Constructing Click Models for Mobile Search

Jiaxin Mao, Cheng Luo, Min Zhang*, and Shaoping Ma
Department of Computer Science and Technology,
Beijing National Research Center for Information Science and Technology,
Tsinghua University
Beijing, China
maojiaxin@gmail.com,z-m@tsinghua.edu.cn

ABSTRACT

Users' click-through behavior is considered as a valuable yet noisy source of implicit relevance feedback for web search engines. A series of click models have therefore been proposed to extract accurate and unbiased relevance feedback from click logs. Previous works have shown that users' search behaviors in mobile and desktop scenarios are rather different in many aspects, therefore, the click models that were designed for desktop search may not be as effective in mobile context. To address this problem, we propose a novel Mobile Click Model (MCM) that models how users examine and click search results on mobile SERPs. Specifically, we incorporate two biases that are prevalent in mobile search into existing click models: 1) the click necessity bias that some results can bring utility and usefulness to users without being clicked; 2) the examination satisfaction bias that a user may feel satisfied and stop searching after examining a result with low click necessity. Extensive experiments on large-scale real mobile search logs show that: 1) MCM outperforms existing models in predicting users' click behavior in mobile search; 2) MCM can extract richer information, such as the click necessity of search results and the probability of user satisfaction, from mobile click logs. With this information, we can estimate the quality of different vertical results and improve the ranking of heterogeneous results in mobile search.

CCS CONCEPTS

• **Information systems → Web search engines**; *Users and interactive retrieval*; *Retrieval on mobile devices*;

KEYWORDS

Click Model; Mobile Search; Web Search

ACM Reference Format:
Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Constructing Click Models for Mobile Search. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210060

*Corresponding author

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210060

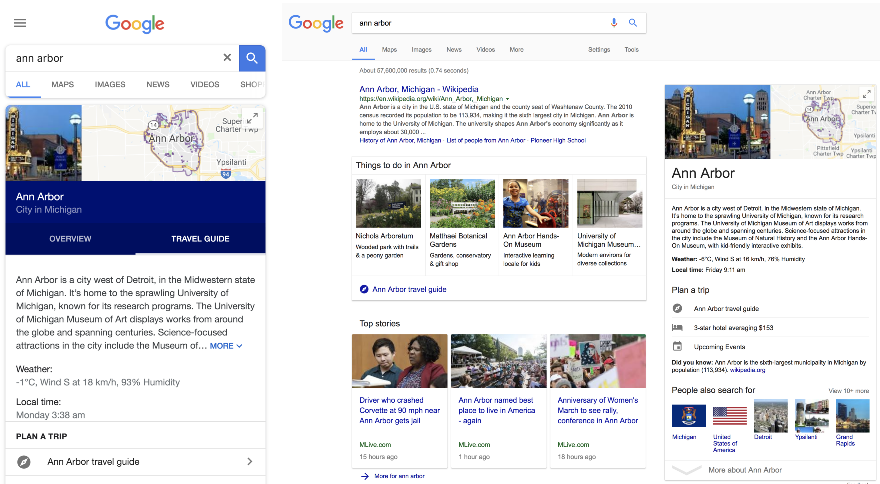
1 INTRODUCTION

Previous studies showed that user clicks can be used as implicit relevance feedback to improve the ranking of search results [13]. However, clicks on a result are inherently stochastic and systematically biased by factors such as the position [6, 13] and presentation style [2, 24] of the result. Therefore, a number of click models (see [3] for an overview) have been proposed to model users' click behavior as a stochastic process and obtain unbiased relevance feedback from the biased click logs.

The performance of a click model depends heavily on making correct assumptions on users' search behavior. By assuming a user will examine and click the results on the search engine result page (SERP) in a certain way, a click model can estimate how different kinds of biases affect users' click actions and derive unbiased relevance feedback from click logs. However, users' search behavior in the mobile environment are different from those in the desktop context. For example, previous studies suggest that users will pay more attention to the top-ranked results and scan fewer results on a small screen [16]; relevance judgments for documents are also affected by search devices [23]. Therefore, the existing click models originally designed for the desktop environment may not be as effective in the mobile search context. We need to refine the existing behavioral assumptions of click models to adapt to the shift from desktop to mobile.

One of the factors that may alter users' behavior in the mobile environment is the heterogeneity of search results. Today's search engines return richer results than the homogeneous ten blue links on both mobile and desktop. The heterogeneous results have a larger impact on users' interaction behavior on mobile SERPs because: 1) Compared to desktop search, direct answer and knowledge card results are federated into mobile SERPs more frequently. In many circumstances, these results present useful information on the SERP and users do not need to click the hyperlinks to visit the corresponding landing pages. While loading a page on mobile devices may take a longer time than on desktop devices, this strategy helps to reduce users' interaction costs as well as data usage on mobile. 2) Due to the limit of screen size, the heterogeneous results are usually injected into the main ranking list and often occupies a large proportion of user viewport. In a recent study, Luo et al. [19] showed these two factors may affect users' behavior in mobile search and proposed to incorporate them in the evaluation of mobile search engines.

As an example, we show two SERPs for the same query, *ann arbor*, on mobile and desktop in Figure 1. Compared with the desktop SERP in Figure 1b that displays the knowledge graph result on the right side, the knowledge graph result is placed at the first position in the mobile SERP (Figure 1a) and occupies almost the whole initial viewport. This result is highly likely to be examined by



(a) Mobile SERP (b) Desktop SERP

Figure 1: Examples of SERPs on (a) mobile and (b) desktop from Google. Only the content in the initial viewport is shown.

users and affect their following actions. The knowledge graph result contains a brief introduction to the city, as well as the information about the weather and local time. A user who wants to gather some basic information about Ann Arbor will find the knowledge graph result relevant and useful even without clicking it. She may even feel satisfied and leave the SERP just after examining the first knowledge graph result. In this case, an existing click model will: 1) mistakenly regard the skipping (i.e. no click) behavior on the first result as a negative relevance feedback; 2) ignore the cut off effect [18], that the user can be satisfied with the non-clicked knowledge graph result, but still assume the user will scan the following results.

While some studies [2, 4, 24] tried to incorporate the heterogeneity of search results into click models in the desktop environment, they mainly focused on modeling the *presentation bias* or *attetion bias* but ignored the *click necessity bias*. Chen et al. [2] found that users are more likely to examine the vertical result and more likely to be satisfied if they click it. Chuklin et al. [4] assumed that the probability that a user will examine a result is determined by her intent and the type of the result. Based on the findings in an eye-tracking study, Wang et al. [24] further assumed that the vertical result will affect not only the examination probabilities but also the examination order of search results. However, none of these existing efforts considered the situations where some results provide sufficient information on SERPs and thus are less likely to be clicked.

To address this problem in the mobile search context, we propose a novel click model named Mobile Click Model (MCM). The proposed MCM assumes that: 1) Some types of search results (e.g. the knowledge graph and direct answer results) have lower click necessity than others, which means that they can fulfill users' information needs without requiring any clicks (click necessity bias); 2) A user can be satisfied after *examining* a search result with low click necessity because this kind of results are designed to satisfy users' common information needs directly on SERPs (examination satisfaction bias). We will further introduce how we incorporate these two biases into the proposed model in Section 3.

