# Investigating Examination Behavior in Mobile Search

Yukun Zheng†, Jiaxin Mao†, Yiqun Liu†*, Mark Sanderson‡, Min Zhang†, and Shaoping Ma†
† Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
‡ RMIT University, Melbourne VIC 3000, Australia
zhengyk13@gmail.com,yiqunliu@tsinghua.edu.cn

## ABSTRACT

Examination is one of the most important user interactions in Web search. A number of works studied examination behavior in Web search and helped researchers better understand how users allocate their attention on search engine result pages (SERPs). Compared to desktop search, mobile search has a number of differences such as fewer results on the screen. These differences bring in mobile-specific factors affecting users' examination behavior. However, there still lacks research on users' attention allocation mechanism via viewports in mobile search. Therefore, we design a lab-based study to collect user's rich interaction behavior in mobile search. Based on the collected data, we first analyze how users examine SERPs and allocate their attention to heterogeneous results. Then we investigate the effect of mobile-specific factors and other common factors on users allocating attention. Finally, we apply the findings of user attention allocation from the user study into click model construction efforts, which significantly improves the state-of-the-art click model. Our work brings insights into a better understanding of users' interaction patterns in mobile search and may benefit other mobile search-related research.

## KEYWORDS

Mobile search, examination behavior, eye tracking, mobile viewport

**ACM Reference Format:**
Yukun Zheng†, Jiaxin Mao†, Yiqun Liu†*, Mark Sanderson‡, Min Zhang†, and Shaoping Ma†. 2020. Investigating Examination Behavior in Mobile Search. In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM '20), February 3–7, 2020, Houston, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3336191.3371797

## 1 INTRODUCTION

In the past decades, the development of Web search is inseparable from the understanding of user interactions. Among all user interaction behaviors, examination is one of the most important. Better understanding how users examine search engine result pages (SERPs) can greatly improve the effectiveness of search. Eye-tracking is a commonly used technique for analyzing user examination behavior in Web search. Eye gaze behavior is usually considered as an important signal for user attention [18, 19]. By collecting eye-tracking data, a number of existing works [6, 9, 12, 17, 28, 30] studied users' examination behavior in Web search, but most of them focused on the desktop search environment.

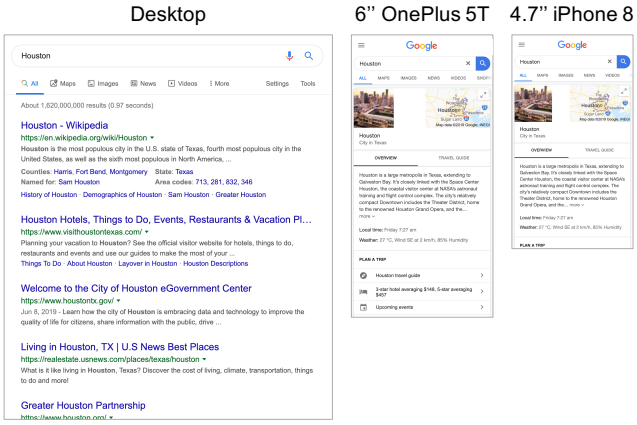



**Figure 1: Examples of Google SERPs on desktop and two mobile phones with different screen sizes. The query is "Houston" and only the initial viewports are shown.**

However, there exist great differences in search behavior on mobile, tablet and desktop devices [32]. Figure 1 shows the Google SERPs of the same query on desktop and mobile with different screen sizes, where a OnePlus 5T is the mobile phone used in our user study. Compared to desktop search, mobile search takes place on a much smaller screen where fewer results can be examined in the viewport[1]. Recently, search results in mobile search became more heterogeneous and have various presentation styles. Some results with rich information in their snippets may help users easily obtain relevant information without any click, which causes so-called *click necessity bias* [27]. All these differences make the patterns of user examination behavior found in desktop search potentially inapplicable in mobile search.

Several existing works already paid attention to the examination behavior in mobile search. Based on experiments with eye-tracking devices, Kim et al. [18] compared the differences in user behaviors

*Corresponding author.


*WSDM '20, February 3–7, 2020, Houston, TX, USA*

ACM ISBN 978-1-4503-6822-3/20/02…$15.00
https://doi.org/10.1145/3336191.3371797

[1]The visible portion of SERP.

on large and small screens, including fixation duration, click pattern, scanning path and etc., but they simulated the phone by shrinking the desktop browser window and used the same SERPs on both large and small screens. Lagun et al. [19] conducted a user study on a 4.7-inch mobile phone and found that the user attention in mobile search is generally focused on the top half of the screen. Based on a similar experimental setting, Lagun et al. [20] also tried to infer user gaze from viewport data using a linear model and some non-parametric methods with hand-crafted features. Nowadays, the screen sizes of mobile phones become bigger and are stable at around 6 inches because the phone sizes are now close to users' limit of one-handed operation. For example, the iPhone Xs has a 5.8-inch screen and the Samsung S10 has a 6.1-inch screen. With such a large mobile screen, the user can examine more area of SERPs in a viewport than before, as shown in Figure 1. To the best of our knowledge, there is still a lack of research to investigate users' examination behavior on a relatively larger mobile screen. Therefore, we propose our first research question:

- **RQ1:** How do users examine SERPs and allocate attention when searching on a mobile device with a large screen?

Search is a complex process of information cognition and seeking, during which user attention is usually biased by a number of factors. According to the previous works in both desktop and mobile search, the factors include result position [18], result presentation style [20, 24], result quality [3], result relevance [23] and etc. Besides these factors, the differences between mobile and desktop search bring more potential factors into our sights, such as the position in a viewport (e.g., located in the upper, middle or lower part of a viewport) and the click necessity[2] [25] of results. We would like to investigate the effect of these new factors as our second research questions:

- **RQ2:** What factors affect users' attention allocation mechanism in mobile search?

