Is This Your Final Answer? Evaluating the Effect of Answers on Good Abandonment in Mobile Search

Kyle Williams[†], Julia Kiseleva[‡], Aidan C. Crook[Γ], Imed Zitouni[Γ],
Ahmed Hassan Awadallah[Γ], Madian Khabsa[Γ]
[†]The Pennsylvania State University, University Park, PA 16802, USA
[‡]Eindhoven University of Technology, Eindhoven, NL
[Γ]Microsoft, One Microsoft Way, Redmond, WA 98052, USA
kwilliams@psu.edu, j.kiseleva@tue.nl,
{aidan.crook, izitouni, hassanam, madian.khabsa}@microsoft.com

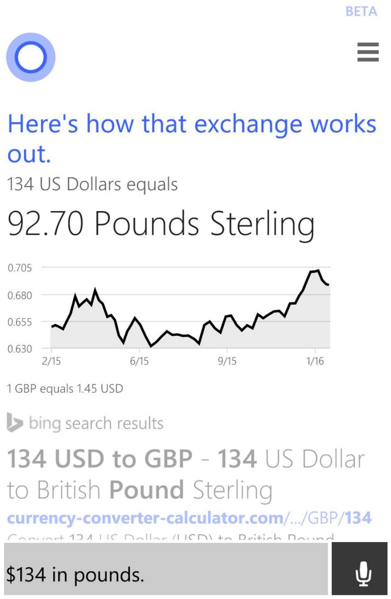
ABSTRACT

Answers on mobile search result pages have become a common way to attempt to satisfy users without them needing to click on search results. Many different types of answers exist, such as weather, flight and currency answers. Understanding the effect that these different answer types have on mobile user behavior and how they contribute to satisfaction is important for search engine evaluation. We study these two aspects by analyzing the logs of a commercial search engine and through a user study. Our results show that user click, abandonment and engagement behavior differs depending on the answer types present on a page. Furthermore, we find that satisfaction rates differ in the presence of different answer types with simple answer types, such as time zone answers, leading to more satisfaction than more complex answers, such as news answers. Our findings have implications for the study and application of user satisfaction for search systems.

1. INTRODUCTION

Mobile search has seen explosive growth in recent years. For instance, in 2013 it was estimated that 63% of Americans used their mobile phones to go online in comparison to only 31% in 2009 [5]. With this growth in mobile use, search engines have had to adapt to better suit user needs and behavior, which have been shown to be different on mobile devices [8, 10]. For instance, previous research has shown that mobile users may formulate queries in such a way so as to increase the likelihood of them being directly satisfied by the SERP [10] and that mobile queries differ from traditional desktop queries in being shorter and in their intents [8].

Search engines have had to adapt in order to accommodate these differences and one way they have done this is by showing answers on the Search Engine Results Page (SERP), such as answers about weather, the time, and sports scores. These answers typically appear in small boxes that appear on the SERP and contain factual information. For instance, Figure 1 shows a mobile answer for the query *$134 in pounds*. This mobile answer has the ability to satisfy a user who is interested in performing a currency conversion without having to click on, say, the first search result, which would take them to a currency conversion webpage. Thus, one of the effects of including these answers on a SERP is that users may no longer need to click on search results in order to satisfy their information need. However, most approaches to modeling and measuring user satisfaction have been based on user click behavior [6, 7] and, traditionally, a lack of clicks on a SERP has been seen as a negative indication of search result quality and the phenomenon has been labeled query abandonment [11]. However, recently there has been increasing awareness that query abandonment can be good [1, 3, 10, 11] in what is referred to as "good abandonment." In these cases, the user abandoned the query not because they were dissatisfied with the results but because the SERP satisfied the user without them needing to click on any search results. Previous research has estimated an upper bound for good abandonment on mobile devices of 54.8%, compared to only 31.8% on PC devices [10].



Here's how that exchange works out.

134 US Dollars equals

92.70 Pounds Sterling

1 GBP equals 1.45 USD

bing search results

134 USD to GBP - 134 US Dollar to British Pound Sterling

currency-converter-calculator.com/.../GBP/134

Convert 134 US Dollar (USD) to British Pound

$134 in pounds.

