Foundations and Trends® in Information Retrieval
Vol. XX, No. XX (2016) 1–27
© 2016 now Publishers Inc.
DOI: 10.1561/XXXXXXXXXX

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 information retrieval (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 such as mobile phone and conversational agents, 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 new IR evaluation paradigms which are 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. 1–27, 2016.

DOI: 10.1561/XXXXXXXXXX.

1

Introduction

In this chapter, we survey the area and lay conceptual foundation for the rest of the paper. We first provide an overview of different approaches to IR evaluation. We then focus on offline evaluation, explaining traditional approaches and recent trends. Finally, we introduce a conceptual framework and the outline for the rest of this paper. (15-20 pages)

1.1 Evaluation Paradigms in IR

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

Offline evaluation Sanderson [2010] vs. Online evaluation Katja Hofmann [2016]
- Signals from human judges vs. users

- Early-stage vs. production-stage iteration
- Explicit judgment vs. implicit behavior
- Experimental control vs. realistic environment

Offline evaluation vs. Counterfactual online evaluation Chuklin et al. [2015] Li et al. [2015, 2010]
- Label-based vs. Behavior-based
- Independent metric of quality vs. predictor of online evaluation results

Offline evaluation vs. User study Kelly [2009]
- Focus: system-to-system evaluation vs. understanding interaction/user behavior
- Difference in scale and richness
- Blurred distinction recently

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

The new role of offline evaluation
- Evaluation at the early stage of development
- Experimental control and resolution
- Reusability across experiments

1.2 Offline Evaluation for IR

1.2.1 Traditional Approaches in Offline Evaluation

Conceptual Model
- Labels/Metrics based on Query-URLs
- Test collections
- Concept of relevance

History
- TREC and related evaluation venues Sanderson [2010]
- Refer to Borlund [2003] Cleverdon [1967] Voorhees and Harman [2005]

1.2.2 Recent Trends in Offline Evaluation

Need for User-centric Evaluation
- Definition: Aiming for user satisfaction, or other user-visible measure such as engagement or task completion (Scholer et al. [2013b]), evaluation based on models of user behavior
- Traditional metrics seem to not agree much with user behavior/satisfaction Al-Maskari et al. [2007]
- Cross-metric studies btw. online and offline evaluation Radlinski and Craswell [2010]

Need methodologies to better estimate user satisfaction and behavior
- Metric design Yilmaz et al. [2010], Carterette et al. [2011], Chapelle et al. [2009]
- Judgment design Verma and Yilmaz [2016b], Verma et al. [2016b]

Extending the realms of evaluation
- Whole-page evaluation Zhou et al. [2012]
- Session-level evaluation Kanoulas et al. [2011a], Carterette et al. [2014]
- Desktop vs. Mobile / Typed vs. Spoken IR Verma et al. [2016b]

Online-Offline Hybrid approaches
- Log-based offline evaluation Li et al. [2015] Li et al. [2010]
- Collecting feedback directly from users Kim et al. [2016]
- Crowdsourcing / Agile Experiment

1.3 General Framework for Offline Evaluation

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

Organization of this paper: A pipeline for offline evaluation

- Select an audience (who you want to talk to: end users, accountants, sysadmins, advertisers).

- Collect appropriate data: documents, tasks, queries, judgments. Chapter 2 deals with gathering judgments, which need to be created for the purpose.

- Choose a metric based on your tasks and on likely user models. Aggregate judgments, if needed, to compute the metric. Chapter 3 considers this.

- Examine the metric to draw some conclusion. This is covered in Chapter 4.

2

Collecting Human Judgments

The first step in offline evaluation is collecting labels from human judges. In this chapter, we describe various considerations in collecting high-quality labels from human judges at scale. We first discuss the method for collecting search tasks, followed by the design of a judging method. We then discuss the collection of actual judgments, which is an non-trivial task to perform at scale. We also cover the trade-off and in using different types of judging resources – in-house vs. crowd judges. (20-25 pages)

2.1 Collecting Search Tasks

Collect hypothetical search tasks
- Examples from user study papers

Sample search tasks from existing system
- Which sampling methods to use? Baeza-Yates [2015]

2.2 Designing a Judging Method

Judging Unit: URL vs. 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 [2016a]

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. [2016a]
- Relevance vs. Usefulness-based evaluation

2.3 Collecting Judgments

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] Aroyo and Welty [2013b] Aroyo and Welty [2013a]

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. [2013a]
- 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]

2.4 Open Issues

- Collecting labels for contextual / personalized search results

- Collecting labels for whole SERP / non-document results
- Collecting labels for non-traditional endpoints (i.e., conversational agent)

