Models and Metrics:
IR Evaluation as a User Process

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

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

Paul Thomas
ICT Centre,
Canberra
CSIRO
paul.thomas@csiro.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. The former has the benefit of directly assessing the actual goal of the system, namely the user's ability to complete a search task; whereas the latter approach has the benefit of being quantitative and repeatable. Each given effectiveness metric is an attempt to bridge the gap between these two evaluation approaches, since the implicit belief supporting the use of any particular metric is that user task performance should be correlated with the numeric score provided by the metric. In this work we explore that linkage, considering a range of effectiveness metrics, and the user search behavior that each of them implies. We then examine more complex user models, as a guide to the development of new effectiveness metrics. We conclude by summarizing an experiment that we believe will help establish the strength of the linkage between models and metrics.

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. OVERVIEW

Information retrieval (IR) systems are measured in two quite different ways. The *efficiency* of an IR system is quantified in terms of CPU, memory and disk resources required, as functions of the volume of data in the system, the rate at which queries must be processed, and the semantics of each query (and hence the query processing modality), and is ultimately demonstrated by experimental measurement of an instrumented implementation. The *effectiveness* of an IR system – the ease with which users of the system can carry out information-seeking tasks and satisfy information needs – is a more subtle concept. We focus on the latter in this paper.

Two approaches to quantifying effectiveness have emerged over the years. The first is the use of what are generically called *user studies*, in which a pool of experimental subjects are given one or more search tasks to carry out, and their actions and behaviors during the prosecution of those tasks are monitored, analyzed, and reported. Provided that the information-seeking tasks used during the experiment are ones that the experimental subjects are able to empathize with, a well-designed user study can provide rich information about all aspects of the system being evaluated, including the interface ("why does it do that when I do this"), robustness ("wow, it crashed again"), and underlying system effectiveness ("it's a bit strange that it didn't put that one on the first page of results").

But user studies are expensive to plan and run, both in terms of actual money, and in terms of time. The planning is costly because of the need to fix all variables and then seek institutional ethics board clearance for a particular experiment, and then recruit subjects; and carrying experiments out is costly because of the need to provide supervision while the subjects are undertaking the specified search tasks. These costs mitigate against continuous user studies in all but the very largest of organizations.

Instead, a second type of effectiveness investigation is common. In a *batch evaluation* (or a *test collection* evaluation), a document collection is compiled; a set of topics or information needs is formulated that can be answered out of that collection; and some or all of the documents in the corpus are *judged* against the topics, to determine whether or not each such document is *relevant* to the specified topic. Those relevance judgments can then be repeatedly used to *score* the outputs of the IR system using a chosen *effectiveness metric* to convert the system outputs into a numeric score. In such an environment, experimental turnaround can be measured in minutes rather than weeks, and a large number of program modifications or parameter settings can be evaluated in a relatively short span of time, and at relatively low cost.

While repeatable, and hence convenient, batch mode experimental evaluation has potential drawbacks:

- Performing comprehensive – or even moderately wide – relevance judgments is a significant initial cost that can only be recouped over a period of time and through multiple uses.
- Effectiveness metrics are typically evaluated over individual queries when used in batch evaluations, whereas a user may

ADCS'12 December 5-6, 2012, Dunedin, New Zealand.

pose multiple queries as part of a session of activity during an information-seeking activity.

- The metric used might not correlate with the "user experience", meaning that differences in metric scores do not necessarily translate into measurable differences in the user's ability to carry out the desired search task.

Section 6 gives an overview of how previous researchers have addressed these various issues. Our purpose in this paper is explore the relationship between effectiveness metrics and user behavior that is alluded to in the final point, and hence shed further light on the extent to which batch evaluation scores can be argued as having been inspired (or even merely informed) by user behavior.

Section 2 introduces a range of established effectiveness metrics, and for each describes a *user model* that corresponds to the metric. The common thread that links these metrics and models is that they are *static*, and are based on predefined probability distributions. Some of the models are unappealing, in that they do not intuitively resonate with anticipated user behaviors; that reaction can, of course, be interpreted as a suggestion that the metric in question is not particularly appropriate.

In Section 3 we examine *adaptive* models in which the relevance of the documents being inspected comes into play as well as depth in the ranking. Corresponding adaptive effectiveness metrics are introduced for each of the adaptive user models, and their drawbacks considered. Section 4 then asks how models can be compared, and how the choice of an evaluation metric affects the outcomes from comparative IR experiments. Section 5 introduces a new model that better reflects the actions undertaken by typical users, and defines a corresponding effectiveness metric. Section 7 then describes an experiment that might allow confirmation of that user model.

2. STATIC USER MODELS

In this section we describe a sequence of user models. Each of them corresponds to an evaluation metric that can be applied post-hoc to runs and relevance judgments, to obtain numeric scores. A *run* is a ranking of documents or snippets, generated by a information retrieval system is response to a query; and a set of judgments (sometimes called a *qrels* file) is a record of which documents have been judged relevant for that query. Note that there is no requirement that relevance must be binary, and throughout our discussion it is assumed that relevance is (possibly quantized values selected from) a continuous scale $0 \le r \le 1$, with $r = 0$ meaning "no relevance at all" and $r = 1$ meaning "highly relevant".

Another way of thinking about relevance is that it is the *utility* the user gains if or when they view that document in the ranking. The goal of the user is to gain utility at the highest possible rate, where the unit of cost expended is a document viewing. Hence, a system that more successfully places highly relevant documents amongst the first ones viewed by the user will be a more effective system; and this is what an effectiveness metric should reflect.

