Simulating Simple User Behavior for System Effectiveness Evaluation

Ben Carterette*, Evangelos Kanoulas‡, Emine Yilmaz†
carteret@cis.udel.edu, e.kanoulas@shef.ac.uk, {eminey@microsoft.com, eyilmaz@ku.edu.tr}

* Department of Computer & Information Sciences, University of Delaware, Newark, DE
‡ Information School, University of Sheffield, Sheffield, UK
† Microsoft Research, Cambridge, UK & Koc University, Istanbul, Turkey

ABSTRACT

Information retrieval effectiveness evaluation typically takes one of two forms: batch experiments based on static test collections, or lab studies measuring actual users interacting with a system. Test collection experiments are sometimes viewed as introducing too many simplifying assumptions to accurately predict the usefulness of a system to its users. As a result, there is great interest in creating test collections and measures that better model user behavior. One line of research involves developing measures that include a parameterized user model; choosing a parameter value simulates a particular type of user. We propose that these measures offer an opportunity to more accurately simulate the variance due to user behavior, and thus to analyze system effectiveness to a simulated user population. We introduce a Bayesian procedure for producing sampling distributions from click data, and show how to use statistical tools to quantify the effects of variance due to parameter selection.

Categories and Subject Descriptors: H.3.4 [Information Storage and Retrieval] Performance Evaluation

General Terms: Experimentation, Measurement

Keywords: information retrieval, test collections, evaluation, simulation, statistical analysis

1. INTRODUCTION

There are two broad classes of information retrieval system effectiveness evaluation: the *systems-based* approach, using a test collection comprising canned information needs and static relevance judgments to compute evaluation measures such as precision and recall; and *user studies*, in which actual users are observed and measured in controlled interactions with a system. Both have strengths and weaknesses. Systems-based evaluations are fast, repeatable, and relatively inexpensive (since the data can be reused many times), but they make many simplifying assumptions about tasks, relevance, and so on. User studies can answer questions about user behavior and user interaction with systems that cannot be touched by systems-based evaluation, but because there is so much variance in user populations they tend to be either expensive and difficult to analyze, or require simplifying assumptions about the system and the types of information needs and tasks it will be used for.

There is increasing interest in better modeling user needs and user interaction with an engine in systems-based effectiveness evaluations [11]. Some recent directions based on test collections include the TREC Web track's diversity task, which models different intents for a query and penalizes redundancy in the retrieved results, and the TREC Sessions track, which attempts systems-based evaluation of a sequence of query reformulations. Test collection work dovetails with the development of evaluation measures such as *rank biased precision* (RBP) that use test collections and relevance judgments but also incorporate explicit models of user behavior [12]. RBP models a user stepping down a ranked list and deciding whether to stop with probability p. The probability that a user stops at rank k is then $(1-p)^{k-1}p$, which is the geometric probability density function. RBP itself is computed as an expectation over ranks for some value of p.

These measures are typically evaluated with fixed parameter values. If we are to see them as simulating users, using fixed values is akin to a user study with just one user at one fixed point in time. Using a geometric distribution captures the idea that the user might choose different stopping points for different needs (at random in the case of RBP), but using a single parameter value cannot capture the idea that the distribution of stopping points might vary by user, or that even a single user might have a different distribution at some times. In other words, though these measures are intended to model users better, in practice they simplify the user population so much that there is little difference between them and traditional systems-based measures except that they provide a tunable weighting of ranks.

One way to think about the distinction between systems-based evaluation and user studies is in terms of the *bias-variance tradeoff*: a simpler model has less variance but more bias; a more complex model has more variance but less bias. Systems-based evaluations make many simplifying assumptions that reduce variance—allowing statistical analysis with high power to find subtle differences between systems—but increase bias, so that they may be measuring something quite different from the user experience (some

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

work suggests this is definitely the case [15]). By using actual users, user studies have less bias, but they have substantially more variance because users can differ in how they interact with a system in so many ways.

In this work we propose to bring systems-based evaluations slightly closer to user studies through better simulation of a population of users. We do this by evaluating system performance with model-based measures over varying parameter values. We still simplify many of the sources of variance in real users by reducing them to a parameterized probability distribution, but we have a more complex model with more sources of variance than traditional systems-based approaches.

We first present well-known measures based on user models and user simulation in Section 2. We then describe how we can use logged user interactions to inform sampling distributions for model parameters (Section 3). In Section 4, we show how we will analyze the results of a study simulating users by varying parameter values—since traditional tools like the t-test are not suitable for this. Section 5 analyzes TREC test collections by this simulation procedure.

2. PREVIOUS WORK

Effectiveness measures based on user models nearly all take the same form: a user progresses down a ranked list one document at a time, deriving utility from relevant documents (with the relevance judgment coming from an assessor that may not be a user, as is typical in the systems-based setting), and stopping at some point modeled by a random variable. This work focuses on one such measure: *rank biased precision* (RBP). We describe that first, then briefly discuss some other similar measures as well as other approaches to selecting parameter values.

2.1 Rank biased precision

We introduced RBP [12] earlier. Its user model is that described above; it is implemented by the use of a "patience parameter" p that describes whether, at any given rank, the user decides to stop scanning results (with probability p) or to go on to the next rank (with probability $1-p$). The probability that a user ends up at rank k is then $P(k|p) = (1-p)^{k-1}p$, i.e. the probability that the user made $k-1$ decisions in a row to go to the next document followed by one decision to stop. RBP is calculated as a sum over ranks:

$$RBP = \sum_{i=1}^{n} rel_i (1-p)^{i-1} p$$

RBP has one parameter p that models user patience; p is drawn from the range $[0, 1]$. In our formulation, higher values of p indicate *less* patience, i.e. greater probability of stopping early. Probability density curves for selected patience parameters are shown in Fig. 1.

2.2 Other measures and models

The most commonly-used model-based measure is *discounted cumulative gain* (DCG) [8]. DCG is typically computed as a sum over ranks 1 to k, with each document contributing a gain according to its relevance discounted by a log function of the rank.

$$DCG@k = \sum_{i=1}^{k} \frac{gain(rel_i)}{\log_b(i+b-1)}$$

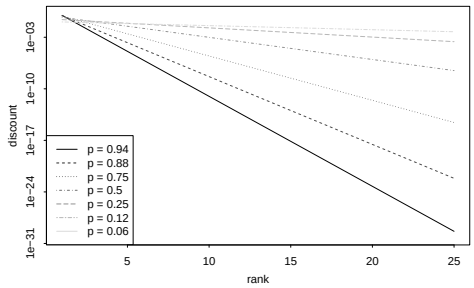

Figure 1: RBP discount $(1-p)^{k-1}p$ curves for different values of parameter p.

