Drawing Sound Conclusions from Noisy Judgments

David Goldberg
eBay
San Jose
California
dgoldberg@ebay.com

Andrew Trotman
University of Otago
Dunedin
New Zealand
andrew@cs.otago.ac.nz

Xiao Wang
eBay
San Jose
California
xwang2@ebay.com

Wei Min
CreditX
Shanghai
PRC
vera_jack312@126.com

Zongru Wan
Evolution Labs
Shanghia
PRC
zwan@evolutionlabs.com.cn

ABSTRACT

The quality of a search engine is typically evaluated using hand-labeled data sets, where the labels indicate the relevance of documents to queries. Often the number of labels needed is too large to be created by the best annotators, and so less accurate labels (e.g. from crowdsourcing) must be used. This introduces errors in the labels, and thus errors in standard precision metrics (such as P@k and DCG); the lower the quality of the judge, the more errorful the labels, consequently the more inaccurate the metric. We introduce equations and algorithms that can adjust the metrics to the values they would have had if there were no annotation errors.

This is especially important when two search engines are compared by comparing their metrics. We give examples where one engine appeared to be statistically significantly better than the other, but the effect disappeared after the metrics were corrected for annotation error. In other words the evidence supporting a statistical difference was illusory, and caused by a failure to account for annotation error.

CCS Concepts

•**Information systems → Presentation of retrieval results;**

Keywords

Precision, Statistical Significance, Standard Error

1. INTRODUCTION

New ranking algorithms are often compared in forums like TREC using manual assessments of the relevance of documents to queries. These manual assessments are treated as the gold standard ground truth, and are used to compute precision metrics which are then used to determine whether differences between algorithms are significant.

This process relies on the gold standard being accurate (and there being one unambiguous assessment of a document with respect to the query regardless of the user). If the ground truth is inaccurate then the statistical tests used to detect significance are inaccurate, and the outcome of an experiment comparing two ranking functions might produce an invalid result. We ask:

Is it possible to perform sound statistical significance tests when the assessments contain errors?

We show that it is. We assume two levels of judging, bronze and gold. Bronze judges are inexpensive but inaccurate. Gold judges are a scarce resource, but have a high level of accuracy. Bronze judge accuracy is estimated by taking a set of bronze judge assessed query-document pairs and asking the gold judge to carefully re-assess the pairs. This provides two accuracy rates for the bronze judges: a rate on relevant documents (relevant according to gold judges), and a rate on irrelevant documents.

We derive equations that directly include these accuracy rates and give the values that P@1, P@k and DCG would have if there were no bronze judge errors. We also show how to compute the standard errors of these corrected metrics and use them in turn to compute a p-value when doing a t-test to compare two ranking algorithms.

We apply our adjustments to a large-scale commercial eCommerce search engine and show that an incorrect conclusion would have been drawn if the accuracy rate were not considered—showing that our test is harder to pass than tests not incorporating accuracy rates. We further apply our techniques to the only TREC collection that mirrors our environment.

2. RELATED WORK

Our investigation centers on robust evaluation in an environment of errorful assessments—which others have already demonstrated exists. Bailey et al. [2] examined three quality levels of judge: gold, silver and bronze. They found low levels of agreement with the silver and bronze judges, and consequently small but consistent variation in system scores and rankings. Most interestingly, Bailey et al. used bronze judges for 50 topics but the gold judges for only 33 topics. A decision that may have been motivated by cost (highly

ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052570

trained judges cost more than TREC participants). This is exactly our motivation. We cannot afford to train and use gold judges for all evaluations. So we use bronze judges to assess and then and gold judges to evaluate the quality of the bronze judges..

Sanderson et al. [12] investigate the cause of judge errors and show that the relevance of a document to the judge is not independent of previously seen documents. Scholer et al. [13] attribute high levels of judgment error in TREC assessments to many factors including how far through the assessment process a given judge is (i.e. tiredness).

Many have turned to crowdsourcing in an effort to reduce the high cost of generating assessments. Our work applies to both crowdsourcers and more dedicated (but imperfect) assessors. Alonso et al. [1] suggested using either a weighted sum or a voting scheme to reduce errors. Snow et al. [15] estimate judge accuracy using examples labeled by an expert judge. They observe that with enough judges the accuracy of each judge can be inferred from the others and that a weighted average of all judges could be used as an aggregated assessment. Vuurens et al. [17] examine spam in crowdsourced assessments. Tang & Lease [16] combine expert assessments with crowdsourced assessments and show that adding small amounts of expert assessments approaches the accuracy of assessments from experts only.

Dawid & Skene [4] compute the error of a clinician taking discrete observations of a patient. When the answer is known (e.g. diagnosis of a break before an X-ray), the error is easily computed. If the gold standard is not known then they use the EM algorithm to compute the error in each cell of a confusion matrix. Passonneau & Carpenter [10] extend this algorithm by adding a Dirichlet prior. Zhao et al. [20] are the first to explore both false positive (type I) errors and false negative (type II) errors. They observe: voting is unreliable as the majority can be wrong. We also use type I and type II errors.

Comparing ranking functions in the light of noise is not new, but previous work considered only noise created by picking a different set of queries or a different set of documents. Moffat & Zobel [9] and Webber et al. [19] introduce Rank-Biased Precision (RBP) and Rank-Biased Overlap (RBO) respectively, both of which compute a precision score along with a residual. In other words, RBP and RBO give an upper bound and a lower bound on the precision score at any point in the results list. Their uncertainty is uncertainty in the remainder of the results list (assuming no judging errors).

Our contribution is to analyze the effect of judge error and the additional noise it introduces (since judge error is only imperfectly known). We show that ignoring judge error affects conclusions about actual ranking functions on actual data sets.

Noise is usually addressed through the use of a significance test such as the t-test or Wilcoxon signed rank test. Smucker et al. [14] examine the use of different significance tests for use with ranking functions and conclude that the t-test is as good as bootstrap (and randomization), and better than Wilcoxon's signed rank test. Cormack & Lynam [3] propose the use of the bootstrap for computing the standard error. We build on this work by showing how to apply the bootstrap to compute the standard error in the light of errorful assessments and how to perform Welch's unequal variances t-test on the precision score and the standard error.

There is prior work on evaluating the accuracy of assessments. Passonneau & Carpenter [10] observe that just because two judges agree, it does not make the agreement accurate. If one judge is usually wrong and the second is sometimes wrong then no claim can be made about an individual assessment without a confidence interval—so they compute confidence intervals for each assessment. Joglekar et al. [7] compute both the error rate of crowdsource workers, and also confidence intervals on their work. From this it's possible to estimate the accuracy of any answer set, and to eliminate bad workers. None of this previous work showed how assessment error affects decisions about the relative quality of two ranking functions.

