Detecting Success in Mobile Search from Interaction

Qi Guo, Shuai Yuan, and Eugene Agichtein
Mathematics and Computer Science, Emory University, Atlanta, GA 30322, USA
{qguo3, syuan3, eugene}@mathcs.emory.edu

ABSTRACT

Predicting searcher success and satisfaction is a key problem in Web search, which is essential for automatic evaluating and improving search engine performance. This problem has been studied actively in the desktop search setting, but not specifically for *mobile search*, despite many known differences between the two modalities. As mobile devices become increasingly popular for searching the Web, improving the searcher experience on such devices is becoming crucially important. In this paper, we explore the possibility of predicting searcher success and satisfaction in mobile search with a smart phone. Specifically, we investigate client-side interaction signals, including the number of browsed pages, and touch screen-specific actions such as *zooming* and *sliding*. Exploiting this information with machine learning techniques results in nearly 80% accuracy for predicting searcher success – significantly outperforming the previous models.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*selection process, search process.*

General Terms

Algorithms, Experimentation, Human Factors, Measurement.

Keywords

Detecting success and satisfaction, mobile search behavior.

1. INTRODUCTION

Predicting searcher success and satisfaction is a key problem in Web search, which is essential for search engines to automatic evaluate and diagnose performance on a large-scale, as well as providing real-time intervention and assistance if dissatisfaction can be detected at an early stage [3]. Some previous research has attempted this problem [3][5][7] in a general Web search setting where computers were considered as the primary platform.

Recently, as mobile devices, such as smart phones, have become an increasingly popular platform for browsing and searching the Web, it is becoming crucial that the Web search experience on a mobile phone is satisfactory. However, to the best of our knowledge, no work has been done to understand how the behavioral patterns on a mobile device can reveal the success and satisfaction of a search task. There are many differences in the usage of computers and mobile devices [2] [8], and so it is unclear whether the models of searcher success in satisfaction developed for the desk-top setting (e.g., using fine-grained user signals, such as scrolling and mousing [1][4][6]) would translate to the mobile search setting.

As an attempt to close this gap, we present, as far as we know, the first study of automatically predicting *searcher success and satisfaction in mobile Web search* from behavior on mobile phones with touch screens, laying the groundwork for more extensive future research. In addition to previously studied behavioral signals of search success, we investigate *client-side interaction features*, including the number of pages browsed, zooming, scrolling/sliding, and orientation changing - increasingly common in modern smart phones - and show that these features significantly improve prediction accuracy.

2. METHODOLOGY

We conducted a controlled user study to collect data with known search success outcome. Ten subjects were recruited (6 male, 4 female, average age 26.2 ± 3.2). All subjects were undergraduate and graduate students or staff at the Emory University, and had some experience with Web search and smart phones. We designed the search tasks (some examples are given in Table 1) for this study to be representative of common Web search tasks on mobile devices. The tasks have varying difficulty and topics, and highlight geographical intents that have been identified as a significant portion of mobile information needs [2].

Table 1. Example task descriptions *(initial queries)*

Task 2: Find the **MARTA routes and schedules** from Georgia Tech to Emory on Tuesdays after 7PM (marta schedules)
Task 5: Find the **closest zipcar location** to MathCS (zipcar locations)
Task 7: Find **the earliest show times** of "social network" after 7:00PM today in the closest movie theater (*social network*)

The user study proceeded as follows: before the tasks began, the participants were given a tutorial of using the phone, including opening bookmarks, clicking, zooming, scrolling, and changing the physical device orientation. Next, a warm-up task was given to each participant to familiarize them with the task procedure. To begin each task, the participants were presented a task description and an initial query. For each task, the participants were instructed to first open the bookmark with the Google search engine result page (SERP) of the initial query. Once they reached the SERP, they could click the search results and/or reformulate the query if needed until the information was found or it took too long than they would spend in reality. After each task, the participants were asked a few questions, including whether they have successfully completed the task and how satisfied they were about the search experience during the task. Following the warm-up, eight search tasks were given to each participant.

