On A Few Responsibilities of (IR) Researchers (Fairness, Awareness, and Sustainability): A keynote at ECIR 2023

Tetsuya Sakai
Waseda University
Japan
tetsuyasakai@acm.org

Abstract

On the ECIR 2023 Virtual Day, I talked about three things: fairness, awareness, and sustainability. This extended abstract summarises the 50-minute talk and provides related resources.

Date: 31 March 2023.

1 Introduction

On the ECIR 2023 Virtual Day (March 31, 2023), I talked about three things: fairness, awareness, and sustainability [Sakai, 2023]. The slide deck is available at https://waseda.box.com/ecir2023keynote. My talk featured three different (but related) topics; the following three sections cover them in turn.

2 Fairness

Bias breeds bias. Search engine and recommender companies should ensure fair exposure to the items being ranked or their stakeholders. In this context, I advertised the ongoing NTCIR-17 FairWeb-1 task,[1] an English web search task that evaluates search engines in terms of both relevance and group fairness. The task adopts the GFR (Group Fairness and Relevance) evaluation framework [Sakai et al., 2023] (See Figure 1), which can handle ordinal groups as well as intersectional group fairness.

Table 1 compares the GFR evaluation framework with AWRF (Attention-Weighted Rank Fairness), the group-fair ranking measure that was used in the TREC 2021 and 2022 Fair Ranking tracks (Task 1, which dealt with single rankings) [Ekstrand et al., 2022, 2023]. Here are some additional remarks on the table.

[1] http://sakailab.com/fairweb1/

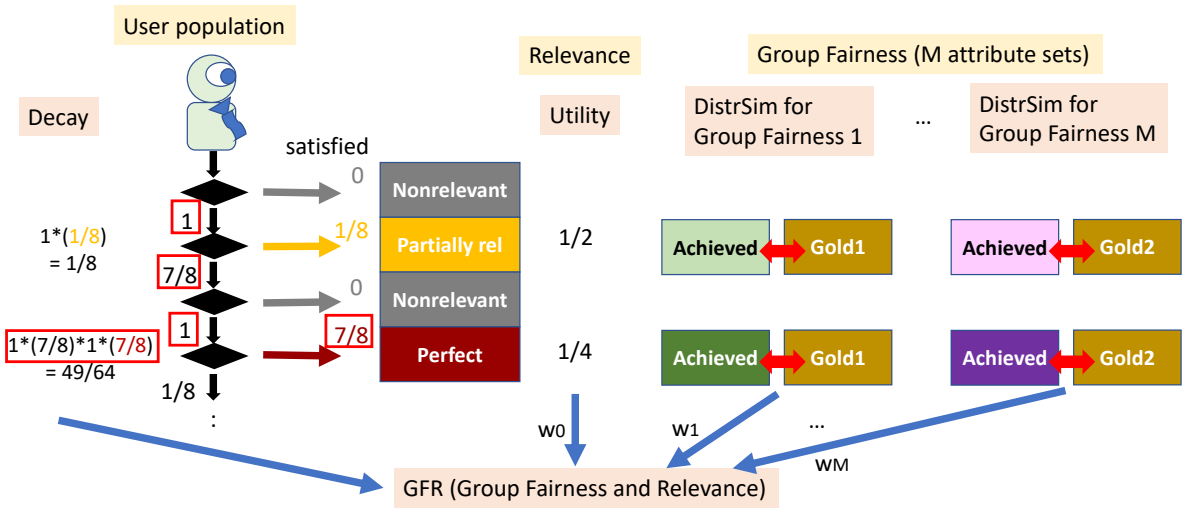

Figure 1. The GFR (Group Fairness and Relevance) evaluation framework, duplicated from Sakai et al. [2023]. In the figure, it is assumed that the Decay and Utility components are based on ERR (Expected Reciprocal Rank) [Chapelle et al., 2009; Sakai, 2014b]. DistrSim stands for Distribution Similarity. Here, the highest relevance grade is assumed to be 3 (Perfect), and its gain value is set to $(2^3 - 1)/2^3 = 7/8$.

