Full text available at: http://dx.doi.org/10.1561/1500000051

Online Evaluation for Information Retrieval

Katja Hofmann
Microsoft
katja.hofmann@microsoft.com

Lihong Li
Microsoft
lihongli@microsoft.com

Filip Radlinski
Microsoft
filip.radlinski@microsoft.com

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

K. Hofmann, L. Li, and F. Radlinski. *Online Evaluation for Information Retrieval.* Foundations and Trends® in Information Retrieval, vol. 10, no. 1, pp. 1–117, 2016.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-162-7

Foundations and Trends® in Information Retrieval
Volume 10, Issue 1, 2016

Editorial Board

Editors-in-Chief

Editors

Full text available at: http://dx.doi.org/10.1561/1500000051

Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2016, Volume 10, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in Information Retrieval
Vol. 10, No. 1 (2016) 1–117

DOI: 10.1561/1500000051

now
the essence of knowledge

Online Evaluation for Information Retrieval

Katja Hofmann
Microsoft
katja.hofmann@microsoft.com

Lihong Li
Microsoft
lihongli@microsoft.com

Filip Radlinski
Microsoft
filip.radlinski@microsoft.com

Contents

Abstract

Online evaluation is one of the most common approaches to measure the effectiveness of an information retrieval system. It involves fielding the information retrieval system to real users, and observing these users' interactions in-situ while they engage with the system. This allows actual users with real world information needs to play an important part in assessing retrieval quality. As such, online evaluation complements the common alternative offline evaluation approaches which may provide more easily interpretable outcomes, yet are often less realistic when measuring of quality and actual user experience.

In this survey, we provide an overview of online evaluation techniques for information retrieval. We show how online evaluation is used for controlled experiments, segmenting them into experiment designs that allow absolute or relative quality assessments. Our presentation of different metrics further partitions online evaluation based on different sized experimental units commonly of interest: documents, lists and sessions. Additionally, we include an extensive discussion of recent work on data re-use, and experiment estimation based on historical data.

A substantial part of this work focuses on practical issues: How to run evaluations in practice, how to select experimental parameters, how to take into account ethical considerations inherent in online evaluations, and limitations. While most published work on online experimentation today is at large scale in systems with millions of users, we also emphasize that the same techniques can be applied at small scale. To this end, we emphasize recent work that makes it easier to use at smaller scales and encourage studying real-world information seeking in a wide range of scenarios. Finally, we present a summary of the most recent work in the area, and describe open problems, as well as postulating future directions.

K. Hofmann, L. Li, and F. Radlinski. *Online Evaluation for Information Retrieval.* Foundations and Trends® in Information Retrieval, vol. 10, no. 1, pp. 1–117, 2016. DOI: 10.1561/1500000051.

1

Introduction

Information retrieval (IR) has a long and fruitful tradition of empirical research. Since early experiments on indexing schemes, and the development of the Cranfield paradigm, researchers have been striving to establish methodology for empirical research that best supports their research goals – to understand human information seeking, and to develop the most effective technology to support it.

In the past decade, IR systems, from large-scale commercial Web search engines to specialized analysis software, have become ubiquitous. They have transformed the way in which we access information, and are for many an integral part of their daily lives. This shift towards everyday, ubiquitous IR systems is posing new challenges for empirical research. While it was previously possible to substantially improve IR systems by measuring and optimizing reasonably objective criteria, such as topical relevance, this is no longer sufficient. IR systems are becoming increasingly contextual and personal. They take into account information about their users' current situation as well as previous interactions, and aim to predict their users' requirements and preferences given new contexts. No longer can users or experts be asked to provide objective assessments of retrieval quality for such complex scenarios.

Online evaluation for IR addresses the challenges that require assessment of systems in terms of their utility for the user. The current state of the art provides a set of methods and tools, firmly grounded in and informed by the tradition of controlled experimentation. Giving an overview of these methods and their conceptual foundations, as well as guiding the reader on how to run their own online evaluations are the purposes of this survey.

