation as to the volume of relevance they hope to find, and is correlated with both the complexity of the information need and their persistence when viewing the SERP.

INST. In the penultimate row of Table 1 a third quantity is introduced: T_i, the as yet unsatisfied appetite for relevance. It is also argued to have a positive correlation with $C(i)$, making the corresponding metric INST *adaptive* in terms of the SERP that is being assessed [9, 10].

3 THE JIANG-ALLAN PROPOSAL

Jiang and Allan [7] also consider the question of varying user behavior in response to different SERPs. Here we focus on their proposal for modifying RBP, but note that our comments equally apply to their modifications to other weighted-precision metrics.

Variable, yet fixed. Jiang and Allan propose that the RBP parameter ϕ be variable across rankings, but be fixed for any given ranking. They model the persistence of a universe of users as a linear combination of a set of trainable coefficients, shared across SERPs, and the relevance vector of the particular SERP being scored. A constant value w_0 is learned, plus a set of $(1 + gmax) \times d$ rank-grade weights $w_{i,g}$, where d is an estimation depth that is chosen prior to the fitting process, and the relevance grades g are between zero and $gmax$. If the current SERP is $\mathcal{S} = \langle r_1, r_2, \ldots r_k \rangle$, where each $0 \leq r_i \leq gmax$ is a relevance grade, then $\phi(\mathcal{S})$ is a summation over components, including the flag-fall value w_0, computed as $\phi(\mathcal{S}) = w_0 + \sum_{i=1}^{d} w_{i,r_i}$. Coefficients w_0 and $w_{i,g}$ are estimated from training data consisting of relevance vectors and either matching gaze-tracking data, or matching click-through information [7]. Table 2, taken from Jiang and Allan [7, page 751], shows the weights computed for one dataset with $d = 5$ and $gmax = 2$. Once $\phi(\mathcal{S})$ has been determined, RBP is calculated, with relevance grades $0 \leq g \leq gmax$ converted to gain values using any suitable mapping.

Since one $w_{i,g}$ value must be selected from each row i, the smallest value of ϕ that can be assigned from Table 2 is $0.544 + 0.047 + 0.049 + 0.048 + 0.042 + 0.052 = 0.782$, for a ranking with relevance grades $[0, 0, 0, 0, 0]$; similarly the largest possible for ϕ is 0.982, for a ranking with relevance grades $[1, 1, 1, 2, 1]$.

Ordering inconsistency. Consider the SERPs $\mathcal{S}_1 = [0, 1, 1, 1, 1]$ and $\mathcal{S}_2 = [0, 1, 1, 2, 1]$. with all further documents in both having

Metric	Reference	Param.	User model properties
Precision	—	k	User examines every document in the first k. Expected evaluation depth is k.
RBP	[8]	p	User focuses more on earlier documents, continuing from one document to next with constant conditional continuation probability $C(i) = \phi$ where $\phi < 1$. Amount of uncertainty in score can be quantified into *residual* value. Expected evaluation depth is $1/(1 - \phi)$.
INSQ	[9]	T	Similar to RBP, except that $C(i)$ increases with depth i, reflecting increased user commitment after increasing amounts of sunk effort. Expected evaluation depth is $2T + 0.5$.
INST	[10]	T	Similar to INSQ, except that $C(i)$ decreases (relative to INSQ at the same depth i) as relevance is accumulated out of first i documents. Metric is *adaptive* to the aggregate volume of relevance in the viewed ranking. Expected evaluation depth is between $T + 0.25$ and $2T + 0.5$.
INST-BA	*this paper*	T	Similar to INST, except that $C(i)$ *decreases* (relative to INST at the same depth i) if egregiously non-relevant documents are observed in the first i ranks.

Table 1: A sequence of refinements in user models and hence in weighted-precision evaluation metrics.

$w_0 = 0.544$	w_{i,r_i}		
	$r_i = 0$	$r_i = 1$	$r_i = 2$
$i = 1$	0.047	0.088	0.059
$i = 2$	0.049	0.084	0.061
$i = 3$	0.048	0.096	0.050
$i = 4$	0.042	0.054	0.098
$i = 5$	0.052	0.072	0.070

Table 2: Parameter values computed by Jiang and Allan for SERPs of length five [7, page 751] for graded relevance, $r_i \in \{0, 1, 2\}$.

grade zero. Table 2 then yields $\phi(\mathcal{S}_1) = 0.897$ and $\phi(\mathcal{S}_2) = 0.941$, and computing scores using those two parameters assuming the gain mapping $(2^g - 1)/(2^{g_{max}} - 1)$ (that is, with gains of $\{0, 1/3, 1\}$ for the three relevance grades) we then get $\text{RBP}(\mathcal{S}_1, \phi(\mathcal{S}_1)) = 0.105$ and $\text{RBP}(\mathcal{S}_2, \phi(\mathcal{S}_2)) = 0.101$. But this is counter-intuitive, since $\mathcal{S}_2$ is strictly a superior ranking to $\mathcal{S}_1$, given that it has an additional high-grade document.

