

Investigating Result Usefulness in Mobile Search

Jiaxin Mao*, Yiqun Liu*, Noriko Kando†,
Cheng Luo*, Min Zhang*, Shaoping Ma*

*Tsinghua University, Beijing, China

†National Institute of Informatics, Tokyo, Japan
yiqunliu@tsinghua.edu.cn

Abstract. The existing evaluation approaches for search engines usually measure and estimate the utility or usefulness of search results by either the explicit relevance annotations from external assessors or implicit behavior signals from users. Because the mobile search is different from the desktop search in terms of the search tasks and the presentation styles of SERPs, whether the approaches originated from the desktop settings are still valid in the mobile scenario needs further investigation. To address this problem, we conduct a laboratory user study to record users' search behaviors and collect their usefulness feedbacks for search results when using mobile devices. By analyzing the collected data, we investigate and characterize how the relevance, as well as the ranking position and presentation style of a result, affects its user-perceived usefulness level. A moderating effect of presentation style on the correlation between relevance and usefulness as well as a position bias affecting the usefulness in the initial viewport are identified. By correlating result-level usefulness feedbacks and relevance annotations with query-level satisfaction, we confirm the findings that usefulness feedbacks can better reflect user satisfaction than relevance annotations in mobile search. We also study the relationship between users' usefulness feedbacks and their implicit search behavior, showing that the viewport features can be used to estimate usefulness when click signals are absent. Our study highlights the difference between desktop and mobile search and sheds light on developing a more user-centric evaluation method for mobile search.

Keywords: Mobile Search, Evaluation, User Behavior Analysis

Introduction

With the rapid growth of mobile search, the evaluation of the mobile search engine is becoming an important research topic in Information Retrieval. Previous research has shown that the mobile search is different from desktop search in several aspects including search intents [23], user interfaces (e.g. a much smaller screen), and users' search behavior patterns including querying [22, 13], SERP scanning [14], and relevance assessment [24]. Recently, to further reduce user's interaction cost on mobile devices, a larger number of search results that aim to satisfy users without requiring them to click, such as knowledge graphs [16] and direct answers [26], are incorporated into mobile SERPs, making the mobile search results become even more diverse. Due to these differences, whether the existing *system-oriented* and *user-oriented* evaluation methods that were developed for desktop search are as effective in mobile search needs further investigation.

For the system-oriented evaluation, the Cranfield-like evaluation paradigm [2] is widely used. In this paradigm, we measure the effectiveness of search systems by computing some *evaluation metrics* such as MAP and NDCG [10], based on a set of *relevance judgments*. While Verma and Yilmaz [24] found that the relevance judgments on desktop and mobile are different, some recent studies [11, 15, 20] in desktop search have spotted a gap between the relevance annotations from assessors and the *usefulness* [4] feedbacks from users and showed that usefulness feedback has a stronger correlation with user satisfaction. However, to what extent the relevance annotation can reflect the result-level user-perceived usefulness and be adopted to estimate the query-level user satisfaction in mobile search has not been extensively studied yet.

In the user-oriented evaluation, user's click [12] and post-click dwell time [5] on landing pages have been widely used as implicit feedbacks to measure the user satisfaction in Web search. However, to reduce user's interaction cost, modern mobile search engines often present search results in the form of information cards [16, 26], which aim to meet user's information needs on SERPs, without requiring further clicks. Therefore, the click-based online evaluation methods may not be as reliable in mobile search. To address this problem, some recent studies (e.g. [16, 17]) proposed to utilize the viewport¹ changes on mobile devices to capture user's viewing behavior and estimate their attention in mobile search. These studies suggested that user's viewing behavior captured by viewport changes is valuable in measuring user satisfaction in mobile search. But to the best of our knowledge, no existing research has systematically investigated the relationship between user's viewing behavior and their explicit usefulness feedbacks *per result* in mobile search.

To fill these two research gaps, we conducted a laboratory user study to address the following research questions:

- **RQ1:** What factors may affect the user-perceived usefulness of a search result in mobile search?
- **RQ2:** How do the user-perceived usefulness of search results correlate with the query-level user satisfaction in mobile search?
- **RQ3:** How can we use search behavior features to estimate user-perceived usefulness in mobile search?

