On Fuhr's Guideline for IR Evaluation

Tetsuya Sakai
Waseda University, Tokyo, Japan
tetsuya@waseda.jp

Abstract

In the December 2017 issue of SIGIR Forum, Fuhr presented ten "Thou Shalt Not"s (i.e., warnings against bad practices) for IR experimenters. While his article provides a lot of good materials for discussion, the objective of the present article is to argue that not all of his recommendations should be considered as absolute truths: researchers should be aware that there are other views; conference programme chairs and journal editors should be very careful when providing a guideline for evaluation practices.

1 Introduction

I was rather surprised to see Fuhr's December 2017 SIGIR Forum article [Fuhr, 2017] mentioned in the ECIR 2019 CFP as follows:[1]

> Additionally authors should consider the ECIR guidelines when crafting their paper and see Fuhr's guide to avoid common IR evaluation mistakes.

The ECIR 2020 CFP looked similar:[2]

> Authors should consult ECIR's paper guidelines and Fuhr's guide to avoid common IR evaluation mistakes, for the preparation of their papers.

I was surprised, because the article provides recommendations from *one* researcher (okay, one *great* researcher who won a SIGIR Award!), not from some editorial board or working group. When I read the article, I mentally applauded some of his recommendations but not the others; I was happy just to agree to disagree then. Or is everyone expected to agree?

In January 2018, Fuhr tweeted that SIGIR 2018 authors and reviewers should read his article.[3], and the official SIGIR 2018 twitter account (`@sigir2018`) retweeted it. I asked the SIGIR 2018 chairs whether that means they officially agree with all of his views; I was assured that "retweeting does not imply endorsement."[4] Then, in January 2020, Fuhr posted another tweet apparently criticising the SIGIR 2020 chairs for not taking heed of his guideline.[5]

[1] http://ecir2019.org/call-for-papers/

[2] https://ecir2020.org/call-for-full-papers/

[3] https://twitter.com/NorbertFuhr/status/955467819671605248

[4] https://twitter.com/kevynct/status/955591413667958784

[5] https://twitter.com/NorbertFuhr/status/1212653760658497536

The objective of the present article is to argue that not all of Fuhr's recommendations should be considered as absolute truths. The remainder of this article examines Fuhr's "Thou Shalt Not"s one by one, with a heavy emphasis on those that I disagree with.

2 Thou Shalt Not?

2.1 Thou Shalt Not Compute MRR nor ERR?

I for one disagree. In Fuhr's article, it is argued: "*one cannot compute the mean for an ordinal scale!*" Yes, Stevens said something like that in 1946 [Stevens, 1946] but other researchers have disagreed. Sauro and Lewis [Sauro and Lewis, 2016, pp.250-254] provide arguments from both camps; I mentioned this book in a tweet I posted in January 2018.[6] However, whether I agree or disagree is not really important; I am but one researcher. What is important is to be aware of the fact that there is a controversy. For example, multipoint scale ratings from user studies are ordinal, but they *are* often averaged, and even tested with parametric tests such as the t-test [Sauro and Lewis, 2016], although the results should then be interpreted with care. "*One cannot compute the mean for an ordinal scale*" is by no means a universal truth, and researchers should be aware of this.

The article also discusses Systems A and B being evaluated with MRR over three queries: A has its first relevant document at Ranks 1,2, and 4 for the three queries, respectively, while B has its first relvant document at Rank 2 for all three queries. It is argued there that it is strange that A is rated higher than B (where the MRR for A is $(1/1 + 1/2 + 1/4)/3 = 0.58$ and that for B is $(1/2 + 1/2 + 1/2)/3 = 0.50$). I do not find this particularly strange; MRR rewards A heavily for the first query.

Perhaps more importantly: it is also not clear to me whether RR really cannot be considered as an interval-scale measure. [7] Fuhr's article points out that if Systems X, Y, Z have their first relevant documents at ranks $1, 2, \infty$, respectively, RR says that the delta between X and Y and that between Y and Z are both 0.5. However, this does not immediately mean that RR is not on an interval scale; it means that RR *considers* the two differences to be of equal magnitude according to its user model. Consider Systems X', Y', Z' that each retrieve exactly one relevant document (with a gain value of 1) at ranks $15, 31, 100$, respectively. According to Discounted Cumulative Gain (DCG) [Järvelin and Kekäläinen, 2002] that uses $1/\log_2(r+1)$ for discouting at rank r, the difference between X' and Y' is $1/\log_2(16) - 1/\log_2(32) = 0.25 - 0.20 = 0.05$; that between Y' and Z' is also $1/\log_2(32) - 1/\log_2(101) \approx 0.20 - 0.15 = 0.05$. How is this situation any different from the above concerning RR with X, Y, Z?

