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, 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. 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. This submission is based on [1].

Keywords

Information retrieval evaluation, Interleaving, Experimental design, Power Analysis

Summary of the Research

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 [3] and Interleaving [2] – 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 [4, 3]. 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 [3], 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 using click models trained on a real click log data. We determine the winner system based on the clicks (a system wins if the user clicks on a document in the interleaved list that is coming from the ranking provided by the system). We repeat this procedure several times and count the number of wins of each system. Then, 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.

Based on the proposed experiment, the results and analysis show that (1) there is 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) there is 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.

1. REFERENCES

[1] H. Azarbonyad and E. Kanoulas. Power analysis for interleaving experiments by means of offline evaluation. ICTIR '16, 2016.

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

[3] 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.

[4] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. CIKM '08, 2008.