The laboratory user study enables us to get explicit feedbacks from users, record rich behaviors, and control the undesired variabilities. In particular, we use the explicit result-level usefulness feedbacks and query-level satisfaction feedbacks from the participants to measure the user-perceived usefulness of a search result and the query-level user satisfaction in this study (See Section Data Collection for more details).

Data Collection

The data was collected through a laboratory user study with 43 participants.

Search Tasks 20 search tasks were adopted in the user study. Each search task is defined by a query selected from the query log of a commercial mobile search engine. Among these 20 search tasks, 13 are informational, 6 are transactional, and only 1 of

¹ the region on the display screen for viewing the content of Web pages.

them is navigational (i.e. finding the official website of a university). The informational tasks covers a variety of topics such as QA, news, healthcare etc. and the transactional tasks are about finding specific videos, images, and mobile games. The authors further created a background story according to each sampled query to reduce the potential ambiguity of a single query. For each search task, we crawled four SERPs from four popular mobile search engines on one day of October, 2016, using the corresponding query. Because the search tasks cover different topics, the search results on these SERPs cover a variety of vertical types such as Image, Video, News, QA, and Knowledge Graph. In this way, we collected $20 \times 4 = 80$ SERPs for 20 search tasks from 4 different mobile search engines.

Participants We hired 43 undergraduate students (20 females and 23 males, aged from 19 to 23) from our university as participants via emails and online social networks. In the pre-experiment questionnaire, most participants reported that they were familiar with search engines (Mean=5.68 in a 7-point Likert scale from *not familiar at all* to *very familiar*) and smart phones (Mean=5.79 in a 7-point Likert scale), which indicates that they had adequate search expertise to complete the mobile search tasks.

Apparatus We implemented a Web-based experimental system to host the crawled SERPs and used an Android smartphone which has a 5-inch touch screen with a resolution of 1280×720 . Using the `WebView` widget provided in Android SDK, we developed an experimental mobile browser which can record rich user interaction logs including the content of visited pages, scrolling, touch gestures, clicks, and switchings between pages.

The widths of the crawled SERPs is equal to the width of the viewport and zooming was not allowed. Therefore, all the scrolling actions in the collected log are in vertical directions. Depending on the heights of search results, the initial viewport usually contains the first 2-4 results of each SERP.

User Study Procedure Each participant were required to complete 20 search tasks. For each search task, only one of the four SERPs from four different search engines would be shown to a participant. To balance the sources of the SERPs, we divided the 20 search tasks into 4 groups (task 1-5 as the first group, 6-10 as the second group, and etc.) and used a Latin square of size 4 to assign the SERPs from one of the four search engines to each group. In this way, we created 4 different settings for assigning the sources of SERPs to search tasks. The participants were assigned to the four settings in a balanced way, therefore, each SERP was shown to 10 or 11 participants. To control the order effects of search tasks, we also rotated the 20 search tasks using a Latin square of size 20.

For each search task, we first showed the task description (i.e. a query and the background story) to the participant. Then the participant was required to search with the query and complete the search task using the experimental mobile browser. The instruction given to the participants was the following:

“Assuming you have the information need described in the background story, please search with this query in our system as you usually do with a mobile search engine.”

Because we only crawled the first SERP for each search task, query reformulations and paginations were not allowed.

After completing the search task, the participant would give usefulness feedback for the search results and satisfaction feedback for the query. Unlike previous studies in desktop settings [20] that only require the participant to give usefulness feedbacks for the *clicked* results, we asked the participants to select all the results they had *examined* on the mobile SERPs and give usefulness feedbacks for all these *examined* results. The collected result-level usefulness feedback in our study should reflect whether the presented snippet on the SERP was useful or not if the result was not clicked, or whether both the content on the SERP *and* the content on the landing page were useful if the result was clicked. We use the 4-level graded usefulness feedback and corresponding feedback instruction ($U \in \{1, 2, 3, 4\}$) that were adopted by Mao et al. [20].

In this way, we collected participants' self-reported examination feedbacks (E) and their explicit usefulness feedbacks (U) for the examined results simultaneously. E is a binary variable and $E = 1$ means the participant reported that she examined the result. We further assume that the unexamined results did not contribute to the completion of the search task, therefore, their usefulness feedbacks U were set to 1: not useful at all.

