Power Analysis for Interleaving Experiments by means of Offline Evaluation

Hosein Azarbonyad
University of Amsterdam, Amsterdam,
The Netherlands
h.azarbonyad@uva.nl

Evangelos Kanoulas
University of Amsterdam, Amsterdam,
The Netherlands
e.kanoulas@uva.nl

ABSTRACT

Evaluation in information retrieval takes one of two forms: collection-based offline evaluation, and in-situ online evaluation. Collections constructed by the former methodology are reusable, and hence able to test the effectiveness of any experimental algorithm, while the latter requires a different experiment for every new algorithm. Due to this a funnel approach is often being used, with experimental algorithms being compared to the baseline in an online experiment only if they outperform the baseline in an offline experiment. One of the key questions in the design of online and offline experiments concerns the number of measurements required to detect a statistically significant difference between two algorithms. Power analysis can provide an answer to this question, however, it requires an a-priori knowledge of the difference in effectiveness to be detected, and the variance in the measurements. The variance is typically estimated using historical data, but setting a detectable difference prior to the experiment can lead to suboptimal, upper-bound results. In this work we make use of the funnel approach in evaluation and test whether the difference in the effectiveness of two algorithms measured by the offline experiment can inform the required number of impression of an online interleaving experiment. Our analysis on simulated data shows that the number of impressions required are correlated with the difference in the offline experiment, but at the same time widely vary for any given difference.

Keywords

Information retrieval evaluation, Interleaving, Experimental design, Power Analysis

1. INTRODUCTION

Evaluation is key to building effective, efficient, and usable information retrieval (IR) systems as it enables their effectiveness to be quantified. For decades the primary approach to conduct IR evaluation was the Cranfield approach, that makes use of test collections, allowing systematic and repeatable evaluations to be carried out in a controlled manner. The Cranfield approach has often been critisized for not quantifying the actual user experience when served by the search algorithm under testing. Online evaluation on the other hand – that is A/B testing [7] and Interleaving [6] – is performed on the basis of user interactions with the ranked list of results produced by the algorithms being tested.

Designing an experiment that enables the discovery of statistical significant effects is of paramount importance in both the online and the offline setting. Power analysis allows to determine the sample size (i.e. the number of measurements collected) required to detect an effect of a given size with a given degree of confidence. Conducting a power analysis however requires setting the effect size to be detected prior to the experiment [7, 12]. This often leads to suboptimal results: the actual effect size observed after having run the experiment is larger than the one specified, and therefore required fewer measurements than the ones dictated by the power analysis.

In this paper we make use of the *funnel approach* often used in information retrieval evaluation [7], according to which, an experimental algorithm is compared to a baseline/production algorithm in an online experiment only if it outperforms the production algorithm in an offline experiment. The question we answer in this work is whether the evaluation results of the offline experiment can inform the number of impressions required, and hence the duration, of an online interleaving experiment. In other words, given that two systems have $\Delta M = \alpha$, with M being any evaluation measure in an offline experiment, can we detect the number of impressions required for an interleaving experiment to conclude that the two systems are statistically significantly different? To study this question we first construct all the different pairs of relevance label rankings for which $\Delta M = \alpha$; then we interleave the rankings and simulate user clicks. We use the proportion of wins for the winning algorithm to calculate the number of impressions required for the experiment to consider that proportion statistically significant, and by repeating the experiment for different values of α try to establish a correlation between the number of required impressions and ΔM.