Figure 1: An example of a mobile SERP, showing an answer triggered in response to a currency conversion query.

SIGIR '16, July 17-21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07...$15.00
DOI: http://dx.doi.org/10.1145/2911451.2914736

Furthermore, previous research has also shown that factual answers on a SERP were responsible for 56% of the satisfaction from abandoned queries in mobile search [12]. Thus, it is clear that a relationship exists between mobile answers, abandonment, and satisfaction. However, to date, most research investigating answers on mobile devices have not differentiated among answer types and investigated how the answer type affects abandonment behavior. We hypothesize that it may be useful to take the answer type into consideration when developing a metric to assess abandonment in the presence of answers on a mobile SERP based on the belief that not all answer types contribute equally to good abandonment. Thus, in this study, we choose to empirically investigate the relationship between answer types and abandonment behavior. In doing so, we seek to answer the following two research questions:

> **RQ1:** *How does the presence of different answer types affect user click behavior and abandonment?*
>
> **RQ2:** *How do different answer types affect satisfaction in abandoned queries?*

To answer our research questions, we conduct a large scale analysis of the logs of a commercial search engine and also analyze satisfaction ratings gathered in a controlled user study. To our knowledge, this is the first work to investigate the relationship between answer types and good abandonment in mobile search.

2. RELATED WORK

This paper extends a long line of work investigating satisfaction in search [1, 3, 7, 11]. Previous work has shown how gesture features can be used to differentiate between good and bad abandonment in mobile search [12]. However, while the cited work showed that answers were a large driver of good abandonment, it did not investigate the effect that different answers types have on satisfaction as we do in this study. Chilton and Teevan [2] show how users interact with different types of answers in desktop search, where interaction specifically focused on click behavior. It was found that, as expected, answers reduced interaction with the rest of the SERP, which the authors refer to as *cannibalization*. Our work is similar to the cited work in that it seeks to understand the relationship between answers and user interactions. However, whereas the cited work focuses on the desktop, we focus on mobile search. Furthermore, we go further than the cited work in that we empirically evaluate the relationship between answer types and satisfaction. In [10] it was shown how mobile users often construct queries in such a way so as to be directly satisfied by the SERP. Furthermore, in [4] it was shown that high quality SERPs lead to increased satisfaction and good abandonment and answers are one way of potentially increasing the quality of SERPs and thus good abandonment. Lagun et al. [9] use viewport and eye-tracking to measure user engagement in mobile search and studied the effect of having relevant/irrelevant answers on a mobile SERP. Our work is similar in studying the effect of answers on a mobile SERP; however, it differs in that we attempt to understand how different answer types affect satisfaction and abandonment behavior.

3. ANSWERS AND BEHAVIOR METRICS

As previously mentioned, we hypothesize that the answer type, such as weather, time, etc., has an effect on abandonment and satisfaction. Thus, we begin by describing the answer types that we investigate in this study and also describe the metrics that we use.

3.1 Answer Types

Answers are triggered by the search engine in response to certain query intents and query types. In this study, we consider the following set of answer types, which were selected due to their frequency in search impressions, their variety and their use in previous studies evaluating answers on non-mobile SERPs [2].

1. **Math** An answer to a math question, such as 2×4
2. **Currency** A currency conversion answer
3. **Dictionary** A dictionary definition
4. **Finance** Financial information about a company
5. **Flight Status** The status of a flight
6. **News** News related to the query
7. **Package Tracking** Tracking information for the query
8. **Phonebook** Contact information
9. **Reference** An inline reference fact
10. **Show Times** Show times related to the query, e.g., movie show times
11. **Sports** Information about sports teams, such as scores
12. **Timezone** The time in a specified time zone
13. **Twitter** Posts from Twitter
14. **Translation** A translation of the query
15. **Weather** The weather forecast in the specified region

