Effective Online Evaluation for Web Search

Alexey Drutsa
Yandex; Moscow, Russia
adrutsa@yandex.ru

Gleb Gusev
Yandex; Moscow, Russia
gleb57@yandex-team.ru

Eugene Kharitonov
Facebook AI Research; Paris, France
eugene.kharitonov@gmail.com

Denis Kulemyakin
Yandex; Moscow, Russia
kulemyakin@yandex-team.ru

Pavel Serdyukov
Yandex; Moscow, Russia
pavser@yandex-team.ru

Igor Yashkov
Yandex; Moscow, Russia
excel@yandex-team.ru

ABSTRACT

We present you a program of a balanced mix between an overview of academic achievements in the field of online evaluation and a portion of unique industrial practical experience shared by both the leading researchers and engineers from global Internet companies. First, we give basic knowledge from mathematical statistics. This is followed by foundations of main evaluation methods such as A/B testing, interleaving, and observational studies. Then, we share rich industrial experiences on constructing of an experimentation pipeline and evaluation metrics (emphasizing best practices and common pitfalls). A large part of our tutorial is devoted to modern and state-of-the-art techniques (including the ones based on machine learning) that allow to conduct online experimentation efficiently. We invite software engineers, designers, analysts, and managers of web services and software products, as well as beginners, advanced specialists, and researchers to learn how to make web service development effectively data-driven.

CCS CONCEPTS

• **General and reference** → **Metrics**; **Evaluation**; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Information systems** → **Web log analysis**.

KEYWORDS

online evaluation; A/B tests; online metrics; interleaving

ACM Reference Format:
Alexey Drutsa, Gleb Gusev, Eugene Kharitonov, Denis Kulemyakin, Pavel Serdyukov, and Igor Yashkov. 2019. Effective Online Evaluation for Web Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3331184.3331378

Evaluation in the online setting constitutes one of the crucial steps in the development process of IR systems that are shipped for real use [50]. In particular, online evaluation allows to make data-driven decisions and is adopted by most global search engines (e.g., Bing [43], Google [32], Yandex [8, 24, 37], etc.), social networks [5, 67], and media providers [66]. Application of online evaluation techniques grew considerably over the last years: from 200 run experiments per day in 2013 (Bing [44]) to more than 1000 ones in 2015 (Google [32]). The number of smaller companies that use A/B testing in the development cycle of their products grows as well. The development of such services strongly depends on the quality of the experimentation platforms. This tutorial is intended to overview the state-of-the-art methods underlying the everyday evaluation pipelines and to reveal how to effectively conduct online evaluation of a web service during its development.

The objectives of our tutorial are:

(1) an introduction to online evaluation that positions the topic among related areas and gives necessary knowledge from mathematical statistics;
(2) description of main evaluation methods such as A/B testing, interleaving, and observational studies;
(3) sharing of rich industrial experiences on constructing of an experimentation pipeline and evaluation metrics (including best practices and pitfalls);
(4) discussion of advanced state-of-the-art techniques (including the ones based on machine learning) that are used to improve online experimentation and to conduct it efficiently.

By the end of the tutorial, attendees will be familiar with (a) state-of-the-art online evaluation approaches; (b) methodologies used to develop and effectively manage a production large-scale experimental pipeline; (c) techniques to construct, evaluate, and improve online metrics; and (d) machine learning approaches to improve efficiency of online evaluation.

This is the third version of the tutorial that have already been presented at The Web Conference (former WWW) 2018 [6] and KDD 2018 [7] by 5 of the co-authors of this tutorial. The tutorial is improved by new fresh practical examples/techniques and w.r.t. obtained feedback. There are also tutorials on A/B testing conducted by Microsoft at SIGIR 2017 and KDD 2017 [20]. In contrast to those tutorials, we (a) discuss machine learning methods to make online evaluation effective; (b) present also interleaving (an alternative online evaluation approach used in web search); and (c) provide management methodologies used in a production large-scale experimental pipeline (such as pre-launch checklists and a team of Experts on Experiments).

The tutorial materials (slides) are available at https://research.yandex.com/tutorials/online-evaluation/sigir-2019. The list of the most **relevant references** is presented below.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331378

REFERENCES

[1] Vineet Abhishek and Shie Mannor. 2017. A nonparametric sequential test for online randomized experiments. In *WWW'2017 Companion.* 610–616.

[2] Olga Arkhipova, Lidia Grauer, Igor Kuralenok, and Pavel Serdyukov. 2015. Search Engine Evaluation based on Search Engine Switching Prediction. In *SIGIR'2015*. 723–726.

[3] Susan Athey and Guido Imbens. 2015. Machine Learning Methods for Estimating Heterogeneous Causal Effects. *arXiv preprint arXiv:1504.01132* (2015).

