Learning to Account for Good Abandonment in Search Success Metrics

Madian Khabsa[1], Aidan Crook[1], Ahmed Hassan Awadallah[1], Imed Zitouni[1],
Tasos Anastasakos[1], Kyle Williams[2]
[1]Microsoft
[2]The Pennsylvania State University
{madian.khabsa, aidan.crook, hassanam, izitouni, tasoanas}@microsoft.com
kwilliams@psu.edu

ABSTRACT

Abandonment in web search has been widely used as a proxy to measure user satisfaction. Initially it was considered a signal of dissatisfaction, however with search engines moving towards providing answer-like results, a new category of abandonment was introduced and referred to as *Good Abandonment.* Predicting good abandonment is a hard problem and it was the subject of several previous studies. All those studies have focused, though, on predicting good abandonment in offline settings using manually labeled data. Thus, it remained a challenge how to have an online metric that accounts for good abandonment. In this work we describe how a search success metric can be augmented to account for good abandonment sessions using a machine learned metric that depends on user's viewport information. We use real user traffic from millions of users to evaluate the proposed metric in an A/B experiment. We show that taking good abandonment into consideration has a significant effect on the overall performance of the online metric.

1. INTRODUCTION

In traditional web search, a search engine responds to user queries by returning a list of relevant links from which the user is assumed to click on the most relevant one . In the case when the user leaves the search engine result page (SERP) without clicking on any result, it is typically referred to as query *abandonment.* Abandonment has long been assumed to indicate that the user has failed to find what he is looking for, therefore it became an indication of dissatisfaction [5, 3]. However, search engines have evolved beyond just returning 10 blue links to presenting direct answers and rich results that can satisfy the users information needs without clicks. Examples of these scenarios include: weather information, business hours and phone numbers, package tracking, and many others. In these scenarios, it is rather desired for the query to be abandoned, and such type of abandonment is referred to as *Good abandonment.*

With the growing share of mobile search among all web search, users on mobile devices are more likely to issue queries that can result in good abandonment [12]. In fact, it was estimated that 27% of of searches were performed with predetermined goal of having the search satisfied by the content on the SERP itself without the intention to click on any results [14].

Good abandonment presents a challenge for search engine metrics that rely heavily on user clicks to detect satisfaction. That is mainly because the users are less likely to leave traces that may be interpreted as explicit signals of satisfaction. On the other hand, when a user clicks in a scenario that is supposed to be answered by a direct answer, this click may be used as a signal of dissatisfaction, or at least an incomplete satisfaction of the user's need. Nevertheless, satisfaction is rather hard to capture in this scenario.

The problem is further exacerbated on mobile devices where 54.8% of the queries could potentially lead to good abandonment [12]. More recently, the surge in mobile search usage was culminated as the volume of queries originating from mobile devices exceeded those originating from desktops [1]. Thus, the sheer volume of abandoned queries is larger than ever. This presents a challenge for commercial search engines to identify queries with abandonment intent, and later populate a SERP with the desired information need, for failing to present the required information will risk degrading the user experience that may translate into dissatisfaction.

Previous work on good abandonment has mainly focused on either identifying queries with potential good abandonment [12] or detecting good abandonment on offline data that was manually labeled by judges [15, 12]. However, in real scenarios of web search engines, online evaluation on real traffic is imperative to identify improvements of the system. In this realm, we describe the challenges associated with devising a good abandonment aware metric, and show how a machine learned metric can be built using viewport information.

Our main contribution in this work is the introduction of a method to modify a standard search engine success metric to account for good abandonment. The resulting augmented metric is shown to be more sensitive than current state-of-the-art click based counterparts. The metric is validated using online data of millions of users through A/B experiments. Furthermore, our proposed metric relies on attention modeling to identify sources of success, which was shown to be a good indicator of satisfaction [7, 11].

CIKM'16 , October 24-November 28, 2016, Indianapolis, IN, USA
© 2016 ACM. ISBN 978-1-4503-4073-1/16/10. . . $15.00
DOI: http://dx.doi.org/10.1145/2983323.2983867

The remainder of this paper is organized as follows. Section 2 describes related work. In Section 3 we describe the good abandonment model. While online evaluation is presented in Section 4. We then conclude the paper in Section 5.