3.2 Metrics

We define three metrics that we use to describe user behavior on the SERP in the presence of answers. We are specifically interested in how answers affect good abandonment and thus we define metrics that focus on user click behavior or lack thereof. We denote the set of answer types as A. For each answer type, $a \in A$, we define the click rate (CR) for answer type a as as:

$$CR_a = \frac{\sum_s^S \mathbf{1}_{\boldsymbol{S_a}}(s) click(s)}{\sum_s^S \mathbf{1}_{\boldsymbol{S_a}}(i)}, \tag{1}$$

where S is the set of SERP impressions sampled from the log, S_a is the set of SERPs containing answer type a, $\mathbf{1}_{\boldsymbol{S_a}}(s)$ is an indicator function indicating membership of s in $\boldsymbol{S_a}$, and $click(s)$ is equal to 1 if SERP s received a click and 0 otherwise. This metric captures the rate that users click on SERPs containing answer type a. From this, we define the abandonment rate (AR) for answer type a as:

$$AR_a = 1 - CR_a. \tag{2}$$

This metric captures the rate at which users abandon SERPs containing answer type a.

We also define an engagement rate (ER) for answer type a as the rate at which the actual answer on the SERP is clicked on. First, we calculate the average number of clicks that SERPs with answer type a receive as:

$$AvgClicks_{\boldsymbol{S_a}} = \frac{\sum_s^S \mathbf{1}_{\boldsymbol{S_a}}(s) TotalClicks(s)}{\sum_s^S \mathbf{1}_{\boldsymbol{S_a}}(s)}, \tag{3}$$

where $TotalClicks(s)$ is the number of clicks on SERP s. We also calculate the average number of direct clicks that answer type a receives on the answer itself as:

$$AvgClicks_a = \frac{\sum_s^S \mathbf{1}_{\boldsymbol{S_a}}(s) TotalAnswerClicks(s)}{\sum_s^S \mathbf{1}_{\boldsymbol{S_a}}(s)}, \qquad (4)$$

where $TotalAnswerClicks(s)$ is the number of clicks on the clickable components of the answer box in SERP s. We then define the ER for answer type a as:

$$ER_a = \frac{AvgClicks_a}{AvgClicks_{\boldsymbol{S_a}}}. \qquad (5)$$

Thus the engagement rate captures the extent to which page clicks are engagements with the answer.

4. ANSWER EFFECT ON BEHAVIOR

The previous section described the answer types that we consider in this study when evaluating abandonment and also defined the metrics that we consider. In this section, we use these metrics to assess how user behavior in terms of clicks, abandonment, and engagement differs in the presence of different answer types.

4.1 Large Scale Log Sample

We perform our analysis on a large scale sample of mobile search logs from a commercial search engine. We sample over 20 million mobile impressions from a week during June 2015. These impressions come from about 9 million sessions and 1 million users. Each impression is associated with anonymized information about the query and session, information about the SERP elements that were visible, such as answers and organic search results, and information about the click behavior of the user. We use this dataset to perform a large-scale analysis of behavior for different answer types.

4.2 Results

Figure 2 shows the click rate (CR) and abandonment rate (AR) for different types of answers on the mobile SERP. In the figure we observe different click and abandonment behavior depending on the answer type. For instance, we notice from Figure 2 that pages that contain answers, such as *package tracking*, *phonebook*, *news*, *math*, and *show times* have relatively high click rates ranging from 59% to 93%, meaning that people often click on pages containing these types of answers. This is perhaps not surprising since information needs related to news, package tracking, and show times, often require the user to perform additional navigation in order to fully satisfy their information need. By contrast, it can also be seen from Figure 2 that pages containing answers related to *currency*, *finance*, *dictionary*, and *time zones* experience relatively low click rates ranging from 14% to 32% implying high abandonment rates from 68% to 86%. As was the case with answer types that led to high click rates, the fact that pages with these types of answers experience relatively high abandonment rates is not surprising since the answers satisfy simple and straightforward information needs. The evidence here suggests that, when evaluating abandonment, it is useful to take into consideration the type of information present on the SERP. For instance, when abandonment happens on a SERP that usually has a high click rate, that may suggest that the user was not satisfied. By contrast, abandonment on a SERP that usually has a high abandonment rate may not indicate dissatisfaction.

