SIGIR Keynote: Proof By Experimentation? Towards Better IR Research

Norbert Fuhr
University of Duisburg-Essen, Germany
norbert.fuhr@uni-due.de

Abstract

Most IR experiments lack both internal and external validity. Top performance is mostly an illusion, given the lack of solid statistical evidence. Thus, reviewers should ignore performance numbers in their judgement, and pay more attention to proper methodology. Researchers need to focus more on performance analysis and prediction.

1 Introduction

Empirical foundation in the form of experiments plays an important role in IR research. Scientific standards require such experiments to be both internally and externally valid. The former demands that claims are supported by the data, while the latter refers to the extent to which results of a study can be generalised. Unfortunately, many IR experiments have serious flaws, as pointed out by Fuhr [2017]. Here we give an updated version of some recommendations from this paper and propose further actions, also with regard to external validity[1].

2 Improving validity of IR experiments

Don't use measures like RR or AP. Many IR measures are not interval-scaled [Ferrante et al., 2017], and thus computation of averages or applying significance tests requiring an interval scale is not appropriate. As recent own experiments have shown [Ferrante et al., 2021], average precision (AP) and reciprocal rank (RR) are the worst of these measures, since more than 10% (more than 30% in the case of RR) of the significance tests change their outcome when they are mapped onto a proper interval scale. Moreover, we found that measures using a recall base suffer from the problem that measurement scales for different queries are incompatible with each other; thus, any comparison or aggregation of their values (e.g. max, mean, median) is invalid. Other measures like RBP or (n)DCG are also affected by these problems, but to a lesser extent.

[1] Slides of the talk can be downloaded from https://www.is.inf.uni-due.de/bib/pdf/talks/Fuhr_20ta.pdf

Instead of relative improvements, regard effect size. Having computed the (mostly invalid) arithmetic means for different systems, many authors compute percentage of changes. As described in [Fuhr, 2017], this contradicts basic statistical principles and would be admissible only in the case of geometric means[2], for which ratio scales are required. A more feasible method is the computation of effect size, which also considers the variance of results. Statistical tests like e.g. Student's t-test also consider effect size, but in combination with sample size. This has the disadvantage that even tiny effect sizes might be interpreted as being significant, if the sample size is just big enough.

Do not perform multiple tests without correction. When significance tests are applied in a paper, there are often multiple tests performed on the same data set. However, this leads to the problem that the probabilities of the type I errors[3] of the tests add up. To avoid this problem, the significance levels of the individual tests must be corrected, e.g. with the Bonferroni or the Bonferroni-Holm method. If Student's t-test is applicable (interval scale!), then one can use Tukey's honestly significant difference (HSD) test instead, which performs a proper test on all possible pairs. Especially for evaluation campaigns, this is the most appropriate method and includes the necessary correction of the significance level.

There are no significant improvements for re-used test collections. In most evaluations, test collections originating from evaluation campaigns are used. This means that other researchers have performed significance tests on the same data sets before – the participants of the original campaign plus other researchers working with this data. Thus, for proper testing, one should at least consider the number of the (published) experiments using the same data set (still ignoring the much higher number of unpublished work this collection has been used for). Thus, one would have to correct the significance level, taking into account at least dozens, maybe hundreds or thousands of other experiments employing this data set. As shown by Carterette [2015], this leads to the situation that the best results for test collections are completely random! However, even if one would consider all the previous experiments of a test collection, there is still the sequential learning problem, i.e., that a researcher develops her models and methods with the outcomes of previous experiments on the same data set in mind. This, however, contradicts the basic assumption of statistical testing that we know nothing about the performance of similar methods on the same data set. Thus, in the end, the only conclusion is that as soon as we learn about the results of other experiments on the same collection, we can no longer perform significance tests on this data. The outcomes of such tests are meaningless.

The current form of leaderboards is too naïve. In the case of leaderboards, only the maintainer can perform runs on the test collection, so the total number of experiments for this collection is known. As the term 'leaderboard' suggests, the focus is on the top performing runs; however, no statistical testing is applied. The least thing to do here would be the application of tests with appropriate corrections of the significance levels (still ignoring the sequential learning

[2] Let $E(.)$ denote the expectation and A, B be random variables, then $E(A-B) = E(A) - E(B)$, but $E(A/B) \neq E(A)/E(B)$. For geometric means, we have $E(\log A/B) = E(\log A - \log B) = E(\log A) - E(\log B)$.

[3] Rejection of a true null hypothesis, i.e. declaring a significant difference when there actually is none.

problem), which would inform the community if there are actually significant differences between (usually groups of) runs.

Conferences and journals need to accept papers with negative results. Reviewers use the performance of a method studied as a major criterion for judging about the quality of a paper. However, as the previous remarks about the statistical validity of experimental results have shown, the performance numbers are often just random, and focusing on positive results only will lead to the acceptance of a large fraction of papers whose findings cannot be reproduced. Moreover, if negative results are not reported, other researchers might waste their time on investigating the same ineffective methods. Several studies have also shown that many authors use weak baselines for comparison; so we must assume that results for collections lacking baselines from independent research are even worse in that respect. In other disciplines, for instance in psychology, there are journals where papers are reviewed before the study described is actually carried out. Taking all these aspects together, the best solution for IR would be to review papers without seeing the actual results – just based on the underlying concepts and the methodology used.

Evaluation initiatives should run only tracks using proper measures and methods. As a consequence of the problems with multiple and sequential testing described above, the only way for achieving solid numbers is via evaluation initiatives; here no participant has seen the test collection before, nor learned about the outcomes of other approaches applied to the same data. This puts the responsibility for internal validity on the track organisers. So they must choose proper evaluation measures and statistical testing methods to be applied. As tracks are usually approved by the program committee of the evaluation initiative, this committee is in the position to ensure that only tracks with proper methodology are admitted – otherwise the scientists participating in the track will waste their time producing meaningless results.

External validity requires performance analysis and prediction. There is very little work on external validity in IR. The ultimate goal here is to make predictions about the performance of methods on new collections, before experiments are run on them. Ferro et al. [2018] describe a framework model for this issue, outlining gaps in current research. As a starting point, researchers should carry out thorough performance analyses of their methods, pointing out their strengths and weaknesses and identifying the major factors affecting performance.

3 Conclusion

Here we have outlined some important issues regarding the validity of IR experiments. Current research focuses too much on top results for overall performance, which is an illusion given the (in most cases) non-existent statistical evidence. Program committees of conferences, journals and evaluation initiatives should put more emphasis on internal and external validity, in order to stop researchers from wasting their time with questionable results.

References

Norbert Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, 2017. URL http://sigir.org/wp-content/uploads/2018/01/p032.pdf.

Marco Ferrante, Nicola Ferro, and Silvia Pontarollo. Are IR evaluation measures on an interval scale? In *Proc. ICTIR*, page 67–74, 2017. URL https://doi.org/10.1145/3121050.3121058.

Marco Ferrante, Nicola Ferro, and Norbert Fuhr. Towards meaningful statements in IR evaluation. mapping evaluation measures to interval scales. Submitted for publication, 2021.

Ben Carterette. The best published result is random: Sequential testing and its effect on reported effectiveness. In *Proc. SIGIR*, pages 747–750, 2015.

Nicola Ferro et al. The Dagstuhl Perspectives Workshop on performance modeling and prediction. *SIGIR Forum*, 52(1):91–101, 2018. URL https://doi.org/10.1145/3274784.3274789.