Users Versus Models: What Observation Tells Us About Effectiveness Metrics

Alistair Moffat
The University of Melbourne, Australia
ammoffat@unimelb.edu.au

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

Falk Scholer
RMIT University, Australia
falk.scholer@rmit.edu.au

ABSTRACT

Retrieval system effectiveness can be measured in two quite different ways: by monitoring the behavior of users and gathering data about the ease and accuracy with which they accomplish certain specified information-seeking tasks; or by using numeric effectiveness metrics to score system runs in reference to a set of relevance judgments. In the second approach, the effectiveness metric is chosen in the belief that user task performance, if it were to be measured by the first approach, should be linked to the score provided by the metric.

This work explores that link, by analyzing the assumptions and implications of a number of effectiveness metrics, and exploring how these relate to observable user behaviors. Data recorded as part of a user study included user self-assessment of search task difficulty; gaze position; and click activity. Our results show that user behavior is influenced by a blend of many factors, including the extent to which relevant documents are encountered, the stage of the search process, and task difficulty. These insights can be used to guide development of batch effectiveness metrics.

Categories and Subject Descriptors

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

Keywords

Retrieval experiment; evaluation; system measurement.

1. OVERVIEW

There has been a tremendous amount of work undertaken in evaluating retrieval system effectiveness. Although there are many alternatives – direct observation of users, log files, or diary studies for example – by far the most common approach is to use one or more batch-evaluation metrics.

The traditional batch-evaluation metrics of precision and recall have been extended by a raft of alternatives, including average precision (AP); discounted cumulative gain (DCG) and normalized discounted cumulative gain (NDCG) [13]; rank-biased precision (RBP) [18]; reciprocal rank (RR); expected reciprocal rank (ERR) [7]; BPref [3]; time-biased gain [21]; plus many more. Carterette [5] gives a framework in which many metrics can be seen as being related; and Moffat [16] categorizes metrics according to their numeric properties.

Underlying all metrics is the assumption that the retrieval system returns a ranked list of documents, and that each of the documents retrieved can be scored for *relevance*, a real value $0 \le r_i \le 1$, with $r_i = 1$ indicating that the ith document in the ranking is highly (or even perfectly) relevant, $r_i = 0$ indicating that the i th document is completely irrelevant, and gradations in between these extremes.

Each effectiveness metric can then be regarded as having a corresponding user model, describing how users interact with the ranked list. For example, Prec@k models each user as inspecting exactly k documents, and by computing $(1/k)\sum_{i=1}^{k} r_i$, generates a score that represents the average rate at which a user accrues relevance.

In this paper we explore the connection between models, metrics, and user behaviors. We begin by establishing a framework in which each metric can be identified with three explicit parts of a user model: weights on each document rank; conditional probabilities of a user continuing to read past each rank; and probabilities of a user stopping at a given rank. Given this formalism, it is natural to then ask: what is a realistic model? Which metrics instantiate this?

Observations from a user study with 34 participants and three types of search task provide concrete data, and we demonstrate that there are a number of factors which contribute to a user's reading behavior which are not considered in present models. Moreover, it is both plausible and possible to include them, and doing so should result in metrics that are more accurate than those in current use.

2. METRICS AND MODELS

Effectiveness evaluations in IR commonly use batch evaluation metrics. For any one query, issued over a set of documents, a relevance score r_i is assigned to each document. Considering the ranked list of documents returned by a search system, and the relevance of each, any number of alternative metrics can then be calculated. For example, precision amongst the first k documents (Prec@k) can be computed with

$$\text{Prec@}k(\vec{\mathbf{r}}) = \sum_{i=1}^{\infty} \mathrm{W}(i) \cdot r_i, \quad \mathrm{W}(i) = \begin{cases} 1/k & \text{when } 1 \le i < k \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

as the inner product of a weight function $\mathrm{W}()$ and the relevance vector $\vec{\mathbf{r}}$. We do not consider the question of how r_i is determined, but note that it is a human process and hence may be the subject of imprecision, and also that it might be context dependent, varying according to what documents have been observed by the user earlier in the ranking. That is, relevance $\vec{\mathbf{r}}$ is due to document and situation, and is not in our power to change. Note that there is no requirement

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2507665.

that relevance be binary, and r_i can be thought of as a fractional value to support graded relevance.

Equation 1 can be generalized to an arbitrary metric M:

$$\mathrm{M}(\vec{\mathbf{r}}) = \sum_{i=1}^{\infty} \mathrm{W}_\mathrm{M}(i) \cdot r_i \tag{2}$$

where $\mathrm{W}_\mathrm{M}()$ is a probability distribution, with $\sum_{i=1}^{\infty} \mathrm{W}_\mathrm{M}(i) = 1$. Metrics have also been suggested in which the sum is not 1, but provided that $\sum_{i=1}^{\infty} \mathrm{W}_\mathrm{M}(i)$ is bounded, these can be normalized into an equivalent set of probabilities. When $\sum_{i=1}^{\infty} \mathrm{W}_\mathrm{M}(i)$ does not converge, normalization is not possible, and truncation at some limiting depth k is required. The issues that arise from truncation are discussed below.

There are two interpretations that can be placed on $\mathrm{W}_\mathrm{M}()$ when it is a probability distribution. In the first, $\mathrm{W}_\mathrm{M}(i)$ is the likelihood that document i is the one being inspected at any given moment by the person examining the ranking; that is, their document inspections constitute a sequence of random selections from $\mathrm{W}_\mathrm{M}()$. The alternative interpretation of $\mathrm{W}_\mathrm{M}()$ is that users examine documents sequentially from the top of the ranked answer list, starting with the first-ranked document. Once they have reached depth i in the ranking, they proceed to depth $i+1$ with conditional probability

$$\mathrm{C}_\mathrm{M}(i) = \frac{\mathrm{W}_\mathrm{M}(i+1)}{\mathrm{W}_\mathrm{M}(i)}\,. \tag{3}$$

Hence, for the metric Prec@k,

$$\mathrm{C}_\mathrm{Prec}(i) = \begin{cases} 1 & \text{when } 1 \le i < k \\ 0 & \text{otherwise}\,. \end{cases}$$

which is to say the user always reads from rank 1 down to k, and then stops.

The relationship between $\mathrm{W}_\mathrm{M}()$ and $\mathrm{C}_\mathrm{M}()$ means that they can be computed from each other. Equation 3 shows how $\mathrm{C}_\mathrm{M}()$ can be derived from $\mathrm{W}_\mathrm{M}()$; the reverse is accomplished by noting that

$$\mathrm{W}_\mathrm{M}(i) = \mathrm{W}_\mathrm{M}(1) \cdot \prod_{j=1}^{i-1} \mathrm{C}_\mathrm{M}(j)\,, \text{ and hence that}$$

$$\mathrm{W}_\mathrm{M}(1) = 1 \Big/ \left(\sum_{i=1}^{\infty} \prod_{j=1}^{i-1} \mathrm{C}_\mathrm{M}(j) \right).$$

There is a third equivalent way of specifying an effectiveness metric. Define $\mathrm{L}_\mathrm{M}(i)$ to be the probability that the ith document in the ranking is the last one observed by the user, that is,

$$\mathrm{L}_\mathrm{M}(i) = \frac{\mathrm{W}_\mathrm{M}(i) - \mathrm{W}_\mathrm{M}(i+1)}{\mathrm{W}_\mathrm{M}(1)}\,.$$

This function is also a probability distribution. For example, $\mathrm{L}_\mathrm{Prec}(i)$ is simply 1 when $i = k$, and 0 otherwise.