To capture the client-side interactions, including the number of browsed pages, zooming, and sliding, we developed a modified version of the Chrome browser application[1] for the Android phone. The events are encoded in a string and sent to the server as HTTP requests for analysis.

For each of the success and satisfaction dimensions, we formulate our prediction task as binary classification: that is, we classify each search task into two classes: successful/satisfied (user selected "very successful/ satisfied" in the post-task questionnaire) and unsuccessful/unsatisfied.

We then represented each search task as a feature vector, with values corresponding to the server-side features and client-side

Copyright is held by the author/owner(s).
SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

[1] Available at http://ir.mathcs.emory.edu/intent/data/sigir2011/

features. The server-side features include the number of queries, clicks, clickthrough rate (CTR), average query length and task duration. The client-side features include the numbers of all browsed pages, search engine result pages (SERP), and non-SERP pages, as well as the event counts on these pages (e.g., scrolling, scaling). Additional details about the features are available on the project website referenced above.

Our intuition was that these client-side features can provide additional insights about the searcher success and satisfaction level. For example, a small number of queries and clicks with moderate task duration might indicate a successful search task, as it seems that the user did not need to spend too much time and effort to complete the task; however, if the user, during this task, browsed a large number of pages and had to intensively scroll and rescale on the browsed pages, she might be actually not satisfied and not even successful, since she actually spent a lot of efforts and was keeping searching without finding the relevant information.

We experimented with various classification algorithms, including Bayes Network (BN), Support Vector Machines (SVMs), decision trees, and others. We report classification results for BN only, as it performed best in this setup, even though other algorithms achieved similar performance.

3. RESULTS AND DISCUSSION

To simulate the real scenarios in mobile search, where searchers are less likely to attempt time-consuming and complex searches, we did not assign overly difficult tasks to our participants. As a result, the means of success and satisfaction ratings are 3.25 (std=1.0) and 3.0 (std=1.2) respectively, on a 5-point scale (i.e., 0 represents very unsuccessful/unsatisfied, 4 represents very successful/satisfied). Interestingly, success rating has a noticeably higher mean and the correlation between these two are high but not perfect (R=0.81), which makes sense since users might feel unsatisfied about the experience even if they end up finding the information successfully. Also, the noticeably higher variance of the satisfaction ratings for each task suggests higher subjectivity of making satisfaction judgments.

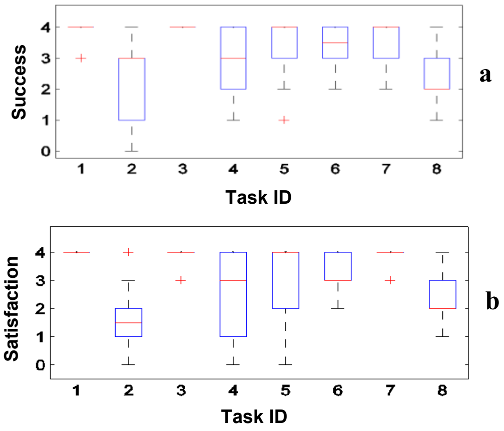

Figure 1. Success (a) and Satisfaction (b) by Task ID

As we can see in Figure 1, some tasks are better solved, while other tasks have more room for improvement. Interestingly, for easier tasks, the variance of both satisfaction and success ratings across users is smaller. In contrast, for more difficult tasks, the variance of both satisfaction and success ratings is larger, which suggests significant opportunities for personalization, or of exploiting successful "expert" mobile searchers to help unsuccessful "novice" mobile search users.

Table 2. Results of predicting success and satisfaction. Significance of differences is indicated between models and: Baseline: $^{\triangle}$ p<.05, $^{\blacktriangle}$ p<.01; Server: $^{\circ}$ p<.05, $^{\bullet}$ p<.01.