Accurately modeling user attention during a search process can help improve a number of IR-related tasks, such as UI design of search engines [29], user satisfaction prediction [19], result ranking [13], click prediction [38], and etc. We would like to know whether our findings of user examination patterns can be applied in practical mobile search tasks. Thus, we propose our third research question:

- **RQ3:** Can our findings of users' attention allocation mechanism be adopted in improving practical search applications?

To address these three research questions, we conduct a lab-based user study to collect rich user interaction data and various feedback annotations in mobile search. Based on the collected data, we first analyze the patterns of user attention allocation within a viewport. Different from previous works, we find that users' attention focuses on different parts of the screen in different stages of the search process. Second, we investigate the impact of several important factors on user attention during a search process, including position, result presentation, click necessity, viewport coverage, result exposure and etc. Finally, we choose the click prediction task as the practical application to answer RQ3 and show the effectiveness of our findings in improving the performance of click models.

[2]How much necessary it is for a user to click on a result to obtain its relevant information.

## 2 RELATED WORK

Examination in Web search always attracts much attention in Information Retrieval (IR) research. Researchers first studied user examination behavior in desktop search. Granka et al. [12], Richardson et al. [30] and Joachims et al. [17] looked into user's basic eye movements and scan patterns during search tasks with eye-tracking devices. Dumais et al. [9] focused on searchers' interactions with the whole search engine result page instead of individual components and investigated individual differences in gaze patterns for web search. Diaz et al. [8] tried to use cursor movements to estimate user visual attention on the components of SERPs. Wang et al. [34] found result presentation styles have different effects on user examination behavior for vertical results and for the whole result list. Liu et al. [23] proposed a two-stage examination model for Web search consisting of a "from skimming to reading" stage and a "from reading to clicking" stage. Liu et al. [24] investigated the influence of vertical results in Web search examination and revealed the existence of a vertical attraction effect, an examination cut-off effect and an examination spill-over effect. Li et al. [22] conducted a user study to investigate reading attention and proposed a two-stage reading model for relevance judgment. Existing works found several user behavior biases in Web search, such as position bias [7], attractiveness bias [1, 37] and domain bias [16].

In the research line of user examination behavior in mobile search, Lagun et al. [19] is one of the first researchers to apply eye-tracking into mobile search. They found eye gaze behavior has a strong correlation with user satisfaction in mobile search and proposed to utilize viewport metrics as an alternative of user attention in estimating satisfaction. Lagun et al. [20] then looked into sponsored search and found that rich ad formats can improve the user experience and tried several methods to infer user attention from viewports based on a per-element basis. When searching on a mobile device, it can be considered that users are examining the SERP by the trail of viewports [25]. Wang et al. [35] investigated examination behavior in mobile search based on large-scale search logs with viewport information and found that click positions mostly happen in the top two-third portion of the viewport.

User behavior is widely used in improving mobile search. Guo and Agichtein [14] found that post-click behavior, including mouse movement and scrolling, can help better estimate document relevance. Guo et al. [15] showed that touch interaction data on mobile devices can effectively predict result relevance. White et al. [36] adopted the information of SERP visual layout, cursor movement and viewport changing behavior to predict real-time document prefetching decisions. Shokouhi and Guo [31] used viewport duration and user clicks to infer result relevance in proactive systems such as Microsoft Cortana.

A more accurate estimation of examination usually brings improvements to click models [5]. Mao et al. [27] proposed that user search behavior is biased by click necessity and examination satisfaction in mobile search and modeled the findings into the Mobile Click Model (MCM). Zheng et al. [38] proposed a Viewport Time

Click Model (VTCM) to incorporate the viewport time information into the modeling of users' click behavior in mobile search. In VTCM, the modeling of viewport time takes examination behavior, user clicks and examination satisfaction into account and adopts four independent conditional probabilities for different conditions as follows:

$$P(V_i = t_i|E_i = 0) = f_{v_i}^{E=0}(t_i) \quad (1)$$

$$P(V_i = t_i|E_i = 1, C_i = 0, S_i^E = 0) = f_{v_i}^{E=1,C=0,S^E=0}(t_i) \quad (2)$$

$$P(V_i = t_i|E_i = 1, C_i = 1, S_i^E = 0) = f_{v_i}^{E=1,C=1,S^E=0}(t_i) \quad (3)$$

$$P(V_i = t_i|E_i = 1, C_i = 0, S_i^E = 1) = f_{v_i}^{E=1,C=0,S^E=1}(t_i) \quad (4)$$

where $C_i$ and $V_i$ are the click and viewport time events of the $i$-th result in a search session, $t_i$ is the observed viewport time of this result, $E_i$ is its examination event, $S_i^E$ is the examination satisfaction event, and $v_i$ is its vertical type, $f$ is a probability density function of a certain distribution. They follow Lagun et al. [19] to calculate the viewport time for each result using the ratios of *viewport coverage* and *result exposure* as weights:

$$t = \sum_{j=1}^{n} t_j * \frac{h_j^e}{h_j^v} * \frac{h_j^e}{h_j^r} \quad (5)$$

where $t$ is the weighted viewport time of a result in a search session, $n$ is the viewport number in the session, $t_j$ is the duration of the $j$-th viewport, $h_j^e$ is the visible height of the result exposed in the $j$-th viewport, $h_j^r$ is the total height of the result, $h_j^v$ is the height of the $j$-th viewport. $h_j^e/h_j^v$ represents how much the result occupies the viewport (viewport coverage) and $h_j^e/h_j^r$ represents how much the result is visible to the user (result exposure).

## 3 USER STUDY

To investigate users' examination behavior during searching on mobile devices, we conduct a lab-based user study with 60 search tasks. In this section, we describe the details of the user study and the dataset we collected. Table 1 shows the statistics of the user study dataset[3]. The dataset consists of tasks, user behavior, user explicit feedback, and crowdsourcing annotations.

