Search Engine Evaluation based on Search Engine Switching Prediction

Olga Arkhipova Lidia Grauer Igor Kuralenok Pavel Serdyukov
Yandex LLC
{olycha, lidia, solar, pavser}@yandex-team.ru

ABSTRACT

In this paper we present a novel application of the search engine switching prediction model for online evaluation. We propose a new metric pSwitch for A/B-testing, which allows us to evaluate the quality of search engines in different aspects such as the quality of the user interface and the quality of the ranking function. pSwitch is a search session-level metric, which relies on the predicted probability that the session contains a switch to another search engine and reflects the degree of the failure of the session. We demonstrate the effectiveness and validity of pSwitch using A/B-testing experiments with real users of search engine Yandex. We compare our metric with recently proposed SpU (sessions per user) metric and other widely used query-level A/B metrics, such as Abandonment Rate and Time to First Click, which we used as our baseline metrics. We observed that pSwitch metric is more sensitive in comparison with those baseline metrics and also that pSwitch and SpU are more consistent with ground truth, than Abandonment Rate and Time to First Click.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

Keywords: online evaluation; search engine switching

1. INTRODUCTION

Due to the intense competitiveness between commercial search engines, the correct evaluation of a search engine's quality is critical for its ability to constantly improve its quality and hence maintain its search market share. Since the impact of the tested improvements becomes relatively smaller with the time, sensitivity of evaluation measures becomes increasingly important. Two major approaches are used to evaluate search engine quality: Cranfield-style offline evaluation based on experts judgements [2] and online evaluation such as A/B-testing [6].

Despite the convenience of offline evaluation with document or search session labels provided by a group of experts, this approach has several considerable disadvantages, namely: expert judgements may not reflect the actual relevance, which is often user-specific; the process requires substantial costs and is also time consuming; some standard metrics do not necessarily correlate with user-centric performance measures. In contrast, the online evaluation approach, such as A/B testing, alleviates these disadvantages as it uses the implicit feedback of real users directly. In this case, the major challenge of online evaluation methods is establishing the relationship between the observed search behaviour and the relative quality of the search engines observing that behaviour. The idea of A/B-testing is the following: users are randomly assigned to two groups. During some period of time, the first (control) bucket of users is presented with the original version of the search engine (A), while the second (treatment) bucket of users is presented with the experimental version of the search engine (B). Successful search engine development requires making effective decisions towards improving all of its components, such as the ranking algorithm, the snippet generation algorithm, the ads ranking algorithm or the user interface. A/B testing, by design, allows to measure the effect of innovations in any of these components, but requires highly sensitive and easily interpretable metrics that would detect even subtle differences in the behavior of the users exposed to both versions of the system.

Many studies focused on the problem of predicting frustration or satisfaction of a searcher during her search session [3], but, to the best of our knowledge, there were no attempts to study the utility of such predictive models for the purposes of A/B testing. In this paper, we present a new metric for evaluating the quality of a search engine that can be used in A/B-testing and which is based on the prediction of the proportion of unsuccessful user sessions among all user sessions. We argue that this is possible by predicting if the user had a reason to switch to another search engine for each search session in each bucket. Indeed, in 57% of switches from one search engine to another the cause of the switch is a poor quality of search results [5]. However, low quality search results are not the only reason of switching, there are other reasons, including the need to verify the acquired knowledge or to find additional information (26%) [5]. Actually, any search engine aims to maximize its sufficiency and hence prevent switches to its competitors in all and, especially, in these 83% of cases, what means to improve relevance, authority and trustworthiness of its search results, as well as to increase the size of its document index. Although, search engine switches are good indicators of user dissatisfaction, search engines are able to observe them only for a small

SIGIR '15 August 09 - 13, 2015, Santiago, Chile

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767786.

Table 1: Characters assigned to actions. Characters with * were not presented in the final metric classifier

H	short click on an organic result (< 10 sec)
C	normal click on an organic result (< 60 sec)
L	long click on an organic result (≥ 60 sec)
S	skip of an organic result
O	query with a next action after a short time
V	query with a next action after a normal time
R	query with a next action after a long time
W *	click on a vertical search result
D *	click on an advertisement

share of users that are used to switch and whose web surfing activities can be tracked. To expand this to the entire audience, one needs to predict the potential switches: the switches that indeed occurred but were not detected or the switches that could occur if the users in the corresponding search sessions had the habit to use more than one search engine (there is always a considerable share of users that never uses more than one search engine [10])(see Section 2).