The parameters that define a user are the rank cut-off k, the log base b (which can be seen as modeling patience similar to RBP), and the gain values for each grade of relevance. By varying all these parameters together we could investigate the effect of sampling from a user space.

A number of other measures and models have been proposed in recent literature. These include *expected reciprocal rank* (ERR), which uses geometric distributions with non-zero probabilities only at the ranks of relevant documents [4]; α-nDCG, which penalizes redundancy according to a parameterized model [5]; the *intent aware* family of measures, which probabilistically model a diverse set of possible intents for a query [2]; and others [14, 19, 18]. All of these have at least one free parameter. Carterette organizes them into a framework partly based on features of their user modeling distributions [3].

In this work we focus exclusively on RBP, for two reasons: first, with just one parameter it is the simplest of all measures that have been described in the literature, and therefore a good starting point for this work. Second, because it uses proper probability distributions, it is very amenable to the type of analysis we would like to do. The methods we present here can be generalized to other measures (assuming data is available); we reserve this for future investigation.

2.3 Choosing parameter values

There has been previous work on the choice of parameter value and the analysis of the effects of making different choices. Nearly every paper that proposes such a measure spends at least some time justifying the existence of the parameter, comparing the models that use it to user data, and describing a way to choose a value.

Kanoulas & Aslam investigated the choice of discount and gain function for nDCG to minimize variance in an evaluation [9]. Their goal was to take advantage of a free parameters to improve the robustness of systems-based evaluation, whereas ours is to intentionally decrease robustness in order to draw conclusions about a wider (simulated) user base.

Zhang et al. develop a model of "gaps" between the ranks of clicks in a user log and use this model to find an expected patience parameter p for RBP [20]. We will similarly model the relationship between patience and clicks, but instead of using the expected value we will use the entire distribution. Zhang et al. also looked at the sensitivity of RBP to choice of parameter value using a bounding approach [21], with the goal of showing that RBP is not sensitive to parameter selection. We actually hope RBP *is* sensitive to parameter selection, in that we would like to be able to detect differ-

ences in effectiveness between systems on the basis of how users might respond to them.

The work described above tends to treat the existence of a free parameter as a problem that needs to be addressed or solved. We view it instead as an *opportunity* to model an additional source of variance that we would not have access to without a larger user study. Instead of using one fixed value, however, we vary it in a way that is informed by user data. Through this approach we are able to evaluate systems in the presence of noise due to a simulated user. In the best case, we can learn something about systems and their utility to a user population that we could not discover with a traditional systems-based evaluation or with a single fixed parameter value.

3. USER SAMPLING DISTRIBUTIONS

We first need a way to sample parameter values in a way that simulates sampling users. To this end we introduce a Bayesian model that starts with uniform distributions and updates them based on logged user data. The advantages of a Bayesian approach are that it produces a posterior distribution from which we can sample, and that we can use it regardless of whether we have very large amounts of user data (as in web search settings) or very little data (as in many domain-specific search settings). When a lot of data is available, the posteriors will closely track the data; when little is available, we are still able to use it to model users better than having no data at all while maintaining uncertainty due to the lack of data.

3.1 Patience distribution for RBP

Our goal in this section is to develop a way to compute a posterior distribution $P(p|E)$ for RBP's patience parameter p given a uniform prior distribution $P(p)$ and user log evidence E consisting of user queries and ranks at which clicks occurred. We will do this using Bayesian methods, so we start by applying Bayes' rule:

$$P(p|E) \propto P(E|p)P(p)$$

We will need a way to model click evidence given the patience parameter p. While there are many models of clicks in the literature (e.g. [20, 6]), none that we are aware of are directly applicable to our purpose. We believe this model—and our reason for using it—is novel.

In the language of Bernoulli trials, which are used to model coin flips and other binary random variables, we can treat a click as a "success" and the absence of a click as a "failure". Counting the number of failures before the first success gives a random variable that has a geometric distribution. For a fair coin, for instance, the probability of seeing zero tails before the first head is 0.5; the probability of seeing one tail before the first head is 0.25; two tails before the first head is 0.125; and so on. Similarly, if we model clicks as geometrically distributed with patience parameter p, the probability of zero unclicked documents before the first click (that is, the probability that the first click is at rank 1) is p; the probability of one unclicked document before the first click (the probability that the first click is at rank 2) is $(1-p)p$; and so on. Thus we can model the rank of the first click with a geometric distribution.

The geometric distribution is a special case of the negative binomial distribution, which can be used to model the total number of failures before a target number of successes is reached. If there are c clicks, the negative binomial distribution parameterized by p and the target number of failures r is defined as:

$$P(c|p,r) = \binom{c+r-1}{c}(1-p)^r p^{c-1}$$

When $r = 1$, $P(c|p, r = 1)$ reduces to the geometric distribution: the probability of $c = 1$ clicks is $(1-p)$; $c = 2$ clicks is $(1-p)p$; and so on. As our user model assumes a user progressing down a ranked list and deciding whether or not to stop, we will assume that the number of failures r is equal to the rank of the last click for a search minus the total number of clicks.

Since a negative binomial distribution is parameterized by the number of failures r and the probability p, we will return to $P(p|E)$ and treat it as marginal over the number of viewed but unclicked documents (assuming that all documents above the rank of the last click were seen):

$$P(p|E) = \sum_{r=0}^{\infty} P(p|r,E)P(r|E)$$

$P(r|E)$ is the probability that some user skips r documents in a search. $P(p|r,E)$ is the posterior distribution of patience parameter values for a given number of unclicked documents, and $P(p|r,E) \propto P(E|p,r)P(p|r)$. The prior $P(p|r)$ is still uniform. We compute $P(E|p,r) = P(c|p,r)$ using the negative binomial distribution above. One way to estimate $P(p|E)$, then, is to iterate over r, for each one sampling p from a uniform distribution and then computing $P(c|p,r)$ for every search in the click log for which there were r failures. Repeating this many times produces an estimate of the posterior distribution.