To complete the circular relationship between $\mathrm{W}_\mathrm{M}(i)$, $\mathrm{C}_\mathrm{M}(i)$, and $\mathrm{L}_\mathrm{M}(i)$, note that

$$\mathrm{C}_\mathrm{M}(i) = \left(\sum_{j=i+1}^{\infty} \mathrm{L}_\mathrm{M}(j) \right) \Big/ \left(\sum_{j=i}^{\infty} \mathrm{L}_\mathrm{M}(j) \right).$$

The expected number of documents inspected is then given by:

$$\sum_{i=1}^{\infty} i \cdot \mathrm{L}_\mathrm{M}(i) = \sum_{i=1}^{\infty} i \cdot \frac{\mathrm{W}_\mathrm{M}(i) - \mathrm{W}_\mathrm{M}(i+1)}{\mathrm{W}_\mathrm{M}(1)} = \frac{1}{\mathrm{W}_\mathrm{M}(1)}\,.$$

Weighted precision metrics can be characterized by any of $\mathrm{W}_\mathrm{M}(i)$, $\mathrm{C}_\mathrm{M}(i)$, or $\mathrm{L}_\mathrm{M}(i)$: that is, the definition of any one of those three functions completely specifies the metric. Since the three functions describe user behavior, specifying a metric this way also specifies a user model. We illustrate these ideas next.

2.1 Static User Models

We first consider a range of static user models, that is, user models in which the conditional continuation probabilities are a function of rank position alone.

Precision Precision at depth k, or Prec@k, was outlined above. The corresponding user model is that the user examines the first k elements in the ranking and then stops, with each of the k items equally-weighted. This metric is top-weighted in that it assigns zero weight to all documents beyond rank k, but it does not discriminate between ranks within the top k.

Discounted Cumulative Gain Järvelin and Kekäläinen [13] observe that top-weightedness is desirable, and propose a metric they call *discounted cumulative gain*, or DCG@k. They specify DCG in terms of a non-convergent infinite weighting vector. To obtain a probability distribution it is necessary to truncate at some depth, and use a *scaled discounted cumulative gain* metric, defined as:

$$\mathrm{C}_\mathrm{SCDG}(i) = \begin{cases} \log_2(i+1)/\log_2(i+2) & \text{when } 1 \le i < k \\ 0 & \text{otherwise}\,. \end{cases}$$

There may be situations in which truncation at depth k is not acceptable. But if the unrestricted function

$$\mathrm{C}_\mathrm{DCG}(i) = \log_2(i+1)/\log_2(i+2) \approx \frac{i \log_e i - 1}{i \log_e i}$$

is regarded as being the conditional continuation probability, as is implicit in the original proposal of Järvelin and Kekäläinen, then $\mathrm{W}_\mathrm{DCG}(i) \approx 0$ for all values i, and the user is assumed to inspect an unbounded number of items.

Rank-Biased Precision To avoid the discontinuity in behavior at depth k, and to allow for infinite distributions while still giving a gradated response within the top k, Moffat and Zobel [18] suggest an effectiveness metric they call *rank-biased precision* (RBP):

$$\mathrm{C}_\mathrm{RBP}(i) = p\,,$$

where p is a "persistence" parameter that describes the propensity of the user to step from one document to the next. For example, when $p = 0.7$, if the user has examined the ith document in the ranking, there is a 30% probability that they will abandon their search and not proceed to document $i+1$. The weights $\mathrm{W}_\mathrm{RBP}(i)$ form a geometric sequence, $\mathrm{W}_\mathrm{RBP}(i) = (1-p)p^{i-1}$, and the expected number of objects examined is thus $1/\mathrm{W}_\mathrm{RBP}(1) = 1/(1-p)$. It is also straightforward to show that for RBP, $\mathrm{L}_\mathrm{RBP}(i) = \mathrm{W}_\mathrm{RBP}(i)$, meaning that for this metric there is a further interpretation possible – the score assigned to a ranking is numerically equal to the expected relevance of the last document inspected [5].

The proposal by Moffat and Zobel explicitly connects an effectiveness metric based on a convergent infinite distribution with a user model that does not limit the depth to which documents may be accessed. But, while it gives better top-weightedness behavior than does precision, RBP is "stateless", in that the user is envisaged as having exactly the same behavior at depth 100 in the ranking as at depth 1, and the same behavior after observing a relevant document as after observing an irrelevant one.

Inverse Squares Moffat et al. [17] propose the use of a different convergent sequence. Their *inverse squares* metric INSQ is parameterized by a value T, the target number of relevant documents the

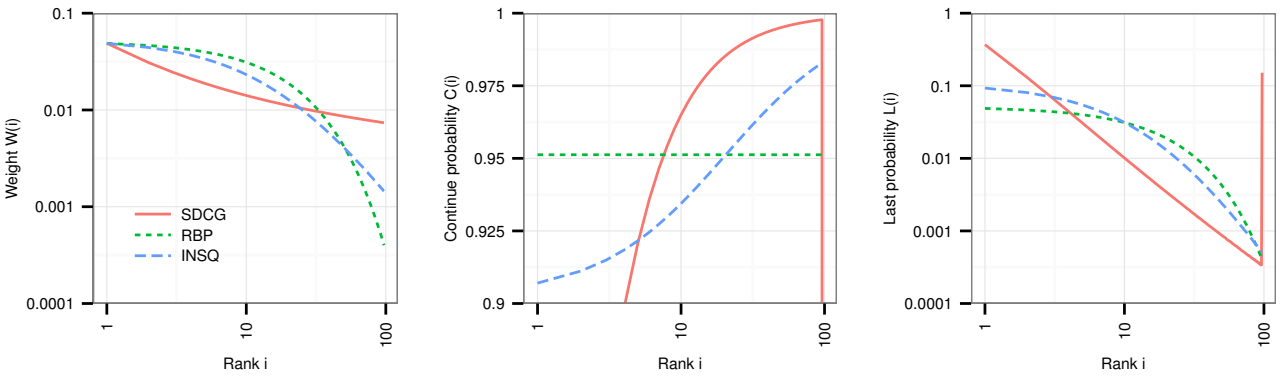


Figure 1: Weights $W(i)$, conditional continuation probabilities $C(i)$, and last probabilities $L(i)$, for weighted-precision metrics INSQ with $T = 10$; SDCG@k with $k = 97$, and RBP with $p = 0.9512$. The latter two parameters are chosen so that $W_{INSQ}(1) = W_{RBP}(1) \approx W_{SDCG}(1)$, giving the same expected number of documents accessed for all three metrics, approximately 20.5. Note the differing vertical scales.

user wishes to identify, and is defined by

$$C_{INSQ}(i) = \frac{(i+2T-1)^2}{(i+2T)^2}; \text{ which then leads to}$$

$$W_{INSQ}(i) = \frac{1}{S_{2T-1}} \cdot \frac{1}{(i+2T-1)^2}, \tag{4}$$

with $S_m = (\pi^2/6) - (\sum_{j=1}^{m} 1/j^2)$. Moffat et al. [17] note that the expected number of documents processed for INSQ is approximately $2T+0.5$, and that T serves the same role as the parameter p associated with RBP. The RRG metric identified by Carterette [5] also uses an inverse squares weight distribution, but does not have an equivalent of the parameter T.

