Users: Can’t Work With Them, Can’t Work Without Them?

Alistair Moffat
The University of Melbourne
Australia
ammoffat@unimelb.edu.au

ABSTRACT

If we could design the ideal IR “effectiveness” experiment (as distinct from an IR “efficiency” experiment), what would it look like? It would probably be a lab-based observational study [3] involving multiple search systems masked behind a uniform interface, and with hundreds (or thousands) of users each progressing some “real” search activity they were interested in. And we’d plan to (non-intrusively, somehow) capture per-snippet, per-document, per-SERP, and per-session annotations and satisfaction responses. The collected data could then be compared against a range of measured “task completion quality” indicators, and also against search effectiveness metric scores computed from the elements contained in the SERPs that were served by the systems.

That’s a tremendously big ask! So we often use offline evaluation techniques instead, employing test collections, static qrels sets, and effectiveness metrics [6]. We *abstract* the user into a deterministic evaluation script, supposing for pragmatic reasons that we know what query they would issue, and at the same time assuming that we can apply an effectiveness metric to calculate how much usefulness (or satisfaction) they will derive from any given SERP. The great advantage of this approach is that aside from the process of collecting the qrels, it is free of the need for users, meaning that it is repeatable. Indeed, we often do repeat, iterating to set parameters (and to rectify programming errors). Then, once metric scores have been computed, we carry out one or more paired statistical tests and draw conclusions as to relative system effectiveness.

But that process of abstraction also represents a compromise that weakens the validity of the measurements that are obtained. This presentation first describes user-motivated models of searcher behavior that suggest ways of developing scrutable effectiveness metrics [4, 5, 7, 8]. It then explores the ramifications of user-derived query variations [1, 2, 4] in terms of offline experimental protocols. Both of these directions can be thought of as seeking to better capture what *would* be measured if we had the resources needed (including suitable numbers of demographically diverse subjects) for a large-scale lab-based observational study out of which we *could* seek to compare system effectiveness. Hence the question of the title: to what extent can we carry out user-inspired effectiveness studies without needing to recruit any users?

CCS CONCEPTS

• **Information systems** → **Task models**; **Retrieval effectiveness**; **Presentation of retrieval results**.

KEYWORDS

User browsing model; effectiveness metric; offline evaluation

BIOGRAPHY

Alistair Moffat has been a faculty member at the University of Melbourne since 1986, and been active in the area of information retrieval since 1990. During that period he has coauthored two books (*Managing Gigabytes*, 1994 and 1999; *Compression and Coding Algorithms*, 2002) and more than 250 research papers on topics spanning text and integer compression, index representations and structures, processing algorithms for top-k queries, and mechanisms for retrieval effectiveness evaluation; a body of work for which he was inducted into the SIGIR Academy in 2021. During his four decades as an academic Alistair has also introduced more than 25,000 undergraduates to the joys and frustrations of programming.

ACKNOWLEDGMENTS

This thread of work has been supported by the Australian Research Council over an extended period, including via grants DP0772078, DP110101934, LP150100252, and DP190101113. Funding has also been received from Seek.com. A large number of colleagues and students have helped obtain and then spend that money, including Alfan Wicaksono, Andrew Turpin, Falk Scholer, Guido Zuccon, Joel Mackenzie, Justin Zobel, Lawrence Cavedon, Leif Azzopardi, Mark Sanderson, Matthias Petri, Melika Ebrahim, Nick Craswell, Oleg Zendel, Paul Thomas, Peter Bailey, Rodger Benham, Shane Culpepper, William Webber, Xiaolu Lu, and Ziying Yang.

REFERENCES

[1] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV100: A test collection with query variability. In *Proc. SIGIR*, pages 725–728, 2016. Public data: http://dx.doi.org/10.4225/49/5726E597B8376.

[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. Retrieval consistency in the presence of query variations. In *Proc. SIGIR*, pages 395–404, 2017.

[3] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. & Trends in IR*, 3(1-2):1–224, 2009.

[4] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Inf. Sys.*, 35(3):24:1–24:38, 2017.

[5] A. Moffat, J. Mackenzie, P. Thomas, and L. Azzopardi. A flexible framework for offline effectiveness metrics. In *Proc. SIGIR*, 2022.

[6] M. Sanderson. Test collection based evaluation of information retrieval systems. *Found. & Trends in IR*, 4(4):247–375, 2010.

[7] A. F. Wicaksono and A. Moffat. Metrics, user models, and satisfaction. In *Proc. WSDM*, pages 654–662, Feb. 2020.

[8] A. F. Wicaksono and A. Moffat. Modeling search and session effectiveness. *Inf. Proc. & Man.*, 58(4):102601, 2021.

SIGIR ’22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3532787