3.1.1 Fast computation

There is a simple way to obtain $P(p|r,E)$ with just one pass over the click data in E. We will define $P(p|r)$ as having a Beta distribution parameterized by α, β. Beta distributions are defined over the range $[0,1]$. When $\alpha = \beta$, they are symmetric. When $\alpha > \beta$, they are skewed left. When $\beta > \alpha$, they are skewed right. When $\alpha = \beta = 1$, the distribution is uniform. We will write:

$$P(p|r) = Beta(p|\alpha,\beta)$$

to denote a parameter with a Beta prior. Formally,

$$Beta(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{Z}$$

where Z is a normalization constant.

The Beta distribution is the *conjugate prior* of the negative binomial distribution. This means that if we start with the uniform prior, then sample from a negative binomial distribution (by looking at clicks), when we use those samples to update the prior the resulting posterior distribution is from the same family but with updated parameter values. For a negative binomial random variable parameterized by r total failures and probability p of success, if we have m distinct sampled instances and c_i is the number of successes in instance i, and if the prior of p is $Beta(\alpha,\beta)$, the posterior of p is $Beta(\alpha+\sum_{i=1}^{m} c_i, \beta+rm)$. Thus this gives us an easy way to update the distribution of p given data.

To estimate $P(p|r,E)$ from data, then, we look at all instances (searches) for which there were r unclicked doc-

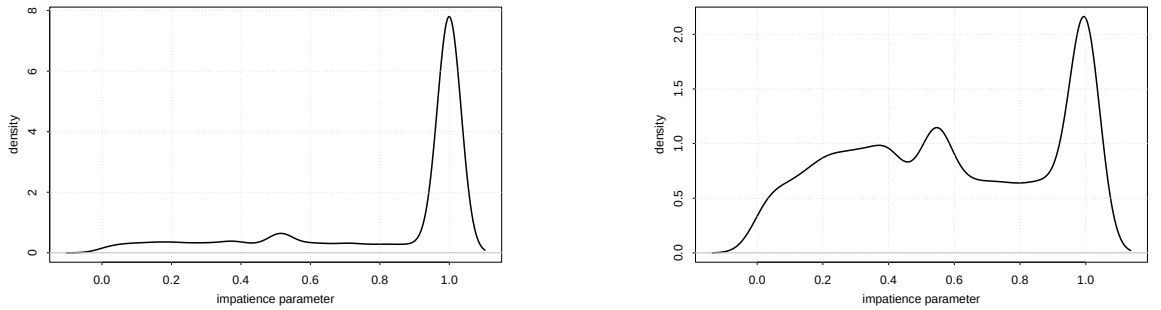

Figure 2: Empirical patience profiles for navigational (left) and informational (right) queries. Users exhibit much less patience for navigational needs, nearly always clicking only the first rank (if they click anything).

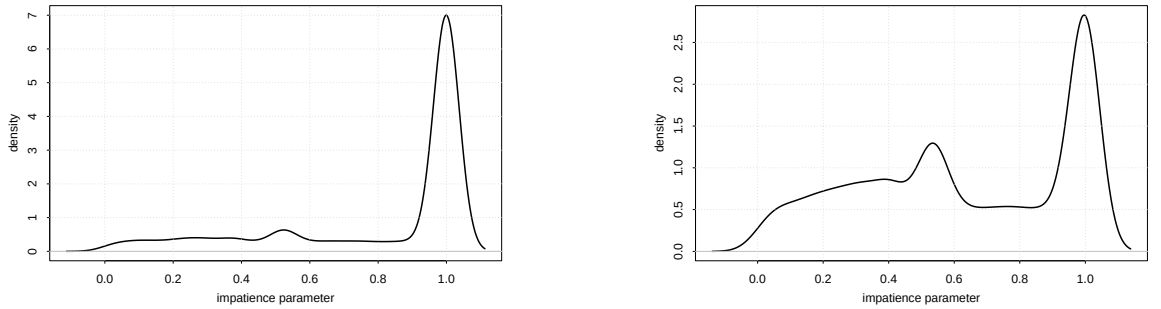

Figure 3: A second set of empirical patience profiles for navigational (left) and informational (right) queries. These are significantly different from those shown in Figure 2.

uments. For example, suppose $r = 1$ and there are m_1 searches with one unclicked document (and varying numbers of clicked documents). The posterior $P(p|r = 1, E)$ is $Beta(\alpha + \sum_{i=1}^{m_r} c_i, \beta + rm_r)$, where c_i is the number of clicks for instance i.

Therefore the full set of equations required is:

$$P(p|E) = \sum_{r=0}^{\infty} P(r|E)P(p|r, E)$$

$$P(r|E) = \frac{m_r + 1}{\sum_{i=0}^{\infty} m_i + 1}$$

$$P(p|r, E) = Beta\left(p|\alpha + \sum_{i=1}^{m_r} c_i, \beta + rm_r\right)$$

where m_r is the number of searches with r unclicked documents (estimated as the rank of the last click minus the total number of clicks), α, β are hyperparameters that we will set to 1, and c_i is the total number of clicks for the ith search with r unclicked documents. $P(r|E)$ is an empirical estimate with simple plus-one smoothing, and $P(p|r, E)$ is calculated with the Beta density function. Note that when no documents go unclicked ($r = 0$), there are two cases: either all ranks from 1 to the stopping rank were clicked, or nothing was clicked at all. These two cases are treated differently; for the former, the clicks can be used in the Beta distribution as usual, but for the latter there is no evidence with which to update priors. Thus there is a special case $P(r = 0, c = 0|E)P(p|r = 0, c = 0)$ with $P(p|r = 0, c = 0)$ modeled as a uniform distribution.

By training these distributions we obtain an empirical "patience profile". We can define this profile over all data, or over subsets of it. For instance, we could compute a patience profile for a particular user by just looking at that user's queries and clicks. We could compute a patience profile for classes of queries (such as informational and navigational) by just looking at the click data for queries in that class. Patience profiles for this case (with queries labeled "informational" or "navigational" by human assessors) are shown in Fig. 2. As expected, users exhibit much more patience for informational queries. Fig. 3 shows the same distributions learned from a different log (from the same search engine); they are similar in shape, though there are differences visible by inspection.