I find both Reciprocal Rank (RR) and Expected Reciprocal Rank (ERR) [Chapelle et al., 2009] useful; especially the latter. These measures depend entirely/heavily on the first relevant document on the ranked list and are useful for evaluating navigational searches. They complement informational search measures such as normalised DCG (nDCG) [Järvelin and Kekäläinen, 2002].

Robertson [2008] pointed out that RR is an instance of the Normalised Cumulative Precision (NCP) family; an NCP is a combination of a *probability distribution* over document ranks rep-

[6] https://twitter.com/tetsuyasakai/status/958222185084076032

[7] This point was raised to me by Stephen Robertson in a private discussion.

resenting how many users from the population will stop scanning the ranked list at each rank r, and the *utility function* (namely, precision) for each user group at r. RR assumes that all users abandon the ranked list at r_1, the rank of the first relevant document found; for this user group, the utility (i.e., precision) is clearly $1/r_1$. On the other hand, as acknowledged in Chapelle et al. [2009], ERR is an instance of the Normalised Cumulative Utility (NCU) family [Sakai and Robertson, 2008, Sakai, 2014a, Sakai and Zeng, 2019] which generalises NCP; the abandoning probability distribution of ERR is arguably more realistic than those of other existing measures, in that the relevance of the documents from rank 1 to rank $(r - 1)$ influences how many users will reach as far as rank r.

At any rate, I do not see why researchers should be *forbidden* to use RR and ERR.

2.2 Thou Shalt Not Use MAP?

In Fuhr's article, it is also argued that Average Precision (AP) should not be used, on the grounds that its stopping probability distribution (See Section 2.1) is not realistic: according to AP's user model as explained by Robertson [2008], there is a user population and an infinite ranked list in which all relevant documents are ranked, and the users are equally likely to stop at any of these relevant documents. The article recommends Rank Biased Precision (RBP) [Moffat and Zobel, 2008] and DCG [Järvelin and Kekäläinen, 2002].

Zobel, Moffat, and Park [Zobel et al., 2009] also argue against AP for a different reason, namely, that AP relies on the total number of (known) relevant documents; they also argue against nDCG because the ideal list of nDCG also requires all (known) relevant documents. RBP is their preferred measure. Personally, I have no problem with normalisation based on all known relevant documents if we already know these relevant documents, but this is not the place to debate it; if one objects to it, score standardisation is available as an alternative [Sakai, 2016, Urbano et al., 2019, Webber et al., 2008a].

Compared to the aforementioned user model of ERR, I find that of RBP less realistic as it assumes that the probability of transitioning from rank i to rank $(i + 1)$ is a constant (e.g. $p = 0.85$): the relevance of the document at rank i does not affect the users. I also pointed out that the maximum possible value of RBP depends heavily on the value of p [Sakai, 2014a]; in Fuhr's article, it is argued that this is a minor issue. Possibly.[8]

At any rate, AP relies on a *model* of the users' abandoning probability distribution, where users are spread evenly across all relevant documents. RBP relies on another *model*, which assumes that p% of users will continue on from rank i to rank $(i + 1)$ regardless of document relevance. "*All models are wrong but some are useful*" [Box, 1979]. So which ones are useful? Let us look at some data.

Table 1 shows how different IR evaluation measures agree with the users' SERP preferences. The users' SERP preferences are from Sakai and Zeng [2019]: from the NTCIR-9 INTENT task [Sakai and Song, 2013], 1,127 topic-SERP-SERP triplets were extracted, and each triplet was judged independently by 15 assessors. Each SERP contains exactly 10 URLs. The question for the assessors was "*Overall, which SERP is more relevant to the query?*" and each assessor accordingly chose from "LEFT is better," "RIGHT is bettter," and "Both are equally irrelevant."

[8] Score standardisation can certainly solve the problem, but the fact is, it is seldom used.

