System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation

Ben Carterette
Department of Computer & Information Sciences
University of Delaware, Newark, DE, USA 19716
carteret@cis.udel.edu

ABSTRACT

There is great interest in producing effectiveness measures that model user behavior in order to better model the utility of a system to its users. These measures are often formulated as a sum over the product of a discount function of ranks and a gain function mapping relevance assessments to numeric utility values. We develop a conceptual framework for analyzing such effectiveness measures based on classifying members of this broad family of measures into four distinct families, each of which reflects a different notion of system utility. Within this framework we can hypothesize about the properties that such a measure should have and test those hypotheses against user and system data. Along the way we present a collection of novel results about specific measures and relationships between them.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]; H.3.4 [**Systems and Software**]: Performance Evaluation

General Terms: Experimentation, Measurement

Keywords: information retrieval, evaluation, user models

1. INTRODUCTION

There has always been interest in producing effectiveness measures that model user behavior for systems-based evaluations with test collections. These measures frequently involve summing over the product of a discount function of ranks and a gain function mapping relevance assessments to numeric utility values, i.e.

$$M = \sum_{k=1}^{K} gain(rel_k) \cdot discount(k)$$

The widely-used *discounted cumulative gain* (DCG) measure, for instance, is typically formulated with an exponential gain function and a log-harmonic discount [10], while *rank-biased precision* (RBP) uses binary relevance and a geometric discount [12]. *Expected reciprocal rank* (ERR) maps graded relevance judgments to probabilities and discounts dynamically according to the relevance judgments at previous ranks [6]. Many traditional measures can be interpreted this way as well [21].

The discount function is often viewed as modeling a user that scans down a ranked list, growing less interested with each successive rank; the gain function models the utility the user derives from each document. This interpretation hides a great deal of diversity in the choices one can make in such a measure. Some use a probability density function as the discount; others do not. Some discount dynamically depending on relevance; others use a static discount. Depending on these choices, the actual user model can vary in subtle ways. In this work we aim to formulate a *conceptual framework*—a way to organize and describe these choices so as to provide structure for reasoning about properties of these measures in general.

To illustrate this point, consider DCG and RBP: click log analysis has suggested that RBP matches "reality" (in the sense of being more closely correlated to observed click behavior) much closer than DCG [6, 20, 21]. Despite that, DCG continues to be far more widely-used in research and development. We may hypothesize why: inertia? familiarity? or is it possible that DCG is actually modeling something quite different than RBP, and what it measures as a result is more useful to developers? We lean towards the latter explanation, and with this work we hope to provide a framework within which to test it.

Though we use RBP and DCG as motivators, our interest is not specifically in them but in model-based measures in general. The primary contribution of this work is increased understanding of effectiveness measures based on explicit user models. We define our framework in Section 2; this framework generates many possible measures (Section 3). Given the measures generated by the framework, we formulate specific hypotheses about qualities such a measure should have and test them in data (Section 4). Finally, we explore differences in modeling document utility independently of rank discounting (Section 5). Along the way we prove a number of novel results about individual measures and relationships between them.

2. MODELS

We argue that model-based measures are actually composed from three distinct underlying models:

1. a *browsing model* that describes how a user interacts with results;

SIGIR'11, July 24–28, 2011, Beijing, China.
Copyright 2011 ACM 978-1-4503-0757-4/11/07 ...$5.00.

2. a model of *document utility*, describing how a user derives utility from individual relevant documents;
3. a *utility accumulation model* that describes how a user accumulates utility in the course of browsing.

Decisions about each component model can be made independently of the others. This establishes a framework for evaluating the outcomes of those decisions.

By far the most well-developed browsing model is that of a user scanning down ranked results one-by-one and stopping at some rank k. In this work we focus entirely on that model, comparing different ways of modeling the stopping rank. Similarly, binary and graded relevance are the two most common ways to model document utility. For simplicity, throughout this section and the next we only consider binary relevance. Graded relevance (and other types of relevance or utility judgments) can be included in measures in relatively straightforward ways discussed in Section 5 below.

Thus the scope of this section and the next is limited to describing four utility accumulation models that we see in existing measures, as well as looking at different choices in modeling the stopping rank k in the simple browsing model. While there are other, more accurate browsing models [20, 11, 2], we believe they can be studied within some framework similar to the one we present here.

2.1 Model 1: Expected utility

Consider a model of a user as progressing down a ranked list, looking at each document, and stopping at some rank k. The probabilistic component of the model is a distribution $P(k)$. Since a user can only stop at exactly one rank, $\sum_{i=1}^{\infty} P(i) = 1$. To model the cost of browsing, we usually constrain $P(k)$ so that $P(1) \geq P(2) \geq ... \geq P(n)$.

Given a probability distribution, a measure reflecting this model has the form:

$$M_1 : \sum_{k=1}^{n} rel_k P(k) \qquad (1)$$

This expression can be understood as the expected relevance of the document at the stopping rank. This follows from the fact that the events of stopping at each rank k are mutually exclusive, and an expectation is computed as the sum of event probabilities times a numeric value of the event (in this case document relevance).

An exemplar for this model is *rank-biased precision* (RBP) [12], in which k is assumed to be geometrically distributed with "persistence" or "patience" parameter θ:[1]

$$P_{RBP}(k) = (1-\theta)^{k-1}\theta$$

θ is a value between 0 and 1 reflecting the patience of users for progressing down the ranked list. It can be thought of as an *a priori* probability of quitting at any given rank; the probability that a user will stop at rank $k = 2$ is the probability that they do not stop at rank $k = 1$ times the probability that they do stop at rank $k = 2$: $(1-\theta)\theta$.