We give an explicit equation for the effect of assessment error on precision along with the standard error of precision. Together they give a range of possible precision scores, and enable what we believe is a better way of comparing two ranking functions.

3. BINARY RELEVANCE ASSESSMENTS

3.1 Precision at 1

This section examines the scenario in which the search engine is given a set of n queries and produces a single result for each query. This result is either relevant or non-relevant. The scenario is not atypical, it is precision at 1 document (P@1)—one of the many measures used to indicate the quality of a commercial search engine.

More formally, the probability that this search engine will produce a relevant result is estimated by the average $\bar{j}$,

$$\bar{j} = \sum_{s=1}^{n} \frac{j_s}{n} \tag{1}$$

where n is the number of queries used in the estimate, and j_s is the relevance (1 for relevant, 0 for non-relevant) of the s-th returned document. This is an estimate because n is finite and j_s could differ with a different set of n queries.

The confidence in $\bar{j}$ is given by the standard error σ_j, estimated as $\widehat{\sigma}_j$, using

$$\widehat{\sigma}_j = \sqrt{\frac{\bar{j}(1-\bar{j})}{n}} \tag{2}$$

We use j to emphasize that these are relevance estimates of a *judge* (the bronze judge). Equation (1) and Equation (2) assume that the judge's assessments are accurate—which we know they are not.

We assume that there is a well-defined relevance for each document, and that there is a gold judge who can reliably determine that relevance value. Assume that across the entire assessment set (1 assessment per query) the bronze judge gives an assessment that matches that of the gold judge with a probability m_J (we use m as it is a mean, see section 3.2).

Prior studies [11] have shown that the time to assess a document is different for relevant and non-relevant documents, so we make no assumption about the probability of a match on documents considered relevant by the gold judge, m_{J_R}, being equal to the probability of a match on documents considered not-relevant m_{J_N}.

			Bronze	
			Relevant	Non-Relevant
			m_J	$1 - m_J$
Gold	Relevant	m_G	m_{J_R}	$1 - m_{J_R}$
	Non-Relevant	$1 - m_G$	$1 - m_{J_N}$	m_{J_N}

Table 1: Confusion matrix showing possible outcomes

Table 1 presents the 2×2 confusion matrix in grey. The first row is for the case when the document is considered relevant by the gold judge, which occurs with probability m_G. The second row is when the document is not considered relevant by the gold judge, which occurs with probability $1 - m_G$. Similarly the first column is when the bronze judge assesses as relevant, and the second column is for non-relevant. The values in the 2×2 matrix are *conditional* probabilities; the probability of the bronze judge's decision given the gold judges's decision. For example the upper left box, m_{J_R}, is the probability that the bronze judge assesses a document relevant, given that the gold judge says it is relevant, and the lower right box, m_{J_N}, is the probability the bronze judge assesses not relevant when the gold judge says it is not relevant.

We are interested in computing m_G, the probability that the one result is considered relevant by the gold judge. Unfortunately, m_G cannot be observed directly. But m_J, the *unconditional* probability that the bronze judge rates the document as relevant can be observed. The value of m_J is given by

$$m_J = m_G m_{J_R} + (1 - m_G)(1 - m_{J_N}) \tag{3}$$

In words, m_J is the probability that the one returned document is considered relevant by the gold judge and the bronze judge agrees, $m_G m_{J_R}$, plus the probability the document is not relevant according to the gold judge but the bronze judge, none-the-less, assesses it as relevant, $(1 - m_G)(1 - m_{J_N})$.

Equation (3) can easily be solved for m_G, the probability that the gold judge considers the document as relevant, giving

$$m_G = \frac{m_J - 1 + m_{J_N}}{m_{J_R} + m_{J_N} - 1} \tag{4}$$

The variables m_{J_R} and m_{J_N} are the hidden accuracy rates of the bronze judge but can be estimated from assessments as $\widehat{m}_{J_R}$ and $\widehat{m}_{J_N}$.

An estimate of m_J, $\widehat{m}_J = \overline{j}$, is given by Equation (1). From these an estimate of m_G, $\widehat{m}_G$, can be computed,

$$\widehat{m}_G = \frac{\overline{j} - 1 + \widehat{m}_{J_N}}{\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1} \tag{5}$$

Equation (1) should be compared to Equation (5) as they are both estimates of P@1. In the former the bronze judge is assumed to be faultless. The latter uses the former along with estimates of the judge's accuracy rates to give a more accurate score. Equation (5) *is* consistent with Equation (1): if $\widehat{m}_{J_R} = \widehat{m}_{J_N} = 1$ then $\widehat{m}_G = \overline{j}$. In other words, if the bronze judge is faultless then both equations give the same result. At the other extreme, if $\widehat{m}_{J_R} = \widehat{m}_{J_N} = 0.5$ then the bronze judge is performing an action equivalent to a coin toss and we can derive no useful information because the denominator of Equation (5) is 0. $\widehat{m}_G$ and $\overline{j}$ are very different. $\overline{j}$ is easy to compute but inaccurate, whereas $\widehat{m}_G$ is accurate but harder to compute.

3.2 Standard Error

In this section we compute the variance of $\widehat{m}_G$, (our estimate of P@1). The square root of variance is the standard error, and statistical significance can be computed from $\widehat{m}_G$ and its standard error. Two different methods for computing the variance are presented: the first is the parametric bootstrap; the second is an explicit equation.

To get the variance of $\widehat{m}_G$ we need to write it as a random variable, which we do by introducing random variables to represent the parts of Equation (5): $\overline{j}$, $\widehat{m}_{J_N}$, and $\widehat{m}_{J_R}$. If J is the random variable which is 1 when a bronze judge assesses a document relevant and 0 otherwise, then $\overline{J} = (J_1 + \cdots + J_n)/n$ is the random variable representing $\overline{j}$. Similarly if J_R is 1 when the bronze judge assesses as relevant a document considered relevant by the gold judge, then $\overline{J}_R = ((J_R)_1 + \cdots + (J_R)_{n_R})/n_R$ represents $\widehat{m}_{J_R}$, similarly for $\widehat{m}_{J_N}$.

The random variables J, J_R and J_N are linked to the variables in the previous section via $\mathbb{E}[J] = \mathbb{E}[\overline{J}] = m_J$, $\mathbb{E}[J_R] = \mathbb{E}[\overline{J}_R] = m_{J_R}$ and $\mathbb{E}[J_N] = \mathbb{E}[\overline{J}_N] = m_{J_N}$, where $\mathbb{E}[J]$ is the expected value of J. This explains the notation m (for mean) introduced in the previous section.