Precision

In this simplest scenario, imagine a user who without variation inspects the first k proposed answers in the result listing; and, once they have done so, makes use of the subset of them that are relevant. That is, the user performs k units of work, and gains some utility as a result. Taking $\mathsf{Rel}(k) = \sum_{i=1}^{k} r_i$ as the sum of the relevance scores of those first k documents, where r_i is the relevance score of the document in the ith position in the ranking, gives a measure of that utility, and hence $\mathsf{Rel}(k)/k$ is the rate at which utility has been attained. If the relevance judgments are binary, then $\mathsf{Rel}(k)$ is the number of relevant documents, and $\mathsf{Rel}(k)/k$, is just the standard definition of *precision at depth* k, or $\mathsf{P@}k$. That is, the metric $\mathsf{P@}k$ has as a corresponding model that the user always even-handedly inspects exactly k documents in the result listing of each and every query that they pose to the retrieval system.

It is also possible to interpret the previous scenario in a probabilistic sense, and infer a uniform probability distribution over the k documents and note that $\mathsf{P@}k$ is the expected relevance that accrues from a user selecting and inspecting a single random document according to that distribution, spending one unit of work as they do:

$$\mathsf{W}_{\mathsf{Prec}}(i) = \begin{cases} 1/k & \text{when } 1 \le i \le k \\ 0 & \text{otherwise}. \end{cases}$$

With this definition, the effectiveness score computed for a ranking can be thought of as being the inner-product of a pre-defined weighting vector and a relevance vector $r = \langle r_i \rangle$. That is,

$$\mathsf{P@}k = \sum_{i=1}^{k} r_i \cdot \mathsf{W}_{\mathsf{Prec}}(i) = \sum_{i=1}^{\infty} r_i \cdot \mathsf{W}_{\mathsf{Prec}}(i),$$

where the sum can be extended to infinity because of the zeros in $\mathsf{W}_{\mathsf{Prec}}(i)$. With this formulation in place, any other probability distribution over the integers $1 \ldots \infty$ can also be used as the basis for a *weighted precision* effectiveness metric.

Scaled Discounted Cumulative Gain

Järvelin and Kekäläinen [8] observe that top-weightedness of evaluation metrics is desirable, writing "...the greater the ranked position of a relevant document ... the less likely it is that the user will ever examine it", and describe an inner-product metric they call *discounted cumulative gain*, or $\mathsf{DCG@}k$. In their description, Järvelin and Kekäläinen [8] make use of a vector of weights that in fact is not a probability distribution, multiplying the relevance of the ith item in the ranking by $1/\max\{1, \log_b i\}$; that initial formulation has since evolved in use to become $1/\log_2(i+1)$. Note that the inverse logarithmic sequence is not bounded, and that raw DCG effectiveness scores have no upper limit. To generate a probability distribution, and hence ensure that effectiveness scores are in the range $[0,1]$, the evaluation depth k must be fixed, and a truncated and scaled weight vector employed:

$$\mathsf{W}_{\mathsf{SDCG}}(i) = \begin{cases} (1/S(k)) \cdot (1/\log_2(i+1)) & \text{when } 1 \le i \le k \\ 0 & \text{otherwise}. \end{cases}$$

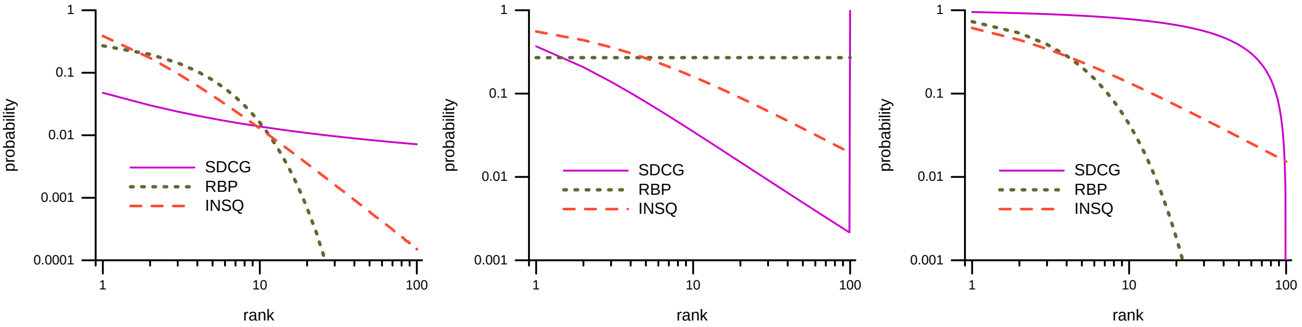
where

$$S(k) = \sum_{i=1}^{k} \frac{1}{\log_2(i+1)}$$

is the necessary scaling constant. We denote the resultant effectiveness metric as *scaled discounted cumulative gain*,

$$\mathsf{SDCG@}k = \sum_{i=1}^{\infty} r_i \cdot \mathsf{W}_{\mathsf{SDCG}}(i).$$

The corresponding user model represents users as having determined in advance that they will examine exactly k items in the result listing, and within that set of k documents, will be somewhat biased in favor of those near the top of the ranking, but also with a non-trivial interest in all of the answers through to the kth. Figure 1a shows the distribution that arises when $k = 100$, with item weight plotted a function of depth in the ranking. As can be seen, $\mathsf{SDCG@}100$ is somewhat top-weighted. But the bias is relatively small, and the document in the first position in the ranking is only seven times more likely to be examined than the document in position 100. Put another way, the model defined by $\mathsf{SDCG@}100$



(a) Weight $\mathsf{W}_{\mathsf{M}}()$ (b) Halting probability $\mathsf{H}_{\mathsf{M}}() = 1 - \mathsf{C}_{\mathsf{M}}()$ (c) Residual $\mathsf{R}_{\mathsf{M}}()$

Figure 1: Weights, halting probabilities, and residuals as a function of rank, for weighted-precision metrics SDCG@100, RBP with $p = 0.73$, and INSQ. All scales are logarithmic.

suggests that around one in seven searches reaches depth 100, but that no searches ever go to position 101 and beyond.

