Model-Based Inference About IR Systems

Ben Carterette

Dept. of Computer & Info Sciences, University of Delaware, Newark, DE, USA
carteret@cis.udel.edu

Abstract. Researchers and developers of IR systems generally want to make inferences about the effectiveness of their systems over a population of user needs, topics, or queries. The most common framework for this is statistical hypothesis testing, which involves computing the probability of measuring the observed effectiveness of two systems over a sample of topics under a null hypothesis that the difference in effectiveness is unremarkable. It is not commonly known that these tests involve *models* of effectiveness. In this work we first explicitly describe the modeling assumptions of the t-test, then develop a Bayesian modeling approach that makes modeling assumptions explicit and easy to change for specific challenges in IR evaluation.

1 Introduction

Arguably the fundamental problem in IR is modeling the relevance of information to users. Vast amounts of effort have gone into developing features, model families, and optimization methods that can model relevance in a way that produces systems that are useful to users. Nearly as important is modeling the actual utility of these systems to the users that are supposed to benefit from them. This is the *effectiveness evaluation* problem, and it is traditionally based on a combination of effectiveness measures to estimate utility and statistical hypothesis testing to make inferences about the relative utility of different systems. But while IR research on relevance modeling looks into all aspects of the problem, much of the IR literature on hypothesis testing defers to basic work on statistics that is written conservatively, with the goal of providing solutions that make the fewest and weakest assumptions so as to be applicable to the widest range of cases. We argue that better inference is possible if we tailor our tests to the particular challenges of evaluating IR.

Just as relevance and retrieval models are based on features (of queries, of documents, of query/document pairs, of users, etc), evaluation by statistical hypothesis test is based on features as well. But while relevance models can be extraordinarily complex, evaluation models have remained very simple. Every significance test in wide use models an evaluation measure with at most two "features" along with an intercept and residual error. The models and features used in evaluation models are almost always hidden from the practitioner, however. Unless one has a deep familiarity with the t-test—a familiarity beyond what is presented in introductory textbooks—one may not be aware that it is

equivalent to performing inference in a simple linear model with categorical system ID and topic ID features. There is much room to tailor this model to specific evaluation scenarios in IR, but no statistics textbook will explain how to do that.

This paper proposes an explicitly model-based Bayesian framework for hypothesis testing and evaluation in general. Bayesian models have of course been used to model relevance, but to the best of our knowledge they have not been used in IR effectiveness evaluation. The advantage of the Bayesian framework is that it allows construction of tailored models for evaluation largely free from the details of how to perform inference in them (unlike the t-test). Bayesian inference is arguably more intuitive as well, as it comes down to the probability of a hypothesis being true rather than the probability of observing data given a null hypothesis (the p-value).

We begin in Section 2 by describing the use of models in traditional IR evaluation. We then present the Bayesian approach in Section 3, with Bayesian versions of the t-test along with a new Bayesian method for direct inferences about system effectiveness using relevance generated by a user model. In Section 4 we empirically analyze the Bayesian framework.

2 Traditional Model-Based Inference

In this section we summarize some previous work on models in evaluation and show how we use modeling assumptions in evaluation even though they may not be explicitly stated. Broadly speaking, models come into play at two points in evaluation: first, in effectiveness measures that model user utility or satisfaction, and second, in statistical tests of the significance of differences. A third way models come into play is in the use of data mined from interaction logs for evaluation, but that is outside the scope of this work.

2.1 Model-based evaluation measures

There has always been interest in using effectiveness measures to model and approximate utility to a user. Recent work along these lines often involves constructing an explicit probabilistic user model, then combining it with relevance judgments to summarize utility [2]. Some examples are *rank-biased precision* (RBP) [9], *discounted cumulative gain* (DCG) [7], *expected reciprocal rank* [3], *expected browser utility* (EBU) [16], α-nDCG [4], and others [12, 1]. In addition, traditional measures have had user models backfit to their mathematical expression [11, 17], showing that most measures at least suggest a user model.