2.2 Model 2: Expected total utility

The model above captures a user stopping at a particular rank, but not that the user looked at the documents above that rank. No matter where a user chooses to stop, they will see the first document with probability 1. They will only see the second if they do not stop at rank 1, so the probability of viewing document 2 is $1 - P(1)$. Continuing this way, it becomes apparent that the viewing probability at rank k is simply the cumulative probability of stopping at all ranks from k to n. Define this cumulative probability $F(k)$ as:

$$F(k) = \sum_{i=k}^{n} P(k)$$

Given a distribution, we can define a measure with the form:

$$M_2 : \sum_{k=1}^{n} rel_k F(k) = \sum_{k=1}^{n} rel_k \sum_{i=k}^{n} P(i) \qquad (2)$$

The underlying user model becomes more apparent after algebraic manipulation. Arranging summands $rel_k P(i)$ in an $n \times n$ matrix, we obtain the following:

$$\begin{bmatrix} rel_1P(1) & rel_1P(2) & rel_1P(3) & \cdots & rel_1P(n) \\ & rel_2P(2) & rel_2P(3) & \cdots & rel_2P(n) \\ & & rel_3P(3) & \cdots & rel_3P(n) \\ & & & \cdots & \cdots \\ & & & & rel_nP(n) \end{bmatrix}$$

Summing each row first, then summing the results, gives the expression above. Alternatively, summing each column first shows that:

$$\begin{aligned} M_2 : \sum_{k=1}^{n} rel_k \sum_{i=k}^{n} P(i) &= \sum_{k=1}^{n} P(k) \sum_{i=1}^{k} rel_i \\ &= \sum_{k=1}^{n} R_k P(k) \end{aligned}$$

where $R_k = \sum_{i=1}^{k} rel_i$, i.e. the number of relevant documents retrieved from rank 1 to rank k. This reveals an alternative user model: a user picks a stopping rank k, and then derives utility from *all* of the relevant documents from ranks 1 through k. The measure is the expected total utility.

We claim that *discounted cumulative gain* (DCG) [10] is an exemplar of this model, with $F(k) = 1/\log_2(k+1)$. Define a stopping probability:

$$P_{DCG}(k) = \frac{1}{\log_2(k+1)} - \frac{1}{\log_2(k+2)}$$

Then:

$$\begin{aligned} \sum_{k=1}^{n} R_k P_{DCG}(k) &= \sum_{k=1}^{n} \left(\frac{1}{\log_2(k+1)} - \frac{1}{\log_2(k+2)} \right) \sum_{i=1}^{k} rel_i \\ &= \sum_{k=1}^{n} rel_k \sum_{i=k}^{n} \left(\frac{1}{\log_2(i+1)} - \frac{1}{\log_2(i+2)} \right) \\ &= \sum_{k=1}^{n} rel_k \left(\frac{1}{\log_2(k+1)} - \frac{1}{\log_2(n+2)} \right) \end{aligned}$$

For large enough n, $1/\log_2(n+2)$ becomes negligible, and the expression reduces to:

$$\sum_{k=1}^{n} rel_k \frac{1}{\log_2(k+1)} = DCG@n$$

This means that DCG can be interpreted as modeling a user that picks a stopping rank k with probability $P_{DCG}(k)$,

[1] RBP was not originally defined as the expected relevance of one document, but it is a valid (and we argue more parsimonious) way to backfit to the expression. We discuss this further in Section 3.2.

then derives utility from all relevant documents to that rank. While others have treated DCG's discount as a viewing probability rather than a stopping probability [6], as far as we know this overall interpretation of DCG is novel, and clearly different from the usual interpretation which is closer to M_1 above. Furthermore, the stopping probability distribution is shaped similarly to the geometric distribution used by RBP (Fig. 1), meaning that DCG may encode a more realistic browsing model than previously thought.

2.3 Model 3: Expected effort

Rather than compute expected *utility*, as M_1 and M_2 do, we could compute the expected *effort* a user must put forth to achieve a particular amount of utility.

$$M_3 : \sum_{k=1}^{n} f(k)P(k) \tag{3}$$

Here $P(k)$ is conditional on relevance judgments (the "gain times discount" formulation is recovered from the use of rel_i in $P(k)$). Unlike the previous two models, this is an expectation of effort modeled by rank rather than an expectation of relevance. If $f(k) = k$, it is the expected stopping rank; if $f(k) = 1/k$ it is the expected reciprocal stopping rank. The presumption is that $f(k)$ can be defined in a way that accurately reflects user effort.

The *expected reciprocal rank* (ERR) [6] measure is exemplar for this class. Let $f(k) = 1/k$ and

$$P_{ERR}(k) = \theta_{rel_k} \prod_{i=1}^{k-1} (1 - \theta_{rel_i})$$

Here θ_{rel_i} is a parameter indicating the probability that a document with relevance rel_i would be useful to the user. With binary relevance ($rel_i = 0$ or 1), and with $\theta_0 = 0$ and $\theta_1 = \theta$, this can be rewritten as:

$$P_{ERR}(k) = rel_k (1 - \theta)^{R_k - 1} \theta$$

revealing that P_{ERR} is geometric over total accumulated relevance, with non-zero probability only at ranks at which relevant documents appear. Here θ can still be understood as a patience parameter, as in RBP, but now the probability of stopping at a nonrelevant document is zero while the probability of stopping at a relevant document is θ. Thus we could reasonably re-express this model as:

$$M_3 : \sum_{R_k=1}^{R} f(k)P(R_k)$$

This makes explicit the idea of evaluating the effort the user expends to find R_k relevant documents (or in general to achieve a given amount of utility).

2.4 Model 4: Expected average utility

The final model combines the utility- or reward-based models M_1 and M_2 and the effort-based model M_3.

$$M_4 : \sum_{k=1}^{n} rel_k \sum_{i=k}^{n} f(i)P(i) \tag{4}$$

Intuitively, this models a user that, after each relevant document, considers the expected effort of further browsing. As with M_2, the underlying user model becomes more transparent after algebraic manipulation:

$$\begin{aligned} \sum_{k=1}^{n} rel_k \sum_{i=k}^{n} f(i)P(i) &= \sum_{k=1}^{n} f(k)P(k) \sum_{i=1}^{k} rel_i \\ &= \sum_{k=1}^{n} f(k) R_k P(k) \end{aligned}$$

and if $f(k) = 1/k$,

$$\sum_{k=1}^{n} \frac{R_k}{k} P(k) = \sum_{k=1}^{n} prec@k \cdot P(k)$$

Thus the user model is based on average gain per document viewed: a user stops at a rank k and gains R_k total utility over k documents for an average of *prec@k* each.

Average precision (AP) is the exemplar for this model (note the similarity to Robertson's AP model [13]). If

$$P_{AP}(k) = \frac{rel_k}{R}$$

where R is the total number of relevant documents in the corpus, we recover AP from the expression. Like $P_{ERR}(k)$, this models a user that will never stop at a nonrelevant document, but unlike any of our other distributions, it uses a simple uniform distribution over all possible stopping points. AP assumes knowledge of the complete set of relevant documents, but this is not required for measures in this family.