Rank-Biased Precision

As an alternative way of addressing the non-convergence of the inverse logarithmic sequence, Moffat and Zobel [10] suggest the use of the infinite *geometric distribution* to construct a metric they call *rank-biased precision*, or RBP, specified by:

$$\mathsf{W}_{\mathsf{RPB}}(i) = (1-p)p^{i-1}\,,$$

where p is a *persistence* parameter. Moffat and Zobel [10] also describe the user model that accompanies this probability distribution, supposing that the user always views the first answer in the ranking, and, having viewed the document at rank i, views the document at rank $i+1$ with a fixed conditional probability p. On average, a user will thus examine $1/(1-p)$ documents in the ranking.

A benefit of the use of the geometric distribution is that it converges, and hence the RBP@k metric is monotonic as the depth of evaluation k is increased. Neither P@k nor SDCG@k have this property. In both of them, as the depth of evaluation increases from k to k', scores for P@k and SDCG@k do not provide lower bounds for scores P@k' and SDCG@k'. The fact that the sequence of weights used in RBP converges also means that at any given depth of evaluation an upper bound on the eventual metric score can also be computed, based on the sum of the tail of the distribution [10]. Hence, it is possible to drop the "@k" part of the metric and refer to it as RBP; as a result, all of P, SDCG, and RBP are metrics with a single parameter each.

Inverse Squares

Any other infinite convergent distributions can also be employed, suitably normalized so that the sum is 1.0. One such alternative is given by inverse squares of ranks:

$$\mathsf{W}_{\mathsf{INSQ}}(i) = \frac{1}{S} \cdot \frac{1}{(i+1)^2}\,, \tag{1}$$

with

$$S = \frac{\pi^2}{6} - 1 \approx 0.6449\,,$$

which is a probability distribution because of the properties of the Riemann function, $\zeta(2) = \sum_{i=1}^{\infty}(1/i^2) = \pi^2/6$. Figure 1a includes the infinite weighting functions $\mathsf{W}_{\mathsf{RBP}}(i)$, plotted with $p = 0.73$, and $\mathsf{W}_{\mathsf{INSQ}}(i)$. Both are more heavily top-weighted than is SDCG.

Halting and continuing

In these metrics it is assumed that the user scans the items in the result listing from top to bottom, and stops at some point and abandons that query. That assumption allows another probability distribution to be used to characterize each of the models [3]: the probability (according to metric M) that each item in the ranking is the last one inspected, computed as:

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

which describes a probability distribution because the sequence of weights is decreasing, and because $\mathsf{W}_{\mathsf{M}}(1)$ is the largest weight. For example, $\mathsf{L}_{\mathsf{Prec@100}}(100)$ is 1.0 and $\mathsf{L}_{\mathsf{Prec@100}}(i) = 0.0$ at all other points i. The 100 th item in the ranking is always the last one inspected in this metric.

Another set of values can be derived from weight distribution $\mathsf{W}_{\mathsf{M}}(i)$ associated with metric M – the conditional probability of viewing the $i+1$ th item in the ranking, given that the i th has just been examined:

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

For example, in the user model associated with RBP, the conditional probability of viewing the $i+1$ st item in the result listing, given that the i th item has just been viewed, is always p. Figure 1b plots $\mathsf{H}_{\mathsf{M}}(i) = 1 - \mathsf{C}_{\mathsf{M}}(i)$, the conditional probability at depth i of halting the search at that point. Because SDCG@100 uses a truncated distribution, the conditional halting probability is 1.0 at depth 100. On the other hand, the weight, last, and halting probabilities for RBP and INSQ are smooth distributions.

Residuals

If any one of the four distribution $\mathsf{W}_{\mathsf{M}}()$, $\mathsf{L}_{\mathsf{M}}()$, $\mathsf{C}_{\mathsf{M}}()$, or $\mathsf{H}_{\mathsf{M}}()$ is provided for some metric M, the other three can be inferred. In addition, note that, by construction,

$$\mathsf{R}_{\mathsf{M}}(k) = \sum_{i=k+1}^{\infty} \mathsf{W}_{\mathsf{M}}(i) = \prod_{i=1}^{k} \mathsf{C}_{\mathsf{M}}(i)\,.$$

That is, the *residual* – the sum of the weight of the non-included tail at depths $k+1$ and beyond for metric M – is given by the product of the first k conditional continuation probabilities. The residual represents the score uncertainly that arises when relevance assessments r_i are only known for the first k elements in the ranking. For example, $\mathsf{R}_{\mathsf{SDCG@100}}(100) = 0.0$, by construction; whereas $\mathsf{R}_{\mathsf{INSQ}}(100) \approx 0.0151$. With RBP and $p = 0.73$ the same level of residual is achieved earlier, because of the steeper drop-off in the weight distribution compared to INSQ. With larger values of p – representing more persistent searching – that relationship alters. Figure 1c plots residual functions $\mathsf{R}(i)$ for SDCG, RBP with $p = 0.73$, and INSQ.

Are static user models realistic?

All of these four static metrics – P@k, SDCG@k, RBP, and INSQ – can be criticized. For example, P@k is not top-weighted, and SDCG@k only moderately so. Moreover, the truncated weight distributions used by P and SDCG prohibit user access beyond depth k, an aspect of their structure that is unlikely to be reflected in user performance. Similarly, RBP can be criticized because the conditional halting probability is constant at all depths, whereas it seems reasonable to suppose that a user who reaches depth 42 is less likely to stop at that point than is a user who has just examined the 2 nd document in the ranking; and a user at depth 92 is even less likely to not look at another document. That is, it seems natural for halting probabilities $\mathsf{H}_{\mathsf{M}}(i)$ to decrease as a function of depth.

In addition, when values of p below 0.9 or so are used, RBP underweights the deep part of the distribution (with $p = 0.73$, the document at depth 100 has a weight of just 6×10^{-15}); but when p is closer to 1, the distribution probably underweights the top part of the distribution (at $p = 0.95$, $\mathsf{W}_{\mathsf{RBP}}(1) = 0.05$).