In this work we will focus on just two measures: precision at rank k, modeling a user that will stop at rank k and derive utility from the relevant documents appearing about that rank; and RBP, modeling a user that steps down a ranked list, deriving decreasing utility from each subsequent relevant document. RBP can be expressed as

$$RBP = \sum_{k=1}^{\infty} y_k \theta^{k-1}(1-\theta)$$

The reason for focusing on these two is that they have simple, clear user models, and that they have light computational requirements as compared to measures like average precision (AP). Computational requirements are an unfortunate drawback to some of the methods we propose, particularly those in Section 3.2.

2.2 Statistical hypothesis tests

Statistical hypothesis tests use the idea that a set of topics is a sample from some larger population to test the hypothesis that a difference between two systems is "real" and cannot be ascribed to random chance. Many different tests are used by IR researchers; the most common are the sign test, the Wilcoxon signed rank test, the t-test, ANOVA, the randomization (exact) test, and the bootstrap test [13, 18]. Every test is in one way or another based on a model; every model has assumptions. The model and its assumptions are usually hidden to the practitioner, but understanding the model is key to understanding the test.

For this work we will focus on the t-test. We assume the basics of the t-test are well-known and not reiterate them here. Our interest is in its modeling assumptions: it is actually based on a linear model of a measure of the effectiveness y_{ij} of system j on topic i as a linear combination of an intercept μ, a "system effect" β_j, a "topic effect" α_i, and random error ϵ_{ij} that is assumed to be normally distributed with variance σ^2 [10]. The t-test model is therefore:

$$y_{ij} = \mu + \beta_j + \alpha_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$

In this notation y_{ij} is equal to a sum of effects, one of which (the errors ϵ_{ij}) are drawn from a normal distribution with mean zero and variance σ^2 (as indicated by the $\sim N(0, \sigma^2)$ notation). We can equivalently express this model as:

$$\widehat{y_{ij}} = \mu + \beta_j + \alpha_i$$
$$y_{ij} \sim N(\widehat{y_{ij}}, \sigma^2)$$

and even more compactly as:

$$y_{ij} \sim N(\mu + \beta_j + \alpha_i, \sigma^2)$$

Note that this is exactly the same linear model that is the basis of linear regression, and in fact it is the same linear model that is the basis of ANOVA as well. ANOVA is a special case of linear regression, and the t-test is a special case of ANOVA. This is not well-known among IR practitioners; we refer to Monahan [10], Gelman et al. [6], and Venables & Ripley [14] for deeper treatment of linear models from different perspectives.

Performing a t-test therefore involves estimating parameters $\mu, \beta_j, \alpha_i, \sigma^2$. In practice, a paired t-test only requires estimates the magnitude of the difference between two system effects ($\beta_1 - \beta_2$) and the error variance σ^2. The maximum likelihood estimates of $\beta_1 - \beta_2$ and σ^2 are the mean difference and variance of

differences in measure values respectively. If topics are sampled independently and identically (i.i.d.), the Central Limit Theorem says the estimate of $\beta_1 - \beta_2$ can be treated as having a normal distribution, and therefore $(\beta_1 - \beta_2)/\sqrt{\sigma^2/n}$ has a Student's t distribution with $n - 1$ degrees of freedom.

This statement of the t-test as a linear model makes its assumptions explicit:

1. errors ϵ_{ij} are normally distributed with mean 0 and variance σ^2 (normality);
2. variance σ^2 is constant over systems and topics (homoskedasticity);
3. effects are additive and linearly related to y_{ij} (linearity);
4. topics are sampled i.i.d. (independence).

The first three of these assumptions are almost certainly *false* in typical IR experiments. The reason is that IR effectiveness measures are discrete-valued and bounded in the range $[0, 1]$. Consider each assumption in turn:

1. Normality: normal distributions are unbounded, so our error distribution will give non-zero probability to values outside the range of the measure.
2. Homoskedasticity: very bad and very good systems necessarily have lower variance than average systems, simply because as the measure approaches 0 or 1 there are fewer ways it can vary.
3. Linearity: $\widehat{y_{ij}}$ can be outside the range $[0, 1]$ because there is no bounding of the linear combination. Also, there is at least one measure that is surely non-linearly related to topic effect (recall) and one that is non-additive (GMAP).

In this work we are not terribly concerned with the effect of these violations—in fact, the t-test is quite robust to them. Our point is to state them clearly so that we can begin to form alternative models for evaluation that more precisely capture aspects of IR effectiveness that are not captured by the t-test, and so that we can state exactly how the models differ from each other.