3. MEASURES

The previous section establishes our framework with four utility accumulation models for one common browsing model. Each accumulation model has a well-known exemplar measure, and those four measures have four different ways to model the stopping rank for browsing:

$$\begin{aligned} P_{RBP}(k) &= (1 - \theta)^{k-1} \theta \\ P_{DCG}(k) &= \frac{1}{\log_\theta(k + \theta - 1)} - \frac{1}{\log_\theta(k + \theta)} \\ P_{ERR}(k) &= rel_k (1 - \theta)^{R_k - 1} \theta \\ P_{AP}(k) &= \frac{rel_k}{R} \end{aligned}$$

For compactness we will assume base $\theta = 2$ for P_{DCG}. For simplicity of exposition we still assume binary relevance, which is where the simplified P_{ERR} comes from. R_k is the total number of relevant documents retrieved at ranks 1 through k. We will add two more distributions to those:

$$P_{RR}(k) = \frac{1}{k(k+1)} \qquad P_{RRR}(k) = \frac{rel_k}{R_k(R_k + 1)}$$

These distributions are interesting because their cumulative forms are reciprocal ranks—hence the names P_{RR} (for reciprocal rank) and P_{RRR} (for reciprocal relevant rank)—and because they arise naturally as a mixture of the geometric distributions in both RBP and binary-ERR (Section 3.4).

The six distributions can be characterized by whether they are *static*, i.e. independent of relevance judgments, or *dynamic*, i.e. dependent on a specific ranking (cf. Yilmaz et al. [20]). P_{RBP}, P_{DCG}, and P_{RR} are static; P_{ERR}, P_{AP}, and P_{RRR} are dynamic.

All six distributions and their cumulative forms $F(k)$ are illustrated in Figure 1. The three static distributions have

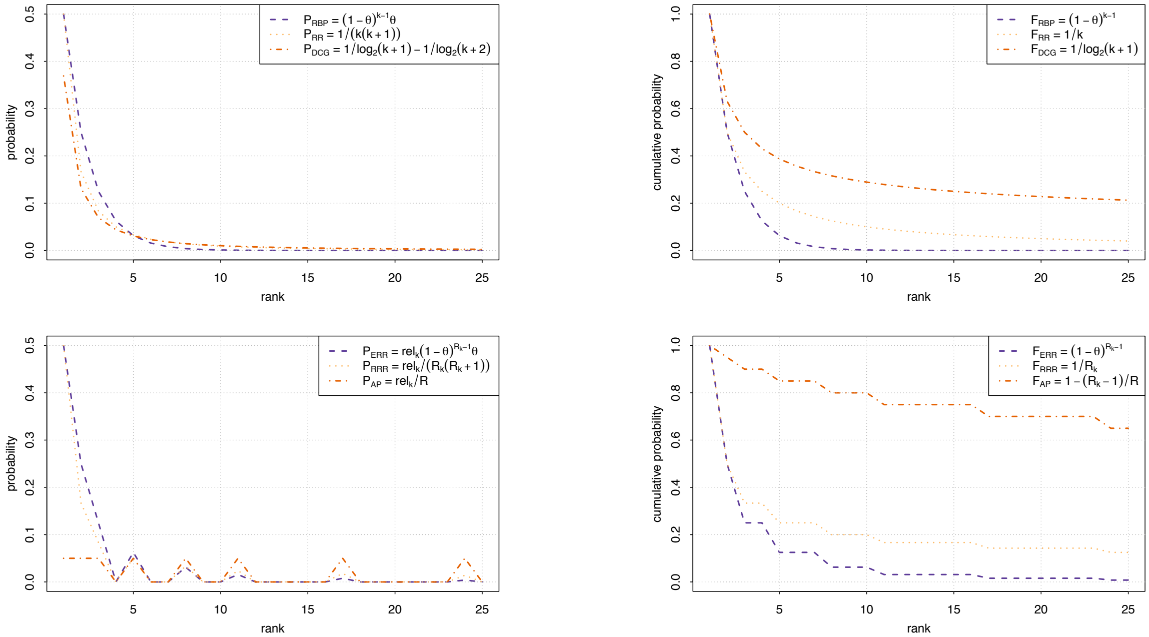

Figure 1: Comparison of stopping probability density functions. The left two plots compare stopping probabilities; the right two compare viewing probabilities (cumulative distributions of stopping probabilities). Upper left: static density functions; upper right: static cumulative density functions. Lower left: dynamic density functions (for a ranking with relevant documents at ranks 1, 2, 3, 5, 8, 11, 17, 24); lower right: dynamic cumulative density functions.

similar shape, but P_{DCG} and P_{RR} have much fatter tails than P_{RBP}—while this is difficult to see by inspection of Figure 1, it is very clear on a log-scale plot. Furthermore, P_{DCG} has a substantially fatter tail than P_{RR}, which becomes clear when looking at their cumulative densities. The dynamic distributions all give zero stopping probability to ranks at which nonrelevant documents appear. P_{ERR} and P_{RRR} have a similar shape over relevant documents, but P_{RRR} has a fatter tail (visible by inspection and very clear in the cumulative density). P_{AP} is uniform over relevant documents, so its cumulative density decreases slowly. For RBP and ERR, we used $\theta = 0.5$; these properties hold for other common values.

3.1 Alternative measures

The choice of a probability distribution produces up to four different measures, one for each model. We have mixed and matched models and distributions to produce eleven measures in addition to our four exemplars (Table 1). We have attempted to name them in a way that makes clear both the choice of model and distribution.