### 3.1 Tasks and Participants

We manually sample 60 queries with high or intermediate frequency from search logs of a commercial search engine, *Sogou.com*, including 10 navigational, 25 informational and 25 transactional queries. We don't sample long-tail queries because they may contain user privacy and require much context information or knowledge to understand the search intent. For each query, we construct a description of search background to make the search intent more clear and unambiguous. We crawl the first two SERPs of these queries from *Sogou.com* on May 8th, 2019. We remove the query suggestion and sponsored results. About 22 results are reserved for a search task on average. Pagination and query reformulation are not allowed. We randomly divide the 60 tasks into three equal groups.

[3]We publicly released this dataset at http://www.thuir.cn/data-wsdm20-UserStudy/

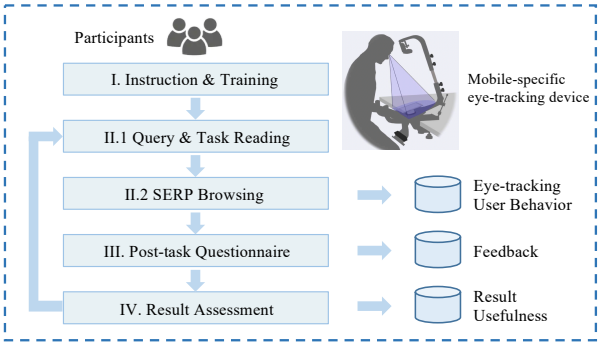



**Figure 2: The procedure of our user study.**[4]

**Table 1: The statistics of the dataset in our user study.**

| #Users | #Tasks | #Valid sessions | #Unique results |
|---|---|---|---|
| 13 | 60 | 478 | 1308 |

We recruit 13 undergraduate and graduate students (6 females and 7 males, aged from 18 to 26) from a university via online forums and social networks. All participants are familiar with mobile devices and have daily mobile search experience. To ensure the validity of collected eye movements, we screen the applicants based on their eyesight. Each participant needs to accomplish one to two groups of search tasks. It takes about 1 hour to complete one search group. For their involvement, the participants will be paid about 15 dollars.

### 3.2 Procedure

The procedure of our user study is shown in Figure 2. In the beginning, we introduce the experiment to participants by guiding them to complete two training tasks. Next, they are required to accomplish one or two groups of search tasks independently. The system shows 20 tasks of a certain group to the participant one by one in random order. The procedure for each search task is as follows:

**Searching with the given task.** The participants are first presented with the search query and background description of the task. After reading, they are required to search with the given query on the provided mobile phone. The phone is fixed on a stand to allow eye tracking. Participants can hold the phone but can't move or rotate it. Then the pre-processed SERP will be shown to participants. They can freely examine and interact with the results. At the same time, their eye movements are collected by the eye tracker and other interactions are recorded by our system. We collect the eye-tracking including both fixations and saccades, clicks, viewport information, and etc. Once the participants feel satisfied or the provided SERP can not satisfy the information needs, they can exit by clicking on a "finish searching" button.

**Post-task questionnaire.** After browsing the provided SERP, the participants are required to complete a post-task questionnaire.

[4]The diagram of mobile-specific eye-tracking device is from Tobii support document (https://www.tobii.com)

We collect explicit feedback on the experience of searching, including the binary question *is_successful* and a five-grade *satisfaction* response.

**Result assessment.** In this step, the SERP is presented to participants again and they are asked to annotate usefulness for the search results. The usefulness labels are on a four-level scale (1: useless, 2: somewhat useful, 3: fairly useful, 4: extremely useful).

Since the analysis of this work doesn't involve users' feedback from the post-task questionnaire and the usefulness annotations for results, we don't describe the details of the post-task questionnaire and the result assessment step here.

## 3.3 Experiment System and Platform

We conduct the user study on an Android smartphone, OnePlus 5T. It is equipped with a 6-inch screen and the resolution is 412 x 824 in density-independent pixels, which is the mainstream specification of smartphones in recent years. We develop a mobile app using Java, through which participants can log in to the system of our user study and complete search tasks. We use a backend database to record participants' interaction behavior including click, scrolling, etc. We use a Tobii X2-30 eye tracker to record the eye movements of participants as they browse the SERPs. To ensure that the eye movements are recorded accurately, a calibration process is taken for each participant before they begin the user study.

## 3.4 Search Result Annotation

After collecting user search behavior and explicit feedback from the participants, we further annotate the *page relevance*, *snippet relevance*, *click necessity*, and *result type* for each search result through crowdsourcing. The page relevance is measured based on the whole landing page of the result and the snippet relevance is judged according to the snippet of the result in the SERP. For page relevance and snippet relevance, we use a four-grade relevance judgment scale (1: irrelevant, 2: marginally relevant, 3: relevant, 4: highly relevant) according to the TREC criterion [33]. For click necessity, we use the three-grade necessity judgment (1: not necessary, 2: fairly necessary, 3: definitely necessary) [25]. For result type, we categorize all results into 18 result types according to their vertical types and presentation styles, such as news, encyclopedia, video, direct answer and so on. Each annotation is judged by three professional workers and we perform quality inspection on the annotations to ensure the reliability of these annotations. The values of Fleiss' $\kappa$ [11] of page relevance, snippet relevance, click necessity and result type are 0.847, 0.725, 0.628, and 0.912 respectively, all of which reach a substantial or almost perfect interpersonal agreement [21]. For page relevance, snippet relevance and click necessity, if there is a disagreement between judgments of a result, we use the median as the final judgment. For result type, we use majority voting to determine the result type. If all three annotations are different from each other, an external expert will be involved to judge and make a decision.

# 4 EXAMINATION IN MOBILE SEARCH

Based on the collected data in the user study, we would like to investigate the users' attention allocation patterns in mobile search.