In the next section, we define key concepts and terminology used throughout this survey. Then, we closely examine the motivations for online evaluation, and provide example use cases. Finally, we outline the scope and organization of the remainder of this survey.

1.1 Terminology

For the purpose of this survey, we adopt the following definition of *online evaluation*.

Definition 1.1. Online evaluation is evaluation of a *fully functioning* system based on *implicit measurement* of *real users'* experiences of the system in a *natural* usage environment.

The first key to the definition is *implicit measurement*, which we take to include any measurements that can be derived from observable user activity that is part of users' *natural* or *normal* interaction with the system [Kelly and Teevan, 2003]. Implicit measurements can range from low-level and potentially noisy signals, such as clicks or dwell-times, to more robust signals, such as purchase decisions. The key distinction we make between implicit and explicit measurements is that implicit measurements are a by-product of users' natural interaction, while *explicit* ones are specifically collected for feedback purposes. Both types of measures can also be combined into composite metrics capturing higher-level concepts, such as user satisfaction. These considerations give rise to a wide range of metrics, as discussed in Chapter 3.

We specifically include methods for *offline estimation*, *i.e.*, the estimation of online evaluation metrics based on past observations of users' behavior, in Chapter 4. Such estimation substantially increases the flex-

ibility of online evaluation and facilitates theoretically well-founded end-to-end evaluation of system components.

1.2 Motivation and Uses

Online evaluation is often seen as a set of methods that are particularly applicable in industry and industrial research. In these settings, a fully functioning IR system is typically available and in need of constant innovation. These factors have significantly contributed to the rapid adoption of online evaluation techniques in these settings. In industry, online evaluation approaches such as AB tests (*c.f.*, Section 2.4) and interleaved comparisons (Section 2.6) are now the state of the art for evaluating system effectiveness [Kohavi et al., 2009, Radlinski and Craswell, 2010, Li et al., 2011, Bendersky et al., 2014].

However, it is important to recall that much of the initial work on online evaluation originated in academic settings. Important motivations here were the need for reliable measurement of search quality of specialized search services [Radlinski et al., 2008c]. This line of work originated in the tradition of interactive IR. The fruitful exchange of ideas between applications and research continues today. On the one hand, practical challenges from IR applications motivate the development of online evaluation methodology; Chapter 2 gives a few examples. On the other hand, lessons learned in practical applications make their way into the state-of-the-art methodological tool set of IR researchers.

In the context of both practical applications and basic research, a key aspect of online evaluation is its reliance on controlled experiments. This allows the researcher to answer explanatory questions, which can explain causal relations in observed phenomena. In practical settings, this is crucial for correctly attributing observed changes in user behavior to system behavior. In research, this allows the development of theory in terms of causal concepts. More details on controlled experiments for online evaluation are provided in Chapter 2.

Finally, in Chapter 5, we discuss pros and cons of online evaluation, compared with more traditional offline evaluation methodology. This will help guide the reader to understand when an online evaluation is suitable, and when it is not.

1.3 This Survey

Online evaluation comprises a specific set of tools and methods that we see as complementary to other evaluation approaches in IR. In particular, online evaluation addresses questions about users' experience with an IR system that are quite distinct from those answered by *offline* evaluation using a test-collection-based approach, surveyed by Sanderson [2010]. Test-collection-based evaluation models users at varying levels of abstractions, and uses explicit assessments and offline metrics to assess system performance under these abstractions. Questions that are more appropriate for offline evaluation are those for which reliable and unbiased judgments can be collected from assessors (be they trained experts or crowdsourced representative users), but would be hard to infer from user interactions; an example being the quality of a document. Vice versa, online evaluation is more appropriate when the opposite is the case: for example, which of two topically relevant documents users find more interesting.

This survey does not discuss test-collection-based approaches in any detail, but points out conceptual differences when deemed appropriate. Furthermore, Chapter 5 focuses on online evaluation and test-collection-based approaches along a few dimensions.

