Uncertainty in Rank-Biased Precision

Laurence A. F. Park
Centre for Research in Mathematics
School of Computing, Engineering and Mathematics
Western Sydney University, Australia
lapark@scem.westernsydney.edu.au

ABSTRACT

Information retrieval metrics that provide uncertainty intervals when faced with unjudged documents, such as Rank-Biased Precision (RBP), provide us with an indication of the upper and lower bound of the system score. Unfortunately, the uncertainty is disregarded when examining the mean over a set of queries. In this article, we examine the distribution of the uncertainty per query and averaged over all queries, under the assumption that each unjudged document has the same probability of being relevant. We also derive equations for the mean, variance, and distribution of Mean RBP uncertainty. Finally, the impact of our assumption is assessed using simulation. We find that by removing the assumption of equal probability of relevance, we obtain a scaled form of the previously defined mean and standard deviation for the distribution of Mean RBP uncertainty.

1. INTRODUCTION

When evaluating an Information Retrieval system, we require a document collection for the given domain, a sample set of queries, a set of manual relevance judgements for each query over the set of documents, and an evaluation function to summarise the accuracy of the system rankings. The documents and queries can be sampled from the domain in which the retrieval system is intended for, but since relevance judgements are dependent on the documents and queries, they must be created by manually assessing the relevance of each document for each query.

Assessing documents for relevance is a tedious process, therefore many resort to a method of importance sampling, where the top ranked documents supplied by a retrieval system are assessed for relevance and the rest of the documents are left unjudged for that query [6].

Information retrieval evaluation metrics have historically treated unjudged documents as being irrelevant to the given query [1]. Recent metrics, such as Rank-Biased Precision [2], have allowed for the uncertainty introduced by unjudged documents. Unfortunately, the uncertainty information is unused when examining the mean over a set of queries.

In this article, we examine the distribution of uncertainty for RBP and derive equations for the mean and variance of the Mean RBP uncertainty. Our contributions are:

- Derivation of the mean and variance of the distribution of Mean RBP uncertainty and identification of its distribution, under the assumption of constant probability of relevance (Section 2.2),
- A document ranking simulation model based on Wallenius' non-central hypergeometric distribution (Section 3.1).
- A comparison of the mean and standard deviation of Mean RBP when removing the assumption of constant probability of relevance (Section 3.2).

The article proceeds as follows: Section 2 examines the distribution of RBP and Mean RBP uncertainty, and Section 3 examines the effect of the assumptions from Section 2.

2. UNCERTAINTY IN RBP

Rank-biased precision [2] is an Information Retrieval evaluation function that assigns a score contribution based on the position of each relevant document in a ranking, where the score contribution decreases based on a geometric sequence:

$$\text{RBP} = (1-p)\sum_{i=1}^{D} r_i p^{i-1} \tag{1}$$

where $0 \leq p \leq 1$ is the user persistence constant, r_i is the set of relevance judgements, where r_i is 1 if the ith ranked document is relevant, otherwise it is 0.

If the ith ranked document is missing a relevance judgement, then RBP provides a margin of uncertainty of $(1-p)p^{i-1}$. If there are multiple missing relevance judgements then the uncertainty increases by adding the components of uncertainty for each rank. If we define two sets U and J, where J contains the set of documents with relevance judgements, and U contains the set of documents without relevance judgements, we have:

$$\text{RBP} = (1-p)\sum_{i\in J} r_i p^{i-1} + (1-p)\sum_{i\in U} r_i p^{i-1} \tag{2}$$

Since we are missing the set of relevance judgements r_i for $i \in U$, the second term of equation 2 represents the uncertainty in the RBP score, which we call υ. To account for the uncertainty, RBP scores are usually presented as an interval, where the lower bound is the score where all r_i for $i \in U$ are assumed 0, and the upper bound where they are assumed 1. Unfortunately, when computing the mean over a set of queries, we are unable to use the interval to obtain a mean uncertainty interval. In the following sections we will examine the mean and variance of the uncertainty term, which we will use to identify the distribution of the *Mean RBP uncertainty* over a set of queries.