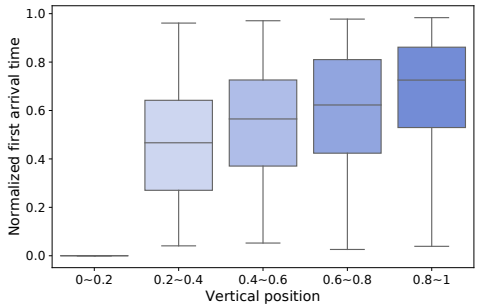



**Figure 3: The normalized first arrival time at different vertical positions of SERPs.**

**Table 2: The statistics of viewport and examination behavior at different viewport offsets.**

| Viewport offset | 0 | 0∼0.5H | 0.5∼1H | 1∼2H | >2H |
|---|---|---|---|---|---|
| #Avg. viewports per session | 2.19 | 1.18 | 0.89 | 0.85 | 1.53 |
| #Avg. visible results per viewport | 3.31 | 3.70 | 3.63 | 3.86 | 4.13 |
| #Avg. examined results per viewport | 1.41 | 1.07 | 0.99 | 0.94 | 0.90 |
| Avg. duration per viewport (ms) | 2427 | 1344 | 1095 | 1089 | 977 |
| Avg. reading time per viewport (ms) | 1568 | 895 | 713 | 662 | 581 |

Since users allocate attention via the trail of viewports, we separate **RQ1** into two sub-questions:

- **RQ1a:** What are the patterns of users to examine mobile SERPs?
- **RQ1b:** How do users pay attention in the viewports of SERPs?

## 4.1 Examination Behavior Patterns

To address **RQ1a**, we investigate examination order by measuring the first arrival time at different vertical positions of the SERP, see Figure 3. We normalize the vertical positions of a SERP by the max exposed depth in a session and normalize the first arrival time at different vertical positions by the total session duration. It shows that the first arrival time increases with the vertical position, indicating that on average, the user browses the SERP in a top-down order, which is consistent with the assumption of Mobile Click Model (MCM) [27]. Further, we examine the viewport transition behavior. By recording the viewport offset in the SERP, we can get the viewport transition directions, including *forward* to view the results with lower ranks and *backward* to a higher position of the SERP, and calculate the sliding distances between two adjacent viewports. The ratio of forward is 86.9%, which is much larger than that of backward (13.1%), which verify that users tend to examine SERPs from the top to the bottom. In our user study, the height of a SERP viewport is 789 dp, which we label it as $H$. The average sliding distances of forward and backward are 459.1 dp ($SD$=1183.1) and 611.6 dp ($SD$=1106.3) respectively, indicating that users usually slide forward for about a half viewport height on average to view the next results and sometimes move backward for a rather long distance to visit relatively higher-rank results.

Table 2 shows the statistics of the viewport and eye-tracking behavior in the user study. We view fixation duration as reading time and treat a result with more than 200 ms reading time as an examined result according to [23, 26]. The result with less than 200ms reading time is regarded as not examined. As the viewport

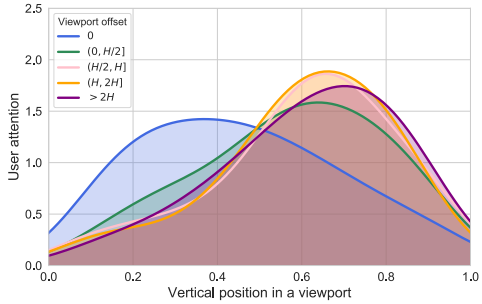



**Figure 4: The vertical viewport attention distributions at different offsets in the SERPs, where $H$ is the height of a viewport.**

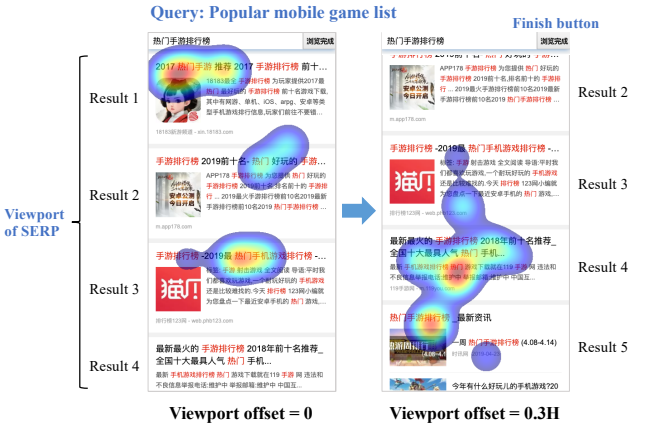



**Figure 5: The interface of our system and the user attention heatmaps of two viewports in the same search session, where $H$ is the height of a viewport.**

offset increases, the average viewport number in a session decreases. Users stay at the top of the SERP for a longer time than other viewports and spend more time reading more results. After, users examine about one result per viewport on average and allocate diminishing attention when browsing deeper in the SERP. The average number of examined results per session is 4.47 ($SD$=3.43), while the average number of unexamined results before the last examined result is 1.13 ($SD$=2.46), indicating that users may skip some results during the search processes.

## 4.2 Viewport Attention

To address **RQ1b**, We examine *viewport attention* during search tasks. Figure 4 shows the vertical attention distributions within a viewport with different viewport offsets. The attention distribution with zero viewport offset is significantly different from others with $p < 0.01$ using an unpaired t-test, while the attention distributions with more than zero viewport offsets are highly similar. With the finding of user sequential browsing in mobile search, we can see that after the user examines the initial viewport, the focus of attention transfers from the top to the bottom half. This indicates that the viewport attention is biased by two factors at the same time, the viewport offset and the vertical position in a viewport. Figure 5 shows the interface of our system and the eye-gaze heatmaps in two viewports of the same search session. In the left viewport with zero offset, the user mainly examined the first three results and didn't pay attention to the fourth result. In the right viewport whose offset is about 0.3$H$, the user spent more time to examine the new results at the bottom half of the viewport. On the one hand, the results at the bottom half are usually new to users. On the other hand, the bottom half is more convenient for users to click or make other interactions if they want with their fingers when they are holding the phone.

## 4.3 Summary

To answer **RQ1a** and **RQ1b**. We found that, during mobile search, the user tends to first pay attention to the top half of the screen and carefully examines several top-ranked results. Then, if they would like to continue browsing, they usually scroll down and focus on the bottom half of the screen to view one new result in the next viewport and repeat this examination action with occasional revisiting or skipping some results until entering another page or leaving.