Engagement rates for different answer types are shown in Figure 3. As can be seen from the figure, some answer types dominate the SERP page clicks or, to use the language of [2], the answers *cannibalize* the clicks from the rest of the SERP. For instance, the answer engagement rate on pages containing a *package tracking* answer is 95%, indicating that users likely had to engage with the answer to satisfy their information needs. By contrast, for answer types that had high abandonment rates, the engagement rates are relatively low indicating that not much information is gained by engaging with the answer. Thus we have an answer to **RQ1:** *How does the presence of different answer types affect user click behavior and abandonment?* The data shows that user behavior differs depending on answer types and suggests that it is worth further investigating the relationship between answer presence and satisfaction since some answers require clicks for satisfaction, whereas other answers are able to satisfy a user without them needing to click. We investigate this further in the next section.

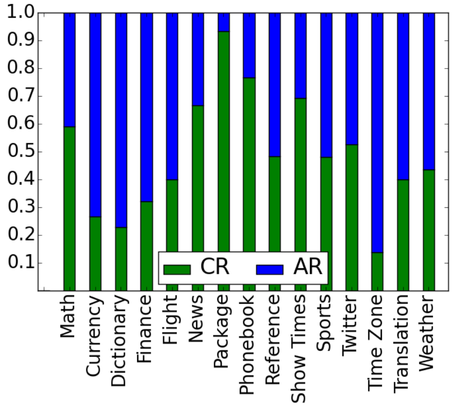


Figure 2: Click rate (CR) and abandonment rate (AR) for different answers on mobile SERPs.

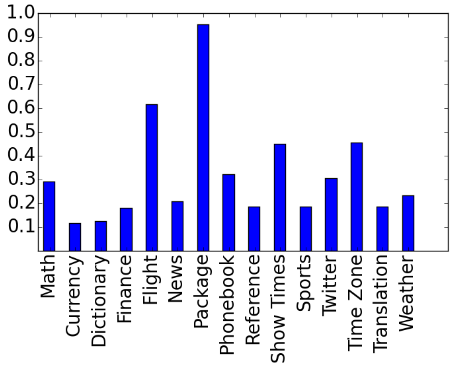


Figure 3: Engagement Rate (ER) for different answers on mobile SERPs.

5. ANSWER EFFECT ON ABANDONMENT

The previous section showed how different answer types on the SERP lead to different abandonment behavior. However, are these abandonment differences good or bad? In this section, we analyze labeled data in order to try and answer this question.

5.1 Dataset

We perform our analysis on a dataset gathered through a controlled user study. We briefly describe the dataset in this section with a more complete description presented in [12]. 60 participants were recruited from the United States of which 25% were female and the remainder male. The mean age of participants was

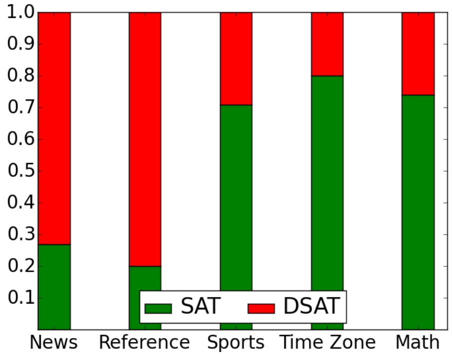


Figure 4: SAT and DSAT rating associated with the different answer types gathered in the user study.

25.5 (±5.4) years and the user study included 5 information seeking tasks, which were designed to increase the likelihood of good abandonment [12]. At the end of the tasks, the users were asked to provide a satisfaction ratings. The dataset contains a total of 607 queries of which 576 were classified as abandoned since they received no clicks. Since this dataset involved a controlled user study experiment, not all of the answer types described in Section 3.1 were present. Thus, we only focus on the following subset of answer types: *news, reference, time zone, sports, math* with frequencies of 62, 5, 5, 14, and 19, respectively. After filtering out queries that did not trigger these answer types, we retained 105 queries.