Through extensive experiments on a large-scale mobile search log from a popular commercial search engine in China, we show that the proposed model can effectively infer the parameters for click necessity and examination satisfaction, along with the parameters for relevance and click satisfaction, from users' interaction logs with heterogeneous mobile SERPs. With these parameters learned from logs, we can: 1) improve the ranking of heterogeneous results in mobile search; 2) analyze how users interact with a certain type of vertical results. The experiment results also show that MCM achieves better performance in both click prediction and relevance estimation tasks than the baseline click models which are not specifically designed for the mobile environment.

The rest of the paper is organized as follows: We first provide an overview of the background of mobile search and click models in Section 2. In Section 3, we will formally introduce MCM model and compare it with existing click models. We then present the experiment setup and results in Section 4. Finally, we conclude the paper and discuss directions for future work in Section 5.

2 RELATED WORK AND BACKGROUND

2.1 Search Behavior on Mobile

With the rise of mobile search, understanding users' search behavior on mobile devices becomes increasingly important. Existing research has characterized the differences between desktop search and mobile search in various aspects.

First, compared to desktop search, mobile search is often conducted to fulfill different types of information needs, in diverse contexts. Yi et al. [27] and Kamvar et al. [15] are among the first who spotted a difference in the distribution of query categories across difference search devices. Song et al. [22] further found that the information needs of mobile searchers varied at the different time of the day. They also showed that a mobile user tended to search at different locations and users' click preferences changed with the search devices. Recently, Harvey and Pointon [11] suggested that users often used mobile devices to search in an "on the go" context, where they might be interrupted or distracted. They conducted a user study to assess the impact of these "fragmented attention" situations on the users' search behavior and performance.

The differences in search contexts and information needs on mobile and desktop suggest that the mobile search engine should return different results to satisfy mobile searchers. Therefore, it is crucial to develop new methods to extract relevance feedbacks from mobile search logs.

Second, the user interface (UI) of mobile search is very different from that of desktop search. Unlike a desktop PC with a large display (13 to 30 inches) as well as a mouse and a keyboard as input devices, a mobile phone usually has a much smaller screen (4 to 5 inches) and responds to a variety of touch interactions, including swiping, zooming, and on-screen text input. Previous works studied how the differences in UIs affect users' search behavior on mobile and desktop. Regarding the differences in input interactions, Kamvar and Baluja [14] and Song et al. [22] showed that while the query length was not significantly different on mobile and desktop, the mobile searcher tended to issue fewer queries in a session than the desktop searcher; Guo et al. [10] proposed to use the mobile touch interactions as features to estimate the relevance of mobile search results and identified some similarities and differences between user's fine-grain interactions on the landing pages in both desktop and mobile environments. On the other hand, the difference in screen size may impose more efforts for the mobile searchers to gather the same amount of information. Kim et al. [16] conducted an eye-tracking study to compare users' SERP scanning patterns on small screens and large screens. They found that on small screens, users put more attention to top-ranked results and exhibited a more linear scanning pattern. Recently, Ong et al. [20] found that users used different search strategies to adapt to the SERPs with varying Information Scent Levels and Information Scent Patterns [26] on mobile and desktop. These studies showed that users' search behavior on mobile devices was different from that in traditional desktop settings, therefore the click models that were originally designed to model users' click behavior in desktop search need to be adapted for mobile environment.

Third, today's mobile search engines will return more diverse results to cope with some specific information needs (e.g. checking the weather forecast or looking for a restaurant nearby) and reduce users' interaction cost in mobile environment. These heterogeneous *vertical* results may alter users' search behavior on mobile. For desktop search, Liu et al. [18] conducted a dedicated eye-tracking study to analyze the effects of different types vertical results on users' examination and click behavior on SERPs. For mobile search, Lagun et al. [17] studied how knowledge graph results affected users' attention and satisfaction. Their results showed that when a relevant knowledge graph result was presented, the user would pay less attention to the results below it, spend less time on the whole SERP, and feel more satisfied in the search. They also used an eye-tracker to measure users' gaze time on each search result and found that users paid more attention to the second and third results than the first results in mobile search, which is different from the findings in the eye-tracking studies conducted in desktop search settings (e.g. [8, 13]). Williams et al. [25] found that in mobile search, the direct answer results often led to *good abandonment*, where the user was directly satisfied by the SERP without clicking any hyperlinks, and proposed a gesture model to predict user satisfaction for the abandoned queries. These findings emphasized the importance of modeling the heterogeneity of search results in building click models for mobile search.

2.2 Click Models for Web Search

In this section, we will first present some definitions and notations used in this paper and introduce some existing click models, along with their corresponding behavioral assumptions, in these notations. We will also introduce existing research on click models that has considered the heterogeneity of search results.

When a user submits a *query* q to the search engine in a *session* s, an SERP that consists of M ranked *search results*, $(d_1, d_2, \ldots, d_M)$, will be returned to the user. Usually, M is set as 10 because there are usually ten results on the first page. d_i denotes the search results ranked at position i. d_i can be an *organic* result or one of different types of *vertical* results. We use v_i to denote the *type* of d_i. M binary random variables $(C_1, C_2, \ldots, C_M)$ are used to indicate whether the user *click* d_i ($C_i = 1$) or *skip* d_i ($C_i = 0$). C_i can be observed in the search log. A click model is usually a probabilistic generative model of the click sequence $(C_1, C_2, \ldots, C_M)$ that models the joint distribution $P(C_1, C_2, \ldots, C_M)$.

Originally, click models were proposed to explain the position bias that users are more likely to click top-ranked results because these results are more likely to be *examined*. To model this bias caused by differences in examination likelihood at different ranks, the *Examination Hypothesis* was formulated by Richardson et al. [21] in predicting the click-through rate of ads and Craswell et al. [6] in modeling the position bias in web search. This hypothesis assumes that a user will click a search result if and only if she examined the result and was attracted by it:

$$C_i = 1 \iff E_i = 1 \land A_i = 1 \tag{1}$$

E_i and A_i are binary random variables. Unlike C_i, they are *latent* variables that can not be observed directly from search logs. $E_i = 1$ indicates the user examined d_i and otherwise $E_i = 0$. $A_i = 1$ means the search result can attract the user's click whenever she examines it. A_i is usually considered as fully determined by the relevance between query q and result d_i:

$$P(A_i = 1) = \alpha_{q,d_i} \tag{2}$$

Therefore, A_i is independent of E_i and the click probability of d_i can be computed as:

$$P(C_i = 1) = P(E_i = 1) \cdot P(A_i = 1) \tag{3}$$

A series of click models have different implementations of $P(E_i)$. For example, the *cascade model* proposed by Craswell et al. [6] assumes a user will examine the search results sequentially from top to bottom until she clicks a result. Therefore, $P(E_i = 1) = 1, \forall i \le j$, where j is the rank of last clicked results in the session. Guo et al. [9] extended the cascade model to multi-click sessions by assuming that the user will continue to examine next results after clicking a result at position i with a probability of λ_i. Dupret and Piwowarski [7] proposed User Browsing Model (UBM), which assumes that $P(E_i)$ depends on the current position i and its distance d to a previously clicked result:

$$P(E_i = 1) = \gamma_{i,d} \tag{4}$$

Chapelle and Zhang [1] used additional binary variables S_i to denote the user's *satisfaction* after clicking a result. If d_i is clicked ($C_i = 1$), S_i only depends on the query q and result d_i and is considered as an additional signal for relevance.

