The Impact of Result Abstracts on Task Completion Time

Rehan Khan
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
rkhan@google.com

David Mease
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
dmease@google.com

Rajan Patel
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
rajan@google.com

ABSTRACT

This paper examines the impact of web search result abstracts on task completion time. The dataset we analyze comes from a controlled study in which paid participants carry out assigned tasks using a popular search engine. We compare task completion times with the abstracts removed to the task completion times with the abstracts present. We conclude that the removal of the abstracts leads to longer task completion times. The statistical significance of the result verifies that task time is a potentially useful metric for evaluating abstract quality. For comparison, we also have participants carry out the same tasks with the abstracts present but with the first five search results removed. This leads to even longer task completion times, and gives us a reference point for quantifying the value of the search result abstracts relative to the value of the first five search results themselves.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Representation**]: Miscellaneous

General Terms

World Wide Web, Search, Human Factors

Keywords

World Wide Web, Search, User Study, Usability, Snippets, Query Logs

1. INTRODUCTION

A major factor in users' interactions with search results is the nature of result abstracts provided. At a minimum, the abstract provides a pointer to the indexed document, but ideally it must provide enough information about the document that users may form a meaningful judgment of its likely relevance. Understanding how users interact with abstracts has been studied by a number of authors. Presenting users with sentences extracted from documents that match their query has been shown to increase user interaction with results [15, 6]. Eyetracking studies of user interaction with search results have shown that users spend most of their gaze time interacting with the top few results [6, 8, 9], and suggest that eye-tracking metrics are a useful way to measure the impact of different types of abstracts. The use of click-through data to measure the quality of varying abstract generation algorithms has been investigated by a number of authors. Clarke [5] found that readability, inclusion of query terms in the abstract, and query length all increased the likelihood of user clickthrough on a result. Joachims [10, 11] investigated the use of implicit feedback as an indicator of the judged relevance of documents based on their abstracts. Finally, Dumais [7] investigated alternatives to the standard relevance-sorted lists of results, and found that grouping results by category and revealing category information in abstracts improved users ability to complete tasks. This study used a task-based approach most similar to ours.

In this paper we seek to quantify the value of search result abstracts relative to the value of the ranking of the search results themselves. This is challenging because search result ranking evaluation is commonly carried out using metrics such as precision, recall, Mean Average Precision (MAP) [3] and binary preference (bpref) [4] which ignore the abstract entirely. More holistic methods of search engine evaluation are necessary.

Some researchers have had success in using task-based metrics such as task completion time. Allan et al. [2] demonstrated that task time had a statistically significant negative relationship with system retrieval accuracy as measured by bpref. Al-Maskari et al. [1] observed a statistically significant decrease in the geometric mean time to find the first relevant document when comparing a system with high average precision to a system with low average precision. Su [13] identified time as the most frequently mentioned reason given by users as contributing to their rating of the overall success of the IR system. Turpin and Scholer [14] found a negative relationship between precision at rank 1 and the time users took to find their first relevant document. Xu and Mease [16] confirmed that task completion time was negatively correlated with task satisfaction and that task completion time could be used to separate two ranking algorithms of different quality.

The use of task completion time as a metric is attractive for our purposes since it provides a standard yardstick with which we can evaluate both the search result retrieval quality as well as the abstract quality. In this paper we will use the methodology developed in Xu and Mease [16] to compare task completion times under (1) standard conditions to those with (2) the abstracts removed and (3) the abstracts present but the quality of the search results greatly reduced. Specifically, the search results and abstracts are taken from

WWW2009, April 20-24, 2009, Madrid, Spain.

a popular search engine. In order to reduce the quality for (3) we simply remove the top 5 search results. From the quantitative analysis of the task completion time data we can compare the value of search result abstracts relative to the value of the search result ranking.