For query-level satisfaction (SAT), a 5-level graded scale [18] was used and the instruction was:

*“Are you satisfied with your search experience with the query and search results?
1: not satisfied at all - 5: very satisfied”.*

Data Annotation To investigate what factors affect the user-perceived usefulness of mobile search results, we further hired professional assessors to assess *relevance* and *click necessity* [19] for all the search results. Because previous research [24] shows that the relevance annotation will be affected by the device used in the annotation process, we required the assessors to make annotations on the same smartphone that was used in the user study.

We used a typical 4-level graded relevance following the TREC criteria [25]. Each search result was annotated by three assessors. The Fleiss' κ of relevance annotation is 0.388, which demonstrates a fair agreement between the assessors.

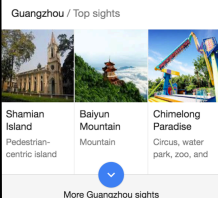
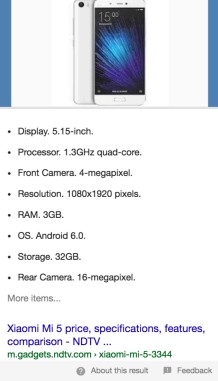
First Viewport for Query: <i>place to visit in Guangzhou</i>	Click Necessity	First Viewport for Query: <i>Mi 5 specs</i>	Click Necessity
 <p>Guangzhou / Top sights</p> <p>Shiamian Island Pedestrian-centric Island</p> <p>Baiyun Mountain</p> <p>Chimelong Paradise Circus, water park, zoo, and</p> <p>More Guangzhou sights</p>	2	 <ul style="list-style-type: none"> • Display. 5.15-inch. • Processor. 1.3GHz quad-core. • Front Camera. 4-megapixel. • Resolution. 1080x1920 pixels. • RAM. 3GB. • OS. Android 6.0. • Storage. 32GB. • Rear Camera. 16-megapixel. <p>More items...</p> <p>Xiaomi Mi 5 price, specifications, features, comparison - NDTV ...</p> <p>m.gadgets.ndtv.com • xiaomi-mi-5-3344</p> <p>About this result Feedback</p>	1
<p>10 Best Places to Visit in Guangzhou (2017) - TripAdvisor</p> <p>https://www.tripadvisor.in • Attractions-g...</p> <p>Hotels near Canton Tower. Hotels near Shiamian Island. Hotels near Chimelong Safari Park. Hotels near Chen Clan Ancestral Hall-Folk Craft Museum. Hotels near Baiyun Mountain. Hotels near Chimelong Paradise. Hotels near Yuexiu Mountain. Hotels near Pearl River (Zhujiang)</p>	3		
<p>10 Best Places to Visit Around Southern China - EscapeHere</p> <p>m.escapehere.com • destination • 10-best...</p>	3	<p>Xiaomi Mi 5 - Full phone specifications - ...</p>	3

Fig. 1: Examples of results with different click necessity (1: Not Necessary; 2: Possibly Necessary; 3: Definitely Necessary).

Different types of vertical results are federated into the SERP of mobile search engines. The contents and the presentation styles of these heterogeneous results may have an effect on how the user interacts with them. Traditionally, such effect is investigated by categorizing the results into different vertical types, such as knowledge graph [16], direct answer[26] as well as weather, travel, finance, and etc. [8]. However, the search results in our dataset were crawled from four different search engines and have many different presentation styles. It is tricky to develop a taxonomy to cover all the presentation styles in our dataset properly. Therefore, in this study, we adopted the *click necessity* measure and the corresponding annotation procedure proposed by Luo et al. [19] to investigate the effect on usefulness brought by the abundant information presented in the snippets of heterogeneous mobile search results.

Similar to relevance annotation, each result was annotated by three assessors. The Fleiss' κ of the click necessity annotation is 0.475, which reaches a moderate agreement level and shows that the click necessity can be annotated reliably by external assessors. We show some examples of results with different click necessity scores in Figure 1. 137 (17.6%) of the unique results were annotated as "1: not necessary", 136 (17.4%) as "2: possibly necessary", and 507 (65.0%) as "3: definitely necessary". Over half of results were annotated as "3: definitely necessary" because organic results constitute a major proportion of the search results.

Collected Data After a thoroughly inspection of the collected dataset, we removed 3 informational search tasks because of the malfunctioning of the experimental apparatus, especially the search behavior logging function. We collected 731 valid search sessions². There are 1,831 clicks and 2,305 usefulness feedbacks in these sessions.