$$P(S_i = 1 | C_i = 0) = 0 \tag{5}$$

$$P(S_i = 1 | C_i = 1) = s_{q,d_i} \tag{6}$$

They also assumed that a user will scan the SERP linearly but they allowed the user to leave the SERP, not examining lower ranked search results, when she is satisfied by a result d_i ($S_i = 1$) or choose to abandon the query with a probability $1 - \gamma$:

$$P(E_1 = 1) = 1 \tag{7}$$

$$P(E_i = 1 | S_{i-1} = 1) = 0 \tag{8}$$

$$P(E_i = 1 | E_{i-1} = 0) = 0 \tag{9}$$

$$P(E_i = 1 | S_{i-1} = 0, E_{i-1} = 1) = \gamma \tag{10}$$

With the emergence of vertical results and federated search, some existing efforts in desktop web search tried to incorporate the influence of different vertical results into click models. Chen et al. [2] considered the *attention bias* that if d_i is a vertical result, it may have a higher examination probability $P(E_i = 1)$ and the *exploration bias* that the user may choose not to examine any organic results if she clicked a vertical with a certain probability $e(s)$ in the session s. Chuklin et al. [4] addressed this problem by assuming that a session is associated with a pre-defined intent $I(s)$. This intent and the type of result v_i will affect the examination probability $P(E_i)$ and click attractiveness $P(A_i)$. One can incorporate the influence of intents and result types into a click model that follows the examination hypothesis (Equation 1). For example, the UBM can be enhanced in the following way:

$$P(E_i = 1) = \gamma_{i,d}(I(s), v_i) \tag{11}$$

$$P(A_i = 1) = \alpha_{q,d_i}(I(s)) \tag{12}$$

The Vertical-aware Click Model (VCM) proposed by Wang et al. [24] further modeled how the presence of different types of verticals affects both the examination probabilities and examination order of the results on SERP. However, VCM can only model the influence of the *first* vertical result on the SERP, making it not suitable for the mobile search scenario where a SERP usually contains multiple vertical results.

These studies all focused on modeling how vertical results affect the examination probability $P(E_i)$ but none of them addressed the problem that some types of vertical results can satisfy users without requiring any clicks. Such skips on good results with low click necessity are rather common in mobile search, which motivates us to propose a new click model to cope with the corresponding click necessity bias and examination satisfaction bias in mobile environment.

3 MOBILE CLICK MODEL

3.1 Modeling Biases

We first formally introduce the click necessity bias and examination satisfaction bias as well as how we incorporate them into click models.

- **Click Necessity Bias**: *Some types of search results (e.g. the knowledge graph and direct answer results) have low click necessity because they can satisfy users' information needs without requiring any clicks, which will lower the click probabilities of these results.*

To model the click necessity bias, we introduce a binary variable N_i for each result d_i. $N_i = 1$ indicates that a user *needs* to click the result to get useful information and $N_i = 0$ indicates that a user can be satisfied directly by reading or interacting with the snippet on the SERP. We extend the examination hypothesis (Equation 1) as:

$$C_i = 1 \iff E_i = 1 \wedge A_i = 1 \wedge N_i = 1 \tag{13}$$

A user will click a search result if and only if: 1) she examined it; 2) it is attractive; and 3) she needs to click it to get useful information.

We further assume that N_i only depends on the type of search results v_i:

$$P(N_i = 1) = \beta_{v_i} \tag{14}$$

We acknowledge that $P(N_i = 1)$ may also be affected by other factors such as user intent and relevance between the query and result, but we choose to use this simplified assumption and leave the exploration of how to model $P(N_i = 1)$ for future work.

By incorporating the click necessity bias, we can avoid the negative feedback caused by the good skips on the results with low click necessity. However, we also need to define a positive signals, other than clicks, for these results. Therefore, we propose the examination satisfaction bias.

- **Examination Satisfaction Bias**: *A user can feel satisfied and leave the SERP after examining a search result that is both attractive and with low click necessity.*

We use a binary variable S_i^E to denote whether the user is satisfied just by *examining* result d_i (examination satisfaction), S_i^C to denote whether the user is satisfied after *clicking* it (click satisfaction). We further use S_i to denote user's *state of satisfaction* after position i. We assume that: 1) a user will stay satisfied once she encountered either an examination satisfaction event ($S_i^E = 1$) or a click satisfaction event ($S_i^C = 1$); 2) if the user is in the satisfied state $S_i = 1$, she will not examine follow-up results. Therefore, we have:

$$S_i = 1 \iff S_{i-1} = 1 \vee (S_i^E = 1 \vee S_i^C = 1) \tag{15}$$

$$P(E_i = 1 | S_{i-1} = 1) = 0 \tag{16}$$

Because $S_i = 1 \Rightarrow S_{i+1} = 1$, we have $\forall j > i, P(E_j = 1 | S_i = 1) = 0$. By adding the satisfaction state variable S_i, we allow the click/examination satisfaction event at position i (S_i^C and S_i^E) to influence all the follow-up examination events (E_j, where $j > i$).

The click satisfaction ($S_i^C = 1$) can happen when a result is clicked while the examination event ($S_i^E = 1$) can only occur when a result is examined ($E_i = 1$), attracts user's attention ($A_i = 1$), and it does not need to be clicked ($N_i = 0$). Similar to DBN, we assume that S_i^E and S_i^C are governed by the parameters that associate with the relevance between q and d_i.

$$P(S_i^C = 1 | C_i = 1) = s_{q,d_i}^C \tag{17}$$

$$P(S_i^E = 1 | E_i = 1, A_i = 1, N_i = 0) = s_{q,d_i}^E \tag{18}$$

By incorporating the examination satisfaction bias, we can give credits to the search results that have low click necessity but can provide relevant and useful information for users when there is no clicks below it. We hope that capturing this signals can help us rank the results with low click necessity more properly in mobile search.

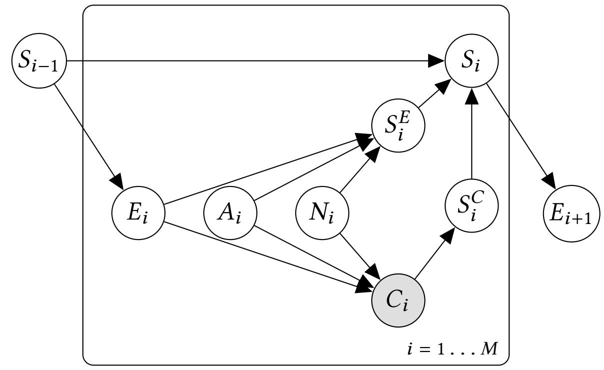


Figure 2: The Bayesian network structure of Mobile Click Model (MCM). C_i is the only observed variable.