RBTR (rank biased total relevance, M_2 with P_{RBP}):

$$RBTR = \sum_{k=1}^{n} R_k (1-\theta)^{k-1}\theta = \sum_{k=1}^{n} rel_k (1-\theta)^{k-1}$$

RBAP (rank biased average precision, M_4 with P_{RBP}):

$$RBAP = \sum_{k=1}^{n} prec@k \cdot (1-\theta)^{k-1}\theta$$

CDG (cumulated discounted gain, M_1 with P_{DCG}):

$$CDG = \sum_{k=1}^{n} rel_k \left(\frac{1}{\log_2(k+1)} - \frac{1}{\log_2(k+2)} \right)$$

DAG (discounted average gain, M_4 with P_{DCG}):

$$DAG = \sum_{k=1}^{n} prec@k \left(\frac{1}{\log_2(k+1)} - \frac{1}{\log_2(k+2)} \right)$$

EPR (expected precision, M_4 with P_{ERR}):

$$EPR = \sum_{k=1}^{n} prec@k \cdot rel_k (1-\theta)^{R_k - 1}\theta$$

ARR (average reciprocal rank, M_3 with P_{AP}):

$$ARR = \sum_{k=1}^{n} \frac{rel_k}{kR}$$

RRG (reciprocal rank gain, M_1 with P_{RR}):

$$RRG = \sum_{k=1}^{n} \frac{rel_k}{k(k+1)}$$

RR (reciprocal rank, M_2 with P_{RR}):

$$RR = \sum_{k=1}^{n} \frac{R_k}{k(k+1)} = \sum_{k=1}^{n} \frac{rel_k}{k}$$

RAP (reciprocal average precision, M_4 with P_{RR}):

$$RAP = \sum_{k=1}^{n} prec@k \cdot \frac{1}{k(k+1)}$$

type	$P(k\|\theta, rel_1...rel_n)$	$M_1 : \sum rel_k P(k)$	$M_2 : \sum R_k P(k)$	$M_3 : \sum \frac{1}{k} P(k)$	$M_4 : \sum prec@k P(k)$
static	$P_{RBP} = (1-\theta)^{k-1}\theta$	**RBP**	RBTR	–	RBAP
	$P_{DCG} = \frac{1}{\log(k+1)} - \frac{1}{\log(k+2)}$	CDG	**DCG**	–	DAG
	$P_{RR} = \frac{1}{k(k+1)}$	RRG	RR	–	RAP
dynamic	$P_{ERR} = rel_k(1-\theta)^{R_k-1}\theta$	–	–	**ERR**	EPR
	$P_{AP} = \frac{rel_k}{R}$	–	–	ARR	**AP**
	$P_{RRR} = \frac{rel_k}{R_k(R_k+1)}$	–	–	RRR	RRAP

Table 1: A model and a probability distribution together specify a measure, or possibly a family of measures parametrized by θ. Cells with – have been deemed uninteresting because they are either constant-valued or isomorphic to traditional recall or precision. Measures in bold font are the chosen exemplars for the model.

RRR (reciprocal relevant rank, M_3 with P_{RRR}):

$$RRR = \sum_{k=1}^{n} \frac{1}{k} \frac{rel_k}{R_k(R_k+1)}$$

RRAP (reciprocal relevant average precision, M_4 with P_{RRR}):

$$RRAP = \sum_{k=1}^{n} prec@k \frac{rel_k}{R_k(R_k+1)}$$

We do not claim that these measures are altogether new to IR—we know that some have been described before in various contexts (particularly our so-called RR, which has frequently been described as "DCG with a reciprocal rank discount"). Furthermore, there are other model-based measures that we do not discuss (e.g. the NCU family [16]). Our interest is less in developing or arguing for any particular measures than in using them to explore hypotheses about model-based measures in general.

3.2 Normalization

Some of the measures require normalization. The M_2 family in particular is heavily dependent on the total number of judged relevant documents for a topic. In addition, some of the probability distributions may not sum to one if the total number of relevant documents is low or if the ranking is truncated. Both of these are confounding effects that normalization helps resolve.

To normalize, we simply divide the value of the measure by the maximum possible value given the judged relevant documents, i.e. the value of the measure on a perfect ranking. This ensures that the maximum achievable value is 1 for all topics and all systems. This is of course a well-known normalization procedure, commonly used with DCG to produce nDCG [10]. We will apply it to all measures in the M_2 family, as well as a few other measures that do not naturally fall between 0 and 1 such as ARR.

Normalization is a somewhat thorny topic. In presenting RBP, Moffat and Zobel started with an unnormalized version that would fit in our M_2 family, but explicitly chose not to normalize by the maximum achievable effectiveness—instead normalizing by the expected number of documents viewed [12] and thereby moving the measure into the M_1 family. There is a tradeoff between preserving the user model and having a measure that can be averaged across topics, and it is not always clear how to resolve it.

3.3 Rank cut-offs

Some measures are calculated to a pre-specified rank cut-off rather than over the full ranking of n documents. This is particularly true of DCG. In the case of a rank cut-off $K \ll n$, the math in Section 2.2 does not work—K must be large enough that $1/\log_2(K+2)$ is close to zero, and that is certainly not the case for the usual values of K.

There is a simple resolution that fits with the user model, though. First, note that we can express DCG@K in terms of $P(k)$ and $F(K+1)$ with some simple algebraic manipulation:

$$\begin{aligned} DCG@K &= \sum_{k=1}^{K} rel_k \frac{1}{\log_2(k+1)} = \sum_{k=1}^{K} rel_k F_{DCG}(k) \\ &= \sum_{k=1}^{K} rel_k \sum_{i=k}^{n} P_{DCG}(i) \\ &= \sum_{k=1}^{K} rel_k \left(\sum_{i=k}^{K} P_{DCG}(i) + \sum_{i=K+1}^{n} P_{DCG}(i) \right) \\ &= \sum_{k=1}^{K} rel_k \left(\sum_{i=k}^{K} P_{DCG}(i) + F_{DCG}(K+1) \right) \\ &= \sum_{k=1}^{K} rel_k \sum_{i=k}^{K} P_{DCG}(i) + \sum_{k=1}^{K} rel_k F_{DCG}(K+1) \end{aligned}$$

We can then apply the same algebraic trick we used in Section 2.2 to complete the expression as:

$$DCG@K = \sum_{k=1}^{K} R_k P_{DCG}(k) + R_K F_{DCG}(K+1)$$

Note that $R_K F_{DCG}(K+1) = \sum_{k=K+1}^{n} R_K P_{DCG}(k)$. Thus the user model is exactly the same. The difference is that calculation of the measure now assumes the worst case for a user that chooses to stop beyond rank K—that the user will not find any new relevant documents, and therefore will only derive utility from those at ranks 1–K.

This argument generalizes to any M_2 measure.Therefore a rank cut-off K solves a problem with non-converging discounts by making a worst-case assumption about the effectiveness of the system below K.

3.4 Reciprocal rank distributions

Here we show how our P_{RR} and P_{RRR} distributions emerge from a mixture of geometric distributions, and that their cumulative forms are reciprocal ranks.