Method	Both		Successful		Unsuccessful	
	Acc	F1	Prec	Recall	Prec	Recall
Baseline	57.5	36.5	57.5	100	0	0
Server	66.3	64.2$^{\blacktriangle}$	67.9$^{\blacktriangle}$	78.3$^{\triangle}$	63.0$^{\blacktriangle}$	50.0$^{\blacktriangle}$
Client	80.0$^{\triangle}$	79.2$^{\blacktriangle}$	80.0$^{\blacktriangle}$	87.0$^{\triangle}$	80.0$^{\blacktriangle}$	70.6$^{\blacktriangle}$
Full	78.8$^{\triangle}$	78.0$^{\blacktriangle}$	79.6$^{\blacktriangle}$	84.8$^{\triangle}$	77.4$^{\blacktriangle}$	70.6$^{\blacktriangle}$
Method	**Both**		**Satisfied**		**Unsatisfied**	
	Acc	F1	Prec	Recall	Prec	Recall
Baseline	53.8	35.0	53.8	100	0	0
Server	71.5$^{\blacktriangle}$	68.1$^{\blacktriangle}$	66.1$^{\blacktriangle}$	95.3	88.9$^{\blacktriangle}$	43.2$^{\blacktriangle}$
Client	76.3$^{\blacktriangle}$	75.6$^{\blacktriangle}$	74.0$^{\blacktriangle}$	86.0$^{\blacktriangle}$	80.0$^{\blacktriangle}$	64.8$^{\blacktriangle}$
Full	78.8$^{\blacktriangle}$	78.5$^{\blacktriangle\circ}$	78.3$^{\blacktriangle\circ}$	83.7$^{\blacktriangle\circ}$	79.4$^{\blacktriangle}$	73.0$^{\blacktriangle\circ}$

Ten-fold cross validation was used: in each fold, tasks of nine participants were used for training and the remaining tasks of the left-out participant were used for testing. We report the average of the results across the folds in Table 2. As we can see, the BN classifiers (our models) significantly outperformed the majority baselines, exhibiting the accuracy of 79% compared to the baseline system (accuracy of 54% for predicting satisfaction and 58% for predicting success). Interestingly, using client-side features achieved better performance than using server-side features, and combining the two achieved optimal performance for predicting search satisfaction, while clients-side achieved the best performance in predicting search success.

To understand the contributions of the various features, we computed the X^2 statistic for each feature with respect to the class. We found that the most significant features include the number of browsed pages, number of non-SERP events, task duration, clickthrough rate, number of clicks and average query length. Generally, the more effort a user spent on searching, the less likely she were to be satisfied or successful – which makes sense in a mobile setting, where the screen size is small, the bandwidth is limited, and each user interaction requires effort.

In summary, we explored the feasibility of predicting searcher success and satisfaction in mobile search. Our experiments show that we can predict search success and satisfaction with accuracy of nearly 80%, by incorporating additional client-side interactions. Our results support the feasibility of the automatic evaluation and improvement of the search engine performance for mobile search.

ACKNOWLEDGMENTS This work was partially supported by the NSF grant IIS-1018321 and the Yahoo! Faculty Research Engagement program.

4. REFERENCES

[1] Buscher G., van Elst, L., and G., Dengel A. Segment-level display time as implicit feedback: a comparison to eye tracking. In Proc. SIGIR 2009.

[2] Church, K., Smyth, B. Understanding the intent behind mobile information needs. In Proc. IUI 2009.

[3] Feild, H., Allan, J., and Jones, R. Predicting searcher frustration. In Proc. SIGIR 2010.

[4] Guo, Q., and Agichtein, E. Ready to buy or just browsing? Detecting web searcher goals from interaction data. In Proc. SIGIR 2010.

[5] Hassan, A., Jones, R., and Klinkner, K.L. Beyond DCG: user behavior as a predictor of a successful search. In Proc. WSDM 2010.

[6] Huang, J., White, R.W., and Dumais, S.T. No clicks, no problem: using cursor movements to understand and improve search. In Proc. CHI 2011.

[7] Liu, J., Liu C., Gwizdka, J., and Belkin N.J. Can search systems detect users' task difficulty? some behavioral signals. In Proc. SIGIR 2010.

[8] Kamvar, M., Kellar, M., Patel, R. and Xu, Y. Computers and iPhones and mobile phones, oh my! A logs-based comparison of search users on different devices. In Proc. WWW 2009.