(a) AWRF uses an nDCG-based attention decay to compute a single achieved distribution over groups for a SERP; hence it assumes that user attention is relevance-independent. In contrast, when relevance assessments are available, GFR adopts an ERR-based user model from the Normalised Cumulative Utility (NCU) framework [Sakai and Robertson, 2008].

(b) Unlike AWRF, GFR chooses a distribution similarity function (for comparing the achieved and target distributions) depending on whether the groups are ordinal or not. Thus, for example, when researchers are grouped based on their h-indexes, GFR can take into account the fact that the gap between Groups 1 and 3 is greater than one between Groups 1 and 2.

(c) AWRF obtains a single achieved distribution from a SERP, where attention decay is reflected within the distribution; this is compared with the gold distribution. In contrast, given relevance assessments, GFR obtains an achieved distribution for each search engine user group that is assumed to abandon the ranked list at a particular relevant document (e.g., the users at ranks 2 and 4 in Figure 1), following the aforementioned NCU framework.

(d) Following prior art, AWRF takes a Cartesian product of multiple attribute sets; however, it should be noted that when ordinal groups are involved, this operation can break their ordinal nature [Sakai et al., 2023]. In contrast, GFR computes group fairness score for each attribute set and *then* combines the scores.

(e) While AWRF and nDCG were multiplied at TREC, the default formulation of GFR is defined as the *expected user experience*, where the user experience for a user group at a particular rank is given as the sum of *Utility* and distribution similarities, as shown in Figure 1.

Table 1. Comparison of AWRF (from the TREC Fair Ranking Tracks) and GFR (used in the NTCIR-17 FairWeb-1 task).

	AWRF	GFR
(a) Decay	nDCG-based (relevance-unaware)	ERR-based (RBP-based if relevance assessments are unavailable)
(b) Divergence	JSD	JSD for nominal groups NMD and RNOD for ordinal groups
(c) When achieved and gold distributions are compared	once per SERP	for every rank with a relevant document (ERR-based) or for every rank (RBP-based)
(d) How intersectional group fairness is handled	takes a Cartesian product of attribute sets	computes a fairness score for each attribute set and then the scores are combined
(e) How relevance and fairness are combined	AWRF and nDCG are multiplied	User experience quantified as Utility + distribution similarities; GFR is then "expected user experience"

Table 2. Attribute sets considered in the NTCIR-17 FairWeb-1 task.

	attribute set with ordinal groups	attribute set with nominal groups
R (Researcher) topics	HINDEX (4 groups based on h-indexes)	GENDER (3 groups based on whether "he" or "she" occurs in the researcher biography)
M (Movie) topics	RATINGS (4 groups based on IMDb ratings)	ORIGIN (8 groups based on "country of origin" from IMDb)
Y (YouTube) topics	SUBSCS (4 groups based on #subscribers of the YouTube creator)	-

Table 2 shows the actual attribute sets that are used in the NTCIR-17 FairWeb-1 task. Nine research groups have registered to participate in the task; it will be concluded at the NTCIR-17 conference in December 2023.

3 Awareness

In the second part of my talk, I discussed the ongoing dispute about whether it is okay to average existing IR measures (Reciprocal Rank in particular) across topics.

My main message was not about whether one camp is right and the other is wrong: it was: *conference chairs and journal editors should not declare just one viewpoint to be truth and enforce that on the entire research community* [Sakai, 2020] and that *researchers should be aware of multiple viewpoints and should make their own informed decisions.*

To help IR researchers make informed decisions regarding the above particular controversy, here I provide some relevance references. In the December 2017 edition of SIGIR Forum, Fuhr

[2017] made several claims,[2] such as "*Thou shalt not compute MRR nor ERR*" and "*one cannot compute the mean for an ordinal scale!*" The ECIR 2019 and 2020 PC chairs apparently accepted his recommendations, for their CFPs included a link to Fuhr's article. I was worried about this trend; hence I published my own SIGIR Forum article in June 2020 [Sakai, 2020]. As a result, the ECIR 2021 CFP included a link to the Fuhr article, *and* a link to mine. I thought: *providing different viewpoints is great!*