Non-parametric tests rely on modeling assumptions as well, though they are typically weaker than those of the linear model. Even tests like the randomization and bootstrap tests rely on modeling assumptions that may be false: both of those tests relax homoskedasticity to a weaker assumption of *exchangeability*, and trade the Gaussian error distribution for an empirical error distribution. They still assume a linear model and independence of topic effects.

3 Bayesian Inference

Our aim is to find a framework for testing hypotheses about systems that can be adopted for the specific challenges of IR evaluation. Our first effort is towards explicit model-based hypothesis testing: we introduce a fully Bayesian version of the linear model we presented above. We then introduce greater and greater complexity to show what the Bayesian framework can do.

3.1 Bayesian linear model

As discussed above, the t-test assumes a linear model with three effects and normally-distributed errors. For our Bayesian version, we will start with the same two assumptions. We will also introduce *prior distributions* for each model parameter. These prior distributions can be used to model any information we already have about the experiment. If we do not wish to make any strong assumptions, we can use non-informative priors. An example non-informative prior might be a uniform distribution over the entire real line. Better is a normal distribution with uncertain variance—i.e. the variance itself is a parameter with a prior distribution.

Thus our first attempt at a fully-Bayesian model is:

$$\begin{aligned}
\widehat{y_{ij}} &= \mu + \beta_j + \alpha_i \\
y_{ij} &\sim N(\widehat{y_{ij}}, \sigma^2) & \sigma &\sim 1/\sigma \\
\mu &\sim N(0, \sigma^2_{\text{int}}) & \sigma_{\text{int}} &\sim 1/\sigma_{\text{int}} \\
\beta_j &\sim N(0, \sigma^2_{\text{run}}) & \sigma_{\text{run}} &\sim 1/\sigma_{\text{run}} \\
\alpha_i &\sim N(0, \sigma^2_{\text{topic}}) & \sigma_{\text{topic}} &\sim 1/\sigma_{\text{topic}}
\end{aligned}$$

In words, each measure of system effectiveness on a topic is a sum of a population effect μ, a system effect β_j, and a topic effect α_i. Since we do not know anything about these effects *a priori*, we put prior normal distributions over them. Since we further do not know anything about the variances of those distributions, we use improper flat priors on the log scale (the non-informative Jeffreys prior).

To make inferences about systems, we need the posterior distribution of system effects: $P(\beta_j|y)$, where y is the effectiveness evaluation data. Obtaining the posterior distributions is best done by simulation, iteratively sampling parameters from their prior distributions, then updating posteriors based on the likelihood of the data. Monte Carlo Markov Chain simulation is a standard technique. We do not provide details here, as there are packages for general MCMC computation of posteriors available.

Once the posteriors have been computed, making inferences about the systems is relatively simple: we estimate the probability of a hypothesis such as $S_1 > S_2$ by estimating $P(\beta_1 > \beta_2)$ from our simulation data. Note that the Bayesian approach actually estimates the probability that a hypothesis is true (conditional on the model and the data) rather than the probability of observing the data under a null model (like the t-test and other traditional tests). We feel this has the additional advantage of being more intuitive than a p-value.

3.2 Direct inference about relevance

Effectiveness measures are themselves summaries of individual measurements on documents—relevance judgments. Instead of testing a hypothesis about a summarization of judgments by an effectiveness measure, the Bayesian framework allows us to model relevance *directly* according to some user model.

Let x_{ijk} be the judgment to the document retrieved by system j at rank k for topic i. We will model the judgments as coming from a Bernoulli distribution with parameter p_{ij}, essentially a coin flip biased by the system and topic. We will model p_{ij} using the linear model with a population effect, a system effect, and a topic effect, filtered through a sigmoid function to ensure the result is bounded between 0 and 1.