The fixed model associated with INSQ rectifies many of these deficiencies – it avoids truncation, the halting probability decreases with depth, and it assigns plausible weights at both the top of the ranking ($\mathsf{W}_{\mathsf{INSQ}}(1) = 0.388$, $\mathsf{W}_{\mathsf{INSQ}}(2) = 0.172$, and $\mathsf{W}_{\mathsf{INSQ}}(3) = 0.097$) as well as further down ($\mathsf{W}_{\mathsf{INSQ}}(100) = 1.5 \times 10^{-4}$). It could also be parameterized through the use of a different power than 2, or a different additive constant than 1, to shift the three curves plotted in Figure 1.

But the real failing of static metrics is that, in terms of a user model, none of them take into account what it is that the user is experiencing as they step down the ranking. That is, static metrics completely ignore the fact that as the user examines documents they either make progress towards their search goal or they do not, and their internal assessment of the task they are working on must be evolving. Indeed, unless the user is completely agnostic as to the outcome of their search session, their behavior must of necessity differ as they do, or do not, get closer to answering the question they sought to answer. For example, they will terminate their search as soon as (or not long after) their information need has been satisfied, regardless of what they have done up until that point. This is a critical failing that has been noted by a number of authors (see, for example, Chapelle et al. [5]).

3. ADAPTIVE USER MODELS

We now consider methods in which the user model is sensitive to the relevance of the documents being examined.

Reciprocal Rank

Using the definitions already established, reciprocal rank, or RR, is given by:

$$\mathsf{L}_{\mathsf{RR}}(i) = \begin{cases} 1 & \text{if } i = \arg\min_j \{r_j \mid r_j = 1\} \\ 0 & \text{otherwise}. \end{cases}$$

The corresponding model is that the user inspects all documents in the ranking down to, and including, the first relevant one; they always end their search at the first relevant document encountered. The score is again a "rate of utility gained per unit of effort spent", since one unit of relevance is gained, out of the $\arg\min_j \{r_j \mid r_j = 1\}$ documents examined during a sequential search.

Average Precision

Reciprocal rank requires knowledge in the ranked list of the position of the first-appearing relevant document; *average precision*, or AP, can be viewed as being a generalization of RR in which knowledge is required of the positions of *all* of the relevant documents. Like RR, it is most conveniently expressed in terms of the last document probability $\mathsf{L}(i)$,

$$\mathsf{L}_{\mathsf{AP}}(i) = \begin{cases} r_i/R & \text{if } R > 0 \\ 0.0 & \text{otherwise}. \end{cases}$$

where $R = \mathsf{Rel}(N) = \sum_{i=1}^{N} r_i$ is the total sum of the relevance for that query over all of the N documents in the collection being ranked. The other distributions, $\mathsf{W}_{\mathsf{AP}}(i)$ and $\mathsf{C}_{\mathsf{AP}}(i)$ can be derived from $\mathsf{L}_{\mathsf{AP}}(i)$, as discussed in the previous section.

The corresponding user model is one in which a user selects at random one of the relevant documents (in the case of multi-grade relevance, with the selection biased by the degree of relevance) and then examines every document down to and including that one in the result listing [11]. The score assigned by the metric is again an expected rate at which utility is gained.

Are adaptive user models realistic?

As was the case with the static models, questions are quick to arise when the plausibility of the user models is considered. The model for RR requires that users scan through to the first relevant document in the ranking, regardless of how deeply it appears; the model for AP is even more contrived, in that it suggests that a user somehow intuits how many relevant documents there are in the ranking, and then scans past (on average) half of them before stopping, regardless of how far through the ranking that might take them, and regardless of how many answers they are interested in finding.

In terms of calculating scores, AP has the additional drawback that a value for the metric cannot be computed until R is known (or somehow approximated), which requires rather more work than (say) just judging the first k documents in the ranking, as is required for P@k and SDCG@k; or scanning the ranking until a relevant document is encountered, as is the case for RR. Another area for concern is that neither RR nor AP are defined if there are no relevant documents for the query. Despite these concerns, AP and RR are widely used in retrieval experimentation. Other adaptive metrics (for example, the recently-proposed ERR expected reciprocal rank metric [5]) have yet to gain traction.

4. COMPARING METRICS

Having compared the various metrics based on philosophical grounds, it is also of interest to determine if they can be compared empirically in some way. One desirable attribute of a metric is the ability to differentiate systems, since we are typically interested in determining which system is obtaining the highest scores.[1] Hence, one way of evaluating effectiveness metrics is to apply them to system runs generated in shared-task experimental regimes, and examine their ability to differentiate between the systems that contributed to the experiment in a statistically significant manner.

For example, in the TREC-10 Web Track a total of 97 system runs were submitted for evaluation, meaning that there are 4,656

[1] Note, however, that this is a somewhat circular argument, since we are only interested in separating systems if the metric is capturing some essence of the systems that is believed to be important to usability and usefulness. A metric shouldn't be chosen purely because it provides consistent system separations. The name of the system gives completely unambiguous system separations, but is clearly not an interesting reflection of retrieval performance and shouldn't be taken to be an effectiveness metric.

	rr	insq	p10	rbp73	sdcg10	p100	rbp95	sdcg100	ap
rr	**55.6**	53.8	51.2	52.9	52.4	49.1	51.6	51.5	50.2
insq	0.0	**63.8**	59.4	62.4	61.5	56.4	60.1	59.9	58.1
p10	0.0	0.0	**64.5**	61.1	62.7	59.1	63.3	62.2	60.4
rbp73	0.0	0.0	0.0	**64.7**	63.2	57.6	61.7	61.2	59.4
sdcg10	0.0	0.0	0.0	0.0	**64.9**	58.5	62.7	62.0	60.0
p100	0.2	0.0	0.0	0.0	0.0	**68.8**	63.8	66.9	64.9
rbp95	0.0	0.0	0.0	0.0	0.0	0.0	**69.1**	67.2	64.3
sdcg100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**71.0**	66.3
ap	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**72.5**