4. METHODS FOR ANALYSIS

Conclusions from both systems-based evaluations and user studies depend on statistical analysis of measurements on systems, topics, and (in the latter case) users. Systems-based evaluations typically use a statistical hypothesis test such as the t-test or a non-parametric alternative such as the Wilcoxon signed rank test to make an inference about the significance of the result. User studies vary tremendously in their design and analysis, but one typical approach is a Latin-squares design in which each user interacts with one or more systems on one or more topics, but never with the same topic twice. The typical statistical analysis for this design is the analysis of variance (ANOVA) or non-parametric alternatives like the Friedman test that analyze variance due to both topics and users.

The t-test and similar tests are not adequate for our goal of introducing user-model variance into a systems-based evaluation. Those tests model a topic sample as a so-called "random effect" that introduces variance to the measure,

but they can only model one random effect. If we are to model users, we will have at least two random effects: the topic sample and the values of the parameters that model users. In other words, the parameters can be viewed as random variables, and we can view the selection of a value as sampling from some distribution, just as a user study would sample users from some population.

These two sources of variation are different in an important way: we generally have access to only one sample of topics and we cannot sample another set without incurring a fairly high one-time cost of relevance judging. But we can sample as many parameter values as we like; there is almost no cost to sampling them. Studying variance due to the latter is therefore somewhat easier than studying variance due to the former, and it is useful to study parameter values in isolation of topic sample variance.

4.1 Marginal distribution analysis

As we have discussed, RBP and other measures in Section 2 are typically computed by selecting one value for the parameter(s), then evaluating all systems and topics with that value. The value may be informed by data, but it usually does not reflect variability in the data. To do that, we could instead sample many values of the parameters from distributions such as those in Figure 2, compute the measure for every value, and produce a *marginal distribution* for that measure. We can compare two of these distributions to investigate the effect of variability in the user space.

For the sake of generality, we will use θ to denote a vector of parameters and $P(\theta)$ to denote a distribution over that vector. For RBP $\theta = p$ and $P(p)$ is a distribution over $[0, 1]$. Sampling from the distribution $P(p)$ simulates sampling a user that interacts with the system as described by the RBP model. Each value we sample from these distributions can be used to compute RBP for a system; over a number of samples we can create a histogram of RBP values.

In general, we can form a marginal distribution over values of a measure M parameterized by θ by averaging over parameter vectors θ:

$$P(M = m) = \int P(M = m|\theta)P(\theta)d\theta$$

where $P(M = m|\theta) = 1$ if $M = m$ for parameters θ and 0 otherwise. We can estimate the distribution by Monte Carlo simulation: sample values of θ and use them to calculate M; over many trials the distribution $P(M)$ emerges.

This basic procedure allows us to form many other marginal distributions. We can ask questions such as:

- Over all possible choices of a parameter, what is the probability that one system is better than another? (i.e. $P(M_1 > M_2)$)?
- Over all possible choices of a parameter, what is the probability that the magnitude of the difference in performance is greater than some threshold (i.e. $P(M_1 - M_2 > t)$)?
- If there are parameters such that $M_1 > M_2$ and others such that $M_2 > M_1$, what is the probability that the magnitude of the former is greater than the magnitude of the latter?
- Over all possible choices of a parameter, what is the probability that the percent difference in performance is greater than some threshold (i.e. $P(\frac{M_1 - M_2}{M_1} > t)$)?

As an example, suppose we are evaluating two systems on a single topic by RBP. The relevance of the documents retrieved by the two are:

$$S_1 = \begin{bmatrix} R & N & N & N & N & N & N & N & N & N \end{bmatrix}$$
$$S_2 = \begin{bmatrix} N & R & R & R & R & R & R & R & R & R \end{bmatrix}$$

Depending on the choice of parameter value, we would draw different conclusions. If $p = 0.2$, we would conclude that S_2 is more than three times better than S_1. If $p = 0.8$, we would conclude that S_2 is four times worse than S_1. For $p = 0.5$, they are about the same.

Let us instead consider their performance over the entire range $p \in (0, 1)$. For now we will assume that $P(p)$ is uniform over that range. Fig. 4 shows three RBP histograms: the first shows the two marginal distributions of RBP over p. The second shows the distribution in the difference in RBP; depending on the parameter value, there is a 50% chance that the first system will be better than the second, but when that system is better it is likely to be a greater magnitude difference than when the reverse is true. The third shows the distribution in the percent difference in RBP; while 50% of the mass is to the right of 1 (meaning S_1 is better), the values of p for which S_2 is better can result in a much greater percent difference in performance. If we believe a user is as likely to be impatient as patient, we may want to deploy S_1 just so those impatient users will see at least one relevant document.

4.2 Mixed-effect models

Marginal distribution analysis looks at variance due to the choice of parameter value. It ignores variance due to the topic sample, but this is the variance we model when we perform statistical hypothesis tests such as the t-test. Analyzing a user study with an ANOVA takes variance from both topics and users into account; we would like to use a similar approach in our analysis. Because our "users" are such simple objects, we can use more powerful tools than ANOVA. A general *mixed-effect model* (of which both ANOVA and the t-test are special cases) can incorporate arbitrary sources of evidence into an analysis [7]; in this section we show how to set up and interpret a mixed-effect model.

Specifically, we would like to model the parameter as a random effect along with the topic sample. This would take into account the idea that users may have varying patience depending on their information need and other random variables. The standard t-test/ANOVA linear model is:

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (1)$$

where y_{ij} is the value of an evaluation measure calculated on topic j for system i, μ is a population effect or model intercept, α_i is the effect of system i, β_j is the effect of topic j, and ϵ_{ij} is the residual error, which subsumes system/topic interaction effects. One way to understand the topic effect is as a linear model with an intercept (but no slope) that is dependent on the topic number j; then fitting the model involves fitting both a topic model to the random topic sample as well as the overall model to the measures.