Figure 1 compares $W_M()$, $C_M()$, and $L_M()$ for SDCG, RBP, and INSQ. The parameters (respectively: k, the SDCG truncation depth; p, the RBP persistence; and T, the INSQ target) are chosen so that all have the same weight $W_M(1)$, and hence all have the same expected value for the number of items inspected by the user.

2.2 Adaptive User Models

Metrics have also been defined in which the user's path through the ranking is *adaptive*, and affected by the relevance of the documents that they see at each inspection. (Note that some authors refer to static models as being "positional", and to the adaptive models described in this section as being "cascade".)

Reciprocal Rank This metric is defined as:

$$C_{RR}(i) = \begin{cases} 1 & \text{if } r_i < 1 \\ 0 & \text{if } r_i = 1\,. \end{cases} \tag{5}$$

That is, RR computes the average precision across the documents down to, and including, the first fully relevant one. In the case of binary relevance judgments, if that first relevant document appears at depth d, then RR $= 1/d$. The corresponding user model is also straightforward: users sequentially examine documents until a fully relevant one is identified, and then end their search.

Average Precision In the case of binary relevance judgments the metric *average precision* is the average of the $R = \sum_{i=1}^{N} r_i$ precision scores attained at the locations in the ranking at which relevant documents appear. Average precision can also be expressed as a weighted precision metric by attributing to each relevant item the total contribution it makes. For example, if a ranking has relevant documents at depths 2, 5, and 6 (only), then there are $R = 3$ relevant documents in total, and the AP score is $(1/2+2/5+3/6)/3 = 0.467$. But the components of that score can be striped across the relevant items that contributed, with ranks 1 and 2 assigned $W_{AP}(1) = W_{AP}(2) = (1/2+1/5+1/6)/3 = 0.289$, and so on. Table 1 completes this example, and adds conditional continuation probabilities $C_{AP}()$ and last probabilities $L_{AP}()$.

i	1	2	3	4	5	6
r_i	0	1	0	0	1	1
$W_{AP}(i)$	0.289	0.289	0.122	0.122	0.122	0.056
$C_{AP}(i)$	1.000	0.423	1.000	1.000	0.455	0.000
$L_{AP}(i)$	0.000	0.577	0.000	0.000	0.231	0.192

Table 1: Average precision as a weighted precision metric. The weights $W_{AP}(i)$ depend on the relevance vector $\vec{\mathbf{r}}$ as described in the text; then $C_{AP}(i)$ and $L_{AP}(i)$ are derived from $W_{AP}(i)$.

Hence, in our terminology AP can be specified as:

$$C_{AP}(i) = \begin{cases} \frac{\sum_{j=i+1}^{\infty}(r_i/i)}{\sum_{j=i}^{\infty}(r_i/i)} & \text{if } \sum_{j=i+1}^{\infty}(r_i/i) > 0 \\ 0 & \text{otherwise}\,. \end{cases} \tag{6}$$

The user model that corresponds to AP, described by Robertson [19], suggests that the user selects one of the relevant documents at random, and then examines every document down to and including that one in the result listing. In the form shown in Equation 6, AP models the user as always knowing how many relevant documents remain in the as-yet-unseen part of the ranking beyond depth i, and also what locations they are in. That is, the user model corresponding to AP is plausible only if the user can be assumed to base their decisions on documents that they have *not* yet seen, rather than on documents that they have.

Equations 5 and 6 indicate that the user continues down the ranking until at least one relevant document is found, while allowing no possibility that a user might exit from their search prior to finding even a single relevant document. As an extreme, neither metric is defined for rankings that do not contain any relevant documents. Despite being adaptive in terms of responding to the relevance of the result listing, neither RR nor AP have the flexibility to cope both with situations in which a single answer document is required (navigational queries) and situations that require many documents to be identified (informational tasks).

Initial expectation	Answer occurrence observed after query issued		
	No answers	Some answers	Many answers
Few answers (navigational) $T \approx 1$	Quickly dissatisfied; early reformulation	Possibly satisfied without needing reformulation	Satisfied quickly; no reformulation
Many answers (informational) $T \gg 1$	Dissatisfied; but will have looked down ranking before reformulating	Partially satisfied; will reformulate after looking down ranking	May be satisfied after first query; if not, will reformulate

Table 2: Hypothesized user search behavior, influenced by two factors: the anticipated number of answers required (T), and the rate at which relevant documents are identified while searching. (Adapted from Moffat et al. [17].)

3. WHAT DO WE THINK USERS DO?

There are several metrics/models described above, either explicitly, or via definition of one or more of $\mathrm{W}(i)$, $\mathrm{C}(i)$, and $\mathrm{L}(i)$; and many more in the literature. It is then reasonable to ask how to choose one over the others. There are several options: we can rely on rhetoric; we can compare outcomes (do they agree with each other? which is more stable? which is more sensitive?), or we can ask about the fidelity of the model.

In the rest of the paper we do the last of these, and ask: how well do the models corresponding to various IR metrics match real behavior? In particular, is there a "right" formulation for $\mathrm{W}(i)$, $\mathrm{C}(i)$, and $\mathrm{L}(i)$? We tackle this question in two ways: first, by listing a set of possible user behaviors and asking if there is a model that encapsulates them all; and then, in the next section, by studying users carrying out search tasks.

Possible User Behaviors There is a range of reasonable hypotheses about user behavior, which should be captured in any user model (and hence metric). The pattern of behavior suggested by these hypotheses is summarized in Table 2, adapted from Moffat et al. [17].

1. *Users undertake searches for different reasons. Some searches start with the goal of identifying a single answer, and others with the goal of finding multiple answers.* This suggests a model must allow the setting of a different number of target documents, a flexibility not offered by RR and AP, but that is allowed for via the various parameters that govern Prec@k, SDCG@k, RBP, and INSQ.

2. *Users may wish to examine documents to arbitrary depth in the ranking, albeit with decreasing probability.* This suggests that $\mathrm{C}(i)$ should never go to zero. The models for Prec@k and SDCG@k do not comply with this goal.

3. *All other factors being equal, users may be more likely to continue their search the more effort (or time) they have invested into it.* If this hypothesis is correct then, all else being equal, $\mathrm{C}(i)$ should tend to go up as i increases. Both Prec@k and RBP fail this expectation.

4. *Users may alter their behavior based on the part of the ranking that has been inspected.* This suggests that none of the static models of Section 2.1 are adequate.

5. *Users may exit from their query without having (fully, or even partially) satisfied their information need.* As already observed, the models behind RR and AP do not support this.

In summary, all of the common weighted-precision metrics can be criticized in some way or another when they are weighed up against hypothesized user behavior.

4. DO USERS DO WHAT WE THINK?

The five suggested behaviors listed above appear plausible, but are just hypotheses. In this section we describe a user study that examined user behavior (including gaze data) on a number of search tasks, seeking to gather evidence for or against them.

4.1 Experimental Design

Subjects were presented with a set of six information need statements (see Table 3), plus one warm-up task (not shown), and asked to use a search engine to find and mark documents that would help them answer the questions. All interactions were undertaken using an instrumented interface that both limited user actions in certain ways, and allowed detailed logging that included click actions and query reformulations. Ethics committee approval for this project was granted at RMIT University.

