Evaluating Mobile Web Search Performance by Taking Good Abandonment into Account

Olga Arkhipova
Yandex LLC,
Saint Petersburg, Russia
olycha@yandex-team.ru

Lidia Grauer
Yandex LLC,
Saint Petersburg, Russia
lidia@yandex-team.ru

ABSTRACT

Usage of mobile devices for Web search grows rapidly in recent years. The common tendency is that users want to receive information immediately results in incorporating rich snippets and vertical results into search engine result pages (SERPs) and in increasing of good abandonment. This article provides an offline metric for quality evaluation of mobile Web search, which takes good abandonment rate into consideration. The metric is the DBN click model that allows the probability to be satisfied directly on the SERP. The model parameters are estimated from the mobile search logs of a controlled experiment. The new metric outperforms traditional ERR metric in terms of the validation dataset built using a SERP degradation technique.

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

Keywords: Click models; evaluation; information retrieval measures; user behavior; mobile web search

1. INTRODUCTION

The number of users that use search engines on their mobile devices grows rapidly due to mobile web search convenience and the opportunity to have "information at your fingertips", that it brings to its users. The increase of satisfaction of mobile web search users became an important task in information retrieval [5]. In mobile web search rich snippets and different vertical results represent a convenient summary of information about the search results at a glance, so a user does not need to click on SERP's results anymore. As the result is the growth of good abandonment rate for mobile devices. According to Li et al. [8], the potential good abandonments in mobiles is more than 50% of all abandoned queries compared with 30% for desktop searches. The common tendency for good abandonment is that its rate will be continuously increasing in the coming years, since modern search engines aim not only to generate good and attractive snippets, but also to show snippets with the actual answers to satisfy the user before any further clicks become necessary.

It was shown that user behaviour of conducting mobile Web search significantly differs from search behaviour on a desktop [7]. In [3] a number of click model-based offline metrics are constructed and compared with traditional offline metrics like ERR, DCG or Precision and online metrics. Also in [3] it was shown that metrics based on click models like Dynamic Bayesian Networks (DBN) [1] and User Browsing Model (UBM) [6] are better correlated with online measurements than traditional metrics. In order to model good abandonment we propose a new metric based on the DBN model, but allowing for the probability that the user finds the relevant information (for example, the correct answer) directly on the SERP (Section 2). To evaluate parameters of our metric we conducted a controlled user experiment on Android devices (Section 3). One of the methods for comparing information retrieval metrics is the swap method proposed by Voorhees and Buckley [10] and extended by Sakai [9]. They evaluate metrics consistency comparing pairs of systems without knowing whether a system is better than another. But this method could prefer a metric inconsistent with some online experiment results. Since the agreement between offline and online evaluation is the desirable result for a new offline metric [3], we are interested whether the offline metrics decision is consistent with the online experiment's results. Unlike Sakai and Voorhees, Buckley, we compare performance of the proposed metric with ERR metric [4] on pairs of systems with predefined system differences using a SERP degradation technique (Section 4).

2. MODEL DESCRIPTION

We propose a click model based on the Dynamic Bayesian Network [1], but which accounts for the probability that the user finds the correct answer directly in a snippet and which also assumes that the probability to get tired and stop search before finding some relevant information depends on the action on the last examined snippet (click or examination). The reasons to consider the probability to get the correct answer from a snippet are the following: at first, queries indicating good abandonment represent a large subset of all abandoned queries [8]. The second, good abandonment rate on mobiles search is significantly higher than on PC search [8]. The proposed model is described below. For a given position i of the SERP the following variables are defined:

- E_i - a label indicating the user's examination of the snippet i
- C_i - a label indicating the user's click on the snippet i
- $End+_i$ - a label indicating that the user found an cor-

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609505.

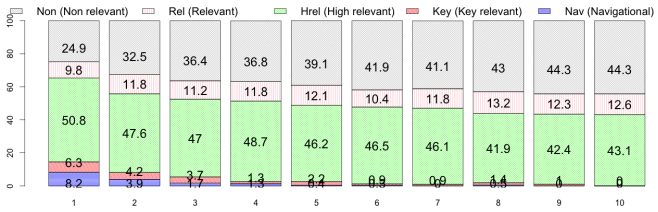

Figure 1: Distribution of relevance labels by rank for four commercial search engines