Influencing Factors of Usefulness Feedback in Mobile Search

Regarding **RQ1**, in this section, we investigate three factors that may influence the result-level usefulness feedback in mobile search: the ranking position of the result, the relevance with the query, and the click necessity of its presentation style.

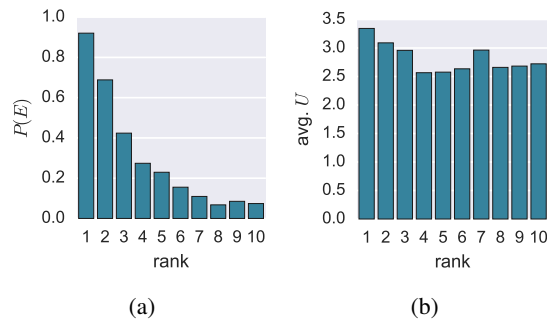


Fig. 2: The effects of rank on user's (a) examination and (b) usefulness feedbacks.

² The dataset will be open to public for research purpose after the double-blind review process.

Effect of Ranking Positions Previous research showed that in desktop setting, the examination of search results is affected by the *position bias* [6]. Higher-ranked results tend to receive more user attention and larger probabilities of examination. Because examination is a prerequisite for usefulness, the rank of a result may also influence its usefulness feedback.

We first show the effect of ranking positions on participants' self-reported examination in Figure 2a. The results confirm that the probability of examination $P(E)$ is decreasing with the rank of results. While 92.0% of the results in the 1st position were examined by participants, only 22.9% of the results in the 5th position were examined.

We then show the average usefulness feedbacks for the examined results in different ranks in Figure 2b. From this figure, we find that: 1) the top-3 results have significantly higher usefulness feedbacks than other results (independent t-test, two-tailed, $p < 0.001$). 2) the usefulness feedbacks of the top-3 results decrease with the rank significantly (one-way ANOVA test, $F(2, 1503) = 22.24$, $p < 0.001$). 3) There is no significant difference in the usefulness feedbacks of the results from 4th to 10th positions³ (one-way ANOVA test, $F(6, 729) = 1.69$, $p = 0.12$).

These observations indicate that the rank of results affects not only user's examination behavior but also their usefulness judgments on examined results. It is also interesting to see that the users treat the results in the initial viewport (i.e. the results in top 3 positions) differently than the other results. They rate the examined results in top 3 positions as more useful than the other examined results. While the position bias affects their usefulness feedbacks in the initial viewport, the position effect seems to become less important when users scroll downwards to examine the results in the 4th to 10th positions.

Effect of Relevance To investigate the relationship between relevance and usefulness in mobile search, we compute the Pearson's r and Cohen's Weighted κ [3] between the relevance annotations from assessors and usefulness feedbacks from participants. For all the displayed results, there are only a weak linear correlation ($r = 0.29$) and a slight agreement ($\kappa = 0.11$) between relevance and usefulness. But if we only consider the examined results (i.e. the results that have usefulness feedbacks from the participants), a moderate linear correlation ($r = 0.50$) and a fair agreement ($\kappa = 0.33$) are detected. The reason for this apparent difference is that many relevant results were not examined by the participants because of the position bias on examination shown in Figure 2a. We also note that the correlation between the relevance annotations and usefulness feedbacks of examined results in our study is stronger than the correlation in desktop search reported by Mao et al. [20] ($r = 0.332$, $\kappa = 0.209$). Compared to their study, the search tasks is more specific and query reformulation is not allowed in our study. Therefore, it is easier for the relevance assessors to guess user's information needs and make relevance judgments that can better reflect the user-perceived usefulness.

Effect of Click Necessity We are also interested in understanding how the click necessity of results affect the user-perceived usefulness.

While the Pearson's r between the 3-level click necessity annotation and 4-level usefulness feedback of the examined results is not significantly different from 0 ($r =$

³ We omitted the results ranked below the 10th position here because some SERPs only contains 10 results.

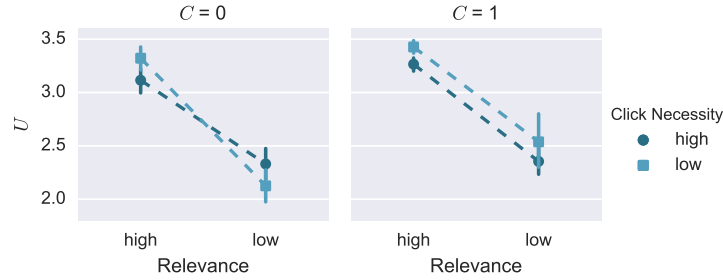


Fig. 3: Effects of click necessity and relevance on the usefulness feedbacks of unclicked results ($C = 0$) and clicked results ($C = 1$).

