

Foundations and Trends® in Information Retrieval
Vol. XX, No. XX (2016) 1–10
© 2016 now Publishers Inc.
DOI: 10.1561/XXXXXXXXXX



Offline Evaluation for Information Retrieval

Jin Young Kim
Microsoft
jink@microsoft.com

Emine Yilmaz
University College London
emine.yilmaz@ucl.ac.uk

Contents

1	Introduction	2
1.1	Overview of Evaluation Paradigms in IR	2
1.2	Offline Evaluation for IR	2
1.3	Recent Trends in Offline Evaluation	3
2	Metrics	4
3	Judging Method	5
4	Crowdsourcing Judgment Collection	6
5	Experiment Design and Analysis	7
6	Evaluation Practices from Industry	8
	References	9

Abstract

now Publishers Inc.. *Offline Evaluation for Information Retrieval*. Foundations and Trends® in Information Retrieval, vol. XX, no. XX, pp. 1–10, 2016.

DOI: 10.1561/XXXXXXXXXX.

1

Introduction

1.1 Overview of Evaluation Paradigms in IR

Online vs. Offline evaluation

- What is it? Why is it important? How is it used?
- How are they different?
(advantages/disadvantages?)

Online-Offline Hybrid approaches

- Log-based offline evaluation (i.e, click models)
- Collecting feedback directly from users (Kim et al.)

Katja Hofmann [2016]

1.2 Offline Evaluation for IR

Traditional Approaches in Offline Evaluation

- Concept of relevance

Components of Offline Evaluation

- Search Task (Query / context)

- Judging Method (Interface / rating scale)
- Metric
- Experiment
- Test collections for offline evaluation (combining all the components)

Sanderson [2010] Borlund [2003] Cleverdon [1967]

1.3 Recent Trends in Offline Evaluation

Need for User-centric Evaluation

- Definition of User-centric
- Aiming for user satisfaction
- Evaluation based on models of user behavior
- Traditional metrics seem to not agree much with online signals, as well as each
 - o Need methodologies to better estimate user satisfaction (metric / judgments)
- How to address this issue?
 - o Metric design
 - o Judgment design

Extending the realms of evaluation

- Whole-page evaluation
- Session-level evaluation

New approaches

- Online/Offline hybrid
- Crowdsourcing / agile experimentation

2

Metrics

- Basic IR evaluation metrics
- Ranking-based metrics (Tau/TauAP)
- Evaluation metrics that are based on explicit models of user behaviour
 - o ERR, EBU, GAP, Time-biased gain, etc.
 - o Alpha-NDCG, IA metrics, etc.

3

Judging Method

- SERP-level evaluation
 - o Side by side / SASI
- Session/Task-based evaluation
 - o User study for search experience
- Effort based judgments
- Usefulness-based evaluation

Thomas and Hawking [2006] Chandar and Carterette [2013] Al-Maskari et al. [2007] Bailey et al. [2010] Carterette et al. [2008]

4

Crowdsourcing Judgment Collection

Crowd judges are closer to the user

- Different components (experiment, interface design, payment)

- Reducing noise in judging

 - o Multiple judgments and majority voting, etc.

 - o Statistics to measure judge agreement/noise

- Incentivising judges and how to make it more attractive (payment / I/F)

- Design decisions that need to be tackled

 - o Trade-off between how many labels per item

(fewer items with many labels versus more items with fewer labels)

Megorskaya et al. [2015] Davtyan et al. [2015]

5

Experiment Design and Analysis

Power analysis
Sensitivity analysis
Informativeness (MaxEnt)

6

Evaluation Practices from Industry

How are the companies doing?

-Google / Bing

-Netflix

-Facebook

Common features

- Online + offline evaluation

Practical tips

Gomez-Uribe and Hunt [2015]

References

- Azzah Al-Maskari, Mark Sanderson, and Paul Clough. The relationship between ir effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR*, SIGIR '07, pages 773–774, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. . URL <http://doi.acm.org/10.1145/1277741.1277902>.
- Peter Bailey, Nick Craswell, Ryen W. White, Liwei Chen, Ashwin Satyanarayana, and S. M.M. Tahaghoghi. Evaluating search systems using result page context. In *Proceedings of the third symposium on Information interaction in context*, IIX '10, pages 105–114, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0247-0. . URL <http://doi.acm.org/10.1145/1840784.1840801>.
- Pia Borlund. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, May 2003. ISSN 1532-2882.
- Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there. In *ECIR*, pages 16–27, 2008.
- Praveen Chandar and Ben Carterette. Preference based evaluation measures for novelty and diversity. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 413–422, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL <http://doi.acm.org/10.1145/2484028.2484094>.
- C. W. Cleverdon. The cranfield tests on index language devices. *Aslib*, 19: 173–192, 1967.

- Martin Davtyan, Carsten Eickhoff, and Thomas Hofmann. Exploiting document content for efficient aggregation of crowdsourcing votes. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 783–790, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3794-6. . URL <http://doi.acm.org/10.1145/2806416.2806460>.
- Carlos A. Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4):13:1–13:19, December 2015. ISSN 2158-656X. . URL <http://doi.acm.org/10.1145/2843948>.
- Filip Radlinski Katja Hofmann, Lihong Li. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 2016.
- Olga Megorskaya, Vladimir Kukushkin, and Pavel Serdyukov. On the relation between assessor’s agreement and accuracy in gamified relevance assessment. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 605–614, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. . URL <http://doi.acm.org/10.1145/2766462.2767727>.
- Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010. ISSN 1554-0669. . URL <http://dx.doi.org/10.1561/15000000009>.
- Paul Thomas and David Hawking. Evaluation by comparing result sets in context. In *Proceedings of the 15th ACM CIKM*, CIKM '06, pages 94–101, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2. .