2. METHOD

Commercial search engines typically take a *funnel approach* in evaluating a new search algorithm [8]: first the experimental algorithm (E) is compared to the production algorithm (P) using an offline test collection; if E outperforms P with respect to the evaluation measure of their interest, the two algorithms are then compared online, e.g. through an interleaving experiment. In an interleaving experiment the ranked lists of results produced by P and E (for some user query) are interleaved in a single ranked list which is then presented to the user. The user interacts with the interleaved results by clicking on them, and the algorithm that receives most of the clicks wins the "dual". The experiment is repeated for a number of times (impressions) and the total wins for P are compared to

ICTIR '16, September 12-16, 2016, Newark, DE, USA

DOI: http://dx.doi.org/10.1145/2970398.2970432

those of E [10]. A Sign Test is then run to assess whether the difference in wins between the two algorithms is statistically significant. Alternatively a Binomial Proportion Test can be used for which one calculates the proportion of times E wins and tests whether this proportion, p, is greater than $p_0 = 0.5$. Prior to running an online experiment one would like to know what is an adequate sample size, i.e. what is the number of impressions required to detect a statistically significant difference between E and P. The proportion of time E winning is not known a priori, i.e. before actually running the experiment. If it was known one could perform a power analysis for the Binomial Proportion Test and find the necessary number of impressions. What is known is the margin by which E outperformed P in the offline experiments.

In this section we propose an experiment to test whether one can determine the sample size (i.e. the number of interleaved impressions) required for an interleaving experiment to identify statistical significant differences, given the margin, ΔM, by which E outperformed P in an offline experiment as measured by an evaluation measure M.

The experiment works as follows:

1. We generate pairs of ranked relevance labels, R_E and R_P for the experimental and production system respectively.
2. We measure the quality of the two rankings by some measure M, and compute the difference $\Delta M = M(R_E) - M(R_P)$.
3. We apply an interleaving algorithm to interleave the produced pairs of rankings.
4. We simulate user clicks over the interleaved ranking; for each of interleaved ranking we run k simulations of user interactions (clicking behaviour).
5. We measure the proportion p of wins for E against P.
6. We run a power analysis for the Binomial Proportion Test to determine the number of impressions required to find the observed proportion of win p statistically significant.

We use p to compute the sample size needed to detect such a proportion in a statistically significant manner. We allow a chance of falsely rejecting the null hypothesis (i.e. concluding that E is better than P, when it is not) of 5%, and a chance of falsely not rejecting the null hypothesis (i.e. not concluding that E is better than P, when it is) of 10%. We use the computed probability p for determining the number of required impressions using the proportion test. We assume that the sampling distribution for proportions can be approximated by a normal distribution [11], and therefore we can use the following equation for computing the minimum sample size [5]:

$$N' \geq (\frac{z_{1-\alpha}\sqrt{p_0(1-p_0)} + z_{1-\beta}\sqrt{p_1(1-p_1)}}{\delta})^2, \quad (1)$$

where we set $p_0 = 0.5$ to reflect that each system wins 50% of times. p_1 is the proportion of times E wins over P out of k simulations. α and β specify the level of significance. We set $\alpha = 0.05$ and $\beta = 0.1$. δ is $|p_1 - p_0|$ and z is the standard normal distribution. Finally, using the continuity correction the minimum sample size is determined as:

$$N = N' + 1/\delta \quad (2)$$

The procedure of determining the minimum number of impressions for an interleaved online experiment given ΔM is shown in Figure 1. Given the rankings produced by E and P and also the difference of their performance in offline evaluation, we simulate k experiments (line 2 to 10). In each simulation, we first interleave the rankings (line 3), then we produce clicks on the interleaved ranking using a click model (line 4). Then, in lines 6 to 9, we determine which system wins based on the number of clicks the documents corresponds to each system receive. Finally, based on the proportion of wins of E over P, we determine the number of required impressions in line 12.

```
1:  procedure POWER ANALYSIS
        Input:
          R_E: ranking produced by E
          R_P: ranking produced by P
          ΔM: the delta of performance of E and P
        Output:
          N: number of impressions
2:      for i ← 1 to k do
3:          R = Interleave(R_E, R_P)
4:          Clicks = ClickModel(R)
5:          if Clicks_E > Clicks_P then
6:              wins_E ← wins_E + 1
7:          else
8:              wins_P ← wins_P + 1
9:          end if
10:     end for
11:     p_1 = wins_E/(wins_E + wins_P)
12:     N = determine sample size using Equation 2
13: end procedure
```

Figure 1: The procedure of determining number of required impressions in an online experiment given the performance of two systems in offline evaluation.