The geometric distribution has a parameter θ that requires the researcher/developer to specify a value *a priori*. Perhaps being unwilling to make any strong statements about user patience, one could instead use several different values of θ and average the results. Taken to the limit, they could obtain $P(k)$ by averaging geometric distributions over all possible values of θ:

$$P(k) = \int_0^1 P(k|\theta)p(\theta)d\theta = \int_0^1 (1-\theta)^{k-1}\theta p(\theta)d\theta$$

where $p(\theta)$ is uniform over the range $[0,1]$. Since it is uniform, we can disregard it; then integration by parts gives:

$$P_{RR}(k) = \frac{1}{k(k+1)}$$

Thus P_{RR} can be seen as an average of infinitely many geometric distributions.

The cumulative distribution is:

$$F_{RR}(k) = \sum_{i=k}^{n} \frac{1}{i(i+1)} = \sum_{i=k}^{n} \frac{i+1-i}{i(i+1)} = \sum_{i=k}^{n} \left(\frac{1}{i} - \frac{1}{i+1}\right)$$
$$= \frac{1}{k} - \frac{1}{n+1}$$

As $n \to \infty$, $\frac{1}{n+1} \to 0$, so for large enough n this is approximated as reciprocal rank $1/k$. It does not require very large n for the effect to be negligible.

The same argument generalizes the binary-relevance version of P_{ERR} to P_{RRR} as an average of infinitely many geometric distributions over ranks of relevant documents.

4. ANALYSIS

We have presented a framework for classifying and generating measures that model system utility to a user. The benefit of a framework is that it poses questions and also provides a guide to answering them. Our goal is not to evaluate the new measures we propose, but to formulate specific hypotheses about model-based measures in general and answer them by appealing to our suite of measures.

Some of the questions the framework raises are:

1. Are utility-based models (M_1, M_2) better than effort-based models (M_3, M_4)? Our hypothesis is that there is no difference on average.
2. Are measures based on stopping probabilities (M_1, M_3) better than measures based on viewing probabilities (M_2, M_4)? Our hypothesis is that the latter are more robust to different sources of variance.
3. What properties should P have to produce a good measure? We hypothesize that it should have a fatter tail (within reason), and that static distributions are more robust than dynamic.

"Better" and "good" are of course qualitative words whose meaning depends on the retrieval task being studied, the users of the system, and a host of other factors. We will look at goodness-of-fit to user data and robustness of evaluation, though there are other ways to evaluate these questions.

4.1 Click logs

One possible definition of a "good" measure is one that more closely models user behavior. We compared our static stopping rank distributions to aggregated clicks from the 2006 AOL log. Of course, processing click log data is itself implicitly model-based. Comparing models to click logs should be seen not as comparing models to reality, but as comparing one model to another under all the assumptions that both models require. In this case we cannot know stopping ranks; we need to model them from recorded clicks.

We tried two different models. The first simply maps each recorded click to a stopping rank to estimate an empirical distribution of stopping ranks, with each rank mutually exclusive of the others. The implicit model is that every click is a new event for a new user/query pair, even if it is in not actually the last click by that user for that query. The second maps only the last recorded click for a user/query to a stopping rank. This may be more realistic, but it results in throwing out all clicks but the last. Other models are possible (e.g. the "gap" model of Zhang et al. [21]).

Both empirical distributions are shown in Figure 2 along with our static distributions. Among the distributions we are considering, P_{RR} is clearly the best fit to both models of the data. Previous work has suggested that DCG and reciprocal rank discounts do not adequately model users [20, 21, 6]. This analysis suggests they might model users well in some data, provided they are considered *cumulative* densities rather than probability densities.

4.1.1 Critique of click log analysis

While click log analysis can provide a guide for evaluating a probability distribution $P(k)$, it is not clear to us how it could be used to evaluate a model of utility accumulation. How can we use click log data to choose between M_1 (in which a user derives utility only from the document at the stopping rank) and M_2 (in which a user derives utility from all relevant documents from rank 1 to the stopping rank)? How can we use it to choose between M_1 and M_3 (in which a user expends a certain amount of effort to achieve a given total utility)? These decisions must be made on the basis of more than just user data.

Furthermore, even if only a small fraction of users are negatively affected by a poor model fit, the cost of not providing those users with the best possible system may be disproportionately large. Suppose that the probability of a user becoming frustrated and quitting the search engine altogether increases over ranks in a reverse geometric way up to rank K, so that

$$P(\text{quit engine}|k) = (1-\theta)^{K-k}$$

This is distinguished from

$$P(\text{stop current search at } k) = (1-\theta)^{k-1}\theta$$

Then:

$$\begin{aligned} &P(\text{stop search and quit engine at } k) \\ &\quad = P(\text{quit engine}|k)P(\text{stop search at } k) \\ &\quad = (1-\theta)^{K-k}(1-\theta)^{k-1}\theta = (1-\theta)^{K-1}\theta \end{aligned}$$

which is completely independent of rank! We may as well use a uniform stopping probability—leading us right back to traditional precision and recall.

We do not advocate that model; we only wish to point out the pitfalls in focusing on behavior evidenced in logs. This analysis suggests to us that fatter-tailed distributions are superior even if they do not fit behavior data, because those distributions are better-equipped for the risk of not satisfying users that are in the tail.

4.2 Robustness and stability

The previous section supposes that one purpose of an effectiveness measure is to model users in order to estimate the utility of the system. Another purpose of evaluation

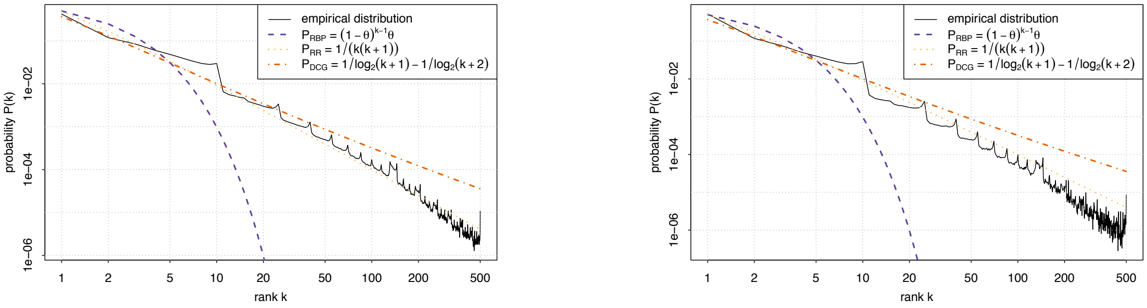

Figure 2: Comparison of empirical click probabilities to static stopping rank models. The left plot uses every recorded click; the right only uses the last recorded click for a user/query pair. It is clear that P_{RR} is the best fit to both distributions (though better fits are possible).

is to choose among different retrieval models, features, and system implementations. For that purpose we would like decisions to be roughly the same whether they are based on a few topics versus many, or extensive relevance judgments versus shallow pools, or one group of assessors versus another. The extent to which conclusions are different depending on differences in the data used to compute the measures reflects the *robustness* and *stability* of those measures. To investigate these properties we will look at how evaluation measure scores and relative rankings of systems change as the underlying data changes.