Abandoning the SERP. Jiang and Allan motivate their mechanism by arguing that poor-quality SERPs lead to earlier abandonment, noting in their abstract that:

> *rational users change their browsing behavior according to the search result page … avoid wasting time (a low persistence level) if the results look apparently off-topic.*

Consider two further SERPs: $\mathcal{S}_3 = [0, 0, 1, 1, 1]$ and $\mathcal{S}_4 = [2, 2, 2, 1, 2]$, with $\mathcal{S}_3$ clearly more off-topic than $\mathcal{S}_4$. But from Table 2 we have $\phi(\mathcal{S}_3) = 0.862 > 0.838 = \phi(\mathcal{S}_4)$, at odds with the proposition that the user would abandon $\mathcal{S}_3$ earlier than $\mathcal{S}_4$.

Clairvoyant users. Perhaps the most substantial issue that we note with the Jiang and Allan approach is that the corresponding user model requires that the SERP as a whole be clairvoyantly "grokked" by the user prior to their inspection of its first document, in order that the parameter ϕ be set. Like Moffat et al. [10], Jiang and Allan denote their arrangement as being "adaptive", but comparing the two approaches, theirs is a substantially weaker form of adaptivity, and does not allow $C(i)$ to be responsive on a per-document basis. Indeed, it might be argued that while the approach of Jiang and Allan has a parameter that is modified on a per-SERP basis, the resultant measurement is not actually adaptive at all.

A further drawback of the setting of the parameter on a per-SERP basis is that it is no longer straightforward to categorize information needs according to perceived task complexity, or in response to the way in which the user adjusts their viewing behavior because of varying anticipated volumes of relevance [9, 10].

4 AN ALTERNATIVE APPROACH

We now briefly sketch a mechanism that incorporates the "bad abandonment" observation that some SERPs may be exited quickly, and also includes the desires for genuine adaptivity and for the evaluation to be parameterized by task complexity. In the context of Table 1, we define

$$C(i) = \left(\frac{f(i) - 1}{f(i)} \right)^2 ,$$

to be the conditional continuation probability at rank i, and now focus on $f(\cdot)$, noting that $C(i)$ is monotonic in $f(i)$. For RBP, $f(i) = 1/(1 - \sqrt{\phi})$, and is a constant related to (only) the user's persistence. For INSQ, that was altered to $f(i) = (i + 2T)$, and meant that $C(i)$ increased with i. For INST, $f(i) = (i + T + T_i)$, and $C(i)$ is also adaptive to T_i, the amount of relevance still being sought at depth i.

Bad abandonment. To accommodate early abandonment of rankings that contain egregiously non-relevant documents (introduced in Section 1), we add a further factor to $f(\cdot)$, and propose

$$f(i) = \frac{i + T + T_i}{1 + E_i} , \tag{1}$$

where E_i is the number of egregiously non-relevant documents observed by the user during their inspection of ranks 1 to i.

The expected length of search for any weighted-precision metric is given by the reciprocal of the weight of the first document:

$$\text{ESL} = \frac{1}{W(1)} = 1 + \sum_{i=1}^{\infty} \left(\prod_{j=1}^{i} C(j) \right) . \tag{2}$$

Table 3 shows expected depths reached in the ranking when Equation 1 is used as the basis for $C(i)$ for three extreme rankings: when all documents result in a gain of 1.0 being generated (that is, $T_i = T - i$ at all times, the "good"); when all documents in the

T	Expected depth		
	good	bad	ugly
1	1.33	2.58	1.12
3	3.27	6.53	1.79
10	10.26	20.51	3.41
30	30.25	60.50	6.21
T	$\approx T + 0.25$	$\approx 2T + 0.5$	$\approx 1.25\sqrt{T}$

Table 3: Expected depths of search when using Equation 1. The column headings refer to the situations when every document in the ranking has a gain of 1.0; when every document in the ranking is plausibly non-relevant; and when every document in the ranking is egregiously non-relevant. (With apologies to Sergio Leone.)