A number of assumptions are being made by the above experiment:

1. *Accurate user behaviour*: We assume that the parametrized click models employed in the simulation are accurate, and representative of actual user bahviour.
2. *No query variability*: We assume that both the experimental and production system result in the exact same ranked list of relevance for all queries in the offline experiment, or in other words, there is a single query in the offline test collection, and hence no variability due to queries. In this way we integrate out any variance in the experiments due to the different nature of queries and only focus on variance due to the differences in user behaviour. We leave the integration of query variability as future work.
3. *Offline/Online query set alignment*: We assume that the queries (query) used in the online experiment is exactly the same as the one(s) used in the offline experiment. We leave the study of sampling online queries for the offline evaluation, and its impact on the power analysis of the online experiment as future work.
4. *Distinct documents*: We assume that the ranked list of documents produced by E and P are distinct, so that the analysis is not overly complicated by accounting for the de-duplication step of the interleaving algorithms.

The aforementioned assumptions essentially lead to a lower bound on the number of required impressions.

3. EXPERIMENTAL SETUP

In this paper we aim at answering two research questions:

RQ1 Are (simulated) interleaving and collection-based evaluation measures in agreement regarding the relative performance of two ranking algorithms?

RQ2 Can the retrieval performance in collection-based evaluation inform the sample size (i.e. inform the required number of impressions) of the interleaving experiment?

RQ1 concerns the correlation of the offline and (simulated) online effectiveness measures. To answer this question, we first generate pairs of rankings of relevance, for a production system P and an experimental system E, respectively. For that we assume a 5-graded relevance, i.e. *rel* $\in \{0, 1, 2, 3, 4\}$, and construct all possible P and E ranking pairs of length 5. We consider only those pairs for which E outperforms P in the offline evaluation. For each ranking pair we calculate $\Delta M = M(R_E) - M(R_P)$. We also calculate p_1, the chance that the experimental system E will outperform the production system P in the online experiment, by using the algorithm described in Figure 1 (line 11 of the algorithm). Finally we plot ΔM against p_1 to test the agreement between offline and online effectiveness measures. The results of experiments corresponding to RQ1 are described in §4.1.

RQ2 concerns the impact of ΔM on the required number of impressions in the online evaluation. To answer this question, again we generate pairs of rankings of relevance for P and E, and calculate the difference $\Delta M = M(R_E) - M(R_P)$. We still consider only those pairs for which E outperforms P and we group the differences in 10 bins/groups. For each bin/group, we calculate the average number of impressions that are required for the interleaving experiment to conclude that two systems are statistically significantly different. For that we use the algorithm described in Figure 1.The result of experiment regarding RQ1 are discussed in §4.2.

The collection-based evaluation measures used in our experiments are the Discounted Cumulative Gain (DCG), the Ranked Biased Precision (RBP), and the Expected Reciprocal Rank (ERR). Team-Draft interleaving [10] is used to interleave the rankings by P and E, and the Simplified Dependent Click Model (SDCM) [3], the User Browsing Model (UBM) [4], and the Dynamic Bayesian Network Model (DBN) [1] to simulate user clicks on the interleaved ranked list (line 4 of the algorithm). The click model parameters were estimated using the Yandex Relevance Prediction challege click log. The number of simulations (k in Figure 1) was arbitrarily set to 50. The open-source implementation of our approach is available[1].

4. RESULTS

In this section, we analyze the correlation of ΔM in offline evaluation with the number of required impressions in the interleaved experiments.