(a) Significance agreements and disagreements, $p = 0.05$

	rr	insq	p10	rbp73	sdcg10	p100	rbp95	sdcg100	ap
rr	—	85.3	77.7	81.8	80.2	65.9	71.4	68.0	62.1
insq	90.0	—	86.8	94.8	92.0	70.9	81.1	77.2	68.4
p10	85.2	92.6	—	90.3	94.2	77.4	89.5	83.1	74.3
rbp73	87.9	97.1	94.7	—	95.5	72.5	84.1	79.3	70.7
sdcg10	87.0	95.6	96.8	97.5	—	74.8	86.8	81.3	72.1
p100	79.2	85.2	88.7	86.3	87.5	—	83.4	90.0	80.4
rbp95	82.8	90.4	94.8	92.2	93.5	92.5	—	90.5	77.7
sdcg100	81.4	88.9	91.9	90.2	91.2	95.7	95.9	—	81.1
ap	78.7	85.3	88.2	86.6	87.3	91.9	90.8	92.5	—

(b) Class agreements, $p = 0.05$

Table 1: TREC-8 Adhoc Track (1999): 129 systems and 8,256 system pairs, evaluated over 50 topics. Details are explained in the text.

	rr	insq	p10	rbp73	sdcg10	p100	rbp95	sdcg100	ap
rr	**49.9**	46.6	43.2	45.9	45.3	35.7	42.2	40.6	40.1
insq	0.0	**55.5**	50.7	54.2	53.2	41.3	49.8	47.4	46.3
p10	0.0	0.0	**58.8**	53.5	55.6	44.5	54.3	50.8	49.3
rbp73	0.0	0.0	0.0	**58.0**	56.2	42.8	52.1	49.2	47.9
sdcg10	0.0	0.0	0.0	0.0	**58.6**	44.1	53.4	50.5	48.8
p100	0.5	0.2	0.2	0.2	0.2	**54.1**	49.8	52.3	49.3
rbp95	0.1	0.0	0.0	0.1	0.1	0.0	**62.7**	56.6	55.1
sdcg100	0.2	0.1	0.1	0.1	0.1	0.0	0.0	**60.0**	53.2
ap	0.5	0.1	0.1	0.1	0.1	0.0	0.0	0.0	**62.5**

(a) Significance agreements and disagreements, $p = 0.05$

	rr	insq	p10	rbp73	sdcg10	p100	rbp95	sdcg100	ap
rr	—	87.1	75.6	82.7	80.4	67.1	68.2	68.6	64.4
insq	88.4	—	84.9	94.1	91.0	70.6	77.5	75.8	69.3
p10	79.5	88.7	—	88.5	92.6	73.0	83.8	79.1	71.4
rbp73	85.2	95.5	91.8	—	95.0	70.5	79.4	76.4	69.3
sdcg10	83.5	93.3	94.8	96.4	—	72.4	81.6	78.8	70.2
p100	69.3	75.6	79.1	76.8	78.5	—	79.3	88.9	78.4
rbp95	75.2	84.4	89.5	86.5	88.0	85.3	—	87.7	79.8
sdcg100	74.2	82.2	85.7	83.5	85.4	91.7	92.3	—	79.3
ap	72.0	78.6	81.4	79.7	80.6	84.6	87.9	86.9	—

(b) Class agreements, $p = 0.05$

Table 2: TREC-10 Web Track (2001): 97 systems, and 4,656 system pairs, evaluated over 50 topics. Details are explained in the text.

"system S1 versus system S2" pairwise system comparisons that can be considered. In addition, if metric A and metric B are both used to score systems S1 and S2, then a total of five different outcomes are possible in terms of confidence indicators from a test for statistical significance, categorized as follows:

SSA Active agreements, where metric M1 and M2 both provide evidence that system S1 is significantly superior to S2, or vice versa on systems;

SSD Active disagreements, where metric M1 says that S1 is significantly better than S2, but metric M2 says that S2 is significantly better than S1, or vice versa on systems;

SN Passive disagreements, where metric M1 provides evidence that system S1 is significantly better than S2 (or vice versa on systems), but metric M2 does not provide evidence in support of the same claim;

NS Passive disagreements, where metric M2 provides evidence that system S1 is significantly better than S2 (or vice versa on systems), but metric M1 does not provide evidence in support of the same claim;

NN Passive agreements, where metric M1 fails to provide sufficient evidence that system S1 is significantly better than S2, and so does metric M2;

Tables 1 and 2 shows the result of such a comparison using the documents, runs, and judgments associated with the TREC-8 Adhoc Track (newspaper articles) and the TREC-10 Web Track (web documents). Similar results were obtained on TREC-9 Adhoc Track and TREC-9 Web Track data; those outcomes are omitted.

In part (a) of each table, the diagonal numbers (in bold) show the discriminative power of the metric in question, calculated as the proportion of all system pairs that are deemed to be significantly different, for that metric. As has been noted by other authors, there is a clear trend whereby metrics that take longer sections of the ranked search results lists into account are able to identify a larger fraction of statistically significant differences between systems. This holds for both collection types. The numbers above the diagonal in part (a) of the tables show the percentage of system pairs for which both metrics agree that one system is significantly superior to another, and both agree which is the better system (category SSA). The numbers below the diagonal show the number of systems for which both metrics show a significant difference between systems, but disagree as to which of the two systems is better (category SSD). Fortunately, this number is generally very small for the web collection, and zero for most pairs of metrics when evaluating the newswire collection.