$$
\begin{aligned}
x_{ijk} &\sim \text{Bernoulli}(p_{ij}) \\
p_{ij} &= \exp(y_{ij})/(1+\exp(y_{ij})) \\
y_{ij} &\sim N(\mu + \beta_j + \alpha_i, \sigma^2) & \sigma &\sim 1/\sigma \\
\mu &\sim N(0, \sigma^2_{\text{int}}) & \sigma_{\text{int}} &\sim 1/\sigma_{\text{int}} \\
\beta_j &\sim N(0, \sigma^2_{\text{run}}) & \sigma_{\text{run}} &\sim 1/\sigma_{\text{run}} \\
\alpha_i &\sim N(0, \sigma^2_{\text{topic}}) & \sigma_{\text{topic}} &\sim 1/\sigma_{\text{topic}}
\end{aligned}
$$

While we still have y_{ij} in the model, it should no longer be thought of as a measure of effectiveness. Now it is a hidden variable that influences the probability that a document appearing at a particular rank is relevant. p_{ij} is a convenience variable that converts the real-valued y_{ij} to a probability in $[0, 1]$ by applying the sigmoid function.

As it turns out, however, y_{ij} can be congruent to an *estimate* of an effectiveness measure. If we restrict ranks to $k \leq K$ (for some constant K) the parameter p_{ij} is an estimate of the precision at rank K of system j on topic i. To see this, think of precision as the expectation that a randomly-chosen document in the set of K is relevant. That expectation is $\sum_{k=1}^{K} x_{ijk} p_{ij}$; the maximum-likelihood estimate of p_{ij} is $1/K \sum_{k=1}^{K} x_{ijk}$, which is precision. Thus our explicit model of the relevance judgments produces precision at rank K, and y is just the log-odds of that precision. Furthermore, we can do inference on precision using the β_j parameters just as we would in a t-test or ANOVA.

Other models give rise to other evaluation measures. Suppose we sample a rank k with probability p_k, and model x_{ijk} as

$$P(x_{ijk}) = p_{ij} p_k$$

Now p_{ij} is still a Bernoulli distribution parameter, but depending on how we define p_k, p_{ij} will be an estimate of utility. If $p_k = 1$ for $k \leq K$ and 0 for $k > K$, p_{ij} estimates precision at K. If p_k is a geometric distribution (that is, $p_k = \theta^{k-1}(1-\theta)$, then p_{ij} will be an estimate of RBP. This fits with our formulation in previous work, meaning many other measures can fit in this framework [2].

3.3 Modeling other evidence

Once the models are explicit, and computation/inference is divorced from model structure and assumptions, we can easily incorporate other sources of evidence without having to find new methods for inference. This is a limitation of traditional methods such as the t-test; the inference is strongly tied to the model structure and assumptions.

In this section we adopt a more "intuitive" notation for our models; rather than express a model as a sum of variables, we express it in words as a sum of effects. Our simple linear models above will be:

$$y_{ij} = \mu + \text{system}_j + \text{topic}_i + \epsilon_{ij}$$

This notation is meant to reduce the Greek letter assignment problem: as we add more sources of variance, we also add interactions between them, and there is a resulting combinatorial explosion in coefficients.

As an example of incorporating another source of variance, suppose we suspect that which assessor is assigned to which topics may explain something about the evaluation. We can incorporate assessor as another effect in the linear model:

$$\begin{aligned} y_{ijk} = \mu &+ \text{system}_j + \text{topic}_i + \text{assessor}_k \\ &+ \text{topic}_i \times \text{system}_j + \text{topic}_i \times \text{assessor}_k + \text{system}_j \times \text{assessor}_k + \epsilon_{ijk} \end{aligned}$$

where k is an integer identifying the assessor that worked on topic i. We also add interaction effects (denoted as $\textit{effect}_1 \times \textit{effect}_2$) to model any possible bias that an assessor might have for a system or topic. We use normal priors with log-normal priors on the variance parameters for all of these interaction effects as well as the assessor effect; these are omitted for space. The interaction between all three effects is subsumed by the errors, so it is not necessary to model it explicitly.

Note that assessor variance is not easy to model in a traditional ANOVA/t-test linear model: we have repeated measures of systems on topics (so we can do paired or within-group analysis), but we generally do not have repeated measures of assessors on topics (so we can only do unpaired or between-group analysis). Combining within-group and between-group analyses requires a whole other generalization of the linear model in classical statistics; in Bayesian statistics there is essentially no issue.