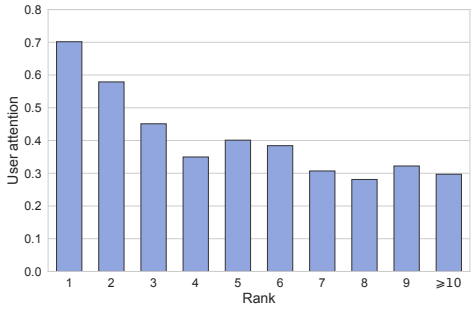



**Figure 6: The average user attention per viewport at different results ranks.**

# 5 VIEWPORT ATTENTION BIAS

To address **RQ2**, we try to investigate what factors affect the viewport attention. Since results that are taller in the SERP are more likely to get more eye gaze, we normalize the result reading time by result height. We use this normalized reading time as the measure of user attention during search tasks. Note that the reading time is in milliseconds (ms) and the result height is in density independent pixel (dp).

## 5.1 Result Rank

Figure 6 shows the average user attention at different results ranks per viewport. We can see that the first two ranks usually attract more attention. The top results are usually highly relevant to a query and can often contain an answer to the query or the result may consist of multiple images and links, making them attractive and causing more user attention. However, from rank three down, the levels of average user attention per viewport become similar with little decay, showing that it takes at least a certain time for users to read a result in the viewport.

**Table 3: The average user attention per viewport on results with different levels of viewport coverage and result exposure.**

| **Coverage** | 0~0.2 | 0.2~0.4 | 0.4~0.6 | 0.6~0.8 | 0.8~1 |
|---|---|---|---|---|---|
| Attention | 0.331 | 0.500 | 0.560 | 0.697 | 0.789 |
| Frequency | 4604 | 7160 | 1249 | 309 | 242 |
| **Exposure** | 0~0.2 | 0.2~0.4 | 0.4~0.6 | 0.6~0.8 | 0.8~1 |
| Attention | 0.123 | 0.223 | 0.309 | 0.330 | 0.554 |
| Frequency | 1066 | 989 | 1034 | 1173 | 9302 |


**Figure 7: Snippet examples of several popular result types shown in SERPs.**

**Table 4: The average user attention per viewport on different result types and the statistics of these result types.**

| Result type | Attention | #Words | Img. coverage | Height (dp) |
|---|---|---|---|---|
| Organic | 0.464 | 67.3 | 0% | 447 |
| Encyclopedia | 0.440 | 104.5 | 9.5% | 550 |
| Direct answer | 0.577 | 125.7 | 14.6% | 912 |
| Video | 0.606 | 301.1 | **28.1%** | 1232 |
| Aggregation | 0.648 | **617.6** | 12.0% | **2453** |
| Application | **1.457** | 207.9 | 2.2% | 937 |

## 5.2 Viewport Coverage and Result Exposure

Second, we examine the ratio of viewport coverage and result exposure (See the definitions in Equation 5). Table 3 shows the average user attention per viewport on results with different levels of viewport coverage and result exposure. We can see that with the increase of results' viewport coverage, users tend to pay more attention to examining them. However, about 86.7% results occupy less than 40% area of the viewport. Results with more than 40% viewport coverage are mostly a long result with a specially-designed snippet and abundant information in the SERP and hence usually attract more user attention. For the aspect of result exposure, we can see that the most of the results in viewports have a more than 80% visible area, which is more likely to be the focus of users in viewports. Our results support the findings in Lagun et al. [19] that it is effective to calculate the viewport time of a result weighted by its viewport coverage and result exposure.

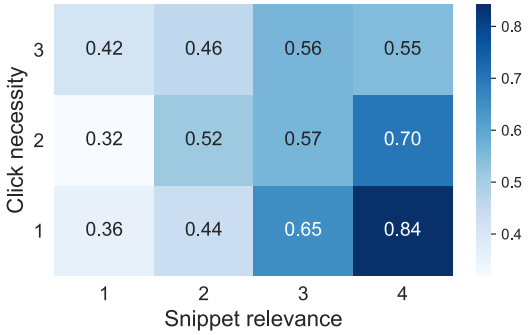



**Figure 8: The average user attention per viewport on results with different levels of snippet relevance and click necessity.**

## 5.3 Result Type and Presentation Style

Current mobile search engines tend to retrieve heterogeneous results from multiple sources for a query and design different presentation styles for these results. We investigate six popular result types with different presentation styles to see how users pay attention to them. Figure 7 shows the snippet examples of these results types. Table 4 shows the average viewport attention on different result types in mobile search. Direct-answer results provide an answer box to directly answer users' question, while aggregation results contain multiple pieces of vertical information which is highly relevant to the query and hence have a rather large result height. Video results consist of several images and have a large ratio of image coverage. Users spend more time examining these result types than organic results which only contain a title and a short snippet text. Application results allow users to directly interact with them on the SERP to get useful information, such as flight inquiry and exchange rate calculation. Therefore, application results attract the most user attention among all result types.

## 5.4 Snippet Relevance and Click Necessity

Figure 8 shows the average user attention on results with different snippet relevance and click necessity. We obtain the snippet relevance and click necessity of a result from crowdsourcing. We can see that users pay more attention to relevant and highly relevant results than irrelevant and marginally relevant results. Results with low click necessity are likely to attract more attention from users than results with high click necessity. In general, the highly relevant results with low click necessity attract the most user attention among all combinations of snippet relevance and click necessity. Since users browse SERPs with the aim to satisfy their information need, they usually look for the relevant results. Results with low click necessity can provide users with abundant information with little interaction effort, so these results attract much user attention when they are relevant to the query.

## 5.5 Summary

In summary, these factors can be classified into two categories: (1) general factors whose effect on user attention allocation have been investigated in desktop search, such as result rank, result type and presentation style and result relevance, and (2) mobile-specific factors including viewport coverage, result exposure and click necessity. Our results show the viewport attention doesn't have an obvious decay from the third rank to lower ranks. In addition, the lower click necessity of a result usually leads results to a higher level of user attention when the result is relevant.