2. RELATED WORK

Good abandonment was formally introduced by Li et al. in [12] where they studied queries for which the information need is potentially satisfied by the content of the SERP itself. They analyzed abandonment across devices, and across different market segments. In the study they found that good abandonment is far more prevalent on mobile devices. In a different study, it was estimated that at least one quarter of web search queries were issued with predetermined intention of finding the information need on the SERP page itself, instead of following any link [14]. There has also been an effort to understand the underlying reason behind abandonment [6], with factoid queries being one class of queries that highly correlate with good abandonment [3].

Chuklin and Serdukov devised a set of features to build an SVM based classifier to detect bad abandonment using topical and linguistic features in conjunction with historical data[4]. The approach was evaluated on a manually labeled offline data. Koumpouri and Simaki conducted a user study to identify classes of queries that lead to good abandonment, and later constructed a decision tree using implicit measures to predict good abandonment[10]. Arkhipova and Grauer's work is the closest to our work where they devised a metric based on Dynamic Bayesian Network with its parameters being estimated on data from a user study. The metric was later tested on a small sample of mobile query sessions after introducing degradation to multiple parts of the system to show that the new metric is more sensitive to degradation [2]. However, they make the limiting assumption that users scan the results sequentially, whereas we make no assumptions about the browsing model. In addition, their dataset consists of multiple tasks that do not necessarily lead to good abandonment, with some of the parameters in their model being manually chosen. In our approach, all the parameters are learned from the data.

3. DETECTING GOOD ABANDONMENT

We propose to extend the search engine success metric to handle good abandonment as follows. Assume the function F to be the current success metric of a search engine, where F does not reward good abandonment. We introduce a function G that will be learned in Section 3.2, which evaluates *an abandoned query impression*, q_i, in order to distinguish good abandonment from bad abandonment. G only returns success when the impression q_i resulted in good abandonment. Using F and G we construct a new success metric S, such that for a given query impression q_j:

$$S(q_j) = \begin{cases} G(q_j) & \text{if } q_j \text{ is abandoned and not reformulated} \\ F(q_j) & \text{otherwise} \end{cases}$$

The query q_j is considered reformulated iff its edit distance with query q_{j+1} is below a given threshold.

Metrics are devised to quantify the performance of a given aspect of the search engine. System designers benchmark improvements by how much gain they attain on a given metric. Furthermore, success metrics become the function for which ranking algorithms are trained to maximize. Hence, there are many constraints imposed on devising new metrics. One such constraint is that metrics are only allowed to utilize *exogenous* signals, those that the system has no direct control over. On the other hand, *endogenous* signals are not allowed, since the system has direct control over them. For example, designing a feature based on whether a certain answer or vertical was shown on the SERP is an *endogenous* signal because the system has control on which verticals to show. A machine learned ranker will eventually learn that the presence of a certain vertical leads to gains in the metric, thus it would show that vertical whether it is relevant to the query or not. In addition to the constraint on the signals, metrics need to be interpretable, and should not be black boxes that merely output numbers. This stems from the need to understand the causes of success and failure in order to debug the system.

Given the aforementioned constraints, the good abandonment part of the metric G is learned through machine learning using the dataset describe in Section 3.1. Whereas we use a standard click based model [8] for the metric F, although it can be replaced with another metric without loss of generality.

3.1 Dataset

The source of the offline data in this study comes from randomly sampled abandoned queries in the mobile logs of a major search engine. Each query, along with the presented results were crowd sourced to 3 judges who were asked to provide satisfaction level on a 5 points scale. The judges were also shown the previous and next query in the session to provide them with the context. Majority voting was used to determine the final label of each query, and later the labels were binarized by considering ratings of 4 and 5 as satisfaction (SAT), similar to the trend in the literature. Overall, we collected a total of 3,895 labeled queries, of which we retained 1,565 queries that were labeled as SAT and 1,924 queries that were labeled DSAT.

3.2 Approach

To compensate for the lack of clicks, user attention can be used as an implicit signal to detect satisfaction. Previous work has shown that user attention, as measured by gaze tracking, is a good predictor of good abandonment [13]. The problem with gaze tracking, however, is that it is not feasible in an A/B setting since it requires special instrumentation that is not available to users at large scale.

Viewport information (information about the visible portion of the screen) was shown to be useful for understanding user intent on mobile devices [7], and it also correlates with gaze time [11]. Furthermore, it can scale with the large number of users whereas more accurate instrumentation such as eye tracking is not scalable.