An online metric has to be applicable for comparing relative quality of modifications of search engines when they differ in different aspects such as ranking or user interface. We propose to use the probability that a session contains a switch to another search engine as such a metric (see Section 4.1).

In order to demonstrate that the proposed online metric is considerably more sensitive than the state-of-the-art online metrics [7, 9], we performed a series of A/B-testing experiments with the users of one of the most popular search engines (see Section 4).

2. SWITCHING PREDICTION MODEL

In this section we consider the problem of search engine switching prediction in a search session. We define a switch as an event of changing one search engine to another in order to continue the current search session. Actually, the fact of switching can be unambiguously detected only in a small part of the search sessions performed by users who installed the browser or the special browser toolbar plug-in developed by a search engine [10]. We used the Yandex browser/toolbar interaction logs collected in October 2014. For each session recorded in these logs, we could exactly detect the fact of switching from the search engine to other search engines. After that, we detected users, who had at least one switch from the search engine to other search engines in a given period. Then, we extracted a random sample of the search sessions of those "switching-tolerant" users from the period under study. Our final data set consisted of 224k search sessions, corresponding to 88k users. Further, we splited users into learn and test sets (1:1). In order to build a prediction model, we used a variation of stochastic gradient boosted decision trees [4]. The quality of the different versions of predictors that we describe in the paper was measured using Area Under Curve (AUC) measure.

Table 2: The performance of feature groups

feature group	**description**	**AUC with**	**AUC without**
Baseline	sequence based features	0.59	-
Baseline + "actions dwell time"	all features about time intervals between actions	0.617	0.728
Baseline + "query text"	query text features	**0.681**	0.673
Baseline + "user session number"	total number of user sessions during the day, current user session №	**0.66**	0.715
Baseline + "vertical results and ads appearance"	features about number of vertical results and ads appearance	0.601	0.727
Baseline + "session-level time"	session dwell time, session duration	0.623	0.727
Baseline + query text + session number	features from two groups	**0.7223**	**0.6547**

3. FEATURES SELECTION

We started from encoding each search session into an ordered set of characters representing a sequence of search actions. The alphabet for encoding is presented in Table 1. For example, using interaction logs, we encode the user session as follows: the user typed the query "toyota" and did not click anywhere. Seven seconds later, she reformulated her query "toyota celica" and, five seconds later, she clicked on the first result: image vertical search result. The next activity of that user happened after 2 minutes and it was a long click on the organic search result the third position. Using the full alphabet from Table 1, such a session would be encoded as (OOWSSL) and using the reduced version (without differentiation of clicked result types) as (OOLSSL). Our algorithm automatically extracts frequent subsequences of the characters that the sessions are represented by. Further, it uses these subsequences as binary features of the sessions. The classifier trained only with such sequence features (AUC = 0.59) was our baseline that we tried to improve with adding more features. Thus, we considered a number of session-level features, such as the duration of the session, the average duration of clicks, the level of advertising shown, etc (56 features). Every feature was included in the dataset with two normalizations, as in [8]. AUC of the classifier trained with the baseline features and all these additional features was 0.73. Feature selection is worth doing not only for evaluating feature performance, but also for selecting the optimal feature set and reducing the amount of the data to be stored during the production process. To select an optimal feature set, we grouped all features into the following groups: "actions dwell time" features, "query text"

Table 3: Top-10 features AUC

Minimum query length (characters) in the session	0.6607
Total number of user sessions during the given period (day)	0.6605
Maximum query length (characters) in the session	0.6481
Average query length (characters) in the session	0.6475
Current user session number during the given period (day)	0.6257
Average query length (words) in the session	0.6235
Minimum query length (words) in the session	0.6229
Maximum query length (words) in the session	0.6216
Time passed from the end of the previous user session before start of the current user session (user's absence time)	0.62
Share of unique clicked URLs among all clicked URLs per session	0.6173

features, "user session number" features, "vertical results and ads appearance" features and "session-level time" features. The descriptions of each feature group are presented in Table 2. We evaluated the utility of each of the feature groups in two ways: we trained a classifier only with the considered additional feature group and without this additional feature group (with baseline features always present). The results are presented in Table 2. Table 3 contains the top-10 most useful features across all the groups in terms of AUC. The most important features for switch prediction are the features from "query text" and "user session number" groups: combining them into one group gives us AUC = 0.722, so we select features only from these two groups to include into the final set of features (26 additional features).