The distribution of the random variable $\overline{J}$ is

$$\overline{J} \approx \frac{\text{Binom}(\overline{j}, n)}{n}$$

To get the distributions of $\overline{J}_R$ and $\overline{J}_N$, select n_R assessments considered relevant by the gold judge and n_N considered non-relevant by the gold judge[1]. If the bronze judge rates r of the n_R items relevant, then $\widehat{m}_{J_R} = \frac{r}{n_R}$ is an estimate of J_R, and the distribution of J_R is

$$\overline{J}_R \approx \frac{\text{Binom}(\widehat{m}_{J_R}, n_R)}{n_R}$$

similarly,

$$\overline{J}_N \approx \frac{\text{Binom}(\widehat{m}_{J_N}, n_N)}{n_N}$$

In all three cases $\text{Binom}(\theta, n)$ is the binomial distribution, the number of heads in n coin tosses, where a coin has probability of θ of coming up heads.

Algorithm 1 Bootstrap Computation of Standard Error

```
function OneBootstrap($\overline{j}, n$)
    $j^* \leftarrow Binom(\overline{j}, n)/n$
    $j^*_R \leftarrow Binom(\widehat{m}_{J_R}, n_R)/n_R$
    $j^*_N \leftarrow Binom(\widehat{m}_{J_N}, n_N)/n_N$
    $\widehat{m}^*_G \leftarrow (j^* - 1 + j^*_N)/(j^*_R + j^*_N - 1)$
    return $\widehat{m}^*_G$
for $i \leftarrow 1$ to $n.iter$ do
    $samples_i \leftarrow OneBootstrap(\overline{j}, n)$
$var \leftarrow \text{Variance}(samples)$
return $\sqrt{var}$
```

To perform the parametric bootstrap, draw many samples of $\overline{J}$, $\overline{J}_R$ and $\overline{J}_N$ from the binomial distributions and for each sample use Equation (5) to compute an instance of $\widehat{m}_G$, $\widehat{m}^*_G$. The standard deviation of these $\widehat{m}^*_G$ is an estimate of the standard error of $\widehat{m}_G$. The square of the standard

[1] This could be a separate set from the n items used to compute $\overline{j}$. That is, it could be a prior process, or even the validation phase in a crowdsourcing experiment.

deviation is the variance, $\mathbb{V}[\widehat{m}_G]$. The algorithm is presented in Algorithm 1.

As well as there being a straightforward algorithm for bootstrapping the variance, there is an explicit equation that adds insight. The derivation of the variance, $\mathbb{V}[\widehat{m}_G]$, the square of the error, is from the standard equation for the variance of a quotient,

$$\mathbb{V}\left[\frac{A}{B}\right] \approx \frac{\mathbb{V}[A]}{\mathbb{E}[B]^2} + \mathbb{V}[B]\,\frac{\mathbb{E}[A]^2}{\mathbb{E}[B]^4} - 2\frac{\mathbb{E}[A]}{\mathbb{E}[B]^3}\mathbb{V}[A,B]$$

where $\mathbb{V}[A,B]$ is the covariance of A and B. To compute $\mathbb{V}[\widehat{m}_G]$ set (from Equation (4))

$$A = \overline{J} - 1 + \overline{J}_N$$

and

$$B = \overline{J}_R + \overline{J}_N - 1$$

We explain below why it is reasonable to assume that $\overline{J}$, $\overline{J}_R$, and $\overline{J}_N$ are uncorrelated (e. g. $\mathbb{V}\left[\overline{J}, \overline{J}_R\right] = 0$).

Now,

$$\mathbb{V}[A,B] = \mathbb{V}\left[\overline{J}_N\right]$$

because

$$\begin{aligned}
\mathbb{V}[A,B] &= \mathbb{V}\left[\overline{J} - 1 + \overline{J}_N, \overline{J}_R + \overline{J}_N - 1\right] \\
&= \mathbb{V}\left[\overline{J}, \overline{J}_N\right] + \mathbb{V}\left[\overline{J}, \overline{J}_N\right] + \mathbb{V}\left[\overline{J}_N, \overline{J}_R\right] + \mathbb{V}\left[\overline{J}_N, \overline{J}_N\right] \\
&= 0 + 0 + 0 + \mathbb{V}\left[\overline{J}_N, \overline{J}_N\right] \\
&= \mathbb{V}\left[\overline{J}_N\right]
\end{aligned}$$

so

$$\begin{aligned}
\widehat{\sigma}_G^2 \;=\; \mathbb{V}[\widehat{m}_G] \;\approx\; & \frac{\mathbb{V}\left[\overline{J}\right] + \mathbb{V}\left[\overline{J}_N\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2} \\
&+ \frac{(\mathbb{V}\left[\overline{J}_R\right] + \mathbb{V}\left[\overline{J}_N\right])(\bar{j} - 1 + \widehat{m}_{J_N})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} \\
&- 2\frac{(\bar{j} - 1 + \widehat{m}_{J_N})\mathbb{V}\left[\overline{J}_N\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^3} \qquad (6)
\end{aligned}$$

This can be rewritten in a more instructive form via some algebraic manipulations (see the Appendix) as:

$$\begin{aligned}
\widehat{\sigma}_G^2 \approx \frac{\mathbb{V}\left[\overline{J}\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2} + \mathbb{V}\left[\overline{J}_R\right]\frac{(\bar{j} - 1 + \widehat{m}_{J_N})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} \\
+ \mathbb{V}\left[\overline{J}_N\right]\frac{(\bar{j} - \widehat{m}_{J_R})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} \qquad (7)
\end{aligned}$$

Equation (7) shows explicitly how the bronze judge accuracy affects the standard error of the estimate of precision at 1, $\widehat{m}_G$. If the assessments are random then $\widehat{m}_{J_R} \approx \widehat{m}_{J_N} \approx 0.5$ and so $\widehat{\sigma}_G \approx \infty$, as expected.

Prior models assume there is no judge error and consequently $\widehat{\sigma}_G = \mathbb{V}\left[\overline{J}\right]$. The rewritten form (Equation (7)) shows that the standard error is in fact larger. First,$\mathbb{V}\left[\overline{J}\right]$ is increased by a factor of $(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^{-2}$; and second, uncertainty in $\widehat{m}_{J_R}$, and $\widehat{m}_{J_N}$ also play a role.