0.034, $p = 0.10$), we hypothesize that the click necessity has interaction effects with relevance on user-perceived usefulness and these effects may differ for the clicked results and unclicked results. Therefore, we conduct two 2×2 two-way ANOVA tests, that regard both relevance and click necessity as factors, for both clicked and unclicked results. Binary labels for relevance and click necessity are generated based on the 4-level and 3-level graded annotations. Because 50.4% of results are highly relevant ($R = 4$), we regard the results with $R = 4$ as results with *high relevance* and other results as with *low relevance*. For click necessity, we group the results with annotations “1: not necessary” and “2: possibly necessary” as results with *low click necessity* ($n = 273$, 35.0%) and the results with annotation “3: definitely necessary” as results with *high click necessity* ($n = 507$, 65.0%).

The interaction effects are shown in Figure 3. From the left part of the figure, we observe that the click necessity and relevance of unclicked results has an interaction effect on usefulness feedback. The ANOVA test shows that the interaction effect is statistically significant ($F(1, 811) = 9.62, p = 0.002$). Presenting highly relevant information directly on the SERP can bring more usefulness even when the result is not clicked. From the right part of the figure, we find no interaction effect of relevance and click necessity for the clicked documents ($F(1, 1486) = 0.03, p = 0.85$). However, the clicked results with low click necessity is significantly more useful than the clicked results with high click necessity ($F(1, 811) = 13.14, p < 0.001$). The low-click-necessity results usually come from high-quality sources, such as online encyclopedia and online Q&A sites. Therefore, were they clicked, the usefulness feedbacks of them may be higher than the organic results with high click necessity.

Usefulness vs. Satisfaction in Mobile Search

Regarding **RQ2**, we examine the relationship between result-level usefulness and query-level satisfaction by correlating some metrics based on participants’ usefulness feedbacks with their satisfaction feedbacks. We use the same metrics based on relevance annotations as baselines. The results measured in Pearson’s r are shown in Table 1.

We first compute the Discounted Cumulative Gain (DCG) [10] truncated at different positions using relevance annotations and usefulness feedbacks. We assume the unexamined results (i.e. the results without usefulness feedbacks from the participant)

Table 1: Pearson’s r between satisfaction feedbacks and metrics based on relevance annotations and usefulness feedbacks. The darker and lighter shadings indicate the correlation is significant at $p < 0.01$ and 0.05 . * (or **) indicates the difference is significant at $p < 0.05$ ($p < 0.01$), comparing to the same metric based on relevance annotation R .

	Relevance (R)	Usefulness (U)
$DCG@1$	0.147	0.350**
$DCG@3$	0.192	0.381**
$DCG@5$	0.172	0.320**
$DCG@10$	0.122	0.282**
CG_C	-0.087	0.014
MAX_C	-0.062	0.114**
AVG_C	0.030	0.206**
CG_E	-0.057	0.072*
MAX_E	0.088	0.541**
AVG_E	0.252	0.548**

are “not useful at all ($U = 1$)”. From the upper part of Table 1, we observe that: 1) the correlation between usefulness feedbacks and satisfaction is stronger than that between relevance annotations and satisfaction, which is similar to the findings in desktop settings [20]. 2) For both relevance annotations and usefulness feedbacks, $DCG@3$ is the best among the $DCGs$ truncated at different positions in terms of the correlation with satisfaction feedbacks, which indicates the results in the initial viewport play an important role in determining the user experience in mobile search.

We also compute some online metrics and correlate them with satisfaction feedbacks. Because users may acquire useful information without clicking the results in mobile search, we take all the examined results into consideration. The following online metrics are adopted:

- CG_C/CG_E : the sum of all result-level judgments of clicked/examined results.
- MAX_C/MAX_E : the maximum of result-level judgments of clicked/examined results.
- AVG_C/AVG_E : the average result-level judgments of clicked/examined results.

The correlations between these online metrics and users’ query-level satisfaction are shown in the lower part of Table 1. It is observed the examined results, especially MAX_E and AVG_E , have stronger correlations with the query-level satisfaction than those based on only the clicked results, which confirms our hypothesis that the unclicked but examined results also contribute to the mobile search experience. We also see that, while the online metrics based on usefulness feedbacks can better reflect satisfaction, the online metrics based on relevance annotations perform poorly in estimating query-level satisfaction in mobile search.