4.2.1 Data

Our primary data is the TREC-6 ad hoc data consisting of TREC topics 301–350, 72,270 total relevance judgments, and 74 submitted runs over a corpus of about 550,000 documents [18]. This is a small corpus, but its deep judgments make it useful for our study. Furthermore, there is an alternate set of relevance judgments from the University of Waterloo for these topics, allowing us to investigate the effect of assessor disagreement [9, 17].

We also used two more recent test collections:

- The named page retrieval task for the TREC 2006 Terabyte track [3]. This is a high-precision task: there are only a few relevant documents in the corpus. The data consists of TREC topics 901–1081, 2,361 total relevance judgments, and 43 submitted runs over the GOV2 corpus of some 25 million web pages.
- The ad hoc task for the TREC 2009 Web track (category A set) [7]. The data consists of TREC Web topics 1–50, 18,666 relevance judgments, and 37 submitted runs over the ClueWeb09 corpus of 1 billion web pages.

4.2.2 Evaluation

The approach is simple: we select some subset of the data (e.g. a subset of topics or a subset of judgments) and evaluate all systems with all 15 of our measures. We then use Kendall's τ rank correlation to compare the results to the "true" rankings using the full TREC data. Kendall's τ ranges from -1 to 1, with greater values indicating greater correlation. In practice, for meta-evaluation studies like this one Kendall's τ is nearly always over 0.6, and a value of 0.9 would be considered an effectively "perfect" correlation considering the presence of variance [17].

Since our hypotheses are about model families or distributions rather than individual measures, we average Kendall's τ results for each measure within a family or with a particular property. We use the averages to evaluate the hypothesis.

4.2.3 Results

The single clearest fact from the results is that measures in the M_2 family with fatter tails tend to be more robust. However, this is not true in all cases: when the judgment pool is shallow, models with distributions that put more weight on top-ranked documents tend to be more robust. For tasks with few relevant documents, tail fatness does not appear to matter. Detailed results follow below.

Varying assessors: We evaluated all 74 TREC-6 systems with all 15 of our measures over two different sets of relevance judgments. Table 2 shows the τ correlations for every measure between the rankings from the two judgment sets. The final column shows the mean τ for each stopping distribution; it is clear that the fatter-tail distributions P_{DCG} and P_{AP} are most robust w.r.t. assessor disagreement, while the slimmer-tail distributions P_{RBP} and P_{ERR} are least robust. Dynamic distributions actually appear to be more robust than static distributions on average, which we found surprising since they seem to have a greater dependence on which documents have been judged relevant.

The last row shows the mean τ for each model family; M_2 is more robust to assessor disagreement than the others. M_3 and M_4 are about equally robust, with M_3 taking the edge if the outlying fat-tailed P_{AP} is removed. M_1 is least robust to assessor disagreement.

Varying topic sample: We evaluated all 74 TREC-6 systems with all 15 of our measures over increasing topic sample sizes from $N = 5$ to 45. For each level of N, we performed 100 trials with a random subset of topics; each trial used the same topic sample to evaluate all 15 measures.

Figure 3 plots summary results for distribution model family (left) and tail type (right). Again we see that fatter tails result in more stable results, and M_2 provides more stable results than the other models. There are differences from the assessor disagreement results, however. M_3 appears to be least robust to varying the topic sample despite being more robust to assessor disagreement, while M_1 and M_4 look equally robust. The difference in robustness due to tail fatness is less pronounced, with a maximum difference

type	$P(k)$	M_1	M_2	M_3	M_4	mean
static	P_{RBP}	$\tau_{RBP} = 0.813$	$\tau_{RBTR} = 0.816$	–	$\tau_{RBAP} = 0.801$	0.810
	P_{DCG}	$\tau_{CDG} = 0.831$	$\tau_{DCG} = 0.920$	–	$\tau_{DAG} = 0.819$	0.857
	P_{RR}	$\tau_{RRG} = 0.819$	$\tau_{RR} = 0.859$	–	$\tau_{RAP} = 0.812$	0.830
dynamic	P_{ERR}	–	–	$\tau_{ERR} = 0.829$	$\tau_{EPR} = 0.836$	0.833
	P_{AP}	–	–	$\tau_{ARR} = 0.847$	$\tau_{AP} = 0.896$	0.872
	P_{RRR}	–	–	$\tau_{RRR} = 0.826$	$\tau_{RRAP} = 0.844$	0.835
mean		0.821	0.865	0.834	0.835	

Table 2: Kendall's τ correlation between TREC-6 relevance judgments and alternate judgments. The last column shows row means (mean τ for each $P(k)$); the last row shows column means (mean τ for each model).

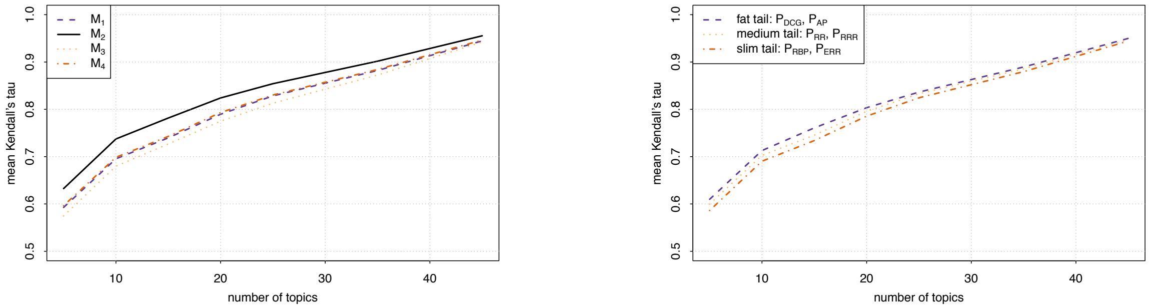

Figure 3: Kendall's τ correlations as topic sample size increases (averaged over 100 samples of size N). All differences are significant with $p < 10^4$. Fat-tailed distributions and the M_2 family clearly offer the greatest stability to topic set size.