ADCS '16, December 05 - 07, 2016, Caulfield, VIC, Australia

ISBN 978-1-4503-4865-2/16/12...$15.00

DOI: http://dx.doi.org/10.1145/3015022.3015029

2.1 Distribution of Uncertainty

To obtain an interval for the Mean RBP uncertainty, we require knowledge of the distribution of RBP uncertainty for a single query. Identifying the distribution of the RBP uncertainty is difficult, since it depends on the number of relevant documents for the given query, that were not judged, and it also depends on the accuracy of the system being evaluated. For example, we would expect the variation of the uncertainty to increase as the number of unjudged documents increases. We would also expect the mean of the uncertainty to increase as the accuracy of the assessed system increases.

To begin our analysis, we will examine the distribution of r_i for each unjudged document. We make the naïve assumption that:

$$r_i = \begin{cases} 1 & \text{with probability } q \\ 0 & \text{with probability } (1-q) \end{cases} \quad (3)$$

for all $i \in U$. Using this probability distribution, we obtain a uniform distribution over all combinations of judgements for the set of unjudged documents if $q = 0.5$. For q greater than 0.5, the probability increases with the number of relevant documents, and if q is less than 0.5, the probability increases with the number of irrelevant documents.

Using the distribution of r_i, we can compute the mean and variance of the RBP uncertainty υ:

$$\begin{aligned} \mathbb{E}[\upsilon] &= \mathbb{E}\left[(1-p)\sum_{i\in U} r_i p^{i-1}\right] \\ &= (1-p)\sum_{i\in U} \mathbb{E}[r_i] p^{i-1} \\ &= (1-p)\sum_{i\in U} q p^{i-1} \\ &= (1-p)q\sum_{i\in U} p^{i-1} \end{aligned}$$

where the mean of r_i is q, and

$$\begin{aligned} \mathrm{Var}(\upsilon) &= \mathrm{Var}\left((1-p)\sum_{i\in U} r_i p^{i-1}\right) \\ &= (1-p)^2\sum_{i\in U} \mathrm{Var}(r_i) p^{2(i-1)} \\ &= (1-p)^2\sum_{i\in U} q(1-q) p^{2(i-1)} \\ &= (1-p)^2 q(1-q)\sum_{i\in U} p^{2(i-1)} \end{aligned}$$

where the variance of r_i is $q(1-q)$. The distribution of RBP uncertainty for different values of q is shown in Figure 1.

2.2 Uncertainty in Mean RBP

Evaluation is usually performed over a large sample of queries and the mean of the evaluation over the set of queries is provided as the overall evaluation score. When computing the Mean RBP score over a set of queries, we are faced with the problem of how to aggregate the uncertainty.

Mean RBP over the set of queries Q is given as:

$$\text{Mean RBP} = \frac{1}{|Q|}\sum_{k\in Q}\left[(1-p)\sum_{i\in J} r_{k,i} p^{i-1} + \upsilon\right] \quad (4)$$

$$= \frac{1}{|Q|}\sum_{k\in Q}(1-p)\sum_{i\in J} r_i p^{i-1} + \Upsilon \quad (5)$$

where $|Q|$ is the cardinality of Q, the first term is the Mean RBP over the judged documents (containing $r_{k,i}$ the relevance judgement for the ith ranked document from query k) and the second term:

$$\Upsilon = \frac{1}{|Q|}\sum_{k\in Q} \upsilon \quad (6)$$

is the contribution to the mean from the unjudged documents. If we assume that υ is independent of the query and the query size is large (at least 30 queries), then according to the Central Limit Theorem, Υ is approximately Normal, with mean equal to $\mathbb{E}[\upsilon]$ and variance equal to:

$$\mathrm{Var}(\Upsilon) = \frac{\mathrm{Var}(\upsilon)}{|Q|} \quad (7)$$

The distribution of Mean RBP uncertainty (Υ) for different values of q is shown in Figure 2. Using this distribution of Υ, we can compute an uncertainty interval for Mean RBP using the quantiles of Υ:

$$\begin{aligned} &\frac{1-p}{|Q|}\sum_{k\in Q}\sum_{i\in J} r_{k,i} p^{i-1} + (1-p)q\sum_{i\in U} p^{i-1} \\ &\qquad \pm z_{1-\alpha/2}\sqrt{\frac{(1-p)^2 q(1-q)\sum_{i\in U} p^{2(i-1)}}{|Q|}} \end{aligned} \quad (8)$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Standard Normal distribution, providing a $100(1-\alpha)\%$ uncertainty interval for Mean RBP[1].