We further investigate the reason for this apparent difference between usefulness and relevance judgments by showing the average values of the online metrics based on them for sessions with different satisfaction level in Figure 4. An increasing gap between the relevance-based online metrics and the usefulness-based ones is spotted in these figures as the decrease of satisfaction level. These gaps suggest the relevance

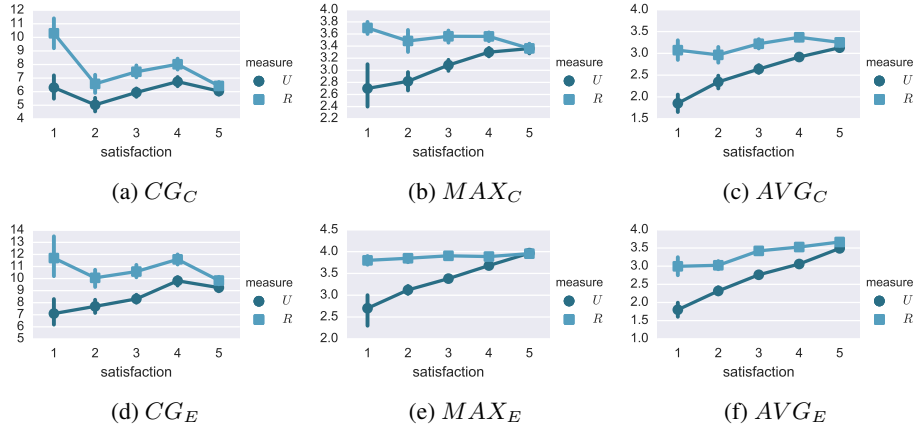


Fig. 4: The average examination-based online metrics of the search sessions with different satisfaction level (SAT).

annotations systematically overestimate the utility of the results in unsatisfied search sessions. This finding in mobile search confirms Mao et al. [20]’s finding in desktop search that relevance is not sufficient for usefulness and satisfaction.

It is also interesting to find non-monotonous relationships between CG_C , CG_E , and satisfaction (SAT). In the sessions with moderate satisfaction level ($SAT=2-4$), CG_C and CG_E based on relevance and usefulness are positively correlated with SAT . However, when the session is highly unsatisfactory ($SAT=1$), the user will compensate for the low result quality by clicking on more results, which results in a higher CG_C of both result-level measures and a higher CG_E of relevance annotation. The users will be satisfied if they can find enough useful information with minimum effort, therefore, the average CG_C and average CG_E of extremely satisfied sessions ($SAT=5$) are lower than those of the sessions with $SAT=4$.

Usefulness vs. User Behavior in Mobile Search

Addressing **RQ3** will help us to infer user’s experience during search using the behavior signals that can be logged passively. Recent studies have proposed to use the viewport time [16] to estimate the attention and satisfaction of users when click signals are absent or at least inaccurate in mobile search. So in this section, we first investigate the relationship between the viewport time and explicit usefulness feedback from the user.

Figure 5 shows the average viewport time on the snippets of clicked and unclicked results. It is interesting to see from the left part of the figure that there is a weak but statistically significant positive correlation ($F(3, 811) = 2.71, p = 0.044$, Pearson’s $r = 0.10$) between the viewport time and usefulness feedback for the unclicked results, while from the right part that there is no such correlation for the clicked results ($F(3, 1386) = 0.89, p = 0.44$). This results suggest that in mobile search, the viewport time can be a useful signal in estimating the usefulness of unclicked results but are not so helpful in inferring the usefulness of the clicked results.

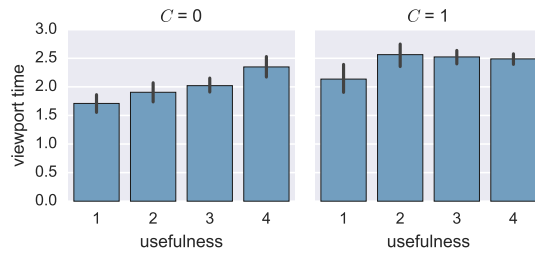


Fig. 5: Average viewport time of the clicked and unclicked results.