3.2 Mobile Click Model

Besides incorporating the click necessity bias and examination satisfaction bias, we can use different functions for $P(E_i)$ and $P(A_i)$, which will be equivalent to incorporating the click necessity bias and examination bias into different click models. In this work, we use UBM's implementation of $P(A_i)$ and $P(E_i)$ (Equation 2 and 4) because: 1) it performs well in the click prediction task; 2) the computation of $P(E_i = 1)$ is fully determined by observable variables $C_j, j < i$, which simplifies the inference of posterior.

We call the derived model Mobile Click Model (MCM) and illustrate it in Figure 2. In this model, only C_i can be observed in logs and only S_i will influence users' further behavior. The conditional probabilities of C_i and the latent variables $\{E_i, A_i, N_i, S_i^E, S_i^C, S_i\}$ are given as follows:

$$P(E_i = 1 | S_{i-1} = 1) = 0 \tag{19}$$
$$P(E_i = 1 | S_{i-1} = 0) = \gamma_{i,d} \tag{20}$$
$$P(A_i = 1) = \alpha_{q,d_i} \tag{21}$$
$$P(N_i = 1) = \beta_{v_i} \tag{22}$$
$$C_i = 1 \iff E_i = 1 \wedge A_i = 1 \wedge N_i = 1 \tag{23}$$
$$P(S_i^E = 1 | \neg(E_i = 1 \wedge A_i = 1 \wedge N_i = 0)) = 0 \tag{24}$$
$$P(S_i^E = 1 | E_i = 1 \wedge A_i = 1 \wedge N_i = 0) = s^E_{q,d_i} \tag{25}$$
$$P(S_i^C = 1 | C_i = 0) = 0 \tag{26}$$
$$P(S_i^C = 1 | C_i = 1) = s^C_{q,d_i} \tag{27}$$
$$S_i = 1 \iff S_{i-1} = 1 \vee (S_i^E = 1 \vee S_i^C = 1) \tag{28}$$

The parameters of MCM are $\{\alpha, \beta, \gamma, s^E, s^C\}$. The maximum likelihood estimates of these parameters can be learned from click logs by using the Expectation-Maximization (EM) algorithm. Please refer to the Appendix for a detailed derivation of the E-step and M-step.

After learning the parameters, we can use them to compute a relevance score for each mobile search result in the logs. This score can be used to rank the search results according to users' implicit relevance feedbacks. We define the relevance score as the probability of becoming satisfied when a user examines a result d_i of type v_i:

$$\begin{aligned} score(q, d_i) &\triangleq P(S_i = 1 | E_i = 1) \\ &= \alpha_{q,d_i} [\beta_{v_i} s^C_{q,d_i} + (1 - \beta_{v_i}) s^E_{q,d_i}] \end{aligned} \tag{29}$$

Table 1: Comparisons between MCM and some existing click models. (*The *exploration bias* found by [2] assumes that after the user clicks a vertical result, she may choose not to click any organic results, which is similar to "satisfaction after click".)

	DBN[1]	UBM[7]	Chen et al.[2]	Chuklin et al.[4]	Wang et al.[24]	MCM
allow skip examination		✓	✓	✓	✓	✓
click satisfaction	✓		✓*			✓
attention bias			✓	✓	✓	
search intent bias				✓		
click necessity bias						✓
examination satisfaction						✓

3.3 Comparisons with Existing Click Models

We compare the behavioral assumptions of MCM and some existing click models in Table 1.

First, we compare MCM with two widely used click models: DBN and UBM. Compared to UBM, MCM takes the click and examination satisfaction into consideration. Therefore, in MCM, the examination probability $P(E_i = 1)$ is not only dependent on users' click behavior on previous results $(C_1, C_2, \dots, C_{i-1})$ but also influenced by the relevance of these results captured by the satisfaction parameters s^C_{q,d_j} and s^E_{q,d_j}, for all $j < i$. Compared to DBN, MCM relaxes the strict *cascade hypothesis* in examination that $E_{i-1} = 0 \Rightarrow E_i = 0$. Instead, MCM allows skips in an examination sequence as the UBM does. From this perspective, MCM can be regarded as an effort to unify these two classic click models.

We also compare MCM with previous efforts on incorporating the heterogeneity of search results into click models [2, 4, 24]. The existing studies in desktop search mainly focused on modeling the influence of heterogeneous results on users' examination behavior (attention bias) and the preference to a certain type of vertical results caused by different search intents (search intent bias). None of them addressed the click necessity bias and examination satisfaction bias that are more common in mobile search.

4 EXPERIMENTS

We conduct a series of experiments on large-scale search logs collected from a popular Chinese mobile search engine to answer the following research questions:

- **RQ1:** Does MCM have better *click prediction* ability in the mobile environment than the baseline models?
- **RQ2:** Can MCM provide better *relevance estimations* of mobile search results than the baseline models?
- **RQ3:** How does MCM model the click necessity and examination satisfaction probability of heterogeneous mobile search results?

Table 2: The statistics of two datasets used in this study.

	Dataset-C	Dataset-R
#unique queries	3,358,199	4,373
#sessions	6,613,393	1,631,756
#unique vertical_ids	2,382	612
#unique URLs	20,548,153	65,827

4.1 Experimental Setup

4.1.1 Datasets. The search logs used in this study were sampled from real mobile search logs of Sogou.com, a popular Chinese search engine. The search log for a session s consists of a query q, ten URLs of search results, a 10-dimensional binary click vector $(C_1, C_2, \ldots, C_{10})$, and ten `vertical_ids` for the search results. In this study, we use the corresponding `vertical_id` to indicate the type (v_i) of a search result (d_i). We note that this is a fine-grain categorization of search results because there are thousands of unique `vertical_ids` in the logs. A different `vertical_id` may mean that the corresponding result has a different presentation style or comes from a different source. Organic results have a set of special `vertical_ids`, so we can also use them to separate organic results from vertical results.

We use two datasets, Dataset-C for the click prediction task (Section 4.3) and Dataset-R for relevance estimation task (Section 4.4), because relevance annotations are needed for the latter task. The detailed statistics for the two datasets are shown in Table 2.

Dataset-C was generated by sampling about 5% sessions from two consecutive weekdays. In the click prediction task, we use the sessions on the first day as the training data and those on the second day as the test data. Dataset-R was generated through the following process: 1) We first randomly sampled 12,000 unique queries from a one-month search log; 2) We then sampled all the sessions associated with those queries in a single day, which may only cover a proportion of those 12,000 queries; 3) We also crawled the SERPs for 12,000 queries for further relevance annotations (See Section 4.4 for more details). In the relevance estimation task, we will use the whole Dataset-R as training set, and evaluate MCM and baseline models using the collected relevance annotations.