3. IMPACT OF ASSUMPTION

While deriving the distribution of Mean RBP uncertainty, we made the assumption that each unjudged document has the same probability q of being relevant. In this section, we will evaluate the impact of this assumption on the distribution of Mean RBP uncertainty, by using a document ranking model to simulate a set of document rankings and comparing the predicted distribution of Mean RBP uncertainty to the actual distribution of Mean RBP uncertainty provided by the model.

3.1 Document ranking model

To examine the accuracy of our uncertainty interval for Mean RBP, we require a large sample of document rankings with complete relevance judgements (giving no uncertainty). Using this sample, we can remove relevance judgements to examine the effect of them being missing on the uncertainly interval. To have control of this experiment, we will simulate a set of document rankings with relevance judgements. Simulating the rankings allows us to control the parameters of the ranking and have full knowledge of the distribution of relevant documents in each ranking.

A document ranking is the ordered set of documents provided by a retrieval system for a given query, where the top ranked document is predicted by the system, as the most relevant to the query. During evaluation, we are only concerned with the relevance at each rank. Therefore, we consider a document ranking to be an ordered set of relevance judgements, where the judgements are binary (either relevant or irrelevant), and they are conditioned on the document retrieval system. The top ranked document is considered the most likely to be relevant by the retrieval system, but may or may not be relevant. A more accurate retrieval system is more likely to rank relevant documents before ranking irrelevant documents. We

[1]If $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96$, giving a 95% interval.

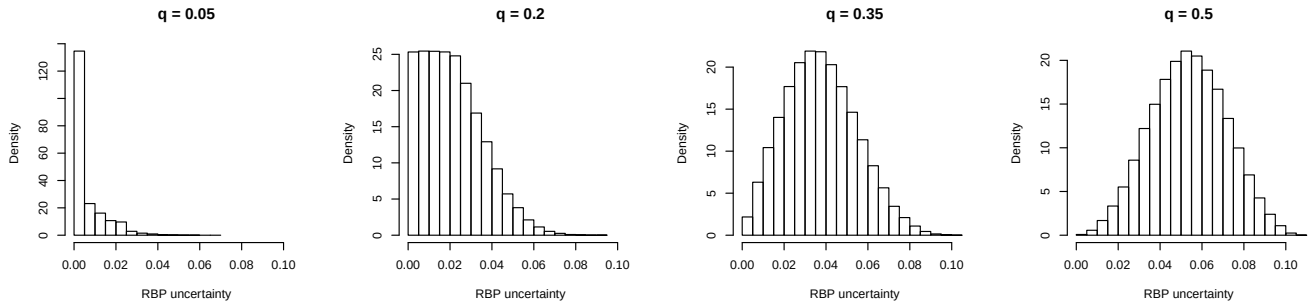

Figure 1: The distribution of *RBP uncertainty* υ, for $p = 0.8$ and $q = 0.05, 0.2, 0.35$ and 0.5, where documents ranked 11 to 100 are unjudged. Due to the symmetrical nature of the uncertainty υ, the distributions for $q = 0.95, 0.8, 0.65$ and 0.5 can be obtained by reflecting the distributions about their vertical centre line.

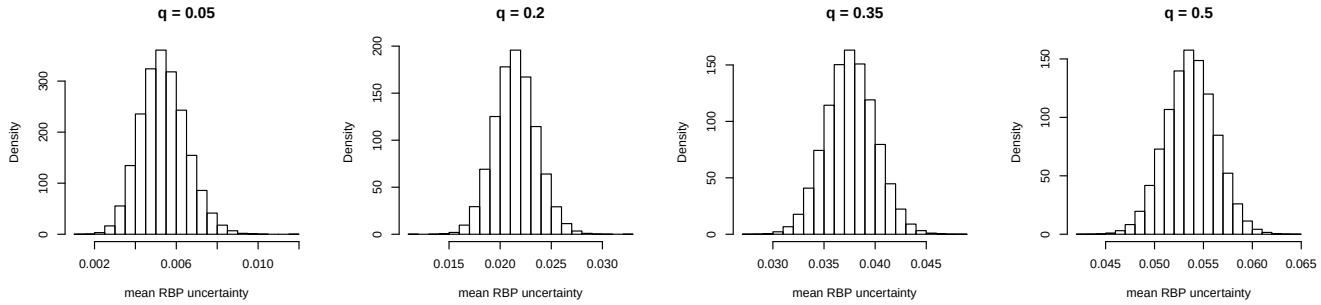

Figure 2: The distribution of *Mean RBP uncertainty* Υ, for $p = 0.8$ and $q = 0.05, 0.2, 0.35$ and 0.5, where documents ranked 11 to 100 are unjudged over 50 queries.

can model this process using random draws from an urn containing black and white balls.