4.1 Correlations of offline and online evaluation measures

Figure 2 shows the proportions of wins of E over P in the simulated interleaving experiments for different values of ΔM. We only considered cases for which $0 < \Delta M \leq 0.1$. As illustrated in the figure, as ΔM increases p_1 also increases. However, it can also be observed that there is a high variance for p_1 for similar values of ΔM. This is inline with past results comparing in-situ interleaved and collection-based experiments [2, 9, 10].

[1] https://github.com/HoseinAzarbonyad/Power-Analysis-for-Interleaving-Experiments

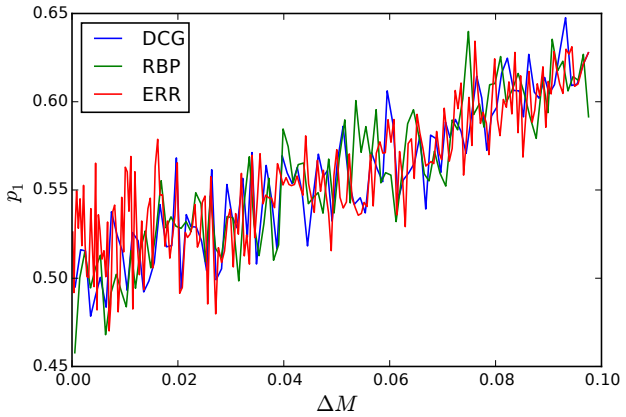


Figure 2: The proportions of wins of E over P for different values of ΔM. SDCM is used for simulating clicks over interleaved lists.

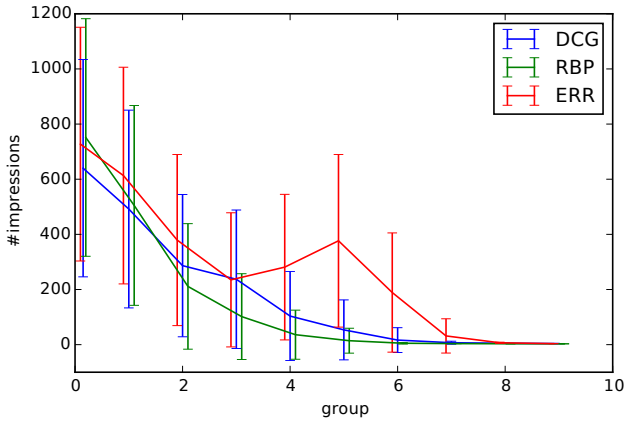
4.2 Analysis on the sample size in online evaluation

Figure 3 shows the number of impressions required by an interleaving experiment given that the difference in retrieval effectiveness as measured by a collection-based experiment is ΔM. The median and the 68% confidence interval (i.e. one standard deviation) is shown in the figure for each ΔM group. As it can be observed, as the value of ΔM increases, the number of required impression decreases, as expected.

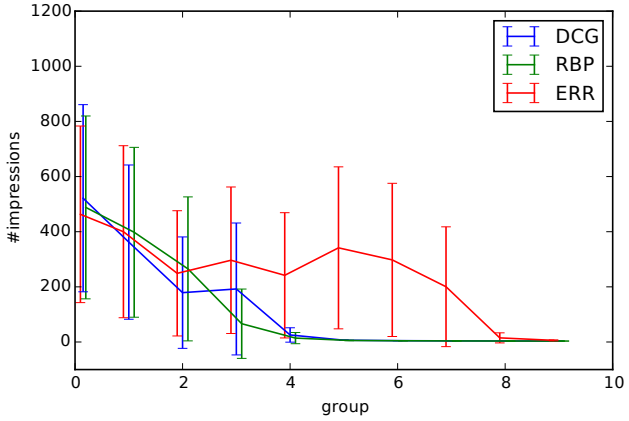
The second observation is that the required number of impressions demonstrates a high variance, in particular for low values of ΔM. This result can be explained by the fact that $\Delta M = \alpha$ can happen in many different ways: relevance differences high in the ranking with small relevance gap, or low in the ranking with high relevance gap. Not not all of these cases however lead to an identical user behaviour. Consider two examples: 1) E returns the following ranked list of relevance labels, (4, 2, 2, 1, 0), while P returns (2, 2, 3, 1, 0), and 2) E returns (1, 1, 1, 3, 0) and P returns: (1, 2, 1, 0, 0). In both cases ΔDCG is about 0.2, however in the first example, if we simulate clicks using SDCM and estimate the number of required impressions using the algorithm described in Figure 1, we need 277 impressions, while the number for the second example is 381. In expectation interleaving will be able to identify that E outperforms P, however this will require less number of impressions for the former case than the latter.