The paper is organized as follows. Section 2 describes the experimental design and the data collection. Section 3 describes the resulting data using some simple graphs and summary statistics. Section 4 gives the details of the statistical models used to fit the data, and Section 5 discusses the results from those models. Finally Section 6 gives some concluding remarks.

2. METHODOLOGY

We will use similar methodology as Xu and Mease [16]. We describe this in the following paragraphs and highlight any important differences.

The first step was to generate tasks. For this we had 200 paid participants each describe a difficult search task that they had recently attempted. Specifically, part of the instructions read

> We are looking for a story about some information which you have recently tried to find on the internet but had a difficult time in doing so.

We provided a number of other guidelines including

> . . . the information needed to answer the question must be publicly accessible on the internet.

From the resulting 200 tasks, 32 were removed for various reasons such as they contained personal information or were not specific enough. The fraction removed here was substantially larger than in Xu and Mease [16]. The remaining 168 were used for the study, although many of these were edited for content and grammar. The most common edits made were to convert multiple part questions to single part questions. For example, a task asking for the name of a book and where to purchase the book would be changed to simply ask either one of the two parts. Two examples from the 168 which are representative are given below.

> Example Task #1:
> I have a cat whose eyes are starting to exude puss. I'm afraid it might have an infection, and she can't open her right eye. I want to find what this is and how to treat it.
>
> Example Task #2:
> Jon Voight has a brother who writes songs and sings country music but I can't think of his name. What is it?

After these tasks were obtained, a different group of 328 paid participants were assigned to carry out the tasks. We will refer to these paid participants as "users" from here on. Each time a user was assigned a task, he or she was assigned to one of the 3 treatments for that task: (1) control, (2) no abstracts or (3) no top 5 search results. The control provides standard search results from the popular search engine. The second treatment differs from the control only in that the abstract text is removed. However, the document titles and URLs are retained. For simplicity we refer to this condition as "no abstracts" throughout, but it is important to keep in mind that the document titles and URLs do provide valuable clues to the users as to the content on the page. Finally, the third condition differs from the control in that the top 5 search results are removed, and the search engine's results 6 though 15 are shown instead of the usual search results 1 through 10. In all three conditions the page was branded to appear very similar to the usual search engine, but a noticeable non-white background was used to differentiate it so that users would not accidentally leave the experiment and go to the public page for the search engine. Users were not permitted to click on advertisements, and the "next page" button was disabled so users were only allowed ten results per query.

Each of the 168 tasks was carried out by 20 users for each of the 3 treatments giving a total of 168 x 20 x 3 = 10,080 observations. Users were permitted to keep acquiring tasks until they desired to stop or until there were no more remaining, so there was no uniformity among the number of tasks carried out by the different users. For example, one of the 328 users did 136 of the 168 tasks, while 6 users did only a single task. This design was chosen only for practical reasons. That is, by allowing each user to acquire as many tasks as he or she desires, we were able to accumulate the total 10,080 observations more quickly than if we had required more uniformity.

In the instructions for the project, the users were told to read the task and then to click a "Start searching" button which would begin the search session by opening the appropriate modified version of the search engine in a new browser window. The users were instructed to keep searching until a total of 7 minutes had passed or until they had successfully completed the task. Note that this is slightly different from Xu and Mease [16] in which no specific time restriction was given, and as a result the data in this paper must be analyzed with respect to this right-censoring. A clock was displayed to the users while they carried out their tasks so that they could keep track of the time. When finished, the users were asked to click a "Finish searching" button so that we could record the total task time.

3. DATA

The boxplots in Figure 1 show the distribution of the resulting task times for the 3 treatments. The censored values are included and set to be equal to the censoring time of 420 seconds for the purpose of this graph only. It can be seen that the median time without abstracts is slightly more than with the abstracts present, and that the median time without the top 5 search results is the largest of the three. It should also be noted that there is a considerable amount of variability in the task times. Much of this is due to task variation and user variation as discussed in Xu and Mease [16]. For any single user or any single task there is much less variation. These two sources of variation will be controlled for in our statistical model in the following section.