Queries were executed via the Yahoo! API, but answer pages were presented to the subjects without any branded identification. Documents could be viewed via a pop-up window that obscured and de-activated the main search listing until it was closed again. To close the document pop-up, subjects were required to indicate whether viewing that document was "useful" or "not useful". Documents that for some reason were redisplayed collected a second subject-generated relevance judgment. We do not have relevance labels for documents which users did not view; we aimed to disrupt natural behaviors as little as possible, and a participant who is asked to label every document may well process a result list differently.

On closing a document, the result listing was redisplayed, with a brief color-coded highlighting of the link that had just been viewed (green for "useful", red for "not useful"). Users were free to access further result pages for each query, and to issue fresh queries for the task, but were not able to open pages in browser tabs, or to open new windows. In addition to the instrumented browser logging, gaze-tracking hardware was used throughout each trial, so that implicit user interactions could also be captured.

The experimental sessions proceeded as follows. First, the participant was asked to complete a brief demographic survey. Next they were shown descriptions of some information seeking tasks, similar to those used later in the operational part of the session, and asked to estimate the number of useful documents they thought they would need to find in order to address the information need. The answers to these questions provided a basis for estimating T.

Once a participant had completed the survey, they embarked on the operational study, and after working through the warmup query, were shown the first of six information needs. Participants were expected to complete each task before moving on to the next one, and were required to make their own decision as to when a task was "done", with no time limit applied. Their instructions were:

You will spend approximately one hour doing a sequence of seven web search tasks. For each task, you'll be given a question and you should use our search engine to help answer it. ... There is no time limit on each of the tasks, and no minimum time limit overall either.

	Information specification	Starter query
1	(*remember*) You recently watched a show on the Discovery Channel, about fish that can live so deep in the ocean that they're in darkness most or all of the time. This made you more curious about the deepest point in the ocean. What is the name of the deepest point in the ocean?	deepest ocean point
2	(*remember*) You recently attended an outdoor music festival and heard a band called Wolf Parade. You really enjoyed the band and want to purchase their latest album. What is the name of their latest (full-length) album?	wolf parade
3	(*understand*) Your nephew is considering trying out for an Australian Rules football team. His parents are supportive of the idea, but you think the sport is dangerous and are worried about the potential health risks. Specifically, what are some long-term health risks faced by football players?	australian rules football health risks
4	(*understand*) You recently became acquainted with one of the farmers at the local farmers' market. One day, over lunch, they were on a rant about how people are ruining the soil. They were clearly upset, so you're interested in finding out more. What are some human activities that degrade soil fertility?	damage soil fertility
5	(*analyze*) Your sister is turning 25 next month and wants to do something exciting for her birthday. She is considering some type of extreme sport. What are some different types of extreme sports in which amateurs can participate? What are the risks involved with each sport?	extreme sport
6	(*analyze*) You recently heard someone claim that identity theft in Australia is on the rise. This has made you concerned about protecting your own identity. How easy or difficult is it for a stranger to open a credit card under your name? What essential information about you is needed and what are some effective ways in which you can protect your identity in the future?	identity theft and credit cards

Table 3: Search tasks used in user experiments, together with task complexity category (shown in parentheses) and associated starter query.

So spend what feels to be an appropriate amount of time on each task, until you have collected a set of answer pages that in your opinion allow that information need to be appropriately met, and then move on to the next task.

Search tasks are of differing levels of complexity, and it is reasonable to expect that user behavior differs too. The six topics used in the study were modeled on the classes and topics proposed by Wu et al. [25]; although users were not able to choose topics freely, this allowed us to control task complexity. In particular, participants were presented with two topics in each of the *remember*, *understand*, and *analyze* categories, representing tasks of increasing levels of cognitive complexity. The first page of results shown to each user for each topic was generated by a uniform "starter query", also shown in Table 3.

Task order was a controlled variable for each participant, presented in a Graeco-Latin permuted order. A second controlled experimental variable was the quality of the search results; for half of the searches, the list of answers returned from the commercial service was interleaved with related-but-incorrect snippets [15]. In this paper we only consider data obtained for searches in the first, unadulterated, half.

Gaze records from the tracker were first reduced to *fixations* – series of records lasting at least 75ms and within a 5-pixel radius – to remove saccades and glances too short to indicate reading. Sequences of fixations on the same snippet, with no intervening clicks, were then further amalgamated. These records were combined with logs from our search software and logs from the browser to produce a complete record of each user's interactions.

4.2 Demographics

A total of 37 participants were recruited, consisting of research students and staff from the Australian National University. Due to eye-tracking calibration and recording quality issues, the data of three individuals could not be included for analysis. Of the remaining $n = 34$ participants, 8 were female and 26 were male, with an average age of 26 years. All were fluent in English, although 50% indicated that it was not their first language. The participants all held or were working towards degrees in the areas of computing, engineering, information science or mathematics. There was a high level of familiarity with searching across the participants: all indicated that they carry out a search using a web search engine several times a day, with a median 11 years of experience with online searching. No participants indicated that they were color-blind.

4.3 Measurements

We now present some of the collected data. Section 5 analyzes what this data implies in terms of the five conjectures listed earlier.

Estimating *T* Before carrying out any searches, participants were shown three sample information need statements – one of each of the three task types *remember*, *understand*, and *analyze*, but tasks which were not used in the remainder of the experiment, to minimize anchoring effects – and asked to respond to the statement: "*I'd expect to need to find* $\boxed{nn}$ *useful web pages to answer this*". The distributions of responses for T are shown in Figure 2a. Given the increasing complexity of the task types, we expected that the estimated number of needed documents would increase from top to bottom. However, the responses did not follow this trend, and the only pairwise difference that was significant was between the *understand* and *analyze* categories. The deviation from the expected outcome may be a consequence of the particular three example queries that were shown, or of the fixed ordering in which they were presented (*understand*, *analyze*, then *remember*). Followup experimentation is required in which a broader palette of scenarios is provided, and presented in a varied ordering.

Details of the documents that participants saved as being "useful" were also collected. Figure 2b shows the distribution of documents saved per user, by task type. While the anticipated trend that more documents would be needed for tasks of increasing complexity is present, the actual numbers are small. Even for the two *analyze* tasks, the number of documents saved to "allow that information need to be appropriately met" is relatively low.