4.1.2 Baseline Models. We use three *basic* click models that do not take the type of search results into consideration, and two *vertical-aware* click models originated from desktop search, as the baseline models. We refer to Chuklin et al. [4] for the implementations of the baseline models and make some necessary modifications to adapt them for a fair comparison on our dataset.

The three basic click models are the following:

- UBM: User Browsing Model proposed by Dupret and Piwowarski [7] (See Equation 2–4 in Section 2.2).
- DBN: Dynamic Bayesian Network model proposed by Chapelle and Zhang [1] (Equation 5-10).
- DCM: Dependent Click Model proposed by Guo et al. [9].

Two vertical-aware baseline models are:

- EB-UBM: UBM with the exploration bias modification proposed by Chen et al. [2].
- UBM-layout: UBM that has different $\gamma_{i,d}$ parameters for search results with different layouts (i.e. different types of results). This model was designed by Chuklin et al. [4]. Note that here we can not use UBM-IA model proposed in the same paper because we do not have a pre-define classification of the search intent for each query. The original UBM-layout model only considers two types of results: *fresh* (i.e. news verticals) and *web* (i.e. organic results). We modify the model to make it work with arbitrary number of result types defined by the `vertical_ids`. This modification improves UBM-layout model's performance in click prediction and relevance estimation. Therefore, we only report the performance of the modified UBM-layout model in this paper.

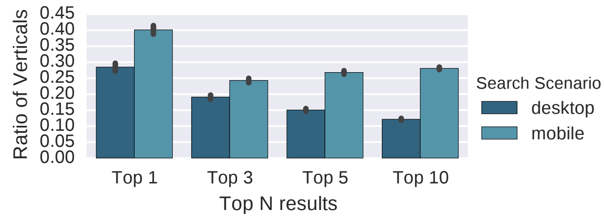


Figure 3: The distribution of vertical results on desktop and mobile.

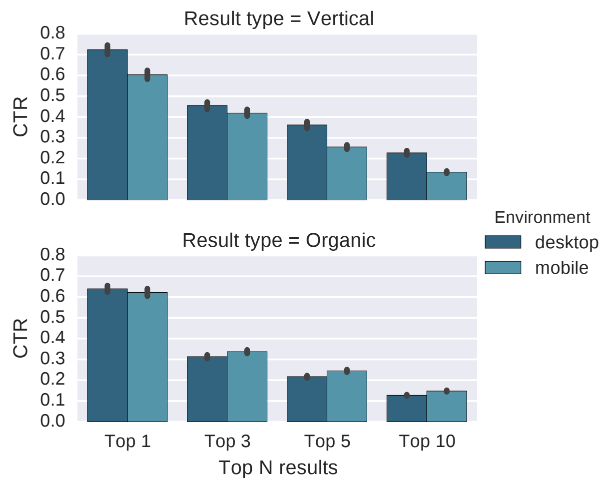


Figure 4: The Click Through Rate (CTR) of vertical and organic results on desktop and mobile.

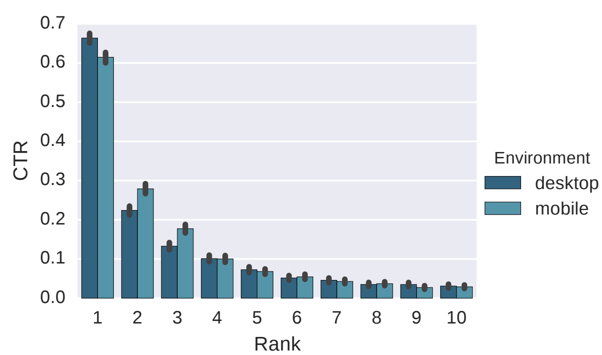


Figure 5: The Click Through Rate (CTR) of Top 10 results on desktop and mobile.

4.2 Comparison between Mobile and Desktop

Before training click models on the mobile search logs, we want to empirically demonstrate the differences between users' click behavior in mobile and desktop search. So we randomly sampled 10,000 mobile search sessions from Dataset-C and 10,000 desktop search sessions from the same commercial search engine to conduct a comparison analysis.

We first show the ratios of vertical results among the Top 1, 3, 5, 10 results in Figure 3. We can see that the ratios of vertical results in mobile search are higher than those in desktop search (all the differences are statistically significant at $p < 0.01$ level, independent t-test, two-tailed). On mobile SERPs, 28.1% of Top 10 results and over 40% of the first search results are vertical, showing a prevalence of heterogeneous results in mobile search.

The click-through rates (CTRs) of both vertical and organic results are shown in Figure 4. For organic results, the click-through rates on mobile and desktop are comparable. For Top 1 results, the click-through rates are not significantly different ($p = 0.14$). For Top 3, Top 5, and Top 10 results, the click-through rates on mobile are slightly but significantly (all $p < 0.01$) higher than those on desktop with the absolute differences of 2.3%, 2.8%, and 2.0%. However, for vertical results, the click-through rates in mobile search are significantly lower than (all $p < 0.01$) those in desktop search with relatively large margins of 12.0%, 3.5%, 10.5%, and 8.3% for Top 1, Top 3, Top 5, and Top 10 results respectively. The differences in click-through rates on vertical results in mobile and desktop search imply that a large number of vertical results on mobile SERPs are designed to directly satisfy users without being clicked, which emphasizes the importance of modeling the click necessity bias in mobile context.

We also compare the position biases of click-through rates on mobile and desktop. From Figure 5, we can see that: 1) the click-through rate for the first mobile results is lower than that for the first desktop results, which can be explained by that over 40% of the first mobile results are vertical and a large proportion of them can satisfy users without clicks. 2) the click-through rates for the second and third results on mobile are higher than those on the desktop. This finding is consistent with Lagun et al.'s finding that mobile users tend to have a longer gaze time for the second and third results [17]. It can also be explained by the *spill-over* effect [18] that a user will pay more attention to the results below a visually attractive vertical result.

4.3 Click Prediction

Following the convention of previous works [2, 7, 24], we use two evaluation metrics, log-likelihood (*LL*) and average perplexity (*AvgPerp*), to evaluate models' performance in the click prediction task. The average perplexity is the mean of perplexity $Perp_i$ over ten positions:

$$AvgPerp = \frac{1}{10}\sum_{i=1}^{10} Perp_i$$

$$Perp_i = 2^{-\frac{1}{|S|}\sum_{s\in S} C_i^s \log_2(q_i^s)+(1-C_i^s)\log_2(1-q_i^s)}$$

Here S is the set of all search sessions in test set and q_i^s is the predicted click probability of result i in session $s \in S$, given by the click model. A larger LL indicates a better performance, and the relative improvement of LL_1 over LL_2 is given by $\exp(LL_1 - LL_2) - 1$). While the perfect prediction at position i will have a perplexity $Perp_i = 1.0$, a smaller value of *AvgPerp* indicates better prediction accuracy. The relative improvement of perplexity value p_1 over p_2 is computed as $(p_2 - p_1)/(p_2 - 1)$.