3

Designing Evaluation Metrics

The second step in offline evaluation is designing a meaningful evaluation metric. This is essentially the question of how to combine labels to meaningful numbers. For traditional IR evaluation where the labels are collected at query-URL level, combining labels to a metric requires quite a few assumptions, or even a user model. In this chapter, we go over the various considerations of IR metric design, as well as the user models behind these metrics. We briefly survey some established metrics but spend more time on recent developments: explicit models of user behavior, deriving metrics from these, and open issues including session-level measurement, dealing with variation, and considering rich SERPs. (20-25 pages)

3.1 Basic IR evaluation metrics

- Metrics based on absolute judgments (e.g. Cooper [1973])

- Metrics based on preference-based judgments, including e.g. aggregated in-situ side-by-side Thomas and Hawking [2006]

- Ranking-based metrics (Tau/TauAP)

3.2 Metrics based on simple aggregation of labels/qrels

- Set-based: P, R
 - Rank-based: P@k, AP, RR
 - Criticisms: what tasks and behaviors are modeled here?

3.3 Models of behavior

Evaluation metrics that are based on explicit models of user behavior
 - The cascade model and variants
 - Weights, the C/L/W framework [Moffat et al., 2013]
 - ERR, EBU, GAP, Time-biased gain, etc.
 - Weighted precision metrics such as RBP, INST; notion of residual [Moffat and Zobel, 2008, Moffat et al., 2015]
 - α-NDCG, IA metrics, etc.
 - Cost-based/economic models and the prospects of metrics from these
 - Session-level metrics Kanoulas et al. [2011b] Järvelin et al. [2008]

3.4 Model fitting

Fit of metrics to models; estimating the distribution of parameters/metric values based on user data
 Carterette et al. [2011], Moffat et al. [2013]

3.5 Open issues

Open issues in behavior models and the corresponding metrics
 - Whole-page quality
 - Caption effects
 - Variation between users: behaviors, learning styles, cognitive styles, topic expertise, search system expertise, expectations of the system, query variation, ...
 - Duplication in SERPs
 - Learning (?)
 - Non-traditional tasks

- Choosing between metrics; sensitivity; finding evidence any of them correlates with user behavior or other important dependent variables

- Measuring things outside the SERP: query formulation, source/engine selection

4

Designing and Analyzing Experiments

Experiments is defined as a set of labels and metrics defined on top of them. We first look over many considerations in order to design an experiment given a budget and time constraint. We then focus on a set of analyses we can perform once the data is collected, followed by the ways of reporting experimental results. ($\approx$ 15 pages)

4.1 Designing an Experiment

- How to select queries?
    - How many queries? Sakai [2014]
    - How many documents? Carterette et al. [2009a]
    - How to distribute judgment efforts across queries and documents? Carterette et al. [2009b], Yilmaz and Robertson [2009]

4.2 Analysis of Experimental Results

Drawing conclusions from metrics
    - Hypothesis Testing Dinçer et al. [2014]
    - Comparison of different types of significance tests Smucker et al. [2009]

Various analysis methods
- Power analysis Sakai [2014]
- Failure analysis
- Sensitivity and Reliability analysis Urbano et al. [2013]
- Informativeness (MaxEnt) Aslam et al. [2005]
- ETC Bron et al. [2013] Boytsov et al. [2013] Robertson and Kanoulas [2012]

Reporting results
- Effect sizes and distributions, vs point estimates and p values

4.3 Open Issues

- Reusability for SERP/task-level evaluation
- Beyond significance testing – bayesian alternatives?
- Reusability / Generalizability of experimental results

5

IR Evaluation in Practice

In this chapter, we review evaluation practices happening in both academia and industry. We first cover evaluation practices from academia, including recent TREC tracks, data generation efforts. We also look at evaluation efforts in related area such as recommendation systems and conversational agents. We then turn to evaluation practices from industry including major players in search and recommendation based on published papers and articles.

5.1 Evaluation Practices from Academia

Emerging TREC tracks

- Task track
- Microblog track
- Live QA track
- Contextual suggestions track

Dataset generation efforts

- Living labs for IR [1]

[1] http://living-labs.net/

- Data set shared by industry [2]

Evaluation in related domains
- Aggregate search Zhou et al. [2013]
- Recommendation systems Gunawardana and Shani [2015]
- Conversational agents

5.2 Evaluation Practices from Industry

How are the practitioners doing? (≈15 pages)
- Google [3] [4]
- Bing [5]
- Netflix Gomez-Uribe and Hunt [2015] [6]
- Facebook [7]

Common features: combine online and offline evaluation
- Offline evaluation for early iteration
- Online evaluation for final ship decisions

[2] http://jeffhuang.com/search_query_logs.html

[3] How Search Works (Google) https://www.google.com/insidesearch/howsearchworks/thestory/

[4] Updating Our Search Quality Rating Guidelines https://webmasters.googleblog.com/2015/11/updating-our-search-quality-rating.html

[5] The Role of Content Quality in Bing Ranking (Bing) http://bit.ly/1T1BaYN

[6] The Netflix Tech Blog: Learning a Personalized Homepage http://techblog.netflix.com/2015/04/learning-personalized-homepage.html

[7] Who Controls Your Facebook Feed (Slate) http://slate.me/1T1BbvU

6

Conclusions

In this chapter we conclude this survey by providing the summary of contents so far. We also provide a brief outlook toward the future of offline evaluation for IR.