Closely related to online evaluation is the long tradition of interactive IR (IIR) and the experimental framework developed for it, as surveyed by Kelly and Gyllstrom [2011]. We see online evaluation as a continuation of the IIR tradition, with considerable overlap. However, online evaluation extends to the specific requirements, limitations, and opportunities afforded by the scale, natural settings, and levels of control that are available in online settings. Generally speaking, IIR approaches, such as lab studies, are more appropriate for answering questions that require a high level of experimental control: for example, which tasks or queries a study participant is asked to solve. Conversely, online evaluation is preferred when researchers aim to study natural interactions at scale. This survey necessarily overlaps with some of the material that is relevant in the IIR setting, and we endeavor to point out connections as much as feasible. Our main focus will be on methodological questions that are specific to online evaluation settings.

Throughout this survey, we consider IR in a broad sense, including for instance recommender systems and advertisement placement. Many aspects of online evaluation are shared across these areas. For example, early work on using historical information for estimating online performance focused on ad placement [Langford et al., 2008] and news recommendation [Li et al., 2011]. We cover work in all these areas, and emphasize work that is specific to IR, such as search result ranking evaluation, as appropriate.

We have also highlighted particular places in the text with tips (such as the one below) that may be particularly useful for experimenters performing online evaluation without having access to very large user bases. While at first glance online evaluation may appear to be best suited to settings such as commercial search engines, in fact it has been widely used in academic settings as well.

Tip for small-scale experiments #1

Online evaluation can also be used with just tens of users, or hundreds of queries. Particular tips for experiments with few users are highlighted in the text with a box like this one.

1.4 Organization

We start in Chapter 2 by motivating the need for controlled experiments and detailing common experiment designs used in online evaluation, with a focus on experimentation methodologies that are particularly useful for IR. Following this, Chapter 3 gives an extensive overview of the variety of metrics that have been proposed for different tasks and research questions. Considering how to re-use online measurement data, Chapter 4 details offline estimation of online metrics from historical data. Turning to more practical issues, Chapter 5 discusses advantages and limitations of online evaluation, while Chapter 6 discusses practical issues around running online experiments. Finally, Chapter 7 concludes this survey with an outlook on emerging trends and open challenges.

References

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Miro Dudík, John Langford, Lihong Li, Luong Hoang, Dan Melamed, Siddhartha Sen, Robert Schapire, and Alex Slivkins. Multi-world testing: A system for experimentation, learning, and decision-making, 2016. Microsoft whitepaper. URL: http://mwtds.azurewebsites.net.

Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, and Xuanhui Wang. Click shaping to optimize multiple objectives. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 132–140, 2011.

Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, and Xuanhui Wang. Personalized click shaping through Lagrangian duality for online recommendation. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 485–494, 2012.

Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, 2011.

Rakesh Agrawal, Alan Halverson, Krishnaram Kenthapadi, Nina Mishra, and Panayiotis Tsaparas. Generating labels from clicks. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 172–181, 2009.

Omar Alonso. Implementing crowdsourcing-based relevance experimentation: An industrial perspective. *Information Retrieval*, 16:101–120, 2013.

Omar Alonso and Stefano Mizzaro. Can we get rid of TREC assessors? using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, volume 15, page 16, 2009.

Natalia Andrienko and Gennady Andrienko. *Exploratory analysis of spatial and temporal data: A systematic approach*. Springer Science & Business Media, 2006.

Olga Arkhipova, Lidia Grauer, Igor Kuralenok, and Pavel Serdyukov. Search engine evaluation based on search engine switching prediction. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 723–726, 2015a.

Olga Arkhipova, Lidia Grauer, Igor Kuralenok, and Pavel Serdyukov. Search engine evaluation based on search engine switching prediction. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 723–726, 2015b.

Arthur Asuncion and David J. Newman. UCI machine learning repository, 2007. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Leif Azzopardi. Modelling interaction with economic models of search. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 3–12, 2014.

Leif Azzopardi, Maarten De Rijke, and Krisztian Balog. Building simulated queries for known-item topics: An analysis using six european languages. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 455–462, 2007.

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008.