Part (b) of each table shows two types of class agreement, again as percentages: $2\mathsf{SSA}/(2\mathsf{SSA}+\mathsf{SN}+\mathsf{NS})$ above the diagonal, and $2\mathsf{NN}/(2\mathsf{NN}+\mathsf{SN}+\mathsf{NS})$ below the diagonal. Numbers above the diagonal show the percentage agreement when both metrics report significant differences, while numbers below the diagonal show the percentage agreement where no significant difference between runs is reported.

These agreement scores represent the outcomes of "real" batch-mode IR experiments. In particular, the numbers on class agreement for significance (above the diagonal in part (b) of each table) show cases where a researcher would have concluded that the performance of one algorithm is substantially better than another, with a real effect that was highly unlikely to have been due to chance variation (at the 95% confidence level). For example, consider the column for AP, perhaps the most widely reported effectiveness metric in IR studies. The agreement between AP and other metrics ranges from 62% to 81% across the two collections. Hence, a researcher who conducted the same IR experiment, but measured the outcomes using a metric other than AP, would have rejected the results as being not significant (and hence uninteresting) around 19% to 38% of the time.

The gap between metric behaviors is problematic because, as discussed, there is currently no principled way in which to choose one evaluation metric over another. While there may be broad agreement in the community that certain metrics are more appropriate for certain task types (for example, RR is considered more appropriate for navigational searches than AP), the real differences between metrics are not well understood. The choice between AP, RBP (with a high p value), and SDCG (with a high k value) for

Initial expectation	Answer occurrence observed after query issued		
	No answers	Some answers	Many answers
Few answers (navigational)	Quickly dissatisfied, early reformulation	Possibly satisfied without needing reformulation	Satisfied quickly, no reformulation
Many answers (informational)	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 3: Hypothesized user search behavior, as influenced by two factors: the anticipated number of answers required, and the extent to which relevant documents are identified while searching. If the query is reformulated, the user's expectation in the followup query will be adjusted to account for relevance carried forward.

evaluating informational searches is largely arbitrary, and yet can lead to different experimental conclusions. It is therefore vital that a better understanding of metrics be developed, and one particular aspect that can help to determine the suitability of a metric is how closely they match real searcher behavior.

5. USER-INSPIRED ADAPTATION

Having argued that existing static and adaptive metrics are flawed in various ways, an obvious question is whether any metric exists that meets all of the design goals that were advocated in Sections 2 and 3. Such a metric should:

1. Be computable based on properties of a ranking, without requiring properties of the whole collection to be established.
2. Be top-weighted, but retain non-negligible weight $\mathsf{W_M}(i)$ at ranks of $i \geq 100$ and beyond, and be, as far as possible, a smoothly varying function of i, without being truncated.
3. Have a conditional halting probability $\mathsf{H_M}(i)$ that decreases with depth.
4. Adapt to relevant documents in the answer ranking.
5. Be parameterized in accordance with the user's initial rationale for undertaking the search.

To motivate the fifth of these goals, note that it is now accepted that there are different types of information-seeking tasks, including *navigational* interactions, where the purpose is to identify a single answer; and *informational* interactions, in which the user may be seeking to synthesize a new document by drawing on a range of a dozen or more existing ones. Legal and medical search are extreme examples of the latter; and in those disciplines a user commencing an information-seeking task might anticipate spending many hours carrying out a sequence of searches, with a view to identifying scores or even hundreds of relevant documents. That is, we believe that users commence different types of task with different expectations as to how many answers they anticipate finding, and that this expectation affects their search behavior.

To develop a user model we suggest that *the conditional probability of a user continuing their search having reached some depth i in the ranking is a combination of three factors: the depth in the ranking that has been reached; the anticipated number of answers; and the number of answers that have been identified so far through to that depth.* That is, we hypothesize that the conditional continuation function $\mathsf{C_M}(i)$ is positively related to T, the anticipated number of answers, and inversely correlated with $\mathsf{Rel}(i) = \sum_{j=1}^{i} r_i$, the amount of relevance identified down to depth i in the ranking.

For example, consider a user undertaking an informational query, with an initial (unvoiced and unexpressed) anticipation of finding perhaps 10 documents. If the first few documents in the ranking are not relevant, the user remains likely to continue looking down the ranking – after all, they were never going to stop after just one document. Alternatively, if relevant documents are encountered early, the user's mental state changes, and they are now (still unvoiced and unexpressed) anticipating finding further answers relatively quickly, after the early wins already attained.

A user that issues a navigational query has quite different behavior. They commence with the expectation that one answer will suffice, and are likely to stop as soon as a relevant document is found. Moreover, they are relatively impatient for that to happen. If the first and second documents are not relevant, they might reformulate even before looking at the third. Table 3 outlines the hypothesized mixture of behaviors.

To formalize these ideas, suppose that at the moment a user issues a query they anticipate needing T relevant documents. To capture their subsequent behavior, we envisage an effectiveness metric that has *two* components – a depth-based *background* conditional continuation probability $\mathsf{C_M}(i)$ that models (as a function of depth) the user's actions in the absence of any relevant documents appearing in the ranking; and a *discounting* modification that is used to adjust that probability as $\mathsf{Rel}(i)$ increases relative to T. Together they yield an *adjusted* continuation probability $\mathsf{C'_M}(T,i)$ that incorporates the required influences.