When measuring the click prediction performance on the test set, we filter out all the queries that have a query frequency less than 10 in the training set. We first show the overall click prediction performance on the remaining 915,521 sessions in Table 3. From the results, we can see that MCM has better click prediction ability than both the basic and vertical-aware baselines on mobile search logs, with all the differences in *LL* and *AvgPerp* being significant at $p < 0.001$ level.

We further compare different models for queries with different query frequencies. From Table 4, we can see that while the click performance measured in *LL* increases as the query frequency for all models, MCM performs consistently better than all the baseline models. It is also interesting to see that the relative improvements of MCM over baseline models are larger for queries that have a frequency over 100 in the training set. A possible reason for this phenomenon is that there are more vertical results designed for and federated into the SERPs of hot queries.

We are also interested in the prediction performance at each ranking position. So in Figure 6, we plot the relative perplexity gains of MCM over two basic baseline models, UBM and DBN, as well as one best performing vertical-aware baseline, UBM-layout. It is worth noting that two basic click models, UBM and DBN, behave differently in mobile search. While all the models have comparable performance at position 1, MCM has larger gains over UBM for positions 2-4 and over DBN for positions 5-10. UBM performs worse at positions 2-4 because it can not adjust the examination probability accordingly when some top-ranked results already satisfy the user. DBN's performance drops as the rank increases because the skip examination behavior is more common at the lower positions, which violates DBN's assumption. MCM overcomes these disadvantages by incorporating examination/click satisfaction and allowing skip examination. Therefore, it has a consistent improvement over UBM and DBN at positions 1-9. We speculate that MCM is worse than UBM at position 10 because the irregularity of click-through rate at the last position can be easily captured by UBM, but for MCM, the estimation of examination probability at position 10 ($P(E_{10} = 1)$) is dominated by the satisfaction probability ($P(S_9 = 1)$). The vertical-aware UBM-layout model has a performance pattern similar to UBM. Because UBM-layout can capture the attention bias on examination probability, it consistently outperforms UBM across ten positions.

Regarding **RQ1**, we find that MCM has better click prediction ability than the baseline models in the mobile search environment. The improvement of MCM is consistent for queries with different frequencies and for almost all positions in the first page. These results suggest that incorporating the click necessity bias and examination satisfaction bias is effective in modeling users' click behavior in mobile search.

4.4 Relevance Estimation

To address **RQ2**, we train MCM and baseline models on Dataset-R and rank the results according to the predicted relevance score provided by each model. For example, for MCM, we use Equation 29 to compute the predicted relevance score. The ranking results

Table 3: The overall performance of click prediction measured in log-likelihood (LL) and average perplexity (AvgPerp.) (all improvements are statistically significant at $p < 0.001$ level, pairwise t-test, two-tailed, $n = 915,521$).

	LL	MCM impr.	*AvgPerp*	MCM impr.
MCM	**-1.866**	-	**1.218**	-
UBM	-1.928	+6.38%	1.226	+3.81%
DBN	-1.944	+8.09%	1.226	+4.00%
DCM	-1.976	+11.68%	1.232	+6.64%
EB-UBM	-1.935	+7.15%	1.227	+4.32%
UBM-layout	-1.902	+3.64%	1.222	+2.22%

Table 4: Log-likelihood of each model and MCM's improvement over each baseline model for different query frequency in training set (all improvements are statistically significant at $p < 0.001$ level, pairwise t-test, two-tailed, n = #Sessions).

Query Freq.	[10, 100)		[100, 1000)		[1000, inf)	
#Sessions	512,322		285,277		117,922	
Model	*LL*	Impr.	*LL*	Impr.	*LL*	Impr.
MCM	**-2.168**	-	**-1.709**	-	**-0.935**	-
UBM	-2.220	+5.31%	-1.789	+8.39%	-0.995	+6.23%
DBN	-2.260	+9.62%	-1.780	+7.40%	-0.967	+3.26%
DCM	-2.250	+8.57%	-1.839	+13.97%	-1.119	+20.17%
EB-UBM	-2.218	+5.14%	-1.802	+9.76%	-1.028	+9.78%
UBM-layout	-2.192	+2.47%	-1.762	+5.51%	-0.977	+4.26%

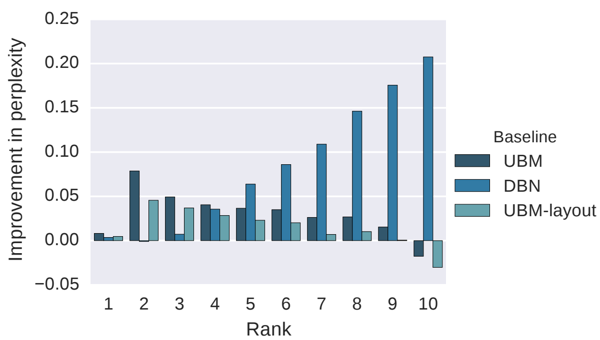


Figure 6: Perplexity gains of MCM for different ranking positions compared to UBM, DBN, and UBM-layout.

can be evaluated by standard IR evaluation metrics. In this study, we use $NDCG@3$ and $NDCG@5$ [12] as the evaluation metrics for the relevance estimation task.

To compute NDCG, we randomly sampled 775 queries from the crawled SERPs in Dataset-R and used the service provided by Baidu Zhongbao[1], a Chinese crowdsourcing platform, to collect relevance labels for the top 5 results of these queries. Because we assume that a user can be satisfied directly on the SERP, besides collecting relevance labels for the landing pages (Rel_{page}), we also collect relevance labels for the snippets ($Rel_{snippet}$) of mobile results by showing the query and a snapshot of the snippet to the crowdsourcing workers. With Rel_{page} and $Rel_{snippet}$ labels, we compute $NDCG_{page}$ and $NDCG_{snippet}$. We also compute an average of

[1]http://zhongbao.baidu.com/

Table 5: Relevance estimation performance measured in NDCG@3. (*/** indicates the difference with MCM are significant at $p < 0.005/0.001$ level, pairwise t-test, two-tailed, $n = 775$.)

Model	$NDCG_{snippet}@3$	$NDCG_{page}@3$	$NDCG_{avg}@3$
UBM	0.807^{**}	0.665^{**}	0.731^{**}
DBN	0.818^{**}	0.668^{**}	0.738^{**}
DCM	0.802^{**}	0.658^{**}	0.724^{**}
EB-UBM	0.816^{**}	0.672^{*}	0.739^{**}
UBM-layout	0.815^{**}	0.667^{**}	0.736^{**}
MCM	**0.834**	**0.692**	**0.759**

Table 6: Relevance estimation performance measured in NDCG@5. (*/** indicates the difference with MCM are significant at $p < 0.005/0.001$ level, pairwise t-test, two-tailed, $n = 775$.)