5. CONCLUSION AND FUTURE WORK

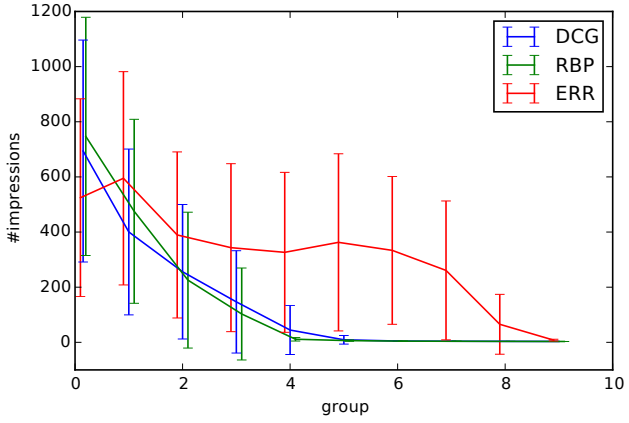
In this work, we first posed the question of whether the performance difference between two ranking algorithms in an offline experiment can inform the required number of impressions for an online experiment to identify statistically significant results and proposed a simulation experiment to answer this question. Based on the proposed experiment, we (1) illustrated a strong correlation between the margin of the quality difference in offline evaluation and the required number of impressions in online evaluation – something expected – , but also (2) a large variability in the number of required expressions for the same margin. This indicates that the results of an offline experiment can inform the design of an online experiment, and in particular the power of the experiment to evaluate a new experimental system, it is beneficial to first conduct an offline evaluation and then based on the results achieved in offline evaluation, an estimate about the duration of the online experiment can be made.



(a) SDCM



(b) UBM



(c) DBN

Figure 3: Number of required impressions for online evaluation based on different groups of values of ΔM

The current work is simply a step forward towards better understanding the relation between offline collection-based evaluation, and online in-situ evaluation. In the future, we intent to extend the current work in a number of directions: (1) *Account for query variability*: The current work assumes that all offline/online queries are exactly the same. In other words, it integrates out any variability coming from queries, and only consider variability due to user interactions with the SERP; we intent to extend the experiment to account for query variability as well. (2) *Explore A/B Testing*: The current work only focuses on interleaved experiments, integrating out between group variability. We intent to extend the experiment to account for this variability, present in A/B testing experiments. (3) *Closed-form solution*: The current work explores the power of an interleaved comparison between two ranking algorithms experimentally. We seek for a closed-form solution to the problem. (4) *Online experiments*: So far we have studied the problem using simulations; we intent to validate our findings using actual online experiments.

Acknowledgments

This work was supported in part by the Google Faculty Research Award scheme. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

REFERENCES

[1] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, 2009.

[2] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst.*, 30(1), 2012.

[3] A. Chuklin, I. Markov, and M. d. Rijke. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2015.

[4] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, 2008.

[5] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[6] T. Joachims. Evaluating retrieval performance using clickthrough data. In *Text Mining*, pages 79–96. 2003.

[7] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 2008.

[8] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. Online controlled experiments at large scale. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 1168–1176, 2013.

[9] F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 667–674, New York, NY, USA, 2010. ACM.

[10] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, 2008.

[11] D. Wackerly, W. Mendenhall, and R. L. Scheaffer. *Mathematical statistics with applications*. Nelson Education, 2007.

[12] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, 2008.