User Gaze Behavior A key assumption in our discussion has been the broadly accepted claim that users scan search result pages from top to bottom, viewing snippets $1, 2, 3, 4$, and so on. Figure 3a plots the mean first arrival time, measured by the number of previous views of other ranks, at each of the top 10 rank positions, following the methodology of Joachims et al. [14], and shows that on average the first viewing of the snippet at rank i is indeed correlated with

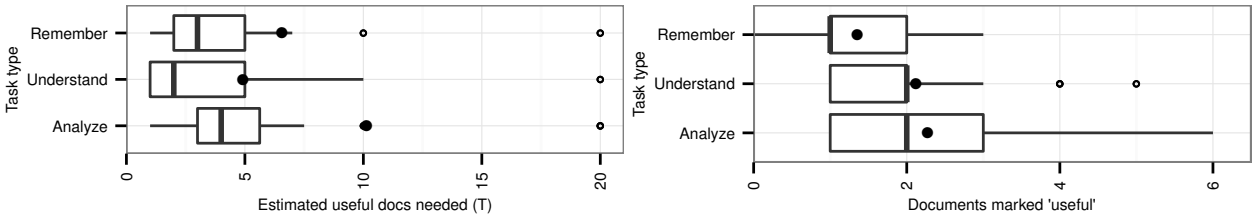
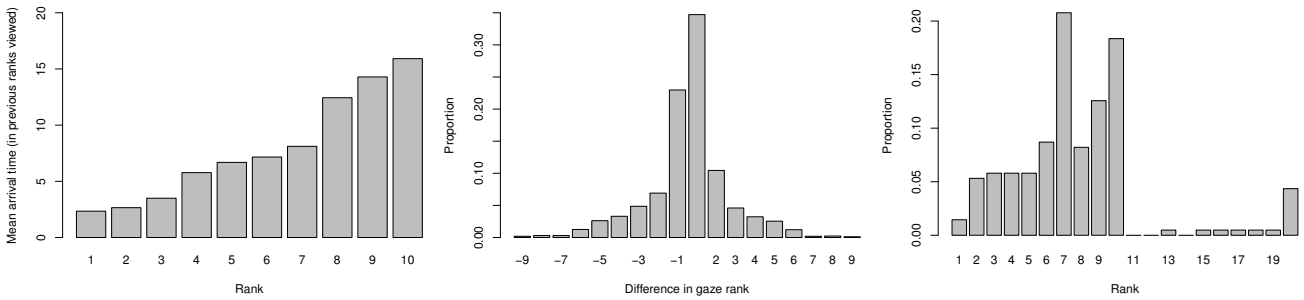


Figure 2: Subject-estimated T, by task type (left), and documents actually marked as useful (right).



(a) Mean arrival time at each result rank (b) Jumps between consecutive fixations (c) Rank of final fixation

Figure 3: Aggregated gaze behaviors, across users and tasks.

i. This is consistent with previous work [14], and has often been interpreted as evidence that users scan links from top to bottom.

However, that outcome needs to be treated carefully. The gaze position of any individual user is much more volatile than Figure 3a would suggest, and the fixation point both sometimes moves backward and sometimes advances by more than one: a viewing sequence might well be $1,2,4,3,1,2$, for example. To quantify this tendency, the sequence of fixation points for each user was processed into a set of "jumps": $+1,+2,-1,-2,+1$ for the same example. Figure 3b shows the resulting distribution. Jumps of $+1$ (one step down the ranked list) dominate, but a significant fraction of the fixation shifts are also by -1 and $+2$ and larger jumps also occur.

We also see effects due to screen layout. Figure 3c shows the distribution of last-viewed ranks (that is, the ranks viewed just before a search ended). There are distinct peaks at ranks 7 (the bottom of the screen) and 10 (the bottom of the first page). Taken together, there is evidence for a wide variety of reading behaviors. Reading top-to-bottom is common, but not universal, and there are definite discontinuities. We have investigated this further in other work [23].

Continue Probability Our primary purpose in the experimentation was to explore factors that affected $\mathrm{C}(i)$, the user's probability of continuing their inspection of documents after viewing the document at position i in the ranking.

To determine an experimental value for $\mathrm{C}(i)$, a "did continue" indicator variable DC was associated with each snippet fixation in the gaze log, taking the value zero if this was the last snippet viewed for this result page, and the value one if it wasn't. For example, for the sequence of snippet views $1,2,5,1,2,4,2$ the inferred set of DC observations, categorized into groups according to rank position, would be $\mathrm{DC}(1) = \langle 1,1\rangle$, $\mathrm{DC}(2) = \langle 1,1,0\rangle$, $\mathrm{DC}(3) = \langle\rangle$, $\mathrm{DC}(4) = \langle 1\rangle$, and $\mathrm{DC}(5) = \langle 1\rangle$. Laplace smoothing was then applied to allow empirical values to be estimated; for the same example data, to compute $\mathrm{C}_{\mathrm{est}}(1) = 3/4 = 0.75$, $\mathrm{C}_{\mathrm{est}}(2) = 3/5 = 0.60$, $\mathrm{C}_{\mathrm{est}}(4) = 2/3 = 0.67$, $\mathrm{C}_{\mathrm{est}}(5) = 2/3 = 0.67$.

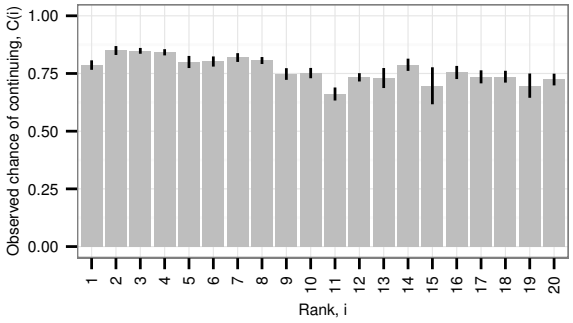


Figure 4: Observed $\mathrm{C}(i)$, averaged across queries and users.

Figure 4 plots estimated $\mathrm{C}(i)$, averaged over all users and all queries. When presented this way, $\mathrm{C}(i)$ appears almost constant, with a value of approximately 0.75. However, this seemingly-consistent gross behavior is an amalgam of many contributing factors. The next section examines the composition of $\mathrm{C}(i)$ in detail.

5. DO USER MODELS MODEL USERS?

Section 2 noted that weighted precision metrics, both static and adaptive, can be specified by any of the interchangeable functions $\mathrm{W}(i)$, $\mathrm{C}(i)$, or $\mathrm{L}(i)$. The closer $\mathrm{C}(i)$ matches observed user behavior, the more confident we can be in the corresponding metric.

For the users and tasks measured, Figure 4 shows $\mathrm{C}(i)$ to be approximately constant when aggregated across ranks. If that is really the case, then the simple model behind RBP is applicable. If not, what other factors explain the variation? For example, if $\mathrm{C}(i)$ varies only with rank, the static model behind a metric like SDCG@k or INSQ may be useful for evaluation. On the other hand,

Factor	Effect
(intercept)	11.70
User	0.11–10.95
Proportion of T collected	0.34
Proportion of docs viewed that are relevant	0.50
Gaze sequence	0.97
Rank i	1.06
Query count, in task	1.10

Table 4: Factors in a fitted model of DC. Users become more persistent (have higher $C(i)$) as they issue more queries and look further down each result page; and become less persistent as they accumulate relevant documents.

if $C(i)$ varies with the relevance of each document, we might be better off with an adaptive model such as that of RR or AP.

5.1 Explaining C(i)

We used logistic regression to model DC, the fixation-by-fixation continuation indicator variable, as a response to a number of potential explanatory variables including indicators of user, task, task type, and search progress. Model selection was performed by considering specifications between the full model (including all explanatory variables) and the minimal model including only the intercept. The final model was selected to minimize the Akaike information criterion (AIC) [1], which combines the likelihood of the model with a penalty for each term which is included. Evaluation was carried out using R's `stats::step.glm` method. The full list of variables evaluated was: user, task, task type, gaze sequence, gaze rank i, judged relevance of the current document r_i, number of relevant documents found, relevant documents viewed as a proportion of documents viewed, proportion of T collected, amount of T remaining, and query number within the task. We used participants' own estimates of T, according to the task type (see Section 4.3 and Figure 2a).