Model	$NDCG_{snippet}@5$	$NDCG_{page}@5$	$NDCG_{avg}@5$
UBM	0.874^{**}	0.771^{*}	0.824^{**}
DBN	0.880^{**}	0.770^{**}	0.825^{**}
DCM	0.870^{**}	0.764^{**}	0.818^{**}
EB-UBM	0.881^{**}	0.774^{*}	0.828^{**}
UBM-layout	0.880^{**}	0.769^{**}	0.825^{**}
MCM	**0.890**	**0.785**	**0.839**

these two labels $Rel_{avg} = (Rel_{page} + Rel_{snippet})/2$ [2], and use it to compute $NDCG_{avg}$. A 4-level scale (1: not relevant, 2: somewhat relevant, 3: fairly relevant, and 4: perfectly relevant) was used for both Rel_{page} and $Rel_{snippet}$. Each snippet and landing page were annotated by at least three workers. The weighted κ [5] for Rel_{page} and $Rel_{snippet}$ are 0.58 and 0.39, indicating an acceptable quality for relevance annotations.

Table 5 and Table 6 show the ranking performance of each model in $NDCG@3$ and $NDCG@5$ separately. From the results, we can see that the vertical-aware models are generally better than the basic models in relevance estimation, especially if we compare UBM with EB-UBM and UBM-layout. These results emphasize the importance of considering the heterogeneity of search results in the mobile context. MCM has better performance than the two vertical-aware baselines, because in mobile environment, the click necessity bias may have a stronger influence on users' click behavior than the attention bias.

To sum up, regarding **RQ2**, we show that MCM has a better performance in estimating the relevance of mobile search results than the baseline models, with all the differences being significant at $p < 0.005$ or $p < 0.001$ levels.

4.5 Parameters Learned by MCM

To see how MCM models the click necessity bias and examination satisfaction bias in mobile search (**RQ3**), we analyze the click necessity parameters β and examination satisfaction parameters s^E learned by MCM.

We first compute the mean of click necessity parameter β for all organic results and vertical results. $M(\beta_{vertical})$ on Dataset-C is 0.464 ($SD = 0.245$), which is significantly lower than $M(\beta_{organic}) = 0.654$ ($SD = 0.111$). This confirms our observation in Section 4.2 that the vertical results in mobile search are more likely to have a low

[2]We also tested the geometric and harmonic mean. The results were similar to the arithmetic mean, so we only report the results using arithmetic mean in the paper.

Query: 双色球开奖结果 (The result of "双色球", a popular lottery in China)	**Query:** wifi查看密码 (viewing wifi password)	**Query:** 焦虑症 (Anxiety disorder)
Result: A direct answer result showing the result of the latest lottery and the payment of the prizes.	Result: A vertical result that provides step-by-step guidance of how to view the wifi password on a PC.	Result: An medical knowledge graph result showing the symptom, diagnose, and treatment of Anxiety disorder on the SERP.
$\beta = 0.0001, s^E = 0.904$	$\beta = 0.0002, s^E = 0.5$	$\beta = 0.0003, s^E = 0.322$

Figure 7: Examples of search results that have the lowest click necessity according to MCM (β).

click necessity. However, the mean of examination satisfaction parameter s^E for the vertical results $M(s^E_{vertical})$ is $0.253(SD = 0.114)$, smaller than $M(s^E_{organic})$, which equals to 0.276 $(SD = 0.131)$. This result suggests that only a small proportion of vertical results can directly lead to examination satisfaction.

We further conduct a case study to inspect the relationship between the model parameters and the search results. Three types of vertical results with lowest β shown in Figure 7. For each vertical type, we select an result and the corresponding query from the logs as an example[3]. We can see that these examples all demonstrate useful information directly in the snippet. A user can get information from them without click, which is captured by the β parameter of MCM. We also show the estimation of s^E for each query-result pair. We can see that the learned s^E can reflect the examination satisfaction to some extent. Users are likely to be satisfied by the result in the first example, if they just want to know the latest lottery result. This can be reflected by a high s^E of 0.904. On the other hand, if a user wants to get sufficient information about *Anxiety Disorder*, although the medical knowledge graph result in the last example can provide a good overview, it is less likely for the user to be satisfied by this single result. Therefore, the corresponding s^E estimated by MCM is only 0.322.

We acknowledge that these examples also reveal a limitation of MCM, which is that s^E may not be a valid relevance indicator in the complex, informational tasks. Merely incorporating the examination satisfaction bias can not fully solve the problem of assigning positive feedbacks to the results that were not clicked. In future work, we can further explore the possibilities of using viewport time and gestures as additional features in estimating the relevance of the results with low click necessity.

5 CONCLUSIONS AND FUTURE WORK

Observing that in mobile search, some vertical results, such as direct answer results and knowledge graph results, have low click necessity, and therefore, will be discriminated by most existing click models, we propose a simple yet effective Mobile Click Model (MCM) to incorporate the related click necessity bias and examination satisfaction bias in mobile search. Theoretically, the proposed MCM extends the examination hypothesis and can be regarded as a unified generalization of two most widely-used click models, DBN and UBM. Empirically, extensive experiments on large-scale mobile search logs demonstrate that MCM achieves substantial performance gains over the baseline models in both the click prediction task and relevance estimation task.

In terms of future work, we note that MCM can be further extended in many ways. First, we can use richer behavioral features, such as the viewport time and gestures, to better calibrate the estimation of the behavioral latent variables E_i, S^E_i, and S^C_i. Second, while the click necessity parameters β are fully learned from click logs in this study, we can introduce external knowledge into MCM to further improve its effectiveness by defining the prior of β for each type of vertical result accordingly. Finally, as we mentioned in Section 4.5, the current definition of examination satisfaction may fail to reflect the relevance of results in complex search tasks. Instead of always attributing satisfaction to the last-clicked or last-examined result, we can explore new ways to properly measure the contribution of every search result. It is also worth noting that while this study is motivated by the difference between mobile and desktop search, the click necessity bias and examination satisfaction bias may exist in desktop search, too. We will try to adopt MCM to the desktop environment in future work.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61472206) and National Key Basic Research Program (2015CB358700). We thank Sogou.com for the anonymized mobile search log used in this work.

REFERENCES

[1] Olivier Chapelle and Ya Zhang. 2009. A dynamic bayesian network click model for web search ranking. In *WWW '09*. ACM, 1–10.

[2] Danqi Chen, Weizhu Chen, Haixun Wang, Zheng Chen, and Qiang Yang. 2012. Beyond ten blue links: enabling user click modeling in federated web search. In *WSDM '12*. ACM, 463–472.

[3] 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.

[4] Aleksandr Chuklin, Pavel Serdyukov, and Maarten De Rijke. 2013. Using Intent Information to Model User Behavior in Diversified Search.. In *ECIR*. Springer, 1–13.

[3]The snippet of result is crawled by us, which may be different from the result viewed by the user in the search log, even if they share the same URL and vertical_id.