Yoram Bachrach, Sofia Ceppi, Ian A. Kash, Peter Key, and David Kurokawa. Optimising trade-offs among stakeholders in ad auctions. In *Proceedings of the ACM Conference on Economics and Computation (EC-14)*, pages 75–92, 2014.

Eytan Bakshy and Eitan Frachtenberg. Design and analysis of benchmarking experiments for distributed internet services. In *Proceedings of the International World Wide Web Conference (WWW)*, 2015.

Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. Designing and deploying online field experiments. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 283–292, 2014.

Krisztian Balog, Liadh Kelly, and Anne Schuth. Head first: Living labs for ad-hoc search evaluation. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, 2014.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

Michael Bendersky, Lluis Garcia-Pueyo, Jeremiah Harmsen, Vanja Josifovski, and Dima Lepikhin. Up next: Retrieval methods for large scale related video suggestion. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1769–1778, 2014.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 129–138, 2009.

Susanne Boll, Niels Henze, Martin Pielot, Benjamin Poppinga, and Torben Schinke. My app is an experiment: Experience from user studies in mobile app stores. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 3(4):71–91, 2011.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis Xavier Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research (JMLR)*, 14:3207–3260, 2013.

Justin Boyan, Dayne Freitag, and Thorsten Joachims. A machine learning architecture for optimizing Web search engines. In *AAAI Workshop on Internet Based Information Systems*, pages 1–8, 1996.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Giuseppe Burtini, Jason Loeppky, and Ramon Lawrence. A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*, 2015.

Georg Buscher, Susan T. Dumais, and Edward Cutrell. The good, the bad, and the random: An eye-tracking study of ad quality in web search. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 42–49, 2010.

Donald T Campbell and Julian C Stanley. *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin Company, 1966.

Ben Carterette. Statistical significance testing in information retrieval: Theory and practice. In *Proceedings of the International Conference on The Theory of Information Retrieval (ICTIR)*, 2013.

Ben Carterette and Rosie Jones. Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 217–224, 2007.

Ben Carterette, Evgeniy Gabrilovich, Vanja Josifovski, and Donald Metzler. Measuring the reusability of test collections. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 231–240, 2010.

George Casella and Roger L. Berger. *Statistical Inference*. Cengage Learning, 2nd edition, 2001.

Sunandan Chakraborty, Filip Radlinski, Milad Shokouhi, and Paul Baecke. On correlation of absence time and search effectiveness. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1163–1166, 2014.

Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. *Journal of Machine Learning Research - Proceedings Track*, 14:1–24, 2011.

Olivier Chapelle and Ya Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 1–10, 2009.

Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. Large-scale validation and analysis of interleaved search evaluation. *Transactions on Information System (TOIS)*, 30(1):6:1–6:41, 2012.

Ye Chen, Yiqun Liu, Ke Zhou, Meng Wang, Min Zhang, and Shaoping Ma. Does vertical bring more satisfaction?: Predicting search satisfaction in a heterogeneous environment. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 1581–1590, 2015.

Abdur Chowdhury and Ian Soboroff. Automatic evaluation of world wide web search services. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 421–422, 2002.

Aleksandr Chuklin, Anne Schuth, Katja Hofmann, Pavel Serdyukov, and Maarten de Rijke. Evaluating Aggregated Search Using Interleaving. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, 2013a.

Aleksandr Chuklin, Pavel Serdyukov, and Maarten de Rijke. Click model-based information retrieval metrics. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 493–502, 2013b.

Aleksandr Chuklin, Anne Schuth, Ke Zhou, and Maarten de Rijke. A comparative analysis of interleaving methods for aggregated search. *Transactions on Information System (TOIS)*, 2014.

Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. *Click Models for Web Search. Synthesis Lectures on Information Concepts, Retrieval, and Services*, volume 7. Morgan & Claypool Publishers, 2015.

Charles L. A. Clarke, Eugene Agichtein, Susan Dumais, and Ryen W. White. The influence of caption features on clickthrough patterns in web search. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 135–142, 2007.

Cyril Cleverdon. The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6):173–194, 1967.

Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 87–94, 2008.