4. SWITCH BASED EVALUATION METRIC

4.1 Metric description

Since search engine switching can be interpreted as an indication of the failure of the user's session, it can serve as a foundation for a search quality measure. We propose a new online metric for measuring and comparing the quality of the search engines via A/B-testing. Our metric is based on the prediction of the proportion of unsuccessful user sessions among all user sessions. Using the classsifier built in Section 3 we can predict the event of a switch for any user session and estimate the proportion of unsuccessful user sessions by the probability of switching averaged over all sessions in the experiment. We call our metric *pSwitch*:

$$pSwitch = \frac{1}{N}\sum_{i=1}^{N}\tilde{p}_{switch}(i)$$

Consider an A/B-testing experiment and suppose that the mean predicted switch probability in the experimental group is statistically significantly higher than the mean predicted switch probability in the control group: that indicates that the users in the experimental group are less satisfied. In order to measure the significance of the mean difference we used the bootstrap test [1] with sampling by users. We considered results to be significant in the case of p-value < 0.05.

4.2 Experimental setup

To analyze pSwitch, we conducted 8 online experiments testing different new features of search engine Yandex. All the experiments were manually inspected and labeled by a group of experts as degrading or improving (it means that we expected each feature to either degrade or improve the user experience). We regarded these labels as the ground truth that we expected our metrics to agree with. We also ran our experiments on 3 A/A-testing experiments [6] to control the validity of pSwitch: reliable online evaluation metrics must not signal about any differences in such experiments, as otherwise these metrics are not valid. All the A/A and A/B experiments were also evaluated by SpU (Sessions per User) [9], Abandonment Rate and Time to First Click, which we used as our baseline metrics. The users from the treatment buckets of the A/B-testing experiments were exposed to the following configurations of the search engine.

Experiments with vertical results (2 experiments) group of experiments consisted of one improvement and one degradation of vertical results.

Swap (1 experiment) In this case, we made the following changes of the original ranking: two random documents from positions 2-9 were swaped. We believe that the search engine generates rankings which are better than a random ordering on average, so we could assume that such swapping will lead to a degradation of the search quality.

Old ranking algorithm(1 experiment) In this case, we used a 1-year-old ranking algorithm instead of the current production algorithm. We treat it as a degradation of the ranking algorithm.

Ranking improvements (4 experiments) This group of experiments consisted of 4 improvements of the ranking algorithm.

In each experiment, the changes were substantial and had to be detected by an online evaluation metric that is supposed to be sensitive to the changes in the target aspects of search engine quality.

4.3 Experimental results

We evaluate the performance of the two pSwitch metrics: one based on the classifier with the full aplhabet and another based on the classifier with the reduced alphabet (without click type differentiation). The performances of the two types of pSwitch metric, SpU, Abandonment Rate and Time to First Click on the A/B dataset and A/A control dataset are presented in Table 4. P-values for SpU metric and pSwitch metrics were measured using the bootstrap test [1]. P-values for Time to First Click and Abandonment Rate are measured by Wilcoxon signed rank test. None of the metrics detected a significant difference in A/A experiments, so all metrics proved their validity for online evaluation. As we can see from Table 4, pSwitch metric with the reduced alphabet is much more sensitive than SpU metric and is more consistent with the ground truth then Time to First Click and Abandonment Rate. However, the outcome of the pSwitch metric based on the classifier with the full alphabet produced the result opposite to the ground truth for vertical results improvement experiment. The differences for other experimental groups are consistent with the ground truth. The reason for this contradiction is the inclusion into

Table 4: Experiment Results. Bootstrap by users. For each metric improvements are marked in green, degradations - in red, according to the metric diff. Expected color (the ground truth) is in the first column. "!" means that the metric diff is opposite to the ground truth. * - significance at 0.05, ** - at 0.01, * - at 0,001**