[4] Juliette Aurisset, Michael Ramm, and Joshua Parks. 2017. Innovating Faster on Personalization Algorithms at Netflix Using Interleaving. https://medium.com/netflix-techblog/interleaving-in-online-experiments-at-netflix-a04ee392ec55.

[5] Eytan Bakshy and Dean Eckles. 2013. Uncertainty in online experiments with dependent data: An evaluation of bootstrap methods. In *KDD'2013*. 1303–1311.

[6] Roman Budylin, Alexey Drutsa, Gleb Gusev, Eugene Kharitonov, Pavel Serdyukov, and Igor Yashkov. 2018. Online Evaluation for Effective Web Service Development: Extended Abstract of the Tutorial at TheWebConf'2018.

[7] Roman Budylin, Alexey Drutsa, Gleb Gusev, Pavel Serdyukov, and Igor Yashkov. 2018. Online evaluation for effective web service development. In *arXiv preprint arXiv:1809.00661*. Tutorial at KDD'2018.

[8] Roman Budylin, Alexey Drutsa, Ilya Katsev, and Valeriya Tsoy. 2018. Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 55–63.

[9] Sunandan Chakraborty, Filip Radlinski, Milad Shokouhi, and Paul Baecke. 2014. On correlation of absence time and search effectiveness. In *SIGIR'2014*. 1163–1166.

[10] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. 2012. Large-scale validation and analysis of interleaved search evaluation. *TOIS* 30, 1 (2012), 6.

[11] Shuchi Chawla, Jason Hartline, and Denis Nekipelov. 2016. A/B testing of auctions. In *EC'2016*.

[12] Dominic Coey and Michael Bailey. 2016. People and cookies: Imperfect treatment assignment in online experiments. In *WWW'2016*. 1103–1111.

[13] Thomas Crook, Brian Frasca, Ron Kohavi, and Roger Longbotham. 2009. Seven pitfalls to avoid when running controlled experiments on the web. In *KDD'2009*.

[14] Alex Deng. 2015. Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments. In *WWW'2015 Companion*. 923–928.

[15] Alex Deng and Victor Hu. 2015. Diluted Treatment Effect Estimation for Trigger Analysis in Online Controlled Experiments. In *WSDM'2015*. 349–358.

[16] Alex Deng, Tianxi Li, and Yu Guo. 2014. Statistical inference in two-stage online controlled experiments with treatment selection and validation. In *WWW'2014*.

[17] Alex Deng, Jiannan Lu, and Shouyuan Chen. 2016. Continuous Monitoring of A/B Tests without Pain: Optional Stopping in Bayesian Testing. In *DSAA'2016*.

[18] Alex Deng and Xiaolin Shi. 2016. Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned. In *KDD'2016*.

[19] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *WSDM'2013*.

[20] Pavel Dmitriev, Somit Gupta, Ron Kohavi, Alex Deng, Paul Raff, and Lukas Vermeer. 2017. A/B Testing at Scale. https://exp-platform.com/2017abtestingtutorial/

[21] Pavel Dmitriev and Xian Wu. 2016. Measuring Metrics. In *CIKM'2016*. 429–437.

[22] Alexey Drutsa. 2015. Sign-Aware Periodicity Metrics of User Engagement for Online Search Quality Evaluation. In *SIGIR'2015*. 779–782.

[23] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Engagement Periodicity in Search Engine Usage: Analysis and Its Application to Search Quality Evaluation. In *WSDM'2015*. 27–36.

[24] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Future User Engagement Prediction and its Application to Improve the Sensitivity of Online Experiments. In *WWW'2015*. 256–266.

[25] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2017. Periodicity in User Engagement with a Search Engine and its Application to Online Controlled Experiments. *ACM Transactions on the Web (TWEB)* 11 (2017).

[26] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2017. Using the Delay in a Treatment Effect to Improve Sensitivity and Preserve Directionality of Engagement Metrics in A/B Experiments. In *WWW'2017*.

[27] Alexey Drutsa, Anna Ufland, and Gleb Gusev. 2015. Practical Aspects of Sensitivity in Online Experimentation with User Engagement Metrics. In *CIKM'2015*. 763–772.

[28] Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

[29] Artem Grotov and Maarten de Rijke. 2016. Online learning to rank for information retrieval: Tutorial. In *SIGIR*.

[30] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. 2015. Network a/b testing: From sampling to estimation. In *WWW'2015*. 399–409.

[31] Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. 2011. A probabilistic method for inferring preferences from clicks. In *CIKM'2011*. 249–258.

[32] Henning Hohnhold, Deirdre O'Brien, and Diane Tang. 2015. Focusing on the Long-term: It's Good for Users and Business. In *KDD'2015*. 1849–1858.