Alex Deng. Objective Bayesian two sample hypothesis testing for online controlled experiments. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 923–928, 2015.

Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 123–132, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

Fernando Diaz, Ryen White, Georg Buscher, and Dan Liebling. Robust models of mouse movement on dynamic web search results pages. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 1451–1460, 2013.

Abdigani Diriye, Ryen White, Georg Buscher, and Susan Dumais. Leaving so soon? Understanding and predicting web search abandonment rationales. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 1025–1034, 2012.

Anlei Dong, Jiang Bian, Xiaofeng He, Srihari Reddy, and Yi Chang. User action interpretation for personalized content optimization in recommender systems. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 2129–2132, 2011.

Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 256–266, 2015a.

Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. Engagement periodicity in search engine usage: Analysis and its application to search quality evaluation. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 27–36, 2015b.

Alexey Drutsa, Anna Ufliand, and Gleb Gusev. Practical aspects of sensitivity in online experimentation with user engagement metrics. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 763–772, 2015c.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1097–1104, 2011.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Sample-efficient nonstationary-policy evaluation for contextual bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 247–254, 2012.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pages 563–587, 2015.

Georges Dupret and Mounia Lalmas. Absence time and user engagement: Evaluating ranking functions. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 173–182, 2013.

Georges Dupret and Ciya Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 181–190, 2010.

Georges E. Dupret and Benjamin Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 331–338, 2008.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman & Hall, 1993.

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 ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 871–880, 2012.

Henry Allen Feild, James Allan, and Rosie Jones. Predicting searcher frustration. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 34–41, 2010.

Henry Allen Feild, James Allan, and Joshua Glatt. Crowdlogging: distributed, private, and anonymous search logging. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 375–384, 2011.

Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan T. Dumais, and Thomas White. Evaluating implicit measures to improve Web search. *Transactions on Information System (TOIS)*, 23(2):147–168, 2005.

Hiroshi Fujimoto, Minoru Etoh, Akira Kinno, and Yoshikazu Akinaga. Web user profiling on proxy logs and its evaluation in personalization. In *Proceedings of the Asia-Pacific web conference on Web technologies and applications (APWeb)*, pages 107–118, 2011.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2006.

Andrew Gelman et al. Analysis of variance – why it is more important than ever. *The Annals of Statistics*, 33(1):1–53, 2005.

Laura A Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 478–479, 2004.

Artem Grotov, Shimon Whiteson, and Maarten de Rijke. Bayesian ranker comparison based on historical user interactions. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 273–282, 2015.

Zhiwei Guan and Edward Cutrell. What are you looking for? An eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 407–416, 2007a.

Zhiwei Guan and Edward Cutrell. An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 417–420, 2007b.

Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. Network A/B testing: From sampling to estimation. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 399–409, 2015.

Fan Guo, Chao Liu, Anitha Kannan, Tom Minka, Michael J. Taylor, Yi-Min Wang, and Christos Faloutsos. Click chain model in Web search. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 11–20, 2009a.

Fan Guo, Chao Liu, and Yi-Min Wang. Efficient multiple-click models in Web search. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 124–131, 2009b.

Qi Guo and Eugene Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 569–578, 2012.

Qi Guo, Shuai Yuan, and Eugene Agichtein. Detecting success in mobile search from interaction. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1229–1230, 2011.

Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 153–162, 2013a.

Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. Towards estimating web search result relevance from touch interactions on mobile devices. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 1821–1826, 2013b.

Jarrod D Hadfield et al. Mcmc methods for multi-response generalized linear mixed models: the mcmcglmm r package. *Journal of Statistical Software*, 33(2):1–22, 2010.

Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.

David Hardtke, Mike Wertheim, and Mark Cramer. Demonstration of improved search result relevancy using real-time implicit relevance feedback. *Understanding the User – Workshop in conjunction with SIGIR*, 2009.

Ahmed Hassan and Ryen W. White. Personalized models of search satisfaction. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 2009–2018, 2013.

Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 221–230, 2010.

Ahmed Hassan, Yang Song, and Li-Wei He. A task level user satisfaction metric and its application on improving relevance estimation. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 125–134, 2011.

Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 2019–2028, 2013.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition, 2009. 7th printing 2013 edition.

William Hersh, Andrew H. Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. Do batch and user evaluations give the same results? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 17–24, 2000.

Keisuke Hirano, Guido W. Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Katja Hofmann, Bouke Huurnink, Marc Bron, and Maarten de Rijke. Comparing click-through data to purchase decisions for retrieval evaluation. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 761–762, 2010.

Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Balancing exploration and exploitation in learning to rank online. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, volume 6611 of *Lecture Notes in Computer Science*, pages 251–263, 2011a.

Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. A probabilistic method for inferring preferences from clicks. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 249–258, 2011b.

Katja Hofmann, Fritz Behr, and Filip Radlinski. On caption bias in interleaving experiments. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 115–124, 2012a.

Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Estimating interleaved comparison outcomes from historical click data. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 1779–1783, 2012b.

Katja Hofmann, Anne Schuth, Shimon Whiteson, and Maarten de Rijke. Reusing historical interaction data for faster online learning to rank for IR. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 183–192, 2013a.

Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval. *Information Retrieval*, 16(1):63–90, 2013b.

Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. Fidelity, soundness, and efficiency of interleaved comparison methods. *Transactions on Information System (TOIS)*, 31(4), 2013c.

Katja Hofmann, Bhaskar Mitra, Filip Radlinski, and Milad Shokouhi. An eye-tracking study of user interactions with query auto completion. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 549–558, 2014.

Henning Hohnhold, Deirdre O'Brien, and Diane Tang. Focusing on the long-term: It's good for users and business. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1849–1858, 2015.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(6):945–960, 1986.

Yuan Hong, Jaideep Vaidya, Haibing Lu, Panagiotis Karras, and Sanjay Goel. Collaborative search log sanitization: Toward differential privacy and boosted utility. *IEEE Transactions on Dependable and Secure Computing*, 12(5):504–518, 2015.

Edward L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Bernard J. Jansen. Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 28(3):407 – 432, 2006.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *Transactions on Information System (TOIS)*, 20(4):422–446, 2002.

Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. Automatic online evaluation of intelligent personal assistants. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 506–516, 2015a.

Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. Automatic online evaluation of intelligent assistants. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 506–516, 2015b.

Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W. White. Understanding and predicting graded search satisfaction. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 57–66, 2015c.

Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *Transactions on Information System (TOIS)*, 25(2), 2007.

Ramesh Johari, Leo Pekelis, and David J. Walsh. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv Preprint arXiv:1512.04922v1 [math.ST]*, 2015.

Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 205–214, 2011.

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2): 1–224, 2009.

Diane Kelly and Leif Azzopardi. How many results per page? a study of SERP size, search behavior and user experience. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 183–192, 2015.

Diane Kelly and Karl Gyllstrom. An examination of two delivery modes for interactive search system experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1531–1540, 2011.

Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

Diane Kelly, Filip Radlinski, and Jaime Teevan. Choices and constraints: Research goals and approaches in information retrieval. *Tutorial at Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2014.

Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. Using historical click data to increase interleaving sensitivity. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 679–688, 2013.

Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. Optimised scheduling of online experiments. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 453–462, 2015a.

Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. Generalized team draft interleaving. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 773–782, 2015b.

Eugene Kharitonov, Aleksandr Vorobev, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. Sequential testing for early stopping of online experiments. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 473–482, 2015c.

Youngho Kim, Ahmed Hassan, Ryen White, and Imed Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 193–202, 2014a.

Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. Comparing client and server dwell time estimates for click-level satisfaction prediction. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 895–898, 2014b.

Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Imed Zitouni, Aidan C Crook, and Tasos Anastasakos. Predicting user satisfaction with intelligent assistants. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, 2016.

Jon Kleinberg. Temporal dynamics of on-line information streams. In Minos Garofalakis, Johannes Gehrke, and Rajeev Rastogi, editors, *Data Stream Management: Processing High-Speed Data Streams*. Springer, 2004.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 786–794, 2012.

Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. Online controlled experiments at large scale. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1168–1176, 2013.

Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. Seven rules of thumb for web site experimenters. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1857–1866, 2014.

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology.* SAGE Publications, 2012.

Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 113–122, 2014.

John Langford, Alexander Strehl, and Jennifer Wortman. Exploration scavenging. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 528–535, 2008.

John Lawson. *Design and Analysis of Experiments with R.* Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2014.

Jane Li, Scott Huffman, and Akihito Tokuda. Good abandonment in mobile and PC Internet search. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 43–50, 2009.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 661–670, 2010.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 297–306, 2011.

Lihong Li, Shunbao Chen, Ankur Gupta, and Jim Kleban. Counterfactual analysis of click metrics for search engine optimization: A case study. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 929–934, 2015a.

Lihong Li, Jin Kim, and Imed Zitouni. Toward predicting the outcome of an A/B experiment for search relevance. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 37–46, 2015b.

Lihong Li, Remi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 608–616, 2015c.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer-Verlag, 2001. ISBN 0387763694.

Winter Mason and Siddharth Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, 2012.

Nicolaas Matthijs and Filip Radlinski. Personalizing web search using long term browsing history. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 25–34, 2011.

T. Minka, J.M. Winn, J.P. Guiver, S. Webster, Y. Zaykov, B. Yangel, A. Spengler, and J. Bronskill. Infer.NET 2.6, 2014. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

Taesup Moon, Wei Chu, Lihong Li, Zhaohui Zheng, and Yi Chang. An online learning framework for refining recency search results with user click feedback. *Transactions on Information System (TOIS)*, 30(4), 2012.

Masahiro Morita and Yoichi Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 272–281, 1994.

Susan A. Murphy, Mark van der Laan, and James M. Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Guillermo Navarro-Arribas, Vicenç Torra, Arnau Erola, and Jordi Castellà-Roca. User k-anonymity for privacy preserving data mining of query logs. *Information Processing & Management*, 48(3):476–487, 2012.

Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

Olivier Nicol, Jérémie Mary, and Philippe Preux. Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 172–180, 2014.

Kirill Nikolaev, Alexey Drutsa, Ekaterina Gladkikh, Alexander Ulianov, Gleb Gusev, and Pavel Serdyukov. Extreme states distribution decomposition method for search engine online evaluation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 845–854, 2015.

Daan Odijk, Ryen W. White, Ahmed Hassan Awadallah, and Susan T. Dumais. Struggling and success in web search. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 1551–1560, 2015.

Umut Ozertem, Rosie Jones, and Benoit Dumoulin. Evaluating new search engine configurations with pre-existing judgments and clicks. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 397–406, 2011.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Ashok Kumar Ponnuswami, Kumaresh Pattabiraman, Desmond Brand, and Tapas Kanungo. Model characterization curves for federated search using click-logs: Predicting user engagement metrics for the span of feasible operating points. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 67–76, 2011.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 759–766, 2000.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, 2010.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, 2008.

Filip Radlinski and Nick Craswell. Comparing the sensitivity of information retrieval metrics. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 667–674, 2010.

Filip Radlinski and Nick Craswell. Optimized Interleaving for Online Retrieval Evaluation. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 245–254, 2013.

Filip Radlinski and Thorsten Joachims. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 1406–1412, 2006.

Filip Radlinski, Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. Optimizing relevance and revenue in ad search: A query substitution approach. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 403–410, 2008a.

Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 784–791, 2008b.

Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 43–52, 2008c.

Andrea Rotnitzky and James M. Robins. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4):805–820, 1995.

Mark Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

Falk Scholer, Milad Shokouhi, Bodo Billerbeck, and Andrew Turpin. Using Clicks as Implicit Judgments: Expectations Versus Observations. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pages 28–39, 2008.

Anne Schuth, Floor Sietsma, Shimon Whiteson, Damien Lefortier, and Maarten de Rijke. Multileaved comparisons for fast online evaluation. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 71–80, 2014.