5.2 Results

For each answer type that appeared for the queries in the data described above, we measure the SAT and DSAT rates. As can be seen in Figure 4, we observe different SAT and DSAT rates for different answer types. For instance, the presence of *time zone, sports*, and *math* answer types are all associated with relatively high SAT rates in the data, with all being above 70%. By contrast, the SAT rates in the presence of news and reference answers are both below 30%. Thus, there is evidence that the presence of different answer types may affect satisfaction. Though the data are drawn from different sources and under different circumstances, it is also interesting to observe the relationship between the SAT rates in Figure 4 and the abandonment rate in Figure 2. For instance, mobile SERPs containing *sports* and *reference* answers both have abandonment rates of 52%, but very different SAT rates of 71% and 20%, respectively. Similarly, mobile SERPs containing *math* and *news* answers have somewhat similar abandonment rates of 41% and 33%, but very different SAT rates of 74% and 27%, respectively. This is in contrast to, say, SERPs containing *timezone* and *math* answers, which have very different abandonment rates of 86% and 41%, but similar SAT rates of 80% and 74%, respectively. As an example of bad abandonment, SERPs with *reference* answers experience an abandonment rate of 52% in Figure 2 and a satisfaction rate of only 20%. Thus most of the abandonment is bad. The findings in this section allow us to begin answering **RQ2:** *How do different answer types affect satisfaction in abandoned queries?* It is clear that user behavior and whether it indicates good or bad abandonment is influenced by the answer types seen by users. The reason for this is that, as shown in Figures 2 and 3, different answers lead to different user behavior and some answers require clicks, whereas others do not. Thus, we argue that any evaluation of satisfaction and abandonment in the presence of mobile answers should take the answer type and its properties into consideration.

6. CONCLUSIONS AND FUTURE WORK

We investigated the effect that different types of mobile answers have on good abandonment. Similar to the desktop case [2], it was shown how user behavior differs in the presence of different answer types in terms of clicks and answer engagement. Furthermore, it was shown how the rates of satisfaction differ for answer types that have similar abandonment rates, thereby showing that the answer type influences whether abandonment is good or bad.

It is interesting to hypothesize why this may be the case; we propose two areas for further investigation. (1) Ambiguity in query intent – if the query intent is unambiguous (as is the case in a math or time zone queries) then an informational answer is more likely to satisfy the user than if the query intent is ambiguous, such as a query about a place or celebrity, where the intent could vary from factual information to topical news. (2) The ability of an answer to fully address the interpreted intent - e.g., answering an inquiry on the height of Mount Everest can be more succinctly presented than a query for the latest news on the election. Modeling the likelihood of good abandonment in terms of properties of the query intent, as opposed to the rendered answer types, will be important in enabling future experimentation and improvement of answers that attempt to satisfy the underlying intent without necessitating a click.

Our study does have some limitations. For instance, user behavior in response to an answer is largely related to answer design in terms of display, ranking, etc. By considering answer type we effectively seek to capture some of these properties but acknowledge the limitation. Also, the frequency of the reference and time zone answers was relatively low in the dataset presented in Section 5.1.

In summary, this paper has shown that answer types are relevant in SAT measurement and we believe it will be useful to account for the underlying reason that different answer types appear to influence SAT on abandoned pages. However, answers should be considered alongside other features, such as gestures and features from the query and session. This is especially the case for answer types that achieve relatively equal levels of clicks and abandonment.

References

[1] M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz. Direct answers for search queries in the long tail. In *CHI*, pages 237–246, 2012.

[2] L. B. Chilton and J. Teevan. Addressing people's information needs directly in a web search result page. In *WWW*, pages 27–36, 2011.

[3] A. Chuklin and P. Serdyukov. Potential good abandonment prediction. In *WWW*, pages 485–486, 2012.

[4] A. Chuklin and P. Serdyukov. Good abandonments in factoid queries. In *WWW*, pages 483–484, 2012.

[5] M. Duggan and A. Smith. Cell Internet Use 2013, 2013. URL http://www.pewinternet.org/2013/09/16/cell-internet-use-2013/.

[6] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2):147–168, 2005.

[7] A. Hassan, R. Jones, and K. L. Klinkner. Beyond dcg: User behavior as a predictor of successful search. *WSDM*, pages 221–230, 2010.

[8] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and iphones and mobile phones, oh my! In *WWW*, pages 801–810, 2009.

[9] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. *SIGIR*, pages 113–122, 2014.

[10] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR*, pages 43–50, 2009.

[11] Y. Song, X. Shi, R. White, and A. H. Awadallah. Context-aware web search abandonment prediction. In *SIGIR*, pages 93–102, 2014.

[12] K. Williams, J. Kiseleva, I. Zitouni, A. C. Crook, A. H. Awadallah, and M. Khabsa. Detecting good abandonment in mobile search. In *WWW*, pages 495–505, 2016.