Foundations and Trends® in Information Retrieval
Vol. XX, No. XX (2016) 1–16

DOI: 10.1561/XXXXXXXXXX

now
the essence of knowledge

Offline Evaluation for Information Retrieval

Jin Young Kim
Microsoft
jink@microsoft.com

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

Paul Thomas
Microsoft
pathom@microsoft.com

Contents

Abstract

Offline evaluation provides characterization of an IR system based on human judgments without relying on actual users in real-world environment. Offline evaluation, notably test collection based evaluation, has been dominant approaches in IR evaluation. It is no exaggeration that shared evaluation efforts such as TREC has defined the IR research over the years. The reason for this success lies in the ability to compare retrieval systems in a reusable manner.

Recently, there has been several trends which necessitates the change in the role and method of offline evaluation. First and foremost, online search engines with large-scale user base has become commonplace, enabling online evaluation based on user behavior. Also, there are many endpoints for search beyond desktop web browser, and the types of search results has diversified beyond the list of web documents to include other results types and direct answers. Finally, crowdsourcing has provided ways for human judgments of any kind to be collected at an large scale. The overall outcome of this trend is the advent of IR evaluation paradigm which is more user-centric, diverse and agile.

This survey aims to provide an overview of recent research in IR evaluation pertaining to the trends above. We first introduce offline evaluation for IR, focusing on how it relates to other evaluation paradigms such as online evaluation. We also overview traditional offline evaluation for IR, and how recent trends have shaped the research so far. We then review research in offline evaluation mainly on three levels: human judgment, evaluation metric and experiment design. This organization will allow readers to follow recent developments in research from micro-level (human judgment) to macro-level (experiment). Finally, we discuss evaluation practices from industry, which has been a major driving force in research and development in IR.

now Publishers Inc.. *Offline Evaluation for Information Retrieval*. Foundations and Trends® in Information Retrieval, vol. XX, no. XX, pp. 1–16, 2016.

DOI: 10.1561/XXXXXXXXXX.

1

Introduction

1.1 Evaluation Paradigms in IR

New Landscape in IR Evaluation Research
- More endpoints / models
- User-centric view (understanding user)
- Online evaluation (industry)
- Agile experimentation (crowdsourcing)

Online Katja Hofmann [2016] vs. Offline evaluation Sanderson [2010]
- What is it? Why is it important? How is it used?
- How are they different?

Offline evaluation vs. Log study (Click Modeling) Chuklin et al. [2015]
- Label-based vs. Behavior-based
- Experimental control(?)

Offline evaluation vs. User study Kelly [2009]
- Focus: system-to-system evaluation vs. understanding interaction/user behavior
- Scale(?) / Richness(?)

- Blurred distinction recently Bron et al. [2013] Liu et al. [2014] Shah and González-Ibáñez [2011]

1.2 Offline Evaluation for IR

Traditional Approaches in Offline Evaluation
- Concept of relevance
- Labels/Metrics based on Query-URLs
- Test collections

Borlund [2003] Cleverdon [1967] Voorhees and Harman [2005]

General Components of Offline Evaluation
- Search Task (Query / context)
- Judging Method (Interface / rating scale)
- Metric
- Experiment

1.3 Recent Trends in Offline Evaluation

Definition of User-centric:
- Aiming for user satisfaction
- Evaluation based on models of user behavior

Need for User-centric Evaluation
- Cross-metric studies btw. online and offline evaluation Radlinski and Craswell [2010]
- Traditional metrics seem to not agree much with user behavior/satisfaction Al-Maskari et al. [2007]

Need methodologies to better estimate user satisfaction and behavior
- Metric design
- Judgment design

Extending the realms of evaluation
- Whole-page evaluation
- Session-level evaluation
- Desktop vs. Mobile / Typed vs. Spoken IR

Online-Offline Hybrid approaches

- Log-based offline evaluation Li et al. [2015] Li et al. [2010]
- Collecting feedback directly from users (Kim et al.)
- Crowdsourcing / Agile Experiment

2

Human Judgment

Considerations in collecting high-quality labels from human judges at scale

2.1 Judgment Design

SERP-level evaluation:
- Preference Judgment Chandar and Carterette [2013] Carterette et al. [2008]
- Side by side Thomas and Hawking [2006] Kim et al. [2013]
- SASI Bailey et al. [2010]

Judgment for Desktop vs. Mobile environment Verma and Yilmaz [2016]

Judgment based on Query vs. Intent Description Yilmaz et al. [2014a]

Session/Task-based evaluation Moraveji et al. [2011] Xu and Mease [2009]