To further test whether the viewport time can help in estimating user-perceived usefulness in mobile search, we build regression models to predict the 4-level usefulness feedbacks of the examined results ($n = 2,305$). We compute four viewport features for each result: 1) viewport time of the snippet (viewport time); 2) viewport time divided by the session time (viewport time%); 3) viewport time divided by the area of the snippet (viewport time per pixel); 4) the number of snippets that have been covered by the viewport (#snippet in viewport). We combine the viewport features with existing click and dwell time based features (e.g. the features used in [20]) to train the Gradient Boosted Regression Tree (GBRT) to predict user-perceived usefulness and test whether the viewport features can improve the prediction performance via 10-fold cross-validations. We randomly shuffle the dataset for three times and apply the 10-fold cross-validation for each shuffled dataset, which generates $3 \times 10 = 30$ test folds.

Table 2: The results of usefulness prediction for examined results ($n = 2,305$).

	MSE	MAE	Pearson's r
Click&Dwell time	0.865	0.748	0.374
Click&Dwell time + Viewport	0.849**	0.747	0.394**

The average performance of the models on the 30 test folds, measured in the Mean Squared Error (MSE), Mean Absolute Error (MAE), and Pearson's r between the predicted values and true values of the usefulness feedbacks, are shown in Table 2. By comparing the performance of these two models, we find that adding viewport features brings a small but statistically significant improvement, measured in MSE and Pearson's r . This result suggests that the viewport time is a valid signal for usefulness.

Related Work

Mobile Search Existing research show that mobile search is different from traditional desktop search in the following aspects: (1) Mobile search is often conducted in different contexts compared to desktop search: people are more likely to search for news, location-based information and etc. with “fragmented attention” [9]. This observation is also confirmed by analysis on commercial search engines’ query logs [23]. (2) The screen space of mobile devices is much smaller than a particular desktop display. Thus mobile users have to incur more effort to read the same amount of information. This will also impact users’ behavior pattern and experience [14, 16, 19,

17, 21]. (3) Since modern mobile devices are often equipped with a touch screen, users usually interact with SERPs with Multiple Touch Interactions [7], which provides a new opportunity to model users' search processes in a finer grain. These differences between mobile and desktop environment motivate the study of the evaluation of mobile search engines.

Usefulness as an Evaluation Criteria for Search Search evaluation sits at the center of IR studies. While the *system-oriented* evaluation methods aim to build reusable test collections to evaluate the effectiveness of search systems, the *user-oriented* evaluation methods try to measure user's experience during the information seeking process.

Since Belkin et al. [1] and Cole et al. [4] has proposed usefulness as a criteria for the evaluation of interactive information retrieval, some recent effort has been put into filling up the gap between relevance judgment from assessors and usefulness feedback from searchers [11, 15, 20]. They found that usefulness feedback has a stronger correlation with user satisfaction. Although these studies have already gained much success in modeling user satisfaction, the effectiveness of usefulness in the context of mobile search has not been extensively investigated.

Discussions and Conclusions

To summarize, via a carefully designed user study, we collect users' search behaviors along with their explicit usefulness feedbacks for both the clicked and unclicked results in mobile search. Using the collected data, we investigate the relationships between usefulness feedbacks, ranking positions, relevance annotations, and click necessity annotations to address **RQ1**. We find that the ranking positions have an effect on the usefulness feedbacks of the results in initial viewports. While a moderate linear correlation is found between usefulness feedbacks and relevance annotations, we find that the presentation style of results, reflected by their click necessity, is a moderating factor of the relationship between relevance and usefulness. Regarding **RQ2**, we correlate the result-level measures with query-level user satisfaction and confirm in mobile environment that the usefulness feedbacks have a stronger correlation with user satisfaction than the relevance annotations, showing a potential limitation of system-oriented evaluation in estimating actual user satisfaction. Regarding **RQ3**, we use ANOVA tests and regression models to examine the relationships between usefulness feedbacks and user's search behaviors, especially the viewport time of snippets on mobile SERPs. The results suggest that the viewport time can be a useful feature in estimating usefulness in mobile search. Because the usefulness feedback can: 1) better reflect user satisfaction in mobile search; 2) be estimated by search behavior features, it is promising to be adopt in the user-oriented evaluation of mobile search engines.