Experiment name	SpU pvalue	pSwitch pvalue (classifier with reduced alphabet)	pSwitch pvalue (classifier with full alphabet)	Abandonment Rate pvalue	Time to First Click pvalue
improvement 1	0.2662	0.0302 *	0,0228 *	0.04 (**!***)	0.43
improvement 2	0.0834	0.25	0,313	0.02 *	0.82
improvement 3	0.16	0.33	0,0604	0.35	0.01 (**!***)
improvement 4	0.0388 *	0.006 **	0,0016 **	0.12	0.78
vertical result improvement	0.0298 *	0.048 *	0,0144 (**!***)	0 ***	0 ***
vertical result degradation	0.6918	0.65	0 ***	0.73	0 (**!*****)
swap	0.2788	0.0016 **	0.0002 ***	0.001 **	0.18
old formula	0.4498	0.007 **	0.0168 *	0.001 **	0.91
A/A control 1	0.67	0.22	0.236	0.54	0.23
A/A control 2	0.48	0.63	0.62	0.14	0.23
A/A control 3	0.89	0.68	0.67	0.85	0.58
correct decisions ratio	0.25	**0.625**	0.625	0.5	0.125
incorrect decisions ratio	0	**0**	0.125	0.125	0.25

the alphabet those actions whose appearance in the sequence depends not only on the users preferences, but is also directly affected by the experiment's design. Clicks on vertical results (W) and ads (D) are good examples of such actions. For example, if there are no vertical results on the SERP by design, a user cannot click on them and therefore there will be no (W) action in the user action sequence, that could lead to wrong results. Our baseline metrics Time to First Click and Abandonment Rate also had experiments with results, which are opposite to the ground truth (two experiments for Time to First Click and one - Abandonment Rate). SpU and pSwitch (classifier with reduced alphabet) metrics did not contradict the ground truth in any of the experiments.

5. CONCLUSIONS

In this paper, we proposed a new metric called pSwitch for evaluating the quality of search engines based on search engine switching prediction. To predict the probability of switching in a particular session, we presented the classification model that used a rich set of aggregated session features and sequences of search actions as features as well. We selected an optimal feature set and evaluated the performance of groups of features. We also presented a new direction of using models predicting of user frustration and a success during search sessions in A/B-testing. To analyze our metric, we conducted A/B-testing experiments with real users of search engine Yandex. We draw the attention to the importance of the action coding selection and illustrated this on real experiments. Our findings demonstrated that the new metric can be applied to evaluation of different aspects of a search engine. pSwitch is more sensitive in comparison with the recently proposed SpU metric and is also more consistent with the ground truth than query-level metrics, such as Abandonment Rate and Time to First Click. Since our work is the first to study an online evaluation metric based on search engine switching prediction, a variety of its extensions can be considered in the future. For instance, since our proposed metric can be used for evaluating different types of search engine's changes, not only the ones used in our experiments, we plan to demonstrate such applicability in the future experiments. Another direction of future work is the improvement of the classification model, which serves as the basis of our metric and might benefit from using a broader set of features and alternative classification methods.

6. AKNOWLEDGMENTS

We would like to thank Irina Orlova for her great contribution to this work.

7. REFERENCES

[1] M. R. Chernick and R. A. LaBudde. *An Introduction to Bootstrap Methods with Applications to R*. Wiley Publishing, 1st edition, 2011.

[2] C. Cleverdon. Readings in information retrieval. chapter The Cranfield Tests on Index Language Devices, pages 47–59. Morgan Kaufmann Publishers Inc., 1997.

[3] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *Proc. SIGIR 2010*, pages 34–41. ACM.

[4] J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, Feb. 2002.

[5] Q. Guo, R. W. White, Y. Zhang, B. Anderson, and S. T. Dumais. Why searchers switch: Understanding and predicting engine switching rationales. In *Proceedings SIGIR 2011*, pages 335–344. ACM.

[6] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: Survey and practical guide. *Data Min. Knowl. Discov.*, 18(1):140–181, Feb. 2009.

[7] F. Radlinski and N. Craswell. Comparing the sensitivity of information retrieval metrics. In *Pr. SIGIR 2010*, pages 667–674. ACM.

[8] D. Savenkov, D. Lagun, and Q. Liu. Search engine switching detection based on user personal preferences and behavior patterns. In *Pr. SIGIR 2013*, pages 33–42. ACM.

[9] Y. Song, X. Shi, and X. Fu. Evaluating and predicting user engagement change with degraded search relevance. In *Pr. WWW 2013*, pages 1213–1224. International World Wide Web Conferences Steering Committee.

[10] R. W. White, A. Kapoor, and S. T. Dumais. Modeling long-term search engine usage. In *Pr. UMAP 2010*.