A mixed-effect model based on the ANOVA model but including an additional continuous-valued source of variance from the parameter could be expressed as:

$$y_{ijk} = \mu + \alpha_i + (\beta_j + \phi_j p_k) + (\kappa_{ij} + \gamma_{ij} p_k) + \epsilon_{ijk} \quad (2)$$

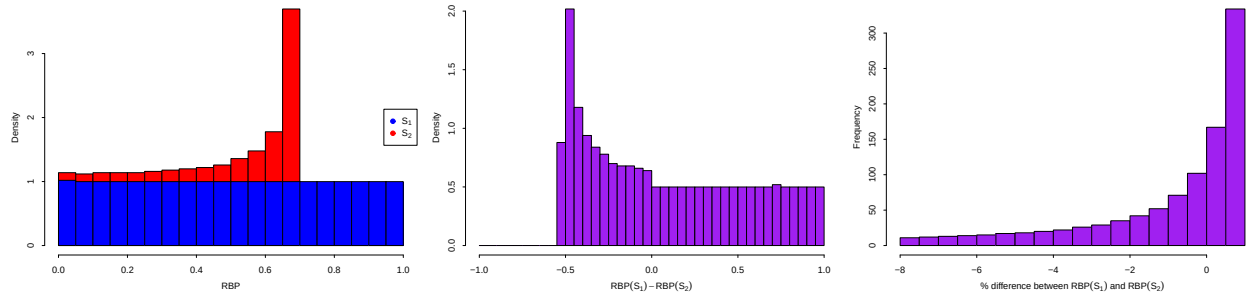

Figure 4: Distributions of RBP (left), difference in RBP (center), and percent difference in RBP (right) over uniform distributions of parameter p.

where y_{ijk} is the measure calculated on topic j, system i with parameter value p_k. In this formulation, the topic effect is itself a linear model with intercept β_j and slope ϕ_j (which vary by topic), and furthermore there is a separate system/topic interaction effect modeled as a line with intercept κ_{ij} and slope γ_{ij} (which vary by topic and system together). If topic effects dependent on p or system/topic interaction effect dependent on p wash out the system effects independent of p, we will not be able to detect a significant difference between systems.

4.2.1 Fitting mixed-effect models

Here we describe how to fit a mixed-effect model in the statistical programming environment R. While there are formulas that one can use, they are quite complicated and fairly non-transparent. We hope it will be more useful to show how to use a freely-available open source library to fit them and at the same time provide some intuition about the relationship between linear regression, the t-test, ANOVA, and mixed-effect models that is not immediately evident from looking at the formulas.

One can become convinced that the t-test is a special case of ANOVA, which in turn is a special case of linear regression, by the following analysis. First we obtain some data; this may be actual IR experimental data or randomly generated (which is useful for illustration). The data must be in an R data frame with one row for each system/topic pair.

```
data <- data.frame(y = rnorm(25*2),
                   system = as.factor(rep(1:2, each=25)),
                    topic = as.factor(rep(1:25, 2)))
```

This samples 50 numbers from a standard normal distribution to simulate an evaluation being performed for two systems over 25 topics each. Each sampled number is associated with a system number (1 or 2) and a topic number (from 1 to 25).

Once the data has been generated, the following procedures fit the linear model (Eq. 1) and result in equivalent inferences:

```
t.test(y ~ system, paired=T, data)
summary(aov(y ~ system + Error(topic/system), data))
summary(lm(y ~ system + topic, data))
```

The absolute value of the t statistic output by the `t.test` is equal to the square root of the F value output by the `aov` ANOVA analysis; the p-values are identical. The t statistic on the coefficient for `system2` output by the linear regression `lm` is equivalent to the t statistic from the `t.test`. These equivalencies can be observed across any two systems over n topics generated by any process. This demonstrates the utility of the linear model in analyzing evaluation results.

A mixed-effect linear model can be fit using the `lmer` function in the `lme4` package. The following is equivalent to the `t.test`, `aov`, and `lm` above:

```
lmer(y ~ system + (1|topic), data)
```

The t value for the `system2` fixed effect is again identical to the t values produced by the t-test and linear regression and to the square root of the ANOVA F value. The syntax makes a clear distinction between fixed effects and random effects, and furthermore allows much more flexibility in modeling random effects.[1]

Given that a t-test is an instance of a linear model, it is a short leap to including measure parameters in a linear model. When we have multiple measures for each system/topic pair based on using different parameter values, we fit the more complex mixed-effect linear model (Eq. 2):

```
lmer(y ~ system + (p|topic/system), data)
```

This sets up random intercepts for topics and system/topic pairs, and random slopes for both based on p as described above. This model allows us to analyze variance due to the topic sample and variance due to the choice of p within a system/topic pair.

4.2.2 Interpreting mixed-effect models

The output of R's `lmer` function (with two systems evaluated by RBP over 25 topics with 50 different parameter values sampled from a uniform distribution) looks like this:

```
   AIC   BIC logLik deviance REMLdev
 -6243 -6190   3130    -6271   -6261
Random effects:
 Groups       Name        Variance  Std.Dev. Corr
 system:topic (Intercept) 0.0229323 0.151434
              p           0.2788552 0.528067 -0.401
 topic        (Intercept) 0.0000000 0.000000
              p           0.1149877 0.339098   NaN
 Residual                 0.0037814 0.061493
Num obs: 2500, groups: system:topic, 50; topic: 25
```

[1]It is better to think of the random effect as being `(1|topic/system)` for congruence with `aov`'s syntax. However, because of inconsistencies in implementation, `lmer` cannot use that syntax unless there is more than one measure for each system/topic pair.

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.34353    0.02853  12.043
system2      0.01068    0.03949   0.271
```

Consider this output from the bottom to the top. First, as with the linear model, the `lmer` function gives a coefficient and a t-statistic for each system in the experiment in the `Fixed effects` section. We can therefore evaluate the significance of a system with variance due to both topics and modeled users. In this case the t-statistic is quite low, suggesting that the systems are not significantly different. Unfortunately it does not provide the number of degrees of freedom to use to compute a p-value, but as long as the design is fully nested (that is, every system evaluated on every topic with every sampled parameter value), we can use $df = n - m + 1$ where n is the number of topics and m the number of systems. Thus a p-value for the hypothesis that the second system is better is $P(t > 0.271|df = 24) = 0.394$.

Second, the `Random effects` section gives the variance in the measure due to the parameter value in system/topic groups and topic groups. The higher this variance is, and in particular the greater it is relative to residual variance, the greater effect the selection of parameter has on the final results. In this example we can see the variance due to system/topic interaction is greater than residual variance ($0.023 > 0.004$); the variance due to parameter p within system/topic groups is much greater than residual variance ($0.279 > 0.004$); and the variance due to parameter p in topic groups is greater than residual variance ($0.115 > 0.004$). This clearly shows that the parameter is responsible for a great deal of variance in the measure separately from the system and the topic.

Finally, the first two lines give some statistics about the goodness of the model fit. These cannot easily be interpreted absolutely, but they can be interpreted relative to other models. Fitting a model without random slope due to p to the same data gives the following:

```
    AIC    BIC logLik deviance REMLdev
-951.2 -922.1  480.6   -969.2  -961.2
```

The goodness-of-fit for the model including p is much better than the model not including p. The `anova` function tests whether the difference is significant:

```
> anova(m1, m2)
Models:
m2: y ~ system + (1 | topic/system)
m1: y ~ system + (p | topic/system)
   Df     AIC     BIC logLik Chisq Chi Df Pr(>Chisq)
m2  5  -959.2  -930.1  484.6
m1  9 -6253.2 -6200.8 3135.6  5302      4  < 2.2e-16
```

So including p in the model gives a significantly better fit (p-value $< 2.2 \times 10^{-16}$ by the χ^2 test).[2] This is not too surprising but we note the implication that changing p does not just result in constant differences in the magnitude of effectiveness difference.

[2]The goodness-of-fit statistics reported by `anova` are slightly different from those reported by `lmer`. This is because `anova` re-fits the models with a different likelihood function. In practice this does not make much difference.

dataset	task	topics	runs submitted
Terabyte 2005	ad hoc	751–800	58
Terabyte 2005	named page	601–872	42
Terabyte 2006	ad hoc	701–850	61
Terabyte 2006	named page	901–1081	43

Table 1: TREC Terabyte track data for 2005–2006.

5. EXPERIMENTAL ANALYSIS

We will analyze the application of the proposed simulation and analysis to TREC evaluation. The idea is that we simulate users who have TREC-style information needs but who vary in their patience according to the distributions shown in Fig. 2: many will stop after rank 1, but a significant fraction will go to rank 100 or deeper. The mean patience is $p = 0.83$ for navigational queries and $p = 0.59$ for informational queries.

The primary question we are interested in is this: when we perform our simulated user study and analyze results using marginal distribution analysis and mixed-effect models, do we learn anything about systems that we did not already know from purely systems-based experiments with unparameterized evaluation measures and t-tests? If not, then traditional systems-based evaluation is good enough for all intents and purposes. Otherwise, we should consider adding this approach to our evaluation toolkit.

A secondary question is whether evaluation results differ depending on the log used to learn a sampling distribution. The distributions in Figures 2 and 3, learned from two separate logs, are similar by inspection, but they are significantly different by statistical goodness-of-fit tests. Thus the question: if two different logs produce two different distributions, will we draw different conclusions about the evaluation, or are these methods robust?

5.1 Data

For this analysis we focused on evaluating two tasks: an ad hoc retrieval task (meant to model informational needs) and a named-page finding task (to model navigational needs). For this we used data from the TREC Terabyte 2005 and 2006 tracks, described in Table 1. Note that we removed 19 Terabyte 2006 ad hoc manual runs, since they were evaluated over a smaller subset of topics than the automatic runs from the same year.

Our patience profiles are derived from the query log of a commercial search engine collected in January 2009. The log consists of queries, their frequency in the log and the ranks at which clicks occurred for each query. All queries in the log have been labeled as "navigational" or "informational" by human assessors. We randomly divided the queries into two sets of training (used in Fig. 2) and testing (used in Figure 3) with approximately equal size. We form patience profiles from the navigational and informational classes separately; we sample from the navigational profile (Fig. 2, left) to evaluate the named page task and from the informational profile (Fig. 2, right) to evaluate the ad hoc task. We will use Fig. 3 to test the effect of the sampling distribution.

5.2 Marginal distribution analysis

All of our marginal distribution analysis is based on the distribution of a measure over the parameter space. For any pair of systems, we can answer any of the questions in

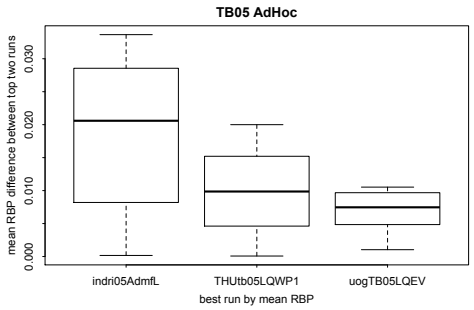

Figure 5: Depending on parameter value, the "best" ad hoc system as measured by RBP is one of these three. Boxes show the distribution of differences between each system (given that it is the best) and the second-best system.

Section 4.1; for analyzing a set of systems we investigate questions about orderings. To do so, we sample a parameter value, then evaluate all systems in the set with that value and rank them. Over many samples we form histograms of measurements on the rankings. We select some interesting results from our four datasets; much more analysis is possible, but space is limited.

Which system provides the best user experience? Typically the "best" system is the one with the greatest value according to some evaluation measure. But with enough variance over the user space, some users may find a different system much more useful to their particular needs. By varying parameters in a model-based measure, we may obtain a *distribution* of "best" systems, with different systems being "best" for different user classes.

When we vary RBP's patience parameter according to the informational distribution, we find that there are three different TB 2005 ad hoc systems that could claim to be best: indri05AdmfL for very patient users in roughly the range $p \in (0, 0.10]$, uogTB05LQEV for patient users in the range $p \in (0.10, 0.82]$, and THUtb05LQWP1 for impatient users—33% of the distribution—in the range $p \in (0.82, 1.00]$. indri05AdmfL has the highest mean average precision of any system, and uogTB05LQEV has the highest precision@5 of any system, but THUtb05LQWP1 is no better than third place for any of the traditional TREC evaluation measures. Thus we have already learned something new: a system that does not rank better than third place by any standard measure actually provides the best experience for a large proportion of our simulated users. It is particularly interesting that it is better for these users than the system with highest precision@5, since that is usually thought to be a good measure to model impatient users.