Our approach relies on viewport information to record implicit interactions of the user with the SERP, such as scrolls, the part of the screen that is visible to the user, how long was each element of the SERP visible to the user, etc. The data from the viewport is processed and features are extracted to capture the main events. We use the same set of features that was described in [15], which was previously introduced in [11].

Given the constraints, we treat the problem as binary classification where the the goal is to classify *abandoned queries*

Table 1: The list of features used building the good abandonment metric G

Id	Feature Description
1	The total distance swiped in pixels
2	Attributed reading time (RT) of the first visible answer
3-5	Max, Min, StdDev attributed RT for answers
6-7	Max, StdDev attributed RT per pixel for answers
8	StdDev attributed RT for organic results
9-10	Max, StdDev shown fraction of organic results

into good and bad classes. First, we perform feature selection by choosing the top K most informative features ranked by information gain from the list of features proposed in [15]. Table 1 lists the choosen featues. Later, these features are used to build a shallow decision tree classifier. Decision tree is chosen because it is easily interpretable, yet effective. The obtained model achieves 66% F1 score on the offline judged data using 10 fold cross validation. Our model achieves a relatively competitive accuracy when considering that a boosted tree model trained using hundreds of features, and trees achieves 75% F1 score on the same dataset [15]. Recall that the goal is not to build the most accurate classifier on the labeled data, although it is desired, rather we seek to devise a satisfaction metric within the given *constraints*, and validate it on online data.

4. EXPERIMENTAL EVALUATION

Our evaluation relies on controlled experimentation, specifically A/B testing [9]. In designing the experiment it is important to keep in mind that certain *answers*, or *verticals* are more likely to appear on good abandoned queries [10]. The more likely the answer to be triggered in good abandonment scenario, the more sensitive the metric should be when degradation in that answer are introduced.

On the other hand, since clicks combined with dwell time remain a desired interaction from the system stand point, the metric should be careful not to over reward abandonment, especially in scenarios where the user is expected to click on results. A desired metric should be able to show sensitivity for good abandonment cases, while at the same time not change the signal when click through is expected.

Based on the discussions above we use an A/B experiment in which the user experience is degraded in the treatment pool by randomly modifying the position of the answer on the SERP. For example, assuming the search engine determined that the weather answer should be shown for a given query with the optimal position being at the top, then the treatment effect would move the answer to a random position on the SERP. It's possible that the final position of the answer remains the same as the one determined by the ranking algorithm. Clearly this is a degradation experiment since it presents a rather random ordering of answers instead of the optimal one.

The new metric is evaluated based on its performance when a given answer was present on the SERP. There are two set of answers that are chosen. The first list of answers is carefully chosen to represent impressions with high likelihood of good abandonment. Whereas the second list is a list of answers for which high click through was previously observed in the search engine logs. Ideally, the metric is expected to show improvements on the first list, while not introducing new significant changes on the second list of answers (some times this is referred to as A/A experiment where the metric is expected not to change). The following answers are chosen to capture potential good abandonment scenarios:

- **Weather answer**: this answer is shown when the query is believed to be weather related
- **Dictionary answer**: an answer that returns the definition of the query word
- **Finance answer**: shows finance related information such as stock and index prices

On the other hand, the following answers were chosen to represent the high click through scenarios:

- **Tracking answer**: an answer that shows tracking information when the query is a package tracking number. Users tend to track on the link in the answer to go to the courier's website.
- **Navigational Queries**: an answer that identifies navigational queries and returns the expected Url.
- **Showtimes**: this answer returns information about movies and theaters showing times. Users typically click on links to multiple movies and theaters when interacting with this answer

For comparison we employ a baseline metric that is common in the literature wherein satisfactions are observed whenever a user clicks on a result and spends at least t seconds on it [8]. Such approaches has already been shown to correlate with satisfaction, but it fails to capture cases with good abandonment. The baseline is appropriate in this experiment as it can show the improvements introduced by a good abandonment sensitive metric, while at the same time serve as guard against aggressive abandonment rewarding by ensuring that both metrics do not deviate when clicks are expected.