There are, of course, many options that suit these requirements. We now propose one arrangement that meets the hurdle of being "reasonable", even though we are not in a position to provide any evidence that it is "right".[2] As an underlying background model for user activity in the absence of any relevant documents, we parameterize the INSQ metric by adjusting it for T, the anticipated number of documents:

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

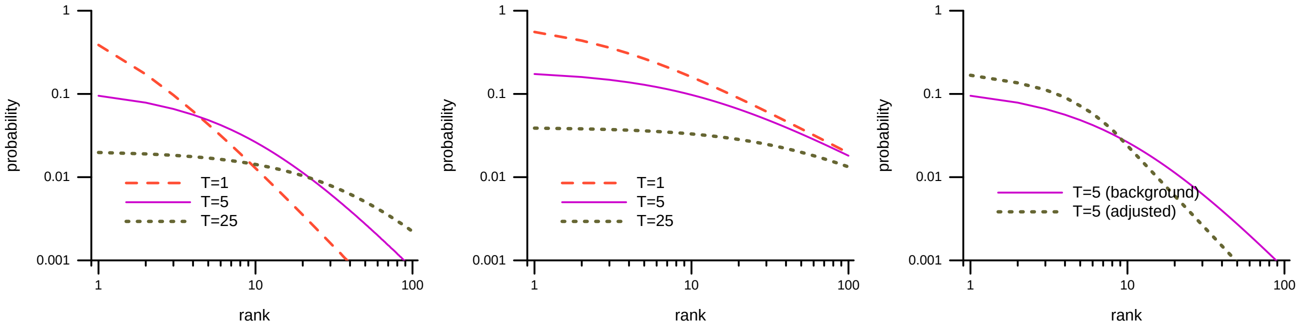
where $S_k = (\pi^2/6) - (\sum_{i=1}^{k} 1/i^2)$ is the normalization constant, and hence that

$$\mathsf{C_{INSQ}}(T,i) = \frac{(i+2T-1)^2}{(i+2T)^2}\,.$$

When $T > 1$, this has the effect of "flattening" the $\mathsf{W_{INSQ}}(i)$ curves, decreasing the weights when i is small, and increasing them when i is large. This effect is shown in Figure 2a, for three values of T; the $T = 1$ curve is the same as the INSQ curve plotted in Figure 1. Figure 2b shows the three corresponding conditional halting probability functions, $\mathsf{H_{INSQ}}(i)$.

In support of this choice for the background user behavior, note that it satisfies requirements 1–3, listed above. In terms of requirement 5, the expected search length for a weight distribution $\mathsf{W_M}(i)$

[2] That is, we exercise artistic licence at this point, and trust that the reader will accept that our intention is to be illustrative rather than prescriptive.



(a) Background $\mathsf{W}_\mathsf{M}(i)$ (b) Background $\mathsf{H}_\mathsf{M}(i)$ (c) Adjusted $\mathsf{W}'_\mathsf{M}(i)$, $T = 5$

Figure 2: Adding parameters to the INSQ metric: (a) the background weight function $\mathsf{W}_\mathsf{INSQ}(T,i)$ for three values of T; (b) the corresponding conditional halting probabilities $\mathsf{H}_\mathsf{INSQ}(T,i)$; and (c) an example showing the background and adjusted weights when $T = 5$ and $r_i = \langle 1,3,4,6,8,12,14,34,37,43,64,82,86,95 \rangle$. All scales are logarithmic.

is given by

$$E = \sum_{i=1}^{\infty} i \cdot \mathsf{L}_\mathsf{M}(i)\,.$$

The properties of the inverse square relationship, specifically that

$$\int_{k}^{2k} \frac{1}{x^2}\,\mathrm{d}x = \frac{1}{2}\int_{k}^{\infty} \frac{1}{x^2}\,\mathrm{d}x\,,$$

mean that the expected search depth E reached in the parameterized INSQ metric described by Equation 2 is approximately $2T + 0.5$. That is, a user seeking T answers is modeled as looking, on average, at around $2T + 0.5$ documents before concluding their search. For example, the expected numbers of documents examined by the parameterized INSQ metric described by Equation 2 and plotted in Figure 2a are 2.58, 10.52, and 50.50 for $T = 1$, $T = 5$, and $T = 25$ respectively. We believe that this relationship between T and expected search length $E \approx 2T + 0.5$ for the adaptive variant of INSQ helps get us to first (or even second) base in terms of "intuitive plausibility".

The default "no relevant documents encountered" behavior embodied in $\mathsf{C}_\mathsf{INSQ}(T,i)$ is then modified (requirement 4) by a discounting factor that, as the user gets closer to their goal of finding T relevant documents, increases the probability of the search terminating at any particular depth. One way this can be done is to note that once the ith document has been inspected, the user is now anticipating finding $T - \mathsf{Rel}(i)$ relevant documents where, as before, $\mathsf{Rel}(i) = \sum_{j=1}^{i} r_i$ is the total relevance achieved through to depth i. A possible formulation for an adjusted conditional continuation probability is to then use

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

as an estimate of the volume of relevance still anticipated, and take

$$\mathsf{C}'_\mathsf{INSQ}(T_i,i) = \frac{(i + 2T_i - 1)^2}{(i + 2T_i)^2}\,. \qquad (3)$$

Figure 2c shows the effect of these changes on an example ranking processed when $T = 5$ answers are anticipated. The particular ranking used is rich in relevant documents near the top, and so, compared to the model established by the background probabilities, the user is more likely to halt early. That propensity translates into an adaptive weighting function $\mathsf{W}'_\mathsf{INSQ}(T,i)$ that can only be computed once the ranking is given. In the case of the example, the computed effectiveness score rises from 0.350 to 0.502 as a result of the adaptation.

	insq5	insq10	rr	insq	p10	rbp95	sdcg100	ap
insq5	**67.5**	65.9	52.4	61.9	63.3	64.9	64.2	62.1
insq10	—	**69.0**	51.8	60.8	63.9	67.6	66.6	64.2

(a) Percentage in SSA category

	insq5	insq10	rr	insq	p10	rbp95	sdcg100	ap
insq5	—	96.6	85.1	94.3	96.0	95.0	92.7	88.7
insq10	—	—	83.1	91.6	95.7	97.9	95.3	90.9

(b) Class agreement for SSA category

Table 4: TREC-8 data: (a) percentages of system pair comparisons in the SSA categories for selected metric combinations; and (b) class agreements. These values correspond to the numbers on and above the diagonals in Tables 1a and 1b respectively.

	insq5	insq10	rr	insq	p10	rbp95	sdcg100	ap
insq5	**60.8**	58.0	44.5	52.6	54.9	55.9	52.0	50.9
insq10	—	**61.6**	42.7	50.7	54.5	58.6	54.4	53.1

(a) Percentage in SSA category

	insq5	insq10	rr	insq	p10	rbp95	sdcg100	ap
insq5	—	94.8	80.4	90.5	91.8	90.6	86.1	82.4
insq10	—	—	76.6	86.7	90.5	94.4	89.5	85.5

(b) Class agreement for SSA category

Table 5: TREC-10 data, other details as for Table 4, and can be compared with the values on and above the diagonals in Table 2.