As another example, suppose we are interested in the effect of different corpus filters on results. We could incorporate that into the model easily:

$$\begin{aligned} y_{ijk} = \mu &+ \text{system}_j + \text{topic}_i + \text{corpus}_k \\ &+ \text{topic}_i \times \text{system}_j + \text{topic}_i \times \text{corpus}_k + \text{system}_j \times \text{corpus}_k + \epsilon_{ijk} \end{aligned}$$

In general, we can add any number of additional effects along with interactions between them. While this leads to a combinatorial explosion in the number of coefficients to estimate, which in turn requires an exponential amount of data in the traditional ANOVA/t-test model, the Bayesian approach does not suffer as the traditional approach would. Given little evidence for estimating a kth-order interaction effect, the Bayesian approach simply falls back to the prior and says that it cannot conclude with any confidence that the interaction effect is not a significant source of variance.

4 Empirical Analysis

Our main proposal in this work is that Bayesian models are a powerful alternative to traditional tools for evaluation and statistical analysis such as the t-test. We cannot prove that Bayesian versions are superior: empirically, there is no gold standard against which we can compare the inferences from different approaches to show that one is more accurate on average; theoretically, both are valid. We can only show when the two approaches agree and when they disagree, and argue from principles about the relative cost of the disagreements.

4.1 Experimental set-up

Our data is retrieval systems submitted to TREC tracks over the years. Since we are concerned with significance testing, it does not really matter what data we use. We picked the 74 runs submitted to the TREC-6 ad hoc track for a large set of runs on a relatively small corpus that can be broken up into more homogeneous chunks for modeling (non-random) corpus effects. TREC-6 also has a second set of *qrels* from work done at the University of Waterloo [5]; we can use this to demonstrate modeling assessor effects.

In all of our analyses below we test a one-sided hypothesis about two systems. The null hypothesis is that the system S_1 is better than S_2 by some measure.

$$H_0 : S_1 \geq S_2$$
$$H_a : S_1 < S_2$$

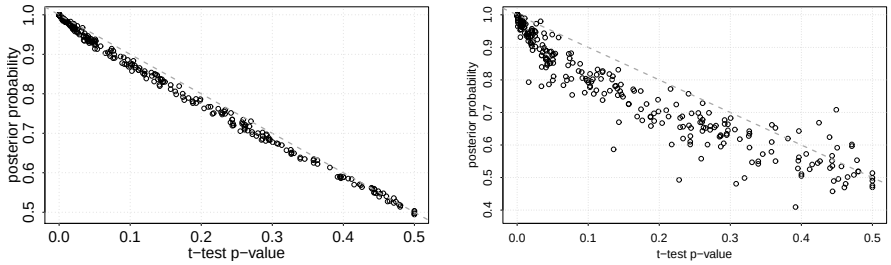
The one-sided null hypothesis is generally the hypothesis practitioners are interested in, and we argue it is more "natural" than the two-sided point null hypothesis of $S_1 = S_2$. We test these hypotheses with two measures: precision at rank 10 and rank-biased precision (RBP) with $\theta = 0.8$. We compare the p-value for rejecting H_0 to the Bayesian posterior probability of H_a being true—this means that *higher* Bayesian probabilities correspond to *lower* p-values.

We use JAGS (Just Another Gibbs Sampler, an open-source implementation of BUGS for MCMC sampling to compute posteriors in Bayesian models) and its R interface `rjags` to implement the models we describe above. JAGS allows a user to write a "model file", a programmatic description of the model which is parsed into a full simulation program. `rjags` executes the simulation, computing posterior distributions conditional on data objects in R. All of our model files and R code can be downloaded from `ir.cis.udel.edu/~carteret/testing.html`.

Because it requires simulation, Bayesian testing is much more computationally-intensive than classical testing. Rather than test all 2,701 pairs of runs, we sampled a subset of 500 pairs to test. All results are based on the same sample.