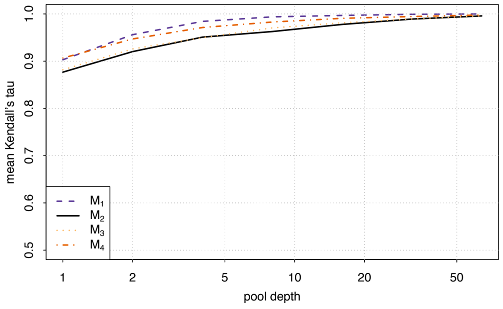

Figure 4: Kendall's τ correlations as pool depth increases. Correlation is already very good with a pool depth of just 1; this suggests that these systems are retrieving many common documents.

of 0.026 between any two points (though this is statistically significant).

Varying pool depth: We formed shallower judgment pools from the original TREC-6 qrels by iterating over systems and pooling only documents that appeared above a particular rank cutoff. For each of these pools we evaluated all 74 systems by all 15 measures and correlated the resulting ranking of systems to the "true" ranking using all judgments.

Figure 4 shows increasing τs for each of our models. The M_2 family is actually least robust to missing judgments, while the M_1 family is most robust. This makes sense, as the M_1 family places much more weight on the top-most documents than the others, and the top-ranked documents are the ones that are judged in both datasets. The M_4 family is more robust to missing judgments than the M_3 family, possibly because M_4 discounts lower-ranked documents more. Nevertheless, all four are quite robust on average, most likely because five of our six distributions weight the top-ranked documents very highly. Results for tail fatness are not shown, but here too there is a reversal: slim-tail distributions are more robust to shallow pools than fatter-tail ones. Again, this is most likely because the slimmer tails result in much more weight on the top-ranked documents.

Varying test collection: For our other two collections, we varied topic sample size and calculated τ correlations.

For the TREC 2006 Terabyte named page task—a high-precision task—the distribution does not appear to matter. There is almost no difference between τ correlations with fat- or slim-tailed distributions. M_2 measures are still more robust than other families, though M_3 is a very close second. This supports our intuition that M_3 is particularly useful for tasks where there are only a few highly-relevant documents.

For the TREC 2009 Web ad hoc task, M_3 measures with slim- or medium- tail distributions are most robust to varying topic sample. Fat-tail distributions are quite poor. M_2 measures are least robust, though not by much. This is most likely due to the challenge of evaluating an ad hoc retrieval task with sparse relevance judgments, but it may suggest that ERR is the best measure we have for web evaluation.

5. DOCUMENT UTILITY

To this point our discussion has focused on stopping probability distributions within a common browsing model and models of how users accumulate utility over documents. We simplified the idea of document utility itself to simple binary relevance, but we can investigate alternative models for that independently of stopping probabilities.

5.1 Graded judgments

Some documents are more useful than others. A common way to model this is with judgment grades, such as the ternary scale *nonrelevant, relevant, highly relevant* and the quinary scale *bad, fair, good, excellent, perfect.* A gain function maps grades to numeric values.

Though we have focused on binary judgments, graded judgments fit easily within our framework. DCG and ERR are, of course, explicitly designed with graded judgments in mind. AP can be adapted to graded judgments using a user-modeling distribution mapping grades to probabilities of relevance [14]. At any point where we use rel_k or R_k for binary judgments, we can substitute the mapping function $gain(rel_k)$ or the cumulative gain $CG_k = \sum_{i=1}^{k} gain(rel_i)$ respectively. In ERR, we can write $P(k)$ as:

$$P_{ERR}(k) = \prod_{g \in \text{grades}} (1 - \theta_g)^{G_k - 1} \theta_g$$

where the product is over unique grades, θ_g is a patience parameter for grade g, and G_k is the total number of documents with grade g up to rank k. This is a slightly different formulation that originally presented by Chapelle et al. [6], but we feel it more clearly shows ERR as modeling cumulated gain up to rank k.

5.2 Preference judgments

Another way to model the idea that some documents are more useful is with preference judgments of the form "A is preferred to B for query q". When transitive, such judgments result in a total ordering of documents by utility [15]. Previous work has shown that preferences can be made easier and faster than graded judgments [5]. Kendall's τ rank correlation is based on preferences, and it can be extended to other measures for IR effectiveness [4].

Models M_2 and M_4 can be seen as instances of a more general family of preference-based measures. Consider the following stochastic process applied to a system ranking: sample a rank k with probability $P(k)$. Then sample one or more documents ranked above k. For each of those documents that was preferred by assessors to the document at k, increment a count of total concordant pairs.

Suppose we just have binary relevance and a preference relation stating that A is preferred to B if and only if A is judged relevant. If in the second sampling stage we sample only one document uniformly at random, the expectation of the process is exactly M_4 with $f(k) = 1/k$. If we use all documents above k, the expectation is exactly M_2.

For natural preference relations, uniform $P(k)$ results in the process having expectation proportional to Kendall's τ. This suggests that both M_2 and M_4 can be viewed as weighted versions of τ. This was already known for AP, our exemplar M_4 measure [19]; the fact that DCG can be viewed in this way is novel.

5.3 Novelty and diversity

In the novelty and diversity retrieval setting, document utility is a function of its relevance to different possible *user intents* as well as its redundancy with other documents in a ranking. The so-called "intent aware" (IA) family of measures uses a distribution of intents $P(i|q)$ for a given query q to compute a weighted average of a measure like AP or DCG computed with document judgments for each intent [1]. Any measure that fits in our framework can be turned into an intent aware variant by computing such a weighted average.

Redundancy can be penalized when computing total utility. Up to this point we have computed total utility as the number of retrieved relevant documents R_k. We could instead define the utility of the top k retrieved documents as:

$$U_k = \sum_{j=1}^{k} rel_j F(R_j)$$

where R_j is the number of relevant documents up to rank j and $F(R_j)$ is a redundancy discount taking the form of a cumulative probability density based on $P(R_j)$, the probability that the last relevant document a user would look at is the R_jth. The α-nDCG measure [8] uses $F(R_j) = \alpha^{R_j - 1}$; it is based on the same geometric penalty that ERR uses.