In 2021, Ferrante, Ferro, and Fuhr published an interesting refereed article [Ferrante et al., 2021], whose views naturally have an overlap with the aforementioned article of Fuhr. In my understanding, their concerns include the following, and I called them Concerns A-C in my keynote:

Concern A IR measures like RR, ERR, and nDCG are not equi-spaced (over SERP states) and therefore they are not interval-scale;

Concern B Ordinal measures (that are not interval) should not be averaged;

Concern C Averaging across topics (with different recall levels and different run lengths) is not appropriate either.

Here are my brief *personal* remarks on the above. **Concern A**: I am still not convinced why IR measures should be equi-spaced over possible SERP states. If I view IR measure scores as rough approximations of a subset of real-world SERP utility scores on an interval scale, what's wrong with them, besides them being rough approximations? Existing IR measures *are* useful; they have been useful for over half a century. **Concern B**: researchers should be aware that not everyone agrees with Stevens [1946], as I pointed out in Sakai [2020]. Some applied statisticians average ordinal numbers like multipoint rating scales and even conduct t-tests etc. [Sauro and Lewis, 2016, p.253] **Concern C**: because of how I view the IR measures (see above), I do not see why averaging across topics is flawed just because different topics can take different possible values within the 0-1 range (due to different recall bases etc.). Again, averaged IR measures have been useful for decades, although per-topic analysis is probably more important in evaluation.

In 2022, Moffat [2022] directly responded to Ferrante et al. [2021]: he also disagreed with them, saying: "*all IR effectiveness metrics can be considered to be interval scale measurements, provided only that the mapping from SERP categories to numeric scores has a real-world basis* [$\cdots$]." In the same year, Ferrante, Ferro, and Fuhr responded to Moffat with an arxiv paper [Ferrante et al., 2022]. Other researchers have also commented on this dispute [Craswell et al., 2021; Lin et al., 2021]. These are all very interesting discussions. Please have a look!

Let me reiterate: given multiple conflicting viewpoints within the research community (whatever the topic of the dispute is), enforcing just one of them on the community is a terrible thing to do.

4 Sustainability

Paul Clough and I are serving as the first sustainability chairs of SIGIR (term of service: 2023-2025). We have been gathering best practices from research communities such as SIGPLAN,

[2] I agree with Fuhr's recommendations about multiple comparison procedures and effect sizes: note that my June 2014 SIGIR Forum article already discussed them [Sakai, 2014a].

SIGCHI, and NLP. In my talk, I briefly mentioned the importance of *green research* [Strubell et al., 2019; Schwartz et al., 2020; Bender et al., 2021; Scells et al., 2022] (i.e., those that aim at preventing negative impacts on our planet, global warming [Thunberg, 2022] in particular) and then discussed what the IR community should do to have *green conferences*.

I argued that IR researchers with different roles within the community should take the following actions at least, to make conferences greener.

Conference bidders The sustainability chairs will provide a checklist for bidders so that they can consider co-locating conferences, choose venues that actually care about sustainability, include carbon offset in the budget, and plan a green conference that minimises carbon emissions, waste, etc.

Conference organisers For successful bidders, the sustainability chairs will provide another checklist to implement the planned green conference. Accommodating both in-person and remote conference attendance is a minimum requirement: conferences should not force long-distance air travels on authors.

Program Committees PC chairs may want to consider incorporating rewarding mechanisms for green research/evaluation, for example, as a review criterion, and/or a sustainability award.

Authors Just like an Ethical Consideration section in an NLP paper, it may be useful to have a Sustainability Consideration section in a paper that does not count towards the page limit.[3]

All of us We (especially senior researchers who have already had a lot of fun travelling all over the world to attend conferences) may want to consider reducing long-distance air travels.[4] Besides co-locating conferences, having distributed regional conferences instead of one mega-conference may also reduce total carbon emissions.