# 6 APPLICATION OF THE FINDINGS

To answer **RQ3**, we choose the click prediction task and perform two experiments based on click models. In the first experiment, we try to introduce the patterns of viewport attention distribution into the calculation of results' viewport time and compare the performances of VTCM under different calculation methods of viewport time. In the second experiment, we incorporate the effect of result rank on user attention into the modeling of viewport time in VTCM and compare our method with the original VTCM and other baseline click models.

## 6.1 Dataset

We sample search sessions from search logs of a Chinese commercial search engine, *Sogou.com*, during the first eight days in August, 2018. The search log of a search session contains the query, ten URLs of search results, a 10-dimensional binary click vector, ten vertical type ids of the search results and viewport-related information, such as the position and the height of the viewport, the visible search results in the viewport and the viewport change events with timestamps [35]. The sampling procedure is: (1) We reserve the sessions whose resolution is similar to that of our mobile device (412×824 dp) with the width from 380-500 dp and the height from 680-1000 dp; (2) We remove search sessions with no clicks. After sampling, we split the sessions into two parts with the equal data size, using sessions of the first four days as the training set and those of the latter four days as the test set. Table 5 shows the statistics of the dataset used in the following experiments. We calculate the weighted viewport time for each result as same as [38] (Equation 5).

## 6.2 Click Model

In this study, we use four click models with probabilistic graphical model (PGM) framework and one click model with neural network framework as baselines:

- UBM: User Browsing Model proposed by Dupret and Piwowarski [10].
- DBN: Dynamic Bayesian Network model proposed by Chapelle and Zhang [4].
- MCM: Mobile Click Model proposed by Mao et al. [27].
- VTCM: Viewport Time Click Model proposed by Zheng et al. [38]. We use Weibull distribution to serve as $f$ in Equation 4 to model the viewport time, which is reported the best by the authors.
- NCM: Neural Click Model proposed by Borisov et al. [2]. We implement NCM with the LSTM configuration.

**Table 5: The statistics of the dataset used in click prediction.**

| #Unique queries | #Sessions | #Unique URLs | Date |
|---|---|---|---|
| 156,507 | 212,059 | 1,141,001 | Aug. 1st to 8th, 2018 |

**Table 6: The overall click prediction performance of click models measured in log-likelihood (*LL*) and average perplexity (*AvgPerp*). All differences over VTCM are statistically significant at $p < 0.01$ level, pairwise t-test, two-tailed, $n = 9,590$. $va$ represents that the model is trained and evaluated on the dataset with user attention-based viewport time. *rank* means that the model tasks advantage of result rank information into the modeling of viewport time.**

| Click model | *LL* | Impr. | *AvgPerp* | Impr. |
|---|---|---|---|---|
| DBN | -0.8478 | -4.11% | 1.0960 | -4.58% |
| UBM | -0.8487 | -4.22% | 1.0952 | -3.68% |
| MCM | -0.8209 | -0.80% | 1.0926 | -0.84% |
| NCM | -0.8193 | -0.61% | 1.0924 | -0.69% |
| VTCM | -0.8144 | - | 1.0918 | - |
| $\text{VTCM}_{va}$ | -0.8060 | 1.02% | 1.0911 | 0.79% |
| $\text{VTCM}_{rank}$ | -0.8094 | 0.61% | 1.0913 | 0.54% |
| $\text{VTCM}_{va+rank}$ | **-0.7974** | 2.08% | **1.0899** | 2.06% |

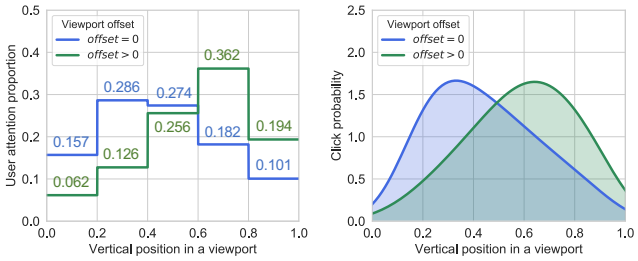
## 6.3 Experiments

*6.3.1 Baseline performance.* We train all click models on the training set until they converge completely and evaluate models on the sessions of the test set whose query appears more than 10 times in the training set. We use log-likelihood (*LL*) and average perplexity (*AvgPerp*) as metrics [27]. Table 6 shows the click prediction performances of baseline models. With the additional viewport time information, VTCM achieves the best performance in both *LL* and *Avgperp* metrics among five baseline models, followed by NCM and MCM, which is consistent with the results in [38].

*6.3.2 Viewport attention distribution.* We find that the peak of viewport attention transfers downward from the top half of the viewport to the bottom half when users browse deeper in a SERP. The reading time of two results with the same result height, viewport coverage and result exposure may consequently be different if they appear at different vertical positions of the viewports. We can utilize this finding to improve the calculation of results' viewport time in mobile search. We extend the calculation of a result's viewport time in a session as:

$$t = \sum_{j=1}^{n} \int_{start_j}^{start_j + h_j^e} W_j(h) * V_j(h) dh \tag{6}$$

where $start_j$ and $h_j^e$ are the offset and the visible height of the result in the $j$-th viewport, $V_j(h)$ is the attention distribution within the $j$-th viewport and $W_j(h)$ is the attention allocation weight for the result in the $j$-th viewport. For the method in Equation 5, $V_j(h)$ and $W_j(h)$ is $1/H$ and $h_j^e/h_j^r$ respectively. We introduce the viewport attention allocation patterns found in the user study into the design




(a) User attention proportion. (b) Vertical click position.