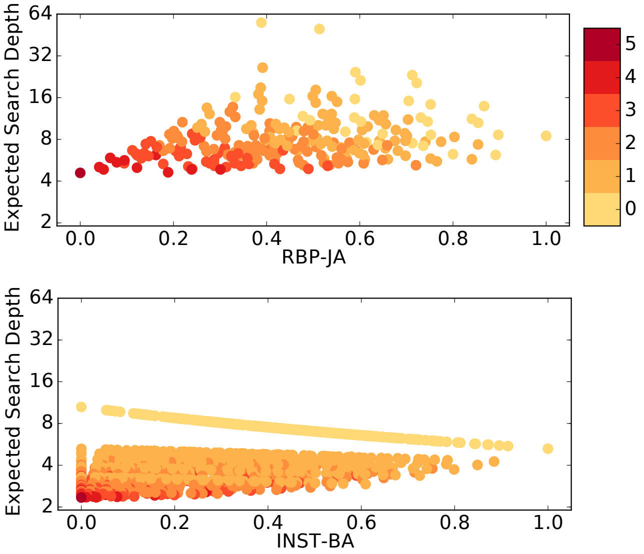

Figure 1: Metric scores and expected search depths for 1024 runs each of 100 items, created from all five-element patterns of the "grades" $\{-1, 0, +1, +2\}$. See the text for further details.

ranking give a gain of zero, but are plausibly non-relevant (that is, $T_i = T$, the "bad"); and when all documents in the ranking are egregiously non-relevant (that is, $E_i = i$ at all times, the "ugly").

Experiment. Figure 1 plots expected search length (Equation 2) as a function of metric score for a set of artificial SERPS. Taking $g \in \{-1, 0, +1, +2\}$ to indicate the grades egregiously non-relevant, plausibly non-relevant, partially relevant, and fully relevant, with corresponding gains of $\{0/3, 0/3, 1/3, 3/3\}$, we constructed all $4^5 = 1024$ possible *patterns* of five grades, and then built a run from each pattern by repeating it 20 times, to obtain a ranking of length 100. Each plotted point represents one run, built from one of the patterns. Note that when weighted-precision metrics are applied to runs of length 100 there is a negligible residual component.

In the first pane, the metric is RBP-JA, with expected viewing depth a function of the computed value for ϕ. In this graph, -1 and 0 are interchangeable, and hence there are $3^5 = 243$ points visible; the colors of the dots represent the number of non-relevant documents in the underlying five-element pattern. While there is a slight trend for stronger runs (to the right on the horizontal axis) to correspond to greater expected depths (higher on the vertical axis), and hence to metric scores that are less heavily top-weighted, it is far from consistent. There are also anomalies that arise, caused by the non-monotonicities in the rows and columns of Table 2. For example, the very high point above 0.4 is for the run based on the pattern $[1, 1, 1, 2, 1]$ and $\phi = 0.982$, as observed earlier.

In the second pane, all 1024 points are plotted, with the colors now representing the number of egregiously non-relevant documents in each pattern, and with the 100-element runs scored using Equation 1 (denoted "INST-BA") at $T = 5$ (chosen to broadly match the average depth of RBP-JA across the 243 patterns over $[0, +1, +2]$). The benefit of including E_i is clear, with the "ugly" runs, containing many egregiously non-relevant documents, being abandoned relatively quickly (the dark-shaded points at lower left); with the "good" runs (light-shaded points, at the right) indicating a viewing depth of approximately T; and with the runs that are free of egregiously non-relevant documents forming the upper line.

5 CONCLUSIONS

By introducing the notion of egregiously non-relevant documents, we have shown that bad abandonment can be reflected in a user model for SERP examination, and hence can also be incorporated into a weighted-precision effectiveness metric. The approach presented here is a proof of concept only, to demonstrate that adjusting $C(i)$ using E_i is feasible. A user study is now the next step, to build evidence in support of Equation 1 or some similar expression that adaptively reduces $C(i)$ as E_i increases.

Finally, note that in very recent work Azzopardi et al. [2] also consider factors that affect the continuation function $C(i)$, concluding that the decision to end viewing a SERP is additionally influenced by the user's implicit expectations in regard to the rate at which gain is accumulated, and providing evidence based on user interactions with 673,000 SERPs in support of that observation.

Acknowledgment. This work was supported under the Australian Research Council's Linkage Projects scheme (LP150100252).

REFERENCES

[1] A. Ashkan and C. L. A. Clarke. On the informativeness of cascade and intent-aware effectiveness measures. In *Proc. WWW*, pages 407–416, 2011.
[2] L. Azzopardi, P. Thomas, and N. Craswell. Measuring the utility of search engine result pages. In *Proc. SIGIR*, 2018.
[3] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proc. SIGIR*, pages 903–912, 2011.
[4] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, 2009.
[5] O. Dan and B. D. Davison. Measuring and predicting search engine users' satisfaction. *ACM Comp. Surv.*, 49(1):18:1–18:35, 2016.
[6] J. Huang, R. W. White, and S. T. Dumais. No clicks, no problem: Using cursor movements to understand and improve search. In *Proc. CHI*, pages 1225–1234, 2011.
[7] J. Jiang and J. Allan. Adaptive persistence for search effectiveness measures. In *Proc. CIKM*, pages 747–756, 2017.
[8] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2.1–2.27, 2008.
[9] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.
[10] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Inf. Sys.*, 35(3):24:1–24:38, 2017.
[11] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.