In a working environment its necessary to compute $\mathbb{V}\left[\overline{J}\right]$, $\mathbb{V}\left[\overline{J}_R\right]$ and $\mathbb{V}\left[\overline{J}_N\right]$. Using variance of $\text{Binom}(\theta, n)/n = \theta(1-\theta)/n$ the variances can be estimated as

$$\begin{aligned}
\mathbb{V}\left[\overline{J}\right] &\approx \frac{\bar{j}(1-\bar{j})}{n} \\
\mathbb{V}\left[\overline{J}_R\right] &\approx \frac{\widehat{m}_{J_R}(1 - \widehat{m}_{J_R})}{n_R} \\
\mathbb{V}\left[\overline{J}_N\right] &\approx \frac{\widehat{m}_{J_N}(1 - \widehat{m}_{J_N})}{n_N} \qquad (8)
\end{aligned}$$

The derivation of Equation (7) assumes that J, J_R and J_N are uncorrelated. They are actually independent, which is even stronger than being uncorrelated. They are independent because knowing a specific value of (say) J_R gives no additional information about the other variables J, and J_N. That's because J is modeled as the bronze judge's assessment of a random document. Knowing that assessment gives no additional information about what happens when a document deemed relevant by the gold judge is randomly selected and rated by a bronze judge (resulting in J_R).

In this section we derived an equation for the standard error of the estimate $\widehat{m}_G$. Unlike others, we work with the error of $\widehat{m}_G$ rather than trying to determine the error in each individual assessment. This choice was made because inferences about the quality of a ranking algorithm require an estimate of $\widehat{m}_G$, not the error in each assement.

3.3 P-Value

In an information retrieval ranking experiment, the experimenter is typically trying to determine whether ranking function a outperforms ranking function b. The p-value approach is to assume the ranking functions return equally relevant documents, and compute the probability that the observed difference in P@1 under this hypothesis is as large as it is. This section outlines how to compute the p-value using Welch's unequal variances t-test (a variant of Student's t-test).

Equation (5) is the equation for $\widehat{m}_G$, the corrected P@1 score. Equation (7) is the equation for the standard error, $\widehat{\sigma}_G$. Given two ranking functions, a and b, there are two precisions $\widehat{m}_{G,a}$ and $\widehat{m}_{G,b}$, and two standard errors, $\widehat{\sigma}_{G,a}$ and $\widehat{\sigma}_{G,b}$.

First presume that the algorithms are not different (the null hypothesis). Then compute the probability that $\widehat{m}_{G,a}$ and $\widehat{m}_{G,b}$ are different.

If the bronze judge always agrees with the gold judge (error-free assessments) this is estimated via Welch's t-test,

$$t = \frac{\bar{j}_a - \bar{j}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} \qquad (9)$$

where for ranking function a, $\bar{j}_a$ is the P@1 score from Equation (1), s_a^2 is the variance in the n_a assessments. Likewise, $\bar{j}_b$, s_b^2 , and n_b for function b.

To compute the p-value for Welch's test, the number of degrees of freedom, ν, is needed. That is given by

$$\nu = \frac{(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b})^2}{\frac{(\frac{s_a^2}{n_a})^2}{n_a - 1} + \frac{(\frac{s_b^2}{n_b})^2}{n_b - 1}}$$

The Welch two-tailed p-value is

$$p\text{-value} = 2\big(1 - \Pr(T_\nu < |t|)\big)$$

where T_ν is a t-distribution with ν degrees of freedom, and p-value is the probability that the difference in the means is as large as observed. The smaller the p-value the more likely we will consider algorithm a to have different relevance from algorithm b.

All the necessary parameters to compute the p-value in the light of uncertain assessments are derived in Section 3.1 and Section 3.2. The denominator of Equation (9) is the square root of the squares of the standard errors—which are given in Equation (7). The numerator is the $\widehat{m}_G$ score from Equation (5); so

$$t = \frac{\widehat{m}_{G,a} - \widehat{m}_{G,b}}{\sqrt{\widehat{\sigma}^2_{G,a} + \widehat{\sigma}^2_{G,b}}} \tag{10}$$

In Equation (9) t grows like $\sqrt{n}$, meaning t is more likely to be significant as n gets larger. The same is true of Equation (10), because $\widehat{\sigma}^2_{G,a}$ and $\widehat{\sigma}^2_{G,b}$ are both $O(\mathbb{V}\left[\overline{J}\right])$ which is $O(1/n)$, so that t is $O(\sqrt{n})$ just as in Equation (9).

In a typical search engine scenario ν is very large and so we use a normal distribution, Z, as an approximation of the t-distribution, T_ν:

$$p = 2\left(1 - \Pr(Z < |t|)\right)$$

giving

$$p = 2\left(1 - \Pr(Z < \frac{|\widehat{m}_{G,a} - \widehat{m}_{G,b}|}{\sqrt{\widehat{\sigma}^2_{G,a} + \widehat{\sigma}^2_{G,b}}})\right) \tag{11}$$

3.4 Precision at k

Our discussion has focused on a single document-query pair which has been averaged over n queries resulting in P@1. An equation to compute standard error and the p-value has been given. This section focuses on a results list—specifically precision at k, P@k.

The P@k score according to the bronze judge, $j@k$, for a single query is the number of relevant documents found in the top k results in the results list divided by k.

$$j@k = \frac{\sum_{t=1}^{k} j^{(t)}}{k}$$

where $j^{(t)}$ is the relevance (1 for relevant, 0 for non-relevant) of the document at position t in the results list.

This score is then averaged over the set of n queries,

$$\overline{j}@k = \frac{\sum_{s=1}^{n} \frac{\sum_{t=1}^{k} j_s^{(t)}}{k}}{n} = \frac{\sum_{t=1}^{k} \frac{\sum_{s=1}^{n} j_s^{(t)}}{n}}{k} \tag{12}$$

where $j_s^{(t)}$ is the relevance of the document at position t of the results list of query s.

Up to this point, J has represented the first result in the results list and the bronze judge has assigned a value of $j = 1$ for relevant and $j = 0$ for non-relevant. For P@k there are k assessments for a results list of length k, represented $J^{(1)}, J^{(2)}, \ldots J^{(k)}$.

Precision at k for a single query can now be defined in terms of those k random variables as

$$\overline{J}@k = \frac{1}{k}\left(J^{(1)} + J^{(2)} + \cdots + J^{(k)}\right) \tag{13}$$

From the expected precision value at each given position in the results list, $\mathbb{E}[J^{(t)}]$, we can compute the expected value of P@k as

$$m_J@k = \mathbb{E}[\overline{J}@k] = \frac{1}{k}\left(\mathbb{E}[J^{(1)}] + \mathbb{E}[J^{(2)}] + \cdots + \mathbb{E}[J^{(k)}]\right)$$

This $\mathbb{E}[\overline{J}@k]$ is computed naively from the bronze judge's assessments. The values $\mathbb{E}[J^{(t)}] = m_J^{(t)}$ on the right hand side can be written using Equation (3) in terms of m_G, m_{J_R} and m_{J_N}. Substituting for $\mathbb{E}[J^{(t)}]$ gives

$$\begin{aligned} m_J@k = \frac{1}{k}\Big(& \big(m_G^{(1)}(m_{J_R} + m_{J_N} - 1) + (1 - m_{J_N})\big) + \cdots \\ & + \big(m_G^{(k)}(q_r + m_{J_N} - 1) + (1 - m_{J_N})\big)\Big) \end{aligned}$$

Note that $m_G^{(t)}$ is position dependent but m_{J_R} and m_{J_N} are position independent. That is, the precision of the search engine at each position in the results list is different, but the accuracy of the judge is position independent.