The boxplots also reveal that the data has a right-skewed distribution which is typical with task time data. For this reason we will choose to model the data with a log-normal distribution. The histogram of the task times for the control condition is shown in Figure 2. The right-censoring task times (at 420 seconds) are not shown in this graph. The curve in the graph is the fitted log-normal distribution with the median estimated to be 150.0 seconds. The parameters of this distribution were estimated using the `survival`

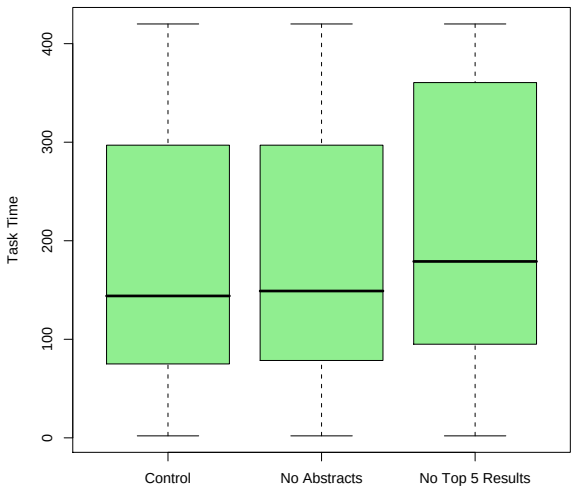

Figure 1: The boxplots show the task times for the 3 treatments. The large variation within the treatments is a result of variation across different users and different tasks. This variation is controlled for by our statistical model.

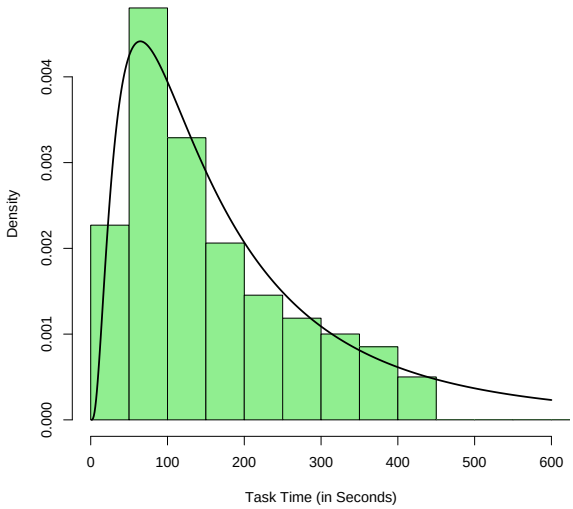

Figure 2: The distribution of the data follows a lognormal distribution. Some of the data is right-censored at 420 seconds and is not shown in this graph. This graph shows only the data for the control group.

package available in R [12], an open source language for statistical computing. Specifically, the "survreg" function was used to fit the log-normal distribution since this function can handle the censored data correctly using maximum likelihood. Without abstracts (condition 2) the estimated log-normal median increases to 153.4 seconds, and without the top 5 search results present (but abstracts included) the estimated lognormal median increases to 182.6 seconds. As with the boxplots, this analysis suggests that both of these conditions result in longer task times with the latter being more severe than the former. We will explore this more carefully in the following sections and check for statistical significance.

4. STATISTICAL MODEL

The maximum likelihood estimation used to fit the distribution as described above is appropriate for the right-censored lognormal data we have, so we extend this technique to compare the three treatments while simultaneously accounting for the variability introduced by both user and task effects. Specifically, we use the "survreg" function to fit a model which assumes that the task time is log-normal (with truncation at 420 seconds) and has a mean that is given by the additive model

$$I + C_k + U_i + T_j \ . \tag{1}$$

Here I is a baseline (intercept) term, C_k $(k = 1, 2, 3)$ are the three conditions including the control, U_i $(i = 1, ..., 328)$ are the user fixed effects and T_j $(j = 1, ..., 168)$ are the task fixed effects. We avoid over-parameterization by fixing the control (C_1) to be zero as well as one arbitrary user and one arbitrary task. Since we have fixed the control to be zero, the estimated effects for the other two treatments and their standard errors are relative to the control. We analyze these values below.