In Table 2, the results of running both the baseline and the new good abandonment sensitive metric on one week of traffic to the mobile segment of a commercial search engine are presented. As mentioned earlier, the experiment introduced a degradation to the user experience, therefore negative delta is expected. The top part of the table shows the answers for which the new metric is expected to show improvement (that is, the new metric should be able to find a negative delta). Whereas the bottom part contains answers for which the delta from both metrics should not differ. In addition to computing the delta between the treatment and the control for each metric, the deltas of both the baseline and the new model are tested to check if their difference is statistically significant at $\alpha = 0.05$, which is denoted in the last column. From the table we find that the new metric is able to detect degradation for both weather and dictionary answers by showing negative delta, whereas the baseline results in a positive delta (that is, the randomization result is better than the optimal result). The behavior of the baseline can be explained by the fact that the randomization is leading to more clicks on SERP results instead of answer-lead abandonment, thus the success metric would increase. While both the baseline and the model fail to detect a degradation

Table 2: The performance of the new metric, S, (under model) compared to the baseline metric, F, categorized by answer types. In the top part of the table, the expected delta is negative, while in the bottom part the new metric should not deviate from the baseline. Pairwise sig denote whether the delta of the baseline is statistically significant from the delta of the new metric

Answer	Baseline		Model		Pairwise Sig
	delta	pValue	delta	pValue	
Weather	0.0423	1.21E-06	-0.0181	6.58E-5	Yes
Dictionary	0.0228	4.55E-07	-0.0111	3.80E-06	Yes
Finance	0.0851	0	0.0355	0	Yes
Tracking	-0.04422	0.3505	-0.0481	0.2829	No
Navigational Query	-0.0001	0.8984	-0.0005	0.4562	No
Showtimes	-0.002	0.698	0.007	0.057	No

for the finance answer, the new proposed metric reduces the delta suggesting it is moving in the right direction.

On the other hand, for the cases where high CTR is expected, we find that the new metric conforms with the baseline in terms of delta signal. Furthermore, the resulting deltas from both metrics can not be considered different at $\alpha = 0.05$. It is worth noting that the insignificance is not due to the size of the data.

5. CONCLUSIONS AND FUTURE WORK

In this work we introduced a new success metric that is capable of detecting good abandonment in mobile search. The metric is used as an extension to click based models, and is based on machine learned classifier. We utilize attention modeling through viewport information of the user to come up with signals of satisfaction. Our metric is evaluated on online data of millions of users of a commercial search engine through A/B experiment where its effectiveness is shown.

In the future, we plan to study good abandonment in desktop search where viewport is of limited value since the SERP can be rendered without the need for scrolling.

6. REFERENCES

[1] Google mobile numbers. http://adwords.blogspot.com/2015/05/building-for-next-moment.html, 2015. Last accessed 2-2-2016.

[2] O. Arkhipova and L. Grauer. Evaluating mobile web search performance by taking good abandonment into account. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1043–1046. ACM, 2014.

[3] A. Chuklin and P. Serdyukov. Good abandonments in factoid queries. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 483–484. ACM, 2012.

[4] A. Chuklin and P. Serdyukov. Potential good abandonment prediction. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 485–486. ACM, 2012.

[5] A. Das Sarma, S. Gollapudi, and S. Ieong. Bypass rates: reducing query abandonment using negative inferences. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 177–185. ACM, 2008.

[6] A. Diriye, R. White, G. Buscher, and S. Dumais. Leaving so soon?: understanding and predicting web search abandonment rationales. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1025–1034. ACM, 2012.

[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*, pages 153–162. ACM, 2013.

[8] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 193–202. ACM, 2014.

[9] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

[10] A. Koumpouri and V. Simaki. Queries without clicks: Evaluating retrieval effectiveness based on user feedback. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1133–1134. ACM, 2012.

[11] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th International ACM SIGIR Conference on Research &#38; Development in Information Retrieval*, SIGIR '14, pages 113–122, New York, NY, USA, 2014. ACM.

[12] 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*, pages 43–50. ACM, 2009.

[13] W. Lu and Y. Jia. Inferring user preference in good abandonment from eye movements. In *Web-Age Information Management*, pages 457–460. Springer, 2015.

[14] S. Stamou and E. N. Efthimiadis. Interpreting user inactivity on search results. In *Advances in Information Retrieval*, pages 100–113. Springer, 2010.

[15] K. Williams, J. Kiseleva, A. C. Crook, I. Zitouni, A. Hassan Awadallah, and M. Khabsa. Detecting good abandonment in mobile search. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.