4.2 Classical t-test vs. Bayesian linear model vs. direct inference

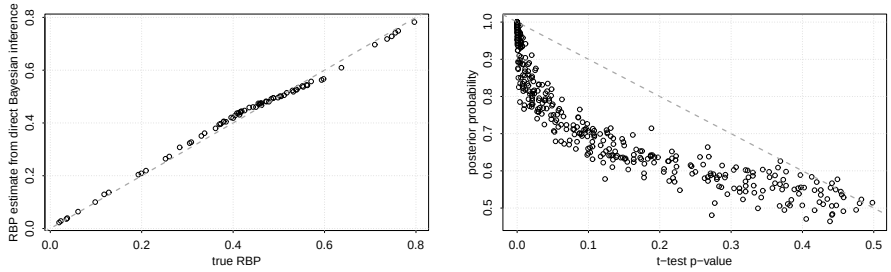
Here we show that p-values from classical t-tests correlate well with posterior probabilities from Bayesian tests.



(a) Comparison to linear model posterior probability.

(b) Comparison to relevance model posterior probability.

Fig. 1. Comparison of one-sided paired t-test p-values for a difference in precision@10 to Bayesian posterior probabilities from two different models: the traditional linear model (left) and the direct model of relevance (right).



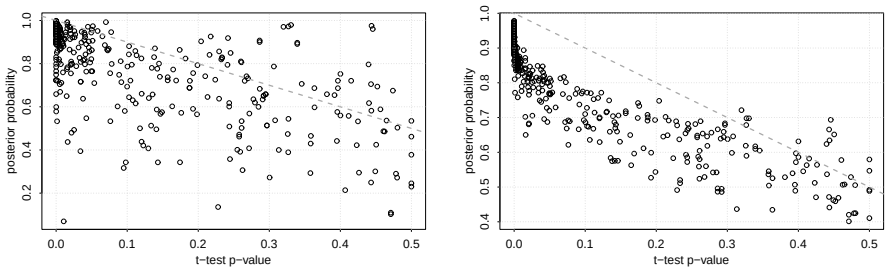
(a) True RBP (with $\theta = 0.8$) agains RBP estimate p_{ij}.

(b) t-test p-values and Bayesian posterior probabilities from direct model.

Fig. 2. Using RBP's user model as part of a direct model of relevance results in accurate estimates of RBP (left), but less confidence in conclusions about H_a.

Figure 1(a) compares p-values from a one-sided paired t-test for a difference in precision to posterior probabilities from the Bayesian linear model we presented in Section 3.1. Note that they are highly correlated, though not quite identical. The Bayesian posterior probabilities are slightly more conservative than the t-test p-values, most obviously as they approach the boundary; this is due to the use of noninformative priors. In the Bayesian framework it takes extraordinary evidence to draw the extraordinary conclusion that one system is better than the other with probability close to 1.

Figure 1(b) compares p-values from the one-sided paired t-test to posterior probabilities from the Bayesian direct inference model we presented in Section 3.2. We now see that tailoring the model to the particular data we have in IR and a particular user model has a stronger effect on inferences. The posterior probabilities are generally much more conservative than the t-test p-values.



(a) Linear model with corpus effects. (b) Linear model with assessor effects.

Fig. 3. Comparison of one-sided paired t-test p-values to Bayesian posterior probabilities from models that include additional effects.

In Section 3.2 we claimed that the parameter p_{ij} can estimate an effectiveness measure based on a user model. Figure 2(a) shows that in the direct inference approach with the RBP user model, p_{ij} indeed gives a good estimate of RBP: the estimates are almost perfectly correlated with true RBP. The posterior probabilities, however, are substantially more conservative than the p-values from a t-test (Figure 2(b)). This suggests that there may be many other reasons for documents to be ranked as they are apart from basic system and topic effects. It also suggests that by giving so much weight to the top-ranked documents, RBP makes it difficult to draw general conclusions about differences between systems.

4.3 Advanced analysis

We simulated evaluation over multiple corpora by splitting each TREC-6 submitted run into separate runs by collection: for each of the Congressional Records, Federal Register, Financial Times, Foreign Broadcast Information Service, and LA Times collections, we created 74 new runs consisting of only documents in that collection (ranked in the same order as the original run). Thus with five collections we have a total of $74 \times 5 = 370$ systems.[1] Since some systems did not retrieve any documents from some collections for some topics, we have unbalanced data—this is a case that is hard for traditional methods to deal with, but the Bayesian approach can solve painlessly.