Given a set of N documents, where M are relevant to a given query, we can simulate an imperfect retrieval system's ranking using N balls in an urn, where M are black and $N-M$ are white. The ranking produced by the system is generated by sampling from the urn without replacement and recording the colour. The colour of the first ball drawn represents the relevance of the first ranked document, the colour of the second ball drawn represents the relevance of the second ranked document, and so on. The position of the relevant documents in the ranking is mainly effected by two properties, the accuracy of the retrieval system and the difficulty of the query. We parameterise the urn model by introducing weights for the set of black and white balls. The weight associated to a ball is proportional to its probability of being drawn. If the weight of each black ball is greater then the weight of each white ball, then the resulting ranking is likely to have more relevant documents ranked ahead of irrelevant documents. Therefore the weight of each black and white ball is a function of the accuracy of the retrieval system and the difficulty of the query. To simplify the model, we assume that all back balls have the same weight w_b and all white balls have the same weight w_w. This model is related to Wallenius' non-central hypergeometric distribution [4], but instead of obtaining the count of black balls after k balls are drawn, we are interested in the order in which the balls are drawn.

Given an urn containing N balls, where M are black and $N-M$ are white, and the probability of drawing a given black ball is w_b, while the probability of drawing a given white ball is w_w; the probability of drawing any black ball (or equivalently, the probability of the first ranked document being relevant to the given query) is:

$$P(D = b) = \frac{w_b M}{w_b M + w_w (N - M)} \tag{9}$$

and the probability drawing a white ball (the probability of the first ranked document being irrelevant to the query) is $1 - P(D = b)$, where $w_b > 0$ and $w_w > 0$. The probability of drawing a black ball, given that c_b black balls have been drawn and c_w white balls have been drawn (the probability of having a relevant document at rank $c_b + c_w + 1$, given that c_b relevant documents have been seen and c_w irrelevant documents have been seen in earlier ranks) is:

$$P(D = b|c_b, c_w) = \frac{w_b(M - c_b)}{w_b(M - c_b) + w_w(N - M - c_w)} \tag{10}$$

where $0 \leq c_b \leq M$ and $0 \leq c_w \leq (N - M)$, and $P(D = w|c_b, c_w) = 1 - P(D = b|c_b, c_w)$. Note that the probability of a given draw is dependent on the number of each colour drawn, but independent of the previous sequence of draws. Dividing the top and bottom of the fraction by w_b allows us to combine the weights into one weight.

$$P(D = b|c_b, c_w) = \frac{(M - c_b)}{(M - c_b) + w_w/w_b(N - M - c_w)} \tag{11}$$

$$= \frac{(M - c_b)}{(M - c_b) + w(N - M - c_w)} \tag{12}$$

where $w > 0$. If we set $w = 1$ for this model, then each ball (document) has the same chance of being selected at each draw, providing a poor ranking. If we set $w \ll 1$ for this model, then the sequence of draws is likely to have black balls ranked ahead of white balls (relevant documents ranked ahead of irrelevant doc-