Tables 4 and 5 extend Tables 1 and 2 respectively, concentrating on the "above the diagonal" values. Two versions of the adaptive INSQ metric defined by Equation 3 are included, with parameter values $T = 5$ and $T = 10$, denoted by "insq5" and "insq10" respectively. The column headed "insq" is the static INSQ metric defined by Equation 1, already compared to other static and adaptive metrics in Tables 1 and 2. As can be seen from the corresponding part (a) segments, in terms of their ability to determine significance, INSQ5 behaves somewhat like the shallow metrics RR and P@10, and INSQ10 is somewhat like the deeper ones. This relationship is as expected. In the corresponding part (b) in each table it is notable that the adaptive INSQ-based metrics have higher class agreements with both deep metrics (AP) and shallow ones than do any of the other metrics considered. This is a very encouraging outcome.

6. RELATED WORK

There has been a great deal of thought given to effectiveness evaluation over the last decade. Järvelin and Kekäläinen [8] introduced the idea of inner-product top-weighted measures and described both DCG and a normalized variant of it called NDCG that we have not considered here; Moffat and Zobel [10] followed up by describing the RBP metric and formalizing the corresponding user model. Zhang et al. [16] considered a range of static weighted-precision metrics, and showed that RBP with $p = 0.73$ was a good fit with the click densities observed in a commercial search engine click log; Carterette et al. [4] also examine the choice of p in RBP, reiterating that it might vary across both users and queries. Robertson [11] provided a user model for AP; Thomas et al. [13] examine the numeric stability of static metrics when applied to perturbed or degraded rankings; they also note that page boundaries can also be handled by altering the continuation probabilities at appropriate intervals. Zhang et al. [16] also consider page boundaries.

Chapelle et al. [5] examine weighted-precision effectiveness metrics, and argue that the history of what the user experiences as they process the answer list affects the way they address the remainder of the list, and discuss ways in which these adaptive cascade models can be structured; we include that critical requirement in the approach described in this paper. Yilmaz et al. [15] also explore metrics in which the probability of continuing the inspection of documents is conditional on the relevance level of the last document inspected.

Carterette [3] analyzes and categorizes a range of effectiveness metrics, grouping them into four classes; and considers the relationships between weights, halting probabilities, and last viewed probabilities that we have also employed in this work. He then explores the implications of the classification using a range of click and TREC data, concluding that DCG has a range of merits.

Most recently Smucker and Clarke [12] 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 [12] 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 propose *time-biased gain* as an effectiveness metric, and argue that it better reflects user search behavior.

In other user-focused work, Al-Maskari et al. [1] question the usefulness of deep evaluation metrics, and find that shallow metrics such as P@10 provide better correlation with the experience reported by users. Doubts have also been expressed about the usefulness of AP as a metric by Turpin and Scholer [14], who measured user task completion using degraded rankings; Huffman and Hochster [7] go further, and compare user satisfaction with a simple depth-three effectiveness metric, and find a strong correlation between them.

There has also been investigation into how best to address the complication of diversity, the fact that a query may have multiple interpretations. We do not consider that literature here; the reader is referred to Kanoulas et al. [6] and Ashkan and Clarke [2].

7. NEXT STEPS

Our key claim is that effectiveness metrics mirror user models, and hence for the scores assigned by a metric to be convincing, the user model must be a plausible one – in particular, that the “cascade” approach to evaluating a ranking must be informed by the user’s intention in issuing the query. With that in mind, we have recently commenced a user study to measure search behavior on a variety of task types, ranging from pure navigational to rich information-seeking ones. We have constructed an instrumented browser, and will monitor explicit user action in terms of queries, click-throughs, and document assessments, in the style also described by Smucker and Clarke [12]; and will be correlating those actions against gaze-tracking behavior captured for each user. In each task users will be presented with answer listings generated via the API of a de-identified commercial search service, with half of the result pages “diluted” by the insertion of attractive but not-relevant documents.

Subjects will also be shown a set of similarly-categorized information needs, and asked (without performing any searching) to estimate the number of documents they think they would need to locate in order to satisfy those information needs.

We believe that this experimental structure will allow testing of our key hypothesis, namely, that as relevant documents are identified, users become more inclined to end their perusal of the answer list, but do so more slowly if they initially sought a high number of answers. We expect to have results early in 2013, including extending the notion of “anticipated relevance remaining” through to multi-query sessions [9], as is hinted at by Table 3.

Acknowledgment

This work was supported by the Australian Research Council.

8. REFERENCES

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

[2] A. Ashkan and C. L. A. Clarke. On the informativeness of cascade and intent-aware effectiveness measures. In *Proc. WWW*, pages 407–416, Hyderabad, India, Apr. 2011.

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

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

[5] 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.

[6] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proc. WSDM 2011*, pages 75–84, Hong Kong, China, 2011.

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

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

[9] E. Kanoulas, B. Carterette, P. D. Clough, and M. Sanderson. Evaluating multi-query sessions. In *Proc. SIGIR*, pages 1053–1062, Beijing, China, July 2011.

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

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

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

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

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

[15] 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.

[16] Y. Zhang, L. A. F. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1), Feb. 2010.