The previous equation can be rewritten as

$$m_J@k = \frac{m_G^{(1)} + \cdots + m_G^{(k)}}{k}(m_{J_R} + m_{J_N} - 1) + (1 - m_{J_N})$$

and pulling all the m_G's to one side of the equation yields

$$m_G@k = \frac{m_G^{(1)} + \cdots + m_G^{(k)}}{k} = \frac{m_J@k - 1 + m_{J_N}}{m_{J_R} + m_{J_N} - 1} \tag{14}$$

The left hand side of Equation (14) is precision at k using the gold judge's judgement of relevance while the right hand side uses the observed P@k using the bronze judge's relevance ratings, $m_J@k$, together with m_{J_N} and m_{J_R}.

3.5 Standard Error and P-Value

The similarity between Equation (14) and Equation (4) is striking, and means that the standard error of $m_G@k$ has the same form as $\widehat{\sigma}^2_G$ in Equation (7). Specifically, replace $\overline{j}$ in Equation (7) with $\overline{j}@k$ from Equation (13). And replace $\mathbb{V}\left[\overline{J}\right]$ with the variance of $\overline{J}@k = (\overline{J}^{(1)} + \cdots \overline{J}^{(k)})/k$. However this variance can no longer be computed using the first equation of (8), because $J@k$ is not binomial. Instead the variance must be computed directly as a standard deviation of the observed P@k values of the different queries.

4. GRADED RELEVANCE ASSESSMENTS

Section 3 discusses the scenario where the judges are asked to mark relevance on a binary scale. In this section we extend from binary assessments to multi-level (i.e. graded) relevance assessments. This extension can then in turn be used to further extend our methods to DCG.

4.1 Multi-Class Assessments

For graded assessment the bronze judge is asked to assign one of k possible labels, or grade names $(A_1, \ldots, A_k)$, to each document-query pair. Each label, A_i, has an (unknown) probability, $m_{G_{A_i}}$ (succinctly, m_{A_i}), that the gold judge assigns that label to the document. Just as m_G was unknown in Table 1, each m_{A_i} is unknown – however the sum of the probabilities m_{A_i} is known to be 1 (every document must be assigned only one label).

Rather than having two accuracy rates, m_{J_R} and m_{J_N}, we now have a matrix of accuracy rates which we denote with a script J, $\mathcal{J}$, where $\mathcal{J}_{ji}$ is the probability that the bronze judge said i when the gold judge said j (the binary case is Table 1).

If $\vec{J}$ is a random vector whose ith component is 1 when the bronze judge assigns the label A_i, then the expected probability of a correct label is

$$\mathbb{E}[\vec{J}_i] = \sum_{j=1}^{k} m_{A_j}\mathcal{J}_{ji} = (m_A^T\mathcal{J})_i \tag{15}$$

where $(\vec{m}_A^T\mathcal{J})_i$ is the ith element of the product of the vector[2] $\vec{m}_A^T = (m_{A_1},\ldots,m_{A_k})$ with the matrix $\mathcal{J}$; while $\vec{J}_i$ is the ith element of the vector $\vec{J}$. In the 2-dimensional case with the label A_1 meaning relevant, $\mathbb{E}[\vec{J}_1] = m_J$ where m_J was defined in Equation (3).

Equation (15) can be rewritten as

$$\mathbb{E}[\vec{J}]^T = m_A^T\mathcal{J} \tag{16}$$

so

$$m_A^T = \mathbb{E}[\vec{J}]^T\mathcal{J}^{-1} \tag{17}$$

As with the binary relevance case, the values of $\mathcal{J}$ are not known and must be estimated. The two dimensional case of Equation (17) reduces to Equation (4) as follows:

$$\begin{aligned}
\mathcal{J} &= \begin{pmatrix} m_{J_R} & 1-m_{J_R} \\ 1-m_{J_N} & m_{J_N} \end{pmatrix} \\
\vec{m}_A &= \begin{pmatrix} m_G \\ 1-m_G \end{pmatrix} \\
\mathbb{E}[\vec{J}]^T &= \begin{pmatrix} \mathbb{E}[J] & 1-\mathbb{E}[J] \end{pmatrix}
\end{aligned}$$

so

$$\begin{aligned}
\mathcal{J}^{-1} &= \frac{1}{m_{J_R}+m_{J_N}-1}\begin{pmatrix} m_{J_N} & m_{J_R}-1 \\ m_{J_N}-1 & m_{J_R} \end{pmatrix} \\
\vec{m}_A^T &= \mathbb{E}[\vec{J}]^T\mathcal{J}^{-1} = \frac{1}{m_{J_R}+m_{J_N}-1}\begin{pmatrix} \mathbb{E}[J]-1+m_{J_N} \\ -\mathbb{E}[J]+m_{J_R} \end{pmatrix}^T
\end{aligned}$$

Note that matrix $\mathcal{J}$ is Table 1, that $\vec{m}_A$ are the row labels in Table 1, and that $(\vec{m}_A)_1$ is Equation (4).

4.2 DCG

The graded assessment analogue of precision at k is Cumulated Gain, CG [6]. Gain is the quantity of relevant material seen by the user, and its exact definition depends on evaluation context. For example, in XCG [8] gain is defined in terms of a quantized score computed for XML document elements assessed on a two dimensional relevance scale—we leave for further work the extension to sub-document relevance and concentrate herein on graded relevance on a whole document scale measured as the user reads down a results list.

The cumulated gain of a user after they have read k documents (in order) from the results of a single query is

$$\text{CG} = \sum_{s=1}^{k} j^{(s)}$$

where $j^{(s)}$ is the relevance score (given by a judge) of a document at position s (often in the range [0..1], where 0 is not relevant and 1 is most highly relevant).

Eye-tracking experiments have shown that users typically examine a results list top to bottom [5] so it is more important to put a highly relevant document at position 1 in the results list than at position 2, than position 3, and so on. Metrics that model such a user model this decrease in utility by increasingly penalizing a document's contribution to the metric the further down the results list it appears.

