ECIR: Checking Methodology

1. Quality metrics

MRR is widely used but difficult to interpret as interval scale, which would be a requirement for computing the arithmetic averages. More importantly, t-tests on MRR have shown that 30% of the tests change their outcome when MRR is replaced by a proper scale (Ferrante et al 2021). A better choice is the means of the first relevant rank (MFR, see 2.1 in Fuhr 2017). MFR has the property of an interval scale by assigning a value of k+1 when no relevant document is found among the first k ranks (like in MRR).

MAP suffers from the same problem as it is a generalization of MRR from the first to all relevant documents. Experiments have shown that it is about half as bad as MRR in terms of t-tests, but worse than a better choice of measures like e.g. RBP and (n)DCG.

2. Relative improvements

Relative improvements of arithmetic means have no meaningful statistical interpretation (see 2.4 in Fuhr 2017). Reporting absolute improvements (e.g., +0.05 nDCG rather than +3%) helps, but a better choice is to report effect sizes as these also take variance into account.

3. Significance testing

Multiple testing on the same dataset leads to valid results only in case a correction of the significance level is performed (using Bonferroni or Bonferroni-Holm, see 2.7 in Fuhr 2017). For comparing all pairs of runs, Tukey's HSD test can be applied. Multiple testing also happens when the actual pairs of runs to be tested are chosen after the experimental results are known: the hypotheses to be tested have to be formulated before the experiments, otherwise (conceptually) all possible pairs are tested.

Testing on re-used test collections is not permissible: First, the total number of tests already performed on this collection should be considered. More important is the sequential learning problem – violating the requirement that we must formulate the hypotheses before we learn about any experimental outcomes from the collection. As a good substitute, effect-size should be regarded.

4. Proofs

Besides the extremely misleading “our experiments prove”, any usage of the word “prove” in a scientific text should be restricted to the cases where universally valid statements are formally proven (the case of existentially quantified statements plays no role in IR).

How to spot these mistakes?

1. *In the result tables, are the claims relying (only) on the use of the problematic measures?*
2. *Do they list relative changes of arithmetic means?*
3. *Do they contain any significance tests? If yes*
    a. *Is this a re-used test collection?*
    b. *Is there more than one significance test (either actually performed or only conceptually, by picking the pairs after the results are known)?*
    c. *If yes, was Bonferroni correction or Tukey's HSD used?*
4. *Does the text, and especially abstract or conclusion, contain the term “prove”?*

Marco Ferrante, Nicola Ferro, Norbert Fuhr (2021). Towards Meaningful Statements in IR Evaluation. Mapping Evaluation Measures to Interval Scales. *IEEE Access.* https://doi.org/10.1109/ACCESS.2021.3116857

Norbert Fuhr (2017). Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum 51(3)*. https://doi.org/10.1145/3190580.3190586

Tetsuya Sakai (2020). On Fuhr's guideline for IR evaluation. *SIGIR Forum 54 (1).* https://doi.org/10.1145/3451964.3451976