Fig. 5 shows the distribution of differences in mean RBP between each of the "best" systems and the second-best system for the same parameter value. When indri05AdmfL is best, the difference between it and the second-best system is relatively large, but it is only best for a very small group of users. uogTB05LQEV has a much smaller difference, though it is useful to a much larger group of users. THUtb05LQWP1 seems to strike the best balance between being useful to a wide range of users while also being likely to be substantially better than the next-best system for those users. Again, this is something that no traditional evaluation paradigm could tell us.

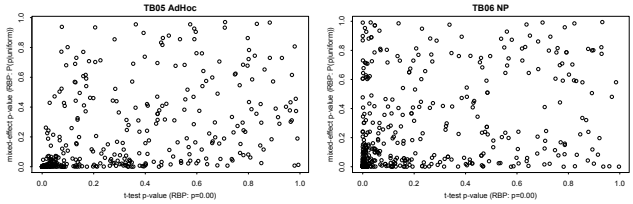

Figure 7: RBP p-values for TB05 ad hoc (left) and TB06 named page (right) from a paired two-sided t-test with a fixed parameter $p = 0.59$ or $p = 0.83$ (respectively) versus the mixed-effect model with 25 parameter values sampled from the respective distributions in Fig. 2.

For TB 2006 ad hoc, there are again three systems that vie for first place: indri06AtdnD ($p \in (0, 0.17]$), uogTB06S50L ($p \in (0.17, 0.99]$), and JuruTWE ($p \in (0.99, 1.00]$). Again we have one system—uogTB06S50L—that covers most user patience profiles, and one—JuruTWE—that covers a large part of the users, including the most impatient, and that is not among the top three systems by any traditional measure. For the named page tasks, there is much more consistency; a single system is best for the vast majority of the user space.

How much does the ranking of systems change with parameter value? Besides the "best" system, we are often interested in an overall ranking of systems by some measure. As the parameter value changes, the ranking will change as well. We computed Kendall's τ rank correlation between the ranking of systems for each parameter value and the ranking by a single point parameter value ($p = 0.59$ for ad hoc; $p = 0.83$ for named page) for the same measure. Histograms of τ values are shown in Fig. 6 for both tracks (with distributions for two tasks superimposed). Both tasks have relatively stable rankings, and the distributions are similarly-shaped for both tasks. But there is enough variability—particularly in the 2005 runs—that τ frequently drops below the usual standard of 0.9 for two rankings to be "equivalent" [16]. In some cases τ can drop nearly as low as 0.5, which reflects a major difference in system ranking.

5.3 Mixed-effect model analysis

We used the mixed-effect model to test for significant differences between pairs of systems as described in Section 4.2.2. We randomly sampled 500 pairs of runs to use in experiments from each of the four settings: TB 2005 ad hoc (1,653 total pairs), TB 2006 ad hoc (1,830 total pairs), TB 2005 named page (861 total pairs), and TB 2006 named page (903 total pairs).

Fig. 7 compares p-values reported by a paired two-sided t-test for RBP (with parameter fixed according to task) to those reported by a mixed-effect model with 25 randomly sampled parameters from the distributions in Fig. 2. The left plot (for TB 2005 ad hoc) has a linear correlation of 0.55; the two tests agree about significance in 81% of the pairs. The mixed-effect model tends to find more pairs significant, with 16.2% of its pairs not significant by a t-test. The right plot (for TB 2006 named page) has a lower linear correlation of 0.38. The two tests agree about the significance of about 70% of the pairs. In this case, 17.2% of the pairs are significant by a t-test but not by the mixed-effect model.

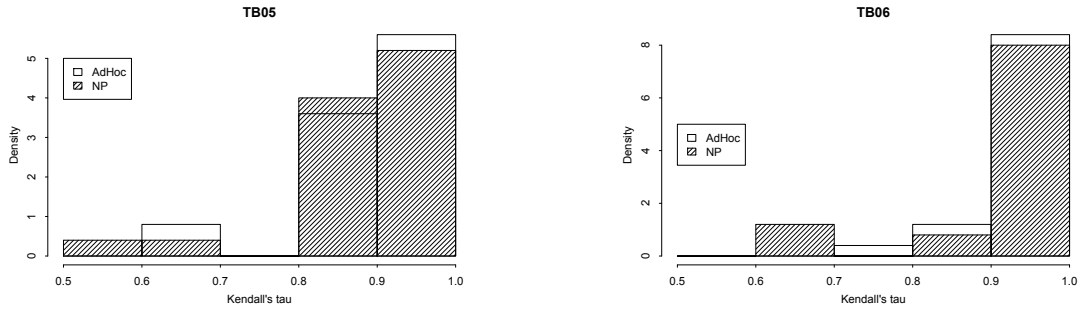

Figure 6: Histograms of Kendall's τ rank correlations for varying RBP patience parameter; Terabyte 2005 (left) and Terabyte 2006 (right) with two histograms for the two tasks. Rankings are relatively stable, but there are some substantial differences.

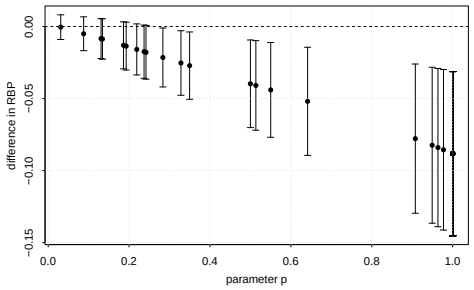

Figure 8: Mean difference and 95% c.i. for RBP for two systems over 50 parameter values sampled from the navigational distribution in Fig. 2. For any given parameter value, the systems are likely to be significantly different (by a t-test), but they are not significantly different when the space of parameters is modeled (with the mixed-effect model).

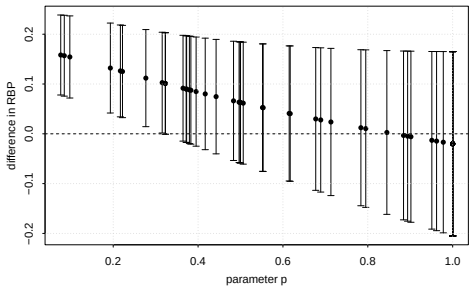

Figure 9: Mean difference and 95% c.i. for RBP for two systems over 50 parameter values sampled from the informational distribution in Fig. 2. For any given parameter value, the systems are not likely to be significantly different (by a t-test), but they are significantly different when the space of parameters is modeled (with the mixed-effect model).

The fact that the t-test and mixed-effect model agree on the majority of pairs is good: it means that we can indeed use this model and trust that it will give results that match previous experiments in general. The fact that they disagree is also interesting, as is the fact that some pairs that are not even close to the significance threshold by one method are very close by the other. These are the pairs that offer the opportunity to learn something new, so we selected some of them to try to identify causes.

Our first pair (NTUNF4 and humTN05pl from TB 2005 named page) is significant by a t-test but not by the mixed-effect model. Fig. 8 shows the difference in mean RBP over topics for each sampled p with 95% confidence intervals. Note that the difference is significant for most individual values, including the most likely in the navigational distribution. Yet these two systems are *not* significantly different by any traditional precision measure; from precision@5 to precision@1000, the t-test p-values are well above the 0.05 threshold. The mixed-effect model in this case reveals that the choice of parameter value contributes more to variability in RBP than differences between systems do.