In Discounted Cumulated Gain, DCG, this penalization is done by dividing the score of the document, $j^{(s)}$, by the log of its position in the results list [18]. There are many subtly different interpretations of this (using different bases to the log, and so on). The interpretation we use is

$$\text{DCG} = \sum_{s=1}^{k} \frac{j^{(s)}}{\log_2(s+1)}$$

for the results of a single query.

In the previous section judges assigned labels $A_1,\ldots A_k$ to queries. For DCG, the labels must be mapped to numerical values, so we introduce the value vector $\vec{v}$ where the label A_j is assigned the numerical value v_j. If $\vec{G}$ is a random vector with a 1 in the t-th coordinate when the gold judge assigns the label A_t, then the dot product $\vec{G}\cdot\vec{v}$ is the random variable representing the gold judge's value of the document. The average judgement $\mathbb{E}[G]$ is likely to vary with the rank of the documents, so we introduce $G^{(s)}$ for gold judge's judgment of documents at the s-th rank, and their value is $\vec{G}^{(s)}\cdot\vec{v}$.

With this notation, the expected DCG according to the gold judge is

$$\mathbb{E}[\text{DCG}_G] = \sum_{s=1}^{k}\frac{\mathbb{E}[\vec{G}^{(s)}]\cdot v}{\log_2(s+1)} = \sum_{s=1}^{k}\frac{\vec{m}_A^{(s)}\cdot v}{\log_2(s+1)}$$

Repeating Equation (17)

$$m_A^T = \mathbb{E}[\vec{J}]^T\mathcal{J}^{-1}$$

and substituting gives

$$\mathbb{E}[\text{DCG}_G] = \sum_{s=1}^{k}\frac{(\vec{m}_A^{(s)})^T v}{\log_2(s+1)} = \left(\sum_{s=1}^{k}\frac{\mathbb{E}[\vec{J}^{(s)}]^T}{\log_2(s+1)}\right)\mathcal{J}^{-1}v \tag{18}$$

which is an estimate of DCG according to the gold judge. Note that $\mathbb{E}[\text{DCG}_G]$ is not computed from $\mathbb{E}[\text{DCG}_J]$. Instead, it is computed based on how the judges classified documents, $\mathbb{E}[\vec{J}^{(s)}]$, rather than how they were scored, $\mathbb{E}[\vec{J}^{(s)}]\cdot v$.

As an example, suppose we give highly-relevant documents a score of 1, not-relevant a score of 0 and partly-relevant a score of 0.5. Then $A =$ (highly-relevant, partly-relevant, not-relevant), and $v^T = (1.0, 0.5, 0.0)$. Suppose the bronze judge gives the ratings shown in Figure 1

Position	Query 1	Query 2
1	*Not-relevant* (0)	*Partly-relevant* (0.5)
2	*Highly-relevant* (1)	*Not-relevant* (0)

Figure 1: Two results lists (relevance grades in braces)

We estimate $\mathbb{E}[\vec{J}^{(1)}]^T$ as the vector $(0, 0.5, 0.5)$ since at rank 1 we never got a rating of highly-relevant, and half the time we got partly-relevant and half not-relevant. Similarly $\mathbb{E}[\vec{J}^{(2)}]^T \approx (0.5, 0, 0.5)$. Then,

$$\begin{aligned}
\mathbb{E}[\text{DCG}_G] &= \left(\frac{\mathbb{E}[\vec{J}^{(1)}]^T}{\log_2(1+1)} + \frac{\mathbb{E}[\vec{J}^{(2)}]^T}{\log_2(2+1)}\right)\mathcal{J}^{-1}v \\
&\approx \left(\frac{1}{\log_2 2}(0, 0.5, 0.5) + \frac{1}{\log_2 3}(0.5, 0, 0.5)\right)\mathcal{J}^{-1}\begin{pmatrix} 1.0 \\ 0.5 \\ 0.0 \end{pmatrix}
\end{aligned}$$

[2] We use column vectors, so $\vec{m}_A$ is a $k \times 1$ (column) vector and the transpose $\vec{m}_A^T$ is a row vector.

The standard error in Equation (18) can be computed using the ordinary non-parametric bootstrap (similarly to Cormack & Lynam [3]). Simply select (query, document) pairs without replacement and use Equation (18) on these bootstrap samples.

5. SIMULATION

This section uses simulation on a specific example to illustrate that the traditional and corrected equations for P@k and its standard error can give dramatically different results.

Imagine a search engine with a true but hidden P@k of 0.4. This is hidden because it is not practical to assess all possible queries. Instead we need to rely on imperfect bronze judges who only assess a subset of queries.

Further imagine that the true but hidden judge accuracy rates are $m_{J_R} = 0.9$ and $m_{J_N} = 0.8$. These are hidden because it is not possible to compute the agreement level over all possible assessments, only a sample of the bronze judge's assessments is given to the gold judge for evaluation.

Take a ranking experiment involving $n = 50$ queries where the search engine returns the top $k = 10$ results per query. In order to estimate $m_{J_R} = 0.9$ and $m_{J_N} = 0.8$, the gold judge assesses documents until they have assessed $n_R = 250$ of the documents as relevant and $n_N = 250$ of the documents as non-relevant. These documents are then given to bronze judges in order to compute the estimates $\widehat{m}_{J_R}$ and $\widehat{m}_{J_N}$.

Algorithm 2 Simulation

```
1:  function ONE_SIMULATION($m_G, m_{J_R}, m_{J_N} n_R, n_J, n, k$)
2:      for $s$ in $1{:}k$ do                    ▷ $g[s,] = 1$ with prob $m_G[s]$
3:          $g[s,] \leftarrow Bernoulli(m_G[s], n)$
4:      $j \leftarrow g$                          ▷ what bronze report
5:      for $(s, t)$ in $(1{:}k, 1{:}n)$ do       ▷ bronze errors
6:          if $g[s,t] = 1$ then
7:              $j[s,t] \leftarrow 0$ with prob $1 - m_{J_R}$
8:          else
9:              $j[s,t] \leftarrow 1$ with prob $1 - m_{J_N}$
10:     for $t$ in $1{:}n$ do                     ▷ judge's P@k for each query
11:         $\bar{j}[t] \leftarrow average\ j[,t]$
12:     $j^* \leftarrow$ ave $\bar{j}$            ▷ average $\bar{j}[t]$ over all $t$
13:     $\widehat{m}_{J_R} \leftarrow Binom(m_{J_R}, n_R)/n_R$
14:     $\widehat{m}_{J_N} \leftarrow Binom(m_{J_N}, n_N)/n_N$
15:     $g^* \leftarrow (j^* - 1 + \widehat{m}_{J_N})/(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)$
16:     return $j^*, g^*$
```