6.1 Summary

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

6.2 Future of Offline Evaluation for IR

Emerging trends in the tech ecosystem
- Mobile-first: different interfaces and information needs
- Natural-language interaction: Bots and Conversational agents
- End-to-end support for task completion: e.g., restaurant reservation

Future of Offline Evaluation
- Evaluation of search agents (as well as engines)
- Evaluation of various information 'cards'
- Evaluation of end-to-end task completion

Future of Offline Evaluation Research
- Need for Academy-Industry collaboration

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.

Lora Aroyo and Chris Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013, 2013a.

Lora Aroyo and Chris Welty. Measuring crowd truth for medical relation extraction. In *2013 AAAI Fall Symposium Series*, 2013b.

Javed A. Aslam, Emine Yilmaz, and Virgiliu Pavlu. The maximum entropy method for analyzing retrieval measures. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, pages 27–34, 2005. . URL http://doi.acm.org/10.1145/1076034.1076042.

Ricardo Baeza-Yates. Incremental sampling of query logs. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1093–1096, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. . URL http://doi.acm.org/10.1145/2766462.2776780.

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.

Ben Carterette, Virgiliu Pavlu, Hui Fang, and Evangelos Kanoulas. Million query track 2009 overview. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*, 2009a. URL http://trec.nist.gov/pubs/trec18/papers/MQ09OVERVIEW.pdf.

Ben Carterette, Virgiliu Pavlu, Evangelos Kanoulas, Javed A. Aslam, and James Allan. If I had a million queries. In *Advances in Information Retrieval, 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings*, pages 288–300, 2009b. . URL http://dx.doi.org/10.1007/978-3-642-00958-7_27.

Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 611–620, 2011. . URL http://doi.acm.org/10.1145/2063576.2063668.

Ben Carterette, Evangelos Kanoulas, Mark M. Hall, and Paul D. Clough. Overview of the TREC 2014 session track. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, 2014. URL http://trec.nist.gov/pubs/trec23/papers/overview-session.pdf.

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.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 621–630, 2009. . URL http://doi.acm.org/10.1145/1645953.1646033.

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.

William S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100, 1973. ISSN 1097-4571. . URL `http://dx.doi.org/10.1002/asi.4630240204`.

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.

Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. *Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, chapter Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions, pages 4–15. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-78646-7. . URL `http://dx.doi.org/10.1007/978-3-540-78646-7_4`.

Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. Evaluating multi-query sessions. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1053–1062, 2011a. . URL `http://doi.acm.org/10.1145/2009916.2010056`.

Evangelos Kanoulas, Ben Carterette, Paul D Clough, and Mark Sanderson. Evaluating multi-query sessions. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1053–1062. ACM, 2011b.

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.

Jin Young Kim, Jaime Teevan, and Nick Craswell. Explicit in situ user feedback for web search results. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16. ACM, 2016.

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`.

Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1), 2008. paper 2.

Alistair Moffat, Paul Thomas, and Falk Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 659–668, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. . URL `http://doi.acm.org/10.1145/2505515.2507665`.

Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. INST: An adaptive metric for information retrieval evaluation. In *Proceedings of the Australasian Document Computing Symposium*, 2015.

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, 2013a. ACM. ISBN 978-1-4503-2034-4. . URL `http://doi.acm.org/10.1145/2484028.2484090`.

Falk Scholer, Alistair Moffat, and Paul Thomas. Choices in batch information retrieval evaluation. In *Proceedings of the Australasian Document Computing Symposium*, 2013b.

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`.

Mark D. Smucker, James Allan, and Ben Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 630–631, 2009. . URL http://doi.acm.org/10.1145/1571941.1572050.

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, 2016a. ISBN 978-3-319-30671-1. . URL http://dx.doi.org/10.1007/978-3-319-30671-1_16.

Manisha Verma and Emine Yilmaz. Characterizing relevance on mobile and desktop. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, pages 212–223, 2016b. . 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, 2016a. ACM. ISBN 978-1-4503-3716-8. . URL http://doi.acm.org/10.1145/2835776.2835840.

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, San Francisco, CA, USA, February 22-25, 2016*, pages 277–286, 2016b. . 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 and Stephen Robertson. Deep versus shallow judgments in learning to rank. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 662–663, 2009. . URL http://doi.acm.org/10.1145/1571941.1572066.

Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1561–1564, 2010. . URL http://doi.acm.org/10.1145/1871437.1871672.

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, Ronan Cummins, Mounia Lalmas, and Joemon M. Jose. Evaluating aggregated search pages. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 115–124, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. . URL http://doi.acm.org/10.1145/2348283.2348302.

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`.