The mention of distributed, regional conferences allowed me to smoothly transition to my advertisement of the new SIGIR-AP (Asia/Pacific) conference series; Xuanjing Huang, Justin Zobel, and myself are serving as PC co-chairs for 2023,[5] with Yiqun Liu and Alistair Moffat as the general chairs as well as steering committee chairs. When I was PC co-chair for SIGIR 2021, I checked the number of authors for accepted full papers and found that there were as many as 471 authors from P.R. China (followed not so closely by the United States with 132 authors). After the launch of SIGIR-AP, will many of these Chinese authors, as well as others based in Asia, prefer SIGIR-AP (which will always be held in Asia/Pacific) over the "global" SIGIR (when it is not held in Asia) so that total long-distance air travels will be reduced? Or will SIGIR-AP just introduce a community divide? We IR researchers like experiments: let's experiment.

[3] In [Benotti and Blackburn, 2022, p.4512], there is an example of an ACL Ethical Consideration section that discusses *environmental costs*. However, it may be better to have a separate sustainability section as a requirement.

[4] Of course, meeting people face-to-face at a conference is extremely important especially for junior researchers for networking, finding job/collaboration opportunities, etc.

[5] http://www.sigir-ap.org/sigir-ap-2023/

Acknowledgements

I thank Professor Cathal Gurrin and the other ECIR 2023 organisers for giving me the opportunity to give this talk. I thank the NTCIR-17 FairWeb-1 task organisers (Sijie Tao, Nuo Chen, Zhumin Chu, Nicola Ferro, Maria Maistro, Ian Soboroff, Hiromi Arai) for running the task with me, and the NTCIR-17 chairs for their support. Last but not least, I thank all the researchers named in this extended abstract.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of ACM FAccT 2021*, pages 610–623, 2021.

Luciana Benotti and Patrick Blackburn. Ethics consideration sections in natural language processing papers. In *Proceedings of EMNLP 2022*, pages 4509–4516, 2022.

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.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. MS MARCO: Benchmarking ranking models in the large-data regime. In *Proceedings of ACM SIGIR 2021*, page 1566–1576, 2021.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. Overview of the TREC 2021 fair ranking track. In *Proceedings of TREC 2021*, 2022.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. Overview of the TREC 2022 fair ranking track. In *Proceedings of TREC 2022*, 2023.

Marco Ferrante, Nicola Ferro, and Norbert Fuhr. Towards meaningful statements in ir evaluation: Mapping evaluation measures to interval scales. *IEEE Access*, 9:136182–136216, 2021.

Marco Ferrante, Nicola Ferro, and Norbert Fuhr. Response to moffat's comment on "towards meaningful statements in ir evaluation: Mapping evaluation measures to interval scales", 2022.

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

Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Significant improvements over the state of the art? a case study of the MS MARCO document ranking leaderboard. In *Proceedings of ACM SIGIR 2021*, pages 2283–2287, 2021.

Alistair Moffat. Batch evaluation metrics in information retrieval: Measures, scales, and meaning. *IEEE Access*, 10:105564–105577, 2022.

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

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

Tetsuya Sakai. On Fuhr's guideline for IR evaluation. *SIGIR Forum*, 54(1):1–8, 2020.

Tetsuya Sakai. On a few responsibilities of (IR) researchers: Fairness, awareness, and sustainability (keynote). In *Proceedings of ECIR 2023, Volume I (LNCS 13980)*, page xxiii, 2023.

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

Tetsuya Sakai, Jin Young Kim, and Inho Kang. A versatile framework for evaluating ranked lists in terms of group fairness and relevance. *ACM TOIS*, 2023.

Jeff Sauro and James R. Lewis. *Quantifying the User Experience (2nd Edition)*. Morgan Kaufmann, 2016.

Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. Reduce, reuse, recycle: Green information retrieval research. In *Proceedings of ACM SIGIR 2022*, pages 2825–2837, 2022.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020.

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

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of ACL 2019*, pages 3645–3650, 2019.

Greta Thunberg. *The Climate Book*. Penguin Press, 2022.