Effort based judgments Yilmaz et al. [2014b] Verma et al. [2016]
- Relevance vs. Usefulness-based evaluation

2.2 Judgment Collection

Choosing Judges:
- Crowd vs. Expert Kazai et al. [2013] Alonso and Mizzaro [2012]
- Query owner vs. non-owners Chouldechova and Mease [2013]

Reducing noise in judging:
- Anchoring bias in judging Shokouhi et al. [2015]
- Multiple judgments and majority voting, etc. Venanzi et al. [2014]

Efficient judgment collection using Crowdsourcing
- Design decisions that need to be tackled Blanco et al. [2011] Kazai et al. [2012] Alonso [2012] Alonso et al. [2015] Scholer et al. [2013]
- Incentivising judges and how to make it more attractive (payment / I/F) Megorskaya et al. [2015] Davtyan et al. [2015] Rokicki et al. [2014] Eickhoff et al. [2012]

3

Evaluation Metrics

How to combine labels to meaningful numbers

Basic IR evaluation metrics

- Ranking-based metrics (Tau/TauAP)

Evaluation metrics that are based on explicit models of user behaviour

- ERR, EBU, GAP, Time-biased gain, etc.
- Alpha-NDCG, IA metrics, etc.
- RBP / INST (notion of residual)

Estimating the distribution of parameters/metric values based on user data

Metrics for other domains of search

- Aggregate search Zhou et al. [2013] - Recommendation systems Gunawardana and Shani [2015]

4

Experiment Design

Drawing conclusions from metrics

- Hypothesis Testing Dinçer et al. [2014]

Analysis of Results

- Power analysis Sakai [2014]
- Sensitivity and Reliability analysis Urbano et al. [2013]
- Informativeness (MaxEnt)
- ETC Bron et al. [2013] Boytsov et al. [2013] Robertson and Kanoulas [2012]

5

Evaluation Practices from Industry

How are the companies doing?

- Google [1]
- Bing [2]
- Netflix Gomez-Uribe and Hunt [2015]
- Facebook [3]

Common features

- Online + offline evaluation

Practical tips

[1] How Search Works (Google) https://www.google.com/insidesearch/howsearchworks/thestory/
[2] The Role of Content Quality in Bing Ranking (Bing) http://bit.ly/1T1BaYN
[3] Who Controls Your Facebook Feed (Slate) http://slate.me/1T1BbvU

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.

Omar Alonso. Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval*, 16(2):101–120, 2012. ISSN 1573-7659. . URL http://dx.doi.org/10.1007/s10791-012-9204-1.

Omar Alonso and Stefano Mizzaro. Using crowdsourcing for {TREC} relevance assessment. *Information Processing {and} Management*, 48(6):1053 – 1066, 2012. ISSN 0306-4573. . URL http://www.sciencedirect.com/science/article/pii/S0306457312000052.

Omar Alonso, Catherine C. Marshall, and Marc Najork. Debugging a crowdsourced task with low inter-rater agreement. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '15, pages 101–110, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3594-2. . URL http://doi.acm.org/10.1145/2756406.2757741.

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*, IIiX '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.

Roi Blanco, Harry Halpin, Daniel M. Herzig, Peter Mika, Jeffrey Pound, Henry S. Thompson, and Thanh Tran Duc. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 923–932, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. . URL `http://doi.acm.org/10.1145/2009916.2010039`.

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.

Leonid Boytsov, Anna Belova, and Peter Westfall. Deciding on an adjustment for multiplicity in ir experiments. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 403–412, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL `http://doi.acm.org/10.1145/2484028.2484034`.

Marc Bron, Jasmijn van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. Aggregated search interface preferences in multi-session search tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 123–132, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL `http://doi.acm.org/10.1145/2484028.2484050`.

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`.

Alexandra Chouldechova and David Mease. Differences in search engine evaluations between query owners and non-owners. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 103–112, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1869-3. . URL `http://doi.acm.org/10.1145/2433396.2433411`.

Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. *Click Models for Web Search*. Morgan & Claypool, 2015. ISBN 9781627056489. .

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.

B. Taner Dinçer, Craig Macdonald, and Iadh Ounis. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 23–32, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. . URL http://doi.acm.org/10.1145/2600428.2609625.

Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 871–880, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. . URL http://doi.acm.org/10.1145/2348283.2348400.

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.

Asela Gunawardana and Guy Shani. Evaluating recommender systems. In *Recommender Systems Handbook*, pages 265–308. Springer, 2015.