Anne Schuth, Robert-Jan Bruintjes, Fritjof Buüttner, Joost van Doorn, Carla Groenland, Harrie Oosterhuis, Cong-Nguyen Tran, Bas Veeling, Jos van der Velde, Roger Wechsler, et al. Probabilistic multileave for online retrieval evaluation. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 955–958, 2015a.

Anne Schuth, Katja Hofmann, and Filip Radlinski. Predicting search satisfaction metrics with interleaved comparisons. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 463–472, 2015b.

William R. Shadish, Thomas D. Cook, and Donald Thomas Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning, 2002.

Fabrizio Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1–2):1–174, 2010.

Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–241, 1951.

Aleksandrs Slivkins, Filip Radlinski, and Sreenivas Gollapudi. Ranked bandits in metric spaces: Learning diverse rankings over large document collections. *Journal of Machine Learning Research (JMLR)*, 14(1):399–436, 2013.

Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 1201–1212, 2013a.

Yang Song, Xiaolin Shi, and Xin Fu. Evaluating and predicting user engagement change with degraded search relevance. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 1213–1224, 2013b.

Alexander L. Strehl, John Langford, Lihong Li, and Sham M. Kakade. Learning from logged implicit exploration data. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 2217–2225, 2011.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 814–823, 2015a.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 3231–3239, 2015b.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation, 2016. arXiv:1605.04812.

B.G. Tabachnick and L.S. Fidell. *Using Multivariate Statistics: Pearson New International Edition*. Pearson Education Limited, 2013.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 17–26, 2010.

Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 1587–1594, 2013.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2380–2388, 2015.

Edward R Tufte and PR Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 11–18, 2006.

Andrew H. Turpin and William Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 225–231, 2001.

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 329–337, 2013.

Antony Unwin. *Graphical Data Analysis with R*, volume 27. CRC Press, 2015.

Gerald van Belle. *Statistical Rules of Thumb*. Wiley-Blackwell, 2008.

Ellen M Voorhees. The effect of sampling strategy on inferred measures. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1119–1122, 2014.

Ellen M Voorhees and Donna K Harman. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press, 2005.

Dylan Walker and Lev Muchnik. Design of randomized experiments in networks. *Proceedings of the IEEE*, 102(12):1940–1951, 2014.

Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ahmed Hassan, and Ryen White. Modeling action-level satisfaction for search task satisfaction prediction. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 123–132, 2014.

Kuansan Wang, Toby Walker, and Zijian Zheng. PSkip: estimating relevance ranking quality from web search clickthrough data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1355–1364, 2009.

Kuansan Wang, Nikolas Gloy, and Xiaolong Li. Inferring search behaviors using partially observable markov (pom) model. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 211–220, 2010.

Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. Detecting good abandonment in mobile search. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 495–505, 2016.

Dragomir Yankov, Pavel Berkhin, and Lihong Li. Evaluation of explore-exploit policies in multi-result ranking systems. Technical Report MSR-TR-2015-34, Microsoft Research, 2015.

Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. Relevance and Effort: An Analysis of Document Utility. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pages 91–100, 2014.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1201–1208, 2009.

Yisong Yue, Yue Gao, Oliver Chapelle, Ya Zhang, and Thorsten Joachims. Learning more powerful test statistics for click-based retrieval evaluation. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 507–514, 2010a.

Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the International World Wide Web Conference (WWW)*, pages 1011–1018, 2010b.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Yuchen Zhang, Weizhu Chen, Dong Wang, and Qiang Yang. User-click modeling for understanding and predicting search-behavior. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1388–1396, 2011.

Dengyong Zhou, Qiang Liu, John C. Platt, and Christopher Meek. Aggregating ordinal labels from crowds by minimax conditional entropy. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 262–270, 2014.

Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 307–314, 1998.

Masrour Zoghi, Shimon Whiteson, and Maarten de Rijke. Mergerucb: A method for large-scale online ranker evaluation. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pages 17–26, 2015.

Masrour Zoghi, Tomáš Tunys, Lihong Li, Damien Jose, Junyan Chen, Chun Ming Chin, and Maarten de Rijke. Click-based hot fixes for underperforming torso queries. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, 2016.