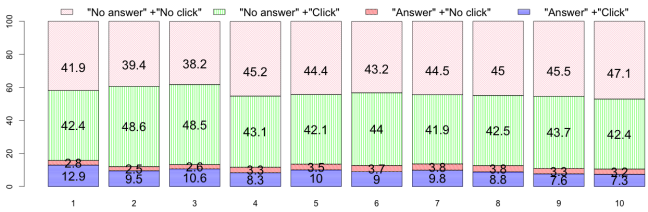

Figure 2: Distribution of attractiveness labels by rank for four commercial search engines

rect answer at i-th position

- $End{-}_i$ - a label indicating that user stops unsatisfied at i-th position

The following equations describe the model:

$$P(E_i = 1 | E_j = 0, i > j) = 0 \quad (1)$$

$$P(E_1 = 1) = 1 \quad (2)$$

$$P(C_i = 1 | E_i = 0) = 0 \quad (3)$$

$$P(End +_i | E_i = 0) = 0 \quad (4)$$

$$P(End +_i | E_i = 1, C_i = 0) = sa_i \quad (5)$$

$$P(C_i | E_i = 1) = ac_i \quad (6)$$

$$P(End +_i | C_i = 1) = s_i \quad (7)$$

$$P(End -_i | E_i = 1, C_i = 0) = 1 - y_1 \quad (8)$$

$$P(End -_i | C_i = 1) = 1 - y_2 \quad (9)$$

Unlike the DBN model our model has variables sa_i (5), which represent the probability of finding the correct answer in the corresponding snippet, and two variables y_1 (8) and y_2 (9), representing the probabilities to continue the search being unsatisfied after only examining a snippet and after a click on a snippet. The inclusion of different y variables in case of click and examination is based on the assumption that navigating to a non-relevant document is more difficult on mobile device and is more irritating for a user compared with only an examination of such a document. As in the DBN model [1] we assume that the user browses SERP from the top to the bottom (1) and she always examines the first snippet of the SERP (2). There is no chance to click on a snippet without its examination (3) and no chance to be satisfied without the snippet examination (4). After the user examines a snippet i, there is a probability sa_i to be satisfied by the answer from this snippet (5) and there is a certain probability ac_i to be attracted by the snippet and to click on it (6). If the user finds the snippet not attractive and decides not to click on it, there is probability $1 - y_1$ to abandon search after examination (8). After the user clicks and visits a document, there is a certain probability s_i that she will be satisfied by this document (7). Once the user is satisfied by the document she has visited, she stops her search. If the user is not satisfied by the current result, there is certain probability $1 - y_2$ to abandon the search(9). Parameters sa_i, ac_i depend on snippet attractiveness and s_i depends on document relevance.

3. PARAMETERS ESTIMATION

3.1 Controlled data collection

To build a dataset for parameters evaluation we conducted a controlled user experiment. The experiment was conducted in the mobile Android devices in May 2013. To collect user data, we developed a plugin for the Dolphin[1] browser for Androids. Ten volunteered (unpaid) participants were recruited for this user study. All volunteers were undergraduate and graduate students and all had some experience with Web search on mobile devices with touch screens. In order to create a collection of mobile search tasks, we randomly sampled 200 unique queries with clicked documents from the Yandex mobile search query logs. Based on them we prepared 200 search tasks similar to mobile Web search users tasks. Below there are some examples of them:

- You want to read a short biography of Henry VIII
- You want to know what Proboscis monkey looks like
- You want to listen to Bob Marley's song "No woman, no cry"

Participants had to complete these search tasks in four commercial search engines: Yandex, Google, Bing, Mail. Every subject had completed from 10 to 186 tasks. To begin each task the participants were presented with a task description. More than one query per task was allowed and there were no restrictions on time or participants actions on search pages. The participant's actions such as queries, clicks, scrolls, touches etc. were logged. Each time the participant completed the task, she was asked (through a popup window) whether or not she solved the task.

Eventually, each query was labeled as the one which led to user's satisfaction or not (End+/End-). If the user issued more than one query to complete her task, the last query could have End+ label, and all previous queries were always labeled as End-. In total we obtained 809 queries with user actions and labels, and 530 clicks were done on the SERPs. The dataset for model parameters estimation was composed of the queries, corresponding SERP documents (including organic and vertical results), users' actions on every document (click or examination) and users' labels to queries. For every SERP from the dataset the relevance labels were also collected. Those labels were separately collected for documents (relevance judgments), their snippets (attractiveness judgments). Navigating convenience on mobile device was taken into consideration in relevance judgements, i.e. relevant but hardly accessible documents were labeled less relevant. The following relevance scale from [4] with the exception of "Junk" label were used: "Nav", "Key", "HRel", "Rel", "Non". Figure 1 shows the distribution of relevance judgements for four commercial search engines by document ranks. To judge every snippet a subject had to answer following questions:

- Is there a correct answer to the query in the snippet? ("Answer", "No answer")
- Would I click on the snippet to find more information? ("Click", "No click")

[1] http://dolphin-browser.com/

Table 1: Percent of times when metric shows significant difference between control group (Initial SERPs) and test group (SERP degradation)