**Figure 9: (a) the proportion of user attention at five-level vertical positions of viewports in our user study and (b) the distribution of vertical click positions within viewports in our practical search logs.**

**Table 7: The average viewport time (in second) learned by $\text{VTCM}_{va+rank}$ and $\text{VTCM}_{va}$ for results with different ranks.**

| Condition | $\text{VTCM}_{va+rank}$ | | $\text{VTCM}_{va}$ |
|---|---|---|---|
| | $rank \leqslant 2$ | $rank > 2$ | |
| $E = 0$ | 2.88 | 0.96 | 0.93 |
| $E = 1, C = 1$ | 6.15 | 3.91 | 5.45 |
| $E = 1, C = 0, S_e = 0$ | 3.83 | 2.42 | 3.61 |
| $E = 1, C = 0, S_e = 1$ | 4.18 | 9.50 | 11.9 |

of $V_j(h)$ and split it into two conditions:

$$V_j(h) = \begin{cases} V_j^0(h) & offset = 0 \\ V_j^{>0}(h) & offset > 0 \end{cases} \tag{7}$$

We follow the method of Lagun et al. [20] to split the viewport into five equal parts according to the vertical position. Different from them, we don't use a learning model to predict the viewport attention distribution because our aim is just to examine the effectiveness of our findings instead of studying how to infer user attention from viewports. So we directly use the statistics of the viewport attention distribution in our user study for $V_j(h)$, as shown in Figure 9(a). We use our method to update the viewport time for all results in the training and test sets, and then train VTCM. The performance of VTCM on the updated dataset is reported in Table 6. We can see with the attention-based viewport time, the performance of VTCM gets significantly improved by 1.02% and 0.79% in *LL* and *AvgPerp* respectively, showing our viewport attention distribution found in the user study is effective in the click prediction task.

We examine the distribution of vertical click positions in the search logs. Figure 9(b) shows the distribution of vertical click positions within viewports in our search logs. We can find user clicks have the same patterns with viewport attention. Clicks occur more often at the top half of the viewport when the viewport offset is zero. When users slide down, more clicks happen at the bottom half of the viewport. Since a click on a result indicates the result was examined in most cases, the highly similar patterns between vertical click position and viewport attention can be used to support our findings as well as the improvement of $\text{VTCM}_{va}$.

*6.3.3 Result rank.* In Section 4, we found that users pay different attention on results with different ranks and results with higher ranks are likely to attract more user attention in mobile search. This inspires us to introduce the result rank information into the modeling of viewport time. We extend these conditional probabilities in Equation 4 by adding the result rank $r_i$ into them:

$$P(V_i = t_i | E_i = 0) = f^{E=0}_{v_i, r_i}(t_i) \tag{8}$$

$$P(V_i = t_i | E_i = 1, C_i = 0, S^E_i = 0) = f^{E=1, C=0, S^E=0}_{v_i, r_i}(t_i) \tag{9}$$

$$P(V_i = t_i | E_i = 1, C_i = 1, S^E_i = 0) = f^{E=1, C=1, S^E=0}_{v_i, r_i}(t_i) \tag{10}$$

$$P(V_i = t_i | E_i = 1, C_i = 0, S^E_i = 1) = f^{E=1, C=0, S^E=1}_{v_i, r_i}(t_i) \tag{11}$$

Here we describe the detailed implementations. From Table 2 and Figure 6, we find that results with higher ranks usually attract more attention, especially the results in the initial viewport of the SERP. Thus, we split the ten results in a session into two groups, the first $k$ results and the latter $10 - k$ results. We set $k$ to 2 in the following experiment with two considerations. On the one hand, Figure 6 shows that users tend to view about 1.41 results in the viewports of the top of SERPs; On the other hand, Figure 6 shows that users usually pay more attention on the first two results. Therefore, $r_i$ is a binary value indicating whether the result is at the first two ranks in the session. Table 6 shows the performance of the new VTCM model with result rank information, $\text{VTCM}_{rank}$. We can see it significantly outperform the original VTCM with 0.61% and 0.54% improvements in *LL* and *AvgPerp* metrics. We then train and evaluate $\text{VTCM}_{rank}$ on the dataset with attention-based viewport time. We can see that $\text{VTCM}_{va+rank}$ gets the best performance among all the click models in our experiment.

We examine the learned parameters of viewport time models in these models to see whether they learn in an effective way. Table 7 shows the average viewport time of results under different conditions estimated by $\text{VTCM}_{va+rank}$ and $\text{VTCM}_{va}$. These results show that users usually spend more time examining the first two results before clicking on them than the other lower-rank results. Besides, compared to the latter eight results, it takes a shorter time on average for users to feel satisfied after examining the first two results without any click. We consider this phenomena is because the first two results are usually highly relevant with abundant and useful information in the snippets. For the first two results in $\text{VTCM}_{va+rank}$, their mean viewport time in three of the four conditions is larger than that of the other results. From the parameters, we can see that $\text{VTCM}_{va+rank}$ can effectively model the viewport time as our findings.

## 7 CONCLUSION AND FUTURE WORK

In this paper, we investigate users' examination behavior in mobile search by conducing a lab-based user study. By analyzing the eye-tracking behavior and other interaction behavior from the user study as well as result annotations from crowdsourcing, we found several patterns of users' examination behavior on a rather large mobile screen. Users tend to first pay more attention to the top half of the initial SERP viewport to read the top-ranked results. When users start scrolling, the focus of user attention moves to the bottom half of the viewport, which is the first time to be observed in an eye-tracking study. We investigate the effect of several

mobile-specific and other common factors on attention allocation in viewports. We are the first to show the effect of click necessity on user attention allocation. Finally, we introduce the patterns of viewport attention allocation into the modeling of clicks, which improves the performance of a state-of-the-art click model and shows the potential of our findings to benefit the practical applications in mobile search.

In future work, we would like to incorporate more findings into click models or other related tasks to further improve their performance. For example, one may improve click models to model the relationship between the viewport time and the snippet relevance (or click necessity) of a result. We also want to investigate the generality of our findings to other mobile screen sizes. We may also better infer user attention on results or other components in the SERP from the viewport information using learning methods.

## ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700) and Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011).