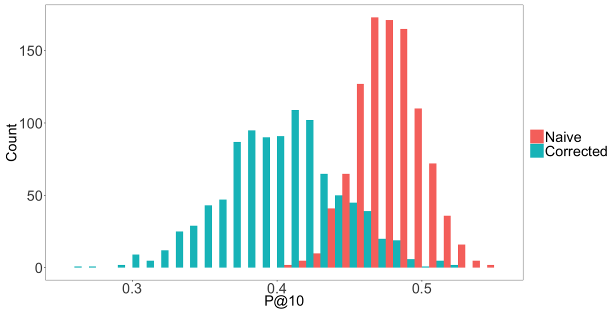

Figure 2: Histogram of simulated P@10 scores and corrected P@k scores. The naive P@K scores computed by averaging over 50 queries (red) are almost always much higher than the true value of 0.4. Algorithm called as ONE_SIMULATION([0.49, 0.47, …0.31]), 0.9, 0.8, 250, 250, 50, 10)

Knowing the true (hidden) values and the number of samples makes it possible to simulate scores for both P@k and the corrected P@k. The algorithm is presented as Algorithm 2. The first input parameter m_G is a vector with $m_G[s]$ the average precision at rank s. Line 3 sets up an array g with $g[s,t] = 1$ when the document of rank s for the t-th query is relevant. Lines 4–9 compute the relevance that the bronze judge will report, taking into account the judge accuracy probabilities m_{J_R} and m_{J_N}. Lines 10–12 compute the bronze judge's estimate of P@k. Lines 13–14 compute the estimates of m_{J_R} and m_{J_N} that will be obtained by asking the bronze judges to assess n_R and n_N documents respectively. Line 15 is the corrected value that would be computed using Equation (14).

This simulation was run 10,000 times and the P@k scores are displayed as a histogram in Figure 2. The red histogram is P@k computed using the traditional (naive) method. The blue histogram is P@k computed according to Equation (14). Even though bronze judge accuracy is high, the naive computation of P@k consistently returns values higher than the true value of 0.4.

Simulation also allows us to compare the naive computation of the standard error with Equation (7). On each simulation we test if the 95% confidence interval about the estimate for P@k contains the true value of 0.4. Using the naive standard error (the standard deviation of the 10 numbers $\bar{j}[t]$ divided by $\sqrt{10}$), the confidence interval includes the true value only 5% of the time. So the confidence interval is much smaller than it should be. Using Equation (7) the interval contained the true value 95% of the time as it should.

In the simulation, the true value of P@k is known to the simulation but not to the experimenter who is evaluating the search engine. Each run of the simulation returns the P@k that would be observed by the experimenter in an actual experiment. In every run the naive observed P@k (red) is larger than the true value, often substantially so.

6. USE IN A LIVE ENVIRONMENT

This work is in use at eBay with its large-scale eCommerce search engine. Unlike a web archive, an eCommerce environment such as ours is highly dynamic: due to inventory changes the document collection is not static; due to changing user needs the query set is not static; due to changes in the code-base the ranking function is not static. We wish to know whether, in this environment, there has been a statistically significant change in precision due to ranking function changes.

We randomly selected queries from the query log weighted by frequency (after removing nonsensical and informational queries). For each query we selected the top 3 results and sent those query-document pairs for assessment by a professional third-party. In turn that third-party used full-time judges paid by the hour (not by the assessment). Three judges examined each pair and made a binary decision (relevant or not). We use the majority vote as the final assessment of the query-document pair according to the bronze judge. In total more than 10 judges were used.

Ten days later the process was repeated (reselection of queries and top-3 results, but same third party who might have used different judges), resulting in two data sets, a, and b. The characteristics of which are:

$$n_a = 10278 \quad \bar{j}_a = 0.6260 \quad s_a = 0.414$$
$$n_b = 20604 \quad \bar{j}_b = 0.6385 \quad s_b = 0.402$$

$\bar{j}$ is the average P@3 as measured by the judges[3] and s_a, s_b are the standard deviations of the P@3 scores. The naive estimate of the p-value comes from Equation (9):

$$\begin{aligned} t &= \frac{\bar{j}_a - \bar{j}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} \\ p\text{-value} &= 2\big(1 - \Pr(T_\nu < |t|)\big) \approx 0.011 \end{aligned}$$

suggesting a significant (in the statistical sense) change in performance. To verify this we randomly sampled 143 assessments from the judges and gave those to an in-house expert who acted as the gold judge. Of the 59 the gold judge found relevant, the bronze judge agreed on 43 ($\widehat{m}_{J_R} = 0.729$). Of the 84 the gold judge found non-relevant, the bronze judge agreed on 67 ($\widehat{m}_{J_N} = 0.798$).

Using Equation (5) to compute corrected P@3,

$$\begin{aligned} \widehat{m}_{G,a} &= \frac{\bar{j}_a - 1 + \widehat{m}_{J_N}}{\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1} = 0.805 \\ \widehat{m}_{G,b} &= \frac{\bar{j}_b - 1 + \widehat{m}_{J_N}}{\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1} = 0.828 \end{aligned}$$

Note that $|\widehat{m}_{G,a} - \widehat{m}_{G,b}| > |\bar{j}_a - \bar{j}_b|$, suggesting that after correction the change between a and b might be more significant than previously thought. However, using Equation (7) to update the standard errors (see section 3.5)

$$\begin{aligned} \widehat{\sigma}^2_{G,a} = &\left(\frac{s_a^2}{n_a}\right) \frac{1}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2} \\ &+ \left(\frac{\widehat{m}_{J_R}(1 - \widehat{m}_{J_R})}{n_R}\right) \frac{(\bar{j}_a - 1 + \widehat{m}_{J_N})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} \\ &+ \left(\frac{\widehat{m}_{J_N}(1 - \widehat{m}_{J_N})}{n_N}\right) \frac{(\bar{j}_a - \widehat{m}_{J_R})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} \approx 0.0903^2 \end{aligned}$$

and likewise, $\widehat{\sigma}_{G,b} \approx 0.0923$.

The corrected estimate of the p-value from Equation (11),

$$p\text{-value} = 0.841$$

suggests that there was no statistically significant change in performance over the 10 day period of experimentation.

In summary, when the accuracy rates were not included in the evaluation we were lead to the incorrect conclusion that a statistically significant change had occurred, but by including the accuracy rates the reality proves otherwise. We believe this example demonstrates one compelling reason why it is essential to include accuracy rates from now on.

7. USE ON TREC DATA

In 2007 one of the TREC Enterprise Track tasks was finding key documents that might be useful in generating an overview of a given subject. The assessments were community generated by bronze judges. The organizers used gold judges (topic originators and experts in the field) to re-assess 33 of the 50 topics [2]. There were 3004 reassessed query-document pairs. This appears to be the only TREC collection where we can evaluate our techniques.

