Machine Learning Powered A/B Testing

Pavel Serdyukov
Yandex
Lva Tolstogo 16
Moscow, Russia
pavser@yandex-team.ru

ABSTRACT

Online search evaluation, and A/B testing in particular, is an irreplaceable tool for modern search engines. Typically, online experiments last for several days or weeks and require a considerable portion of the search traffic. Despite the increasing need for running more experiments, the amount of that traffic is limited. This situation leads to the problem of finding new key performance metrics with higher sensitivity and lower variance. Recently, we proposed a number of techniques to alleviate this need for larger sample sizes in A/B experiments.

One approach was based on formulating the quest for finding a sensitive metric as a data-driven machine learning problem of finding a sensitive metric combination [2]. We assumed that each single observation in these experiments is assigned with a vector of metrics (features) describing it. After that, we learned a linear combination of these metrics, such that the learned combination can be considered as a metric itself, and (a) agrees with the preference direction in the seed experiments according to a baseline ground truth metric, (b) achieves a higher sensitivity than the baseline ground-truth metric.

Another approach addressed the problem of delays in the treatment effects causing low sensitivity of the metrics and requiring to conduct A/B experiments with longer duration or larger set of users from a limited traffic [1]. We found that a delayed treatment effect of a metric could be revealed through the daily time series of the metricâĂŹs measurements over the days of an A/B test. So, we proposed several metrics that learn the models of the trend in such time series and use them to quantify the changes in the user behavior.

Finally, in another study [3], we addressed the problem of variance reduction for user engagement metrics and developed a general framework that allows us to incorporate both the existing state-of-the-art approaches to reduce the variance and some novel ones based on advanced machine learning techniques. The expected value of the key metric for a given user consists of two components: (1) the expected value for this user irrespectively the treatment assignment and (2) the treatment effect for this user. The expectation of the 1st component does not depend on the treatment assignment and does not contribute to the actual average treatment effect, but may increase the variance of its estimation. If we knew the value of the first component, we would subtract it from the key metric and obtain a new metric with decreased variance. However, since we cannot evaluate the first component exactly, we propose to predict it based on the attributes of the user that are independent of the treatment exposure. Therefore, we propose to utilize, instead of the average value of a key metric, its average deviation from its predicted value. In this way, the problem of variance reduction is reduced to the problem of finding the best predictor for the key metric that is not aware of the treatment exposure. In our general approach, we apply gradient boosted decision trees and achieve a significantly greater variance reduction than the state-of-the-art.

KEYWORDS

Online metrics, online evaluation, A/B testing

1 BIO

Pavel Serdyukov is the Head of Research Projects at Yandex, where he manages a team of researchers working at the intersection of web search/mining and machine learning. He has published extensively in top-tier conferences on these topics (with over 10 papers on online evaluation among them published since 2015) and, in particular, co-authored papers that received Best Student Paper Awards at SIGIR 2015 and 2016. In the past, he co-organized a number of workshops at SIGIR, was a co-organizer of the Entity track at TREC 2009-2011, and co-organized a series of workshops at WSDM in 2012-2014 on search logs mining. He was also the General Chair of ECIR 2013 in Moscow and, recently, the Industry track chair at WWW 2017. Before joining Yandex in 2011, he was a postdoc at Delft University, got his PhD from Twente University (2009) and his MSc from Max-Planck Institute for Computer Science (2005).

REFERENCES

[1] 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'17*.

[2] Eugene Kharitonov, Alexey Drutsa, and Pavel Serdyukov. 2017. Learning Sensitive Combinations of A/B Test Metrics. In *WSDM'17*.

[3] 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'17*.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5022-8/17/08.
https://doi.org/10.1145/3077136.3096468