One-sided t-test p-values and Bayesian posterior probabilities from the model in Section 3.3 are shown in Figure 3(a). Although the relationship looks random by inspection, agreement is actually quite high—the linear correlation is -0.7, meaning the posterior probability of H_a is high when the chance of rejecting H_0 is high. But there are many cases in which taking corpus into account substantially changes the inference about systems. The most extreme case is the point in the lower left; the t-test concludes that S_2 is better, while Bayesian analysis taking

[1] We note the implicit assumption that the systems ranked documents for each collection using corpus statistics computed from all collections together. This is not very realistic, but we think the example is still illustrative.

corpus effects into account concludes that S_1 is better, and both inferences have high confidence. The first system actually has much better retrieval results on each corpus individually, but managed to interleave results in such a way that its final results are much worse. This is a formative conclusion that traditional statistical analysis could not tell us.

To test whether assessors had any effect, we evaluated all systems and topics using both sets of judgments available for the TREC-6 topics. Figure 3(b) shows the relationship between t-test p-values and posterior probabilities when assessor set is part of the model. As in Fig. 1(a), we still "believe" H_a is true in most cases—meaning assessors have little effect on whether we reject or accept H_a, confirming previous work [15]—but we have significantly less confidence. This is because there are more parameters to estimate in the model, and therefore less confidence in the hypothesis with the same amount of data.

5 Conclusion

We have introduced a Bayesian framework for inferences about IR systems. The advantage of this framework is that *all* models—from the user model in the effectiveness measure to the topic population model in the significance test—are made explicit, revealing all assumptions and opening them to refinement or correction. Since computation is largely divorced from model structure and assumptions, assumptions can be changed easily without developing new methods for inference. We showed how an evaluation model can be seamlessly combined with a user model for more user-centered system-based evaluation, and how many more factors affecting effectiveness can be incorporated into the evaluation model; both of these subjects are too big for a detailed treatment here, but we intend to follow up on both in future publications.

Because models are explicit, using this framework in a variety of evaluation scenarios is mostly a matter of building the model. For low-cost evaluation settings, we can model missing judgments. For settings with graded judgments, we can use multinomial distributions instead of Bernoulli trials, or user models that probabilistically map grades to binary judgements [12]. Tasks such as novelty/diversity [4] or sessions [8] simply involve creating new models of user utility. Furthermore, the models can be directly informed by logged user data by using that data to compute posterior distributions.

The tradeoff of increased transparency and power is decreased clarity. We concede that it can be difficult to look at the models in Section 3.2 and easily understand what is being modeled and how. Furthermore, computation is much more arduous (not to mention less transparent), and inferences are subject to simulation error, much like randomization and bootstrap tests.

Nevertheless, the framework is so powerful and flexible that we believe these tradeoffs are worthwhile. The inferences are close enough that the practitioner can still use t-tests for the basic paired experiments that are common in IR. But when more advanced analysis is required, our Bayesian model-based framework seems to be the solution.

References

1. Rakesh Agrawal, Sreenivas Gollapudi, Halan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of WSDM '09*, pages 5–14.
2. Ben Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of SIGIR*, 2011. To appear.
3. Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expceted reciprocal rank for graded relevance. In *Proceedings of the Annual International ACM Conference on Knowledge and Information Management (CIKM)*, 2009.
4. 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 '08*, pages 659–666.
5. 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.
6. Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2004.
7. Kalervo Jarvelin and Jaana Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.
8. Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. Evaluation over multi-query sessions. In *Proceedings of SIGIR*, 2011. To appear.
9. Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Info. Sys.*, 27(1):1–27, 2008.
10. John F. Monahan. *A Primer on Linear Models*. Chapman and Hall/CRC, 1st edition, 2008.
11. Stephen E. Robertson. A new interpretation of average precision. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 689–690, 2008.
12. Stephen E. Robertson, Evangelos Kanoulas, and Emine Yilmaz. Extending average precision to graded relevance judgments. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610, 2010.
13. Mark Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of CIKM*, pages 623–632, 2007.
14. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 4th edition, 2002.
15. Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of SIGIR*, pages 315–323, 1998.
16. Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. Expected browsing utility for web search evaluation. In *Proceedings of the ACM International Conference on Knowledge and Information Management*, 2010. To appear.
17. 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.
18. Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of SIGIR*, pages 307–314, 1998.