## REFERENCES

[1] Judit Bar-Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. 2009. Presentation bias is significant in determining user preference for search results—A user study. *Journal of the American Society for Information Science and Technology* 60, 1 (2009), 135–149.

[2] Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A Neural Click Model for Web Search. In *WWW '16*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 531–541.

[3] Georg Buscher, Susan T. Dumais, and Edward Cutrell. 2010. The Good, the Bad, and the Random: An Eye-tracking Study of Ad Quality in Web Search. In *SIGIR '10*. ACM, New York, NY, USA, 42–49.

[4] Olivier Chapelle and Ya Zhang. 2009. A Dynamic Bayesian Network Click Model for Web Search Ranking. In *WWW '09*. ACM, New York, NY, USA, 1–10.

[5] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 7, 3 (2015), 1–115.

[6] Michael J. Cole, Chathra Hendahewa, Nicholas J. Belkin, and Chirag Shah. 2014. Discrimination Between Tasks with User Activity Patterns During Information Search. In *SIGIR '14*. ACM, New York, NY, USA, 567–576.

[7] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-bias Models. In *WSDM '08*. ACM, New York, NY, USA, 87–94.

[8] Fernando Diaz, Ryen White, Georg Buscher, and Dan Liebling. 2013. Robust models of mouse movement on dynamic web search results pages. In *CIKM '13*. ACM, New York, NY, USA, 1451–1460.

[9] Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual Differences in Gaze Patterns for Web Search. In *IliX '10*. ACM, New York, NY, USA, 185–194.

[10] Georges E. Dupret and Benjamin Piwowarski. 2008. A User Browsing Model to Predict Search Engine Click Data from Past Observations.. In *SIGIR '08*. ACM, New York, NY, USA, 331–338.

[11] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[12] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking Analysis of User Behavior in WWW Search. In *SIGIR '04*. ACM, New York, NY, USA, 478–479.

[13] Zhiwei Guan and Edward Cutrell. 2007. An Eye Tracking Study of the Effect of Target Rank on Web Search. In *CHI '07*. ACM, New York, NY, USA, 417–420.

[14] Qi Guo and Eugene Agichtein. 2012. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *WWW '12*. ACM, 569–578.

[15] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. 2013. Mining touch interaction data on mobile devices to predict web search result relevance. In *SIGIR '13*. ACM, 153–162.

[16] Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. 2012. Domain Bias in Web Search. In *WSDM '12*. ACM, New York, NY, USA, 413–422.

[17] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data As Implicit Feedback. In *SIGIR '05*. ACM, New York, NY, USA, 154–161.

[18] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2015. Eye-tracking Analysis of User Behavior and Performance in Web Search on Large and Small Screens. *J. Assoc. Inf. Sci. Technol.* 66, 3 (March 2015), 526–544.

[19] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards Better Measurement of Attention and Satisfaction in Mobile Search. In *SIGIR '14*. ACM, New York, NY, USA, 113–122.

[20] Dmitry Lagun, Donal McMahon, and Vidhya Navalpakkam. 2016. Understanding Mobile Searcher Attention with Rich Ad Formats. In *CIKM '16*. ACM, New York, NY, USA, 599–608.

[21] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[22] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. 2018. Understanding Reading Attention Distribution During Relevance Judgement. In *CIKM '18*. ACM, New York, NY, USA, 733–742.

[23] Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, and Shaoping Ma. 2014. From Skimming to Reading: A Two-stage Examination Model for Web Search. In *CIKM '14*. ACM, New York, NY, USA, 849–858.

[24] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2015. Influence of Vertical Result in Web Search Examination. In *SIGIR '15*. ACM, New York, NY, USA, 193–202.

[25] Cheng Luo, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Evaluating Mobile Search with Height-Biased Gain. In *SIGIR '17*. ACM, New York, NY, USA, 435–444.

[26] Barry R Manor and Evian Gordon. 2003. Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of neuroscience methods* 128, 1-2 (2003), 85–93.

[27] Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Constructing Click Models for Mobile Search. In *SIGIR '18*. ACM, New York, NY, USA, 775–784.

[28] Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. 2013. Measurement and Modeling of Eye-mouse Behavior in the Presence of Nonlinear Page Layouts. In *WWW '13*. ACM, New York, NY, USA, 953–964.

[29] Rachana S Rele and Andrew T Duchowski. 2005. Using eye tracking to evaluate alternative search results interfaces. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 49. SAGE Publications Sage CA: Los Angeles, CA, 1459–1463.

[30] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting Clicks: Estimating the Click-through Rate for New Ads. In *WWW '07*. ACM, New York, NY, USA, 521–530.

[31] Milad Shokouhi and Qi Guo. 2015. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *SIGIR '15*. ACM, 695–704.

[32] Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang. 2013. Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance. In *WWW '13*. ACM, New York, NY, USA, 1201–1212.

[33] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.

[34] Chao Wang, Yiqun Liu, Min Zhang, Shaoping Ma, Meihong Zheng, Jing Qian, and Kuo Zhang. 2013. Incorporating Vertical Results into Search Click Models. In *SIGIR '13*. ACM, New York, NY, USA, 503–512.

[35] Xiaochuan Wang, Ning Su, Zexue He, Yiqun Liu, and Shaoping Ma. 2018. A Large-Scale Study of Mobile Search Examination Behavior. In *SIGIR '18*. ACM, New York, NY, USA, 1129–1132.

[36] Ryen W White, Fernando Diaz, and Qi Guo. 2017. Search Result Prefetching on Desktop and Mobile. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 23.

[37] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond Position Bias: Examining Result Attractiveness As a Source of Presentation Bias in Clickthrough Data. In *WWW '10*. ACM, New York, NY, USA, 1011–1018.

[38] Yukun Zheng, Jiaxin Mao, Yiqun Liu, Cheng Luo, Min Zhang, and Shaoping Ma. 2019. Constructing Click Model for Mobile Search with Viewport Time. *ACM Trans. Inf. Syst.* 37, 4, Article 43 (Sept. 2019), 34 pages.