References

1. Belkin, N.J., Cole, M., Liu, J.: A model for evaluation of interactive information retrieval. In: Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation. pp. 7–8 (2009)
2. Cleverdon, C., Mills, J., Keen, M.: Aslib cranfield research project: factors determining the performance of indexing systems (1966)
3. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin 70(4), 213 (1968)

4. Cole, M., Liu, J., Belkin, N., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., Zhang, X.: Usefulness as the criterion for evaluation of interactive information retrieval. *Proc. HCIR* pp. 1–4 (2009)
5. Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T.: Evaluating implicit measures to improve web search. *ACM TOIS* 23(2), 147–168 (2005)
6. Granka, L.A., Joachims, T., Gay, G.: Eye-tracking analysis of user behavior in www search. In: *SIGIR'04*. pp. 478–479. ACM (2004)
7. Guo, Q., Jin, H., Lagun, D., Yuan, S., Agichtein, E.: Mining touch interaction data on mobile devices to predict web search result relevance. In: *SIGIR'13*. pp. 153–162. ACM (2013)
8. Guo, Q., Song, Y.: Large-scale analysis of viewing behavior: Towards measuring satisfaction with mobile proactive systems. In: *CIKM'16*. pp. 579–588. ACM (2016)
9. Harvey, M., Pointon, M.: Searching on the go: The effects of fragmented attention on mobile web search tasks. In: *SIGIR'17*. pp. 155–164. ACM (2017)
10. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)* 20(4), 422–446 (2002)
11. Jiang, J., He, D., Kelly, D., Allan, J.: Understanding ephemeral state of relevance. In: *CHIIR'17*. pp. 137–146. ACM (2017)
12. Joachims, T.: Optimizing search engines using clickthrough data. In: *KDD'02*. pp. 133–142. ACM (2002)
13. Kamvar, M., Baluja, S.: A large scale study of wireless search behavior: Google mobile search. In: *SIGCHI'06*. pp. 701–709. ACM (2006)
14. Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., Yoon, H.J.: Eye-tracking analysis of user behavior and performance in web search on large and small screens. *JASIST* 66(3), 526–544 (2015)
15. Kim, J.Y., Teevan, J., Craswell, N.: Explicit in situ user feedback for web search results. In: *SIGIR'16*. pp. 829–832. ACM (2016)
16. Lagun, D., Hsieh, C.H., Webster, D., Navalpakkam, V.: Towards better measurement of attention and satisfaction in mobile search. In: *SIGIR'14*. pp. 113–122. ACM (2014)
17. Lagun, D., McMahan, D., Navalpakkam, V.: Understanding mobile searcher attention with rich ad formats. In: *CIKM'16*. pp. 599–608. ACM (2016)
18. Liu, Y., Chen, Y., Tang, J., Sun, J., Zhang, M., Ma, S., Zhu, X.: Different users, different opinions: Predicting search satisfaction with mouse movement information. In: *SIGIR'15*. pp. 493–502. ACM (2015)
19. Luo, C., Liu, Y., Sakai, T., Zhang, F., Zhang, M., Ma, S.: Evaluating mobile search with height-biased gain. In: *SIGIR'17*. ACM (2017)
20. Mao, J., Liu, Y., Zhou, K., Nie, J.Y., Song, J., Zhang, M., Ma, S., Sun, J., Luo, H.: When does relevance mean usefulness and user satisfaction in web search? In: *SIGIR'16*. pp. 463–472. ACM (2016)
21. Ong, K., Järvelin, K., Sanderson, M., Scholer, F.: Using information scent to understand mobile and desktop web search behavior. In: *SIGIR'17*. pp. 295–304. ACM (2017)
22. Shokouhi, M., Jones, R., Ozertem, U., Raghunathan, K., Diaz, F.: Mobile query reformulations. In: *SIGIR'14*. pp. 1011–1014. ACM (2014)
23. Song, Y., Ma, H., Wang, H., Wang, K.: Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In: *WWW'13*. pp. 1201–1212. ACM (2013)
24. Verma, M., Yilmaz, E.: Characterizing relevance on mobile and desktop. In: *European Conference on Information Retrieval*. pp. 212–223. Springer (2016)
25. Voorhees, E.M., Harman, D.K., et al.: *TREC: Experiment and evaluation in information retrieval*, vol. 1. MIT press Cambridge (2005)
26. Williams, K., Kiseleva, J., Crook, A.C., Zitouni, I., Awadallah, A.H., Khabsa, M.: Detecting good abandonment in mobile search. In: *WWW'16*. pp. 495–505 (2016)