Table 1: Mean Agreement Rate (MAR) over 1,127 topic-SERP-SERP triplets, each with SERP preferences from 15 assessors. Based on a randomised Tukey HSD test [Carterette, 2012, Sakai, 2018] with $B = 10,000$ trials, p-values of all statistically significant differences obtained at $\alpha = 0.05$ are also shown. For example, a p-value shown in the ">Prec" column means that a measure statistically significantly outperforms Prec. The two-way ANOVA residual variance obtained from the 1,127×10 triplet-by-measure matrix is $V_{E2} = 0.0182$; for each pair of measures, the effect size can be obtained from this table as $\Delta MAR/\sqrt{V_{E2}}$.

		p-value (< 0.05)		
	MAR	>Prec	>AP	>ERR
nDCG	0.7424	0	0	0
iRBU ($p = 0.99$)	0.7409	0	0	0
RBP ($p = 0.85$)	0.7354	0	0	0
RBP ($p = 0.99$)	0.7337	0.0001	0.0001	0
Q	0.7317	0.0006	0.0005	0
iRBU ($p = 0.85$)	0.7266	0.0088	0.0077	0
EBR	0.7248	0.0273	0.0240	0
Prec	0.7054			0.0018
AP	0.7052			0.0020
ERR	0.6814			

More details about the SERP preferences can be found in the aforementioned paper. The evaluation measures were computed based on the adhoc qrels from Sakai and Zeng [2019]; as it features 5-point graded relevance assessments, the gain values for graded relevance measures were set to 15,7,3,1 for 4,3,2,1-relevant documents, respectively.

Given a topic-SERP-SERP triplet, an evaluation measure says either "LEFT is better," "RIGHT is better," or "They are equally effective." The agreement rate (AR) is defined as the proportion of users' preference labels that agree with the measure. That is, AR means "the measure's SERP preference agrees with what percentage of users?" The Mean AR (MAR) is the ARs averaged over the entire set of triplets. Table 1 ranks the measures by MAR[9].

First, a quick note to say that although ERR is ranked at the bottom of these tables, this is as expected. As I mentioned earlier, ERR is suitable for navigational searches, but recall that the user SERP preferences were collected based on the *overall relevance* of each SERP. The results do *not* imply that ERR is a bad measure.

Second, it can be observed that AP (not reported in Sakai and Zeng [2019]) performs poorly compared to graded relevance measures shown in the tables; it does not even outperform Precision, a set retrieval measure. Moreover, a randomised Tukey HSD test [Carterette, 2012, Sakai, 2018] shows that AP statistically significantly underperforms nDCG, iRBU, RBP, Q, and EBR (See Sakai and Zeng [2019] for definitions of these measures) in terms of MAR. The results strongly suggest that graded relevance measures should be used if graded relevance assessments are available. That said, however, we *cannot* conclude from these results that AP should be banished from the grounds: from Table 1, when it comes to judging which of the given two SERPs is more

[9] This evaluation method is different from that used in Sakai and Zeng [2019]. In that paper, the 15 preferences were consolidated to form a single gold preference, but this loses the information regarding how much the assessors (dis)agreed for each triplet.

relevant, AP agrees with about 70% of users on average. That's not so bad![10]

In the early 1990s, AP was identified as a reliable measure for evaluating participating systems at TREC, thanks to the efforts by Buckley and Voorhees [Buckley and Voorhees, 2005] along with others. AP *was* useful for identifying effective term weighting schemes (such as BM25 [Robertson et al., 1995]) and pseudo relevance feedback schemes. By the 2000s, it was recognised that the statistical stability of AP made it suitable as a target measure for training IR systems even when the final test measure is something less stable such as Precision at 10 [Robertson, 2006, Webber et al., 2008b]. Thank you AP!

In summary, although AP is probably not the best choice when graded relevance assessments are available, I do not see why it should be banned altogether.

2.3 Thou Shalt Not Overstate the Precision of Your Results?

Agreed.

2.4 Thou Shalt Not Compute Relative Improvements of Arithmetic Means?

I do not know if this should be forbidden. But I prefer reporting effect sizes, more specifically, standardised mean differences such as Hedge's g and Glass's Δ [Sakai, 2018, pp.85-88] to reporting relative improvements. More on this in Section 2.8.

2.5 Thou Shalt Not Apply the Simple Holdout Method?

I agree to the extent that researchers should be aware that different training/validation/test splits may yield different results and that trying different splits is generally recommended.

2.6 Thou Shalt Not Formulate Hypotheses after the Experiment?

Agreed.

2.7 Thou Shalt Not Test Multiple Hypotheses without Correction?

Agreed. Note that I used a randomised Tukey HSD test in Section 2.2. See also Section 2.8.

2.8 Thou Shalt Not Ignore Effect sizes?

Agreed. Note that Table 1 enables the reader to work out effect sizes (standardised mean differences). For example, the effect size that corresponds to the difference between nDCG and ERR is $(0.7424 - 0.6814)/\sqrt{0.0182} = 0.0610/0.1349 = 0.4526$. That is, the two measures are about half a (common) standard deviation apart.

[10] Note that this evaluation method is insensitive to normalisation since each measure applies the same normalisation factor to both SERPs for every SERP pair. That is, AP is equivalent to the sum of precisions at relevant ranks, and nDCG is equivalent to DCG. Again, this paper's aim is not to justify normalisation.