	N=200		N=500		N=800		N=1000	
SERP degradation	ERR	Psat	ERR	Psat	ERR	Psat	ERR	Psat
50% of queries with snippet from middle	0.0	19.1	0.0	51	0.0	75.1	0.0	87.7
50% of queries with random snippet	0.0	12.4	0.0	36.,9	0.0	62.4	0.0	74
80% of queries with snippet from middle	0.0	35.2	0.0	79.5	0.0	96.6	0.0	99
80% of queries with random snippet	0.0	13	0.0	47.1	0.0	70.2	0.0	80.5
100% of queries with snippet from middle	0.0	38.5	0.0	85.8	0.0	97.8	0.0	99.1
100% of queries with random snippet	0.0	14.5	0.0	44.5	0.0	68.7	0.0	81.3
bad vertical image result snippet	0.0	57	0.0	97.5	0.0	100.0	0.0	100,0
bad vertical image result snippet + document	**96.0***	90,4	100.0	100.0	100.0	100.0	100.0	100.0
removing vertical image results	96.4	**100.0***	100,0	100.0	100.0	100.0	100.0	100.0
remove all vertical results	63.7	**99.6***	99.2	**100.0***	100.0	100.0	100.0	100.0
for 50% of queries remove all vertical results	16.5	**84.9***	56,8	**100.0***	82,2	**100.0***	89,7	**100.0***
swap[2;4][5;7] (1 document)	**31.2**	26.5	77.9	**91.9***	91.2	**99.7***	**99.7**	99.1
swap[2;4]	27.6	**29.6**	71.3	**79.2***	91.9	**95.4***	96.6	**98.1**
removing answer from snippet		100.0		100.0		100.0		100.0

Figure 2 shows distribution of snippet judgements for four commercial search engines by document rank. In comparison to relevance judgements, distribution of snippet labels per document rank does not so clearly depend on the rank itself. Interestingly, judgements "Answer" represent only 10% of all judgements.

3.2 Psat Measure

For every SERP from the dataset we have a sequence of clicks, the user's success label and documents and snippets relevance judgements. Using the assumption that browsing goes from the top to the bottom and the fact that user cannot be satisfied with the answer from the snippet labeled as "No answer", at the same time we know the meaning of all possible user action sequences on a judged SERP. By taking this fact into consideration we estimated the parameters of the model using MLE method. For example, in the case of a three-document SERP with the following snippet judgements: "No answer" + "Click", "Answer" + "No click" and "No answer" + "No click", the user's label "End+" and a click on the first document we could have only these two variants of user behavior on the SERP:

- $E_1 \rightarrow C_1 \rightarrow E_2 \rightarrow End+_2$ - user clicks on the first result, finds it non-relevant, continues search and finds a correct answer in the second snippet
- $E_1 \rightarrow C_1 \rightarrow End+_1$ - user clicks on the first result and finds a correct answer in the document

Due to the lack of data for estimation we set $y_1 = 0.9$ as in [1] and $y_2 = 0.8$. 8-fold cross-validation and bootstrap method were used to evaluate the rest of the parameters and the corresponding confidence limits. Using the estimated parameters of the proposed model we construct a new metric - the measure of user satisfaction Psat at rank k.

$$Psat@k = \sum_{i=1}^{k} P(E_i = 1)P(End+_i = 1)$$

$$P(E_{i+1} = 1) = P(E_i = 1)(1-sa_i)[(1-ac_i)y_1 + ac_i(1-s_i)y_2]$$

$$P(End+_i = 1) = sa_i + (1 - sa_i)ac_i \cdot s_i$$

Parameters sa_i and ac_i depend on the label of snippet i (6 parameters: 4 ac_i for snippet labels from the Figure 2 and 2 sa_i for "Answer" labels) and s_i depends on the label of the document i (5 parameters for document labels from Figure 1).

4. PSAT METRIC EVALUATION

4.1 Validation Dataset

To validate the new Psat metric we conduct a number of offline experiments. For this purpose we collected a metric validation dataset based on the results of online experiments. These online experiments were A/B experiments with a control bucket of users served with the results generated by a high quality results and another bucket of users served with degraded snippet generation algorithm, degraded vertical results generation algorithm or a degraded ranking algorithm. These degradations were verified as to indeed degrade search experience by different online metrics (dwell time, CTR, time to first click, measured in the scope of the above-mentioned online expeirments). For the offline validation dataset we sampled 1056 SERPs of Android users from a Yandex interaction log collected in in November 2013. This set of SERPs was "control SERP group" for the offline experiment. For this set of SERPs we have made the same verified degradations as in the online experiments and every set of degraded SERPs was a "test SERP group". The set of resulting control and test group pairs formed our offline validation dataset. In the dataset there are 7 snippet degradations, 5 vertical results degradations and 2 ranking algorithm degradations. For the control and all test SERP groups the same types of judgements as in Section 3 were collected. We use this dataset to compare our metric with ERR metric [3] to evaluate their ability to detect degradations of different types and intensity.