A full intent-aware, redundancy-penalizing novelty/diversity measure in the M_2 family could then be given as:

$$M = \sum_{i \in I_q} P(i|q) \sum_{k=1}^{n} U_{ik} P(k|i)$$

where I_q is the set of intents for query q and U_{ik} is defined as the utility to intent i (using relevance judgments distinguished by intent). Measures in the M_3 and M_4 families follow straightforwardly. Note that M_1 cannot truly model redundancy penalization in a natural way, since it does not model accumulated utility.

This suggests two directions for diversity evaluation:

- Consider the use of a stopping probability conditional on intent, i.e. $P(k|i)$. For an ambiguous query like "cardinals" for instance, some intents may be "navigational" (e.g. finding the home page for the St. Louis Cardinals baseball team) and others may be "informational" (e.g. finding information about Catholic cardinals). A user with the former need might be better modeled by a steeply-decreasing density function, while the latter might be better modeled by a more gentle decrease. Existing diversity measures assume that the same model will be used for all intents.
- Consider alternative models of decrease in utility due to redundancy. α-nDCG uses a geometric decrease. As we have seen in this work, a geometric discount by rank is quite harsh compared to user behavior; it may be a harsh penalty for redundancy as well. Our definition of U_k above allows a researcher or developer to plug in any probability density function to model the probability that information is still useful after having seen it R_j times.

6. DISCUSSION AND CONCLUSION

On one level this work is a collection of novel observations about common evaluation measures and user models. At that level, these observations emerge primarily from algebraic manipulation.

- The discounts in DCG and RBP are used for different purposes. RBP's is the probability of stopping at a particular rank, while DCG's is the probability of viewing a particular rank while progressing down. This leads to different interpretations of the two, as well as the understanding that their discounts are not directly comparable to each other. DCG's discount should instead be understood as arising from a stopping probability distribution that much more closely models user behavior.

- Nearly any measure that can be formulated in terms of a relevance gain times a rank discount can also be formulated in terms of a stopping probability distribution. The only constraint is that $disc(1) \geq disc(2) \geq \cdots \geq disc(n)$. Zhang et al. have shown that many traditional measures can be expressed this way [21].
- The reciprocal-rank discount arises naturally as the cumulative density of an infinite mixture of geometric discounts. It also fits user behavior better than other discounts when properly interpreted.
- M_2 measures with a rank cut-off are equivalent to using no rank cut-off, but making a worst-case assumption about relevance below a certain rank.
- DCG and other measures in both the M_2 and M_4 families can be understood as weighted rank correlations.
- Our measure families generalize to families of measures for novelty and diversity with new questions about modeling distributions.

At a deeper level, these observations all emerged from a conceptual framework in which we describe a measure in terms of its browsing model, specific attributes of its browsing model, and utility accumulation model. This alone shows the value of the framework: it led us to discover things that were not previously known about measures. We went further and used the framework to formulate specific hypotheses about models and measures. Some of our hypotheses turned out to be true in most cases: that fatter-tail distributions would be more robust; that measures in the M_2 family would be more robust. Others turned out to be wrong: that measures in the M_4 family would be more robust; that M_3 would not be significantly better than M_1; that dynamic distributions would be less robust than static distributions. These results open the door to additional hypotheses and discoveries about evaluation based on user models.

On a personal note, we will confess to beginning this study with an unease about DCG—we felt that the discount was perhaps too flat to model real users, that its user model was ad hoc, and that it was unclear what it really measured. We come out of it with a newfound appreciation of DCG. The stopping probability may still be ad hoc (it is not a formal probability distribution), but it is much more clear to us that the discount is not too flat and that it actually measures something very useful. Furthermore, it is a highly robust measure; within our framework, the fact that it uses M_2 and a fat-tail distribution predicts that it would be. To answer the question we posed in Section 1, perhaps this is why DCG has continued to find such wide use: not due to inertia or familiarity, but because it really is a useful user-centered measure of system effectiveness.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under grant number IIS-1017026. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Halan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of WSDM*, pages 5–14, 2009.

[2] Leif Azzopardi, Kalervo Jarvelin, Jaap Kamps, and Mark D. Smucker, editors. *Proceedings of the SIGIR 2010 Workshop on the Simulation of Interaction: Automated Evaluation of Interactive IR*, 2010.

[3] Stefan Buettcher, Charles L.A. Clarke, and Ian Soboroff. The TREC 2006 Terabyte Track. In *Proceedings of TREC*, 2006.

[4] Ben Carterette and Paul N. Bennett. Evaluation measures for preference judgments. In *Proceedings of SIGIR*, 2008. To appear.

[5] Ben Carterette, Paul N. Bennett, D. Maxwell Chickering, and Susan T. Dumais. Here or there: Preference judgments for relevance. In *Proceedings of ECIR*, pages 16–27, 2008.

[6] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expceted reciprocal rank for graded relevance. In *Proceedings of CIKM*, 2009.

[7] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. Preliminary report on the trec 2009 web track. In *Proceedings of Text Retrieval Conference (TREC-2009)*, 2009.

[8] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of SIGIR*, pages 659–666, 2008.

[9] Gordon V. Cormack, Christopher R. Palmer, and Charles L.A. Clarke. Efficient construction of large test collections. In *Proceedings of SIGIR*, pages 282–289, 1998.

[10] Kalervo Jarvelin and Jaana Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[11] Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. Evaluation over multi-query sessions. In *Proceedings of SIGIR*, 2011. To appear.

[12] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Info. Sys.*, 27(1):1–27, 2008.

[13] Stephen E. Robertson. A new interpretation of average precision. In *Proceedings of SIGIR*, pages 689–690, 2008.

[14] Stephen E. Robertson, Evangelos Kanoulas, and Emine Yilmaz. Extending average precision to graded relevance judgments. In *Proceedings of SIGIR*, pages 603–610, 2010.

[15] Mark E. Rorvig. The simple scalability of documents. *JASIS*, 41(8):590–598, 1990.

[16] Tetsuya Sakai and Stephen Robertson. Modelling a user population for designing information retrieval metrics. In *Proceedings of EVIA*, pages 30–41, 2008.

[17] Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of SIGIR*, pages 315–323, 1998.

[18] Ellen M. Voorhees and Donna Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 1–24, 1997. NIST Special Publication 500-240.

[19] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of SIGIR*, pages 587–594, 2008.

[20] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. Expected browsing utility for web search evaluation. In *Proceedings of CIKM*, 2010. To appear.

[21] Yuye Zhang, Laurence A. Park, and Alistair Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Inf. Retr.*, 13:46–69, February 2010.