Table 4 summarizes the factors in the built model. The column labeled "Effect" is the coefficient assigned to each of the listed factors in terms of the odds of continuing: for example, a user currently at rank $i+1$ is 1.06 times more likely to continue to the next document than a user currently at rank i. Effects greater than one represent an increased chance of continuing, that is they increase $C(i)$; effects less than one decrease $C(i)$ as the corresponding factor increases. The outcomes in the table lead to several observations.

First, there is a large effect due to user – some users are simply more likely to keep reading than are others. The variability due to user is large, with a base value for $C(1)$ varying from 0.57 to 0.99.

Second, when per-user variance is allowed for, there are two strong effects arising from the relevance of the documents already seen. The odds of a user continuing decrease sharply as they accumulate relevant documents towards their target T. By the time a user has seen as many relevant documents as they thought they would need, the odds of their continuing have dropped by two thirds. Carterette et al. [6] note that user patience – the p parameter in RBP – varies with task type; the results here are similar, but show dependence not on task type in isolation, but rather, on the user's notion of how much information they need.

The left-hand graph of Figure 5 illustrates this effect. It plots $C(i)$ as the proportion of T gathered is varied, and all other factors are held constant. The lines are estimates from the model for three hypothetical users chosen to match the median, first-quartile, and third-quartile (among the experimental subjects) of the base $C(1)$ values. The shaded areas mark a 95% confidence interval for $C(i)$.

Model	Features	Δ_{AIC}
RBP-like	Intercept only	42
SDCG-like	Intercept plus i	51
RR-like	Intercept plus r_i	37
Best learned	As per Table 4	0

Table 5: Quality estimates for models of DC from three representative families, plus the model developed in Table 4. The values listed for Δ_{AIC} are the difference in AIC values relative to the set of factors listed in Table 4; lower values represent better models.

Note that the logistic regression model estimates the change in odds; the reason that the factors have such different effects on the $C(i)$ of particular users is therefore due to the different starting probabilities. A similar effect, but less pronounced, arises in connection with the proportion of documents viewed which have been judged useful (recall that in our protocol, every document which was viewed was also judged). Users reading "information-heavy" rankings have a reduced $C(i)$, as was anticipated in Table 2.

Third, there are two effects due to query behavior. As users look at more results in each query, they are slightly less inclined to continue (factor "gaze sequence", with effect 0.97). Counteracting this, the users were more inclined to continue the more queries they issued against a particular task (effect size 1.10).

Finally, there is indeed a component of $C(i)$ attributable to i. The effect is slightly above 1.0, meaning that all other things being equal, users are a little more likely to keep reading from a deeper result than from a shallower one – perhaps deeper results are not as good, and users are likely to look back to a better result before finishing. The effect due to rank is illustrated in the plot on the right side of Figure 5, for the same three hypothetical users.

We did not observe an effect due to task: that is, the task itself does not seem to affect $C(i)$, except indirectly via the participants' estimates of T. Models which included task type and task instance, as well as the combination, did not improve on the model in Table 4. In fact using those factors alone did not improve on the simplest model which holds $C(i)$ constant.

The effects of query count, and those due to relevance, accumulate over a search session. Given a better understanding of users, and more extensive data, we might expect that other effects of this kind could exist – spanning result pages, queries, and longer durations. Session-level or longer-term effects are not included in the user models behind any of the commonly-used metrics.

5.2 Metrics and Models, Revisited

The model summarized in Table 4 is interesting, but also more complex than the models in Section 2. It is important to ask whether it is needed, or if the simpler models suffice. To resolve this question, models were also constructed for each of three simple classes, representing three families of effectiveness metrics.

Table 5 compares the fit of these models, based on the Akaike information criterion (AIC). The values listed as Δ_{AIC} are the differences between the AIC value for each model and the AIC value for the best learned model (Table 4): higher values indicate a less parsimonious model, amongst the set of models presented. Differences above about 10 can be interpreted as indicating a model having "essentially no empirical support" [4], so we can have a great deal of confidence that the fitted model improves on the alternatives.

The "RBP-like" model computed in this way just assigns $C(i) = 0.93$, and it is the simplest model. The static "SDCG-like" model allows $C(i)$ to vary as a function of i, but with a different effect

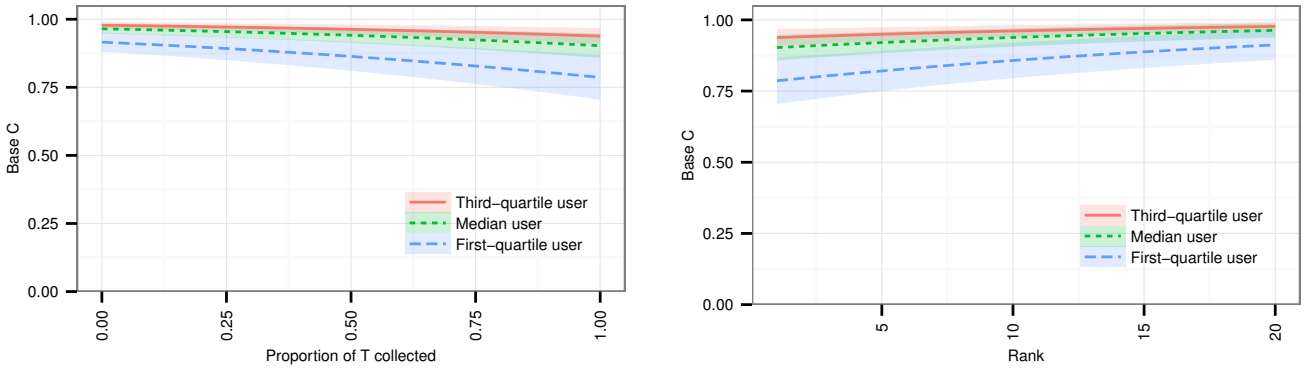


Figure 5: Illustrative effects of relevance and rank on $C(i)$ for three hypothetical users. Except for proportion of target relevance T (on the left) and depth i in the ranking (on the right), all other factors are nominally constant. The shaded area is the 95% confidence interval for $C(i)$.

allowed for each i so that $C(i)$ can take any shape. This differs from SDCG itself, but is more flexible: in particular, it allows for discontinuities at the fold and the end of the page. This, however, produces a relatively poor fit, worse that the RBP-like approach. Finally, the adaptive "RR-like" model allows $C(i)$ to vary with r_i, and this produces the lowest AIC (that is, the best fit) of the three families. However, none of the three are as powerful as the learned model. (It was not possible to produce an "AP-like" model, as Equation 6 requires relevance judgments of all documents in the result set including those which are never viewed. Post-hoc judgments were not performed, so the required data was not available.)

Since per-user effects are so large, we also fitted variants of the RBP-like, SDCG-like, and RR-like models which included a feature for user. This was not sufficient to improve the three models ($\Delta_{\text{AIC}} = 51$–52).

The AIC-based analysis confirms that better fidelity could be attained by using not just adaptive models, but models which expressly allow for differing relevance targets T, and which adapt their behavior as T is approached. Indeed, adding "proportion of T" as a factor to the RBP-like model explains DC better than the rank-varying SDCG-like or the relevance-based RR-like models ($\Delta_{\text{AIC}} = 28$).