In fact, I discussed multiple comparisons and effect sizes in the June 2014 issue of SIGIR Forum [Sakai, 2014b]. See also Sakai [2018] for more discussions on multiple comparison procedures and effect sizes.

2.9 Thou Shalt Not Forget about Reproducibility?

Agreed. Working on it [Clancy et al., 2019, Ferro et al., 2019, Sakai et al., 2019]...

2.10 Thou Shalt Not Claim Proof by Experimentation?

Agreed.

3 Summary

I hope I have convinced at least some researchers in the IR community that not all of the above "Ten Commandments" are completely unarguable. Researchers should be aware of different views; conference programme chairs and journal editors should be very careful when providing a guideline for evaluation practices. Let's agree to disagree.

Acknowledgements

I thank Ben Carterette, Stephen Robertson, and Ian Soboroff for their positive feedback on a draft of this paper: I also thank Norbert Fuhr for this opportunity for a good open discussion on important topics.

References

Norbert Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, 2017.

S. S. Stevens. On the theory of scales of measurement. *Science, New Series*, 103(2684):677–680, 1946.

Jeff Sauro and James R. Lewis. *Quantifying the User Experience: Practical Statistics for User Research (2nd Edition)*. Morgan Kafmann, 2016.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of ACM CIKM 2009*, pages 621–630, 2009.

Stephen Robertson. A new interpretation of average precision. In *Proceedings of ACM SIGIR 2008*, pages 689–690, 2008.

Tetsuya Sakai and Stephen Robertson. Modelling a user population for designing information retrieval metrics. In *Proceedings of EVIA 2008*, pages 30–41, 2008.

Tetsuya Sakai. Metrics, statistics, tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*, pages 116–163, 2014a.

Tetsuya Sakai and Zhaohao Zeng. Which diversity evaluation measures are "good"? In *Proceedings of ACM SIGIR 2019*, pages 595–604, 2019.

Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM TOIS*, 27(1), 2008.

Justin Zobel, Alistair Moffat, and Laurence A.F. Park. Against recall: Is it persistence, cardinality, density, coverage, or totality? *SIGIR Forum*, 43(1):3–8, 2009.

Tetsuya Sakai. A simple and effective approach to score standardisation. In *Proceedings of ACM ICTIR 2016*, pages 95–104, 2016.

Julián Urbano, Harlley Lima, and Alan Hanjalic. A new perspective on score standardization. In *Proceedings of ACM SIGIR 2019*, pages 1061–1064, 2019.

William Webber, Alistair Moffat, and Justin Zobel. Score standardization for inter-collection comparison of retrieval systems. In *Proceedings of ACM SIGIR 2008*, pages 51–58, 2008a.

G.E.P. Box. Robustness in the strategy of scientific model building. In Robert L. Launer and Graham N. Wilkinson, editors, *Robustness in Statistics*, pages 201–236. Academic Press, 1979.

Tetsuya Sakai and Ruihua Song. Diversified search evaluation: Lessons from the NTCIR-9 INTENT task. *Information Retrieval*, 16(4):504–529, 2013.

Ben Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, 30(1), 2012.

Tetsuya Sakai. *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. Springer, 2018.

Chris Buckley and Ellen M. Voorhees. Retrieval system evaluation. In Ellen M. Voorhees and Donna K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3, pages 53–75. The MIT Press, 2005.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. 1995.

Stephen Robertson. On GMAP – and other transformations. In *Proceedings of ACM CIKM 2006*, pages 78–83, 2006.

William Webber, Alistair Moffat, Justin Zobel, and Tetsuya Sakai. Precision-at-ten considered redundant. In *Proceedings of ACM SIGIR 2008*, pages 695–696, 2008b.

Tetsuya Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, 2014b.

Ryan Clancy, Nicola Ferro, Claudia Hauff, Jimmy Lin, Tetsuya Sakai, and Ze Zhong Wu. The SIGIR 2019 open-source IR replicability challenge (OSIRRC 2019). In *Proceedings of ACM SIGIR 2019*, pages 1432–1434, 2019.

Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. CENTRE@CLEF2019: Sequel in the systematic reproducibility realm. In *Proceedings of CLEF 2019 (LNCS 11696)*, pages 287–300, 2019.

Tetsuya Sakai, Nicola Ferro, Ian Soboroff, Zhaohao Zeng, Peng Xiao, and Maria Maistro. Overview of the NTCIR-14 CENTRE task. In *Proceedings of NTCIR-14*, pages 494–509, 2019.