Filip Radlinski Katja Hofmann, Lihong Li. Online evaluation for information retrieval. *Foundations and TrendsÂő in Information Retrieval*, 2016.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178, 2012. ISSN 1573-7659. . URL http://dx.doi.org/10.1007/s10791-012-9205-0.

Gabriella Kazai, Emine Yilmaz, Nick Craswell, and S.M.M. Tahaghoghi. User intent and assessor disagreement in web search evaluation. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 699–708, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. . URL http://doi.acm.org/10.1145/2505515.2505716.

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1âĂŤ2): 1–224, 2009.

Jinyoung Kim, Gabriella Kazai, and Imed Zitouni. Relevance dimensions in preference-based ir evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 913–916, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL `http://doi.acm.org/10.1145/2484028.2484168`.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Lihong Li, Jin Young Kim, and Imed Zitouni. Toward predicting the outcome of an a/b experiment for search relevance. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 37–46, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3317-7. . URL `http://doi.acm.org/10.1145/2684822.2685311`.

Chang Liu, Jingjing Liu, and Nicholas J. Belkin. Predicting search task difficulty at different search stages. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 569–578, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. . URL `http://doi.acm.org/10.1145/2661829.2661939`.

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`.

Neema Moraveji, Daniel Russell, Jacob Bien, and David Mease. Measuring improvement in user search performance resulting from optimal search tips. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 355–364, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. . URL `http://doi.acm.org/10.1145/2009916.2009966`.

Filip Radlinski and Nick Craswell. Comparing the sensitivity of information retrieval metrics. In *SIGIR*, pages 667–674, 2010.

Stephen E. Robertson and Evangelos Kanoulas. On per-topic variance in ir evaluation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 891–900, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. . URL `http://doi.acm.org/10.1145/2348283.2348402`.

Markus Rokicki, Sergiu Chelaru, Sergej Zerr, and Stefan Siersdorfer. Competitive game designs for improving the cost effectiveness of crowdsourcing. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1469–1478, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. . URL http://doi.acm.org/10.1145/2661829.2661946.

Tetsuya Sakai. Designing test collections for comparing many systems. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 61–70, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. . URL http://doi.acm.org/10.1145/2661829.2661893.

Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and TrendsÂő in Information Retrieval*, 4(4):247–375, 2010. ISSN 1554-0669. . URL http://dx.doi.org/10.1561/1500000009.

Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S. Lee, and William Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 623–632, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL http://doi.acm.org/10.1145/2484028.2484090.

Chirag Shah and Roberto González-Ibáñez. Evaluating the synergic effect of collaboration in information seeking. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 913–922, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. . URL http://doi.acm.org/10.1145/2009916.2010038.

Milad Shokouhi, Ryen White, and Emine Yilmaz. Anchoring and adjustment in relevance estimation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 963–966, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. . URL http://doi.acm.org/10.1145/2766462.2767841.

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. .

Julián Urbano, Mónica Marrero, and Diego Martín. On the measurement of test collection reliability. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 393–402, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL http://doi.acm.org/10.1145/2484028.2484038.

Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 155–164, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2744-2. . URL `http://doi.acm.org/10.1145/2566486.2567989`.

Manisha Verma and Emine Yilmaz. *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, chapter Characterizing Relevance on Mobile and Desktop, pages 212–223. Springer International Publishing, Cham, 2016. ISBN 978-3-319-30671-1. . URL `http://dx.doi.org/10.1007/978-3-319-30671-1_16`.

Manisha Verma, Emine Yilmaz, and Nick Craswell. On obtaining effort based judgements for information retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 277–286, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3716-8. . URL `http://doi.acm.org/10.1145/2835776.2835840`.

Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experimentation and Evaluation in Information Retrieval*. MIT Press, 2005.

Ya Xu and David Mease. Evaluating web search using task completion time. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 676–677, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. . URL `http://doi.acm.org/10.1145/1571941.1572073`.

Emine Yilmaz, Evangelos Kanoulas, and Nick Craswell. Effect of intent descriptions on retrieval evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 599–608, New York, NY, USA, 2014a. ACM. ISBN 978-1-4503-2598-1. . URL `http://doi.acm.org/10.1145/2661829.2661950`.

Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. Relevance and effort: An analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 91–100, New York, NY, USA, 2014b. ACM. ISBN 978-1-4503-2598-1. . URL `http://doi.acm.org/10.1145/2661829.2661953`.

Ke Zhou, Mounia Lalmas, Tetsuya Sakai, Ronan Cummins, and Joemon M. Jose. On the reliability and intuitiveness of aggregated search metrics. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 689–698, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. . URL `http://doi.acm.org/10.1145/2505515.2505691`.