4.2 SERP Degradations

In this section we describe the degradations used to form our validation dataset in more detail.

A snippet ranking algorithm generates a ranked list of candidate snippets for each document. The best snippet from this list is the snippet that is shown to user. For every search result from SERPs we collected two variants of bad snippets. First is a random snippet from the list of possible snippets for the document. Second is a snippet from the

middle of candidate snippets ranked by their quality.

We have done following snippet degradations:

- the replacement of the snippets of all organic results for 100%, 80%, 50% of SERPs with "middle" snippets
- the replacement of the snippets of all organic results for 100%, 80%, 50% of SERPs with "random" snippets

We imitated the following document ranking degradations:

- the replacement the 2nd document with 4th - swap[2;4]
- the replacement one random document from positions 2, 3, 4 with one random document from 5, 6, 7 - swap[2;4][5;7]

Also it was shown, that users might be satisfied only by vertical results and end the whole search session [2], therefore removing them from SERP means its degradation. Considering these facts we imitated the following degradations:

- removing all vertical results for 100% and 50% of SERPs
- removing all image vertical results
- spoiled image vertical result snippets with images from 150 page. (All pictures in image vertical result were replaced with images from 150 page of ranked images set.)
- spoiled image vertical results and corresponding snippets. Users see non-relevant images on the SERP and after the click on the image vertical results too.

We mentioned this type of snippet degradation separately, because we did not conduct the corresponding verifying online experiment, but we strongly believe that removing a correct answer from the snippet is a degradation for user. In our SERPs we replaced snippets judged as "Answer" snippets with "No answer, Click" snippets. Group of SERPs without "Answer" were an additional test group in our dataset.

4.3 Results

Our final dataset consists of the set of SERPs with degradations of different types. We choose different sample sizes to look at Psat metric performance and to compare it with ERR performance. Sample of every size was resampled 1000 times with replacement. The percentages of times when metric detects a significant difference between a pair of SERPs (initial control non-degraded SERP vs degraded SERP) are summarized in Table 1. Star (*) indicates a significant difference at level $\alpha = 0.01$ between two metrics. The paired permutation test was used to evaluate the significance of differences. On sample sizes more then 200 Psat outperforms ERR, especially it is evident for the degradations "remove all vertical results" and "for 50% of queries remove all vertical results". It is worth to mention that between two snippets degradations, snippet replacement with a "middle" snippet is stronger worsening (on N=1000: 99% vs 81%) than the replacement with a "random" snippet. It seems to be true, because a random snippet has more chances to be not bad than a snippet from the middle of the snippet ranking set. For every sample size, removing correct answers from snippets is a significant degradation in 100% cases, according to Psat. The sample size of 1000 is enough to detect 100% and 80% snippet degradations with snippet from the "middle".

5. CONCLUSIONS

Good abandonment rate will grow in the future due to the tendency to incorporate reach snippets and vertical results into the SERP. In this paper, we have proposed an offline metric that takes into account the chance that the user finds an answer in snippet without reading the document. Our metric is able to detect both the ranking algorithm's degradations and the snippet algorithm's degradations better than ERR metric. In addition we want to note that our metric Psat can signal about excluding of a correct answer from the snippet. It is worth to mention that the rate of good abandonments is high for desktop web search too, therefore the new metric could also be evaluated for PC web search.

One of our next target is to investigate the dependence of the probability of being tired from the type of the snippet (with or without a correct answer) and the type of queries. There is an assumption that after issuing a difficult query, the user is more patient than after the easy one, because in the second case he expects to find the answer more quickly. The current model is based on the assumption of linear examination (from the top to the bottom) which is very limiting and extending the model to let it deal with non-linear examinations could give better results.

6. REFERENCES

[1] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1–10, New York, NY, USA, 2009. ACM.

[2] D. Chen, W. Chen, H. Wang, Z. Chen, and Q. Yang. Beyond ten blue links: Enabling user click modeling in federated web search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 463–472, New York, NY, USA, 2012. ACM.

[3] A. Chuklin, P. Serdyukov, and M. de Rijke. Click model-based information retrieval metrics. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 493–502, New York, NY, USA, 2013. ACM.

[4] C. Clarke, N. Craswell, I. Soboroff, and G. Cormack. Overview of the trec 2010 web track. In *Proc. of TREC 2010*, TREC '11, 2011.

[5] O. J. Dore. Mobile search moments study. In *Think Insights Skip*, http://www.thinkwithgoogle.com/, 2013.

[6] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 331–338, New York, NY, USA, 2008. ACM.

[7] Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 153–162, New York, NY, USA, 2013. ACM.

[8] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 43–50, New York, NY, USA, 2009. ACM.

[9] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 525–532, New York, NY, USA, 2006. ACM.

[10] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 316–323, New York, NY, USA, 2002. ACM.