5. RESULTS

Fitting the model described above produces the values shown in Figure 3. This graph shows the increase in task time relative to the control for both treatments along with 95% confidence intervals. The removal of abstracts yields an increase in task time of 3.5% with a confidence interval of 0.3% to 6.8%. The removal of the top five search results yields an increase in task time of 23.9% with a confidence interval of 20.0% to 27.9%. Thus, we conclude that both of the effects are statistically significant; however, the impact of the removal of abstracts on task time is only about 1/7th of the impact of removing the top five search results.

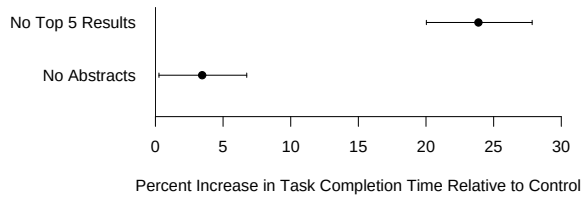

Figure 3: Removal of abstracts causes a 3.5% increase in task time and removal of the top five search results causes a 23.9% increase in task time relative to the control. 95% confidence intervals are shown.

6. CONCLUSIONS

We found that for both of our result manipulations (ablating abstracts and removing the best 5 results) users' task completion times were increased in aggregate. As expected, the effect was much larger for removing the best 5 results than for ablating abstracts (a 23.9% vs. 3.5% increase, roughly a 7-fold difference), although both results were statistically significant. An attractive feature of using task time as a metric is that it allows for comparison of different kinds of search result changes on the same scale, so that their relative impact can be quantified.

The fact that we were able to use this methodology to find a statistically reliable effect due to a change only in abstracts is encouraging and suggests that this methodology could be an effective means to assess variations in abstract generation algorithms, or in abstract organization and user interface changes as in Dumais [7].

An important caveat about any task-based methodology is the procedure by which tasks are generated or selected. The results are only meaningful to the extent that the tasks are representative of the task-stream to which the proposed manipulation would be applied. Highly biased task selection could lead to highly significant results in an experiment, but no impact on a wider set of tasks. Conversely, if a modification affects only a narrow kind of task then experiments performed on a random selection of tasks would tend to dilute the effect of the modification. Developing methods for generating and classifying representative tasks for such studies is an important area for future research.

7. REFERENCES

[1] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: does the test collection predict users' effectiveness? In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 59–66, New York, NY, USA, 2008. ACM.

[2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 433–440, New York, NY, USA, 2005. ACM.

[3] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40, New York, NY, USA, 2000. ACM.

[4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, New York, NY, USA, 2004. ACM.

[5] C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in web search. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 135–142, New York, NY, USA, 2007. ACM.

[6] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 407–416, New York, NY, USA, 2007. ACM.

[7] S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 277–284, New York, NY, USA, 2001. ACM.

[8] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 478–479, New York, NY, USA, 2004. ACM.

[9] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 417–420, New York, NY, USA, 2007. ACM.

[10] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2005. ACM.

[11] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2):7, 2007.

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[13] L. T. Su. Evaluation measures for interactive information retrieval. *Inf. Process. Manage.*, 28(4):503–516, 1992.

[14] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18, New York, NY, USA, 2006. ACM.

[15] R. W. White, I. Ruthven, and J. M. Jose. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 57–64, New York, NY, USA, 2002. ACM.

[16] Y. Xu and D. Mease. Evaluating web search using task completion time. In *SIGIR '09: Proceedings of the 32nd annual international ACM SIGIR conference on Research and development in information retrieval*, (Submitted).