5.3 What We Think Users Do, Revisited

We can also ask whether the model of Table 4, which was built from observations of real users, aligns with the five hypotheses of Section 3, paraphrased as:

1. *Users search for different reasons, with different targets.* In the learned model, $C(i)$ relies in part on T, which is the user's target number of documents.

2. *Users may wish to examine documents to arbitrary depth.* In the learned model, $C(i)$ approaches but never reaches zero – there is always a chance a user will read on.

3. *Users may be more likely to continue the more they have invested.* The evidence for this is mixed. $C(i)$ increases as more queries are issued and as a user reads deeper; however it decreases slightly as more results are read.

4. *Users may alter their behavior based on what they read.* This is certainly consistent with the learned model, where seeing relevant documents reduces $C(i)$ via two factors, proportion of T collected, and fraction of documents seen which are relevant.

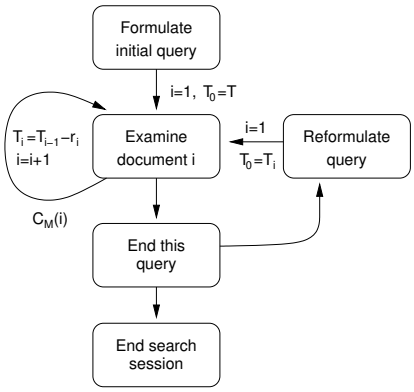


Figure 6: Proposed user model for search. For a given metric M, the quantity $C_M(i)$ is a function of T_i and i.

5. *Users may exit from their query at any time.* Again, $C(i)$ approaches but never reaches unity – there is always a chance a user will bail out.

6. A BEHAVIOR-DRIVEN METRIC

To complete the development, we describe a metric which is inspired by the hypotheses and observations above. We do not suggest that this is the last word – it merely represents one more position in spectrum between simplicity and fidelity. It does, however, embody all five of the user behaviors that were hypothesized in Section 3, and in that sense can be regarded as providing proof-of-concept.

As noted in Section 2, the user model associated with INSQ (Equation 4) meets the first three listed conjectures. In particular, the parameter T in Equation 4 captures conjecture 1, and makes INSQ intent-sensitive, since T can be thought of as being an estimate of the number of relevant documents the user seeks to acquire. For a navigational query, $T = 1$ or $T = 2$ perhaps; and for an informational query, $T = 5$ or $T = 10$ might be more appropriate.

Figures 6 and 7 (the latter derived from Smucker and Clarke [21]) propose an extended model for INSQ that makes the metric adaptive (conjecture 4) and session-based (conjecture 5). When a user examines an answer at rank i in the results list, they first read the snippet. Based on the snippet, a decision is made: to click through and read the underlying document, or to not click. In the latter case, the user is finished with the document. Importantly, from the user's point of view, this means that the document is non-relevant ($r_i = 0$).

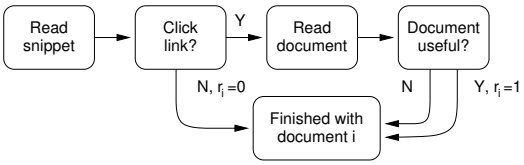


Figure 7: The "Examine document *i*" step in Figure 6.

If the user clicked, they then read the full document, leading to two possible outcomes: it is useful, in which case $r_i = 1$; or it is not, in which case $r_i = 0$. There are two key differences between this proposal and the similar state diagram presented by Moffat and Zobel [18]. The first is, as already discussed, the alteration of the conditional continuation probability at depth i from being a fixed value p (used in RBP) to a variable value $\mathrm{C_M}(i)$; the second is that we introduce the possibility of the user reformulating a revised or substitute query as part of the same search session.

Figure 6 proposes that T be modified as the search proceeds. The user enters the ranked list with $T_0 = T$, their initial intention; but as documents are viewed, their information need is partially or fully satisfied, and their intention evolves. We propose that this evolution be modeled by computing

$$T_i = \max\{0, T - \mathrm{Rel}(i)\}\,,$$

where $\mathrm{Rel}(i) = \sum_{j=1}^{i} r_j$; that is, T_i estimates the "current unmet demand for relevance" after inspecting i documents in the ranking.

In a general metric it is probably not feasible to have a parameter per user, but we do want to incorporate the most important of the user-independent factors from Table 4: the proportion of T collected. Call this $T_{\mathrm{prop}} = \mathrm{Rel}(i)/T$: now as T_{prop} increases, $\mathrm{C}(i)$ should decrease. Equivalently, as $1 - T_{\mathrm{prop}}$ decreases so should $\mathrm{C}(i)$.

To add this notion to INSQ, and thereby create an adaptive version, we use $T + T(1 - T_{\mathrm{prop}})$ rather than $2T$ to compute the conditional continuation probability. Since $1 - T_{\mathrm{prop}} = T_i/T$, we have:

$$\mathrm{C}'_{\mathrm{INSQ}}(T, T_i, i) = \frac{(i + T + T_i - 1)^2}{(i + T + T_i)^2}\,. \qquad (7)$$

The effect is that initially, $T_i = T$ and the model is as in Equation 4. While T_i remains high, so does the conditional continuation probability; then, as T_i decreases and the user's information need is increasingly satisfied, so the likelihood of continuation also decreases, and the expected length of the remaining search decreases. If no relevant documents are accumulated, Equation 7 remains the same as Equation 4.

The probabilistic nature of searching using this model means that the user might exit their search before T_i reaches 0. If this happens, they can be expected to reformulate their query and start inspecting the new ranked list, or switch to a different search service, or just quit. In the case of the search being continued, we suggest that their initial T_0 for the follow-on query is inherited from the T_i value that was attained at the time the previous query was abandoned. The diagram in Figure 6 captures this aspect of the proposed model.

Clearly this is only one of any number of metrics which might be developed based on the model in Table 4 and pending further investigation we offer it as a proof-of-concept. Nevertheless a representative of this metric, modelling DC with i and T_i/T, achieves a Δ_{AIC} of 28: this is substantially more accurate than others in Table 5.

7. RELATED WORK

There is a considerable literature in regard to effectiveness metrics, and in the limited space available, we can at best present a brief overview. The current definition of AP emerged in the 1990s as an evaluation measure associated with the deep rankings (often to depth $k = 1{,}000$) in the TREC project. Järvelin and Kekäläinen [13] then introduced the DCG and NDCG metrics (the latter not considered here), arguing that explicit top-weightedness was preferable to the way it was achieved in AP, and formalizing the notions of graded (that is, non-binary) relevance judgments, and of "gain" as a benefit received by the user. Moffat and Zobel [18] followed up with their RBP proposal and corresponding user model; an important recognition in this work is that it is preferable for the metric to assess the rate at which gain is accrued, rather than the total magnitude of gain that is accrued. Zhang et al. [27] considered a range of static weighted-precision metrics including SDCG and RBP, and show that RBP with $p = 0.73$ is a good match to the normalized document viewing characteristics inferred from the click densities arising from commercial search operations.