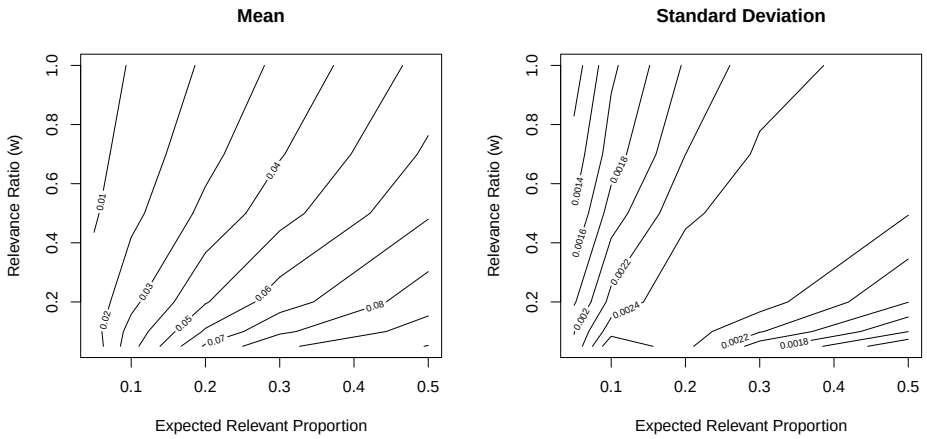

Figure 3: Contour plots of the mean and standard deviation of the uncertainty distribution of Mean RBP using simulated rankings from Section 3.1. Each plot shows the effect of the Relevance Ratio w and Expected Relevant Proportion q on the mean and standard deviation of the distribution of Mean RBP uncertainty.

uments), therefore w represents the system inaccuracy and query difficulty.

3.2 Experiment

We want to examine the effect of our assumption of constant probability of relevance (equation 3), on the mean and variance of the Mean RBP uncertainty distribution. To do this, we use the proposed document ranking model from Section 3.1. For this model, we set the number of documents $N = 100$, RBP persistence $p = 0.8$ (similar to Web user persistence [3, 5]), and randomly assign the number of relevant documents $M \sim \text{Binomial}(N, q)$. By setting the model relevance ratio w to 1, we simulate the case where all documents have an equal probability of relevance, allowing us to obtain the same uncertainty distribution of the Mean RBP from Section 2.2. As we decrease w, we observe the effect that increasing the probability of ranking relevant documents higher than irrelevant documents has on the uncertainty distribution.

We ran the simulation using w = 0.05, 0.1, 0.2, 0.5, 0.7 and 1, and q = 0.05, 0.1, 0.2, 0.3 and 0.5. We assumed the top ten documents were manually judged and so recorded the contribution to RBP made by the remaining (considered unjudged) documents ranked 11 to 100. The Mean RBP was computed over 50 generated ranked lists to simulate 50 queries. This process was replicated 10000 times to obtain an uncertainty distribution for Mean RBP for each (w, q) pair. The distributions for Mean RBP uncertainty is Normal, so we computed the mean and standard deviation of the distribution, and presented them using a contour plot in Figure 3. The results for $w = 1$ are the same as those computed by our model in Section 2.2. As the relevance ratio w decreases, we find that both the mean and the standard deviation of the uncertainty distribution decrease in an almost linear fashion, implying that we can scale the mean and standard deviation of Υ to remove the effect of our assumptions.

4. CONCLUSION

Information retrieval metrics that provide uncertainty intervals due to the uncertainty of relevance of unjudged documents, provide us with an indication of the upper and lower bound of the retrieval system's score. Unfortunately, the uncertainty is disregarded when examining the mean over a set of queries, and the mean over the lower bound is presented as the mean retrieval system score.

In this article, we examined the uncertainty distribution for Rank-Biased Precision (RBP) under the assumption that all unjudged documents have the same probability of relevance q. We found that the distribution of Mean RBP uncertainty is Normal when computing the mean over at least 30 queries, and we also found closed form equations for the mean and variance of the distribution of Mean RBP uncertainty.

To examine the impact of our assumption that unjudged documents have equal probability of relevance, we ran a simulation, controlling the number of relevant documents and their position in the ranked list. We found that removing our assumption leads to a scaled form of our derived mean and variance.

This initial work on examining the distribution of Mean RBP uncertainty has provided us insight into the distribution, but more work is needed on determining the distribution parameters (such as Expected Relevant Proportion and Relevance Ratio) in practice.

References

[1] Chris Buckley and Ellen M Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32. ACM, 2004.

[2] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, December 2008.

[3] Laurence A. F. Park and Yuye Zhang. On the distribution of user persistence for rank-biased precision. In *Proceedings of the 12th Australasian document computing symposium*, pages 17–24, 2007.

[4] Kenneth T Wallenius. Biased sampling; the noncentral hypergeometric probability distribution. Technical report, DTIC Document, 1963.

[5] Yuye Zhang, Laurence A. F. Park, and Alistair Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13(1):46–69, 2010.

[6] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314. ACM, 1998.