[5] Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70, 4 (1968), 213.
[6] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *WSDM'08*. ACM, 87–94.
[7] Georges E Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations.. In *SIGIR'08*. ACM, 331–338.
[8] Laura A Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *SIGIR'04*. ACM, 478–479.
[9] Fan Guo, Chao Liu, and Yi Min Wang. 2009. Efficient multiple-click models in web search. In *WSDM'09*. ACM, 124–131.
[10] 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.
[11] Morgan Harvey and Matthew Pointon. 2017. Searching on the Go: The Effects of Fragmented Attention on Mobile Web Search Tasks. In *SIGIR'17*. ACM, 155–164.
[12] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.
[13] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR'05*. Acm, 154–161.
[14] Maryam Kamvar and Shumeet Baluja. 2006. A large scale study of wireless search behavior: Google mobile search. In *SIGCHI'06*. ACM, 701–709.
[15] Maryam Kamvar, Melanie Kellar, Rajan Patel, and Ya Xu. 2009. Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. In *WWW'09*. ACM, 801–810.
[16] 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. *Journal of the Association for Information Science and Technology* 66, 3 (2015), 526–544.
[17] 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, 113–122.
[18] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2015. Influence of vertical result in web search examination. In *SIGIR'15*. ACM, 193–202.
[19] 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, 435–444.
[20] Kevin Ong, Kalervo Järvelin, Mark Sanderson, and Falk Scholer. 2017. Using Information Scent to Understand Mobile and Desktop Web Search Behavior. In *SIGIR'17*. ACM, 295–304.
[21] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *WWW'07*. ACM, 521–530.
[22] 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, 1201–1212.
[23] Manisha Verma and Emine Yilmaz. 2016. Characterizing Relevance on Mobile and Desktop. In *ECIR'16*. Springer, 212–223.
[24] 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, 503–512.
[25] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Detecting good abandonment in mobile search. In *WWW'16*. 495–505.
[26] Wan-Ching Wu, Diane Kelly, and Avneesh Sud. 2014. Using information scent and need for cognition to understand online search behavior. In *SIGIR'14*. ACM, 557–566.
[27] Jeonghee Yi, Farzin Maghoul, and Jan Pedersen. 2008. Deciphering mobile search patterns: a study of yahoo! mobile search queries. In *WWW'18*. ACM, 257–266.

APPENDIX

In the Appendix, we will first introduce how to update the parameters of MCM $\{\alpha, \beta, \gamma, s^E, s^C\}$ in the M-step using the posterior distributions of the latent variables $\{E_i, A_i, N_i, S_i^E, S_i^C, S_i\}$. After that we will give some details about how to compute the posterior distributions using a forward-backward alogrithm.

M-step

Suppose we have a set of search sessions S and each session $s \in S$ is associated with a query q^s and M search results $(d_1^s, \ldots, d_M^s)$ with types $(v_1^s, \ldots, v_M^s)$. We denote $E^s, A^s, N^s, S^{E,s}, S^{C,s}$ the vector of latent variables in a session s. The updates of the parameters $\{\alpha, \beta, \gamma, s^E, s^C\}$ are as follows:

$$\begin{aligned}\alpha_{q,d} = \operatorname*{argmax}_{\alpha} \sum_{s \in S} \sum_{i=1}^{M} I(q^s = q, d_i^s = d) \\ [P(A_i^s = 1|C^s)\log(\alpha) + P(A_i^s = 0|C^s)\log(1-\alpha)]\end{aligned}$$

$$\begin{aligned}\beta_v = \operatorname*{argmax}_{\beta} \sum_{s \in S} \sum_{i=1}^{M} I(v_i^s = v) \\ [P(N_i^s = 1|C^s)\log(\beta) + P(N_i^s = 0|C^s)\log(1-\beta)]\end{aligned}$$

$$\begin{aligned}\gamma_{r,d} = \operatorname*{argmax}_{\gamma} \sum_{s \in S} \sum_{i=1}^{M} I(i = r, i - Pos(lastClick, i) = d) \\ [P(E_i^s = 1, S_{i-1}^s = 0|C^s)\log(\gamma) \\ + P(E_i^s = 0, S_{i-1}^s = 0|C^s)\log(1-\gamma)]\end{aligned}$$

$$\begin{aligned}s_{q,d}^C = \operatorname*{argmax}_{s^C} \sum_{s \in S} \sum_{i=1}^{M} I(q^s = q, d_i^s = d) \\ [P(S_i^{C,s} = 1, C_i^s = 1|C^s)\log(s^C) \\ + P(S_i^{C,s} = 0, C_i^s = 1|C^s)\log(1-s^C)]\end{aligned}$$

$$\begin{aligned}s_{q,d}^E = \operatorname*{argmax}_{s^E} \sum_{s \in S} \sum_{i=1}^{M} I(q^s = q, d_i^s = d) \\ [P(S_i^{E,s} = 1, E_i^s = 1, A_i^s = 1, N_i^s = 0|C^s)\log(s^E) \\ + P(S_i^{E,s} = 0, E_i^s = 1, A_i^s = 1, N_i^s = 0|C^s)\log(1-s^E)]\end{aligned}$$

E-step

Because MCM assumes that the latent variable in last step S_{i-1} may determine E_i and S_i, we need to use the forward-backward algorithm to infer the posterior distributions of the latent variables in each search session s[4]. We define the following variables:

$$f_i(x) = P(S_i = x, C_1, C_2, \ldots, C_i)$$

$$b_i(x) = P(C_{i+1}, \ldots, C_M|S_i = x)$$

These two variables can be computed recursively:

$$f_{i+1}(x) = \sum_{x' \in \{0,1\}} f_i(x')P(S_{i+1} = x, C_{i+1}|S_i = x')$$

$$b_{i-1}(x) = \sum_{x' \in \{0,1\}} b_i(x')P(S_i = x', C_i|S_{i-1} = x)$$

With $f_i(x)$ and $b_i(x)$ we can compute the posterior distributions needed in the E-step. For example, the posterior distributions needed in the update of $s_{q,d}^E$ can be calculated as follows:

$$\begin{aligned}&P(S_i^E = 1, E_i = 1, A_i = 1, N_i = 0|C_1, C_2, \ldots, C_M) \\ &= \frac{f_{i-1}(0)b_i(1)P(S_i^E = 1, E_i = 1, A_i = 1, N_i = 0, C_i|S_{i-1} = 0)}{\sum_{x \in 0,1} f_i(x)b_i(x)} \\ &= \frac{f_{i-1}(0)b_i(1)}{\sum_{x \in 0,1} f_i(x)b_i(x)} I(C_i = 0)\gamma_{i,d}\alpha_{q,d_i}(1-\beta_{v_i})s_{q,d_i}^E\end{aligned}$$

$$\begin{aligned}&P(S_i^E = 0, E_i = 1, A_i = 1, N_i = 0|C_1, C_2, \ldots, C_M) \\ &= \frac{f_{i-1}(0)b_i(0)P(S_i^E = 0, E_i = 1, A_i = 1, N_i = 0, C_i|S_{i-1} = 0)}{\sum_{x \in 0,1} f_i(x)b_i(x)} \\ &= \frac{f_{i-1}(0)b_i(0)}{\sum_{x \in 0,1} f_i(x)b_i(x)} I(C_i = 0)\gamma_{i,d}\alpha_{q,d_i}(1-\beta_{v_i})(1-s_{q,d_i}^E)\end{aligned}$$

[4] We omit the superscript s here for convenience