We wish to know if there is a statistically significant difference at the 5% level between the top two runs (by P@20).

TREC run	$n_{J,R}$	$n_{G,R}$	$\widehat{m}_{J_R}$	$n_{J,N}$	$n_{G,N}$	$\widehat{m}_{J_N}$	$\bar{j}$@20
DocRun02	17	38	0.447	216	262	0.824	0.527
york07ed4	14	50	0.280	230	285	0.807	0.513

Table 2: TREC Enterprise judge accuracies. The number of documents deemed relevant by the gold judges is $n_{G,R}$. Of those, $n_{J,R}$ are judged relevant by the bronze judges. Columns $n_{G,N}$ and $n_{J,N}$ are for non-relevant documents. The last column is the naive estimate of P@20.

Table 2 presents the data on bronze judge accuracy based on re-assessment by the gold judge. The very low accuracy rate of relevant assessments, $\widehat{m}_{J_R}$, made by the bronze judges is striking. Being less than 50% in both runs, this suggests that $\bar{j}$@20 is a very uncertain estimate of P@20, and that is qualitatively confirmed by Equation (7); which shows that the variance of $\bar{j}$@20 is proportional to a power of $(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^{-1}$, which is large when $\widehat{m}_{J_R}$ is small.

Skipping details, Equations (7), (10 and (11) show that 569,154 queries are needed in order to detect a difference between the two runs significant at $p = 0.05$. Then using Equation (8) shows that 23,027,371 relevant assessments by the gold judge are needed. Both are vastly larger than the 30 topics and 3004 assessments rejudged—hence no statistically sound conclusions about the difference between these two runs can be drawn.

8. CONCLUSIONS

Relevance assessments usually have errors, and it is often not practical to correct all the errors by doing extra assessments or getting more accurate assessors. Our work applies to this common situation. We derive methods that can adjust for assessment errors. We only require that a sample of the assessments be more accurately rejudged.

We show how to adjust P@1, P@k and DCG, along with their standard errors. From these we compute a p-value that can be used to test for statistical significance.

A simulated example demonstrated how our adjustments can recover the value the metric would have had if the assessments did not have errors. We applied the adjustments to our search engine and determined that previously believed improvements were illusory and due to bronze judge inaccuracy. Finally, we applied our metrics to TREC data and showed that insufficient data was available to draw conclusions.

We believe that the extra work necessary to evaluate the accuracy of the bronze judges is worthwhile as it allows us to draw sound conclusions from the necessarily noisy assessments.

APPENDIX

A. DERIVATION OF EQUATION (7)

Start with Equation (6) and collect $\mathbb{V}\left[\overline{J}_N\right]$, giving

$$\widehat{\sigma}_G^2 \approx \frac{\mathbb{V}\left[\overline{J}\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2} + \frac{\mathbb{V}\left[\overline{J}_R\right](\bar{j} - 1 + \widehat{m}_{J_N})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} + S$$

[3] For web search these scores are low, but in eCommerce it is common to see queries for items no longer in the inventory.

where S is

$$\mathbb{V}\left[\overline{J}_N\right]\left(\frac{1}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^2}+\frac{(\overline{j}-1+\widehat{m}_{J_N})^2}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^4}+\right.$$
$$\left.2\frac{\overline{j}-1+\widehat{m}_{J_N}}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^3}\right)$$

Then expand S as

$$S=\frac{\mathbb{V}\left[\overline{J}_N\right]}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^4}\big((\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^2+(\overline{j}-1+\widehat{m}_{J_N})^2$$
$$-2(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)(\overline{j}-1+\widehat{m}_{J_N})\big)$$

or equivalently,

$$S=\frac{\mathbb{V}\left[\overline{J}_N\right]}{(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^4}S_1$$

where

$$S_1=(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)^2-(\overline{j}-1+\widehat{m}_{J_N})^2-2(\widehat{m}_{J_R}+\widehat{m}_{J_N}-1)(\overline{j}-1+\widehat{m}_{J_N})$$

Rewrite S_1 with $z=\widehat{m}_{J_N}-1$ to get

$$S_1=(\widehat{m}_{J_R}+z)^2+(\overline{j}+z)^2-2(\widehat{m}_{J_R}+z)(\overline{j}+z)$$

which simplifies to

$$S_1=\big((\widehat{m}_{J_R}+z)-(\overline{j}+z)\big)^2=(\widehat{m}_{J_R}-\overline{j})^2$$

Putting these pieces together gives Equation (7).

B. REFERENCES

[1] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9, 2008.

[2] P. Bailey, N. Craswell, I. Soboroff, A. P. D. Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *SIGIR 2008*, pages 667–674, 2008.

[3] G. V. Cormack and T. R. Lynam. Statistical precision of information retrieval evaluation. In *SIGIR 2006*, page 533, 2006.

[4] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

[5] L. A. Granka, T. Joachims, and G. Gay. Eye-Tracking Analysis of User Behavior in WWW Search. In *SIGIR 2004*, pages 478–479, 2004.

[6] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *TOIS*, 20(4):422–446, 2002.

[7] M. Joglekar, H. Garcia-Molina, and A. Parameswaran. Evaluating the crowd with confidence. In *KDD 2013*, page 686, 2013.

[8] G. Kazai and M. Lalmas. eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval. *TOIS*, 24(4):503–542, 2006.

[9] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 27(1):1–27, 2008.

[10] R. J. Passonneau and B. Carpenter. The Benefits of a Model of Annotation. *Transactions of the Association for Comptuational Linguistics*, 2:311–326, 2014.

[11] B. Piwowarski, A. Trotman, and M. Lalmas. Sound and complete relevance assessment for XML retrieval. *TOIS*, 27:1:1–1:37, 2008.

[12] M. Sanderson, F. Scholer, and A. Turpin. Relatively relevant: Assessor shift in document judgements. In *ADCS 2010*, pages 60–67, 2010.

[13] F. Scholer, A. Turpin, and M. Sanderson. Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements. In *SIGIR 2011*, pages 1063–1072, 2011.

[14] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM 2007*, pages 623–632, 2007.

[15] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[16] W. Tang and M. Lease. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval*, pages 36–41, 2011.

[17] J. Vuurens, A. de Vries, and C. Eickhoff. How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy. In *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*, pages 21–26, 2011.

[18] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu. A Theoretical Analysis of NDCG Ranking Measures. In *26th Annual Conference on Learning Theory*, pages 1–30, 2013.

[19] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *TOIS*, 28(4):1–38, 2010.

[20] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. *VLDB*, 5(6):550–561, 2012.