A flurry of activity has taken place over the last four to five years. Craswell et al. [9] proposed a *cascade* model of click behavior (which we refer to as "adaptive" here), which posits that users read search results from top to bottom, and at each rank make a decision of whether to click on that document, or to skip it. Once a single document is clicked, the user exits from the search. Unlike previous models of click behavior, the cascade model takes the relevance of answer items that are higher in the results list into account. Craswell et al. demonstrated a closer fit to click data from a commercial search engine than other models of user behavior, such as a position model (where click behavior depends only on the rank of a document, called "static" here), or an examination model (where clicks are a function of rank as well as document-specific factors). Chapelle and Zhang [8] extended the cascade model using a dynamic Bayesian network, allowing for the possibility that a user is not satisfied after a single click and returns to the search results list. In related work, Chapelle et al. [7] further argued that the history of what the user experiences as they process the answer list affects the way they address the remainder of the list. They synthesized a new metric ERR by combining the geometric distribution used in RBP with the approach embedded in RR, adapting to r_i and also, via the choice of p, allowing persistence to be tailored to the search requirement. Yilmaz et al. [26] also explored metrics in which the probability of continuing the inspection of documents is conditional on the relevance level of the last document inspected.

Carterette [5] categorized a wide range of effectiveness metrics, grouping them into four classes; in doing so, the relationships between weights, halting probabilities, and last viewed probabilities was raised, an understanding that we have also employed in this work. Carterette then explored the implications of the classification using a range of click and TREC data, concluding that despite the fact that it is a non-convergent sum, DCG has a range of merits.

Recently Smucker and Clarke [20, 21] have measured the time taken by users to inspect documents, and argued that a more precise unit of "investment" against which utility is assessed should be search time, rather than documents examined. In a user study of search behavior, Smucker and Clarke demonstrate that short documents require less inspection time than do long ones, and that repeated documents can be evaluated very quickly. Based on these, and other factors, they proposed *time-biased gain* as an effectiveness metric, and argued that it better reflects user search behavior.

In other interaction studies, Joachims et al. [14] examined the way in which user gaze fixations can be associated with result listings; Al-Maskari et al. [2] (see also Huffman and Hochster [12]) questioned the usefulness of deep evaluation metrics, and found that shallow metrics such as Prec@10 provide better correlation with the experience reported by users; Turpin and Scholer [24] raised doubts

about the usefulness of AP to predict user task-completion ability; and Thomas et al. [22] examined the numeric stability of static metrics when applied to perturbed or degraded rankings; they also note that page boundaries can be handled by altering the continuation probabilities at appropriate intervals.

Other authors have considered user models for metrics in more abstract settings: as already noted, Robertson [19] described a user model associated with AP, with Dupret and Piwowarski [11] providing elaboration; Dupret [10] has also examined alternative user models for the DCG metric.

8. CONCLUDING REMARKS

Metrics such as Prec, AP, and RR instantiate user models, implicitly or explicitly, and these models can be described by any one of the interchangable functions $\mathrm{W}()$, $\mathrm{C}()$, or $\mathrm{L}()$. This leads us to ask: what do these functions look like for real users? Is one model (or family of models) more or less accurate than others, and can that tell us anything about the associated metrics?

We conducted a user study with the goal of identifying the factors that contribute to $\mathrm{C}(i)$, a user's propensity to continue searching after viewing a document at rank i. Factors we thought would be of possible influence included the rank position i; the relevance r_i of the document in that position; the amount of relevance collected until that point in the viewing sequence; and the fraction of the target relevance that had been collected. Our study of 34 users, each completing six search tasks, provided evidence that all of these factors are indeed contributors to the decision made by the user after viewing each snippet, namely whether to continue or stop.

Unsurprisingly, our investigation has also generated fresh questions to be considered. First, Figure 6 contains provision for query reformulation and/or query reissue. But our sample of user activity contained too few such events for us to try and derive meaningful insights as to what factors drive these decisions, and we can only conjecture that the unanswered information need, $T_i = T - \mathrm{Rel}(i)$ is positively correlated with the conditional probability of a reformulated query being issued. For example, the conditional probability of reformulation, given that the user is not continuing down the ranking, might be modeled as being some kind of $T_i/(T_i+k)$ -like function, where k is a constant. A larger user study would be required for that conjecture to be explored.

Second, we have defined r_i in terms of the user's response to the document – if it was marked as being "useful" to answer the information need, then we regard the corresponding r_i value as being 1. But two users viewing the same ranking make different decisions about what to view, and even if they view the same documents, make different decision about whether to save it. Hence, r_i itself must be thought of as being non-deterministic, introducing a further degree of freedom into any user model. The "starter query" results may yield insights into this issue, and allow estimates to be made of the degree to which users differ.

Acknowledgments This work was supported by the Australian Research Council. We thank Dingyun Zhu for his help running the user experiments.

9. REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, 19(6):716–723, 1974.

[2] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *Proc. SIGIR*, pages 773–774, Amsterdam, The Netherlands, 2007.

[3] C. Buckley and E. M. Voorhees. Retrieval system evaluation. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3, pages 53–75. MIT Press, Cambridge, Massachusetts, 2005.

[4] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: A practical information-theoretic approach*. Springer, New York, 2nd edition, 2005.

[5] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proc. SIGIR*, pages 903–912, Beijing, China, 2011.

[6] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proc. CIKM*, pages 611–620, Glasgow, Scotland, 2011.

[7] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, Hong Kong, China, 2009.

[8] O. Chapelle and Y. Zhang. A dynamic Bayesian network click model for web search ranking. In *Proc. WWW*, pages 1–10, Madrid, Spain, 2009.

[9] N. Craswell, O. Zoeter, M. J. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proc. WSDM*, pages 87–94, Palo Alto, CA, 2008.

[10] G. Dupret. Discounted cumulative gain and user decision models. In *Proc. SPIRE*, pages 2–13, Pisa, Italy, 2011.

[11] 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.

[12] S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *Proc. SIGIR*, pages 567–574, Amsterdam, The Netherlands, 2007.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Information Systems*, 20(4):422–446, 2002.

[14] 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.

[15] T. Jones, D. Hawking, P. Thomas, and R. Sankaranarayana. Relative effect of spam and irrelevant documents on user interaction with search engines. In *Proc. CIKM*, pages 2113–2116, Glasgow, 2011.

[16] A. Moffat. Seven numeric properties of effectiveness metrics. In *Proc. AIRS*, 2013. To appear.

[17] A. Moffat, F. Scholer, and P. Thomas. Models and metrics: IR evaluation as a user process. In *Proc. Australasian Document Computing Symp.*, pages 47–54, Dec. 2012.

[18] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Information Systems*, 27(1):2:1–2:27, 2008.

[19] S. Robertson. A new interpretation of average precision. In *Proc. SIGIR*, pages 689–690, Singapore, 2008.

[20] M. D. Smucker and C. L. A. Clarke. Stochastic simulation of time-biased gain. In *Proc. CIKM*, pages 2040–2044, Maui, Hawaii, 2012.

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

[22] P. Thomas, T. Jones, and D. Hawking. What deliberately degrading search quality tells us about discount functions. In *Proc. SIGIR*, pages 1107–1108, Beijing, 2011.

[23] P. Thomas, F. Scholer, and A. Moffat. What users do: The eyes have it. In *Proc. AIRS*, 2013. To appear.

[24] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. SIGIR*, pages 11–18, Seattle, Washington, 2006.

[25] W.-C. Wu, D. Kelly, A. Edwards, and J. Arguello. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proc. 4th Information Interaction in Context Symp.*, pages 254–257, Nijmegen, The Netherlands, 2012.

[26] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *Proc. CIKM*, pages 1561–1564, Toronto, Canada, 2010.

[27] 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.