Our second pair (TWTB05AD02 and ctfadhocaf1 from TB 2005 ad hoc) is not significant by a t-test but is significant by the mixed-effect model. Fig. 9 shows the difference in mean RBP over topics for each sampled p with 95% confidence intervals. In this case, the difference is *not* significant for values greater than about 0.33. Yet these two systems *are* significantly different by traditional IR measures, including MAP and precision@10. In this case the mixed-effect model can reveal that the selection of parameter value does not matter since the difference in system effects holds once variance due to that value is modeled.

5.4 Differences due to sampling distribution

To determine whether the sampling distribution affects the evaluation, we compared p-values from the mixed-effect model: for each pair of systems in the Terabyte 2005 ad hoc data, we calculated the p-value from a mixed-effect model based on sampling users from the distributions in Fig. 2 and the p-value from a mixed-effect model based on sampling users from those in Fig. 3.

The results are shown in Figure 10. It is clear that there is a high degree of concordance between the results of tests based on the two logs: the linear correlation is 0.97, and the two agree on 97% of the pairs (39% are not significant with both logs; 58% are significant with both logs). For the remaining 3% on which the two logs result in disagreement, the p-values are very close to the 0.05 significance boundary in both cases. From this we tentatively conclude that the log does not make a major difference.

6. CONCLUSION

We have presented a way to simulate simple user behavior for a systems-based evaluation, starting from click data and a simple model of users. We first use click data to create a distribution of parameters modeling a user that steps

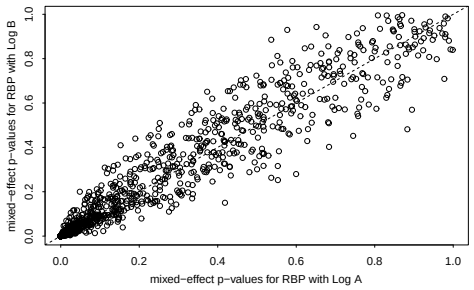

Figure 10: Comparison of p-values from mixed-effect models based on sampling users from distributions in Fig. 2 and Fig. 3. The linear correlation is 0.97.

down search results one rank at a time, deciding whether or not to stop after each one. That distribution can model a single user's behavior changing from search to search, or a wider user population. Sampling from that distribution simulates sampling users; using a value sampled from the distribution together with an evaluation measure based on a user model simulates a user interacting with a system. We demonstrated how to use this to analyze TREC runs and learn something about their potential utility to a user base.

Of course, this simulation procedure is still very far from a true user study. It offers no opportunity for serendipitous discoveries that might occur while exploring data from a user study. The "users" are highly simplified mathematical objects with no will or motivation of their own, and no ability to provide useful feedback that might inform future research directions. We do not believe that we can replace user studies; we only hope to better model the user experience in systems-based evaluations to more thoroughly explore questions of system utility to users.

In the future, we will want to use user data other than clicks, both for informing parameter sampling distributions and for building probabilistic models that measures can be based on. The more we can incorporate user data, the better we can simulate the actual user experience. This data might include dwell time, reformulations, mouse movements, and features such as those used for learning to rank [1, 13, 17]. Measures such as *expected browsing utility* (EBU) [19] and session precision [10] are starting to model much more detailed user behavior for systems effectiveness evaluation.

This work can be extended to measures such as ERR and even more complex measures such as α-nDCG. These will require more information about users—for ERR, we need relevance judgments that we can relate to clicks; for α-nDCG we need to be able to update a redundancy penalty parameter from data. But they will allow us to perform deeper simulations of varying intents, the value of redundant information, the value of better relevant documents, and so on.

Acknowledgments: Research was sponsored in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. We also gratefully acknowledge the support provided by the European Commission grant FP7-PEOPLE-2009-IIF-254562.

7. REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of SIGIR*, pages 19–26, 2006.

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

[3] Ben Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of SIGIR*, 2011.

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

[5] 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.

[6] Georges E. Dupret and Benjamin Piwowarski. A browsing model to predict search engine click data from past observations. In *Proceedings of SIGIR*, pages 331–338, 2008.

[7] Julian J. Faraway. *Extending the Linear Model with R.* CRC Press, 2005.

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

[9] Evangelos Kanoulas and Javed A. Aslam. Empirical justification of the gain and discount function for nDCG. In *Proceedings of CIKM*, 2009.

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

[11] Evangelos Kanoulas, Paul Clough, Ben Carterette, and Mark Sanderson. Session track at trec 2010. In *Proceedings of SIGIR 2010 Workshop on Simulation of Interaction: Automated Evaluation of Interactive IR*, 2010.

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

[13] Filip Radlinski and Thorsten Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of SIGKDD*, pages 239–248, 2005.

[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] Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of SIGIR*, pages 225–231, 2006.

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

[17] Xuanhui Wang and ChengXiang Zhai. Learn from web search logs to organize search results. In *Proceedings of SIGIR*, pages 87–94, 2007.

[18] Yiming Yang and Abhimanyu Lad. Modeling expected utility of multi-session information distillation. In *Proceedings of ICTIR*, 2009.

[19] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. Expected browsing utility for web search evaluation. In *Proceedings of CIKM*, pages 1561–1564, 2010.

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

[21] Yuye Zhang, Laurence A. F. Park, and Alistair Moffat. Parameter sensitivity in rank-biased precision. In *Proceedings of ADCS*, 2008.