[33] Thorsten Joachims. 2002. Unbiased evaluation of retrieval quality using clickthrough data. (2002).

[34] Thorsten Joachims et al. 2003. Evaluating Retrieval Performance Using Clickthrough Data.

[35] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. Peeking at A/B Tests: Why it matters, and what to do about it. In *KDD'2017*. 1517–1525.

[36] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM*. 699–708.

[37] Eugene Kharitonov, Alexey Drutsa, and Pavel Serdyukov. 2017. Learning Sensitive Combinations of A/B Test Metrics. In *WSDM'2017*.

[38] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2015. Generalized Team Draft Interleaving. In *CIKM'2015*.

[39] Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2015. Optimised Scheduling of Online Experiments. In *SIGIR'2015*. 453–462.

[40] Eugene Kharitonov, Aleksandr Vorobev, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. 2015. Sequential Testing for Early Stopping of Online Experiments. In *SIGIR'2015*. 473–482.

[41] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. 2014. Modeling dwell time to predict click-level satisfaction. In *WSDM'2014*. 193–202.

[42] Ronny Kohavi, Thomas Crook, Roger Longbotham, Brian Frasca, Randy Henne, Juan Lavista Ferres, and Tamir Melamed. 2009. Online experimentation at Microsoft. *Data Mining Case Studies* (2009), 11.

[43] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *KDD'2012*. 786–794.

[44] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *KDD'2013*. 1168–1176.

[45] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu. 2014. Seven Rules of Thumb for Web Site Experimenters. In *KDD'2014*.

[46] Ron Kohavi, Randal M Henne, and Dan Sommerfield. 2007. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *KDD'2007*. 959–967.

[47] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Discov.* 18, 1 (2009), 140–181.

[48] Ron Kohavi, David Messner, Seth Eliot, Juan Lavista Ferres, Randy Henne, Vignesh Kannappan, and Justin Wang. 2010. Tracking Users' Clicks and Submits: Tradeoffs between User Experience and Data Loss.

[49] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. 2014. Measuring user engagement. *SLICRS* 6, 4 (2014), 1–132.

[50] Ilya Markov and Maarten de Rijke. 2019. What Should We Teach in Information Retrieval?. In *ACM SIGIR Forum*, Vol. 52. 19–39.

[51] Kirill Nikolaev, Alexey Drutsa, Ekaterina Gladkikh, Alexander Ulianov, Gleb Gusev, and Pavel Serdyukov. 2015. Extreme States Distribution Decomposition Method for Search Engine Online Evaluation. In *KDD'2015*. 845–854.

[52] Eric T Peterson. 2004. *Web analytics demystified: a marketer's guide to understanding how your web site affects your business*. Ingram.

[53] Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. 2017. Some methods for heterogeneous treatment effect estimation in high-dimensions. *arXiv preprint arXiv:1707.00102* (2017).

[54] Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. 2016. Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments. In *KDD'2016*. 235–244.

[55] Filip Radlinski. 2013. Sensitive Online Search Evaluation. http://irsg.bcs.org/SearchSolutions/2013/presentations/radlinski.pdf.

[56] Filip Radlinski and Nick Craswell. 2013. Optimized interleaving for online retrieval evaluation. In *WSDM*.

[57] Filip Radlinski and Katja Hofmann. 2013. Practical online retrieval evaluation. In *ECIR*.

[58] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How does clickthrough data reflect retrieval quality?. In *CIKM'2008*. 43–52.

[59] Filip Radlinski and Yisong Yue. 2011. Practical Online Retrieval Evaluation. In *SIGIR*.

[60] Kerry Rodden, Hilary Hutchinson, and Xin Fu. 2010. Measuring the user experience on a large scale: user-centered metrics for web applications. In *CHI'2010*.

[61] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. 2017. Detecting network effects: Randomizing over randomized experiments. In *KDD'2017*. 1027–1035.

[62] Anne Schuth, Floor Sietsma, Shimon Whiteson, Damien Lefortier, and Maarten de Rijke. 2014. Multileaved comparisons for fast online evaluation. In *CIKM*.

[63] Milad Shokouhi. 2011. Detecting seasonal queries by time-series analysis. In *SIGIR'2011*. 1171–1172.

[64] Yang Song, Xiaolin Shi, and Xin Fu. 2013. Evaluating and predicting user engagement change with degraded search relevance. In *WWW'2013*. 1213–1224.

[65] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: More, better, faster experimentation. In *KDD*.

[66] Huizhi Xie and Juliette Aurisset. 2016. Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix. In *KDD'2016*.

[67] Ya Xu and Nanyu Chen. 2016. Evaluating Mobile Apps with A/B and Quasi A/B Tests. In *KDD'2016*.

[68] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From infrastructure to culture: A/B testing challenges in large scale social networks. In *KDD'2015*.